Application of the Tuned Kalman Filter in Speech Enhancement

Orchisama Das, Bhaswati Goswami and Ratna Ghosh
Department of Instrumentation and Electronics Engineering
Jadavpur University Kolkata, India
Email: orchisama.das@gmail.com, bg@iee.jusl.ac.in, rg@iee.jusl.ac.in

***Abstract*—The Kalman filter has a wide range of applications, noise removal from corrupted speech being one of them. The filter performance is subject to the accurate tuning of its parameters, namely the process noise covariance, Q, and the measurement noise covariance, R. In this paper, the Kalman filter has been tuned to get a suitable value of Q by defining the robustness and sensitivity metrics, and then applied on noisy speech signals. The Kalman gain is another factor that greatly affects filter performance. The speech signal has been frame-wise decomposed into silent and voiced zones, and the Kalman gain has been adjusted according to this distinction to get best overall filter performance. Finally, the algorithm has been applied to clean a noise corrupted known signal from the NOIZEUS database. It is observed that significant noise removal has been achieved, both audibly and from the spectrograms of noisy and processed signals.**

***Index Terms*—Kalman filter, speech enhancement, filter tuning, Kalman gain, robustness metric, sensitivity metric.**

I. Introduction

The purpose of speech enhancement is the suppression of white noise in corrupted speech to make it more intelligible. Noise corruption in speech is unavoidable in many cases, and the quality needs to be improved. A very interesting application of offline speech enhancement is in the recovery and restoration of old archival audio records that have been heavily corrupted by noise, and have degraded with time.

The Kalman filter is a linear minimum mean squared error (MMSE) estimator that was proposed in [1] to predict the unknown states of a dynamic system. Its first application in speech processing was given by [2] where the system model was designed based on the autoregressive nature of speech. In [2], Paliwal and Basu claimed that the Kalman filter performed better than the most famous method of speech enhancement then, i.e., the Wiener filter [3], because it exploited the knowledge of the speech production process. The Wiener filter also assumed speech to be a stationary process, which in reality, it is not. Since then, many improvements to this algorithm have been proposed, including iterative Kalman filtering proposed by Gibson et al. [4]. So et al. [5] used long tapered windows and Kalman gain trajectories to explain reduction of residual noise in output. Additionally, they also applied Dolph-Chebyshev windowing and iterative LPC (linear prediction coefficient) estimation to further enhance the filter performance.

However, filter tuning, or optimum parameter estimation has not been a major focus in previous works on speech enhancement using the Kalman filter. The two most important parameters in the Kalman filter are the process noise covariance matrix, Q and the measurement noise covariance matrix, R [6]. Their proper choice can greatly improve filter performance. The determination of R (or simply, white noise variance in case of speech), is quite often done by the autocorrelation method proposed by Paliwal [7]. In this paper, a different approach has been used based on extraction of silent zones from speech, and consequent calculation of noise variance. Hence, the primary objective becomes that of properly choosing Q. Saha et.al [6] propose to do so by ensuring a balanced root-mean-square performance (RMSE) in terms of robustness and sensitivity. For this, two performance metrics have been defined - the robustness metric and sensitivity metric. A trade-off between the two is established to ensure a suitable choice of the parameter, Q.

In this paper, an improvement of the Kalman filter used for speech enhancement has been proposed. In doing so, the process noise covariance Q of the filter used in an autoregressive model of speech has been judiciously selected using the robustness and sensitivity metrics. Furthermore, the parameter Q has been chosen framewise in order to account for the voicing in the *silent* and *voiced* frames, while the measurement noise covariance R has been determined from the silent frames. The subsequent Kalman gain trajectory and its effect on filter performance have been studied in detail. Finally, experiments have been performed on an Additive Gaussian White Noise (AGWN) corrupted speech signal with SNR 5dB, available from the NOIZEUS [8] speech corpus.

