An Iterative Kalman Filter with Reduced-Biased Kalman Gain for Single Channel Speech Enhancement in Non-stationary Noise Condition

Sujan Kumar Roy and Kuldip K. Paliwal
Signal Processing Laboratory, Griffith School of Engineering, Griffith University, Brisbane, QLD, Australia, 4111
Email: sujankumar.roy@griffithuni.edu.au, k.paliwal@griffith.edu.au

***Abstract*—This paper presents an iterative Kalman filter (IT-KF) with a reduced-biased Kalman gain for single channel speech enhancement in Non-stationary Noise Conditions (NNCs). The proposed IT-KF aims to offset the bias in Kalman gain through efficient parameter estimation leading to improve the speech enhancement performance. To do this, we introduce a Decision Directed (DD) and *a posteriori* SNR based noise variance estimation method controlled through Speech Activity Detector (SAD). The proposed SAD incorporates a majority voting of three distinct SAD fusions. The LPC parameters are computed from the pre-smoothing of noisy speech. With these initial estimated parameters, an IT-KF processes the noisy speech at first iteration. The parameters are re-estimated from the processed speech, re-adjust the Kalman gain, and the process is repeated at second iteration. It is shown that the adjusted Kalman gain enables the IT-KF to minimize the remaining artifacts of the processed speech, yielding the enhanced speech. Extensive simulation results reveal that the proposed method outperforms other benchmark methods in NNCs for a wide range of SNRs.**

***Index Terms*—speech enhancement, kalman filter, non-stationary noise, speech activity detector, pre-smoothing**

I. Introduction

The background noises degrade the speech signal during voice communication over telephone, speech recognition, speech coding, etc. A Speech Enhancement Algorithm (SEA) acts as a front-end tool for these applications by providing an estimate of clean speech. Many SEAs, such as Spectral Subtraction (SS) [1], MMSE [2], Kalman Filter (KF) [3] etc., have been proposed over decades. However, the speech enhancement performance varies over the SEAs and deteriorates when the speech is corrupted by non-stationary noise.

The enhanced speech by SS method suffers from musical noise and distortion due to under/over subtraction of the estimated noise spectrum from the noisy spectrum [4]. Although, the MMSE method [2] shows improvement over SS, the efficiency of this method completely depends on the accuracy of a priori and a posteriori SNR computation in noisy condition. In [5], regional statistics based noise PSD tracking method has been proposed, which could be used to further improve the performance of MMSE method [2]. However, significant background noise still remains in the enhanced speech. Paliwal and Basu for the first-time introduced KF based SEA [3]. Here, the LPC parameters are computed from the clean speech and operates under white noise condition. Gibson et al. introduced an iterative KF (IT-KF) for colored noise suppression [6], where the LPC parameters are estimated by processing the noisy frame with 3-4 iterations. Due to parameter estimation issues, the Kalman gain becomes biased and a significant amount of background noise remains in the enhanced speech. So and Paliwal [7] studied the impact of the Long Tapered Window (LTW) for LPC estimation that influences the KF performance. However, windowing impacts the KF performance to some extent. In [8], a Sub-band (SB) IT-KF has been introduced. It employs an IT-KF to the partially reconstructed High Frequency (HF) SBs among the 16 decomposed SBs, while keeping the Low Frequency (LF) SBs unchanged. The enhanced speech is obtained by adding the HF enhanced speech with the LF SBs. However, the LF SBs could also be affected by additive noise when processing the non-stationary noise corrupted speech. Recently, Roy and Paliwal [9] introduced a NIT-KF based SEA to minimize the biasing effect of Kalman gain through efficient parameter estimation. However, some artifacts still remain in the enhanced speech.

Although, most of the SEAs perform relatively well under white noise condition, the performance becomes degraded in NNCs. The authors in [9] showed that the KF performance deteriorates due to the biased estimate of Kalman gain under NNCs. In this paper, we focus on the further adjustment of the biased Kalman gain through improving the initial estimate of parameters followed by re-estimation of these in subsequent iterations of IT-KF. Specifically, the initial estimated parameters are applied to IT-KF for filtering the noisy speech at first iteration. The parameters are re-estimated to compute the Kalman gain, and the process is repeated again at second iteration. The adjusted Kalman gain in IT-KF is effective in minimizing the remaining artifacts of the processed speech, yielding the enhanced speech. The efficiency of the proposed method with respect to other benchmark

Manuscript received December 24, 2018; revised February 18, 2019.

doi: 10.18178/ijsps.7.1.7-13

SEAs in terms of subjective and objective testing is reported in this paper.

The rest of the paper is organized as follows. Section II describes the conventional KF for speech enhancement and problem statement in II-A. Section III introduces the proposed speech enhancement system followed by parameter estimation in II-A, proposed SAD algorithm in III-A (1), proposed noise variance estimation in III-A(2), estimation of initial LPC parameters in III-A(3), proposed parameter re-estimation method in III-A(4), summary of the proposed IT-KF based SEA in III-B, and optimality comparison of Kalman gain in III-C. Section IV describes the speech enhancement experiment, where the simulation setup is given in IV-A, IV-B deals with the experimental results and discussion. Section V gives some concluding remarks and future research directions.

II. Conventional KF for Speech Enhancement

The noisy speech $y(n)$ (n^{th} sample) captured by a single microphone is represented as

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

where $s(n)$ is the clean speech, $v(n)$ is the additive noise with variance σ_v^2 and uncorrelated with $s(n)$.

The clean speech $s(n)$ in eq. (1) can be represented with a p^{th} order LPCs (a_i's) as [10]

$$s(n) = -\sum_{i=1}^{p} a_i s(n-i) + u(n) \tag{2}$$

where $u(n)$ is a white Gaussian excitation with zero mean and a variance of σ_u^2.

