I J C T A, 10(9), 2017, pp. 371-378 ISSN: 0974-5572
© International Science Press

Enhancement of Noisy Speech Using Spectral Substraction Method

Thottempudi Pardhu* and R. Karthik**

ABSTRACT

The main objective of this paper is to produce an summary of the most important techniques that are proposed for enhancement of noisy speech. This paper also proposes a technique for enhancement of noisy speech. The projected technique is motivated as an effort to reduce the constraints of typical spectral subtraction technique for enhancement of noisy speech. The proposed technique involves 3 steps. in the beginning speech and non-speech regions are detected from the degraded speech signal. Within the second step, for every speech region, noise elements are estimated from the preceding noise regions and are subtracted by standard spectral subtraction technique. in the third step, speech components are increased further from the spectral subtracted speech signal for reducing the musical noise, which is the object of spectral subtraction technique. The processed speech signals from the proposed technique appear to be higher perceptually compared thereto of the spectral subtraction technique.

Keywords: Speech, Spectral Subtraction Method, Enhancement, Noise

1. INTRODUCTION

Various speech processing systems have found their way in our everyday life through their dramatic use in voice communication, speech and speaker recognition, aid for the hearing impaired and numerous other applications. In many situations of practical interests, the speech signal gets corrupted by one or more of the following: additive background noise, reverberation and speech from other speakers [1]. These degradations will lower the quality and intelligibility of speech message. The speech enhancement methods aim to improve the quality and intelligibility of degraded speech. The processed speech signals should be more comfort for listening and also should give better performance in tasks like automatic speech and speaker recognition [2]. The additive background noise is random in nature and also uncorrelated with speech. In case of reverberation, reflections of speech from various objects will be mixed with the speech in a convoluted fashion. Thus degradation in case of reverberation is signal dependent, where as, it is independent in case of additive background noise. Speech from other speakers may also get mixed with desired speaker's speech in an additive fashion. Since the characteristics of degradation are different in each case, degraded speech may need to be processed in different ways. In the present work we focus on speech degraded predominantly by additive background noise. Speech degraded by such degradation is termed as noisy speech and hence in this work, unless specified, degraded speech refers to noisy speech.

Additive background noise is uncorrelated with the speech signal and present in various environment scenarios like offices, cars, city streets, fans, factory environments, helicopters etc. In case of additive background noise the assumptions made for developing enhancement methods are, (*i*) speech and noise signals are uncorrelated at least over a short-time basis, (*ii*) noise is either stationary or slowly varying over several frames of speech, and (*iii*) noise can be represented as zero mean random process [3]. The degradation

* Department of Electronics & Communication Engineering MLR Institute of Technology Hyderabad, India, *Email: pthottempudi2020@gmail.com*

** Department of Electronics & Communication Engineering MLR Institute of Technology Hyderabad, India, *Email: karthik.r@mlrinstitutions.ac.in*

level of additive background noise is normally specified by the measure called Signal to Noise Ratio (SNR) and is defined as the ratio of signal energy to noise energy.

Speech corrupted by additive background noise is termed as noisy speech [4]. Speech enhancement schemes currently available in the literature to enhance noisy speech can be divided into single channel and multi-channel methods, depending on whether the noisy speech from the environment is collected over single or multiple channels. In most common scenario like mobile communication, hearing aids etc. usually a second channel may not be available. Therefore, single channel systems mostly make use of different statistics of speech and unwanted noise. The performance of these methods are limited in presence of non-stationary noise as most of the methods make an assumption that noise is stationary. Multiple channel speech enhancement technique take advantage of availability of multiple signal input to the system, making possible the use of noise references for enhancement.

This paper explains some of the major methods proposed in the literature for processing noisy speech for enhancement. The work also proposes a speech enhancement method for processing noisy speech. The rest of the paper is organized as follows: Section 2 briefly reviews various noisy speech enhancement methods proposed in the literature. In section 3 proposed method for noisy speech enhancement is explained and finally summary of the present work and scope for future work are given in section 4.