The rest of the paper is organised as follows. Section II contains the state space model for speech enhancement and the Kalman filter equations. Section III contains the determination of R, the definition of robustness and sensitivity metrics and the corresponding selection of Q. Section IV contains the details of the proposed algorithm. Section V contains the experiments and results. The conclusion is stated in Section VI.

II. State Space Model and Filter Equations

Speech can be modelled as a pth order autoregressive process, where the present sample, $x(k)$ depends on the last p samples. It is an all-pole linear system, with an additive Gaussian noise as the input. Speech signal at kth sample is given by:

$$x(k) = -\sum_{i=1}^{p} a_i x(k-i) + u(k) \tag{1}$$

where a_i are the linear prediction coefficients (LPCs) and $u(k)$, the process noise, is a zero-mean Gaussian noise with variance σ_u^2. This equation can be represented by the state space model as shown below.

$$\begin{bmatrix} x(k-p+1) \\ x(k-p+2) \\ \vdots \\ x(k) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_p & -a_{p-1} & -a_{p-2} & \cdots & -a_1 \end{bmatrix} \times \begin{bmatrix} x(k-p) \\ x(k-p+1) \\ \vdots \\ x(k-1) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} u(k) \tag{2}$$

or

$$\boldsymbol{X}(k) = \boldsymbol{\phi}\boldsymbol{X}(k-1) + \boldsymbol{G}u(k) \tag{3}$$

where $\boldsymbol{X}$(k) is the *(p×1)* state vector matrix, $\boldsymbol{\phi}$ is the *(p×p)* state transition matrix, $\boldsymbol{G}$ is the *(p×1)* input matrix and *u(k)* is the noise corrupted input signal at the *k*th instant.

When speech is noise corrupted, the output *y(k)* is given as:

$$y(k) = x(k) + w(k) \tag{4}$$

where $w(k)$ is the measurement noise, a zero-mean Gaussian noise with variance σ_w^2. In vector form, this equation may be written as

$$y(k) = \boldsymbol{H}\boldsymbol{X}(k) + w(k) \tag{5}$$

where $\boldsymbol{H}$ is the *(1×p)* observation matrix given by

$$H = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \tag{6}$$

The Kalman filter calculates $\hat{\boldsymbol{X}}$(k|k) which is the estimate of the state vector $\boldsymbol{X}$(k), given corrupted speech samples upto instant *k*, by using the following equations:

$$\hat{\boldsymbol{X}}(k|k-1) = \boldsymbol{\phi}\hat{\boldsymbol{X}}(k-1|k-1) \tag{7}$$

$$\boldsymbol{P}(k|k-1) = \boldsymbol{\phi}\boldsymbol{P}(k-1|k-1)\boldsymbol{\phi}^{\boldsymbol{T}} + \boldsymbol{G}Q\boldsymbol{G}^{\boldsymbol{T}} \tag{8}$$

$$\boldsymbol{K}(k) = \boldsymbol{P}(k|k-1)\boldsymbol{H}^{\boldsymbol{T}}(\boldsymbol{H}\boldsymbol{P}(k|k-1)\boldsymbol{H}^{\boldsymbol{T}} + R)^{-1} \tag{9}$$

$$\hat{\boldsymbol{X}}(k|k) = \hat{\boldsymbol{X}}(k|k-1) + \boldsymbol{K}(k)(y(k) - \boldsymbol{H}\hat{\boldsymbol{X}}(k|k-1)) \tag{10}$$

$$\boldsymbol{P}(k|k) = (\boldsymbol{I} - \boldsymbol{K}(k)\boldsymbol{H})\boldsymbol{P}(k|k-1) \tag{11}$$

