Journal of Computer Science 7 (6): 941-948, 2011
ISSN 1549-3636

Speech Enhancement Algorithm Using Sub band Two Step Decision Directed Approach with Adaptive Weighting factor and Noise Masking Threshold

Deepa Dhanaskodi and Shanmugam Arumugam
Department of Electronics and Communication Engineering,
Bannari Amman Institute of Technology, TamilNadu, India

Abstract: Problem statement: Speech Enhancement plays an important role in any of the speech processing systems like speech recognition, mobile communication, hearing aid. **Approach:** In this work, human perceptual auditory masking effect is incorporated into the single channel speech enhancement algorithm. The algorithm is based on a criterion by which the audible noise may be masked rather than being attenuated and thereby reducing the chance of distortion to speech. The basic decision directed approach is for efficient reduction of musical noise, that includes the estimation of the a priori SNR which is a crucial parameter of the spectral gain, follows the a posteriori SNR with a delay of one frame in speech frames. In this work a simple adaptive speech enhancement technique, using an adaptive sigmoid type function to determine the weighting factor of the TSDD algorithm is employed based on a sub band approach. In turn the spectral estimate is used to obtain a perceptual gain factor. **Results:** Objective and subjective measures like SNR, MSE, IS distance and were obtained, which shows the ability of the proposed method for efficient enhancement of noisy speech **Conclusion/Recommendations:** Performance assessment shows that our proposal can achieve a more significant noise reduction and a better spectral estimation of weak speech spectral components from a noisy signal as compared to the conventional speech enhancement algorithm.

Key words: Decision directed approach, posteriori SNR, Mean Square Error (MSE), noise masking threshold, Signal-To Noise Ratio (SNR), Short-Time Spectral Amplitude (STSA), mobile communication, speech enhancement algorithm, proposed approach

INTRODUCTION

The need for enhancement of single-channel speech signal degraded by noise arises frequently, e.g., in mobile communication, Speech coding and hearing aid applications. Within single-channel speech enhancement, the noise is often assumed additive, i.e., y = s+d with y the noisy speech signal, s the clean speech signal and d the noise realization. Further, it is common to assume that the clean speech signal and the noise process are uncorrelated. Enhancement methods based on Short-Time Spectral Analysis (STSA) have received significant interest, partly due to their relatively good performance and low computational complexity. A vast amount of literature has been published on spectral-domain noise reduction algorithms for noisy speech signals, the best known being the Wiener filter and both the Short-Time Spectral Amplitude (STSA) and the Log-Spectral Amplitude (LSA) estimator. All of these approaches rely on an accurate estimate of the Signal-To Noise Ratio (SNR) in each frequency bin. Thus, given the noisy spectral components, besides the optimal estimator for the clean speech spectral components itself, also an estimator for the local SNR (in time and frequency) is required. A well-know approach is the decision-directed a priori SNR estimation (Cohen, 2004).

In the case of corruption by colored noise, utilizing noise masking properties to adapt a speech enhancement system is beneficial to result in less amounts of annoying residual noise. The performance of enhanced speech is characterized by a trade-off between the amount of noise reduction, the speech distortion and the level of musical residual noise. Among conventional approaches, utilizing Noise Masking Threshold (NMT) as a critical parameter to adjust either thresholds or gain factors with some unknown adjustment factors were proposed. In this study, a perceptual constraint is employed to optimize the gain factor for each sub band. It aims to keep the energy of residual noise lower than the NMT in an enhanced version. If the energy of residual noise is greater than the NMT in a sub band, the gain factor is modified and it becomes smaller to suppress the infecting

Corresponding Author: Deepa Dhanaskodi, Department of Electronics and Communication Engineering,
Bannari Amman Institute of Technology, TamilNadu, India

noise. However, if the energy of residual noise is smaller than the NMT, the corrupting noise cannot be perceived by the human ear. We do not need to change the gain factor for retaining the speech quality. The experimental results show that our approach can also work well in enhancing the noisy speech which is corrupted by various kinds of colored noise.

The proposed approach is to improve the estimated spectra of speech by the Two step decision-directed algorithm, enabling the noise masking threshold to be well estimated. Accordingly, the performance of perceptual gain factor is improved.

