Modified Kalman Filter-based Approach in Comparison with Traditional Speech Enhancement Algorithms from Adverse Noisy Environments

G. Ramesh Babu[1]
Department of E.C.E, Sri Sivani College of Engg.,
Chilakapalem, Srikakulam-532401

Rameshwara Rao[2]
Professor & Dean Faculty of Engineering,
Department of E.C.E, Osmania University,
Hyderabad, India.

Abstract -The paper presents a new speech enhancement approach for a single channel speech enhancement in a noise environment. In this method speech is mixed with real-world noises from babble, car and street environments. In this paper we proposed modified Kalman method for effective speech enhancement. The proposed method is compared to the traditional Spectral Subtraction (SS), Wiener Filter (WF), Minimum Mean Square Error (MMSE) and Wavelet based Filter (WAVELET). Experiments showed that the modified algorithm can give better *SNR* improvement and Subjective evaluation tests demonstrate significant improvement results over classical algorithms, when tested with speech signal corrupted a posterior by various noises at different signal to noise ratios.

Keywords- Speech enhancement, Kalman filtering, SNR, PESQ

I. INTRODUCTION

Speech enhancement methods can be used to increase the quality of the speech processing devices like mobile telephony, digital hearing aids and human-machine communication systems in our daily life and make them more robust under noisy conditions. Speech enhancement includes improving the speech quality, its intelligibility and reducing listener's fatigue. The quality of speech signal is a subjective measure which reflects the way the signal is perceived by listeners. It can be expressed in terms of how pleasant the signal sounds are or how much effort is required to understand the message. Intelligibility, on the other hand is an objective measure of the amount of information that can be extracted by listeners from the given signal. Among various single microphone algorithms for speech enhancement, the spectral subtraction has been mostly employed. Despite its capability of removing background noise, spectral subtraction [1] introduces additional artifacts known as the musical noise, and is faced difficulties in pause detection. This distortion is caused due to the inaccuracies in the short-time noise spectrum estimate.

The spectrum of real world noise does not affect the speech signal uniformly over the entire spectrum. In previous work the fact that is taken into account that the background noise affects the speech spectrum differently at various frequencies and spectral subtraction is performed independently on each band by estimating noise [2]. Other methods focused on masking the musical noise using psychoacoustic models [3] [4]. In recent years, several alternative approaches to this task include traditional methods such as spectral subtraction and Ephraim Malah filtering [6], a drawback of this technique is the necessity to estimate the noise or the signal to noise ratio. This can be a strong limitation when recording with non stationary noise and for situations where the noise cannot be estimated. For spectral subtraction, Wiener filtering, and Ephraim Malah filtering, the signal is divided into 25 ms windows with 12 ms overlap between frames. Wavelet thresholding (shrinking) as a powerful tool in denoising signals degraded by additive white noise and more recently a number of attempts have been made to use perceptually motivated wavelet decompositions coupled with various thresholding and estimation methods Although the application of wavelet shrinking for speech enhancement has been reported in literature [9-11], there are many problems yet to be resolved for a successful application of the method to speech signals degraded by real environmental noise types. The previous reported results in literature shows that they have lower SNR improvement and Mean Opinion Score (MOS).

The main objective of the proposed method is to improve on existing single-microphone schemes for an extended range of noise types and noise levels, thereby making this method more suitable for mobile speech communication applications than the existing. In this paper the signal is modeled as an AR process and a Kalman filter based-method is proposed. The sequential estimators are derived for sub-optimal adaptive estimation of the un-known a priori driving process and additive noise statistics simultaneously with the system state by reformulating and adapting the classical approach used for control applications. First and second-order moments of the noise and driving process are estimated based on state additive and residual noise samples generated in the Kalman filter algorithm. A limited memory algorithm is developed for adaptive correction of the a priori statistics which are intended to compensate for time-varying model errors. The algorithm involves using the innovation sequence to estimate the additive noise variance and the state corrections to estimate the driving process variance.

The estimation of time-varying AR signal model is based on robust recursive least square algorithm with variable for getting factor. The variable for getting factor is adapted to a non stationary signal by a generalized likelihood ratio algorithm through so called discrimination function, developed for automatic detection of abrupt changes in stationarity of signal. The algorithm provides improved state estimates at little computational expense. A distinct advantage of the proposed algorithm is that a VAD (Voice Activity Detection) is not required.

II. BACKGROUND

A. *Spectral Subtraction:*

Spectral subtraction is a method for restoration of the power or the magnitude spectrum of a signal observed in additive noise, through subtraction of an estimate of the average noise spectrum from the noisy signal spectrum. Thus, y(n), the discrete noise corrupted input signal, is composed of the clean speech signal s(n) and the uncorrelated additive noise signal d(n),then the noisy signal can be represented by an equation,

$$y(n) = s(n) + d(n)$$

B. Wiener filter in frequency domain:

The basic principle of the Wiener filter is to obtain an estimate of the clean signal from that corrupted by additive noise. This estimate is obtained by minimizing the Mean Square Error (MSE) between the desired signal $s(n)$ and the estimated signal $\hat{s}(n)$. Transfer Function in frequency domain is given below

$$H(\omega) = Ps(\omega)/ Ps(\omega) + Pv(\omega)$$

where $Ps(\omega)$ and $Pv(\omega)$ are the power spectral densities of the clean and the noise signals, respectively. This formula can be derived considering the signal s and the noise v as uncorrelated and stationary signals.