where

- $\hat{\boldsymbol{X}}$(k|k-1) is the *a priori* estimate of the current state vector $\boldsymbol{X}$(k).
- $\boldsymbol{P}$(k|k-1) is the error covariance matrix of the *a priori* estimate, given by $\boldsymbol{E}[\boldsymbol{e}_k^- \ \boldsymbol{e}_k^{-T}]$ where $\boldsymbol{e_k^-}$=$\boldsymbol{X}$(k)-$\hat{\boldsymbol{X}}$(k|k-1).
- Q is the process noise covariance matrix,which in this case is σ_u^2. Similarly, *R* is the measurement noise covariance matrix, which is σ_w^2.
- $\hat{\boldsymbol{X}}$(k|k) is the *a posteriori* estimate of the state vector. In our case, the last component of $\hat{\boldsymbol{X}}$(k|k) is $\hat{x}(k)$, which gives the final estimate of the processed speech signal.
- $\boldsymbol{P}$(k|k) is the error covariance matrix of the *a posteriori* estimate, given by $\boldsymbol{E}[\boldsymbol{e}_k \ \boldsymbol{e}_k^T]$ where $\boldsymbol{e_k}$=$\boldsymbol{X}$(k)-$\hat{\boldsymbol{X}}$(k|k).
- Let $\hat{\boldsymbol{X}}$(0|0)=[*y(1)* ,$\cdots$, *y(p)*] and $\boldsymbol{P}(0|0) = \sigma_w^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the *(p×p)* identity matrix.
- $\boldsymbol{K}$(k) is the Kalman gain for the *k*th instant. The term *y(k)* - $\boldsymbol{H}\hat{\boldsymbol{X}}$(k|k-1) is known as the *innovation*.

Equations (7) and (8) are known as the *time update* equations whereas (9), (10), (11) are known as the *measurement update* equations. Intuitively, the Kalman filter equations may be explained thus: The gain $\boldsymbol{K}$(k) is chosen such that it minimizes the *a posteriori* error covariance, $\boldsymbol{P}$(k|k). As $\boldsymbol{P}$(k|k-1) decreases, $\boldsymbol{K}$(k) reduces. An inspection of (10) shows that as $\boldsymbol{K}$(k) reduces, the *a priori* state estimate is trusted more and more and the noisy measurement is trusted less.

III. Filter Tuning

A. Determination of R: Silent and Voiced Frames

In case of the AR model for speech, the measurement noise covariance *R* is a scalar quantity, the value of which is simply the variance of the white noise corrupting the speech, i.e, σ_w^2. To determine σ_w^2, the entire signal is divided into 80ms long frames with 10ms overlap as per the method outlined in [5]. The frames are then classified as *silent* or *voiced*. A frame *i* is classified as *silent* if:

$$E(i) < \frac{max(E)}{100} \tag{12}$$

where *E(i)* is the energy of spectral components below 2kHz for the *i*th frame and *E*=[*E(1), E(2),* $\cdots$ *E(n)*] is the set of spectral energy components below 2kHz for all frames. It is to be noted that for noisy speech, both *silent* and *voiced* frames have significant energy. However, for *voiced* frames, the energy of spectral components below 2kHz is significantly larger than that of *silent* frames, because in speech, the most relevant spectral information is observed to be concentrated in frequencies upto 2kHz. Thus, a *silent* frame only contains noise and no speech. Hence, σ_w^2, or *R* is simply the variance of the sample values in this frame. In order to consider a single value of *R* for the total speech signal, the mean of the variances of sample values in all *silent* frames has been taken as *R* in our case.

B. Choice of Q: Robustness and Sensitivity Metrics

The choice of Q is done framewise in this case based on the method given in [6]. Let two terms A_k and *B* be defined for a particular frame as

$$\begin{aligned} A_k &= \boldsymbol{H}(\boldsymbol{\phi}\boldsymbol{P}(k-1|k-1)\boldsymbol{\phi}^{\boldsymbol{T}})\boldsymbol{H}^{\boldsymbol{T}} \\ B &= \boldsymbol{H}(\boldsymbol{G}Q\boldsymbol{G}^{\boldsymbol{T}})\boldsymbol{H}^{\boldsymbol{T}} = \sigma_u^2 = Q_f \end{aligned} \tag{13}$$

In case of the speech model, the term A_k denotes the kth instant of the *a priori* state estimation error covariance while B represents the kth instant estimate of the process noise covariance in the measured output. Furthermore, in our case A_k, B and R are all scalars. As mentioned in Section IIIA, R is constant for all frames because it is the variance of the white noise corrupting the speech signal. However, B, though constant for a particular frame, is varied from frame to frame in order to capture the unmodelled process dynamics. This choice of the *framewise constant* B is done using the performance metrics as discussed hereafter.

