Sensitivity Metric-Based Tuning of the Augmented Kalman Filter for Speech Enhancement

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

***Abstract*—The state-of-the-art robustness metric-based tuning of the augmented Kalman filter (AKF) gives an under-estimated Kalman gain, resulting *distortion* in the enhanced speech during colored noise suppression. This paper introduces a sensitivity metric-based tuning of the AKF for enhancing speech corrupted with different noises. Specifically, we observe that the sensitivity metric-based tuning of the AKF overcomes the under-estimation issues of Kalman gain in the existing method. It is shown that the reduced-biased Kalman gain enables the AKF to restrict the *residual* noise passed to the enhanced speech. It also minimizes the *distortion* in the enhanced speech. Objective and subjective testing on NOIZEUS corpus reveal that the enhanced speech produced by the proposed method exhibits higher quality as well as intelligibility than the benchmark methods in colored and non-stationary noise conditions for a wide range of SNR levels.**

***Index Terms*—Speech enhancement, Augmented Kalman filter, robustness metric, sensitivity metric, LPC.**

I. Introduction

The speech enhancement algorithm (SEA) gives an estimate of clean speech from the noisy signal. It can be used as a front end tool for many speech processing systems, such as voice communication systems, hearing-aid devices, speech recognition. Many SEAs, namely spectral subtraction (SS) [1], [2], MMSE [3], [4], Wiener Filter (WF) [5], [6], Kalman filter (KF) [7] have been introduced over the decades. However, it is still a challenging task to develop an efficient SEA for real-world noise conditions.

The SS based SEA heavily depends on the accuracy of noise power spectral density (PSD) estimates [8]. If the noise PSD gets under/over-estimated, the enhanced speech suffers from *musical* noise and *distortion* [9, Chapter 5]. The efficiency of the MMSE and WF based SEA depends on the accuracy of the *a priori* SNR estimates. In [3], a decision-directed (DD) approach was proposed to compute the *a priori* SNR in practice. Since the DD approach uses the speech and noise power spectrum estimated from the previous noisy speech frame, the computed *a priori* SNR for current frame becomes inappropriate. The biased estimate of the *a priori* SNR in the MMSE based SEA typically introduce *musical* noise and spectral *distortion* in the enhanced speech [9].

In KF based SEA [7], Paliwal and Basu computed the LPC parameters from the clean speech signal. It is capable to enhance the stationary noise corrupted speech only. Gibson *et al.* introduced an augmented KF (AKF) to enhance the colored noise corrupted speech [10]. Typically, the LPC estimates for the current noisy speech frame are computed from the filtered signal of the previous iteration by AKF. Although the enhanced speech (after 2-3 iterations) shows SNR improvement, however, suffering from significant *distortion* as well as *musical* noise. In [11], Roy *et al.* proposed a sub-band iterative KF based SEA. Since it processes the high-frequency sub-bands (SBs) among the 16 decomposed SBs, some noise components may still remain in the low-frequency SBs.

In [12], So *et al.* showed that the poor LPC estimates introduced bias in Kalman gain, resulting significant *residual* noise in the enhanced speech. To mitigate this impact, a robustness metric was used to *offset* the bias in Kalman gain. So *et al.* [13] further showed that the robustness metric gives under-estimated Kalman gain in speech regions, resulting distorted speech. To address this, a sensitivity tuning of Kalman gain has been proposed. Both of the SEAs [12], [13] were limited to operate in stationary noise conditions. For suppressing the colored noises, George *et al.* introduced a robustness metric-based tuning of the AKF [14]. As in [12], the use of robustness metric still introduce *distortion* in the enhanced speech.

The efficiency of the SEAs reported in literature becomes degraded in real-world noise conditions. Since the AKF uses the dynamic model of the additive noise, it can suppress the real-world noises more accurately than the standard KF [7]. In this paper, we further studied and find that the sensitivity metric is able to *offset* the bias in Kalman gain effectively than that of the robustness metric [14]. It is shown that the reduced-biased Kalman gain enables the AKF to minimize the *residual* noise as well as *distortion* in the enhanced speech. The efficiency of the proposed method is evaluated against the benchmark methods in terms of objective and subjective testing on NOIZEUS corpus for a wide range of SNR levels.