Basic principle of two step decision directed approach: In the classical noise model, the noisy speech is given by y(t) = s(t)+d(t). Let S(m,w), D(m,w) and X(m,w) designate the w spectral component of the short time frame p of the speech s(t), the noise d(t) and the noisy speech y(t), respectively. The quasi-stationarity of the speech is assumed over the duration of the analysis frame. The noise reduction process consists of the application of a spectral gain G(m,w) to each short time spectrum value Y(m,w). In practice, the spectral gain requires the evalution of two parameters. The a posteriori SNR is the first parameter given by:

$$\gamma_{post}(m,w)=\frac{|Y(m,w)|^2}{E\left\{|D(m,w)|^2\right\}} \quad (1)$$

where, E is the expectation operator. The a priori SNR, which is the second parameter of noise suppression rule is expressed as:

$$\hat{\gamma}_{priori}(m,w)=E\left\{\left|\hat{S}(m,w)\right|^2\right\}/E\left\{\left|\hat{D}(m,w)\right|^2\right\} \quad (2)$$

and requires the unknown information of the speech spectrum. As Plapous *et al*. (2004; 2006), the a priori SNR can be computed as follows:

$$\hat{\gamma}_{priori}(m,w)=\beta.\hat{\gamma}_{priori}(m-1,w)+(1-\beta).P\left[\gamma_{post}(m,w)-1\right] \quad (3)$$

where, P denotes the half-wave rectification and $\hat{\gamma}_{priori}(m-1,w)$ is the estimated SNR at previous frame and β is the smoothing constant. The estimator of the a priori SNR described in (3) (Cohen, 2004) corresponds to the so called decision directed approach. And the multiplicative gain function G(m,w) is obtained by:

$$G^{DD}(m,w)=\frac{\hat{\gamma}_{priori}(m,w)}{1+\hat{\gamma}_{priori}(m,w)} \quad (4)$$

In the second step, this gain is utilized to estimate the a priori SNR estimator for the frame m using (Plapous *et al*., 2006) the equation reduces to [two step noise]:

$$G^{TSDD}(m,w)=\frac{G^{DD}(m,w).\gamma_{post}(m,w)}{1+\gamma_{post}(m,w)} \quad (5)$$

The estimated spectra of speech signal for evaluating the noise masking threshold is given by:

$$\hat{S}(m,w)=G(m,w).Y(m,w) \quad (6)$$

MATERIALS AND METHODS

Separation of noisy speech signal into bands: Sub band approach is the frequency dependent processing of the noisy speech signal (Deepa and Shanmugam, 2010). From the literature this method has been found to offer better quality of the enhanced speech with reduced residual noise. This approach has been justified due to variation in signal to noise ratio across the speech spectrum. White Gaussian noise has a flat spectrum, where as the real world noise is not flat. The noise spectrum does not affect the speech signal uniformly over the whole spectrum; some frequencies are affected more adversely than others. To take into account the fact that colored noise affects the speech spectrum differently at various frequencies, we use this approach. The speech spectrum is divided into linear non-overlapping bands and the enhancement algorithm is performed independently in each band:

$$Y=[Y_1,Y_2.....Y_n] \quad (7)$$

where, Y represents the noisy speech signal and n represents the number of bands.

TSDD with Adaptive weighting factor: Figure 1 shows the block diagram of the proposed work. In the decision directed model we can see that $\hat{\gamma}_{priori}(m,w)$ is obtained by applying the weighting factor β to the estimate of the a posteriori SNR at the previous frame. Since the weighting factor β is generally chosen very close to 1 the speech spectra estimated in the previous frame are used to estimate the current a priori SNR and the a priori SNR follows the a posteriori SNR with a delay of one frame when the a posteriori SNR exhibits an abrupt increase. This delay is likely to produce

undesired gain distortion and thus generate audible distortion during abrupt transient periods. With a constant weighting factor the a priori SNR fails to follow the shape of the a posteriori SNR during abrupt transients. Thus an adaptive weighting factor would be necessary for the estimation of a priori SNR to follow the shape of the a posteriori SNR and in turn the speech which is shown in the Fig. 3 and 4. To cope with this situation (Chang and Park, 2007) proposed the adaptive weighting factor incorporating the adaptive sigmoid type function. Specifically, the adaptive value based on the sigmoid type function is applied to the weighting factor of the TSDD algorithm according to the transient of the a posteriori SNR:

$$\beta(m,w)=\frac{\eta\exp\left[-\lambda\left(\log\left(\frac{1}{|\Delta\gamma_{post}(m,w)|}\right)-k_o\right)\right]}{1+\exp\left[-\lambda\left(\log\left(\frac{1}{|\Delta\gamma_{post}(m,w)|}\right)-k_o\right)\right]}+\rho \quad (8)$$