The two performance metrics J_1, J_2 and a controlling parameter, n_q as given in [6], are defined in this case as:

$$\begin{aligned} J_1 &= [(A_k + B + R)^{-1}R] = \frac{\sigma_w^2}{A_k + \sigma_u^2 + \sigma_w^2} \\ J_2 &= [(A_k + B)^{-1}B] = \frac{B}{A_k + B} = \frac{\sigma_u^2}{A_k + \sigma_u^2} \\ n_q &= log_{10}(B) = log_{10}(\sigma_u^2) \end{aligned} \quad (14)$$

Any mismatch between the assumed process noise covariance σ_u^2 and the actual process noise covariance is due to error in modelling, hence J_2, which is dependent on σ_u^2 is termed as the robustness metric. Similarly, any mismatch between actual R of the measurement and assumed R adversely affects the *a posteriori* estimate. Since it is reflected in J_1, it is termed as the sensitivity metric.

Let the process noise variance, σ_u^2 for a frame be denoted as Q_f. For each frame of speech, a nominal value of Q_f = Q_{f-nom} is taken for initial calculation. This Q_f is then varied as $Q_{f-nom} \times 10^n$ where $n \in \boldsymbol{Z}$. Hence, $n_q = n \times \log_{10} Q_f$ and so, in this case, the metrics are obtained in terms of changing n instead of n_q. For each value of n, corresponding Q_f, J_1 *and* J_2 values are determined.

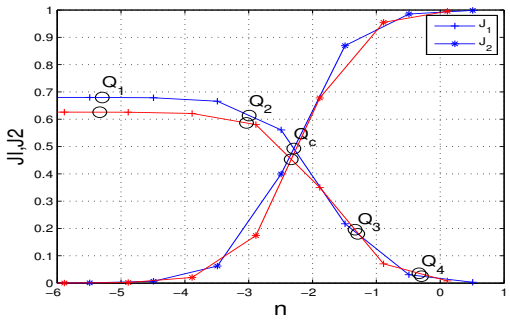


Fig. 1. J_1,J_2 *v/s* n plot for a i) voiced frame (blue) ii) silent frame (red)

The typical plot of the metrics J_1 and J_2 for one voiced frame and one silent frame is shown in Fig 1. It has been observed that J_2 is consistently higher than J_1 for all frames, indicating that the system is robustness prone [9]. If the value of Q_f is increased such that it exceeds R substantially, then from (14), we can say that J_1 reduces to zero while J_2 is high. On the other hand if Q_f is decreased to a small value, then J_2 reduces to zero and J_1 is high, as evident in the graph.

Thus, robust filter performance may be expected for large values of Q_f, whereas small values of Q_f give sensitive filter performance. A trade-off between the two can be achieved by taking the working value of Q_f as the intersection point of J_1 and J_2. In Fig. 1, five values of Q_f have been marked in increasing order, with Q_1 being the lowest and Q_4 being the highest. Q_c is the value of Q_f at intersection of J_1 and J_2. A suitably low choice of Q_f is desirable to keep J_2 low and hence, only a small range around Q_c is likely to be the working range for Q_f. Listening tests and segmental SNR values confirm that performance may be further enhanced by selecting two such values of Q_f, one for *silent* frames and another for *voiced* frames. In this work, the two working values of Q_f taken are, Q_c, that gives equal preference to robustness and sensitivity) and Q_2 ($< Q_c$) that gives greater preference to sensitivity. Two sets of Kalman gains and *a priori* estimates are calculated, each for Q_c and Q_2. The reason for this choice of Q_f is explained in the following section.

