Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 154–174
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Enhancement of noisy speech by temporal and spectral processing

P. Krishnamoorthy [a], S.R.M. Prasanna [b,*]

[a] Samsung India Software Center, Noida 201301, India
[b] Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, Guwahati, Assam 781039, India

Received 22 February 2008; received in revised form 11 August 2010; accepted 12 August 2010

Abstract

This paper presents a noisy speech enhancement method by combining linear prediction (LP) residual weighting in the time domain and spectral processing in the frequency domain to provide better noise suppression as well as better enhancement in the speech regions. The noisy speech is initially processed by the excitation source (LP residual) based temporal processing that involves identifying and enhancing the excitation source based speech-specific features present at the gross and fine temporal levels. The gross level features are identified by estimating the following speech parameters: sum of the peaks in the discrete Fourier transform (DFT) spectrum, smoothed Hilbert envelope of the LP residual and modulation spectrum values, all from the noisy speech signal. The fine level features are identified using the knowledge of the instants of significant excitation. A weight function is derived from the gross and fine weight functions to obtain the temporally processed speech signal. The temporally processed speech is further subjected to spectral domain processing. Spectral processing involves estimation and removal of degrading components, and also identification and enhancement of speech-specific spectral components. The proposed method is evaluated using different objective and subjective quality measures. The quality measures show that the proposed combined temporal and spectral processing method provides better enhancement, compared to either temporal or spectral processing alone.

Keywords: Speech enhancement; Temporal processing; Spectral processing; Temporal and spectral processing

1. Introduction

The problem of enhancing noisy speech received considerable attention and in the literature variety of methods have been proposed. The noisy speech enhancement methods available may be broadly classified into two categories, namely, *spectral and temporal domain enhancement methods*. Generally, the spectral domain processing methods can be classified into two main areas: nonparametric and statistical model-based speech enhancement (Shao and Chang, 2007; Loizou, 2007). The nonparametric methods remove an estimate of the distortion from the noisy features, such as subtractive-type algorithms (Boll, 1979; Berouti et al., 1979; Kim et al., 2000; Kamath and Loizou, 2002; Yamashita and Shimamura, 2005; Yang and Fu, 2005; Lu, 2007) and wavelet denoising (Donoho, 1995; Chang et al., 2007; Senapati et al., 2008). The statistical-model-based speech enhancement (Ephraim and Malah, 1984, 1985; Marzinzik and Kollmeier, 2002; Martin, 2005; Chen and Loizou, 2005, 2007) utilizes a parametric model of the signal generation process. The spectral subtraction algorithm is the oldest one proposed for noise reduction (Boll, 1979). Spectral subtraction is performed by subtracting the average magnitude of the noise spectrum from the spectrum of the noisy speech to estimate the magnitude of the enhanced speech spectrum. The noise spectrum is estimated by averaging short-term magnitude spectra of the non-speech segments. One of the serious drawbacks of this method is that it produces musical noise in the enhanced speech. This noise arises because of randomly spaced peaks in the time frequency plane due to the deviation of the estimated spectrum of noise from the instantaneous noise

* Corresponding author. Tel.: +91 361 2582513; fax: +91 361 2690762.
E-mail addresses: krishna.m1@samsung.com, pkmkicha@gmail.com (P. Krishnamoorthy), prasanna@iitg.ernet.in (S.R.M. Prasanna).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.08.011

spectrum (Seok and Bae, 1999). Several modifications are proposed for the spectral subtraction approach to reduce the effect of musical noise (Loizou, 2007). One of the most popular spectral based noisy speech enhancement is the minimum mean square error (MMSE) estimation of the short time spectral amplitude (STSA) algorithm of Ephraim and Malah (1984). This algorithm is based on a Gaussian statistical model. Accordingly the coefficients of the short time Fourier transform (STFT) of speech and noise are modelled as statistically independent Gaussian random variable. The aim was to enhance degraded speech by minimizing the mean squared error between the STSA of the clean speech and the enhanced speech. This optimality gives very good results in practice, with noticeable reduction in musical noise. A number of non-Gaussian modelling (like Gamma modelling (Marzinzik and Kollmeier, 2002), Laplacian modelling (Chen and Loizou, 2007)) based Ephraim–Malah filters have also been proposed for improving the performance.

Yegnanarayana et al. proposed an enhancement method by exploiting the characteristics of excitation source signal such as linear prediction (LP) residual (Yegnanarayana et al., 1999). The basis for this approach is that human beings perceive speech by capturing features present from the high signal-to-noise ratio (SNR) regions and then extrapolating the features in the low SNR regions (Yegnanarayana et al., 1999). Accordingly, the approach for speech enhancement is to identify the high SNR regions in the noisy speech, and enhance them relative to the low SNR regions, without causing significant distortion in the enhanced speech. A weight function is derived for the residual signal that will reduce the energy in the low SNR regions relative to the high SNR regions of the noisy signal. The residual signal samples are multiplied with the weight function and the weighted LP residual is used to excite the time-varying all-pole filter derived from the noisy speech to generate the enhanced speech. In (Jin and Scordilis, 2006) a speech enhancement algorithm similar to (Yegnanarayana et al., 1999) is proposed. It differs with the former residual weighting scheme in that the weights on the LP residuals are derived based on a constrained optimization criterion. In (Yegnanarayana et al., 2002) authors exploited the use of coherently added Hilbert envelope (HE) for LP residual reconstruction. The feature that the HE has large amplitude at the instant of significant excitation makes it a good indicator of glottal closure (GC), where an excitation pulse takes place. Therefore, applying the HE to the LP residual as a weighting function has the effect of emphasizing the pulse train structure for voiced speech, which leads to an enhanced LP residual signal.

As mentioned, most of the enhancement methods process degraded speech in either temporal or spectral domains for achieving enhancement. The scope of this work is to highlight and demonstrate the merits of combined temporal and spectral processing methods for processing noisy speech. Generally in most of the spectral domain based methods more emphasis is given to suppress the noise components by estimating the noise characteristics from the degraded speech. The merit of this approach is its effectiveness for noise removal. However, information about the noise needs to be continuously estimated, particularly, in non-stationary environments where noise characteristics keep changing. Alternatively, the temporal processing methods that use the characteristics of excitation source information primarily aim at emphasizing the high SNR regions of noisy speech. Therefore no explicit knowledge of characteristics of background noise is required. The limitation of the temporal processing methods is that the level of removal of degradation achieved may not be significant as in the case of spectral based methods. Thus the integration of these two approaches may lead to better suppression of degradation and also enhancement of high SNR speech regions. This may lead to improved performance compared to either temporal processing or spectral processing alone. Further, from the speech production point of view, the temporal and spectral processing methods use independent information from the noisy speech. It will therefore be interesting to study whether they are exploiting different information for processing. If so, then they can be suitably combined to develop robust methods for the speech enhancement. Motivated by these observations, this work proposes a method for the enhancement of noisy speech by the *combined temporal and spectral processing* to provide better noise suppression and also better enhancement in the speech regions.

The various steps involved in the proposed noisy speech enhancement method are illustrated in Fig. 1. The temporal processing involves identifying and enhancing the speech-specific features present at the gross and fine temporal levels. The main objective of the gross level processing is to identify and enhance the speech components at the sound units (100–300 ms) level. In this paper, a method is proposed for detecting high SNR regions using the sum of the ten largest peaks in the discrete Fourier transform (DFT) spectrum, the smoothed HE of the LP residual, and the modulation spectrum values. The objective of the fine level processing is to identify and enhance the speech-specific features at the subsegmental (2–3 ms) level. It is based on the fact that the significant excitation of the vocal tract takes place at the instants of glottal closure and onset of events like burst, frication and aspiration. Depending on the nature of degradation, the LP residual signal will have many other random peaks in addition to the original instants of significant excitation. Temporal processing method identifies the original instants of significant excitation and emphasizes the region around them to obtain the enhanced speech. In this paper for fine level processing, a method is proposed to identify the instants of significant excitation from the noisy speech. The proposed method involves the following: (i) sinusoidal analysis of noisy speech, (ii) convolving the HE of the LP residual of the speech obtained from sinusoidal analysis by the first order Gaussian differentiator (FOGD). Finally, the gross and fine level features are combined to derive a weight

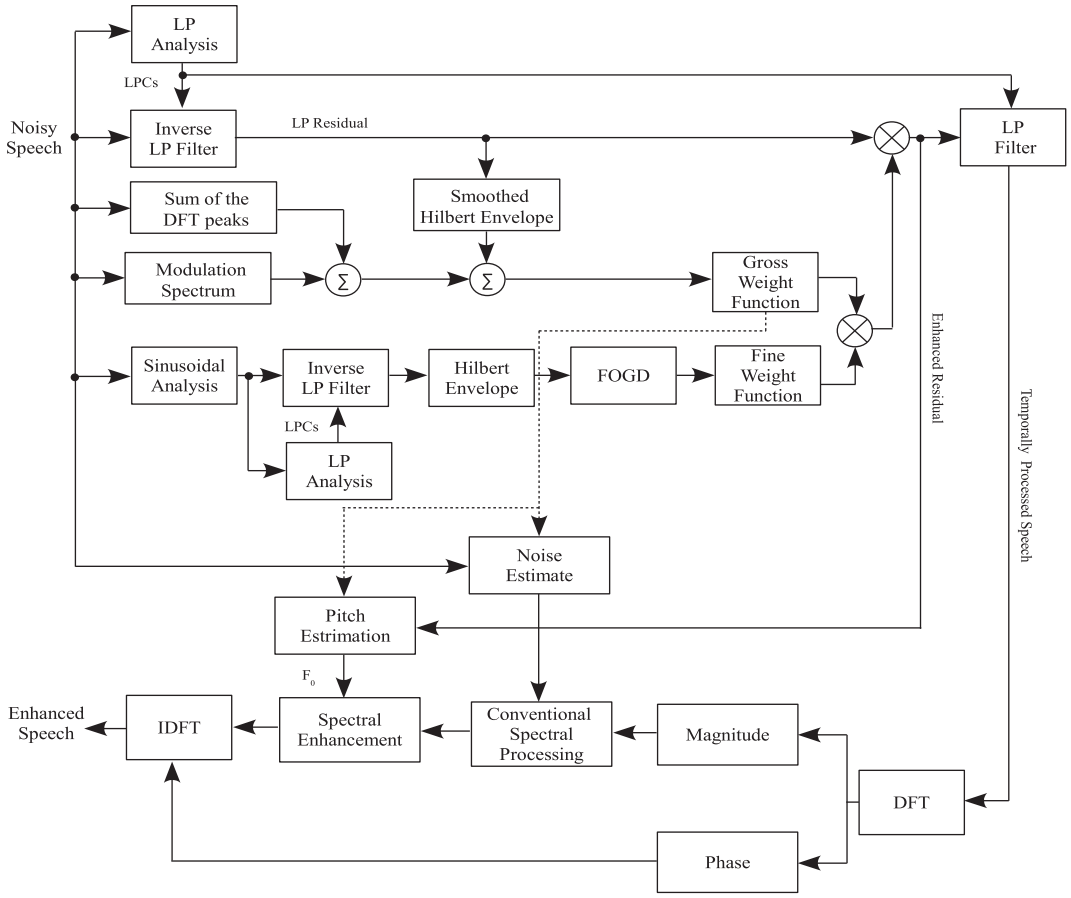

Fig. 1. Block diagram of the proposed noisy speech enhancement method.

function for the excitation source signal which emphasizes the excitation around the instants of significant excitation and deemphasizes the random peaks of background noise. The enhanced excitation signal is used to excite the time-varying all-pole filter derived from the noisy speech to generate the temporally processed speech. The temporally processed speech signal is further subjected to spectral processing. The spectral processing is based on the fact that the spectral values of the degraded speech will have both speech and degrading components. The spectral components of degradation are therefore estimated and removed. Further, there are spectral peaks that are perceptually important that are identified and enhanced. Accordingly in this work spectral processing is performed in two stages: attenuation of spectral characteristics of background noise and enhancement of speech-specific spectral features. In the first stage, the spectral characteristics of the background noise is estimated and attenuated using conventional spectral processing methods based on spectral subtraction or MMSE estimators. In the second stage, the region around pitch and harmonics are enhanced by estimating pitch from the enhanced excitation source signal.

The rest of the paper is organized as follows: Section 2 discusses about the temporal processing of noisy speech signal. The spectral domain processing of noisy speech signal is described in Section 3. Various experimental studies and objective quality measures performed on the individual and the combined processing methods are described in Section 4. Summary and conclusions of this study with scope for future work are discussed in Section 5.

2. Temporal processing of noisy speech

2.1. Gross level temporal processing

The high SNR regions at gross level are identified by using the sum of ten largest peaks in the DFT spectrum, smoothed HE of the LP residual and modulation spectrum values of the noisy speech signal (Krishnamoorthy and Prasanna, 2009). The gross weight function used in this work is similar to voice activity detection (VAD). One can use the existing VAD methods in place of gross level weight function. On the other hand, we would like to emphasize that in this work we have investigated alternate speech-specific features for detection of high SNR regions by exploiting different aspects of speech production. For instance, from speech processing point of view, the peaks in the DFT spectrum predominantly represent the vocal tract information. The LP residual predominantly contains information about the excitation source. The above two features are computed by performing short time analysis (segmental analysis) on the speech signal. The modulation

spectrum represents the long-term (suprasegmental) information of speech production. Since the origin of these three parameters is different, combining them may improve robustness and detection accuracy. Hence we have chosen these parameters in this work. Further, the proposed method relies mainly on the characteristics of speech production process, rather than the characteristics of noise. It does not assume any noise characteristics, and does not depend on parameters estimated from the noise spectra and thereof. Hence, the proposed method can be applied irrespective of the noise level.