C. Wavelet Denoising Method

The wavelet denoising method is a nonlinear denoising method based on the wavelet decomposition. Compared with the traditional low pass filters, the wavelet denoising method can not only realize the function of low pass filter but also maintain the feature of the signal. Among the different methods of wavelet denoising, the wavelet threshold denoising method is applied widely and can meet the needs of real time.

III. NOISY SPEECH MODEL AND KALMAN FILTERING

The speech signal s(n) is modeled as a Pth -order order AR process where

$$S(n) = \sum_{i=1}^{p} (a_i(n)\, S(n-i) + u(n)) \qquad (1)$$

$$y(n) = s(n) + v(n) \qquad (2)$$

where s(n) is the nth sample of the speech signal, y(n) is the nth sample of the observation, and $a_i(n)$ is the i[th] AR parameter. This system can be represented by the following state-space Model

Where

[1] the sequences u(n) and v(n) are uncorrelated Gaussian white noise sequences with the mean $\bar{u}$ and $\bar{v}$ and the variances σ_u^2 and σ_v^2
[2] x(n) is the P x 1 state vector

$$\mathrm{X}(n) = [S(n-p+1) \ldots \ldots s(n)]^T \qquad (3)$$

[3] F(n) is the P x P transition matrix

$$F(n)= \begin{matrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ . & . & . & \ldots & . \\ . & . & . & \cdots & . \\ 0 & 0 & 0 & \ldots & 1 \\ a_p(n) & a_{p-1}(n) & a_{p-2}(n) & \ldots & a_1(n) \end{matrix}$$

[4] G and H are , respectively , the P x 1 input vector and 1 x p observation row vector which is defined as follows

$$\mathrm{H} = G^T = [\,0 \;\; 0 \;\; \cdots\cdots\cdots \;\; 0 \;\; 1\,] \qquad (4)$$

The standard Kalman filter [5] provides the updating state vector estimator equations

$$e(n) = y(n) - H\hat{x}(n/(n-1)) - \bar{v} \qquad (5)$$

$$S(n) = P(n/(n-1))H^T \quad x\, [HP(n/(n-1))\, H^T + \sigma_v^2]^T \qquad (6)$$

$$\hat{x}(n/n) = \hat{x}(n/(n-1)) + K(n)e(n) \qquad (7)$$

$$P(n/n) = [I - K(n)H]\, P(n/n-1) \qquad (8)$$

$$\hat{x}(n+1/n)=F(n)\hat{x}(n/n)+G_{\bar{u}} \quad (9)$$

$$P(n+1/n)= F(n)P(n/n)\, F^T(n) + GG^T\sigma_u^2 \quad (10)$$

where

a) $\hat{x}(n+1/n)$ is the minimum mean square estimation of the state vector X(n) given the past n-1 observations y(1), ……. ,y(n-1)
b) $\tilde{x}(n/n-1) = x(n) - \hat{x}(n/n-1)$ is the predicted state error vector.
c) $P(n/n-1)=E[\tilde{x}(n/n-1)\, \tilde{x}^T(n/n-1)]$ is predicted state error correlation matrix.
d) $\hat{x}(n/n)$ is the filtered estimation of the state vector.
e) $\tilde{x}(n/n) = x(n) - \hat{x}(n/n)$ is the filtered state error vector.
f) $P(n/n)=E[\tilde{x}(n/n-1)\, \tilde{x}^T(n/n)]$ is the filtered state error correlation vector.
g) e(n) is the innovation sequence.
h) K(n) is the Kalman gain.

The estimated speech signal can be retrieved from the state-vector estimator

$$\hat{s}(n) = H\hat{x}(n/n) \quad (11)$$

The parameter estimation (the transition matrix and noise statistics) is presented in the next section.

IV. PARAMETER ESTIMATION

The estimation of the transition matrix, which contains the AR speech model parameters, was made using a adaptation of the robust recursive least square algorithm with variable forgetting factor proposed by Milosavljevic et al. [6]. The estimation of the noise statistics is derived under the assumption of the constant values over N samples by reformulating and adapting the approach proposed in control by Myers and Tapley [7].

A. Estimation of the Transition Matrix

In our approach, getting F(n) requires the AR parameter estimation. The equation (3) can be rewritten in the form

$$S(n) = X^T(n-1)\theta(n) + u(n) \quad (12)$$

where

$$\theta(n) = [\ a_p(n) \ \ a_{p-1}(n) \ \ldots\ldots \ a_1(n)\ ]^T \qquad (13)$$

The robust recursive least square approach estimates the vector ˆθ(n) by minimizing the M-estimation criterion [6] is the Huber influence function and Δ is a chosen constant.

$$J_n = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \lambda^{n-i} \rho[\epsilon^2(i)] \qquad (14)$$

where

$$\varphi(x) = \rho^I(x) = min\left[\frac{|x|}{\sigma_u^2}, \frac{\Delta}{\sigma_u}\right] sgn(x) \qquad (15)$$

The true state vector x(n) used in (12) is unknown but can be approximated by the state-vector estimator ˆx(n/n). In this case the robust recursive least square approach gives the estimation equations

$$\mathrm{e(i)=Hx\hat{}(i/i)\ -x\hat{}}^T\mathrm{(i-1/i-1)}\ \theta\hat{}\mathrm{(i-1)} \qquad (16)$$

$$\mathrm{g(i)} = \frac{\{\mathrm{Q(i-1)}\hat{\mathrm{x}}^T\mathrm{(i-1/i-1)}\}}{\{\lambda\mathrm{(i)}+\varphi^{\mathrm{I}}[\epsilon\mathrm{(i)}]\hat{\mathrm{x}}^T\mathrm{(i-1/i-1)Q(i-1)}\hat{\mathrm{x}}\mathrm{((i-1)/(i-1))}\}} \qquad (17)$$

$$Q(i) = \frac{1}{\lambda\mathrm{(i)}}[\mathrm{Q(i-1)} - g(i)\hat{\mathrm{x}}^T\mathrm{(i-1/i-1)Q(i-1)}\varphi^{\mathrm{I}}[\epsilon\mathrm{(i)}]] \qquad (18)$$

$$\theta\hat{}\mathrm{(i)}= \theta\hat{}\mathrm{(i-1)+Q(i)}\ \hat{\mathrm{x}}\mathrm{((i-1)/(i-1))}\varphi\ [\epsilon\mathrm{(i)}] \qquad (19)$$