II. AKF for colored noise suppression

Assuming the colored noise $v(n)$ to be additive with speech $s(n)$ and uncorrelated each other, at sample n, the noisy speech $y(n)$ is given by:

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

The $s(n)$ and $v(n)$ of eq. (1) can be modeled with p^{th} and

q^{th} order linear predictors as [15]:

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

$$v(n) = -\sum_{j=1}^{q} b_j v(n-j) + u(n) \tag{3}$$

where $\{a_i; i = 1, 2, \ldots, p\}$ and $\{b_j; j = 1, 2, \ldots, q\}$ are the LPCs, $w(n)$ and $u(n)$ are assumed to be white noise with zero mean and variance σ_w^2 and σ_u^2, respectively.

Eqs. (1)-(3) can be used to form the following augmented state-space model (ASSM) of AKF as [14]:

$$\boldsymbol{x}(n) = \boldsymbol{\Phi}\boldsymbol{x}(n-1) + \boldsymbol{d}\boldsymbol{z}(n) \tag{4}$$

$$y(n) = \boldsymbol{c}^T\boldsymbol{x}(n) \tag{5}$$

In the above ASSM,

1) $\boldsymbol{x}(n) = [s(n) \ \ldots s(n-p+1)\, v(n) \ \ldots \ v(n-q+1)]^T$ is a $(p+q) \times 1$ state-vector,
2) $\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi}_s & 0 \\ 0 & \boldsymbol{\Phi}_v \end{bmatrix}$ is a $(p+q) \times (p+q)$ state-transition matrix with:

$$\boldsymbol{\Phi}_s = \begin{bmatrix} -a_1 & -a_2 & \ldots & a_{p-1} & a_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix},$$

$$\boldsymbol{\Phi}_v = \begin{bmatrix} -b_1 & -b_2 & \ldots & b_{q-1} & b_q \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix},$$

3) $\boldsymbol{d} = \begin{bmatrix} \boldsymbol{d}_s & 0 \\ 0 & \boldsymbol{d}_v \end{bmatrix}$, where $\boldsymbol{d}_s = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}^\top$, $\boldsymbol{d}_v = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}^\top$,
4) $\boldsymbol{z}(n) = \begin{bmatrix} w(n) \\ u(n) \end{bmatrix}$,
5) $\boldsymbol{c}^T = \begin{bmatrix} \boldsymbol{c}_s^T & \boldsymbol{c}_v^T \end{bmatrix}$, where $\boldsymbol{c}_s = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}^T$ and $\boldsymbol{c}_v = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}^T$ are $p \times 1$ and $q \times 1$ vectors,
6) $y(n)$ is the noisy measurement at sample n.

Firstly, $y(n)$ is windowed into non-overlapped and short (e.g., 20 ms) frames. For a particular frame, the AKF computes an unbiased and linear MMSE estimate, $\hat{\boldsymbol{x}}(n|n)$ at sample n, given $y(n)$ by using the following recursive equations [14]:

$$\hat{\boldsymbol{x}}(n|n-1) = \boldsymbol{\Phi}\hat{\boldsymbol{x}}(n-1|n-1) \tag{6}$$

$$\boldsymbol{\Psi}(n|n-1) = \boldsymbol{\Phi}\boldsymbol{\Psi}(n-1|n-1)\boldsymbol{\Phi}^T + \boldsymbol{d}\boldsymbol{Q}\boldsymbol{d}^T \tag{7}$$

$$\boldsymbol{K}(n) = \boldsymbol{\Psi}(n|n-1)\boldsymbol{c}(\boldsymbol{c}^T\boldsymbol{\Psi}(n|n-1)\boldsymbol{c})^{-1} \tag{8}$$

$$\hat{\boldsymbol{x}}(n|n) = \hat{\boldsymbol{x}}(n|n-1) + \boldsymbol{K}(n)[y(n) - \boldsymbol{c}^T\hat{\boldsymbol{x}}(n|n-1)] \tag{9}$$

$$\boldsymbol{\Psi}(n|n) = [\boldsymbol{I} - \boldsymbol{K}(n)\boldsymbol{c}^T]\boldsymbol{\Psi}(n|n-1) \tag{10}$$

where $\boldsymbol{Q} = \begin{bmatrix} \sigma_w^2 & 0 \\ 0 & \sigma_u^2 \end{bmatrix}$ is the process noise covariance.

