Noise Suppression Spectral Processing Methods for Degraded Speech Signal for Speech Enhancement

Zeeshan Hashmi Khateeb
Student, M.Tech 4th Semester,
Department of Instrumentation Technology,
Dayananda Sagar College of Engineering,
Bangalore, India

Gopalaiah
Assistant Professor,
Department of Instrumentation Technology,
Dayananda Sagar College of Engineering,
Bangalore, India

***Abstract*— Assessment of clean speech from a noisy speech signal has been a research topic for a long time. This research finds its variety of applications, which includes the present mobile communication also. The most important outcome of this research is the improved quality and reduced listening effort in the presence of an interfering noise signal. In this paper the performance of various noise reduction techniques namely spectral subtraction, wavelet transforms, iterative subtraction, MMSE and Wiener filtering is done. This paper proposes a time-frequency estimator for enhancement of noisy speech signals in the discrete frequency transform domain. In the proposed method the estimation is based on modeling and filtering frequency components of noisy speech signal using Kalman filters. Experimental outcome show that the proposed method provides the better performance as compared to the other Spectral processing approaches.**

Keywords: MMSE, Wiener, Autoregressive and Kalman filter

I. INTRODUCTION

Speech signals in the real world are often corrupted by various types of degradations. Degraded speech is poor in terms of perceptual quality and intelligibility Poor perceptual quality leads to listener fatigue. Poor intelligibility leads to degraded performance in tasks like speech and speaker recognition.

The two basic methods for speech processing are Spectral Processing and temporal processing. In the spectral processing method, degraded speech is processed in frequency domain. In temporal processing method the processing is done time domain. Even though Spectral Processing method leads to artificial distortion due to nonlinear signal processing, leading to a serious deterioration of sound quality, Spectral processing methods are preferred for noise reduction mainly because of their simplicity and effectiveness. Most of the spectral processing techniques relay on the basis that the human speech perception is not sensitive to short–time phase. This is exploited in most of the methods, where only the spectral magnitude associated with original signal is estimated. This algorithm and its advanced versions have been applied to single-channel and multi-channel speech enhancement in speech recognition systems, speech coders, digital hearing-aids, voice activity detectors, and hands-free mobile communication systems. In temporal analysis, speech enhancement is done by exploiting the characteristics of excitation source signal such as LP residual. The basic approach in this method is to identify the high SNR portions in the noisy speech signal and enhance those portions relative to the low SNR portions, without causing significant distortion in the enhanced speech.

In case of noisy speech, the spectral processing methods can be grouped into non parametric and statistical model-based methods. Methods from the first category remove an estimate of the degradation from the noisy features, such as subtractive type algorithms and wavelet de-noising. The statistical model based, such as MMSE estimator uses the parametric model of the signal generation process. Recent studies have focused on a non-linear approach to the subtraction procedure.

To deal with the musical noise difficulty, there have been many investigation of musical noise production in nonlinear signal processing, and methods for its improvement have been proposed. Such conventional musical-noise mitigation methods are, unluckily, designed to reduce musical noise generation at the cost of degrading the noise reduction performance. To achieve both high noise reduction performance and low musical noise generation, an *iterative spectral subtraction* method has recently been proposed. This method is performed through signal processing in which *weak* spectral subtraction processes are iteratively applied to the input signal. The methodology used in iterative spectral subtraction is of great interest to researchers on nonlinear signal processing and machine learning, because it addresses the inherent question of whether or not recursive weak (nonlinear) signal processing can provide better performances.

II. SPECTRAL SUBTRACTION

Spectral Processing methods are the most popular techniques for noise reduction, mainly of their simplicity and effectiveness. Most of the spectral processing techniques [14]-[15] relay on the basis that the human speech perception is not sensitive to short –time phase. This is exploited in these methods, where only the spectral magnitude associated with original signal is estimated .In case of noisy speech; the spectral processing methods can be grouped into non parametric and statistical model-based methods. Methods from the first category remove an estimate of the degradation from the noisy features, such as subtractive type algorithms and wavelet de-noising.