Through the speech enhancement experiments (Rangachari *et al*., 2004), β is obtained using the offset k_0 = -9, the slope parameter λ=0.4, the constants η=-1.5 and ρ=0.99, respectively. Increasing $|\Delta\gamma_{post}(m,w)|$ results in decreasing $\log\left(\frac{1}{|\Delta\gamma_{post}(m,w)|}\right)$ and thus the weighting factor decreases. With this adaptive weighting factor according to the transient, the a priori SNR can appropriately follow the shape of the original speech signal during transients.

The employed weighting factor using the sigmoid type function is plotted in Fig. 2.

Thus the scheme efficiently avoids the limitation of the TSDD algorithm with constant weighting factor and the shape of the proposed a priori SNR estimator resembles the a posteriori SNR during speech onset periods. Thus the a priori SNR is more accurate than the conventional a priori SNR and thus the performance of the enhancement algorithm can be improved.

Speech enhancement based on masking properties: Masking is present because the auditory system is incapable of distinguishing two signals close in the time or frequency domain. This is manifested by an elevation of the minimum threshold of audibility due to a masker signal. Noise masking is a well-known psycho acoustical property of the auditory systems that has already been applied with success to speech and audio coding in order to partially or totally mask the distortion introduced in the coding process (Scalart and Filho, 1996; Ramadan, 2008).

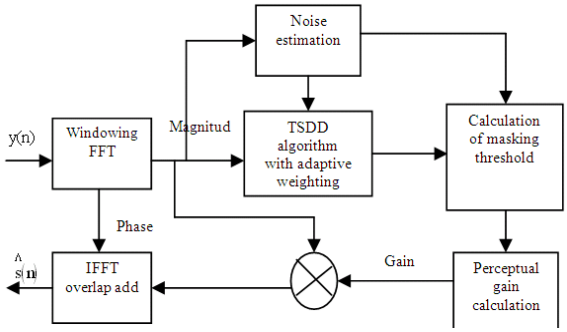

Fig. 1: Block diagram of the proposed speech enhancement method

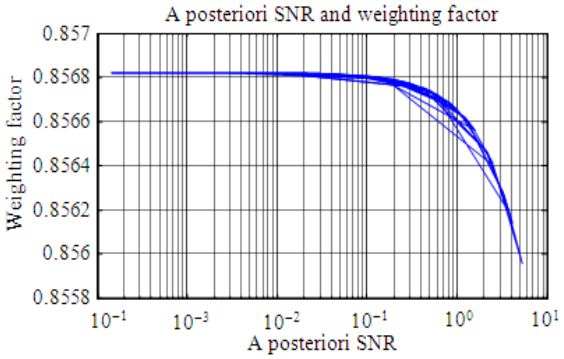

Fig. 2: A posteriori SNR and weighting factor

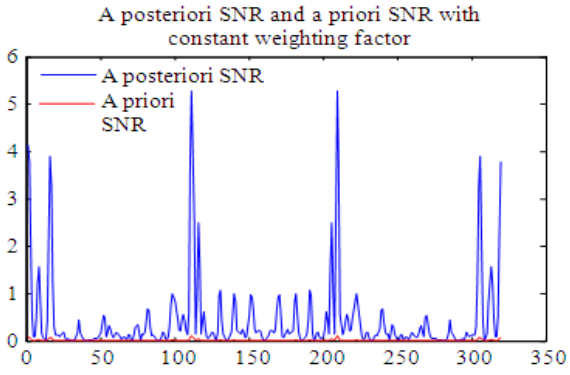

Fig. 3: A posteriori and A priori SNR with constant weighting factor

This study only considers the frequency domain masking, or simultaneous masking: A weak signal is made inaudible by a stronger signal occurring simultaneously. This phenomenon is modeled via a noise masking threshold, below which all components are inaudible.