It is also demonstrated that the excitation source based method perform better than the spectrum based method under babble noise environment. The spectrum based method performs poorly in babble noise environment because of its speech-like spectral properties (Sri Rama Murty et al., 2007). But, the excitation source information and the periodicity of the GC instants are not preserved in the babble noise. On the other hand, the excitation source based method gives relatively less performance than the spectrum based methods under white noise environment. This poor performance is due to the limitations of LP analysis under high degradation due to white noise (Sri Rama Murty et al., 2007). The modulation spectrum essentially represents the syllable rate. Therefore it is independent of noise environment. However, the weak voiced regions may not be correctly identified using modulation spectrum alone due to the use of long-term window for its computation. Therefore combination of these three features improves the identification accuracy for all the noise environments. The speech-specific parameters used in the present work are computed as described below:

(i) Sum of peaks in the DFT spectrum: A voiced speech signal is produced by passing a quasi-periodic excitation signal through a linear time-varying vocal tract system. The quasi-periodic excitation results in a periodic spectrum with peaks at multiples of the pitch frequency. The vocal tract system modulates the excitation source by formant frequencies, which depends on the sound unit being generated. Because of the damped sinusoidal nature of the resonance, the formant frequency appears as a broad resonant peak in the frequency domain (Yegnanarayana and Sri Rama Murty, 2009). As a result the DFT magnitude spectrum of a voiced frame has the same harmonic structure as the excitation source spectrum, but the amplitudes of the harmonics have been shaped according to the frequency response of the vocal tract. The resultant DFT spectrum will have the peaks at pitch and harmonics location and also stronger peaks at formant locations. Hence the sum of amplitudes of the major peak locations will be higher in high SNR regions than low SNR regions (Krishnamoorthy and Prasanna, 2008). This property is exploited in the identification of high SNR regions of the noisy speech. Mathematically, it is expressed as

$$s_d(l) = \sum_{m=1}^{10} |Y(k_m, l)| \tag{1}$$

where l is the frame index, k_m represents the frequency indices of the largest ten spectral peaks and $Y(k,l)$ represents the DFT of noisy speech frame and is computed as

$$Y(k,l) = \sum_{n=0}^{N-1} y(n) w(n - lR) e^{-\frac{j2\pi nk}{N}} \tag{2}$$

where $w(n)$ is a Hamming window, N is the number of points used for computing the DFT and R is the frame shift in samples.

For illustration, the speech data spoken by a female speaker sampled at 8 kHz with 16 bits/sample resolution is taken and white Gaussian noise is added such that the overall SNR of the signal is 3 dB and shown in Fig. 2(a). The sum of the ten largest peaks of the DFT spectrum of the Hamming windowed signal is calculated using a window of 20 ms duration and 10 ms overlap between the frames. The sum of peaks in the DFT spectrum computed for every frame is repeated 80 times (corresponding to frame shift of 10 ms at $F_s = 8$ kHz) to make the indicator length equal to that of the speech signal and plotted in Fig. 2(b).

(ii) Smoothed Hilbert envelope of the LP residual: Speech is produced as a result of excitation of time-varying vocal tract system using time-varying excitation. The instants of significant excitation for speech production correspond to instants of glottal closure (GC) or epochs during voiced speech and onset of events like burst and frication during unvoiced speech. The instants of significant excitation will be quasi-periodic in nature during voiced speech and random in nature during unvoiced speech. Further, the amplitude/energy/strength associated with these instants will be locally high. These are manifested as large errors in the LP residual of the speech (Ananthapadmanabha and Yegnanarayana, 1979). However, locating these instants directly is difficult due to the bipolar nature of the LP residual (Ananthapadmanabha and Yegnanarayana, 1979). This limitation is overcome by computing the Hilbert envelope of the LP residual (Ananthapadmanabha and Yegnanarayana, 1979). The Hilbert envelope of the LP residual $e(n)$ is defined as (Marple, 1999)

$$h_e(n) = \sqrt{e^2(n) + e_h^2(n)} \tag{3}$$

where, $e_h(n)$ is the Hilbert transform of $e(n)$, and is given by:

$$e_h(n) = IDFT[E_h(k)] \tag{4}$$

where,

$$E_h(k) = \begin{cases} -jE(k), & k = 0, 1, \ldots, \left(\frac{N}{2}\right) - 1 \\ jE(k), & k = \left(\frac{N}{2}\right), \left(\frac{N}{2}\right) + 1, \ldots, (N-1) \end{cases}$$

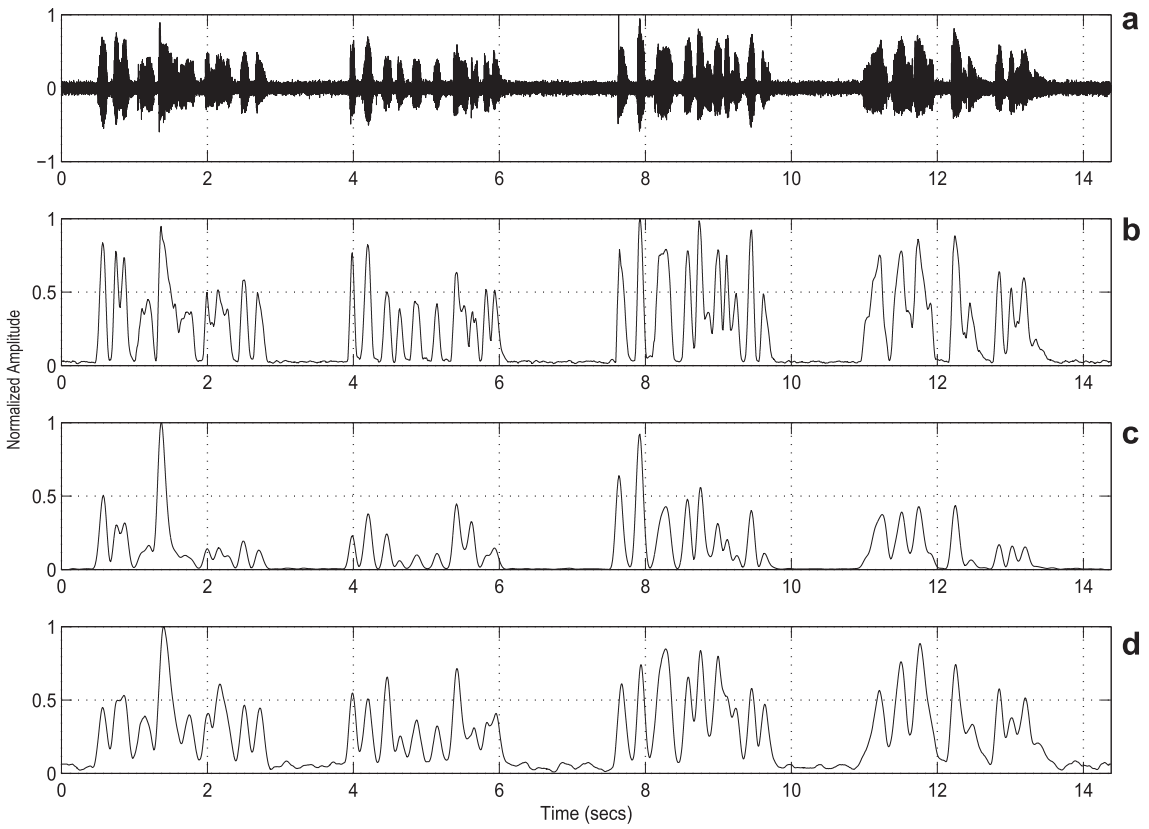

Fig. 2. Gross level features: (a) noisy speech (SNR = 3 dB), (b) sum of the peaks in the DFT spectrum, (c) smoothed HE of the LP residual and (d) modulation spectrum.

where, IDFT denotes the inverse DFT and $E(k)$ is computed as DFT of $e(n)$ and N is the number of points used for computing the DFT.The Hilbert envelope is a unipolar function and shows large amplitudes around the instants where the residual error is large, that is, in high SNR regions compared to the low SNR regions. These values are smoothed by using a mean smoothing filter of 50 ms duration (corresponds to 400 points for speech signal sampled at 8 kHz) to smooth out the smaller variations in the signal. Fig. 2(c) shows the smoothed Hilbert envelope of the LP residual signal shown in Fig. 2(a). The residual signal is derived by inverse filtering of the speech signal, and the inverse filter is obtained using LP analysis (Makhoul, 1975). The LP analysis is performed using 10th order, with a frame size of 20 ms and frame shift of 10 ms.

(iii) Modulation spectrum: The modulation spectrum of speech displays the signal in terms of the distribution of slow modulations across the frequency (Greenberg and Kingsbury, 1997). This representation captures many important properties of the auditory representation of speech. The modulation spectrum represents modulation frequencies in the speech signal between 0 and 8 Hz, with a peak sensitivity at 4 Hz, corresponding closely to the long-term modulation spectrum of speech (Greenberg and Kingsbury, 1997). In case of speech degraded by additive background noise, the speech regions will have strong 4 Hz components compared to non-speech regions.For computing modulation spectrum, the speech signal is analyzed into approximately 18 critical band filters. The filters are trapezoidal in shape, and there is minimal overlap between adjacent bands. In each band, an amplitude envelope signal is computed by half wave rectification and low pass filtering with cutoff frequency of 28 Hz. Each amplitude envelope signal is then downsampled by a factor of 100 and then normalized by the average envelope level in that channel measured over the entire utterance. The modulations of the normalized envelope signals are analyzed by computing the DFT over a 250 ms Hamming window with a frame shift of 12.5 ms in order to capture the dynamic properties of the signal. Finally, the squared magnitudes of the 4 Hz coefficients of the DFTs for each band are taken and all these values are summed and normalized with respect to its maximum value and repeated for frame shift number of times and is shown in Fig. 2(d).

(iv) Gross level weight function: The three parameters described above are due to different aspects of speech production. They exploit different information to provide evidence for the presence of high SNR regions. For instance, the sum of peaks in the DFT

spectrum exploits peaks in the DFT spectrum predominantly due to formants (because only largest 10 peaks) to indicate the presence of high SNR regions. Alternatively, the smoothed Hilbert envelope of the LP residual exploits the strength/energy associated with the instants of significant excitation to indicate the presence of high SNR regions. Therefore all these three parameters may be effectively combined in order to obtain gross weight function, which is robust and also identifies the high SNR regions better compared to the individual parameters.Each of the indicators computed above has different information about high SNR regions as is illustrated by their different shape (Fig. 2(b)–(d)). They may be combined to get robust evidence. The direct combination may not be very effective due to significant variation in their individual values. In the proposed method, the indicators of the high SNR regions are first enhanced and then combined to identify the gross level features. This is achieved with the help of the first order difference (FOD) of the indicators obtained (Krishnamoorthy and Prasanna, 2009). The steps involved in the enhancement of high SNR indicators are explained for the sum of peaks in the DFT spectrum with the help of Fig. 3. Since FOD represents the slope, the positive to negative going zero transition in FOD locates the peaks in the sum of DFT spectrum values. Fig. 3(a) shows first 6 s duration of the speech signal shown in Fig. 2(a). The sum of the DFT spectrum values and its FOD values are shown in Fig. 3(b) and (c), respectively. The positive to negative going zero transition points and the corresponding local peaks are represented by a star (*) symbol in Fig. 3(b) and (c). The unwanted zero crossings that are detected at the low SNR regions are eliminated by finding the sum of absolute FOD values for a duration of 5 ms on either side with reference to each positive to negative going zero crossing point and are given in Fig. 3(d). The peaks with the lower FOD values are eliminated by setting the threshold at 0.5 times the mean value of the FOD. In the next step, if two successive peaks occur within 50 ms, then the peak with lower FOD value is eliminated based on the assumption that it is unlikely that two high SNR regions occur within a 50 ms interval. The star (*) symbols in Fig. 3(e) show the peak locations after eliminating the undesirable peaks. With respect to each of these local peaks the nearest negative to positive going zero transition points on either side are identified and are marked by circles in Fig. 3(e). The regions between the circles are enhanced by taking the normalized value of that particular region and

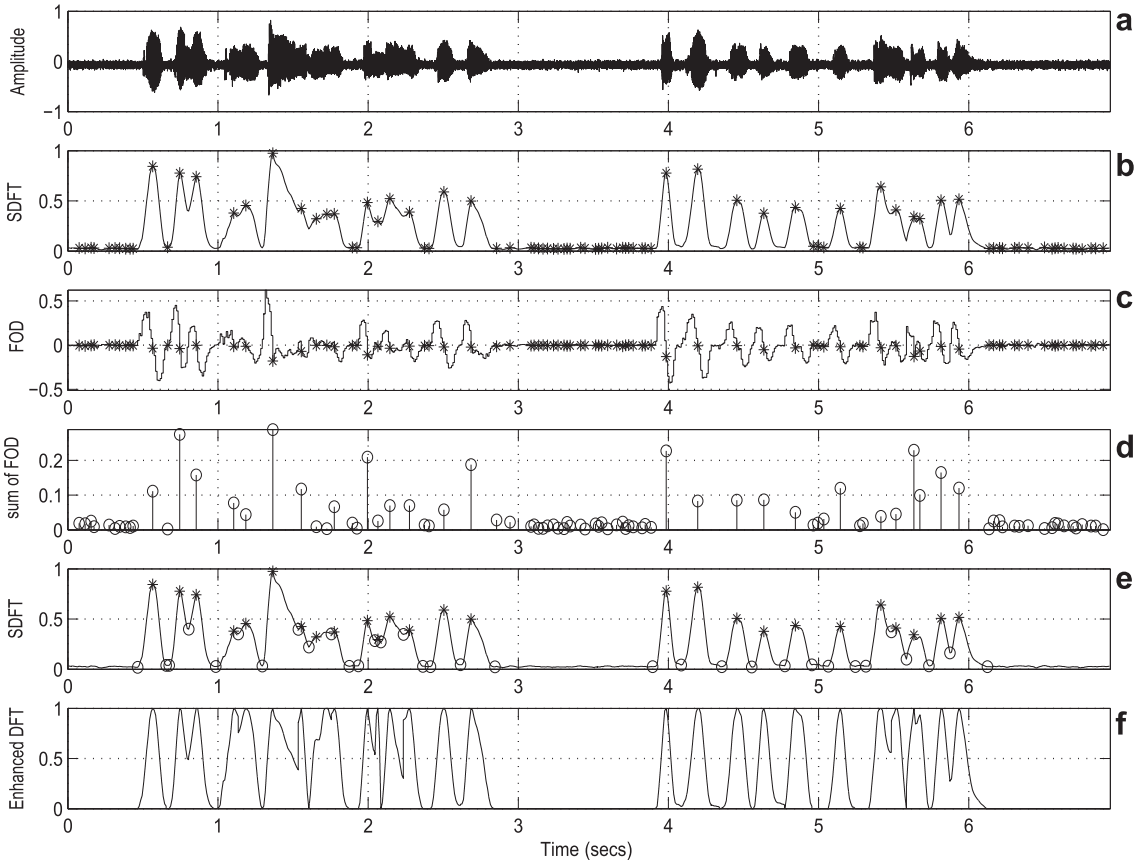

Fig. 3. High SNR regions enhancement: (a) noisy speech, (b) sum of peaks in the DFT spectrum, (c) first order difference (FOD) values, (d) slope values computed at each peak location, (e) sum of peaks in the DFT spectrum and high SNR region locations and (f) enhanced values.