Speech which is “contaminated” by noise can be expressed as y(n) = s(n) + d(n) where x(n) is the speech with noise, s(n) is the “clean” speech signal and d(n) is the noise process, all in the discrete time domain. What spectral subtraction attempts to do is to estimate s (n) from y (n). For a noisy signal, power spectrum is given by

$$| Y(k) |^2 \approx | S(k) |^2 + | D(k) |^2 \qquad (1)$$

The DFT of the *Y (k)* is given by

$$Y(k) = \sum_{n=0}^{N-1} y(n)\, e^{-j\frac{2\pi k n}{N}} = | Y(k) |\, e^{j\varphi(k)} \qquad (2)$$

We cannot calculate noise spectrum *D(k) directly,* hence power spectrum *D(k)* is estimated in time-average scale. Considering that noise is un-correlated with the speech signal, an estimation of the modified speech spectrum is given by

$$| \hat{S}(k) |^2 = | Y(k) |^2 - | \hat{D}(k) |^2 \qquad (3)$$

Spectral Subtraction is accomplished by subtracting the average estimate of the noise from the direct speech spectrum. Errors in computation produce few negative values in the modified spectrum. In order to overcome these negative values *half-wave rectification* process is carried out, where negative values are set to *zero*. With half-wave rectification the modified spectrum can be written as:

$$| \hat{S}(k) |^2 = \begin{cases} | \hat{S}(k) |^2 & if \quad | \hat{S}(k) |^2 > 0 \\ 0 & else \end{cases} \qquad (4)$$

The modified spectrum is combined with the phase information from the noise corrupted signal to reconstruct the time signal by using the Inverse Discrete Fourier Transform (IDFT)

$$\hat{s}(n) = IDFT\left(| \hat{S}(k) |\; e^{j\varphi(k)} \right) \qquad (5)$$

Although spectral subtraction scheme provides an enhancement in terms of noise attenuation, it often produces a new randomly variable type of noise, referred to as *musical noise*. The characteristics of this musical noise have a close resemblance with tone signals, leading to false peaks in the processed spectrum. When the enhanced signal is reconstructed in the time-domain, these peaks result in short sinusoids whose frequencies vary from frame to frame. Even though musical noise is different from the original noise, it leads to disturbance. Speech signal with musical noise has lower perceived quality and lower information content, than the original noisy signal. Lot of research at the present time is focused on ways to overcome the problem of musical noise. It is more or less impossible to decrease musical noise without affecting the speech, and hence there is a tradeoff between the speech distortion and amount of noise reduction. Several modifications for the standard spectral subtraction method have been proposed to alleviate the speech distortion introduced by the spectral subtraction process.

Fig. 1 shows a block diagram of the spectral subtraction method. The extent of the subtraction can be varied by applying a scaling factor α. The values of scaling factor α higher than 1 result in high SNR level of de-noised signal, but too high values may cause distortion in perceived speech quality. After subtraction, the spectral magnitude is not guaranteed to be positive; in order to remove negative values half-wave rectification is used. This introduces musical tone artifacts in the processed signal. Full wave rectification avoids the creation of musical noise, but less effective at reducing noise. After subtraction, *a* root of the Y_k *(w) is* extracted to provide corresponding Fourier amplitude components. An inverse Fourier transform, using phase components directly from Fourier transform unit, and overlap add is then done to reconstruct the speech estimate in the time-domain.

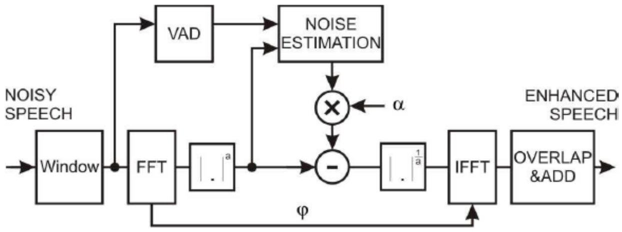

Fig 1: General Spectral Subtraction methodology.

Experimental results for spectral subtractions are shown in fig 4 and fig 5.Input and output SNR are tabulated in Table 1. SNR improvement is more for 5db has compared with other inputs, i.e. spectral subtraction provides a good performance for lower SNR.