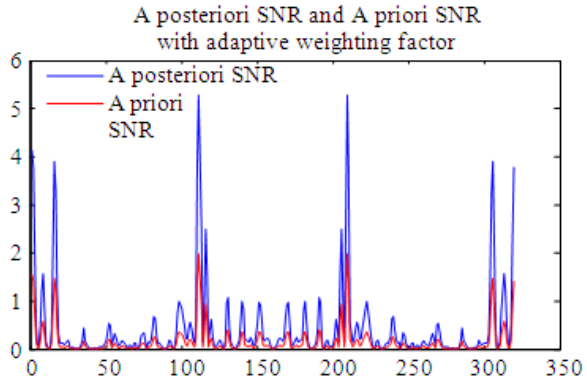

Fig. 4: A posteriori and A priori SNR with constant weighting factor

Critical band analysis: The energies of each critical band are added up. Frequencies Within the same critical band are equally perceived by the human ear. The spectrum is partitioned into critical bands according to (Virag, 1999) and the energy is summed in each band:

$$B_i = \sum_{w=l}^{w=h} \Lambda S(m,w) \tag{9}$$

Where:
l = Lower boundary of the critical band
h = Upper boundary of the corresponding band
B = Energy of the band

Spreading function: The application of a spreading function allows taking into account the masking between signals in different critical bands. A convolution is performed with the spreading function, which operates on a bark scale:

$$C_i = SF_i * Bi \tag{10}$$

The value C_i denotes the spread critical band spectrum A convenient analytical expression of the spreading function has been proposed by Schroeder *et al*. and is given by:

$$SF_i = 15.81 + 7.5(i + 0.474) - 17.5\sqrt{1 + (i + 0.474)^2}\,dB \tag{11}$$

where the frequency variable k is in barks and SF_i in dB.

Calculation of noise masking threshold: In order to determine the noiselike or tonelike nature of the signal, the Spectral Flatness Measure (SFM) is used. The SFM is defined as the ratio of the geometric mean of the power spectrum to the arithmetic mean of the power spectrum. In this use, the SFM is converted to decibels:

$$SFM_{dB} = 10\log_{10}\frac{G_m}{A_m} \tag{12}$$

And further it is used to generate a coefficient of tonality as follows:

$$\tau = \min\left(\frac{SFM_{dB}}{SFM_{dB\,max}}, 1\right) \tag{13}$$

The offset O_i in decibels for the masking energy in each band i is then set as:

$$O_i = \tau(14.5 + i) + (1 - \tau)5.5 \tag{14}$$

In other words the index τ is used to weight geometrically the two threshold offsets, 14.5+i dB for tone masking noise and 5.5 dB for noise masking tone. The threshold offset is then subtracted from the spread critical band spectrum to yield the spread threshold estimate:

$$T_i = 10^{\log 10(C_i) - \left(\frac{Oi}{10}\right)} \tag{15}$$

The simultaneous masking threshold is compared with the absolute-hearing threshold which is frequency dependent and can be closely approximated by the expression:

$$AHT(f) = 3.64f^{-0.8} - 6.5e^{-0.6(f-3.3)^2} + 0.001f^4\,[dB] \tag{16}$$

with f in Kilohertz.

Finally the noise masking threshold Th(ε) is determined by:\

$$\begin{aligned} Th(\varepsilon) &= \max\{AHT(f), Th(\varepsilon)\}Th(\varepsilon) \\ &= \max\{AHT(f), Th(\varepsilon)\} \end{aligned} \tag{17}$$

where, f is chosen to be the central frequency of the εth Critical band. The perceptual gain factor $G_{per}(m,w)$ in the frequency domain can be derived as:

$$G_{per}(m,w) = \frac{1}{1 + \max\left(\sqrt{\frac{|D(m,w)|^2}{Th(\varepsilon)}} - 1, 0\right)} \tag{18}$$

Finally the enhanced speech spectrum can be obtained as:

$$\tilde{S}_{enh}(m,w) = G_{per}(m,w).Y(m,w) \quad (19)$$

Performance measures:

Objective measures for performance evaluation: Objective measures are based on a mathematical comparison of the original and processed speech signals. It is desired that the objective measures be consistent with the judgement of the human perception of speech (Hu and Loizou, 2008). The Signal-to-Noise Ratio (SNR), Itakura-Saito (IS) measure and Mean Square Error (MSE) are the three most widely used objective measures.