2. ENHANCEMENT OF NOISY SPEECH

A noisy speech signal can be modeled as the sum of clean speech and additive background noise.

$$y(n) = s(n) + d(n) \tag{1}$$

where, $y(n)$, $s(n)$ and $d(n)$ denote frames of noisy speech, clean speech and additive background noise, respectively.

Spectral subtraction [3] is a popular frequency domain method to reduce the effect of additive uncorrelated noise in a signal. Spectral subtraction carries out noise reduction by subtracting an estimate of noise spectrum from the noisy speech signal. The noise estimation is obtained from the segments where speech is absent, typically, few 100 ms from the beginning of speech signal, under the assumption that the statistics of noise do not rapidly vary with time. The major drawback of this approach is that it introduces noise with annoying noticeable tonal characteristics referred as musical noise due to short spurious bursts of isolated frequency components that appear randomly across enhanced speech spectrum [5]. Also, the performance degrades severally in case of correlated noise like colored noise and also non-stationary environments. Several modifications have been proposed in the literature on the spectral subtraction method to reduce the effect of musical noise.

Boll [3] proposed few modifications such as magnitude averaging, half wave rectification, residual noise reduction and additional signal attenuation during non-speech activity to reduce the effect of musical noise. Berouti *et al* [6] suggested a method to reduce the musical noise by subtracting an overestimate of the noise power spectrum from the speech power spectrum to minimize the appearance of negative values that generate spectral spikes, and spectral floor to reduce the spectral excursions, however over subtraction was done at the expense of introducing speech distortion. In [7] several supplementary schemes such as spectral smoothing and formant filtering are proposed to improve the performance of the spectral subtraction. In [5] a method is proposed to reduce the musical noise in silence and unvoiced region by dividing each silence and unvoiced frame of spectral subtracted speech into several sub-frames and randomizing the phases of each sub-frame over a uniform interval. Virag [8] proposed a technique based on the masking properties of the human auditory system on the assumption that additive noise is inaudible to the human ear as long as it falls below some masking threshold. In this method enhancement is achieved in two stages, in the first stage conventional spectral subtraction is performed to get an estimate of the masking threshold

and in second stage noise masking threshold is used to adjust the over subtraction factor and spectral floor parameter proposed in [6]. In [9] a method is proposed by using masking property and wavelet transform based on critical band decomposition which converts a noisy signal into wavelet coefficients and enhancement is achieved by subtracting the threshold from the noisy wavelet coefficients. The threshold is estimated from the noise masking threshold and segmental SNR. The performance of these methods is largely dependent on accurate estimation of masking threshold in noise. A parametric formulation of the generalized spectral subtraction method is proposed in [10] to improve the noise suppression performance of the spectral subtraction method, in which two short time spectral amplitude estimators of the speech signal are derived and optimized by minimizing the mean square error between the original spectrum and the parametric spectrum.

Ephraim and Malah [11, 12] proposed a gain function based MMSE STSA estimator based on *priori* and *posteriori* SNRs. This estimator is derived based on the assumption that speech and noise may be modelled as independent, zero-mean Gaussian random variables. Cappe [13] suggested a method to eliminate the musical noise of MMSE STSA estimator by incorporating the nonlinear smoothing procedure to estimate the *priori* SNR when the SNR is low. The enhancement results of these methods mainly depend on estimation of *priori* SNR. Some of *priori* SNR estimation techniques [14-16] are also proposed to improve the performance of these algorithms.