To achieve both high noise reduction performance and low musical noise generation, an *iterative spectral subtraction* method shown in fig 2 has recently been proposed. This method is performed through signal processing in which *weak* spectral subtraction processes are iteratively applied to the input signal. The methodology used in iterative spectral subtraction is of great interest to researchers on nonlinear signal processing and machine learning, because it addresses the inherent question of whether or not recursive weak (nonlinear) signal processing can provide better performances.

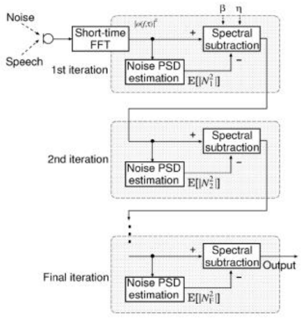

Fig 2: Iterative spectral subtraction

The experimental results for iterative spectral subtraction are shown in fig 6 and 7. Input and output SNR are tabulated in Table 2.It is observed that its performance is comparably good with respect to spectral subtraction for lesser SNR. Other observations made with respect to male and female voice in spectral subtraction hold well in iterative also.

III. MMSE

In spectral subtraction based methods, there were no specific assumptions made about the distribution of the spectral components of either speech or noise. Ephraim and Malah [16] have proposed a system that utilizes the MMSE-short time spectral components of speech and noise signals. MMSE-STSA estimators for speech enhancement aim to minimize *mean square error* between the short time spectral magnitude of the clean and enhanced speech signal. Gaussian statistical model is considered in MMSE-STSA which is, statistically independent and has zero mean value. Spectral gain by MMSE is given by.

$$G(k) = \Gamma(1.5)\frac{\sqrt{(v_k)}}{\gamma_k}\exp\left(-\frac{v_k}{2}\right)\cdot\left[(1+v_k)I_0\left(\frac{v_k}{2}\right)+v_kI_1\left(\frac{v_k}{2}\right)\right], \qquad (6)$$

Where Γ (•) denotes the gamma function, I_0 and I_1 denote the modified Bessel functions of zero and first order, respectively. Also functions in the equation are defined by

$$v_k \triangleq \frac{\xi_k}{1+\xi_k}\gamma_k.$$
$$\xi_k \triangleq \frac{\lambda_s(k)}{\lambda_d(k)},$$
$$\gamma_k \triangleq \frac{R_k^2}{\lambda_d(k)}, \qquad (7)$$

To incorporate perceptually significant information Ephraim and Malah proposed MMSE log-spectral amplitude (MMSE-LSA) estimator [17]. The previous paper of the authors [16] did not consider any of the nonlinear characteristics observable in human perception. To incorporate perceptually important information into the algorithm, the author projected a method [17][18] to minimize the mean square error between the logarithm of the STSA of the clean and enhanced speech i.e. LSA estimator minimizes

$$E\ \{(\log_e Y_k - \text{loge}\ \hat{Y}_k)^2\} \qquad \mathit{(8)}$$

Where Y_k denotes the spectral speech amplitude and $\hat{Y}_k$ is its optimal estimator. This measure of optimality gives high-quality results, with evident reduction in musical noise.

A fundamental assumption made in the MMSE algorithms is that the real and imaginary parts of the clean DFT coefficients can be modeled by a Gaussian distribution. The Gaussian assumption might hold for the DFT coefficients of the noise, typically estimated using relatively short (20-30ms) duration window. Based on this observation, a similar optimal MMSE-STSA estimator using non-Gaussian distributions is proposed. In particular the Gamma or the Laplacian probability distributions are used to model the distributions of the real and imaginary parts of the DFT coefficients.

Erkelens [19]-[21] proposed a method for MMSE estimation of DFT coefficients with general gamma priors, the experimental results for this is shown in fig 7 and 8. Input and output SNR are tabulated in Table 3.It is observed that its performance is better than other two methods discussed above .As similar to that of spectral subtraction MMSE shows good performance for lower SNR, where percentage of SNR improvement is more than 100% for car noise and more than 50% for airport and restaurant noise. Also MMSE shows fine Performance for male voice.