Signal-to-Noise Ratio (SNR): The SNR is a popular method to measure speech quality. As the name suggests, it is a calculated as the ratio of the signal to noise power in decibels. If the summation is performed over the whole signal length, the operation is called global SNR:

$$SNR_{dB} = 10\log_{10}\left(\frac{\sum_{n} S^2(n)}{\sum_{n}\left[S(n)-\hat{S}(n)\right]^2}\right) \quad (20)$$

Itakura-saito distance: The Itakura–Saito distance is a measure of the perceptual difference between an original spectrum and an approximation of that spectrum. Lower the IS distance, better will be the quality of speech (i.e., minimum phase difference between clean and enhanced signal). The average Itakura-Saito measure across all speech frames of the given sentence will be computed to evaluate the spectral noise subtraction algorithm:

$$D_{IS}\left(P(w),\hat{P}(w)\right) = \frac{1}{2\Pi}\int_{-\Pi}^{\Pi}\left[\frac{P(w)}{\hat{P}(w)} - \log\frac{P(w)}{\hat{P}(w)} - 1\right]dw \quad (21)$$

Where:

$P(w)$ = Clean spectrum

$\hat{P}(w)$ = Estimated spectrum

Mean square error: The Mean Square Error (MSE) metric is frequently used in signal processing and is defined as:

$$MSE = \frac{1}{L}\sum_{i=1}^{L}\left(S(i)-\hat{S}(i)\right)^2 \quad (22)$$

where, S(i) denotes the power spectrum of the clean speech signal and $\hat{s}(i)$ denotes the power spectrum of the enhanced speech signal.

Subjective measures for performance evaluation: The subjective quality ratings were obtained using the ITU-TP.835 methodology designed to evaluate the speech quality along three dimensions: Signal distortion, noise distortion and overall quality:

- The speech signal alone using a five-point scale of signal distortion (SIG)
  5-Very natural, no degradation ;4-Fairly natural, little degradation ;3-Somewhat natural, somewhat degraded ; 2-Fairly unnatural, fairly degraded ;1-Very unnatural, very degraded
- The background noise alone using a five-point scale of background intrusiveness (BAK)
  5-Not noticeable ;4-Somewhat noticeable ;3-Noticeable but not intrusive ;2-Fairly conspicuous, somewhat intrusive ;1-Very conspicuous, very intrusive
- The overall effect using the scale of the Mean Opinion Score (OVRL)
  1-bad; 2-poor;3-fair; 4-good; 5-excellent

RESULTS

The test samples are taken from Speech Enhancement Assessment Resource (SPEAR) database of Center for Spoken Language Understanding (CSLU) and the NOIZEUS database. The sampling rate of the noisy speech samples is 8 kHz. And algorithm simulation has been carried out with MATLAB. In the speech enhancement process various noise types have been considered for various speech samples. To evaluate the performance of the proposed method, the noisy samples have been processed as frames of N = 160 samples (i.e., 20 msec of speech) with 40% overlap between successive frames. The estimated frames are overlap-added using Hamming window.

Time domain plot:

Inference: Figure 5 shows that the time plot of the enhanced speech signal using a constant weighting factor results in speech distortions whereas the use of an adaptive weighting factor preserves the speech components.