For a noisy speech frame, the error covariances $\boldsymbol{\Psi}(n|n-1)$ and $\boldsymbol{\Psi}(n|n)$ corresponding to $\hat{\boldsymbol{x}}(n|n-1)$ and $\hat{\boldsymbol{x}}(n|n)$, and the Kalman gain $\boldsymbol{K}(n)$ are continually updated on a samplewise basis, while ($\{a_i\}$, σ_w^2) and ($\{b_k\}$, σ_u^2) remain constant. At sample n, $\boldsymbol{g}^\top\hat{\boldsymbol{x}}(n|n)$ gives the estimated speech, $\hat{s}(n|n)$, where $\boldsymbol{g} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \end{bmatrix}^\top$ is a $(p+q) \times 1$ column vector. As in [14], $\hat{s}(n|n)$ is given by:

$$\hat{s}(n|n) = [1 - K_0(n)]\hat{s}(n|n-1) + K_0(n)[y(n) - \hat{v}(n|n-1)], \tag{11}$$

where $K_0(n)$ is the 1^{st} component of $\boldsymbol{K}(n)$, given by [14]:

$$K_0(n) = \frac{\alpha^2(n) + \sigma_w^2}{\alpha^2(n) + \sigma_w^2 + \beta^2(n) + \sigma_u^2} \tag{12}$$

where $\alpha^2(n)$ and $\beta^2(n)$ are the transmission of *a posteriori* error variances (of the speech and measurement noise samples) by the augmented dynamic model from the previous time sample, $n-1$ [14]. In eq. (12), the $[\alpha^2(n) + \sigma_w^2]$ and $[\beta^2(n) + \sigma_u^2]$ are termed as the total *a priori* prediction error of speech and noise dynamic model [14].

A. Problem Statement

According to the eq. (11), the enhanced speech at sample n, i.e., $\hat{s}(n|n)$ is given by a sum of predicted speech, $\hat{s}(n|n-1)$ and innovation, $[y(n) - \hat{v}(n|n-1)]$ weighted by the $K_0(n)$. Therefore, it is evident to say that the temporal trajectory of the scalar $K_0(n)$ is a useful indicator of the AKF performance in speech enhancement context. To interpret this, we conduct an experiment with utterance sp05 ("*Wipe the grease off his dirty face*") of NOIZEUS corpus [9, Chapter 12] corrupted with 5 dB *factory2* noise, where Fig. 1 (a)-(b) shows the clean speech and noisy speech, respectively.

In ideal case, ($\{a_i\}$, σ_w^2) and ($\{b_j\}$, σ_u^2) are computed from the clean speech and noise signal, respectively. Therefore, during silent activity of observed noisy speech, as there is no speech, it is expected that $a_i = 0$ for $i = 1, 2, \ldots, p$ leading to $[\alpha^2(n) + \sigma_w^2] = 0$ (e.g., 0-0.15 s or 1.8-2.19 s of Fig. 2 (a)), which turns $K_0(n) = 0$ (according to eq. (12)). While during speech activity of noisy speech, it is observed that $[\alpha^2(n) + \sigma_w^2] >> [\beta^2(n) + \sigma_u^2]$ (e.g., 0.16-0.33 s or 0.9-1.06 s of Fig. 2 (a)). According to eq. (12), this condition causes $K_0(n)$ approaching 1 (Fig. 1 (e)). Thus, we can say that the AKF operates in robust-mode for speech regions and very sensitive in speech pauses of the noisy speech.