The forgetting factor λ(i) is a data weighing factor that is used to weight recent data more heavily and thus to permit tracking slowly varying signal parameters. If a non stationary signal is composed of stationary sub signals the estimation of the AR parameters can be given by using a forgetting factor varying between λmin and λmax. The modified generalized likelihood ratio algorithm issued for the automatic detection of abrupt changes in stationarity of signal. This algorithm uses three models of the same structure and order, whose parameters are estimated on fixed length windows of signal. These windows are [i − N + 1, i], [i + 1, i + N] and [i − N + 1, i + N], and move one sample forward with each new sample. In the first step of this algorithm is calculated the discrimination function

$$D(i,N)=L(i-n+1,i+N)-L(i-N+1,i)-L(i+1,i+N) \qquad (20)$$

where

$$L(a,b) = (b-a+1)\ln\left[\frac{1}{b-a+1}\sum_{i=a}^{d}\epsilon^2(i)\right] \qquad (21)$$

denotes the maximum of the logarithmic likelihood function. In the second step a strategy for choosing the variable forgetting factor is defined by letting λ(i) = λmax when D = Dmin and λ(i) = λmin when D = Dmax, as well as by taking the linear interpolation between these values.

B. Estimation of Additive Noise Statistics

The estimation of additive noise statistics is derived under the assumption of the constant mean and variance over N samples v(n), v(n − 1), · · · , v(n −N +1). Using the equation (4) the samples of the additive noise are given by the equation

$$v(n) = y(n) - Hx(n) \qquad (22)$$

The true states vector x(n) is unknown, so v(n) cannot be determined, but the approximation

$$\alpha(n) = y(n) - H\hat{x}(n/n-1) = H\tilde{x}(n/n\text{-}1) + v(n) \qquad (23)$$

can be used [7]. The samples α(n) are assumed to be representative of v(n) and can be considered independent and identically distributed [7]. Based on the last N samples α(n), α(n − 1), · · · , α(n−N+1) the mean .α(n) and the variance σ2α(n) are estimated. An unbiased estimator for .α(n) is taken as the sample mean and an unbiased estimator for σ2 α(n) is obtained by

$$\hat{\bar{\alpha}} = \frac{1}{N}\sum_{i=0}^{N-1}\alpha(n-i) \qquad (24)$$

$$\hat{\sigma}_{\alpha}^2 = \frac{1}{N-1}\sum_{i=0}^{N-1}[\alpha(n-i) - \hat{\bar{\alpha}}(n)]^2 \qquad (25)$$

The estimation of the additive noise mean is

$$\hat{\hat{v}}(n) = \hat{\hat{\alpha}}(n) \qquad (26)$$

If the samples α(n) are considered independent and identically distributed the expected value of ˆσ2α(n) is Using (25) and (27) an unbiased estimator of σ2v(n) is given by

$$E\{\hat{\sigma}_\alpha^2(n)\} = \frac{1}{N}\sum_{i=0}^{N-1} HP(n/n-1)H^T + \sigma_v^2 \qquad (27)$$

$$\hat{\sigma}_v^2\hat{\sigma}_v^2 = \frac{1}{N-1}\left\{\begin{matrix}\displaystyle\sum_{i=0}^{N-1}[\alpha(n-i)-\hat{\hat{\alpha}}]^2 \\ -\left(\frac{N-1}{N}\right)HP(n-i/n-i-1)(H^T)\end{matrix}\right\} \qquad (28)$$

C. Estimation of Driving Process Statistics

The estimation of driving process statistics is derived under the assumption of the constant mean and variance over N samples u(n), u(n − 1), · · · , u(n − N + 1). Using the state propagation equation (3) the samples of the driving process are given by the equation:

$$u(n) = H[x(n) - Fx(n-1)] \qquad (29)$$

The true state vectors x(n) and x(n − 1) are unknown, so u(n) cannot be determined, but the approximation

$$\beta(n) = H[\hat{x}(n/n) - \hat{x}(n/n-1)] \qquad (30)$$

can be used [7]. The samples β(n) are assumed to be representative of u(n) and can be considered independent and identically distributed [7]. Based on the last N measurements the mean β(n) and the variance σ2β(n) are estimated [9]. An unbiased estimator for β(n) is taken as the sample mean

$$\hat{\hat{\beta}}(n) = \frac{1}{N}\sum_{i=0}^{N-1}\beta(n-i) \qquad (31)$$

and an unbiased estimator for σ2β(n) is obtained by

$$\hat{\sigma}_\beta^2(n) = \frac{1}{N-1}\sum_{i=0}^{N-1}[\beta(n-i) - \hat{\hat{\beta}}(n)]^2 \qquad (32)$$