IV. Proposed Algorithm

A. Overview

i) The speech signal is windowed into 80ms frames with 10ms overlap [5]. The frames are then classified as *silent* or *voiced*, and measurement noise covariance R is calculated, as explained in Sec. III-A.

ii) For each frame, the pth order LPC coefficients are calculated from noisy speech (in this case, p=15). The state transition matrix ϕ is determined from these coefficients. The prediction error covariance from LPC estimation is taken to be the nominal process noise covariance Q_{f-nom}.

iii) Q_f is varied in the frame as $10^n Q_{f-nom}$ as mentioned before. The last *a posteriori* error covariance matrix of the previous frame is taken as $\boldsymbol{P}$(k-1|k-1) for the calculation of A_k. J_1 and J_2 are calculated according to (14). Ideally, to achieve a perfect trade-off between robustness and sensitivity, Q=Q_c should be selected. However, since the system is inherently robust ($J_2 > J_1$) [9], this choice of Q_f will not suppress noise effectively, especially in *silent* frames. It is indeed observed that $Q_f = Q_c$ leaves a lot of residual noise in processed output. Hence for *silent* frames, we require more sensitive performance. To achieve that, we may select $Q_f = Q_2 = 10^{-0.7}Q_c$. This selection of Q comes with its own disadvantage. While overall noise suppression is achieved, too much sensitivity leads to overcompensation, and some of the spectral components of speech are lost, making it less intelligible. To overcome both these shortcomings, we calculate two sets of Kalman gains and *a priori* state estimates, one each for Q_c and Q_2. J_1 at $Q_f = Q_c$ and Q_2 and $\log_{10}Q_f$ values corresponding to Q_c and Q_2 for all frames have been plotted in Fig 2.

iv) It is observed from Fig 2 that J_1 at Q_c is lower than that at Q_2 for all frames. This matches with our

previous deduction, that J_1 decreases for increasing values of Q_f. However, Kalman gain *K(k)* is in direct relationship with Q_f as shown in Fig 3. It is to be noted that *K(k)* for Q_2 is lesser than that for Q_c. The significance of this is explained in the following subsection.

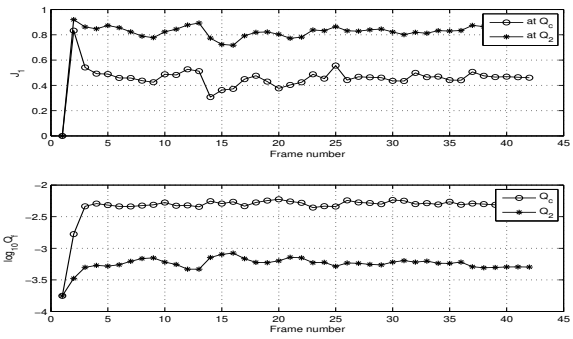


Fig. 2. J_1 and $\log_{10}$ Q_f plot for all frames

B. Kalman Gain Adjustment

For the *k*th sample, let us denote the gain as K_k. Then, K_k and the *a priori* state estimate $\hat{x}$ (k|k-1) are scalars and so, (10) may be rewritten as

$$\hat{x}(k|k) = K_k y(k) + (1 - K_k)\hat{x}(k|k-1) \qquad (15)$$

This simple equation tells us that for a small value of K_k the *a priori* estimate has more weight whereas for a large value of K_k, the noisy measurement is trusted more in calculating the *a posteriori* state estimate. The *silent* frames in processed speech should have a low value of K_k because they depend more on the *a priori* state estimates. In fact, if K_k is high for a *silent* frame, then it borrows heavily from the noise corrupted *silent* frame of noisy signal, and becomes noisy itself. The *voiced* frames, on the other hand, need to depend more on the noisy signal *y(k)*, because it contains valuable spectral information. Hence, K_k needs to be high for *voiced* frames. Another look at the K_k curve in Fig. 3 makes it clear that for a silent frame, K_k corresponding to Q_2 and for *voiced* frames K_k corresponding to Q_c can be expected to yield better overall speech enhancement.