is shown in Fig. 3(f). The same procedure is repeated for both the smoothed HE and the modulation spectrum.Finally, to derive a gross weight function, the enhanced values of all the three evidences are summed and normalized with respect to the maximum value of the sum. The normalized sum values are further processed using a sigmoid non-linear function given by Krishnamoorthy and Prasanna (2009)

$$w_g(n) = \frac{1}{1 + e^{-\lambda(s_i(n) - T)}} \tag{5}$$

where λ is the slope parameter set at 20 and $w_g(n)$ is the nonlinearly mapped values of normalized sum $s_i(n)$ and T is the average value of $s_i(n)$. The $w_g(n)$ is termed as the gross weight function. Note that the value of λ is not very critical, as long as it is in a range which gives desired emphasis and deemphasis. Fig. 4(b)–(d) show the sum of peaks in the DFT spectrum, smoothed HE, modulation spectrum values of the speech signal plotted in Fig. 4(a) (same as Fig. 3(a)) and the corresponding enhanced high SNR indicator plots are shown in Fig. 4(f) and (g), respectively. The normalized sum and the nonlinearly mapped values are given in Fig. 4(h) and (i), respectively.

The performance of the gross weight function detection algorithm is evaluated using the manually marked high SNR and low SNR regions. The manual marking is done independently by three different persons and the final decision about high SNR and low SNR regions is taken based on the majority logic. For comparison, the gross weight function was computed for each of the parameters independently and also for combined one. Table 1 entries show the percentage of correct detection accuracy (P_c) and false alarm probabilities (P_f) for various noise levels and different noises taken from Noisex-92 database (Varga and Steeneken, 1993). Here the value of P_c is computed as a ratio of correctly detected low and high SNR frames to the total number of frames. P_f is the probability that low SNR regions are detected when high SNR regions are present and is computed as the ratio of number of incorrectly classified low SNR frames to the total number of frames. The last column entries of the table indicate the performance of the gross weight function algorithm for the combined one. The proposed method is also evaluated for the speech data recorded in the real noisy environment. The last row entries

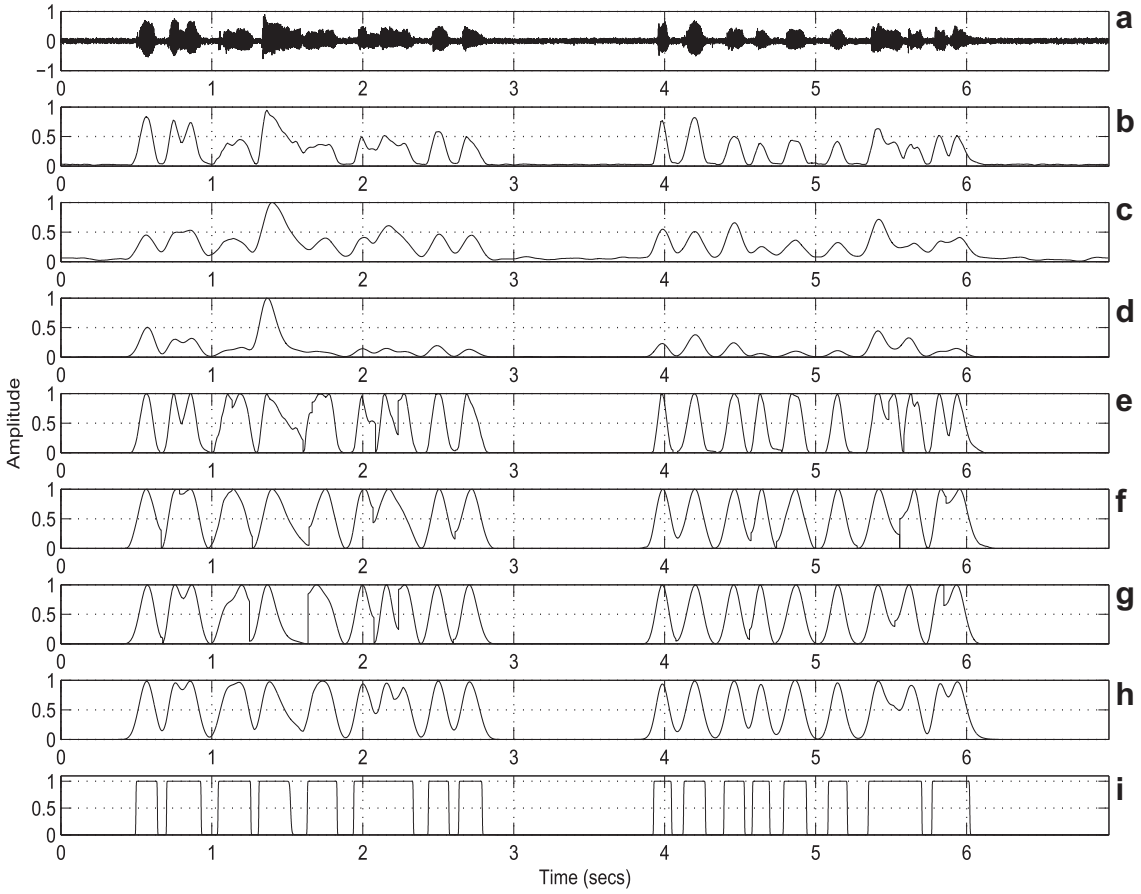

Fig. 4. Gross level identification: (a) noisy speech, (b) sum of peaks in the DFT spectrum, (c) smoothed Hilbert envelope of LP residual, (d) modulation spectrum, (e) enhanced DFT spectrum values, (f) enhanced smoothed Hilbert envelope values, (g) enhanced modulation spectrum values, (h) normalized sum and (i) nonlinearly mapped values.

Table 1
Gross weight function performance. In the table abbreviations DFT, SHE, MS and ALL refers to sum of peaks in the DFT spectrum, smoothed Hilbert envelope of the LP residual, modulation spectrum and combination of all three parameters, respectively.

Noise type	SNR level (dB)	DFT		SHE		MS		ALL	
		P_c	P_f	P_c	P_f	P_c	P_f	P_c	P_f
White Gaussian noise	0	88.09	6.60	77.16	19.39	82.45	9.37	88.58	6.33
	3	87.12	6.33	80.64	16.09	82.59	9.23	88.44	6.73
	6	87.88	3.83	81.55	13.72	81.48	9.76	87.81	3.69
	9	88.65	3.56	83.01	12.66	81.96	9.37	89.07	3.83
Babble noise	0	74.44	22.16	82.45	11.48	82.87	8.05	85.86	8.05
	3	81.55	11.35	82.24	11.21	82.59	8.58	86.35	8.31
	6	85.79	6.60	83.36	9.76	82.38	8.71	86.91	8.05
	9	86.49	6.92	83.22	9.76	82.17	8.97	87.26	6.46
Factory noise	0	81.82	12.14	82.87	11.08	82.66	8.31	87.33	6.99
	3	84.61	9.37	83.50	9.76	82.52	8.44	87.40	7.39
	6	86.91	8.05	82.87	9.76	82.24	8.84	87.53	7.65
	9	87.40	6.33	82.66	9.89	82.17	9.10	87.26	5.67
Pink noise	0	78.34	16.09	79.25	13.06	84.26	7.12	86.63	7.12
	3	83.98	9.37	83.22	9.89	84.26	7.39	87.95	8.39
	6	86.56	7.65	83.36	8.84	83.50	8.18	87.95	7.86
	9	87.60	6.60	83.29	9.10	82.45	8.58	87.40	6.91
Real data		84.47	7.79	84.96	11.53	83.77	9.85	87.49	7.39

in the table show the performance of the gross weight function algorithm for the real noisy speech data. As mentioned earlier, the following observations can be made from the table: (i) smoothed HE performs well under Babble noise, (ii) sum of the DFT spectral peaks performs well under white and pink noise, and (iii) modulation spectrum gives consistent performance under all noise environments. The combined method consistently yields better performance than the individual parameters. It can also be seen from the results that the proposed method shows a maximum of only $\pm 2\%$ deviation across all the SNR levels and hence it can be concluded that the performance of the proposed high SNR identification method is independent of noise level.

2.2. Fine level temporal processing

The basis for the fine level temporal enhancement is that the voiced speech is produced as a result of excitation of quasi-periodic glottal pulses and unvoiced speech is produced as a result of excitation of onset of events like burst, frication and aspiration. The significant excitation in each glottal cycle takes place at the instant of glottal closure (Ananthapadmanabha and Yegnanarayana, 1979). The relative spacing between the GC events is not affected by degradations. Therefore by locating the instants of significant excitation, it is possible to enhance speech around the instants relative to other regions. A weight function is derived for the LP residual from the instants of significant excitation to enhance the excitation source information around these instants relative to other regions. The identification of instants of significant excitation directly from the noisy speech is difficult due to the presence of noise components. If the degraded speech HE envelope is directly used, depending on the magnitude of the peak value, spurious peaks due to the noise components also get detected as the instants of significant excitation. Therefore first the sinusoidal analysis is performed on the noisy speech signal to eliminate most of the noise components.

In the sinusoidal speech model, the excitation signal is represented as the sum of a finite number of sinusoidal parameters (McAulay and Quatieri, 1986). The sine wave parameters are estimated by applying short-term Fourier transform (STFT) to a quasi-stationary part of the speech signal. The STFT of speech will have peaks occurring at all pitch harmonics and formants. Therefore the frequencies of underlying sine waves correspond to the peaks of STFT. The amplitudes and phases are estimated at peaks from the high resolution STFT using a simple peak picking algorithm. If the amplitudes, frequencies, and phases that are estimated for the kth segment are denoted by A_l^k, ω_l^k, and θ_l^k, respectively, the synthetic speech signal $\tilde{s}^k(n)$ can be represented as (McAulay and Quatieri, 1986)

$$\tilde{s}^k(n) = \sum_{l=1}^{L^k} A_l^k \cos\left(\omega_l^k n + \theta_l^k\right) \tag{6}$$

where L^k is the number of sinusoidal components in the frame. In our implementation only eight largest peaks ($L^k = 8$) are considered for synthesizing the speech, so that most of the noise components get eliminated.

An experiment is conducted on ten different male and female speakers to determine the deviation in the degraded speech spectral peak locations with reference to their clean speech spectral peak locations by considering different number of sinusoidal components like 4, 8, 16 and 32. The result of the analysis is given in Table 2. The percent-

Table 2
Percentage of noisy speech peak locations detected at the same locations of clean speech for different number of peaks per frame.

SNR level (dB)	No of peaks/frame White noise				No of peaks/frame Babble noise			
	4	8	16	32	4	8	16	32
0	91.25	78.59	63.28	50.43	93.75	88.91	88.98	80.74
3	93.75	82.97	67.81	53.95	96.25	91.25	90.70	83.24
6	95.00	86.41	72.42	58.48	96.56	93.13	92.34	84.88
9	95.31	89.53	77.11	63.48	97.50	94.53	93.44	86.80
	Factory noise				Pink noise			
0	94.06	90.78	88.98	80.39	90.63	82.97	69.53	57.19
3	95.63	93.44	91.17	83.36	94.69	87.03	75.00	62.58
6	96.88	94.69	92.97	85.35	97.19	89.69	79.06	67.19
9	97.19	95.63	94.53	86.76	97.19	92.03	84.45	71.99

age values show the ratio of total number of noisy speech sinusoidal components (i.e., peak locations) detected at the same locations (allowing the frequency deviation of ± 10 Hz) of clean speech to the number of sinusoidal components. These values are determined by considering only high SNR regions, since the fine weight function is applied only for the high SNR regions of noisy speech. It can be observed that if we consider the lesser number of sinusoidal components per frame, most of the peak locations of degraded speech are not affected with reference to clean speech. Since these peaks mainly represent formants, pitch and its harmonics which has high energy, the effect of noise on these locations will be less. Alternatively, if we consider larger number of sinusoidal components, then more spurious peaks detected from the noisy speech spectra. Therefore only eight peaks are chosen in this study. Note that even though largest four peaks gives best performance, we have chosen eight peaks mainly because the LP residual obtained from four sinusoidal components may not contain evidences about the instants of significant excitation. The reason is that, in majority of the cases the largest four peaks represent first two formants location.