The estimation of the additive noise mean is

$$\hat{\bar{u}}(n) = \hat{\bar{\beta}}(n) \qquad (33\ )$$

If the samples β(n) are considered independent and identically distributed the expected value of ˆσ2β(n) is

$$E\{\hat{\sigma}_{\beta}^{2}(n)\} = \frac{1}{N}\sum_{i=0}^{N-1} E\{[\beta(n-i)]^{2}\} \qquad (34)$$

The analysis reduces to expanding $E\{[\beta(n-i)]^2\}$ in terms of σ_u^2 . We write $\beta(n)$ in term of the filtered state-error vector.

$$\beta(n) = -H\tilde{x}(n/n) + HF\tilde{x}(n-1/n-1) + u(n) - \bar{u} \qquad (35)$$

Since the filtered state-error vectors errors are not independent, the correlation are avoided by writing

$$\beta(n) + H\tilde{x}(n/n) = HF\tilde{x}((n-1)/(n-1)) + u(n) - \bar{u} \qquad (36)$$

The variance of this equation is

$$E\{[\beta(n) + H\tilde{x}(n/n)]^2\} = HPF(n-1/n-1)F^T H^T + \sigma_u^2 \qquad (37)$$

Now we develop $E\{[\beta(n) + H\tilde{x}(n/n)]^2\}$ in terms of $E\{[\beta(n)]^2\}$ and of other computed terms in the Kalman filter

$$E\{[\beta(n) + H\tilde{x}(n/n)]^2\} = E\{[\beta(n)]^2\} + 2E\{\beta(n)\tilde{x}^T(n/n) \ + \ HP(n/n)H^T \qquad (38)$$

$$\tilde{x}(n/n) = [I - K(n)H]\tilde{x}(n/n-1) - K(n)[v(n) - \bar{v} \qquad (39)$$

and the second term in (37) is

$$E\{\beta(n)\tilde{x}^T(n/n)H^T\} \ = -HP(n/n)H^T + \ HP(n/n-1) \ \times [I - K(n)H]^T H^T \qquad (40)$$

Using (32), (34) and (37) an unbiased estimator of $\sigma_u^2(n)$ is given by

$$\hat{\sigma}_u^2 = \frac{1}{N-1}\{\sum_{i=0}^{N-1}\left[\beta(n-i)-\hat{\hat{\beta}}(n)\right]^2 - \frac{N-1}{N}HPF(n-i-1/n-i-1)F^T H^T - \frac{N-1}{N}HP(n-i/n-i)H^T + 2\frac{N-1}{N}HP(n-i/n-i-1)\ [I-K(n)H]^T H^T\} \qquad (41)$$

V. EXPERIMENT RESULTS

For the experiment, the Noizeus database [10] was used. And signal was corrupted by eight different real-world noises at different SNRs 0dB, 5dB, 10dB and 15dB. The sentences were originally sampled at 25 kHz and down sampled to 8 kHz. Noise signals were taken from the AURORA database [11] and included the following recordings from different places: babble (crowd of people), car and street. Figure 1 represents, the noise-free speech signal, the noisy speech signal and the enhanced speech signals of spectral, Wiener, mmse, wavelet & proposed methods respectively. Figure 2 represents spectrograms of all enhanced signals. Improvements of SNR for different real time noises babble, car and street for all the enhanced methods are shown in fig 3.The SNR comparison of BABBLE noise for different enhanced methods shown in table 1. The PESQ score [12] for different methods shown in table 2 and also it is given in bar graph fig 4.

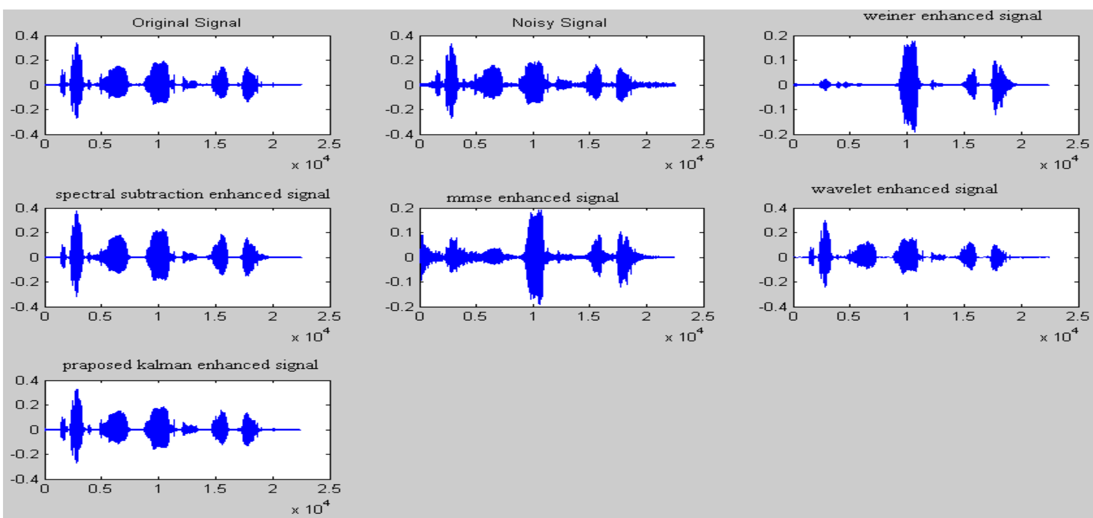

Fig 1. Shows the time domain representation of noise free, noisy and enhanced speech signals.