IV. WAVELET DE NOISING

Most of the speech enhancements algorithms are applied in the frequency domain, using short time Fourier transform which allows analyzing non-stationary speech signals.STFT provides a compromise between time resolution and frequency resolution however once frame length is chosen, the time resolution is the same for all frequency components.

Some of the speech enhancement algorithms [22]-[26] are developed using wavelet transform, which provides more flexible time-frequency representations of speech .one popular technique for wavelet–based signal enhancement shrinkage algorithm. Wavelet shrinkage is a simple denoising method based on the thresh holding of the wavelet coefficients. The estimated threshold defines a limit between the wavelet coefficients of the noise and those of the target signal. However it is not always possible to separate the components corresponding to the target signal from those of noise by simple thresholding. For noisy speech, energies of unvoiced segments are comparable to those of noise. Applying thresholding uniformly to all wavelet coefficients not only suppresses additional noise but also some speech components like unvoiced ones. Consequently the perceptual quality of the filtered speech is affected. Therefore the wavelet transform combined with other signal processing tools like wiener filtering in the wavelet domain and wavelet filter bank have been proposed for speech enhancement. Perceptually motivated wavelet decompositions, coupled with various thresholding and estimation techniques are gaining more importance in recent times.

The experimental results for wavelet are shown in fig 8 and 9. Input and output SNR are tabulated in Table 4.It is observed that its performance is comparatively low with MMSE but better than spectral subtraction and iterative methods. Wavelet de-noising show is very poor for high input SNR As similar to that of spectral subtraction and MMSE shows good presentation for speech signal of SNR 5db.

The comparative chart all the four techniques, for three types of noise are shown in fig 10. In this paper we propose a innovative approach in section IV which provide a good performance than the above four techniques.

V. PROPOSED METHOD

In this paper, Digital filter bank i.e. analysis and synthesis filter bank along with Kalman filters are used for speech enhancement. Analysis and Synthesis filter bank banks complement each other, so that the reconstructed speech signal is free from distortions. These filters divide the noisy speech signal into subbands, Kalman filters are applied for each subband. Kalman filters helps in speech spectrum estimation since these filters makes use of models of the speech and noise construction process. Since the Kalman equations comply with white noise, it is best suitable for speech enhancement for reducing white noise. While arriving at Kalman equations it is usually assumed that the process noise is uncorrelated and has a normal distribution. This assumption leads to paleness character of this noise. In Kalman filtering it is assumed that noisy speech signal is stationary during each framework, i.e. the Auto Regressive model of speech remains the same across the section. There are, diverse methods developed to robust the Kalman approach for colored noises.

For shaping the single-dimensional speech signal to the state space model of Kalman filter state vector *u (k)* is given by:

$$u\,(k) = [u\,(k\text{-}p\text{+}1)\ u(k\text{-}p\text{+}2)\ u(k\text{-}p\text{+}3)\ \ldots\ x\,(k)]^T$$

Where u (*k*) is the speech signal at time *k*. Speech signal is contaminated by additive white noise d(*k*) given by *y(k) i.e. y*(*k*)= u(*k*) + d(*k*). The speech signal could be modeled with an AR process of order *p*.

$$u\,(k) = \Sigma\, a_i\, u(k\text{-}i) + p(k)\quad .\ i = 1 \ldots\ldots p$$

Where a_i's are AR (LP) coefficients and *p (k)* is the prediction error it is assumed to have a normal distribution.

The Kalman filter is known in signal processing for its well-organized structure. Paliwal and Basu [27] used a Kalman filter to enhance speech corrupted by white noise. On a short time base, speech signals were modeled as stationary AR processes and AR parameters were assumed to be known. Gibson, Koo, and Gray considered speech enhancement with Colored noise. They modeled both speech and colored noise as AR processes and developed scalar and vector Kalman filtering algorithms.