Another particular class of noisy speech enhancement methods is the signal subspace approach [17]. The main principle of this approach is each vector of noisy speech is composed of a signal plus noise subspace or simply signal subspace and the noise subspace. The noise subspace contains signal from noise only. Enhancement is achieved by removing the noise subspace and estimating the enhanced signal from the remaining signal subspace. The decomposition of the noisy signal into a signal subspace and a noise subspace can be done using the singular value decomposition (SVD) [18], the Eigen value decomposition (EVD) [19] or Karhuen-Loeve transform (KLT) [20]. The performance of the signal subspace method is further improved by using signal/ KLT approach proposed in [21]. In this method a different enhancement method is used for frames with different segmental SNR. The noisy speech frame is classified into speech dominated frame and noise dominated frames. The signal KLT matrix is used for speech dominated frame and a noise KLT matrix is used for noise dominated frames. In [22] the authors have proposed a subspace method by using the masking properties of human ear. The perceptual based Eigen filter is designed by using the frequency to Eigen domain transformation (FET) to reduce the residual noise effect. A difficult task for these methods is to accurately determine the dimension of the subspaces in the presence of non-stationary noise. The main drawback of signal subspace approach is it requires lot of computational load.

Most of the studies on the speech enhancement discussed above focus on enhancement based on suppression of noise. These methods disturb the spectral balance in speech, resulting in unpleasant distortions in the enhanced speech. In [23] Yegnanarayana *et al*. proposed an enhancement method by exploiting the characteristics of source signal such as Linear Prediction (LP) residual. The basic idea is to derive a weight function from the residual signal, which will reduce the energy in the low SNR regions relative to the high SNR regions of the LP residual of 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. The advantage of this method is that no explicit knowledge of noise characteristics is required. Since there is no direct spectral manipulation is involved, this method does not produce the type of distortions which the spectral subtraction produces. The perceptual quality of this method may be further improved by a better estimation of vocal tract characteristics.

Some of multi-channel based enhancement methods are also proposed for processing noisy speech. In [24] authors proposed a method which makes use of two or more input channels containing correlated signal components but uncorrelated noise components. The various input signals need not be of the same

shape, since adaptive enhancer filters the input before summing them. The output is a best least square estimate of the underlying signal in a chosen input channel. Srinivasan *et al* [25] proposed a parametric model based MMSE estimation of the clean speech signal using microphone array. The estimation is performed using an auto regressive (AR) model. Signal subspace based multi-channel enhancement technique is proposed in [26] by combining both frequency and spatial characteristics of speech and noise. Recently, Nagata et *al* [27] proposed two- channel speech enhancement that is based on auto gain control to eliminate musical noise by using the proper gain.

3. PROPOSED METHOD FOR ENHANCE-MENT OF NOISY SPEECH

The motivation for the proposed method is derived from the observation of the limitations of conventional spectral subtraction method. The limitations of conventional spectral subtraction method are it produces undesirable musical noise, not suitable for non-stationary environments and also not suitable for colored noise case. The proposed method aims to address the first two limitations. Even though similar approaches may be present in the literature, the novelty of the present work lies in developing new methods for speech/ non-speech detection in case of noisy speech and also modifying the spectrum of spectral subtracted speech for reducing musical noise. The proposed method may be termed as noise components subtraction and speech components enhancement method. In [28] we showed that kurtosis and energy values can be used to separate the speech and non-speech regions of noisy speech. In the proposed method first a weight function is derived from the kurtosis and energy values to separate the speech and non-speech regions. The noise components in non-speech regions are attenuated with the help of the weight function. In speech regions, enhancement of speech components is achieved in two stages. In first stage spectral subtraction is performed to remove the noise components and in second stage speech spectral components are further enhanced by adding the spectral contents of pitch and harmonic instants to the spectral subtracted speech spectrum by determining pitch of noisy speech.

The weight function for separating speech and non-speech regions is derived by computing kurtosis and energy values for short segments of speech (frame of 20 ms duration with a frame shift of 10 ms) and each of these values are repeated for frame shift number of times to make length of kurtosis and energy values equal to length of speech signal. Then these values are smoothed and nonlinearly mapped to derive the kurtosis and energy weight functions. Final weight function is derived by considering the maximum of kurtosis and energy weight functions at every sample instant. For more details on this, readers are requested to refer [28]. The noisy speech is weighted by the weight function to attenuate noise components in non-speech regions.