In the next step, to determine the approximate locations of the instants of significant excitation, LP analysis is performed on the speech signal synthesized from the eight sinusoidal components. Then the HE of the LP residual is computed and mean smoothed using 1 ms rectangular window to smooth out the smaller variations. The peaks in the large error regions, representing the instants of significant excitation are detected using the first order Gaussian differentiator (FOGD) (Prasanna and Subramanian, 2005). Because of the anti-symmetric nature of the Gaussian differentiator, it gives a zero-crossing around the peaks in the HE of the LP residual. In discrete time case, the FOGD is defined as (Prasanna and Subramanian, 2005)

$$g_d(n) = \frac{1}{\sigma\sqrt{2\pi}}\left[e^{-\frac{(n+1)^2}{2\sigma^2}} - e^{-\frac{n^2}{2\sigma^2}}\right], \quad 1 \leqslant n \leqslant L_g \tag{7}$$

where L_g is length of Gaussian window and σ is standard deviation. FOGD is obtained from a Gaussian window of length $L_g = 80$ samples and $\sigma = 8$.

The negative of FOGD is convolved with the mean smoothed HE of the LP residual. The zero crossings accompanied by negative to positive transition are detected as the candidates for the instants of significant excitation (Prasanna and Subramanian, 2005). It is experimentally verified that, to detect the impulse train of duration T_d using the FOGD, the value of T_d is $\geqslant 2.35\sigma$. As the pitch period of adult speakers lies between 2.5 and 20 ms (Schroeder, 1970), so the standard deviation of Gaussian window is selected as 8 to detect the events with the minimum interval of 2.5 ms. A fine weight function is derived to enhance the region around the instants of significant excitation by convolving them with the Hamming window of 3 ms duration. In the region around the instant of significant excitation the strength of excitation is large. Hence a small region around it may contain significant information. Therefore one can select a small region around the instant of significant excitation. The width of the region should be of the order of closed phase region, which is less than the period of one glottal cycle (Smits and Yegnanarayana, 1995). We have chosen a width of 3 ms for the experimental studies reported in this paper. We have noticed that any choice of width in the region 2–5 ms does not seem to affect the performance of enhanced speech significantly. Mathematically, the fine weight function $w_f(n)$ is expressed as

$$w_f(n) = \left(\sum_{i=1}^{N_i} \delta(n - a_i)\right) * h_w(n) \tag{8}$$

where N_i represents total number of detected instants, a_i is the approximate location of instants and $h_w(n)$ is the Hamming window of 3 ms duration.

For illustration, a segment of LP residual, its mean smoothed HE, convolved output with the negative FOGD, the detected instants of significant excitation locations and the fine weight function are shown in Fig. 5(a)–(e), respectively.

The final weight function for the noisy speech LP residual is derived by multiplying the gross and fine weight functions. The minimum value of the weight function at the

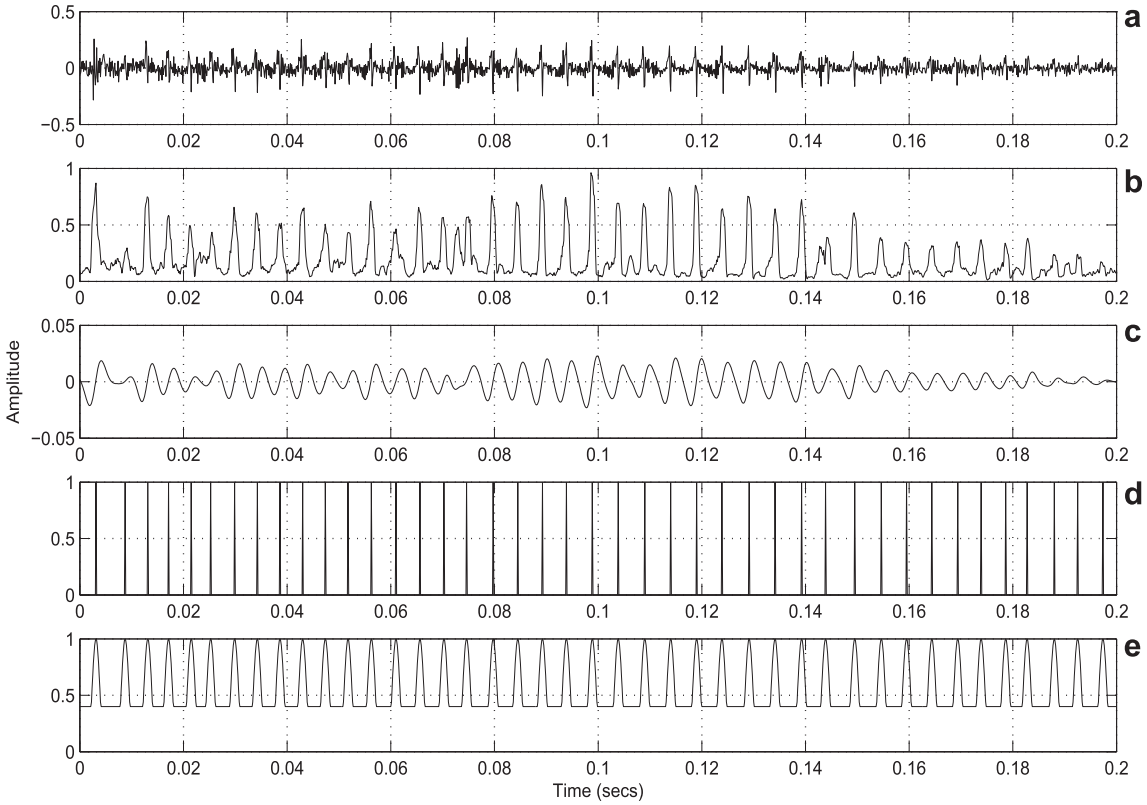

Fig. 5. Determination of fine weight function: (a) LP residual of speech signal synthesized using eight sinusoidal components, (b) mean smoothed HE, (c) convolved output of mean smoothed HE with negative of FOGD operator, (d) instants of significant excitation and (e) fine weight function.

high SNR region is kept as 0.4 to reduce the perceptual distortion. For a segment of degraded speech the nature of the gross, fine and final weight functions are illustrated in Fig. 6. For reference, a segment of clean speech, degraded speech, speech signal synthesized from the eight sinusoidal components, its LP residual and the mean smoothed HE are also shown (see Fig. 6). The noisy speech residual signal samples are then multiplied with the final weight function. The residual samples are weighted rather than the speech samples mainly because the residual samples are relatively less correlated and hence weighting may lead to less perceptual distortion (Yegnanarayana et al., 1999). The weighted residual signal is used to excite the time-varying all-pole filter derived from the noisy speech to generate the enhanced speech which is termed as temporally processed speech. The enhancement in the instants of significant excitation is illustrated in Fig. 7. Fig. 7(a)–(c) respectively show the LP residuals of clean, noisy speech and the enhanced LP residual signal obtained by multiplying the noisy speech residual signal using a weight function and it shows the enhancement in GC locations compared to the degraded speech residual signal.

The performance of the fine weight function detection algorithm is evaluated by computing the deviation in the approximate instants location of the proposed method with respect to their clean speech instants location. First the approximate instants location of the clean speech signal are computed using the FOGD as described earlier. Then the percentage of accuracy (P_a) in determining the instant of significant excitation is found as

$$P_a = \frac{N_{tc}}{N_{ti}} \times 100 \tag{9}$$

where N_{tc} represents the total number of instants detected at the same locations (or within the specified time resolution) of clean speech instants and N_{ti} is the total number of clean speech instants. The results of the analysis is given in Table 3. The table values show the percentage of approximate instants and their deviation with respect to the clean speech instants location. It can be observed that most of the detected instants lie within the 2 ms interval with reference to clean speech instants. As already mentioned, from speech enhancement perspective an approximate location of instants is sufficient. This is because the enhancement is normally performed by emphasizing the residual signal in the regions around the instants of significant excitation.

3. Spectral processing of noisy speech

The temporally processed speech sounds to be perceptually enhanced. This is mainly due to the enhancement of speech-specific features in the noisy speech signal. This includes high SNR regions at gross level and regions around the instants of significant excitation. This is achieved by multiplying the LP residual of the noisy speech signal by the weight function. Even though the speech-specific

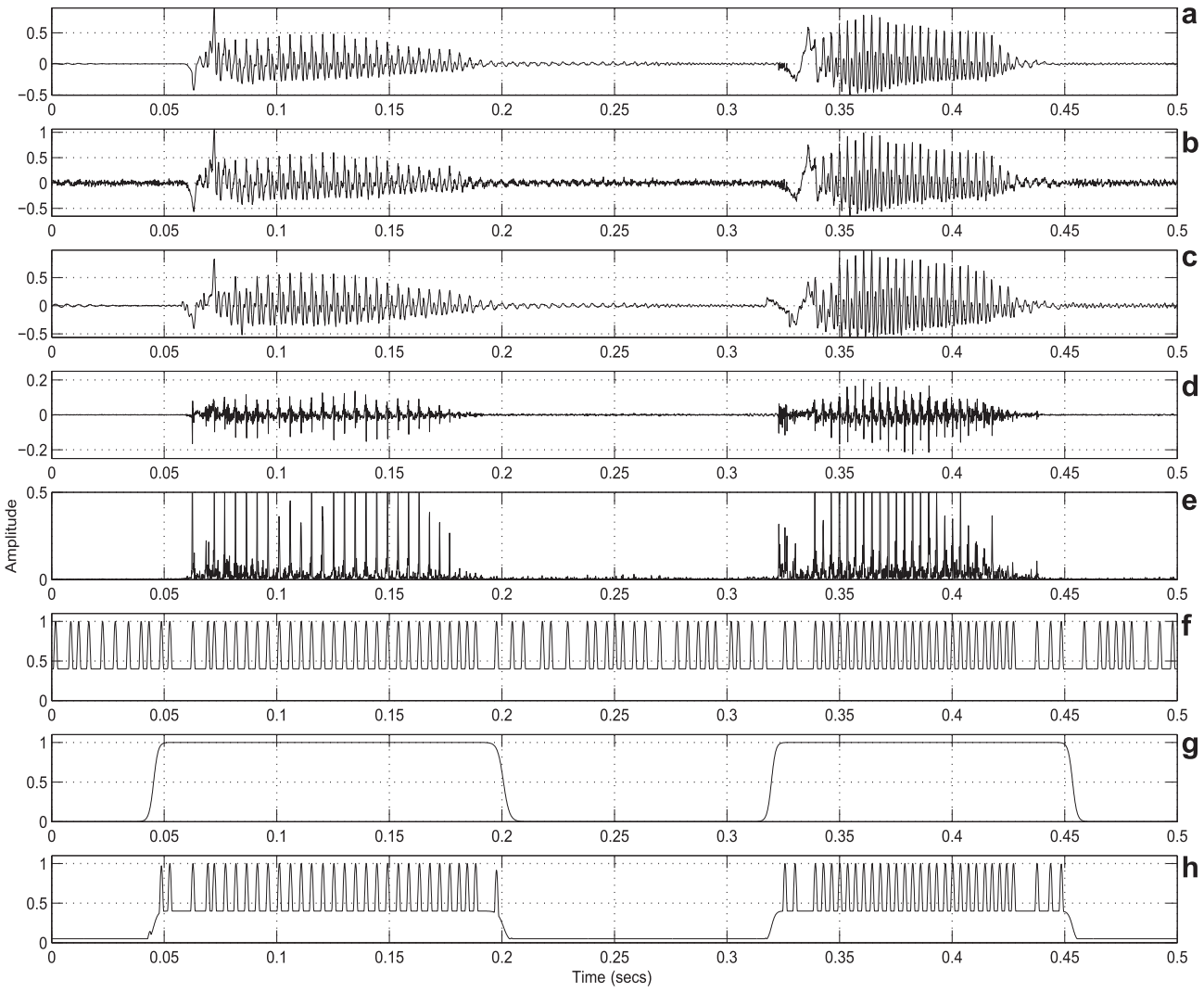

Fig. 6. LP residual weight function determination: (a) clean speech, (b) noisy speech (SNR = 3 dB), (c) speech signal synthesized using eight sinusoidal components, (d) LP residual of signal shown in (c), (e) mean smoothed HE, (f) fine weight function, (g) gross weight function and (h) final weight function.

features are emphasized in the temporally processed speech, the noise suppression is minimal mainly due to the use of all-pole filters derived from the noisy speech. To improve the vocal tract characteristics at the spectral level and to provide better noise suppression, the spectral processing is performed on the temporally processed speech that involve conventional spectral processing and proposed spectral enhancement techniques.

3.1. Conventional spectral processing

Generally, in majority of the conventional spectral processing methods, both short-term magnitude of degradation and degraded speech spectra are estimated first. According to the suppression rule, a spectral gain function is applied to the magnitude spectra of the degraded speech to obtain enhanced speech spectra. The enhanced magnitude and degraded speech phase spectra are then combined to produce an estimate of clean speech. For time domain resynthesis, overlap-add (OLA) method is typically used. In this work, the proposed temporal processing method is combined with four different spectral based speech enhancement algorithms, namely, spectral subtraction (SS) (Boll, 1979), multi-band spectral subtraction (MBSS) (Kamath and Loizou, 2002), MMSE-STSA estimator (Ephraim and Malah, 1984) and MMSE-LSA estimator (Ephraim and Malah, 1985). The spectral gain function of individual methods is given in Table 4.