$$\begin{aligned}\hat{\boldsymbol{X}}(k|k) &= \boldsymbol{K}_{Q_f} y(k) + (\boldsymbol{I} - \boldsymbol{K}_{Q_f})(\boldsymbol{H}\hat{\boldsymbol{X}}_{Q_f}(k|k-1)) \\ \boldsymbol{K}_{Q_f} &= \boldsymbol{K}_{Q_c}, \hat{\boldsymbol{X}}_{Q_f} = \hat{\boldsymbol{X}}_{Q_c} \; for \; voiced \; frames \\ \boldsymbol{K}_{Q_f} &= \boldsymbol{K}_{Q_2}, \hat{\boldsymbol{X}}_{Q_f} = \hat{\boldsymbol{X}}_{Q_2} \; for \; silent \; frames \end{aligned} \qquad (16)$$

The original Kalman gain curve for Q_2 and Q_c and resulting Kalman gain curve after this manipulation is shown in Fig 3. The second curve has sharp transitions and large jumps in gain value while going from *silent* frame to *voiced* frame and vice versa.

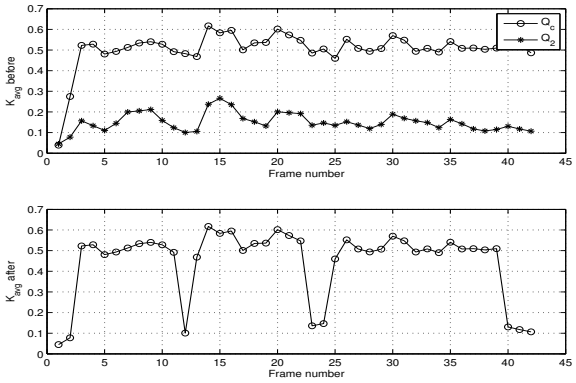
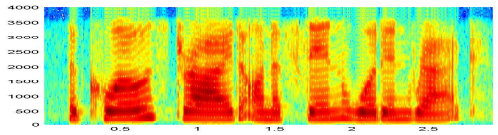


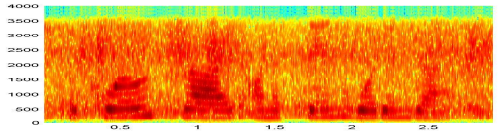
Fig. 3. Kalman gain curve i) before adjustment ii) after adjustment

v) After frame-wise Kalman gain adjustment, the *a posteriori* state estimates $\hat{\boldsymbol{X}}$(k|k) are calculated. These estimates can be referred to as the cleaned signal samples. A fresh set of LPC coefficients are calculated from this cleaned signal. Thereafter, Kalman filtering is done once again on them, without any tuning. This is known as iterative filtering. The resulting set of *a posteriori* state estimates from the second iteration of the Kalman filter form the final processed output samples.

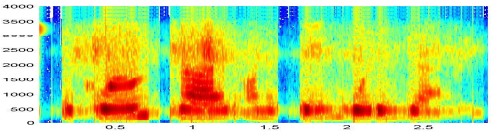
vi) Overlap adding of all the processed frames yields the final enhanced speech output.



(a) Original clean speech spectrum



(b) Noisy speech spectrum



(c) Enhanced speech spectrum

Fig. 4. Results for sp15-" The clothes dried on a thin wooden rack."