Eqs. (1) and (2) are used to form the following SSM (where the **bold** faced letters represent vectors/ matrices).

$$\boldsymbol{x}(n) = \boldsymbol{\Psi}\boldsymbol{x}(n-1) + \boldsymbol{c}u(n) \tag{3}$$

$$y(n) = \boldsymbol{d}\boldsymbol{x}(n) + v(n) \tag{4}$$

where $\boldsymbol{\Psi}$ is the state transition matrix containing the a_i's, $\boldsymbol{x}(n) = [s(n-p+1) \quad s(n-p+2) \quad \dots \quad s(n)]^T$ is the state vector, $\boldsymbol{d} = \boldsymbol{c}^T = [0 \quad 0 \quad 0 \quad 1]$ are the measurement vectors for the excitation and observation noises, respectively.

For a particular frame, KF computes an unbiased and linear MMSE estimate $\hat{\boldsymbol{x}}(n|n)$ of $\boldsymbol{x}(n)$ at time n, given $y(n), y(n-1), \dots, y(1)$ by using the following recursive equations [3]

$$\hat{\mathbf{x}}(n|n-1) = \boldsymbol{\Psi}\hat{\mathbf{x}}(n-1|n-1) \tag{5}$$

$$\boldsymbol{\Sigma}(n|n-1) = \boldsymbol{\Psi}\boldsymbol{\Sigma}(n-1|n-1)\boldsymbol{\Psi}^T + \mathbf{c}\sigma_u^2\mathbf{c}^T \tag{6}$$

$$\mathbf{K}(n) = \boldsymbol{\Sigma}(n|n-1)\mathbf{d}^T(\mathbf{d}\boldsymbol{\Sigma}(n|n-1)\mathbf{d}^T + \sigma_v^2)^{-1} \tag{7}$$

$$\hat{\mathbf{x}}(n|n) = \hat{\mathbf{x}}(n|n-1) + \mathbf{K}(n)(y(n) - \mathbf{d}\hat{\mathbf{x}}(n|n-1)) \tag{8}$$

$$\boldsymbol{\Sigma}(n|n) = (\mathbf{I} - \mathbf{K}(n)\mathbf{d})\boldsymbol{\Sigma}(n|n-1) \tag{9}$$

The estimated speech at time n is given by $\hat{s}(n) = \boldsymbol{d}\hat{\boldsymbol{x}}(n|n)$. The above procedure is repeated for the following frames, yielding the enhanced speech $\hat{s}(n)$.

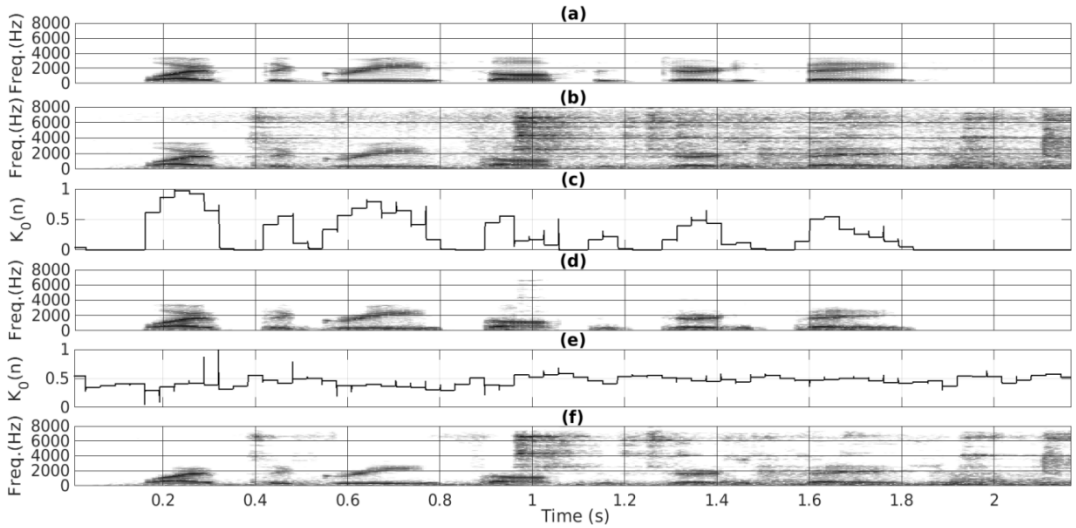

Figure 1. The impact of σ_u^2 and σ_v^2 on biased $K_0(n)$ of NIT-KF: (a) spectrogram of clean speech, (b) spectrogram of noisy speech (corrupted by 0 dB restaurant noise), (c) $K_0(n)$, where σ_u^2 and σ_v^2 are computed in ideal case, (d) spectrogram of enhanced speech (ideal case, PESQ=2.42), (e) $K_0(n)$, where σ_u^2 and σ_v^2 are computed from noisy speech, (f) spectrogram of enhanced speech (noisy case, PESQ=2.07).

Gibson et al. introduced an IT-KF by repeating eqs. (5)-(9) iteratively, where $\boldsymbol{\Psi}$ is formed with the p^{th} and q^{th} order LPCs of $s(n)$ and $v(n)$. The parameters are re-estimated at the end of each iteration leading to increases the computational complexity. To make computationally efficient, Roy et al. showed that the $\boldsymbol{\Psi}$ of IT-KF could be formed with the LPCs of $s(n)$ only and effective for non-stationary noise suppression [8]. Unlike the IT-KF in [8], the proposed IT-KF re-estimates the parameters in the subsequent iterations differently based on SAD.