In noisy conditions, the silent frames are completely filled with noise. Therefore, the computed ($\{b_j\}$, σ_u^2) and ($\{a_i\}$, σ_w^2) during speech pauses of noisy speech gives $[\alpha^2(n) + \sigma_w^2] \approx [\beta^2(n) + \sigma_u^2]$ (e.g., 0-0.15 s or 1.8-2.19 s of Fig. 2 (b)). According to eq. (12), this condition introduces 0.5 bias in $K_0(n)$ (Fig. 1 (c)). With 0.5 biased $K_0(n)$, eq. (11) implies that the 50% of innovation, i.e., $[y(n) - \hat{v}(n|n-1)]$ leaking into the enhanced speech $\hat{s}(n|n)$ as shown in Fig. 1 (d). The poor estimates of ($\{a_i\}$, σ_w^2) may also result biased $K_0(n)$ in speech regions and passes the *residual* noise to $\hat{s}(n|n)$ accordingly. Therefore, in practice, the biased $K_0(n)$ forces AKF to operate in semi-robust mode [14]. In this circumstance, there should

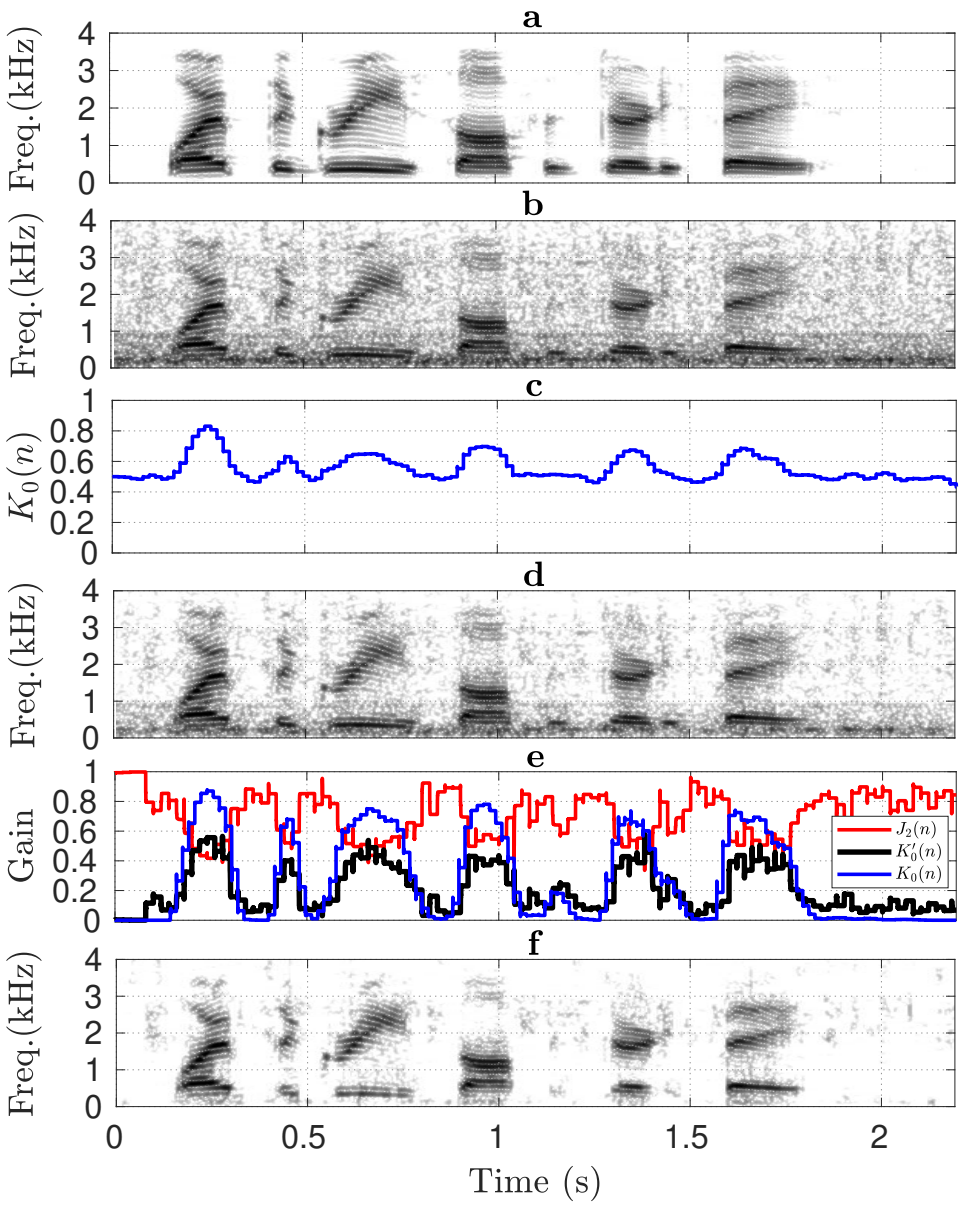

Fig. 1. (a) Clean speech (utterance sp05), (b) noisy speech (corrupt sp05 with 5 dB *factory2* noise), (c) $J_2(n)$, adjusted $K_0'(n)$, and ideal $K_0(n)$, and (d) enhanced speech produced by existing AKF based SEA [14].

have a performance metric or index that could quantify the level of robustness and sensitivity of AKF. Specifically, the robustness and sensitivity metrics, $J_2(n)$ and $J_1(n)$ quantify the level of robustness and sensitivity of AKF, defined as [14]:

$$J_2(n) = \frac{\sigma_w^2}{\alpha^2(n) + \sigma_w^2} \tag{13}$$

$$J_1(n) = \frac{\beta^2(n) + \sigma_u^2}{\alpha^2(n) + \sigma_w^2 + \beta^2(n) + \sigma_u^2} \tag{14}$$