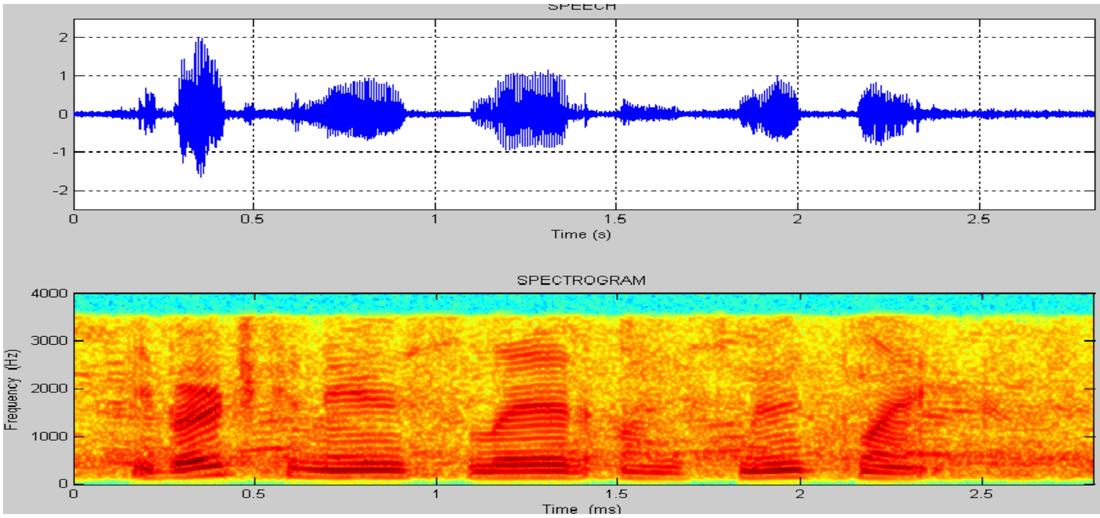

Fig 2. a Spectrograms of original signal

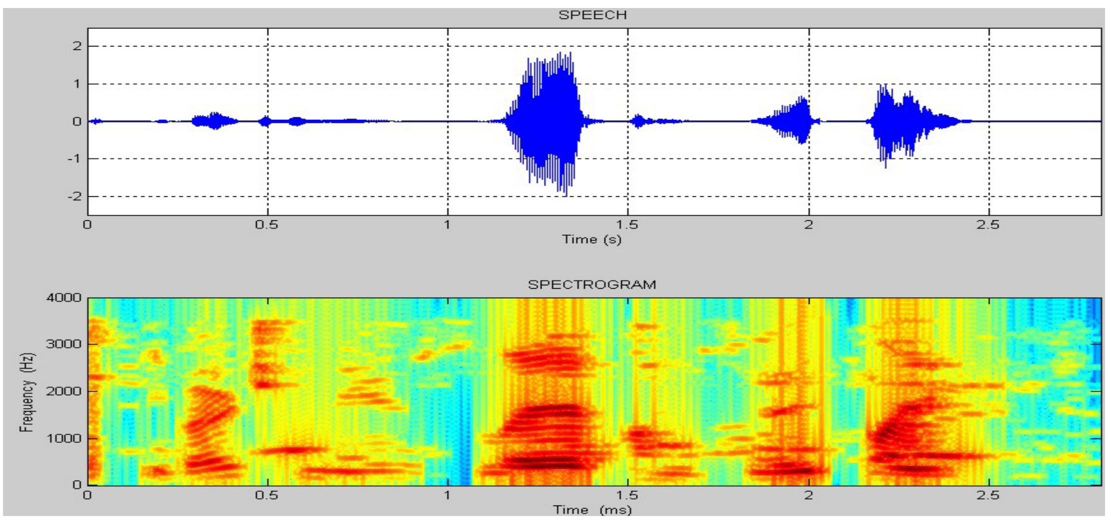

Fig 2.b Spectrograms of Wiener enhanced signal

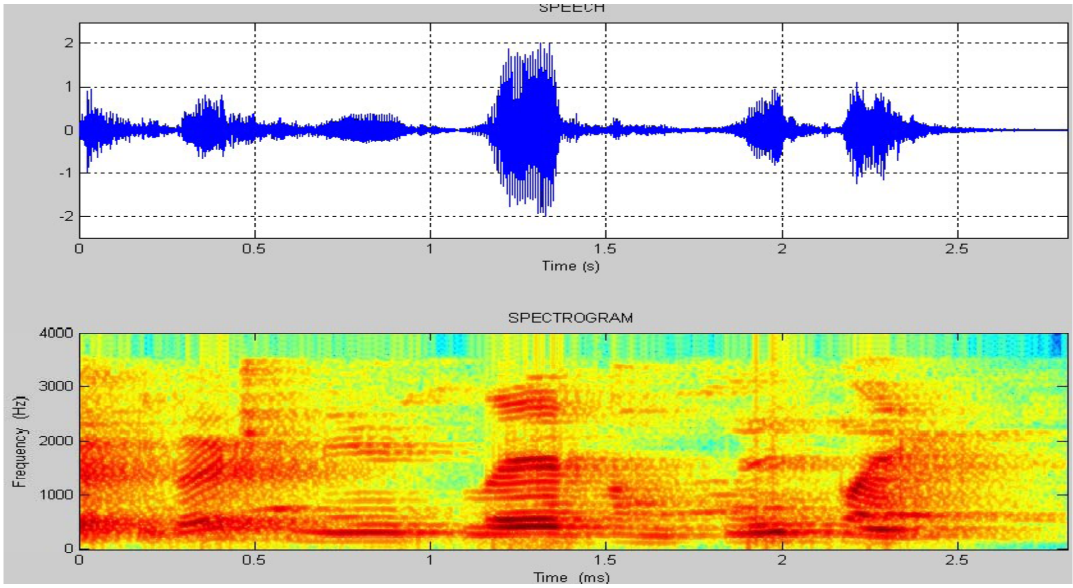

Fig 2.c Spectrograms of spectral enhanced signal

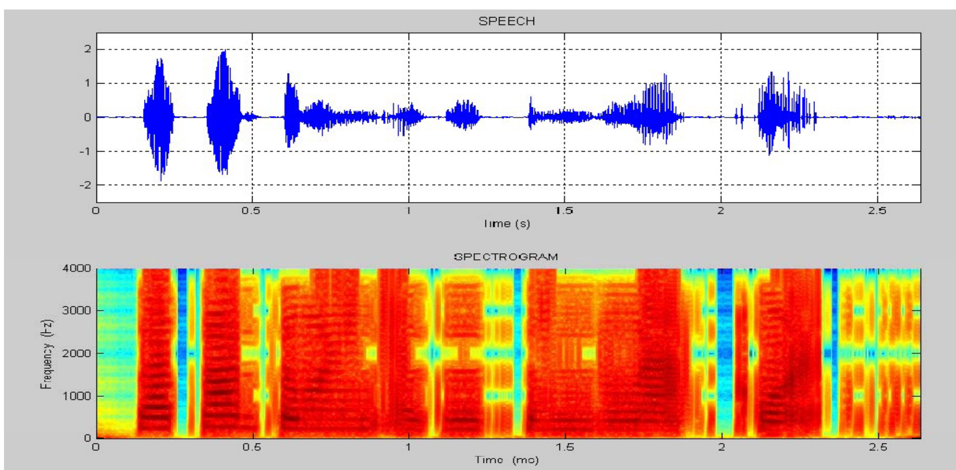