A. Problem Statement

Though the conventional KF works reasonably well for the stationary noise condition, its performance suffers in NNCs. Roy and Paliwal showed that the poor estimates of LPC parameters ($\{a_i\}$ and σ_u^2) and noise variance (σ_v^2) introduce biasing effect to the first component ($K_0(n)$) of Kalman gain $\boldsymbol{K}(n)$, particularly during the silent activity, resulting a significant amount of residual noise in the enhanced speech [9]. We will briefly review the impact

of biased $\boldsymbol{K}(n)$ on SE performance. To do this, we further simplify the $\boldsymbol{K}(n)$ (eq. (7)) to represent $K_0(n)$ as [7]

$$K_0(n) = \frac{\Sigma_0(n|n-1)}{\Sigma_0(n|n-1)+\sigma_v^2} \quad (10)$$

where $\Sigma_0(n|n-1)$ corresponds to prediction error σ_u^2 of the first component of the *a priori* state estimate $\hat{\boldsymbol{x}}(n|n-1)$.

By replacing $\Sigma_0(n|n-1)$ with σ_u^2, eq. (10) could be represented as

$$K_0(n) = \frac{\sigma_u^2}{\sigma_u^2+\sigma_v^2} \quad (11)$$

To examine the impact of the biased $K_0(n)$ on speech enhancement (SE) by KF, we further simplify the *a posteriori* state estimate $\hat{\mathbf{x}}(n|n)$ (eq. (8)) and represent it in a scalar form ($\hat{x}(n|n)$) as [7]

$$\hat{x}(n|n) = K_0(n)y(n) + (1-K_0(n))\hat{x}(n|n-1) \quad (12)$$

In special case of $\sigma_v^2 = 0$, according to eq. (11), the $K_0(n)$ becomes unity and the output is equal to $y(n)$ (eq. (12)). Whereas $\sigma_u^2 = 0$ during the silent activity, then $K_0(n) = 0$ and no corrupting noise from $y(n)$ is passed to the output (enhanced speech).

In ideal case, $\{a_i\}$, σ_u^2, and σ_v^2 are computed from the clean speech and noise, respectively. Thus, the computed $K_0(n)$ in ideal case shows a smooth transition between 0 and 1 depending on the silent/speech activity. The smooth $K_0(n)$ could blend $y(n)$ and $\hat{x}(n|n-1)$ (eq. (12)) in an effective manner, yielding better $\hat{x}(n|n)$. Therefore, the enhanced speech (Fig. 1(d)) obtained by ideal KF is almost identical to the clean speech (Fig. 1(a)).

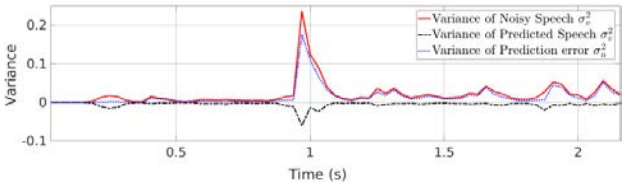

Figure 2. Comparing the variances (σ_v^2) of noisy speech Fig. 1(b) and predicted speech (inverted) with the prediction error variance (σ_u^2) for the same experimental setup used in Fig. 1.

Whereas in practice, the $\{a_i\}$, σ_u^2, and σ_v^2 are computed from the noisy speech. Thus, the predicted σ_u^2 and σ_v^2 become worse and rise up to 1 (normalized form). For example, it can be seen from Fig. 2 that the predicted σ_v^2 and σ_u^2 remain closer from 0.9s to 2.18s. Therefore, the computed $K_0(n)$ in noisy case remains biased around 0.5 between 0.9s to 2.18s (Fig. 1 (e)). The $K_0(n)$ for other regions varies accordingly. With 0.5 biased $K_0(n)$ according to eq. (12), 50% of $y(n)$ is passed to the output.

As a result, the corresponding enhanced speech contains 50% background noise as clearly visible in Fig. 1(f) specifically ranging from 0.9s to 2.18s. This is attributed as the biasing effect of Kalman gain.

This paper aims to offset the bias in $K_0(n)$ through improving the initial estimate of $\{a_i\}$, σ_v^2, and σ_u^2 followed by re-estimation of these in subsequent iterations of IT-KF.

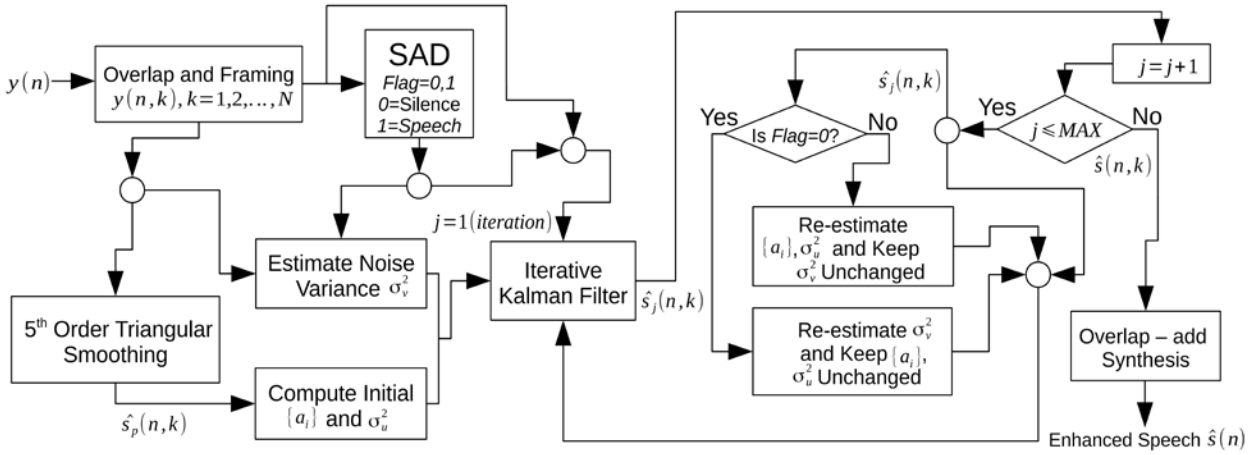

Figure 3. Schematic diagram of the proposed IT-KF based speech enhancement system.