George *et al.* used the robustness metric, $J_2(n)$ to *offset* the bias in $K_0(n)$ under colored noise conditions as [14]:

$$K_0'(n) = K_0(n)[1 - J_2(n)] \tag{15}$$

Eq. (15) implies that $J_2(n)$ needs to hover 1 to make the tuning of $K_0(n)$ effective, which is quite difficult due to the poor estimates of $(\{a_i\}, \sigma_w^2)$. To reduce the colored noise effect, George *et al.* employed a whitening filter $H_w(z)$ to the noisy speech frame prior to compute $(\{a_i\}, \sigma_w^2)$:

$$H_w(z) = 1 + \sum_{j=1}^{q} b_j z^{-j} \tag{16}$$

where $\{b_j; j = 1, 2, \ldots, q\}$ are the whitening filter coefficients as estimated from the initial speech pauses.

It can be seen from Fig. 1 (e) that the $J_2(n)$ hovers 1 in silent regions, resulting $K_0'(n) \approx 0$. While it significantly reduces the values of $K_0'(n)$ in speech regions as compared to

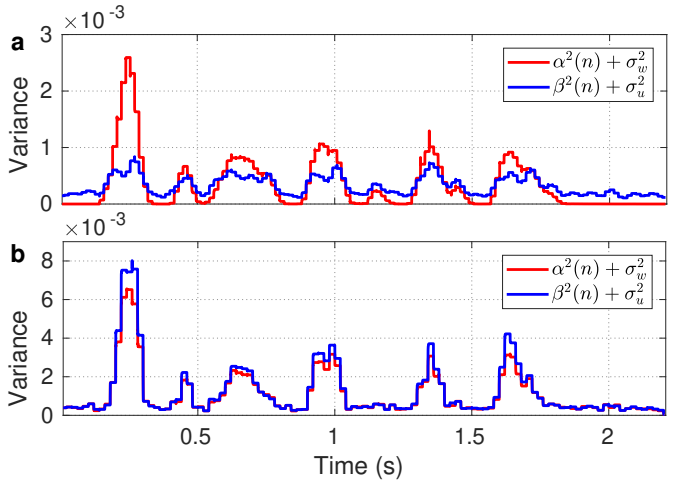

Fig. 2. Comparing the total *a priori* prediction error of speech and noise dynamic model, $[\alpha^2(n)+\sigma_w^2]$ and $[\beta^2(n)+\sigma_u^2]$ for: (a) ideal and (b) noisy cases, where the same experimental setup of Fig. 1 was used.

the ideal $K_0(n)$ (Fig. 1 (e)). As a result, the $K_0'(n)$ may over-suppress the speech components, resulting *distortion* in the enhanced speech (Fig. 1 (f)). In addition, the tuning of $K_0(n)$ using $J_2(n)$ can be affected in noise conditions having time varying amplitudes (e.g., *babble*), and the level of *distortion* as well as *residual* noise of enhanced speech varies accordingly.