At the second level the noise estimate is updated from the original noisy speech signal as we move forward through the non-speech regions. Whenever a speech region is encountered, spectral subtraction is performed using the most recent estimate of noise. This updating of noise estimate regularly in the proposed method makes it suitable for non-stationary environments.

The speech signal can be modelled as the result of convolution between excitation and vocal tract impulse response. The excitation is either a periodic train of impulses for voiced speech or random noise for unvoiced speech [29]. This shows that during the voiced speech interval major portion of speech energy is available at pitch and harmonic instants. The objective of the propose method is to enhance these pitch and harmonic instants in the spectral subtracted speech to reduce the musical noise effect. In this work pitch of the noisy speech is determined by using simplified inverse filter tracking (SIFT) algorithm [30] and the harmonic locations are derived from the estimated pitch information.

After obtaining the pitch, its harmonic frequencies are calculated. Maximum values near the pitch and harmonics frequencies are found. The amplitude spectrum of the desired speech components is constructed by sampling the spectral subtracted speech spectrum at pitch and harmonic instants. The pitch and harmonics are sampled by using the window function of type

$$w(n) = 2Xe^{-axsign(n)}$$

where, L is number of samples which corresponds to the pitch frequency and a is chosen as 0.5 in this study. The sampled spectrum is added with the spectral subtracted speech spectrum. The resultant speech spectra is recombined with the original noisy speech phase spectra and converted back to the time domain by an inverse Discrete Fourier Transform (DFT).

The spectrum of a frame of voiced portion of degraded speech is shown in Fig 1*a* and the determined pitch and harmonics instants are shown by '*'. After determining pitch and harmonic instants, at every instant the maximum value of the spectrum is searched in and around ±30Hz and the instant at which maximum value occurs is considered as the actual harmonic instant and this is indicated by vertical lines in Fig 1*a*. The spectrum of spectral subtracted speech is shown in Fig 1*b*. Fig 1*c* shows the window function used for sampling the spectrum. The sampled spectrum is added with the spectral subtracted speech spectrum and is shown in Fig. 1*d*. Enhanced spectral peaks may be observed at pitch and harmonic instants.