3.2. Proposed spectral enhancement technique

From human perception point of view, the high SNR regions in the temporal domain (instants of significant excitation) and the peaks in the short time spectra, specifically, formants, pitch and its harmonics play central importance in the perception of speech (Lim and Oppenheim, 1979; Smits and Yegnanarayana, 1995; Yegnanarayana et al., 1999; Munkong and Juang, 2008). The temporal processing

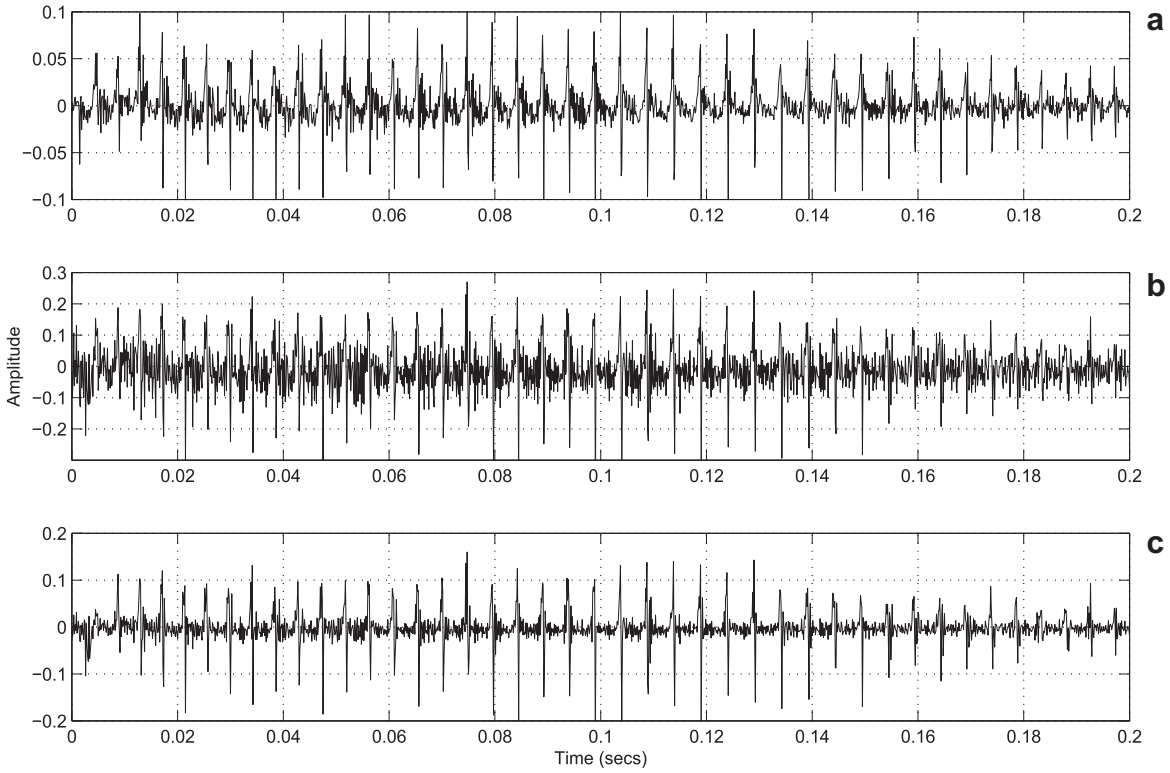

Fig. 7. LP residual weighting : (a) LP residual of clean speech, (b) LP residual of noisy speech (SNR = 3 dB) and (c) LP residual shown in (b) weighted by a weight function shown in Fig. 5(e).

Table 3
Percentage of approximate instants derived for different deviations with respect to clean speech instant locations.

SNR level (dB)	Deviation in time White noise				Deviation in time Babble noise			
	0.5 ms	1 ms	1.5 ms	2 ms	0.5 ms	1 ms	1.5 ms	2 ms
0	49.90	67.10	78.00	84.44	74.37	83.07	88.42	91.98
3	54.69	70.60	80.95	87.05	74.85	83.96	88.97	92.6
6	58.67	73.13	82.25	88.07	77.11	84.99	89.65	93.01
9	61.62	76.56	84.24	89.24	77.31	86.15	89.92	92.67
	Factory noise				Pink noise			
0	72.86	81.97	87.59	91.50	54.35	70.73	79.85	87.18
3	74.64	83.41	88.21	91.71	60.66	73.47	81.91	89.03
6	77.38	85.47	89.58	92.73	64.15	76.90	84.85	91.02
9	78.68	86.36	89.72	92.53	67.99	78.68	85.26	90.82

Table 4
Gain functions of conventional spectral processing methods.

Method	Gain function
SS (Boll, 1979)	$H(k) == \frac{\lvert Y(k)\rvert - \lvert\widehat{D}(k)\rvert}{\lvert Y(k)\rvert}$ $Y(k)$ is the maginitude specrum of noisy speech, and $\widehat{D}(k)$ is the time-average of the magnitude spectrum of noise calculated during silence
MBSS (Kamath and Loizou, 2002)	$H_j(k) == \max\left\{\sqrt{\frac{\lvert Y_j(k)\rvert^2 - \alpha_j\delta_j\lvert\widehat{D}_j(k)\rvert^2}{\lvert Y_j(k)\rvert^2}}, \beta\lvert Y_j(k)\rvert\right\};\ b_j \leqslant k \leqslant e_j$ b_j & e_j are the beginning and ending frequency bins of the jth frequency band, and α_j & δ_j are the over-subtraction and tweaking factor of jth frequency band
MMSE-STSA (Ephraim and Malah, 1984)	$H(k) = \left(\frac{\sqrt{\pi}}{2}\right)\frac{\sqrt{v_k}}{\gamma_k}\exp\left(-\frac{v_k}{2}\right)\left[(1+v_k)I_0\left(\frac{v_k}{2}\right) + v_k I_1\left(\frac{v_k}{2}\right)\right]$ $I_0(\cdot)$ & $I_1(\cdot)$ denote the zero and first order modified Bessel functions, respectively, and $v_k = \frac{\xi_k}{1+\xi_k}\gamma_k$; ξ_k & γ_k are *a priori SNR* and *a posteriori SNR*, respectively
MMSE-LSA (Ephraim and Malah, 1985)	$H(k) = \frac{\xi_k}{1+\xi_k}\exp\left(\frac{1}{2}\int_{v_k}^{\infty}\frac{e^{-x}}{x}dx\right)$

approach enhances the region around the instants of significant excitation and the subsequent spectral processing suppresses the noise spectral components. To further improve the perceptual quality of the speech, this work proposes the spectral enhancement technique on the high SNR regions of the spectrally processed speech (speech processed by the temporal and conventional spectral processing method) that enhances the pitch and its harmonics of voiced speech. The motivation here is that the most of speech energy is concentrated at the harmonics of fundamental frequency, and therefore enhancing these regions further enhance the speech components. This will results in reduction of residual noise.

The proposed spectral enhancement is performed only on the high SNR regions of the spectrally processed speech. This require an estimate of pitch information and is computed from the autocorrelation of the HE of temporally processed LP residual (Prasanna and Yegnanarayana, 2004). In the temporally processed speech the instants of significant excitation of the voiced speech are already gets enhanced and hence the estimation of pitch will be robust. Let, $s_t(n)$ be the enhanced speech signal by temporal processing method and $h(n)$ be the HE of LP residual of $s_t(n)$. For each block of 40 ms with a shift of 10 ms, the autocorrelation sequence is computed as (Proakis and Manolakis, 1996)

$$R(\tau) = \sum_{n=0}^{L-1-l} h_m(n)h_m(n+\tau); \quad \tau = 0, 1, 2, \ldots, L-1 \tag{10}$$

where $L = 320$ for $F_s = 8$ kHz and

$$h_m(n) = h(n) - E\{h(n)\} \tag{11}$$

where $E\{\cdot\}$ denotes the expected value operator. The first major peak with reference to zero time lag is considered as pitch period. The autocorrelation methods need at least two pitch periods to detect pitch and hence frame size of 40 ms is chosen.

The performance of pitch estimation is evaluated in terms of deviation between the estimated (i) pitch frequencies of clean and degraded speech and (ii) pitch frequencies of clean and temporally processed speech. For evaluation, the pitch of the respective signal is estimated from the autocorrelation of the HE of LP residual. Then the accuracy of pitch estimation is measured as

$$P_e = \frac{N_{tp}}{N_{cs}} \times 100 \tag{12}$$

where N_{cs} is the total number of frames in the clean speech and N_{tp} is the total number of frames having $F_{cs} > 0$ & $|F_{cs} - F_{tp}| \leqslant F_r$. The abbreviations F_{cs} and F_{tp} represent pitch (in Hz) of clean and temporally processed speech, respectively. F_r is frequency deviation considered for the evaluation. The results of this evaluation are given in Table 5 for different values of F_r ($F_r = 5$, 10, 15 & 20 Hz). For comparison the same experiment is repeated with reference to clean and degraded speech and the results are tabulated in Table 6. From the results it can be seen that the pitch estimate obtained from temporally processed speech consistently gives superior performance than from the degraded speech. For higher levels of degradation, the estimation error of degraded speech is very high (nearly 50%). On the other hand the pitch estimate obtained from the temporally processed speech shows acceptable performance. As mentioned earlier, this is mainly due to the enhancement of the instants of significant excitation of voiced speech.

After obtaining the pitch, its harmonic frequency locations are derived from the estimated pitch information. The enhanced magnitude spectrum of the desired speech components are constructed by sampling the spectrally processed speech spectrum at pitch and harmonic instants. The double-sided exponential function is used to sample the pitch and its harmonics. That is,

$$w_d(k) = e^{-\upsilon|k|}; \quad -\frac{L_p}{4} \leqslant k \leqslant \frac{L_p}{4} \tag{13}$$

where L_p is the frequency index corresponding to the pitch frequency and the value of υ is experimentally determined as 0.5. The sampled spectra is added with the spectrally processed speech spectra and the resultant spectra is

Table 5
Percentage of accuracy of the pitch estimation of temporally processed speech with reference to clean speech.

SNR level (dB)	Deviation in frequency (Hz)				Deviation in frequency (Hz)			
	±5	± 10	±15	±20	±5	±10	±15	±20
	White noise				Babble noise			
0	75.58	78.47	78.67	78.84	89.69	90.54	90.74	90.84
3	80.16	82.57	82.81	82.98	91.12	91.90	92.13	92.23
6	83.93	85.72	85.93	86.10	92.23	92.85	93.05	93.08
9	86.88	88.57	88.74	88.84	93.22	93.73	93.86	93.90
	Factory noise				Pink noise			
0	85.62	87.15	87.42	87.49	81.45	83.49	83.55	83.69
3	88.40	89.79	89.89	90.00	85.15	86.81	86.94	87.05
6	90.30	91.42	91.45	91.52	87.96	89.32	89.52	89.66
9	91.79	92.54	92.68	92.71	89.66	91.05	91.22	91.32

Table 6
Percentage of accuracy of the pitch estimation of degraded speech with reference to clean speech.

SNR level (dB)	Deviation in frequency (Hz)				Deviation in frequency (Hz)			
	±5	± 10	±15	±20	±5	±10	±15	±20
	White noise				Babble noise			
0	57.95	58.32	58.32	58.32	81.62	82.03	82.10	82.13
3	62.36	62.94	62.94	62.94	85.69	86.06	86.13	86.20
6	68.84	69.35	69.35	69.35	88.34	88.81	88.88	88.91
9	73.89	74.53	74.57	74.57	90.13	90.57	90.67	90.71
	Factory noise				Pink noise			
0	71.85	72.43	72.47	72.47	65.31	65.75	65.75	65.75
3	77.18	77.86	77.89	77.89	70.97	71.62	71.62	71.65
6	81.52	82.16	82.16	82.16	76.36	77.18	77.21	77.21
9	85.89	86.40	86.44	86.47	80.91	81.62	81.69	81.72

recombined with the original noisy speech phase spectra and converted back to the time domain by an IDFT.

The spectral enhancement steps are illustrated in Fig. 8 with reference to the spectral subtraction method. Fig. 8(a)–(c) show the spectrum of a frame of voiced portion of clean, noisy and the spectral subtracted speech, respectively. Fig. 8(d) shows the window function derived from the pitch and harmonic locations used for sampling the spectrum. The sampled spectrum is added to the spectrally subtracted speech spectrum to enhance the spectral peaks at pitch and harmonics and is shown in Fig. 8(e). Enhanced spectral peaks may be observed at pitch and harmonic instants. Finally, the various steps involved in the proposed combined Temporal and Spectral Processing (TSP) method are given in Table 7.

For illustration, the speech data spoken by a female speaker is selected and white Gaussian noise is added to make global SNR of the signal as 3 dB and shown in Fig. 9(a) (last 2.5 s of the signal given in Fig. 2(a)). The degraded signal is processed by the conventional spectral processing and the proposed combined TSP method as described earlier. Fig. 9 show the comparisons of the speech spectrograms obtained by different enhancement methods. All the speech spectrograms presented in this section used Hamming window of 128 samples with an overlap of 64 samples. The spectrogram of processed signal by the proposed method (Fig. 9(j)) shows significant improvement and also noticeable reduction of random peaks compared to conventional spectral processing methods.

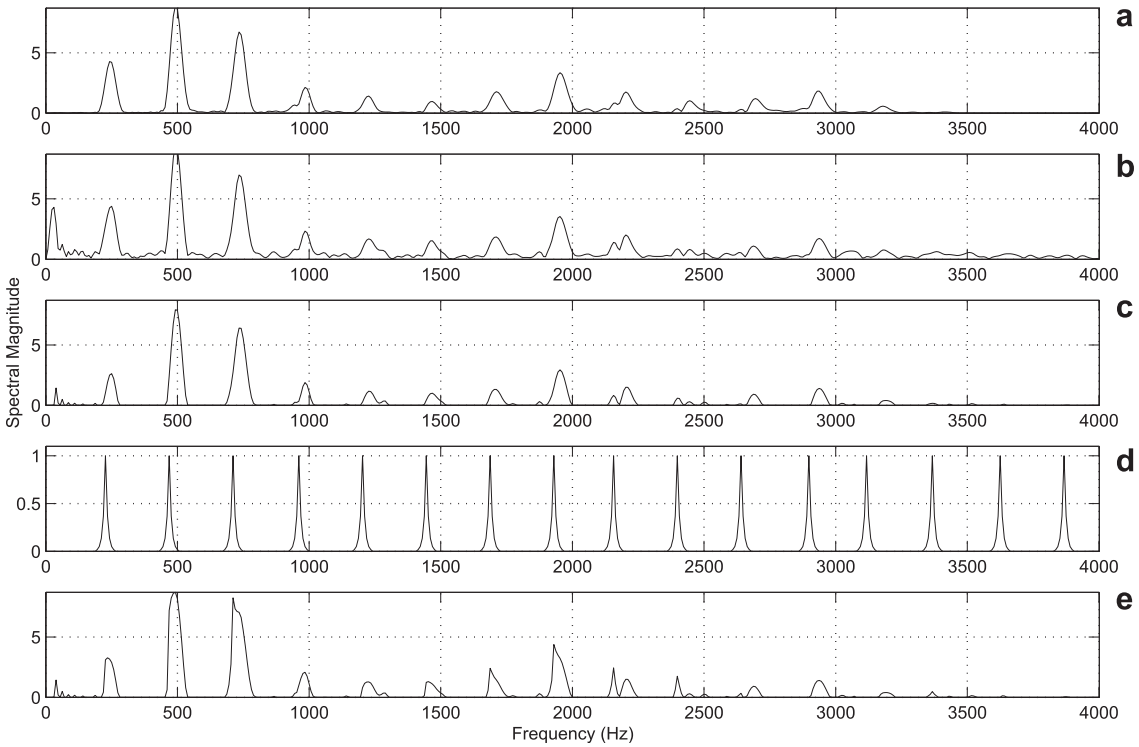

Fig. 8. Speech components enhancement: (a) clean speech spectrum, (b) noisy speech spectrum, (c) spectral subtracted speech Spectrum, (d) window function for sampling the spectral subtracted speech spectrum and (e) enhanced spectrum.