V. Experiment And Results

The algorithm has been tested on a short speech signal from the NOIZEUS speech corpus (http://ecs.utdallas.edu/loizou/speech/noizeus/), which has been corrupted with an Additive Gaussian White Noise (AGWN) of 5dB SNR. Both the noisy

and clean signal are available for analysis. The spectrograms for the original and the noisy signals obtained from the corpus and the enhanced speech signal resulting from proposed method, are shown in Fig 4. Visually, the spectrogram of the processed speech is much cleaner than the one of noisy speech. Cleaning is especially evident in the intermediate silent zones. Audibly, the processed speech obtained using the proposed method sounds least noisy, while keeping its intelligibility intact. The segmental SNR values for processing with various values of Q_f are given in Table I. Schwerin and Paliwal [10] suggest that segmental SNR is more consistent with subjective preference scoring than several other methods and this is corroborated in our results both visually and audibly. As seen from the table, our proposed method gives the highest value of segmental SNR, i.e, the cleanest speech. Listening tests also agree with the above.

TABLE I
TABLE OF SEGMENTAL SNR VALUES FOR FILTERING WITH VARIOUS VALUES OF Q

Q	*Seg SNR Noisy(dB)*	*Seg SNR Processed(dB)*
Q_{nom}	-3.0425	1.4183
Q_1	-3.0425	-0.4577
Q_2	-3.0425	1.2721
Q_c	-3.0425	1.2162
Q_3	-3.0425	1.0813
Q_4	-3.0425	1.0337
Proposed method	-3.0425	2.2548

Overall inspection of Fig 4 and Table I indicates that the proposed algorithm suppresses noise effectively in the *silent* frames while keeping the spectral components intact in the *voiced* frames. In this context, it must be noted that the computational load increases as a result. However, for short segments of speech, and particularly in offline speech enhancement, this increase in execution time or the associated computational complexity is admissible for the sake of better noise removal. In conclusion, in offline speech enhancement where only noisy measured signal is available,our proposed algorithm performs sensitively with respect to the *silent* frames while acting robustly in terms of the noise present in the *voiced* frames.

VI. CONCLUSION

In this paper, Kalman filter tuning and Kalman gain adjustment for offline speech enhancement has been discussed in detail. A new method has been proposed where the Kalman filter is first tuned followed by selection of Kalman gain based on *silent* and *voiced* frame distinction. The sensitivity and robustness metrics and their role in choosing a suitable process noise covariance, Q_f for each frame has been explained. It has been shown that toggling between two values of Q_f for the *silent* and the *voiced* frames gives us the best results. Experiments have been performed on a speech signal, and the results show that the *silent* frames are cleaned and the *voiced* frames retain their spectral components and maintain their intelligibility. In future, this work can be extended by further altering the obtained Kalman gain trajectory and smoothing it to achieve even better results.

REFERENCES

[1] R. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[2] K. Paliwal and A. Basu, "A speech enhancement method based on kalman filtering," in *Proc. ICASSP*, vol. 12, 1987.

[3] J. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," in *Proc. IEEE*, vol. 67, no. 12, 1979.

[4] J. Gibson, B. Koo, and S. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Signal Process.*, vol. 39, no. 8, pp. 1732–1742, 1991.

[5] S. So and K. Paliwal, "Suppressing the influence of additive noise on the kalman gain for low residual noise speech enhancement," *Speech Communication*, vol. 53, pp. 355–378, 2011.

[6] M. Saha, R. Ghosh, and B. Goswami, "Robustness and sensitivity metrics for tuning the extended kalman filter," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 4, pp. 964–971, 2014.

[7] K. Paliwal, "Estimation of noise variance from the noisy ar signal and its application in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 292–294, 1988.

[8] Y. Hu and P. Loizou, "Subjective evaluation and comparison of speech enhancement algorithms," *Speech Communication*, vol. 49, pp. 588–601, 2007.

[9] S. Roy, R.Ghosh, and B. Goswami, "Comparison of cartesian and polar estimates of the extended kalman filter for different choices of the robustness and sensitivity metrics," in *Proceedings of the Third International Conference on Advances in Control and Optimization of Dynamical Systems*, 2014.

[10] B. Schwerin and K. Paliwal, "Using stft real and imaginary parts of modulation signals for mmse-based speech enhancement," *Speech Communication*, vol. 58, pp. 49–68, 2014.