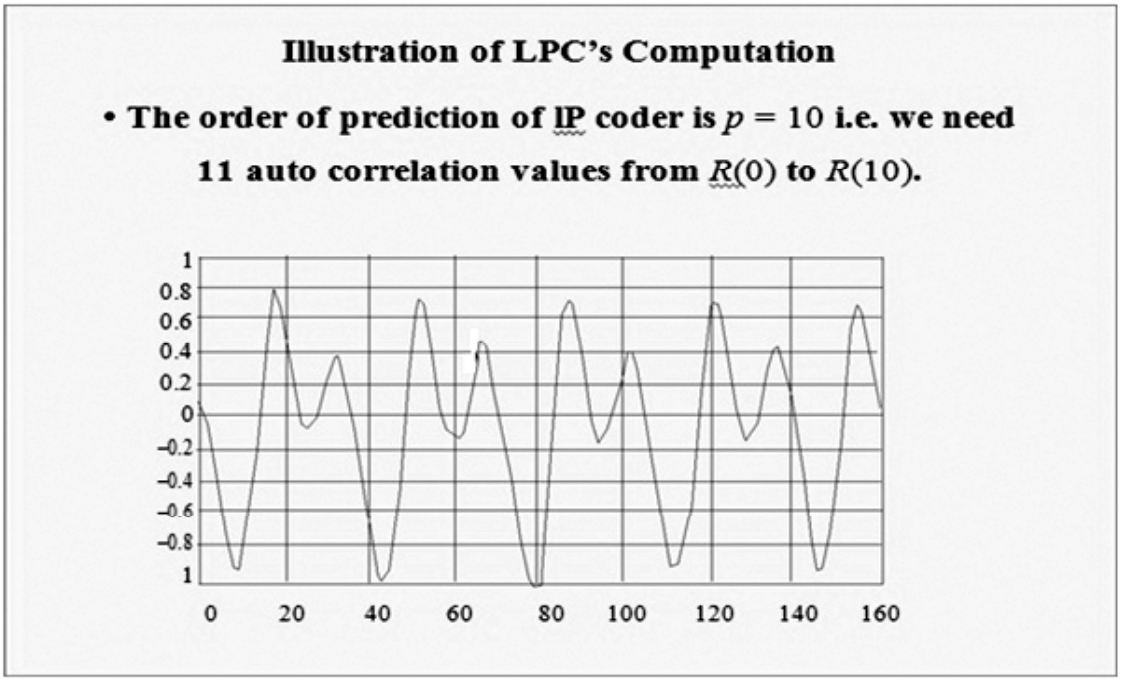

Figure 1: 20 mesc windowed voiced speech segment

The proposed method is tested for different types of data like speech signals from TIMIT database, speech recorded from the laboratory environment and for speech signals collected from the real world environment. The results of proposed method seem to be perceptually better compared to conventional spectral subtraction method. The speech signal spoken by a female speaker sampled at 8 kHz with a resolution of 16 bits/sample is taken from the TIMIT database is shown in Fig 2*a* and 3 dB stationary additive noise is added to it and is shown in Fig 2*b*. The smoothed and nonlinearly mapped kurtosis and energy values are shown in Fig 2*c* and *d*, respectively. The final weight function is shown in Fig 2*e*. The speech processed by the conventional spectral subtraction and the proposed method are shown in Fig 2*f* and *g*, respectively.

Figure 3*a* shows the clean speech signal spoken by a female speaker recorded in the laboratory. The non-stationary noise (additive nose of different noise levels at different times) is added to the clean speech signal and is shown in Fig 3*b*. The nonlinearly mapped kurtosis weight function, energy weight functions are shown in Fig 3*c* and *d*, respectively. Figure 3*e* shows the final weight function. The noisy speech signals enhanced directly by the spectral subtraction method and the proposed method are shown in Fig 3*f* and Fig 3*g*, respectively.