Table 7
Temporal and spectral processing algorithm for noisy speech enhancement.

Temporal processing:
Gross level processing
- Compute LP residual of noisy speech using a frame size of 20 ms, shift of 10 ms and 10th order LP analysis
- Compute the sum of 10 largest peaks in the DFT magnitude spectrum
- Compute the Hilbert envelope of the LP residual and mean smooth using 50 ms rectangular window
- Compute the modulation spectrum of the noisy speech signal
- Enhance the high SNR regions of each of the above parameters using the FOD
- Sum all the enhanced parameters and normalize the sum with respect to maximum value
- Smooth the normalized sum by using a 50 ms hamming window
- Nonlinearly map the normalized sum values by using a sigmoid non-linear function with slope parameter $\lambda = 20$ and T is equal to average value of the normalized sum
- The nonlinearly mapped values is termed as gross weight function

Fine level processing
- Compute the DFT magnitude and phase spectra for the noisy speech using 1024 point DFT
- Pick the largest eight peaks in the DFT magnitude spectrum and corresponding phase values and synthesize the speech
- Calculate LP residual of the signal obtained.
- Compute the Hilbert envelope of the LP residual and mean smooth using 1 ms rectangular window
- Obtain the FOGD operator from Gaussian window of length $N = 80$ samples and $\sigma = 8$
- Convolve the negative of FOGD operator with the mean smoothed Hilbert envelope of the LP residual and determine negative to positive transitions
- Convolve detected instants with 3 ms hanning window. The resultant signal is termed as fine weight function

Final weight function
- Multiply the two weight functions (gross weight function and fine weight function) to generate the final weight function
- Multiply the LP residual signal of noisy speech by the final weight function
- Excite the time-varying all-pole filter using weighted residual to obtain the temporally processed speech

Spectral processing:
- Update the noise magnitude spectrum if five consecutive frames are detected as non-speech (low SNR) regions
- Process the temporally processed speech by any of the conventional spectral processing method
- Determine the pitch period of the high SNR region from the Hilbert envelope of the temporally processed LP residual
- Enhance the pitch and harmonic locations of spectrally processed speech
- Reconstruct the enhanced speech signal using IDFT and overlap-add method

4. Experimental results and performance evaluation

The proposed speech enhancement method is evaluated using the composite objective quality measures that have high degree of correlation with subjective quality (Hu and Loizou, 2006, 2008). This measure evaluates the quality of enhanced speech along three dimensions: signal distortion, noise distortion, and overall quality. The resultant objective score values are in between 1 and 5 like mean opinion score (MOS). This measure rates the quality of the enhanced speech on three different counts. They are (Hu and Loizou, 2006, 2008)

1. The speech signal alone using a five-point scale of signal distortion (C_{sig}) [1-Very unnatural, 2-Fairly unnatural, 3-Somewhat natural, 4-Fairly natural and 5-Very natural].
2. The background noise alone using a five-point scale of background intrusiveness (C_{bak}) [1-very intrusive, 2-somewhat intrusive, 3-Noticeable but not intrusive, 4-Somewhat noticeable and 5-Not noticeable].
3. The overall quality (C_{ovl}) [1 = bad, 2 = poor, 3 = fair, 4 = good and 5 = excellent].

These values are obtained by linearly combining the existing objective measures by the following relations (Hu and Loizou, 2006)

$$C_{sig} = 3.093 - 1.029LLR + 0.603PESQ - 0.009WSS \quad (14)$$

$$C_{bak} = 1.634 + 0.478PESQ - 0.007WSS + 0.063segSNR \quad (15)$$

$$C_{ovl} = 1.594 + 0.805PESQ - 0.512LLR - 0.007WSS \quad (16)$$

where, *LLR*, *PESQ*, *WSS* and *segSNR* represents the log likelihood ratio, perceptual evaluation of speech quality, weighted slope spectral distance and segmental SNR, respectively.

For evaluation, ten different samples (five male and five female) from the TIMIT database (Zue et al., 1990), ten different samples from the NOIZEUS database (Hu and Loizou, 2007) and the data recorded in the laboratory environment under noisy and noise free conditions are used. The TIMIT sentences are first down sampled to 8 kHz before noise is added. The NOIZEUS database contains 30 IEEE sentences spoken by three male and three female speakers. Out of this 10 speech samples (five male and five female) are randomly selected. The sentences were originally sampled at 25 kHz. These signals are also first resampled to 8 kHz. Four different noise sources (white Gaussian noise, babble noise, factory noise and pink noise) from NOISEX-92 database (Varga and Steeneken, 1993) are taken and added to the clean speech to obtain the noisy speech. The energy level of the noise is scaled such that the overall SNR of the noisy speech is maintained at 0, 3, 6 and 9 dB.

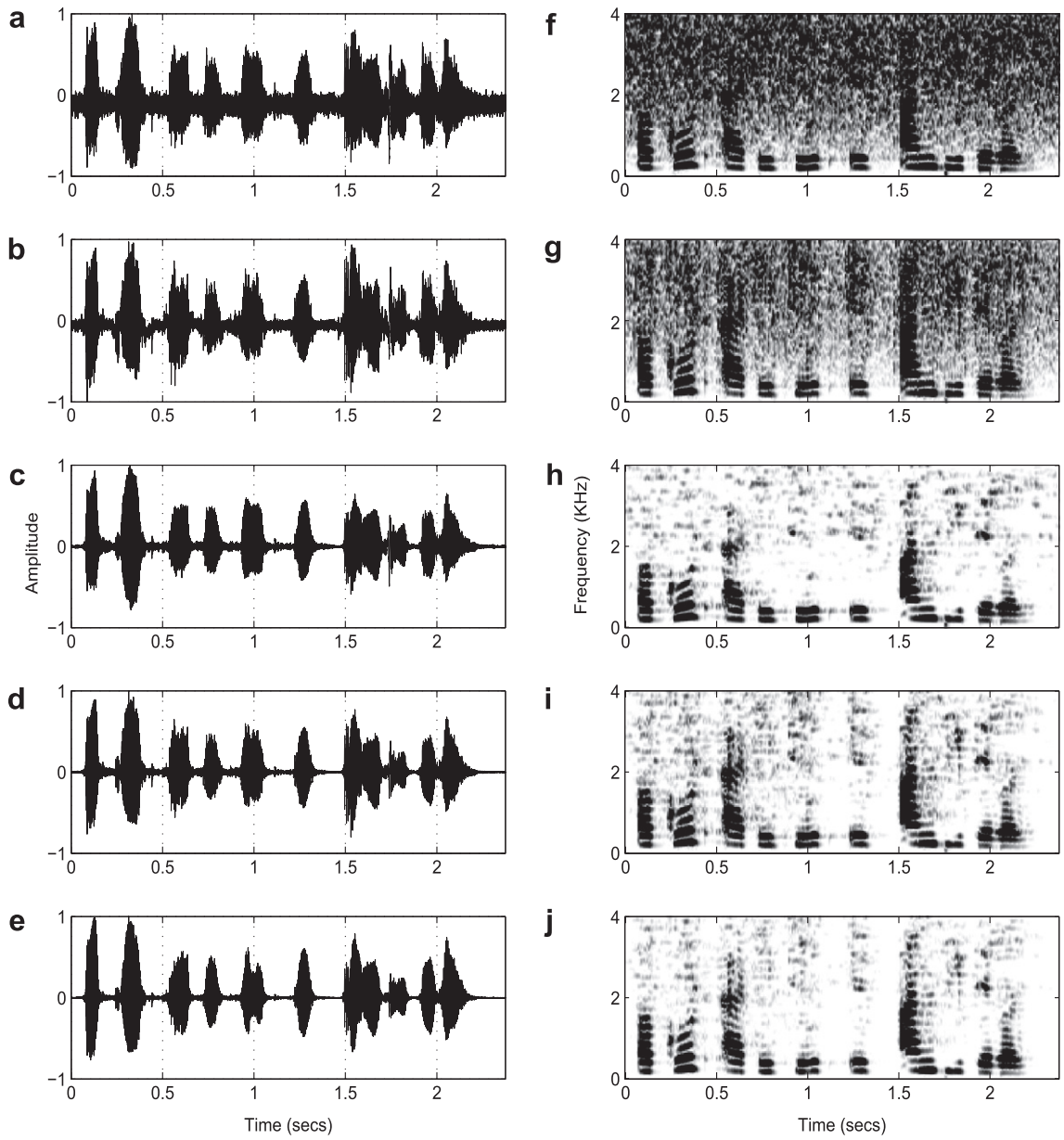

Fig. 9. Results of enhancement of noisy speech by temporal and MMSE-STSA estimator: (a) degraded speech, (b) speech processed by temporal processing, (c) speech processed by spectral processing (MMSE-STSA estimator), (d) speech processed by temporal and spectral processing (e) speech processed by temporal and spectral processing with spectral enhancement and (f)–(j) spectrograms of the respective signals shown in (a)–(e).

Tables 8–10 show MOS for signal distortion level, background noise level and overall objective quality values obtained from the different algorithms. Table 11 shows the percentage of improvement in each of the MOS score with reference to the degraded speech signal. The percentage of improvement in MOS is computed as

$$P_i = \frac{S_d - S_c}{S_d} \times 100 \tag{17}$$

where S_d and S_c respectively represent the degraded/enhanced speech MOS score and clean speech MOS score. These values are calculated for all SNR levels in all four noise cases and finally averaged across the noise types.

Table 12 values show the PSEQ values alone obtained from the different processing methods. The PESQ algorithm is designed to predict the subjective opinion score of a degraded audio sample and it is recommended by ITU-T for speech quality assessment (Rix et al., 2002). In PESQ measure a reference signal and the processed signal are first aligned in both time and level. This is followed by a range of perceptually significant transforms which include Bark spectral analysis, frequency equalization, gain variation equalization and loudness mapping. Then two parameters namely average disturbance value and average asymmetrical disturbance value are computed and then combined in a mapping function to give an estimate of MOS (Rix et al., 2002). For normal subjective test material

Table 8
Signal distortion score for different speech signals of the examples collected from the TIMIT and NOIZEUS database. In the table, abbreviations DEG, TP, SP1, SP2, SP3 and SP4 refer to degraded Speech, temporal Processing, spectral subtraction, multi-band spectral subtraction, MMSE-STSA estimator and MMSE-LSA estimator, respectively. TSPx refers to combined temporal and respective spectral processing. Similarly, TSPxE refers to combined temporal and respective spectral processing with spectral enhancement.

Noise type	SNR level (dB)	DEG	TP	SP1	SP2	SP3	SP4	TSP1	TSP2	TSP3	TSP4	TSP1E	TSP2E	TSP3E	TSP4E
White Gaussian noise	0	2.09	2.43	2.45	2.53	3.02	2.90	2.85	3.00	3.36	3.44	2.98	3.13	3.45	3.47
	3	2.46	2.78	2.84	2.91	3.37	3.28	3.27	3.42	3.63	3.70	3.38	3.51	3.70	3.72
	6	2.83	3.12	3.23	3.27	3.78	3.73	3.60	3.74	3.88	3.95	3.70	3.82	3.95	3.96
	9	3.19	3.43	3.66	3.69	4.06	4.01	3.88	4.01	4.11	4.16	3.97	4.08	4.16	4.17
Babble noise	0	3.03	3.30	3.40	3.46	3.62	3.57	3.54	3.64	3.72	3.71	3.67	3.78	3.81	3.79
	3	3.33	3.56	3.70	3.68	3.91	3.90	3.81	3.81	3.96	3.96	3.92	3.92	4.05	4.03
	6	3.62	3.81	3.99	3.97	4.19	4.19	4.06	4.06	4.19	4.21	4.16	4.16	4.28	4.28
	9	3.90	4.04	4.26	4.24	4.45	4.45	4.29	4.30	4.41	4.43	4.38	4.39	4.48	4.49
Factory noise	0	3.29	3.46	3.73	3.76	3.87	3.87	3.78	3.83	3.97	4.02	3.91	3.95	4.07	4.10
	3	3.56	3.71	4.02	4.04	4.13	4.15	4.04	4.08	4.19	4.26	4.16	4.20	4.29	4.32
	6	3.84	3.96	4.30	4.32	4.37	4.39	4.29	4.33	4.40	4.46	4.40	4.44	4.48	4.52
	9	4.10	4.19	4.56	4.58	4.61	4.63	4.51	4.55	4.59	4.64	4.61	4.65	4.67	4.69
Pink noise	0	2.76	3.01	3.08	3.09	3.41	3.34	3.35	3.44	3.70	3.75	3.48	3.55	3.79	3.80
	3	3.09	3.32	3.42	3.41	3.75	3.70	3.69	3.78	3.93	3.98	3.81	3.89	4.01	4.02
	6	3.41	3.61	3.79	3.78	4.06	4.03	3.97	4.05	4.15	4.20	4.08	4.15	4.23	4.23
	9	3.72	3.88	4.15	4.16	4.32	4.31	4.22	4.30	4.36	4.40	4.32	4.38	4.42	4.42

Table 9
Background noise level score for different speech signals of the examples collected from the TIMIT and NOIZEUS database. In the table, abbreviations DEG, TP, SP1, SP2, SP3 and SP4 refer to degraded speech, temporal processing, spectral subtraction, multi-band spectral subtraction, MMSE-STSA estimator and MMSE-LSA estimator, respectively. TSPx refers to combined temporal and respective spectral processing. Similarly, TSPxE refers to combined temporal and respective spectral processing with spectral enhancement.