Fig 2.d Spectrograms of mmse enhanced signal

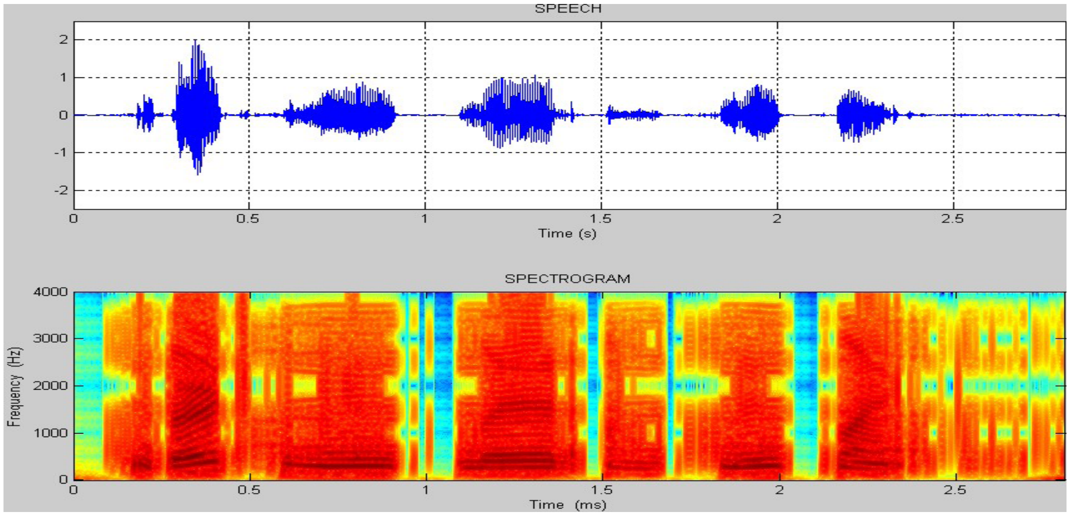

Fig 2.e Spectrograms of wavelet enhanced signal

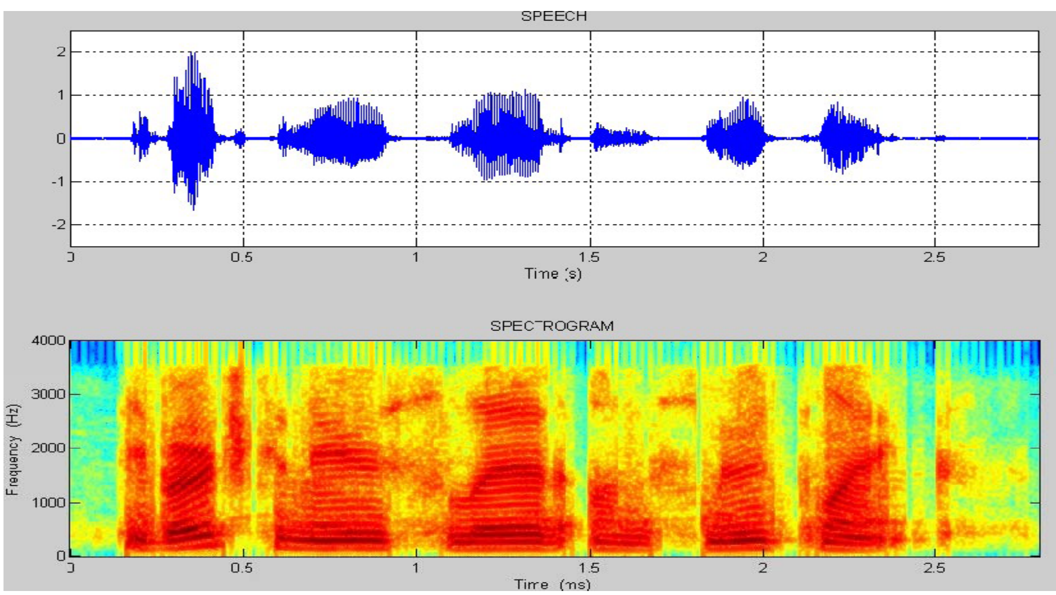

Fig 2.f Spectrograms of proposed method enhanced signal

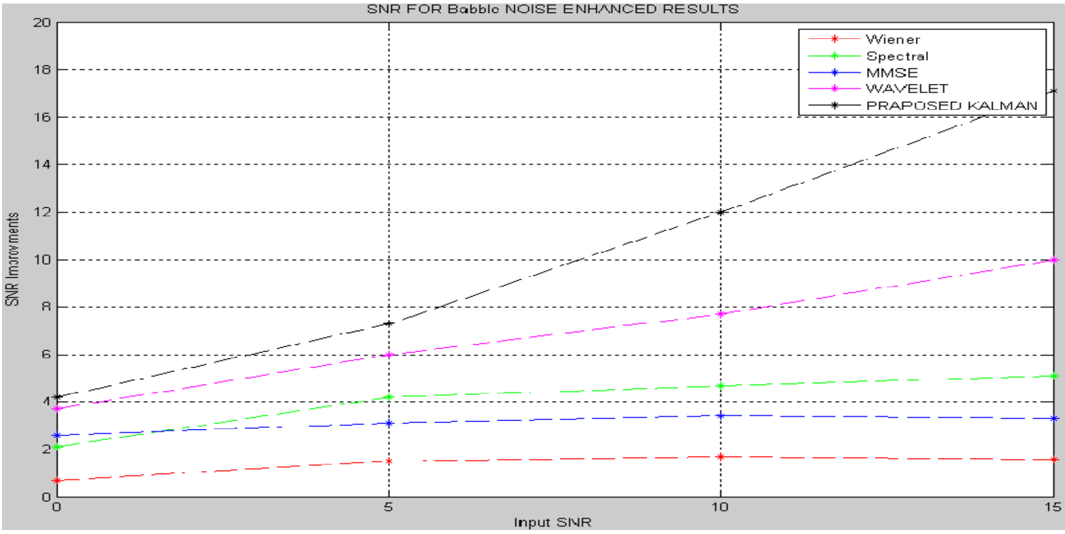