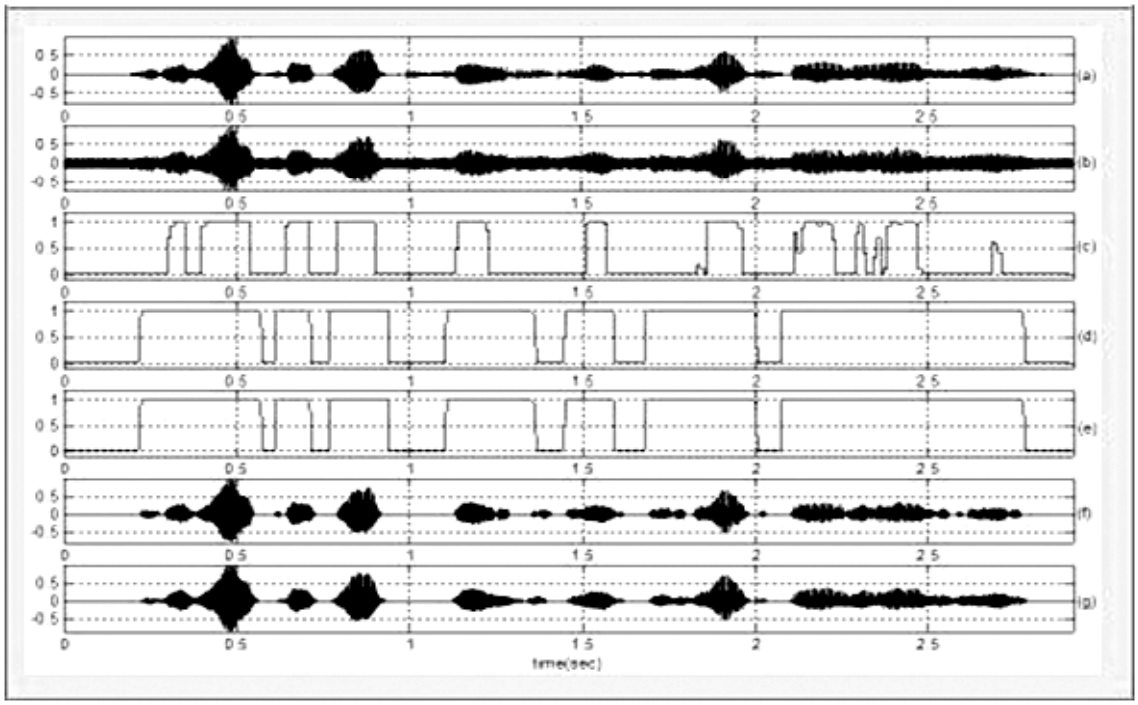

Figure 2: Results of enhancement of female voice degraded by stationary additive background noise. (a) clean speech, (b) degraded speech, (c) smoothed and nonlinearly mapped kurtosis, (d) smoothed and nonlinearly mapped energy, (e) weight function, (f) speech processed using spectral subtraction method, and (g) speech processed using proposed method.