Noise type	SNR level (dB)	DEG	TP	SP1	SP2	SP3	SP4	TSP1	TSP2	TSP3	TSP4	TSP1E	TSP2E	TSP3E	TSP4E
White Gaussian noise	0	1.22	1.47	1.50	1.54	1.90	1.96	1.76	1.84	2.02	2.13	1.87	1.96	2.11	2.20
	3	1.39	1.63	1.70	1.74	2.05	2.11	1.95	2.03	2.16	2.26	2.04	2.13	2.25	2.33
	6	1.57	1.80	1.89	1.92	2.21	2.26	2.11	2.19	2.29	2.39	2.21	2.29	2.38	2.45
	9	1.75	1.96	2.09	2.13	2.36	2.41	2.26	2.34	2.41	2.51	2.36	2.44	2.50	2.57
Babble noise	0	1.52	1.73	1.79	1.88	1.96	1.97	1.95	2.00	2.07	2.11	2.06	2.16	2.17	2.19
	3	1.69	1.89	1.97	1.96	2.13	2.15	2.10	2.11	2.22	2.26	2.22	2.22	2.32	2.34
	6	1.86	2.04	2.14	2.14	2.29	2.32	2.26	2.27	2.36	2.40	2.37	2.38	2.46	2.50
	9	2.03	2.19	2.31	2.31	2.45	2.48	2.40	2.42	2.49	2.54	2.52	2.53	2.60	2.63
Factory noise	0	1.69	1.84	1.98	2.02	2.10	2.15	2.08	2.12	2.21	2.30	2.19	2.23	2.32	2.39
	3	1.85	1.99	2.16	2.19	2.25	2.32	2.24	2.27	2.35	2.44	2.35	2.39	2.46	2.53
	6	2.00	2.14	2.33	2.37	2.39	2.46	2.39	2.43	2.48	2.57	2.51	2.55	2.58	2.66
	9	2.17	2.28	2.50	2.53	2.54	2.61	2.53	2.57	2.60	2.68	2.65	2.70	2.71	2.78
Pink noise	0	1.46	1.66	1.74	1.76	1.95	2.00	1.91	1.97	2.11	2.21	2.02	2.08	2.21	2.29
	3	1.64	1.82	1.91	1.93	2.12	2.17	2.08	2.14	2.24	2.33	2.19	2.25	2.34	2.41
	6	1.81	1.97	2.11	2.12	2.27	2.32	2.24	2.30	2.37	2.45	2.35	2.41	2.46	2.53
	9	1.97	2.13	2.30	2.32	2.42	2.47	2.39	2.45	2.49	2.57	2.51	2.56	2.58	2.63

the PESQ score ranges from 1.0 to 4.5, with higher score indicating better quality (Rix et al., 2002). Some of the observations from the results are as follow:

1. The combined TSP method shows improved performance compared to the individual processing methods.
2. The temporal processing alone gives less performance compared to the conventional spectral processing methods. It is expected mainly because there is no specific attempt made to explicitly remove the background noise. The enhancement is achieved only by processing of speech-specific regions, i.e., instants of significant excitation. From perception point of view also the speech enhanced by the temporal processing method is noisier than the one enhanced by the spectral based methods.
3. In spectral processing method, the order of best performing system in terms of overall objective quality scores are: (i) MMSE-LSA estimator, (ii) MMSE-STSA estimator, (iii) multi-band spectral subtraction, and (iv) conventional spectral subtraction. However with reference to signal distortion score MMSE-STSA estimator performs better than that of LSA estimator.

Table 10
Overall objective quality score for different speech signals of the examples collected from the TIMIT and NOIZEUS database. In the table, abbreviations DEG, TP, SP1, SP2, SP3 and SP4 refer to degraded speech, temporal processing, spectral subtraction, multi-band spectral subtraction, MMSE-STSA estimator and MMSE-LSA estimator, respectively. TSPx refers to combined temporal and respective spectral processing. Similarly, TSPxE refers to combined temporal and respective spectral processing with spectral enhancement.

Noise type	SNR level (dB)	DEG	TP	SP1	SP2	SP3	SP4	TSP1	TSP2	TSP3	TSP4	TSP1E	TSP2E	TSP3E	TSP4E
White Gaussian noise	0	1.82	2.12	2.12	2.17	2.59	2.54	2.47	2.60	2.88	2.99	2.59	2.72	2.97	3.03
	3	2.13	2.41	2.45	2.50	2.90	2.87	2.83	2.96	3.12	3.22	2.93	3.06	3.19	3.25
	6	2.44	2.70	2.78	2.81	3.26	3.27	3.12	3.26	3.34	3.44	3.22	3.34	3.41	3.46
	9	2.74	2.97	3.17	3.21	3.52	3.53	3.37	3.50	3.54	3.63	3.47	3.58	3.61	3.65
Babble noise	0	2.56	2.79	2.87	2.95	3.05	3.04	3.00	3.11	3.14	3.16	3.13	3.22	3.24	3.24
	3	2.83	3.03	3.15	3.14	3.32	3.34	3.24	3.24	3.36	3.39	3.36	3.36	3.46	3.47
	6	3.10	3.26	3.42	3.42	3.58	3.61	3.48	3.48	3.58	3.63	3.59	3.59	3.68	3.70
	9	3.36	3.48	3.68	3.68	3.83	3.86	3.70	3.71	3.89	3.84	3.80	3.81	3.88	3.91
Factory noise	0	2.78	2.93	3.18	3.23	3.32	3.36	3.22	3.27	3.38	3.47	3.35	3.40	3.49	3.55
	3	3.03	3.16	3.46	3.49	3.56	3.61	3.47	3.51	3.59	3.69	3.60	3.64	3.70	3.76
	6	3.28	3.39	3.73	3.77	3.78	3.84	3.70	3.75	3.79	3.89	3.83	3.88	3.89	3.94
	9	3.53	3.62	3.98	4.03	4.01	4.07	3.92	4.04	3.98	4.06	4.04	4.08	4.07	4.11
Pink noise	0	2.31	2.54	2.61	2.62	2.90	2.88	2.85	2.93	3.13	3.22	2.97	3.04	3.23	3.28
	3	2.60	2.82	2.91	2.91	3.20	3.20	3.15	3.24	3.34	3.43	3.27	3.35	3.43	3.48
	6	2.89	3.07	3.25	3.25	3.48	3.50	3.41	3.49	3.55	3.63	3.52	3.60	3.63	3.67
	9	3.16	3.31	3.59	3.60	3.72	3.76	3.65	3.73	3.74	3.82	3.76	3.83	3.82	3.84

Table 11
Percentage of improvement in signal distortion, background noise level and overall objective quality score with reference to the degraded speech. In the table, abbreviations TP, SP1, SP2, SP3 and SP4 refer to temporal processing, spectral subtraction, multi-band spectral subtraction, MMSE-STSA estimator and MMSE-LSA estimator, respectively. TSPx refers to combined temporal and respective spectral processing. Similarly, TSPxE refers to combined temporal and respective spectral processing with spectral enhancement.

SNR level (dB)	TP	SP1	SP2	SP3	SP4	TSP1	TSP2	TSP3	TSP4	TSP1E	TSP2E	TSP3E	TSP4E
Percentage of improvement in signal distortion score													
0	9.85	13.60	15.37	26.29	23.81	22.37	26.18	34.57	36.27	27.16	30.80	37.96	38.35
3	7.89	12.54	13.16	22.94	21.69	20.06	22.59	27.84	29.45	23.82	26.07	30.58	30.92
6	6.12	11.87	12.14	20.54	20.01	16.88	18.96	22.28	23.80	19.97	21.81	24.63	24.98
9	4.40	11.69	11.98	17.49	17.15	13.77	15.63	17.77	18.86	16.33	17.91	19.50	19.76
Percentage of improvement in background noise level score													
0	14.22	19.26	22.50	35.63	38.62	31.61	35.69	44.26	50.22	39.19	44.29	51.09	55.67
3	11.91	18.02	19.30	31.10	34.18	28.11	31.02	37.59	42.57	34.67	37.75	43.70	47.45
6	10.04	17.13	18.24	27.20	29.96	24.79	27.53	31.92	36.28	30.88	33.62	37.19	40.81
9	8.27	16.30	17.47	23.86	26.38	21.32	23.93	26.65	30.63	27.13	29.61	31.70	34.51
Percentage of improvement in overall quality score													
0	10.21	13.99	16.02	26.60	25.96	23.03	27.20	34.49	37.98	28.41	32.28	38.78	40.68
3	8.24	13.11	13.86	23.51	23.75	20.76	23.48	28.04	31.17	25.22	27.84	31.52	33.28
6	6.35	12.61	13.22	21.19	22.16	17.73	20.24	22.69	25.57	21.59	23.89	25.67	27.07
9	4.82	12.89	13.69	18.44	19.50	14.92	17.16	19.02	20.67	18.29	20.21	20.85	21.88

4. In spectral subtractive based algorithms multi-band spectral subtraction performed consistently better across all conditions.
5. The combined TSP method (without spectral enhancement technique) itself gives higher MOS score for signal distortion, background noise level and the overall objective quality compared to the temporal or spectral processing alone. The results show that the additional spectral enhancement technique gives relatively higher improvement in background noise level score as compared to the signal distortion score. This reduction in the background noise level is achieved through the enhancement of speech-specific features (region around pitch and its harmonics) in the spectral domain. As mentioned earlier, the enhancement of the speech-specific spectral amplitudes relatively reduces the noise spectral amplitudes in the high SNR regions. This results in higher MOS for background noise level compared to the combined temporal and spectral processing. However the same amount of relative increment is not evident in the overall quality score. This is theoretically interpreted as follows: Hu and Loizou considered that the overall quality score as the dependent variable and the speech and noise scores as independent variables. By regression analysis they found the relationship between the three scores as (Hu and Loizou, 2007)

Table 12
PESQ objective quality score for different speech signals of the examples collected from the TIMIT and NOIZEUS database. In the table, abbreviations DEG, TP, SP1, SP2, SP3 and SP4 refer to degraded speech, temporal processing, spectral subtraction, multi-band spectral subtraction, MMSE-STSA estimator and MMSE-LSA estimator, respectively. TSPx refers to combined temporal and respective spectral processing. Similarly, TSPxE refers to combined temporal and respective spectral processing with spectral enhancement.

Noise type	SNR level (dB)	DEG	TP	SP1	SP2	SP3	SP4	TSP1	TSP2	TSP3	TSP4	TSP1E	TSP2E	TSP3E	TSP4E
White Gaussian noise	0	1.36	1.60	1.58	1.61	1.93	1.94	1.86	1.96	2.15	2.28	1.96	2.07	2.23	2.33
	3	1.59	1.82	1.83	1.87	2.18	2.21	2.15	2.26	2.34	2.47	2.23	2.34	2.42	2.51
	6	1.82	2.04	2.09	2.12	2.47	2.54	2.38	2.50	2.52	2.65	2.47	2.58	2.60	2.67
	9	2.06	2.25	2.42	2.46	2.68	2.75	2.59	2.71	2.70	2.81	2.68	2.78	2.76	2.83
Babble noise	0	1.87	2.05	2.11	2.19	2.24	2.26	2.20	2.32	2.31	2.35	2.33	2.40	2.41	2.43
	3	2.10	2.24	2.35	2.35	2.47	2.51	2.41	2.42	2.50	2.55	2.53	2.53	2.60	2.63
	6	2.32	2.44	2.58	2.59	2.69	2.75	2.61	2.62	2.70	2.76	2.72	2.73	2.80	2.83
	9	2.55	2.64	2.80	2.82	2.92	2.97	2.81	2.83	3.07	2.95	2.92	2.94	2.98	3.02
Factory noise	0	2.03	2.16	2.37	2.43	2.49	2.58	2.40	2.45	2.53	2.64	2.52	2.58	2.63	2.71
	3	2.24	2.35	2.61	2.67	2.69	2.79	2.61	2.66	2.71	2.83	2.74	2.80	2.81	2.89
	6	2.46	2.55	2.86	2.92	2.89	2.98	2.83	2.88	2.88	3.00	2.96	3.01	2.99	3.06
	9	2.68	2.75	3.09	3.16	3.09	3.19	3.03	3.21	3.06	3.16	3.16	3.20	3.16	3.22
Pink noise	0	1.65	1.85	1.91	1.92	2.15	2.17	2.11	2.17	2.31	2.43	2.22	2.28	2.40	2.49
	3	1.89	2.07	2.16	2.16	2.39	2.44	2.35	2.43	2.49	2.60	2.47	2.54	2.58	2.65
	6	2.12	2.28	2.44	2.45	2.62	2.69	2.58	2.65	2.66	2.77	2.69	2.76	2.75	2.81
	9	2.34	2.48	2.74	2.77	2.83	2.91	2.79	2.87	2.83	2.93	2.91	2.97	2.91	2.95

$$C_{ovl} = -0.0783 + 0.571C_{sig} + 0.366C_{bak}. \quad (18)$$

This shows that the overall quality score has higher correlation with the signal distortion score as compared to the background noise level score. Due to this lower correlation the same amount of relative increment is not seen in the overall quality score. For lower correlation, the reason being stated was listeners seem to place more emphasis on the distortion imparted on the speech signal itself rather than on the background noise, when making judgments of overall quality.