III. Proposed IT-KF Based SEA

Fig. 3 shows the schematic diagram of proposed SEA. Firstly, $y(n)$ is converted through overlap and windowing into frames $y(n,k)$, where n is the time index ($n = 1,2,3,\ldots,M$) and k is the frame index ($k = 1,2,3,\ldots,N$). We have used 50% overlapped Kaiser window with $\beta = 2.5$ for generating $y(n,k)$ as can be found effective in terms of bias reduced $K_0(n)$ [9].

A. Parameter Estimation

The σ_v^2 is computed from the estimated noise spectrum $|\hat{V}(m,k)|$, and the LPC parameters ($\{a_i\}$ and σ_u^2) are computed from a *pre-smoothing* speech $\hat{s}_p(n,k)$ of noisy $y(n,k)$. Whereas $\{a_i\}$, σ_u^2, and σ_v^2 are re-estimated in the subsequent iterations of IT-KF based on SAD. The next sections describe the proposed SAD and parameter estimation methods.

1) Proposed SAD Method: The proposed SAD is implemented through a majority voting (MV) of three distinct SAD fusions corresponding to *Spectral Flatness* (SF), *zero-crossing rate-weighted root mean square energy* (ZCRMS), and *Kaiser-Teager Energy* (KTE). It is observed that the SF approaches 1/0 depending on the silent/speech activity [11]. The degree of speech activity

could be predicted through ZCRMS when it rises to 1, while close to 0 for silent frames. Whereas during the speech activity, KTE rises up to 1 and gives a prominent local peak, while goes down to 0 for silent frames [12].

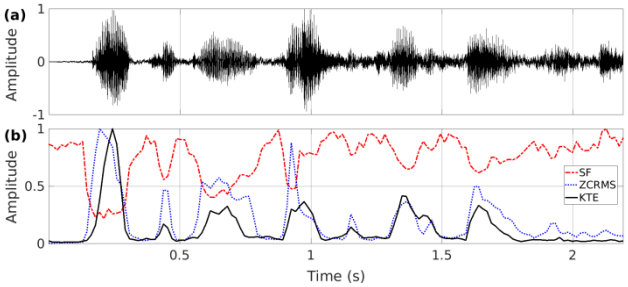

Figure 4. (a) Noisy speech (corrupted by 0 dB restaurant noise), (b) computed SF, ZCRMS, and KTE from (a).

In noisy conditions, Fig. 4 reveals that the SF still varies between 0 to 1 depending on the speech/silent activity. Whereas the ZCRMS and KTE rise up to 1 once the speech activity is present and approaching 0 at silent activity. However, to make the SAD robust against noise, the threshold for each feature continually updated on a framewise basis to implement the corresponding SAD fusions. MV takes the average of SAD fusions $\in \{0,1\}$ (0: *silent* and 1: *speech* activity) for each frame and speech activity is present if $MV > 0.5$, otherwise, silent.

For k^{th} frame, the SF (denoted by η) is computed as [11]

$$\eta(k) = \frac{\sqrt[L]{\prod_{m=0}^{L-1} |Y(m,k)|}}{\frac{1}{L}\sum_{m=0}^{L-1} |Y(m,k)|} \quad (13)$$

where $|Y(m,k)|$ is the magnitude spectrum of $y(n,k)$, m is the acoustic frequency bin index, and $m = 1,2,\ldots,L$.

For k^{th} frame, the ZCRMS (denoted by ζ) is given by [12]

$$\zeta(k) = \frac{\sqrt{\frac{1}{M}\sum_{n=0}^{M-1} y^2(n,k)}}{\frac{1}{2M}\sum_{n=1}^{M} |sign(y(n,k)) - sign(y(n-1,k))|} \quad (14)$$

For k^{th} frame, the KTE (denoted by λ) is given by [12]

$$\lambda(k) = \sqrt{\sum_{n=0}^{M-1} \{y^2(n,k) - y(n-1,k)y(n+1,k)\}} \quad (15)$$

By assuming the 5 starting $y(n,k)$'s are silent, the proposed SAD algorithm is given below (where t_η, t_ζ, and t_λ are the adaptive threshold of η, ζ, and λ, respectively)

Algorithm 1: Proposed SAD Algorithm

1) **Initialization**

$f_\eta = 0, \quad f_\zeta = 0, \quad f_\lambda = 0$

$\mathbb{S}_\eta = \frac{1}{5}\sum_{k=1}^{5} \eta(k), \quad \mathbb{S}_\zeta = \frac{1}{5}\sum_{k=1}^{5} \zeta(k), \quad \mathbb{S}_\lambda = \frac{1}{5}\sum_{k=1}^{5} \lambda(k)$

$FLAG(k) = 0$ for $k = 1,2,3,4,5$

2) **for** $k = 6 \quad to \quad N$ **do** [*framewise processing loop*]

a) **Update Thresholds**

$\mathbb{S}_\eta = \mathbb{S}_\eta + \eta(k), \quad t_\eta = \mathbb{S}_\eta / k$

$\mathbb{S}_\zeta = \mathbb{S}_\zeta + \zeta(k), \quad t_\zeta = \mathbb{S}_\zeta / k$