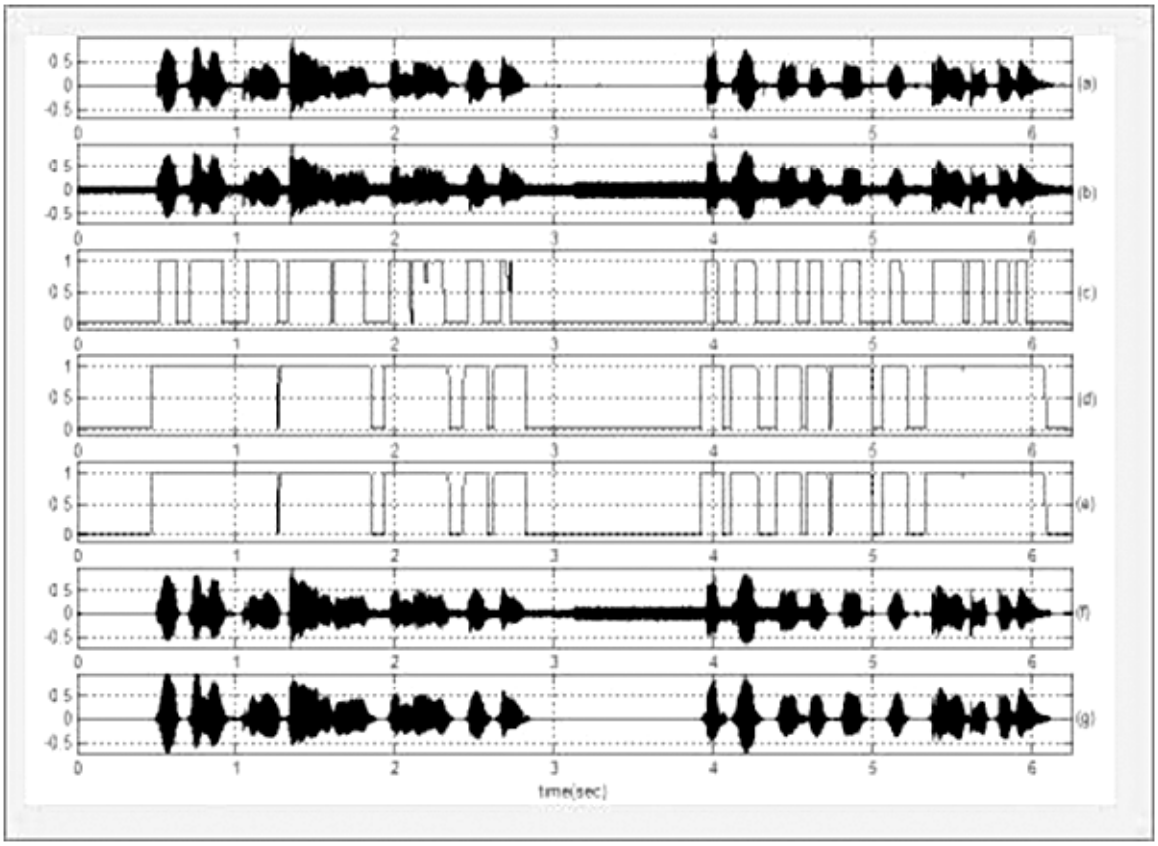

Figure 3:Results of enhancement of female voice degraded by non -stationary noise. (*a*) clean speech, (*b*) degraded speech, (*c*) smoothed and nonlinearly mapped kurtosis, (*d*) smoothed and nonlinearly mapped energy, (*e*) weight function, (*f*) speech processed using spectral subtraction method, and (*g*) speech processed using proposed method

4. SUMMARY

The concepts of major techniques proposed for enhancement of noisy speech are discussed in this paper. In noisy speech enhancement methods the subspace methods provide a mechanism to control the tradeoff between speech distortion and musical noise, but with the cost of a heavy computational load. Frequency domain methods, on the other hand, usually consume less computational resources, but do not have a theoretically established mechanism to control tradeoff between speech distortion and residual noise.

This paper also proposed a method for enhancement of noisy speech which is suitable for non-stationary environments. Further, in the proposed method the musical noise of spectral subtraction is reduced by enhancing the speech components such as pitch and harmonic amplitudes in the spectral subtracted speech spectrum.

The proposed method may be modified to make it suitable in case of colored noise case. Since the proposed method is a frequency domain approach, efforts may be made to combine existing time domain processing methods like LP residual manipulation for developing a robust speech enhancement method.

REFERENCES

[1] S R M Prasanna, Event based analysis of speech, *PhD dissertation*, Indian Institute of TechnologyMadras, Department of Computer Science and Engg., Chennai, India, March 2004.

[2] J Lim & A Oppenheim, Enhancement and bandwidth compression of noisy speech, *Proc IEEE*, vol. 67, pp. 1586-1604, Dec 1979.

[3] S. Boll, Suppression of acoustic noise in speech using spectral subtraction, *IEEE Tans Acoust.,Speech, Signal Processing*, vol ASSP-27, pp. 113-120, April 1979

[4] P Satyanarayana, Short segment analysis of speech for enhancement, *PhD dissertation*, Indian Insititute of Technology Madras, Department of Electrical Engg, Chennai, India, Feb 1999.

[5] J W Seok& K S Bae, Reduction of musical noise in spectral subtraction method using subframe phase randomisation, *IEEE Electronics Letters*, vol. 35, pp. 123-125, Jan 1999.

[6] M Berouti, R Schwartz & J Makhoul, Enhancement of speech corrupted by acoustic noise, in *Proc IEEEInternational Acoustics, Speech, and Signal Processing (ICASSP)*, Washington, DC, April 1979, pp. 208-211.

[7] H T Hu, F J Kuo& H J Wang, Supplementary schemes to spectral subtraction for speech enhancement, *Speech Communication*, vol. 36, pp. 205-218, March 2002.

[8] N Virag, Single channel speech enhancement based on masking properties of the human auditory system, *IEEE Trans Speech Audio Processing*, vol. 7, pp. 126-137, March 1999

[9] C T Lu & H C Wang, Enhancement of single channel speech based on masking property and wavelet transform, *Speech Communication*, vol. 41, pp. 409-427, Oct 2003. B L Sim, Y C Tong, J Chang & C T Tan, A parametric formulation of the generalized spectral subtraction method, *IEEE Trans Speech Audio Processing*, vol. 6, pp. 328-337, July 1998.