This paper proposes a simple technique for speech enhancement using Kalman filters where the Speech signals are first decomposed into subbands. Lower order AR processes is used to model the speech signals in each sub band, so that Kalman filters can be used for modeling. By combining the improved subband speech signals enhanced full band speech signals are obtained. There are quite a few methodologies for extraction of LP model parameters from noisy Speech signals. The strength of Kalman Filter for speech enhancement is utilized by assuming LP parameters without worrying about the extraction of these parameters and the effect of this error caused on the system. The performance analysis of these filters with biased parameters is analyzed measured. It has revealed that accurate estimates of AR coefficients are not required when the driving-noise variance is properly estimated. Results obtained by simulation show that speech enhancement achieved by dividing the speech signals to subband not only reduces the computational difficulty, but also achieve improved performance as compared to the full band domain.

Sub band filtering will reduces the complexity involved in estimation of AR coefficients because the power spectral densities (PSD’s) of subband speech signals are simpler and flatter than their fullband signals, low-order AR models are sufficient and only Kalman filters of lower-order is required. In this case, the Kalman filter involves only scalar operations, thus saving a considerable amount of computation. For identification of AR coefficients, we use a prediction-error filter adapted by the LMS algorithm. The LMS algorithm is well known for its ease and robustness, however, its slow convergence precludes its use in many realistic applications. Since the PSD’s of subband speech signals are relatively flat and there is at most one AR coefficient, the LMS algorithm will thus converge more rapidly.

Sub band filtering is achieved by M-channel filter banks, in which the signal is split into M subband $X_k(n)$ by the M analysis filters $H_k(z)$. Fig 3 shows typical frequency responses of the analysis filters. Each signal $X_k(n)$ is then decimated by M to obtain $V_k(n)$. The decimated signals are eventually passed through M-fold expanders and recombined via the synthesis filters $F_k(z)$ to produce x (n). The analysis filter bank is given by H(z) and synthesis filter bank by F(z) .After using noble identities the filter bank can be implemented in terms of E(z) and R (z) as shown in the fig 3.
In this paper M is chosen as eight, Each sub band signal $X_k(n)$ is Processed using Kalman filters, as shown in fig 4. The estimated output $X_k(n)$ is combined using synthesis filter. The experimental result shows that the proposed method gives better performance than MMSE estimator.

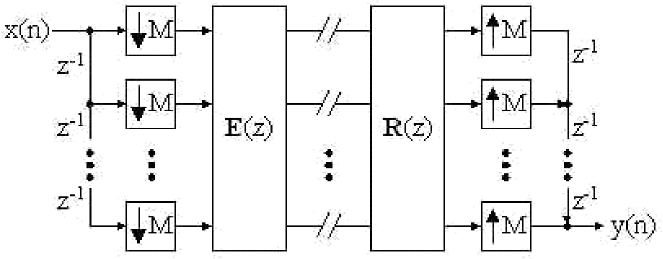

Fig 3: Sub band filtering.

VI. EXPERIMENT RESULTS

In the simulation, noisy speech signals are obtained by adding a clean speech signal with airport, car, and restaurant noise. The noisy speech signals are extracted from NOIZEUS database. The noisy database has thirty IEEE sentences produced by 3 male and 3 female speakers. These signals are corrupted by 8 different types of noises at various SNRs. Four

SNR levels, namely 0 dB, 5 dB, 10 dB are used to evaluate the performance of a speech enhancement system. The sentences where first sampled at twenty five kilo hertz and down sampled to 8 kHz. In our experiment we have used the following speech signals: *"The birch canoe slid on the smooth planks"* and *"The sky that morning was clear and bright blue".*

The experimental results are shown for all the four methodologies discussed in the paper. For each Technique Time domain and Spectrogram results are shown separately for male and female speaker. Experimental results for these techniques are shown in Fig 5-10.SNR improvement in each of the techniques is shown in Table 1-5. In Fig 11-13 all the four techniques are compared for three type of noise. The proposed technique is compared with MMSE in fig 7. Results for the proposed method are shown in Fig 14-15.

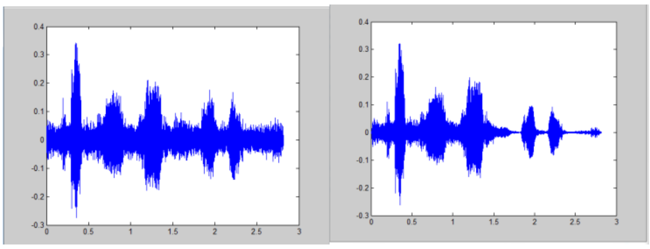

Fig 4a : Speech utterance with background car noise(Top).Effect of Proposed Method (bottom) for Male speaker.