$\mathbb{S}_\lambda = \mathbb{S}_\lambda + \lambda(k), \quad t_\lambda = \mathbb{S}_\lambda / k$

b) **if** $\eta(k) < t_\eta$ **then**

$f_\eta = 1$

elseif $\zeta(k) > t_\zeta$ **then**

$f_\zeta = 1$

elseif $\lambda(k) > t_\lambda$ **then**

$f_\lambda = 1$

end if

c) $MV = (f_\eta + f_\zeta + f_\lambda)/3$

d) **if** $MV > 0.5$ **then**

$FLAG(k) = 1$ [*Speech Activity*]

else

$FLAG(k) = 0$ [*Silent Activity*]

end if

end for

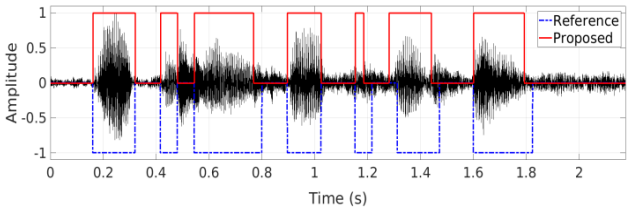

Figure 5. Comparing the reference and detected SAD flags for clean (Fig. 1 (a)) and noisy speech (corrupted by 5 dB restaurant noise).

It can be seen from Fig. 5 that few miss-detections are found between the detected and reference SAD flags. Note that the reference SAD flags are generated by visually inspecting the clean speech (Fig. 1 (a)) frames (0: *silence* and -1: *speech* activity).

2) Proposed σ_v^2 Estimation Method: The initial noise periodogram $|\hat{V}(m,k)|^2$ is computed by assuming the 5 starting $y(n,k)$'s are silent as

$$|\hat{V}(m,k)|^2 = \frac{1}{5}\sum_{k=1}^{5} |Y(m,k)|^2 \quad (16)$$

During silent activity of $y(n,k)$ $(k > 5)$, $|\hat{V}(m,k)|^2$ is updated by using the DD approach as [4]

$$|\hat{V}(m,k)|^2 = G|\hat{V}(m,k-1)|^2 + (1-G)|Y(m,k)|^2 \quad (17)$$

where G is a smoothing parameter and set to 0.9.

In stationary noise conditions, the estimation of $|\hat{V}(m,k)|^2$ during non-speech activity is effective [4]. Since the non-stationary noise is characterized by time varying amplitude, the active speech regions also affected by noise. Therefore, the traditional DD approach is not appropriate to estimate $|\hat{V}(m,k)|^2$ in NNCs. To address this issue, we compute the *a posteriori* SNR (denoted by γ) during speech activity to asses the amount of noise available. For k^{th} frame, $\gamma(k)$ is computed as

$$\gamma(k) = 10\log_{10}\left(\frac{|Y(m,k)|^2}{|\hat{V}(m,k-1)|^2}\right) \quad (18)$$

It is observed that the $\gamma(k)$ becomes lower (mostly negative) if the active speech region is highly affected by additive noise. Thus, we compute an adaptive threshold (t_γ) by taking the average of $\gamma(k)$'s up to frame k during processing the k^{th} frame as

$$t_\gamma = \frac{1}{k}\sum_{i=1}^{k} \gamma(i) \quad (19)$$

During the speech activity of $y(n,k)$, if $\gamma(k) \leq t_\gamma$, $|\hat{V}(m,k)|^2$ is updated by eq. (17), otherwise, keep $|\hat{V}(m,k)|^2$ unchanged. The σ_v^2 is computed from $\hat{v}(n,k)$ (IDFT of $|\hat{V}(m,k)|exp[\angle Y(m,k)]$) as

$$\sigma_v^2 = \frac{1}{M}\sum_{n=0}^{M-1} \hat{v}^2(n,k) \quad (20)$$

3) Initial $\{a_i\}$ and σ_u^2 Computation: The LPC parameters ($\{a_i\}$ and σ_u^2) are very sensitive to noise, specially at low SNRs. The existing IT-KF methods compute these parameters from the noisy speech at first iteration and re-estimate at the subsequent iterations [6, 8]. Due to compute these parameters from noisy speech at first iteration, the σ_u^2 rise up to 1 and introduce biasing effect in $K_0(n)$ (eq. (11)). To address this issue, we employ a 5^{th} order triangular smoothing to noisy $y(n,k)$ for reducing the noise effect, giving a *pre-smoothing* speech $\hat{s}_p(n,k)$ as [13]

$$\hat{s}_p(n,k) = \frac{1}{9}\sum_{i=-(L_s-1)/2}^{+(L_s-1)/2} w[i+(L_s+1)/2]y(n-i,k) \quad (21)$$

where $w = [1 \quad 2 \quad 3 \quad 2 \quad 1]$ is a 5^{th} order triangular smoothing window and L_s is the length of w.

Then compute $\{a_i\}$ and σ_u^2 from $\hat{s}_p(n,k)$ by using the autocorrelation based method [10]. The performance comparison of the initial estimated $\{a_i\}$ is shown in Fig. 6.

4) Re-estimation of $\{a_i\}$, σ_u^2, and σ_v^2 in Proposed IT-KF: The conventional IT-KF [6] re-estimates the $\{a_i\}$ and σ_u^2 from $\hat{s}_j(n,k)$ (processed speech at j^{th} iteration) while no action takes on σ_v^2. Since each iteration gives more refined enhanced speech, the additive noise effect becomes reduced. Therefore, it is also necessary to update σ_v^2 from $\hat{s}_j(n,k)$ as introduced in our proposed IT-KF method. To make the re-estimation of parameters effective, unlike conventional IT-KF [6], we do it in a different manner based on SAD. Specifically, during the silent activity, the $\hat{s}_j(n,k)$ is filled up completely with noise. Thus, σ_v^2 is updated at silent activity of $\hat{s}_j(n,k)$ while keeping the $\{a_i\}$ and σ_u^2 unchanged. During speech activity of $\hat{s}_j(n,k)$, the $\{a_i\}$ and σ_u^2 are re-estimated and keeping σ_v^2 unchanged.