[10] Y Ephraim & D Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator, *IEEE Tans Acoust., Speech,Signal Processing*, vol ASSP-32, pp. 1109-1121, Dec1984.

[11] Y Ephraim & D Malah, Speech enhancement using a minimum mean square error log-spectral amplitude estimator, *IEEE Tans Acoust., Speech, SignalProcessing*, vol ASSP-33, pp. 443-445, April 1985.

[12] O Cappe, Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor, *IEEE Trans Speech Audio Processing*, vol. 2, pp. 345-349, April 1994.

[13] M Hasan, S Salahuddin& M Khan, A modified a priori SNR for speech enhancement using spectral subtraction rules, *IEEE Signal Processing Letters*, vol. 11, pp. 450-453, April 2004.

[14] K Yamashita & T Shimamura, Nonstationary noise estimation using low-frequency regions for spectral subtraction, *IEEE Signal Processing Letters*, vol. 12, pp. 465-468, June 2005.

[15] K Manohar & P Rao, Speech enhancement in nonstationary noise environments using noise properties, *Speech Communication*, vol. 48, pp. 96-109, Jan 2006.

[16] Y Ephraim and H V Trees, A signal subspace approach for speech enhancement, *IEEE TransSpeech Audio Processing*, vol. 3, pp. 251-266, July1995.

[17] S Jensen, P Hansen, S Hansen & J Sorensen, Reduction of broad-band noise in speech by truncated QSVD, *IEEE Trans Speech AudioProcessing*, vol. 3, pp. 439-448, Nov 1995.

[18] J Huang & Y Zhao, An energy-constrained signal subspace method for speech enhancement and recognition in colored noise, *SpeechCommunication*, vol 26, pp 165-181, Nov 1998.F Asano, S Hayamizu, T Yamada & S Nakamura, Signal/ noise KLT based approach for enhancing speech degraded by colored noise, *IEEE TransSpeech Audio Processing*, vol 8, pp. 159-167, March2000.

[19] A Rezayee and S Gazor, An adaptive KLT approach for speech enhancement, *IEEE Trans Speech AudioProcessing*, vol 9, pp. 87-95, Feb 2001.

[20] F Jabloun& B Champagne, Incorporating the human hearing properties in the signal subspace approach for speech enhancement, *IEEE Trans Speech AudioProcessing*, vol. 11, pp. 700-708, Nov 2003.

[21] B Yegnanarayana, C Avendano, H Hermansky& P Murthy, Speech enhancement using linear prediction residual, *Speech Communication*, vol. 28,

[22] E Ferrara & B Widrow, Multichannel adaptive filtering for signal enhancement, *IEEE TransAcoustics, Speech, and Signal Processing*, vol. 29, pp. 766-770, (1981).

[23] S Srinivasan, Aichner, W B Kleijn & W Kellermann, Multichannel parametric speech enhancement, *IEEESignal Processing Letters*, vol. 13, pp. 304-307, May 2006.

[24] S Doclo& M Moonen, GSVD-based optimal ltering for single and multimicrophone speech enhancement, *IEEE Trans Signal Processing*, vol. 50, pp. 2230-2244, Sept 2002.

[25] Y Nagata, T Fujioka & M Abe, Speech enhancement based on auto gain control, *IEEE Trans SpeechAudio Processing*, vol 14, pp. 177-190, Jan 2006.

[26] P Krishnamoorthy& S R M Prasanna, Modified spectral subtraction method for enhancement of noisy speech, in *Proc Third InternationalConference on Intelligent Sensing and Information Processing*, pp 146-150, Bangalore, India, 14-17 Dec 2005.

[27] J Deller, J Hansen & J Proakis, *Discrete TimeProcessing of Speech Signals*, 1st ed, IEEE Press, 1993.

[28] J Markel, The SIFT algorithm for fundamental frequency estimation, *IEEE Trans Audio andElectroacoustics*, vol. 20, pp. 367-377, Dec 1972.