Fig 3.a Improvement of SNR in different methods for Babble noise

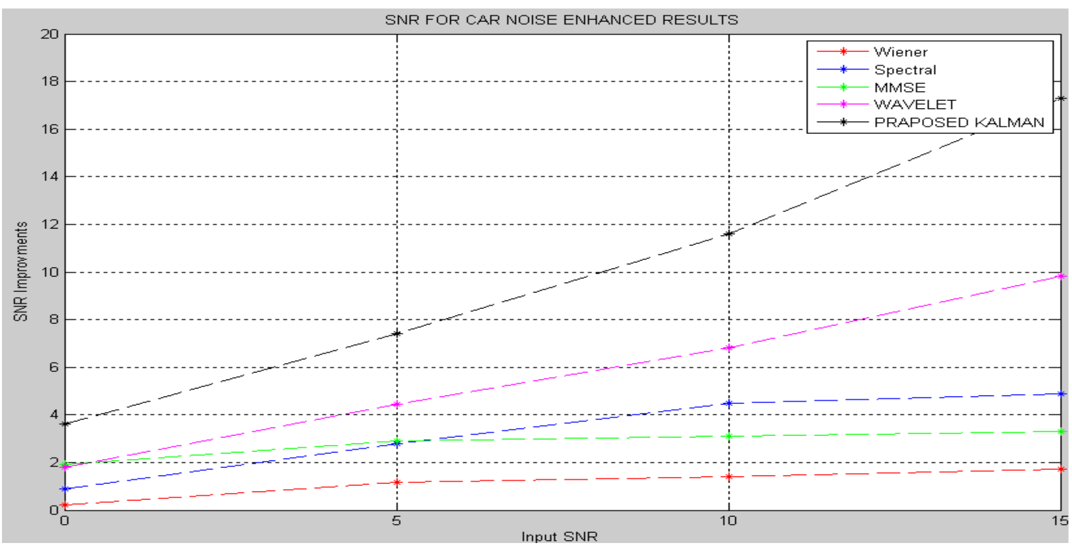

Fig 3.b Improvement of SNR in different methods for Car noise

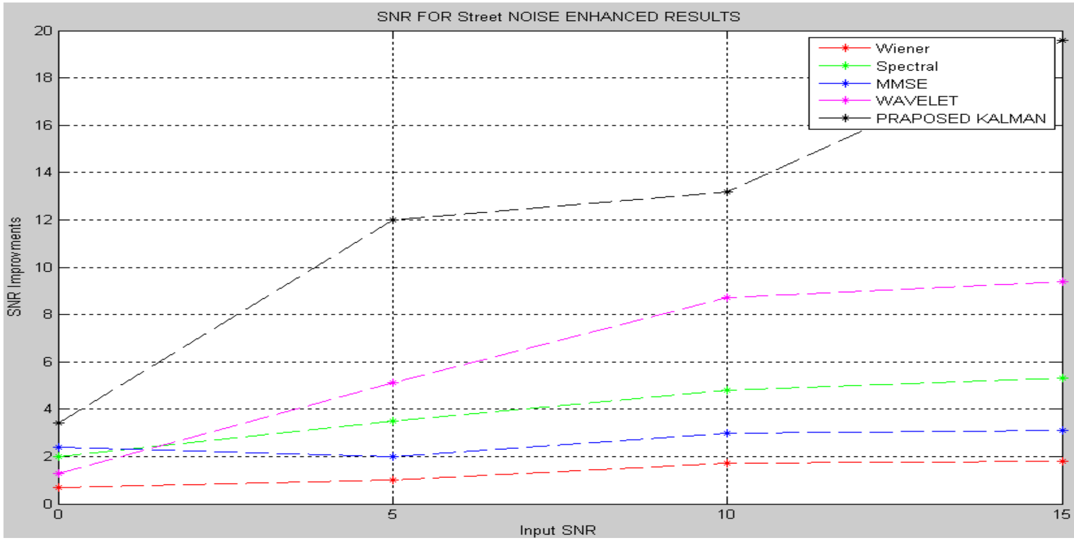

Fig 3.c Improvement of SNR in different methods for Street noise

TABLE1. SNR COMPARISONS (BABBLE NOISE)