B. Summary of the Proposed IT-KF Based SEA

For k^{th} frame, by letting *MAX=2*, the proposed IT-KF based SEA is summarized below.

Algorithm 2: Proposed IT-KF Based SEA

1) **Initialization:**
   a) Extract $FLAG(k)$ from $y(n,k)$ by SAD (3.1.1)
   b) Compute σ_v^2 from $y(n,k)$ (3.1.2)
   c) Compute initial $\{a_i\}$ and σ_u^2 from $y(n,k)$ (3.1.3)
   d) Set $\hat{\mathbf{x}}_1(0|0) = 0$ and $\mathbf{\Sigma}_1(0|0) = [0]_{p\times p}$
   e) Form $\mathbf{\Psi}$ with estimated $\{a_i\}$
   f) Set $\hat{s}_0(n,k) = y(n,k)$
2) **for** $j = 1$ *to* MAX **do** [*iteration loop*]
   a) **for** $n = 1$ *to* M **do** [*samplewise processing loop*]

$$\hat{\mathbf{x}}_j(n|n-1,k) = \mathbf{\Psi}_j\hat{\mathbf{x}}_j(n-1|n-1,k) \quad (22)$$

$$\mathbf{\Sigma}_j(n|n-1,k) = \mathbf{\Psi}_j\mathbf{\Sigma}_j(n-1|n-1,k)\mathbf{\Psi}_j^T + \mathbf{c}\sigma_u^2\mathbf{c}^T \quad (23)$$

$$e_j(n,k) = \hat{s}_{j-1}(n,k) - \mathbf{d}\hat{\mathbf{x}}_j(n|n-1,k) \quad (24)$$

$$\mathbf{K}_j(n,k) = \mathbf{\Sigma}_j(n|n-1,k)\mathbf{d}^T(\mathbf{d}\mathbf{\Sigma}_j(n|n-1,k)\mathbf{d}^T + \sigma_v^2)^{-1} \quad (25)$$

$$\hat{\mathbf{x}}_j(n|n,k) = \hat{\mathbf{x}}_j(n|n-1,k) + \mathbf{K}_j(n,k)e_j(n,k) \quad (26)$$

$$\mathbf{\Sigma}_j(n|n,k) = (\mathbf{I} - \mathbf{K}_j(n,k)\mathbf{d})\mathbf{\Sigma}_j(n|n-1,k) \quad (27)$$

$$\hat{s}_j(n,k) = \mathbf{d}\hat{\mathbf{x}}_j(n|n,k) \quad (28)$$

end for [*end of samplewise processing loop*]
   b) Re-estimate $\{a_i\}$, σ_u^2, and σ_v^2 from $\hat{s}_j(n,k)$ (3.1.4)
end for [*end of iteration loop*]
3) Set $\hat{s}(n,k) = \hat{s}_j(n,k)$ and employ the overlap-add synthesis to $\hat{s}(n,k)$, yielding the enhanced speech $\hat{s}(n)$

C. Optimality Comparison of Kalman Gain

Fig. 6 shows that the re-estimated LPC envelope at 2^{nd} iteration of IT-KF is sharper than the initial LPC envelope and closer to the clean speech envelope. Whereas the LPC envelope computed from the corresponding noisy frame deviates a bit from the clean envelope. Due to improved $\{a_i\}$, the re-estimated σ_u^2 becomes lower than that of the initial estimated σ_u^2. Also, the re-estimation of σ_v^2 during silent activity makes it more effective. Therefore, the σ_u^2 and σ_v^2 offset the bias in $K_0(n)$ effectively. It can be seen from Fig. 7 that the adjusted $K_0(n)$ at 2^{nd} iteration of IT-KF is almost free of biasing effect and shown smooth transition at the edges as like ideal case $K_0(n)$, even at low SNR of 0 dB. Whereas the $K_0(n)$ computed from noisy speech is biased around 0.5 almost the entire trajectory.

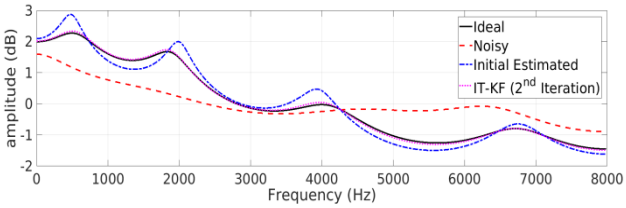

Figure 6. LPC spectrum comparison computed from the clean, noisy, $\hat{s}_p(n,k)$, and IT-KF (2^{nd} iteration) for the same setup used in Fig. 1.

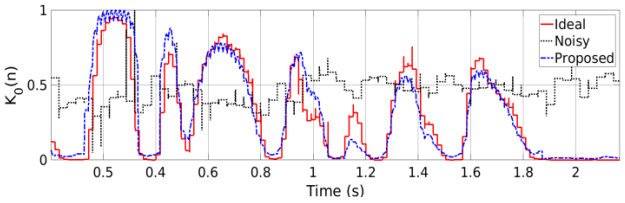

Figure 7. Comparing the trajectory of $K_0(n)$ computed through ideal, noisy, and proposed cases with the same experimental setup used in Fig. 1.