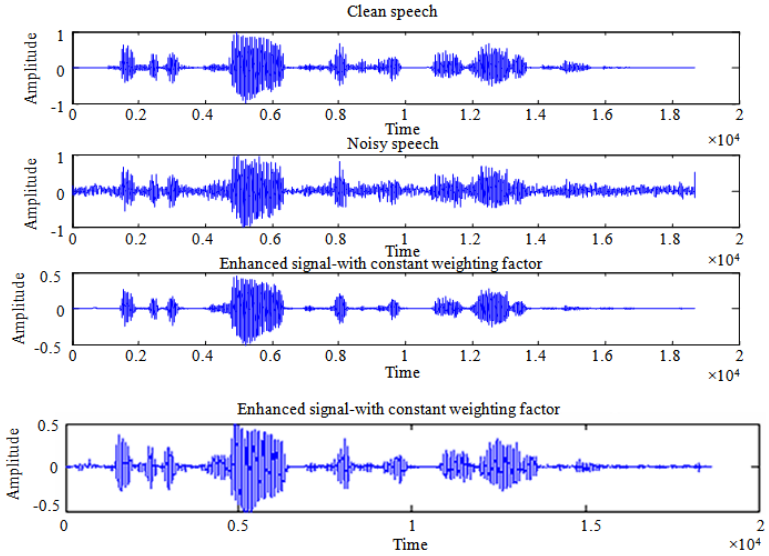

Fig. 5: Time plot: speech signal, speech signal corrupted by 4db-cockpitnoise, enhanced signal with conventional method, signal enhanced with proposed method

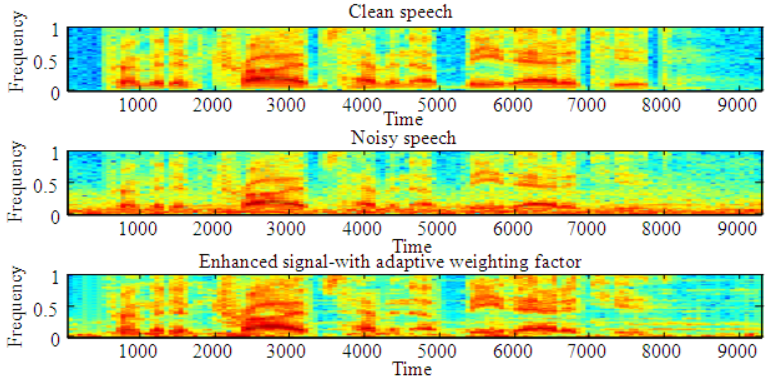

Fig. 6: Spectrogram plot: speech signal, speech signal corrupted by 5db babble noise, signal enhanced using the proposed method

Spectrogram plot:

Inference: It can be observed that the plot of the enhanced speech signal is very close to the clean speech signal and the noise components are significantly removed from the noisy speech.

Comparison of MSE and IS Distance values obtained With Conventional and Proposed Method:

Inference: From the plot it is obvious that the enhancement has not induced any distortions to the signal and the Mean Square Error and Itakura Saito distance values are minimum compared to the conventional method.

D. Comparison of SNR and Subjective measures for proposed and existing method.

Table 1 gives the MOS listening test results which confirm that our proposed masking-based enhancement method leads to the best performance for human listeners.

The values of SNR and IS distance of the conventional method and the proposed method are tabulated in Table 1 and 2. Thus it becomes clear that the proposed method performs better in terns of SNR compared to the conventional decision directed approach. Figure 6 shows the spectrogram comparison and Fig. 7 and 8 shows the MSE and (IS) Itakura-Saito distortion measure for the different noisy speech signals. From these figures, we can see that the proposed method always outperforms the other methods.

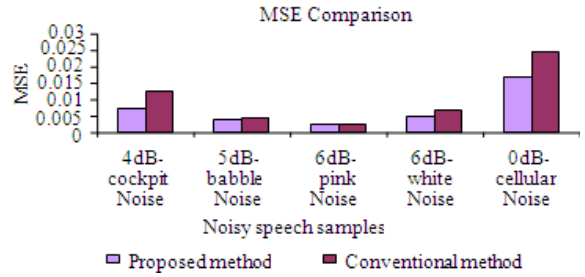

Fig. 7: MSE comparison

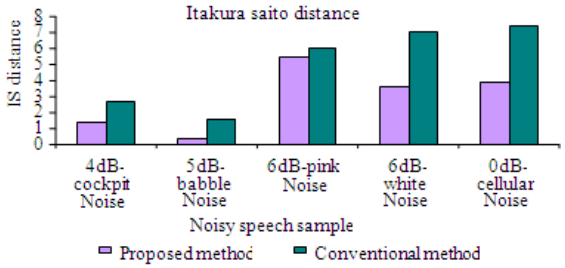

Fig. 8: IS distance comparison

Table 1: Subjective measure for the enhanced signal with sub band TSDD with adaptive noise estimation