III. Proposed Speech Enhancement System

Fig. 3 shows the block diagram of the proposed SEA. Firstly, a 32 ms rectangular window with 50% overlap was considered for converting $y(n)$ (eq. (1)) into frames $y(n,k)$, i.e., $y(n,k)=s(n,k)+v(n,k)$, where $k\epsilon\{0,1,2,\ldots,N-1\}$ is the frame index with N being the total number of frames in an utterance, and M is the total number of samples within each frame, i.e., $n\epsilon\{0,1,2,\ldots,M-1\}$.

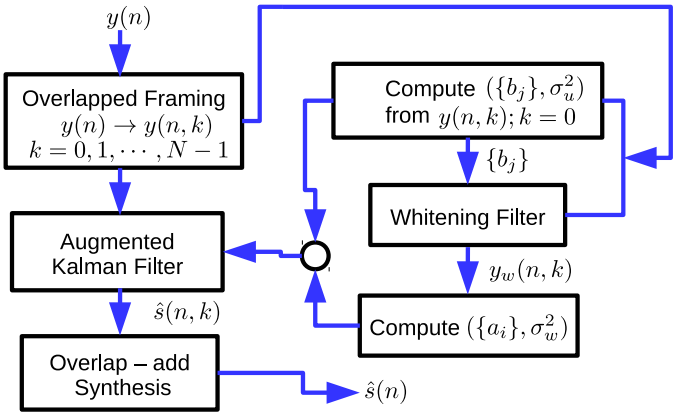

Fig. 3. Block diagram of the proposed AKF based SEA.

A. Proposed $K_0(n)$ Tuning Method

In the proposed SEA, we incorporate the sensitivity metric, $J_1(n)$ to dynamically *offset* the bias in $K_0(n)$ as:

$$K_0'(n)=K_0(n)-J_1(n) \tag{17}$$

By substituting eq. (12) and (14) into eq. (17) and re-arranging yields:

$$K_0'(n)=\frac{[\alpha^2(n)+\sigma_w^2]-[\beta^2(n)+\sigma_u^2]}{\alpha^2(n)+\sigma_w^2+\beta^2(n)+\sigma_u^2} \tag{18}$$

To utilize the sensitivity metric $J_1(n)$ for tuning of biased $K_0(n)$ in eq. (18), the $(\{b_j\},\sigma_u^2)$ $(q=40)$ are computed from the first noisy speech frame, $y(n,0)$ being considered as silent. The whitening filter, $H_w(z)$ (eq. (16)) is then implemented with the estimated $\{b_j\}$. After that the $H_w(z)$ is employed to the noisy speech frame, $y(n,k)$, yielding the pre-whitened speech, $y_w(n,k)$. The $(\{a_i\},\sigma_w^2)$ $(p=10)$ are computed from $y_w(n,k)$ using the autocorrelation method [15].

To justify the validity of adjusted $K_0'(n)$, we analyze the characteristics of estimated total *a priori* prediction error of speech and noise dynamic model, i.e., $[\alpha^2(n)+\sigma_w^2]$ and $[\beta^2(n)+\sigma_u^2]$ as shown in Fig. 4 (a). It can be seen that the $[\alpha^2(n)+\sigma_w^2]\approx[\beta^2(n)+\sigma_u^2]$ is found in silent regions (e.g., 0-0.15 s or 1.8-2.19 s of Fig. 4 (a)). According to eq. (14), this condition results $J_1(n)\approx 0.5$ as shown in Fig. 4 (b) for these typical silent regions. According to eq. (17), the condition $[\alpha^2(n)+\sigma_w^2]\approx[\beta^2(n)+\sigma_u^2]$ minimizes the 0.5 bias in $K_0(n)$ (e.g., 0-0.15 s or 1.8-2.19 s of Fig. 1 (c)), i.e., it causes $K_0'(n)\approx 0$ as desired. While the $[\alpha^2(n)+\sigma_w^2]>>[\beta^2(n)+\sigma_u^2]$ is found in speech regions (e.g., 0.16-0.33 s or 0.9-1.06 s of Fig. 4 (a)), resulting $J_1(n)$ approaching 0. In this circumstance, eq. (17) implies that the $K_0'(n)$ approaching 1.