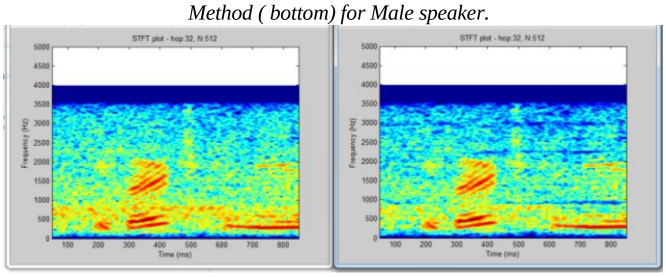

Fig 4b : Spectrogram with background car noise(left). Effect of spectral subraction (right) for Male speaker.

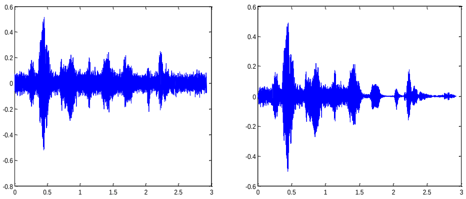

Fig 5a : Speech utterance with background car noise(top).Effect of spectral subraction (bottom) for female speaker.

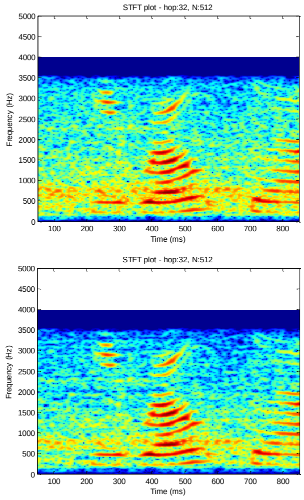

Fig 5b : Speech utterance with background car noise(top).Effect of Proposed Method (bottom) for female speaker.

Speaker	Actual in Decibels	Output in Decibels		
		Car	Airport	Restaurant
Male	10db	11.6811	11.1802	10.1835
	5db	8.0322	6.8405	6.3782
	0db	3.0209	3.0301	2.2521
Female	10db	12.1428	10.8153	11.6376
	5db	7.7168	6.0725	6.4193
	0db	2.5692	2.6655	1.9459

Table 1 : SNR improvement in support of Proposed method for Speech expression with background noise.

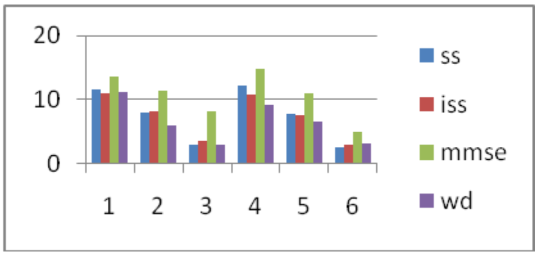

Fig 6 : Comparsion of SS, ISS ,MMSE and WD for speech signal with backround car noise (1: Male speaker for input SNR 10db, 2: Male speaker for input SNR 5db, 3: Male speaker for input SNR 0db, 4: female speaker for input SNR 10db, 5: female speaker for input SNR 5db,6: femalele speaker for input SNR 0db).

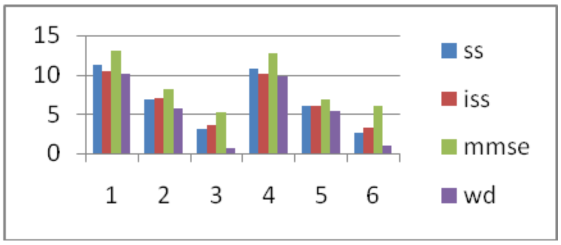

Fig 7: Comparsion of SS, ISS ,MMSE and WD for speech signal with backround airport noise (1: Male speaker for input SNR 10db, 2: Male speaker for input SNR 5db, 3: Male speaker for input SNR 0db, 4: female speaker for input SNR 10db, 5: female speaker for input SNR 5db,6: femalele speaker for input SNR 0db).