IV. Speech Enhancement Experiment

A. Simulation Setup

To evaluate the performance of the proposed SEA, 30 speech sentences belonging to six speakers are taken from the NOIZEUS corpus sampled at 16 kHz [14, Chapter 12]. To perform experiments, we generate a stimuli set that has been corrupted by restaurant and

babble noises for a wide range of SNRs (-5dB to 15dB). The objective quality evaluation was carried out by PESQ and spectrogram analysis [14, Chapter 11]. We have used the quasi-stationary speech transmission index (QSTI) for objective intelligibility testing, which provides a rating in (%) [15]. The subjective evaluation was performed on two sentences (1 male and 1 female) randomly chosen from the stimuli set. Five English speaking listeners rate the quality of the enhanced speech obtained by all methods based on a pre-defined scale as introduced in the mean opinion score (MOS) test [14]. During this test, the listeners have no information about the proposed and benchmark methods to make it unbiased. The efficiency of the proposed method (IT-KF) is carried out by comparing it with other benchmark methods, such as sub-band iterative KF (SBIT-KF) [8], MMSE with regional statistics (MMSE-RS) [5], and long tapered window based KF (LTW-KF) [7].

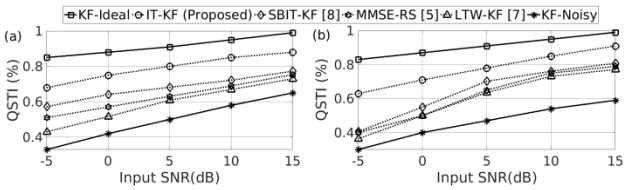

Figure 8. Average QSTI (%) comparison between the proposed and other SEAs on NOIZEUS corpus corrupted with: (a) restaurant and (b) babble noises for SNRs (-5dB to 15dB).

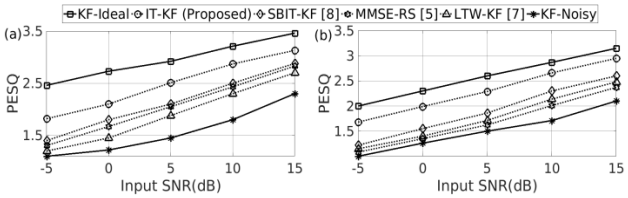

Figure 9. Average PESQ comparison between the proposed and other SEAs on NOIZEUS corpus corrupted with: (a) restaurant and (b) babble noises for SNRs (-5dB to 15dB).

B. Simulation Results and Discussion

The QSTI results for the restaurant noise experiment in Fig. 8 (a) specifies that the IT-KF method gives QSTI of 0.68 to 0.88 at all SNRs, whereas the competitive methods give QSTI ranged between 0.57 to 0.77. The QSTI for the babble noise experiment in Fig. 8 (b) suggests that the proposed IT-KF method yields 0.63 to 0.91 followed by the other methods give average QSTI ranging from 0.4 to 0.8. The high QSTI of the proposed method reveals that the enhanced speech provides better intelligibility than the benchmark methods for a wide range of SNRs.

The average PESQ results for the restaurant noise experiment are shown in Fig. 9 (a). It can be seen from this figure that the proposed method (PESQ between 1.83 and 3.13) is better than the other methods (with PESQ ranging from 1.4 to 2.88). Similar results are obtained for the babble noise experiment as shown in Fig. 9 (b). Note that the high PESQ indicates the enhanced speech provides natural quality of sound, whereas quality degradation for low PESQ. Therefore, the PESQ evaluation results in Fig. 9 reveals that the proposed method ensures better quality in the enhanced speech over benchmark methods.

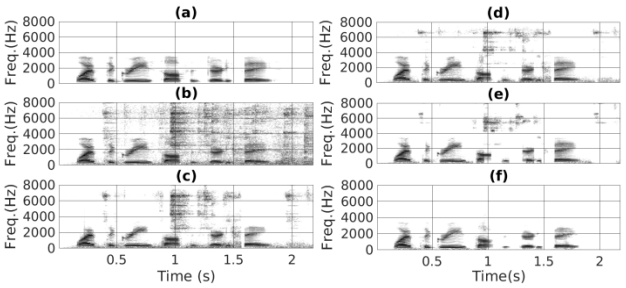

Figure 10. Spectrogram comparison among: (a) clean speech (as in Fig. 2.1(a)), (b) noisy speech (corrupted with restaurant noise at 5 dB SNR), with enhanced speech obtained through (c) LTW-KF [7], (d) MMSE-RS [5], (e) SBIT-KF [8], and (f) IT-KF (Proposed) methods.

These methods also compared in terms of their spectrograms in Fig. 10. Here, it can be seen that the IT-KF enhanced speech is almost free of noise floor, whereas the existing SEAs contain a significant amount of noise floor. The informal listening tests also confirm that the existing methods produce very annoying sounds as compared to the negligible audio artifacts produced by the proposed method. However, when compared with clean speech spectrogram, the proposed method introduced a little bit distortion in the enhanced speech. This may result due to an *over-suppression* of the spectral valleys by the adjusted Kalman gain during the speech activity.

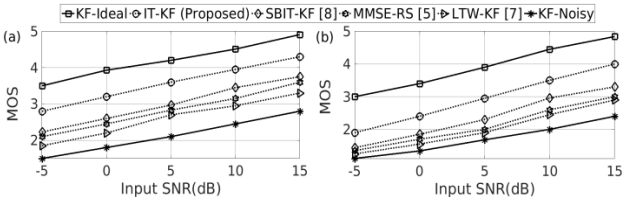

Figure 11. Average MOS comparison between the proposed and other SEAs on NOIZEUS corpus corrupted with: (a) restaurant and (b) babble noises for SNRs (-5dB to 15dB).