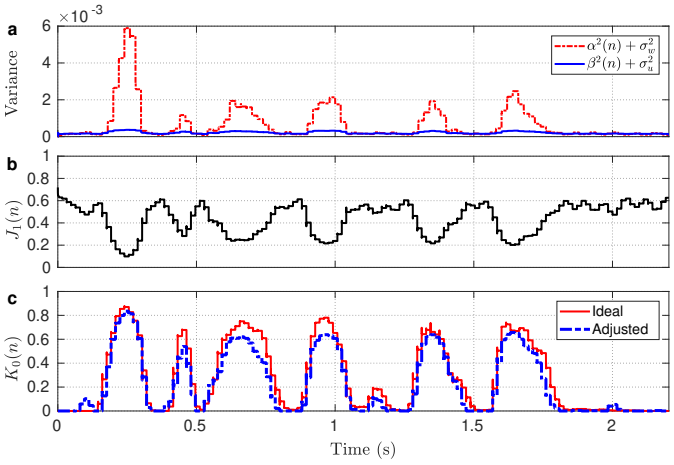

Fig. 4. Estimated: (a) $[\alpha^2(n)+\sigma_w^2]$ and $[\beta^2(n)+\sigma_u^2]$, (b) $J_1(n)$, (c) comparing the ideal and adjusted $K_0(n)$ with the same setup of Fig. 1.

Fig. 4 (c) compares the adjusted $K_0'(n)$ to that of the ideal $K_0(n)$, where the same experimental setup of Fig. 1 is used. It can be seen that the $K_0'(n)$ shows significantly less bias and closely similar to that of the ideal $K_0(n)$. Specifically, it maintains a smooth transition at the edges and the temporal changes in speech regions are closely matched to that of the ideal $K_0(n)$. Since the $K_0'(n)$ does not exceed the ideal $K_0(n)$, it enables the AKF to minimize the *distortion* and *residual* noise in the enhanced speech.