SNR	**WIENER**	**SS**	**MMSE**	**WAVELET**	**PROPOSED KALMAN**
0dB	0.7	2.1	2.6	3.7	4.2
5dB	1.5	4.2	3.1	6.0	7.3
10dB	1.7	4.7	3.4	7.7	12.0
15dB	1.6	5.1	3.3	10.0	17.1

TABLE 2. PESQ SCORE FOR DIFFERENT METHODS

PESQ	0dB	5dB	10dB	15dB
WIENER	0.067	1.210	1.062	1.415
SS	1.745	1.835	1.854	2.069
MMSE	1.716	2.511	2.458	2.416
WAVELET	2.105	1.803	1.724	1.797
PROPOSED KALMAN	2.594	2.335	2.764	2.797

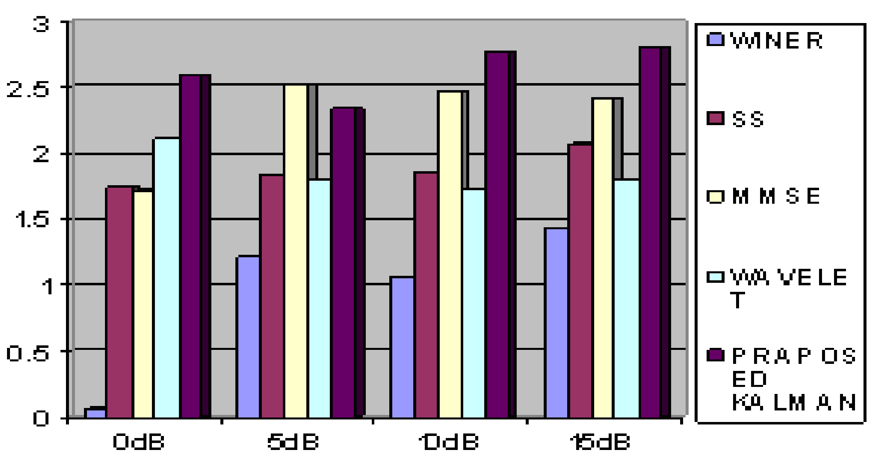

Fig 4. PESQ score for different methods

VI. CONCLUSION

In this paper a new method (Modified Kalman) for speech enhancement was proposed. The proposed method is evaluated by using various criteria of quality and quantity. By using the mentioned criteria, it is presented that this method can compete with other speech enhancement methods. As seen in Fig. (3a), (3b) and (3c) the proposed method for the input SNR between 0 and 15 dB the proposed method provides better SNR results than four previously proposed methods by the author and Gibson's algorithm and it is observed that The improvement of SNR is higher at high Input SNR levels than lower input SNR levels. From the Fig. (4) the Perceptual Evaluation of Speech Quality scores (PESQ, ITU-T p.862) also verifies the efficiency of the proposed method.

REFERENCES

[1] S.F.Boll, "Suppression of acoustic noise In speech using spectral subtraction", IEEE Trans. Acoustics. Speech. Signal process.vol. 27 pp. 13-120, April 1979.

[2] M.G. Sumithra, D. Deepa and K.Thanuskodi "Frequency Dependent Single Channel Speech Enhancement from Additive Background Noise "Proceedings of the International conference on Advanced Communication Systems, pp. 85-90,2007.

[3] D.E.Tsoukalas, J.N.Mourjopoulos and Kokkinakis "Speech enhancement based on audible noise suppression" , IEEE Trans. Speech Audio . Proc.,vol.5, pp. 479–514, Nov.1997.

[4] 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, Mar. 1999.

[5] Y. Ephraim and D. Malah " Speech Enhancement Using a minimum mean-square error log-spectral amplitude estimator", IEEE Trans. Acoustic .Speech Signal Processing ASSP-32(6), 1109-1121,1984.

[6] D.L. Donoho ," De-noising by soft thresholding",IEEE Trans. on Information Theory, vol. 41 no. 3,613-627, May 1995.

[7] W. Seok and K.S.Bae, " Speech enhancement with reduction of noise components in the waveletdomain", in Proceedings of the ICASSP, pp.1323-1326, 1997.

[8] E. Ambikairajah, G.Tattersall and A. Davis, "Wavelet transform based speech enhancement" in Proceedings of ICSLP, 1998.

[9] Yasser Ghanbari, Mohammad Reza Karami Mollaei, "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets", Speech Communication 48, 927-940, 2006.

[10] "Noizeus: A noisy speech corpus for evaluation of Speech enhancement algorithms, "http://www.utdallas.edu/~loizou/speech/noizeus.

[11] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in ISCA ITRW ASR2000, Sept. 2000, Paris, France.ss.

[12] ITU-T P.862, Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrow band telephone networks and speech codecs, ITU-T Recommendation p.862, 2000.