6. In combined TSP method, the relative amount of increase in the performance reduces as the SNR of the noisy speech is increased from 0 to 9 dB. In addition, under white noise environment, combined TSP methods result lower performance than that of other noise environments. This poor performance can be attributed to the limitations of LP analysis under high degradation due to white noise.

Finally, the subjective evaluation is performed for assessing the quality of proposed method with the conventional approaches. The subjective tests are conducted with the help of 20 subjects. The test is conducted in the laboratory by playing the speech signals through headphones. The subjects are asked to listen to examples of speech and rate the perceived quality of speech on a five-point scale. The meanings assigned relative to the five-point scale rating are: 1-Bad, 2-Poor, 3-Fair, 4-Good, 5-Excellent (Deller et al., 1993). For evaluation, two sentences (one male and one female) are selected for each noise level and every noise type. Further, in spectral processing algorithms speech signals enhanced by the multi-band spectral subtraction (MBSS) and the MMSE-LSA estimator are considered for subjective evaluation. These two are the best performing algorithms in spectral subtraction and MMSE category. Table 13 shows the mean and their standard deviations of MOS ratings for different speech signals obtained from different processing methods. It is seen that the MOS values obtained from the proposed method is comparable and better than the individual processing methods.

To further analyze the statistical significance of MOS score values paired T-test is conducted. This is also termed *Students t* test (Press et al., 1992, Chapter 14). Table 14 shows the summary of *Students t* values and the significance values of (i) the spectral processing method over the temporal processing and (ii) the combined temporal and spectral processing method over the individual temporal and

Table 13
Summary of MOS ratings. In table the abbreviations DEG, TP, SP and TSP refer to degraded speech, speech processed by the temporal, spectral and combined temporal and spectral processing, respectively.

SNR (dB)	DEG		TP		SP				TSP			
					MBSS		MMSE-LSA		MBSS		MMSE-LSA	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
0	1.10	0.31	1.55	0.60	1.70	0.73	1.80	0.77	1.80	0.70	1.90	0.79
3	1.15	0.37	1.70	0.73	2.05	0.83	2.15	0.81	2.30	0.92	2.40	1.05
6	1.15	0.49	1.80	0.83	2.35	0.81	2.45	0.94	2.70	0.92	2.65	1.14
9	1.30	0.57	2.20	0.89	2.70	0.86	2.90	0.97	2.90	0.85	3.20	0.95

Table 14
Summary of paired T-test. In table the abbreviations DEG, TP, SP and TSP refer to degraded speech, speech processed by the temporal, spectral and combined temporal and spectral processing, respectively. Further, TP-SP refers to paired T-test between temporal and spectral processing (MBSS or MMSE) methods.

SNR (dB)	TP-SP		TP-TSP		SP-TSP	
	MBSS	MMSE	MBSS	MMSE	MBSS	MMSE
Students t value						
0	1.83	2.52	2.52	3.20	1.45	1.45
3	3.20	3.94	5.34	5.48	2.52	2.52
6	4.82	4.95	9.00	6.47	3.20	2.18
9	4.36	6.66	6.66	13.78	2.18	2.85
Significance value						
0	0.083	0.021	0.021	0.005	0.163	0.163
3	0.005	0.001	0.000	0.000	0.021	0.021
6	0.000	0.000	0.000	0.000	0.005	0.042
9	0.000	0.000	0.000	0.000	0.042	0.010

spectral processing methods. The significance is a number between zero and one, and is the probability that $|t|$ could be this large or larger just by chance, for distributions with equal means. Therefore, a small numerical value of the significance (<0.05) means that the observed difference is very significant (Press et al., 1992, Chapter 14). The significance values obtained from MOS scores infer that the MOS values of combined method are statistically significative compared to individual temporal and spectral processing methods.

5. Summary and conclusions

In this paper for speech degraded by background noise, a combined TSP method is proposed by emphasizing high SNR regions in the temporal domain, and eliminating the degradation and enhancing the speech-specific components in the spectral domain. The main objective of this study is to show that the combined TSP method gives relatively better performance compared to temporal or spectral processing alone. The enhancement of noisy speech is achieved in two stages, namely, temporal enhancement followed by spectral enhancement. In temporal enhancement process, first the sum of the ten largest peaks in the DFT spectrum, the smoothed HE of the LP residual and the modulation spectrum values of the noisy speech are computed to identify the gross level high and low SNR regions. Then the HE of the LP residual is used to derive the information about strength of excitation. A fine weight function is derived using a FOGD to enhance the excitation source information around the GC instants of speech. Because of high SNR nature of the regions around the GC events, the periodicity information is preserved even under high levels of degradation. A weight function for the LP residual is derived from the gross and fine weight functions. The enhanced speech is derived by exciting the time-varying all-pole filter with the LP residual modified by the weight function. In spectral enhancement, first enhancement is achieved by conventional spectral processing technique (spectral subtraction or MMSE estimator) and an additional spectral enhancement is performed by enhancing the pitch and its harmonics to further improve the perceptual quality of the speech. A performance evaluation is conducted using different objective quality measures: signal distortion, noise distortion, overall quality and PESQ. The performance measures showed that the processed speech signals from the proposed method results better MOS compared to temporal or spectral processing alone. The important contribution of the research work reported in this paper is the development of combined TSP methods for noisy speech enhancement. While developing combined TSP method the other contributions are as follows: (i) set of speech-specific features are proposed for the gross level detection of high SNR speech regions, (ii) method to determine the instants of significant excitation in noisy speech is proposed, and (iii) new spectral enhancement technique is proposed for the voiced regions of degraded speech.

There are several improvements that can be made in the proposed methods. The performance of the proposed gross level detection can be increased by adaptively updating the thresholds or combining the proposed features with existing VAD features. For this thorough assessment of the strengths and weaknesses of the individual parameters needs to be done to compare the performance with the existing VAD methods. The proposed gross level detection algorithm requires the maximum value of an individual feature computed across the whole utterance to obtain the normalized values. Hence at present, the proposed algorithm can be used only in offline. The future work should focus on how to make this algorithm suitable for real time operation. In this work no explicit study is made for processing unvoiced regions of degraded speech. Therefore analysis of unvoiced regions can be done both in temporal and spectral processing. In particular, methods need to be developed for (i) identification of unvoiced sounds, (ii) defining and identification of instants of significant excitation for unvoiced sound, and (iii) identification of speech-specific spectral features of unvoiced sounds. In practical conditions, methods can be developed to identify the type of degradation and also the level of degradation. As a result, the temporal weight function can be adaptively updated to improve the performance.

References

Ananthapadmanabha, T., Yegnanarayana, B., 1979. Epoch extraction from linear prediction residual for identification of closed glottis interval. IEEE Trans. Acoust. Speech Signal Process. ASSP-27, 309–319.

Berouti, M., Schwartz, R., Makhoul, J., 1979. Enhancement of speech corrupted by acoustic noise. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing, pp. 208–211.

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. ASSP-27, 113–120.

Chang, J.-H., Gazor, S., Kim, N.S., Mitra, S.K., 2007. Multiple statistical models for soft decision in noisy speech enhancement. Pattern Recognition 40, 1123–1134.

Chen, B., Loizou, P., 2005. Speech enhancement using a MMSE short time spectral amplitude estimator with Laplacian speech modeling. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing, vol. 1, Philadelphia, PA, USA, pp. 1097–1100.

Chen, B., Loizou, P.C., 2007. A Laplacian-based MMSE estimator for speech enhancement. Speech Comm. 49, 134–143.

Deller, J.R., Hansen, J.H., Proakis, J.G., 1993. Discrete Time Processing of Speech Signals, 1st ed. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Donoho, D.L., 1995. De-noising by soft-thresholding. IEEE Trans. Inf. Theory 41 (3), 613–627.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. ASSP-32, 1109–1121.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. ASSP-33, 443–445.

Greenberg, S., Kingsbury, B.E.D., 1997. The modulation spectrogram: In pursuit of an invariant representation of speech. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing, Munich, Germany, pp. 1647–1650.

Hu, Y., Loizou, P.C., 2006. Evaluation of objective measures for speech enhancement. In: Proc. Interspeech, Philadelphia, PA, USA.

Hu, Y., Loizou, P.C., 2007. Subjective comparison and evaluation of speech enhancement algorithms. Speech Comm. 49, 588–601.

Hu, Y., Loizou, P.C., 2008. Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio, Speech, Lang. Process. 16 (1), 229–238.

Jin, W., Scordilis, M.S., 2006. Speech enhancement by residual domain constrained optimization. Speech Comm. 48, 1349–1364.

Kamath, S., Loizou, P., 2002. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing, Orlando, USA.

Kim, W., Kang, S., Ko, H., 2000. Spectral subtraction based on phonetic dependency and masking effects. IEEE Proc. Vision, Image Signal Process. 147 (5), 423–427.

Krishnamoorthy, P., Prasanna, S.R.M., 2008. Temporal and spectral processing of degraded speech. In: IEEE Proc. Internat. Conf. on Advanced Computing and Communications 2008 (ADCOM08), Chennai, India, pp. 112–118.

Krishnamoorthy, P., Prasanna, S.R.M., 2009. Reverberant speech enhancement by temporal and spectral processing. IEEE Trans. Audio, Speech, Lang. Process. 17 (2), 253–266.

Lim, J., Oppenheim, A., 1979. Enhancement and bandwidth compression of noisy speech. Proc. IEEE 67 (12), 1586–1604.

Loizou, P.C., 2007. Speech Enhancement: Theory and Practice, 1st ed. CRC, Boca Raton, FL.

Lu, C.-T., 2007. Reduction of musical residual noise for speech enhancement using masking properties and optimal smoothing. Adv. Pattern Recognition Lett. 28 (11), 1300–1306.

Makhoul, J., 1975. Linear prediction: A tutorial review. Proc. IEEE 63, 561–580.

Marple, J.L., 1999. Computing the discrete-time "analytic" signal via FFT. IEEE Trans. Signal Process. 47 (9), 2600–2603.

Martin, R., 2005. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. IEEE Trans. Speech Audio Process. 13, 845–856.

Marzinzik, M., Kollmeier, B., 2002. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. IEEE Trans. Speech Audio Process. 10, 109–118.

McAulay, R., Quatieri, T., 1986. Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans. Acoust. Speech Signal Process. ASSP-34, 744–754.

Munkong, R., Juang, B.-H., 2008. Auditory perception and cognition. IEEE Signal Process. Mag. 25 (3), 98–117.

Prasanna, S.R.M., Subramanian, A., 2005. Finding pitch markers using first order Gaussian differentiator. In: IEEE Proc. 3rd Internat. Conf. on Intelligent Sensing Information Process., Bangalore, India, pp. 140–145.

Prasanna, S.R.M., Yegnanarayana, B., 2004. Extraction of pitch in adverse conditions. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing, vol. 1, Montreal, Quebec, Canada, pp. I-109–I-112.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1992. Numerical recipes in C : The art of scientific computing, second ed. Cambridge University Press, pp. 615–619, (Chapter 14).

Proakis, J.G., Manolakis, D.G., 1996. Digital Signal Processing-Principles, Algorithms, and Applications, third ed. Prentice Hall.

Rix, A.W., Hollier, M.P., Hekstra, A.P., Beerends, J.G., 2002. Perceptual evaluation of speech quality (PESQ) the new ITU standard for end-to-end speech quality assessment part I – Time-delay compensation. J. Audio Eng. Soc. 50 (10), 755–764.

Schroeder, M., 1970. Parameter estimation in speech: A lesson in unorthodoxy. Proc. IEEE 58 (5), 707–712.

Senapati, S., Chakroborty, S., Saha, G., 2008. Speech enhancement by joint statistical characterization in the Log Gabor Wavelet domain. Speech Comm. 50, 504–518.

Seok, J.W., Bae, K.S., 1999. Reduction of musical noise in spectral subtraction method using subframe phase randomisation. Electron. Lett. 35, 123–125.

Shao, Y., Chang, C.-H., 2007. A generalized time frequency subtraction method for robust speech enhancement based on wavelet filter banks modeling of human auditory system. IEEE Trans. Systems Man Cybernet. Part B 37 (4), 877–889.

Smits, R., Yegnanarayana, B., 1995. Determination of instants of significant excitation in speech using group delay function. IEEE Trans. Speech Audio Process. 3, 325–333.

Sri Rama Murty, K., Yegnanarayana, B., Guruprasad, S., 2007. Voice activity detection in degraded speech using excitation source information. In: Proc. Interspeech, Antwerp, Belgium, pp. 2941–2944.

Varga, A., Steeneken, H.J.M., 1993. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Comm. 12 (3), 247–251.

Yamashita, K., Shimamura, T., 2005. Nonstationary noise estimation using low-frequency regions for spectral subtraction. IEEE Signal Process. Lett. 12, 465–468.

Yang, L.P., Fu, Q.J., 2005. Spectral subtraction-based speech enhancement for cochlear implant patients in background noise. J. Acoust. Soc. Amer. 117, 1001–1004.

Yegnanarayana, B., Sri Rama Murty, K., 2009. Event-based instantaneous fundamental frequency estimation from speech signals. IEEE Trans. Audio, Speech, Lang. Process. 17 (4), 614–624.

Yegnanarayana, B., Avendano, C., Hermansky, H., Satyanarayana Murthy, P., 1999. Speech enhancement using linear prediction residual. Speech Comm. 28, 25–42.

Yegnanarayana, B., Prasanna, S.R.M., Rao, K.S., 2002. Speech enhancement using excitation source information. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing, vol. 1, Orlando, USA, pp. I-541–I-544.

Zue, V., Seneff, S., Glass, J., 1990. Speech database development at MIT: TIMIT and beyond. Speech Comm. 9 (4), 351–356.