Fig. 11 shows the subjective MOS results for a male sentence "*The birch canoe slid on the smooth planks*" and a female sentence "*Bring your best compass to the third class*". It can be seen from this figure that the proposed method was preferred by the listeners effectively and gives superior quality over other methods. Specifically, the restaurant noise experimental results (Fig. 11 (a)) reveals that the proposed method gives average MOS of 2.68 to 4.11 at all SNRs, whereas the competitive methods ranging from 2.23 to 3.75. The proposed method shows continuous improvement over benchmark methods for the babble noise experiment (Fig. 9 (b)). Among the benchmark methods, the listeners preferred the SBIT-KF [8] over other SEAs, apart from the ideal KF.

V. CONCLUSION

In this paper, an IT-KF with reduced-biased Kalman gain has been proposed for single channel speech

enhancement in NNCs. We have introduced an improved noise variance estimation method. The initial LPC parameters are computed from a *pre-smoothed* speech. These initial estimated parameters are used at first iteration of IT-KF to process the noisy speech and they are re-estimated at second iteration from the processed speech. It is shown that the re-estimated parameters offset the bias in Kalman gain and enables the IT-KF to minimize the noise effect, giving better enhanced speech. Experimental results reveal that the proposed method outperforms other benchmark SEAs in NNCs for a wide range of SNRs. An opportunity for further research lies in dynamically offsetting the bias of the Kalman gain under NNCs.

REFERENCES

[1] S. Ball, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113-120, April 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109-1121, Dec. 1984.

[3] K. K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 12, April 1987, pp. 177-180.

[4] N. Upadhyay and A. Karmakar, "Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study," *Procedia Computer Science*, vol. 54, pp. 574-584, Aug. 2015.

[5] X. Li, L. Girin, S. Gannot, and R. Horaud, "Non-stationary noise power spectral density estimation based on regional statistics," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, pp. 181-185.

[6] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Transactions on Signal Processing*, vol. 39, no. 8, pp. 1732-1742, Aug. 1991.

[7] S. So and K. K. Paliwal, "Suppressing the influence of additive noise on the Kalman gain for low residual noise speech enhancement," *Speech Communication*, vol. 53, no. 3, pp. 355-378, March 2011.

[8] S. K. Roy, W. P. Zhu, and B. Champagne, "Single channel speech enhancement using subband iterative Kalman filter," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2016, pp. 762-765.

[9] S. K. Roy and K. K. Paliwal, "A non-iterative Kalman filter for single channel speech enhancement in non-stationary noise condition," in *Proc. 12th International Conference on Signal Processing and Communication Systems (ICSPCS)*, Cairns, Australia, 2018.

[10] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, Hoboken: John Wiley & Sons, Ltd., 2001, pp. 227-262.

[11] Y. Ma and A. Nishihara, "Efficient voice activity detection algorithm using long-term spectral flatness measure," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, pp. 87:1-87:18, July 2013.

[12] C. Shahnaz, W. P. Zhu and M. O. Ahmad, "A multifeature voiced/unvoiced decision algorithm for noisy speech," in *Proc. IEEE International Symposium on Circuits and Systems*, Island of Kos, 2006.

[13] T. O'Haver, *A Pragmatic Introduction to Signal Processing: With Applications in Scientific Measurement*, Create Space Independent Publishing Platform, 2017.

[14] P. C. Loizou, "Speech enhancement: Theory and practice," *Signal Processing and Communications*, 2007.

[15] B. Schwerin and K. K. Paliwal, "An improved speech transmission index for intelligibility prediction," *Speech Communication*, vol. 65, pp. 9-19, Dec. 2014.

Sujan Kumar Roy was born in Kurigram, Bangladesh, in 1983. He received the B.Sc. and M.Sc. degrees in Computer Science and Engineering from the University of Rajshahi, Bangladesh, in 2008 and 2010, respectively. He also received a Master of Applied Science (M.A.Sc) degree in Electrical and Computer Engineering from Concordia University, Canada in May 2016. He is currently a Ph.D candidate in the School of Engineering at Griffith University, Brisbane, Australia. His research interests include speech enhancement.

Kuldip K. Paliwal was born in Aligarh, India, in 1952. He received the B.S. degree from Agra University, India in 1969, the M.S. degree from Aligarh Muslim University, India in 1971, and the Ph.D degree from Bombay University, India in 1978. He has worked at Tata Institute of Fundamental Research, Bombay, India Norwegian Institute of Technology, Trondheim, Norway, University of Keele, U.K., AT&T Bell Laboratories, Murray Hill, New Jersey, U.S.A., AT&T Shannon Laboratories, Florham Park, New Jersey, U.S.A., and Advanced Telecommunication Research Laboratories, Kyoto, Japan. Since July 1993, he has been a professor at Griffith University, Brisbane, Australia, in the School of Microelectronic Engineering. His current research interests include speech recognition, speech coding, speaker recognition, speech enhancement, face recognition, image coding, pattern recognition and artificial neural networks. He has published more than 300 papers in these research areas. He is a Fellow of Acoustical Society of India. He has served the IEEE Signal Processing Society's Neural Networks Technical Committee as a founding member from 1991 to 1995 and the Speech ©2016 Int. J. Sig. Process. Syst. 268 Processing Technical Committee from 1999 to 2003. He was an Associate Editor of the IEEE Transactions on Speech and Audio Processing during the periods 1994-1997 and 2003-2004. He is in the Editorial Board of the IEEE Signal Processing Magazine. He also served as an Associate Editor of the IEEE Signal Processing Letters from 1997 to 2000. He was the General Co-Chair of the Tenth IEEE Workshop on Neural Networks for Signal Processing (NNSP2000). He has co-edited two books: "Speech Coding and Synthesis" (published by Elsevier), and "Speech and Speaker Recognition: Advanced Topics" (published by Kluwer). He has received IEEE Signal Processing Society's best (senior) paper award in 1995 for his paper on LPC quantization. He served as the Editor-in-Chief of the Speech Communication journal (published by Elsevier) during 2005-2011.