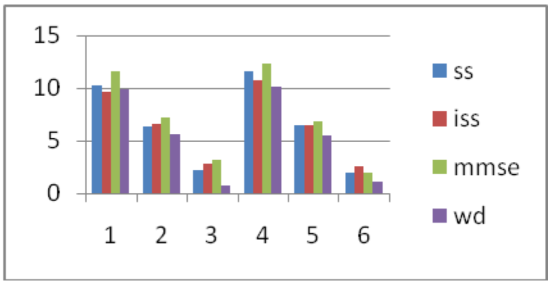

Fig 8 : Comparsion of SS, ISS ,MMSE and WD for speech signal with restaurant noise (1: Male speaker for input SNR 10db, 2: Male speaker for input SNR 5db, 3: Male speaker for input SNR 0db, 4: female speaker for input SNR 10db, 5: female speaker for input SNR 5db,6: femalele speaker for input SNR 0db).

VII. CONCLUSION

In this paper, a novel speech enhancement system has been proposed. It can be optimized into several influential processing techniques that exploit the working of human auditory system to recover premium speech from noise tainted speech. The proposed system consists of two well-designed stages, first sub-band filtering and other one is estimation using Kalman filters. The noisy speech is first decomposed using sub band filtering and then noise reduction is performed for each band using Kalman filtering approach. This modified method takes into account the non-uniform effect of colored noise on the spectrum of speech and cross correlation between back ground noise and speech signal. Experimental outcome show that the proposed method provides the better performance as compared to the other Spectral processing approach. The proposed methodology can be improved by considering the significant bands with non cognitive weighting function.

REFERENCES

[1] S.F.Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.,* Vol 27,pp. 113-120,1979.

[2] P. Lockwood and J. Boudy, "Experiments with a nonlinear Spectral subtractor (NSS), hidden markov models and the projection, for robust speech recognition in cars," *Speech Communication*, Vol. 11, Nos. 2-3, pp. 215-228, 1992.

[3] I. Soon, S. Koh and C. Yeo, "Selective magnitude subtraction for speech enhancement," *Proceedings. The Fourth International Conference/Exhibition on High Performance Computing in the Asia-Pacific Region,* vol.2, pp. 692-695, 2000.

[4] K. Wu and P. Chen, "Efficient speech enhancement using spectral subtraction for car hands-free application," *International Conference on Consumer Electronics*, vol. 2, pp. 220-221, 2001.

[5] C. He and G. Zweig, "Adaptive two-band spectral Subtraction with multi-window spectral estimation," *ICASSP*, vol.2, pp. 793-796, 1999.

[6] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, Vol. 2, pp. 345–349, 1994.

[7] Z. Goh, K.-C. Tan, and B. Tan, "Post processing method for suppressing musical noise generated by spectral subtraction," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 3, pp. 287–292, May 1998.

[8] K. Yamashita, S. Ogata, and T. Shimamura, "Spectral subtraction iterated with weighting factors," in *Proc. IEEE Speech Coding Workshop*, 2002, pp. 138–140.

[9] K. Yamashita, S. Ogata, and T. Shimamura, "Improved spectral subtraction utilizing iterative processing," (in Japanese) *IEICE Trans. A*, vol. J88-A, no. 11, pp. 1246–1257, 2005.

[10] M. R. Khan and T. Hasan, "Iterative noise power subtraction technique for improved speech quality," in *Proc. Int. Conf. Elect. Comput. Eng. (ICECE2008)*, 2008, pp. 391–394.

[11] X. Li, G. Li, and X. Li, "Improved voice activity detection based on iterative spectral subtraction and double thresholds for CVR," in *Proc. 2008 Workshop Power Electron. Intell. Transport. Syst.*, 2008, pp. 153–156.

[12] T. Fukumori, M. Morise, T. Nishiura, and H. Nanjo, "Musical tone reduction on iterative spectral subtraction based on optimum flooring parameters," *IEICE Tech. Rep.*, Vol. 110, pp. 43–48, 2010.