Noisy speech sample	Enhancement method	SIG	BAK	OVRL
Babble noise-5dB	Conventional	3.3	2.42	2.86
	proposed	3.6	2.47	3.06
Pink Noise-6dB	Conventional	3.1	2.39	2.49
	proposed	3.22	2.67	2.76
Cockpit noise-4dB	Conventional	2.52	2.06	2.94
	proposed	3.73	3.12	3.49
Cellular noise-0dB	Conventional	2.11	2.69	2.14
	proposed	3.03	3.14	2.96
White noise-6dB	Conventional	2.56	1.43	2.69
	proposed	3.39	2.96	3.13

Table 2: Comparison of SNR values for conventional and proposed method

Noisy speech samples	Enhancement method	SNR
Babble noise-5dB	Conventional	11.27
	proposed	12.53
Pink noise-6dB	Conventional	13.80
	proposed	17.18
Cockpit noise-4dB	Conventional	13.73
	proposed	22.94
Cellular noise-0dB	Conventional	4.53
	proposed	6.29
White noise-6dB	Conventional	10.22
	proposed	11.84

DISCUSSION

From the above results shown, it is found that the proposed method of sub band approach with adaptive weighting factor and Noise masking threshold gives better results in terms of increased SNR, decreased MSE and with minimum IS distance. The subjective measures also improved compare to the conventional method.

CONCLUSION

In this study, a method that combines the Sub band Two step decision directed approach with adaptive noise estimation and perceptual properties of the human auditory system is proposed. Based on the perceptual properties of the human auditory system, a noise masking threshold is identified which masks the noise rather than attenuating it and thus overcomes the speech distortion problem existing in various algorithms and method is verified by computer simulations. From the simulations results, it is shown that the proposed algorithm outperforms the existing method. In addition, it possesses a good tradeoff in minimizing both speech distortion and residual noise simultaneously. This characteristic is especially useful for the case of weak spectral components of speech signal that is corrupted by noise. With such properties this method can be used as a preprocessing technique in speech recognition, Speech Coding, Mobile Communication and Hearing aid Applications.

REFERENCES

Chang, J.H. and Y.S. Park, 2007. A novel approach to a robust a priori SNR estimator in speech enhancement. IEICE Trans. Commun., 90: 2182-2185. http://search.ieice.org/bin/summary.php?id=e90-b_8_2182

Cohen, I., 2004. On the decision directed approach of ephraim and malah. Proceedings of IEEE International Conference, May 17-21, pp: 293-296. DOI: 10.1109/ICASSP.2004.1325980

Deepa, D. and A. Shanmugam, 2010. Analysis of subband speech enhancement technique for digital hearing aids. Int. J. Comput. Network Secur., 2: 132-135. http://www.doaj.org/doaj?func=abstract&id=653071

Hu, Y. and P.C. Loizou, 2008. Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang. Proc., 16: 229-238. DOI: 10.1109/TASL.2007.911054

Plapous, C., C. Marro and P. Scalart, 2006. Improved signal-to-noise ratio estimation for speech enhancement. IEEE Trans. Audio, Speech Lang. Proc., 14: 2098-2108. DOI: 10.1109/TASL.2006.872621

Plapous, C., C. Marro, L. Mauuary and P. Scalart, 2004. A two-step noise reduction technique. Proceedings of IEEE International Conference, May 17-21, pp: 289-292. DOI: 10.1109/ICASSP.2004.1325979

Ramadan, Z.M., 2008. A three-microphone adaptive noise canceller for minimizing reverberation and signal distortion. Am. J. Applied Sci., 5: 320-327. DOI: 10.3844/ajassp.2008.320.327

Rangachari, S., P.C. Loizou and Y. Hu, 2004. A noise estimation algorithm with rapid adaptation for highly non-stationary environments. Proceedings of IEEE International Conference, May 17-21, Richardson, TX, USA., pp: 305-308. DOI: 10.1109/ICASSP.2004.1325983

Scalart, S. and J.V. Filho, 1996. Speech enhancement based on a priori signal to noise estimation. Proceedings of the IEEE International Conference Acoustics Speech Signal Processing, (ICASSP'96), IEEE Computer Society Washington, DC, USA., pp: 629-632. DOI: 10.1109/ICASSP.1996.543199

Virag, N., 1999. Single channel speech enhancement based on masking properties of human auditory system. IEEE Trans. Speech Audio Proc., 7: 126-137. DOI: 10.1109/89.748118