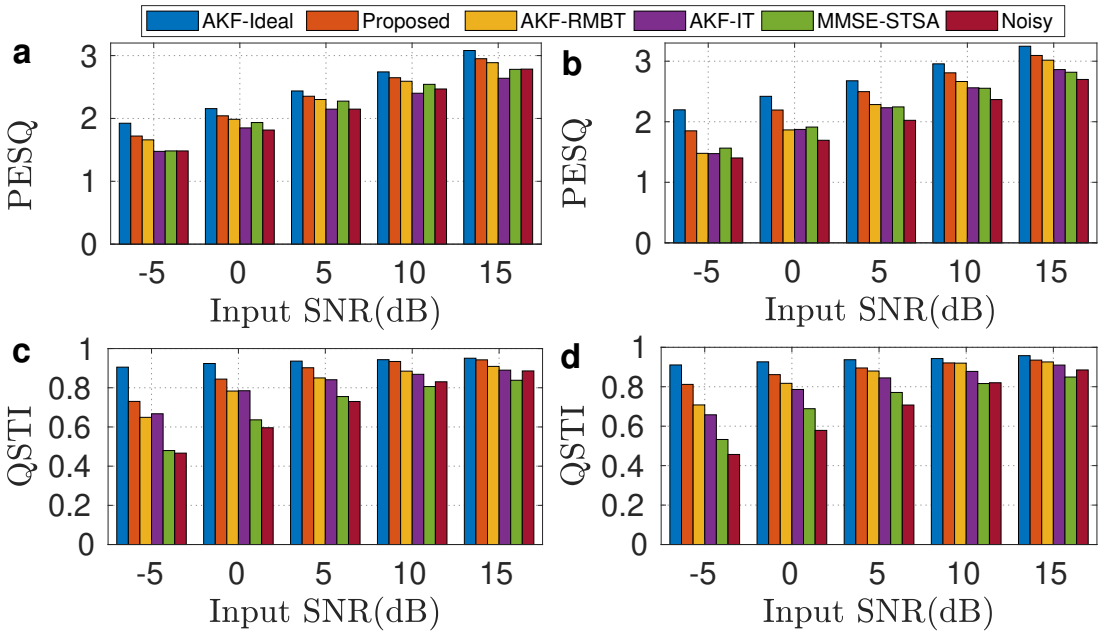

Fig. 5. Comparing the efficiency of proposed SEA with the benchmark SEAs on NOIZEUS corpus in terms of average PESQ: (a) *babble*, (b) *factory2* and average QSTI : (c) *babble*, (d) *factory2* noises.

IV. Speech Enhancement Experiment

A. Simulation Setup

For objective experiments, 30 utterances belonging to six speakers are taken from NOIZEUS corpus sampled at 8 kHz [9, Chapter 12]. We generate a noisy data set that has been corrupted by colored (*factory2*) and non-stationary (*babble*) noises [16] at multiple SNR levels (from -5dB to 15dB).

The objective quality and intelligibility evaluation were carried out through perceptual evaluation of speech quality (PESQ) [17] and quasi-stationary speech transmission index (QSTI) [18] measures. We also analyze the spectrograms of the enhanced speech produced by the proposed and benchmark SEAs in terms of the level of *residual* noise and *distortion*.

The subjective evaluation was carried out through blind AB listening test [19, Section 3.3.4]. It is conducted on sp05 corrupted with 5 dB *factory2* noise. The enhanced speech produced by five SEAs as well as the corresponding clean and noisy speech signals, a total of 42 stimuli pairs played in a random order to each listener, excluding the comparisons between the same method. For each pair, the listener prefers the first or second stimuli which is perceptually better, or a third response indicating no difference is found between them. Where 100% award is given to the preferred method, 0% to the other, and 50% to each method for the similar preference response. Participants could re-listen to stimuli if required. Five English speaking listeners participate in the AB listening tests. The average of the preference scores given by the listeners, termed as the mean preference score (%).

The proposed method is compared with benchmark methods, such as MMSE-STSA [3], AKF-IT [10], robustness-metrics based tuning of AKF (AKF-RMBT) [14], AKF-Ideal (where $(\{a_i\}, \sigma_w^2)$ and $(\{b_k\}, \sigma_u^2)$ are computed from the clean speech and noise signal), and Noisy (noise corrupted speech).

B. Results and Discussion

Fig. 5 (a)-(b) shows that the proposed method consistently shows improved PESQ over the benchmark methods, except AKF-Ideal. Whereas the AKF-RMBT [14] exhibits competitive PESQ with proposed method for *factory2* noise among the benchmark methods (Fig. 5 (a)). For *babble* noise experiment, the efficiency of AKF-RMBT method [14] gets reduced and competitive with other benchmark methods (Fig. 5 (b)).

Fig. 5 (c)-(d) implies that the proposed method also shows consistence QSTI improvement across the noise experiments, apart from AKF-Ideal. The existing AKF-RMBT [14] method is also competitive with the proposed method. Whereas the QSTI of MMSE-STSA [3] and Noisy methods are significantly lower than the AKF-IT [10] method at low SNR levels.

The proposed method exhibits a significant noise reduction (Fig. 6 (f)) than the benchmark methods (Fig. 6 (c)-(e)) and is similar to the AKF-Ideal (Fig. 6 (g)). While the AKF-RMBT method [14] introduce a bit *distortion* and noise-floor. The AKF-IT method [10] produces most distorted speech and significant *residual* noise by MMSE-STSA method [3].

It can be seen from Fig. 7 that the enhanced speech obtained through the proposed method is widely preferred by the listeners (74%) to that of the benchmark methods, apart from the clean speech and AKF-Ideal. While the AKF-RMBT method [14] is found to be the best preferred method (around 62%) among the benchmark methods by the listeners.

V. Conclusion

This paper investigates sensitivity metric-based tuning of the AKF for enhancing speech in different noise conditions.

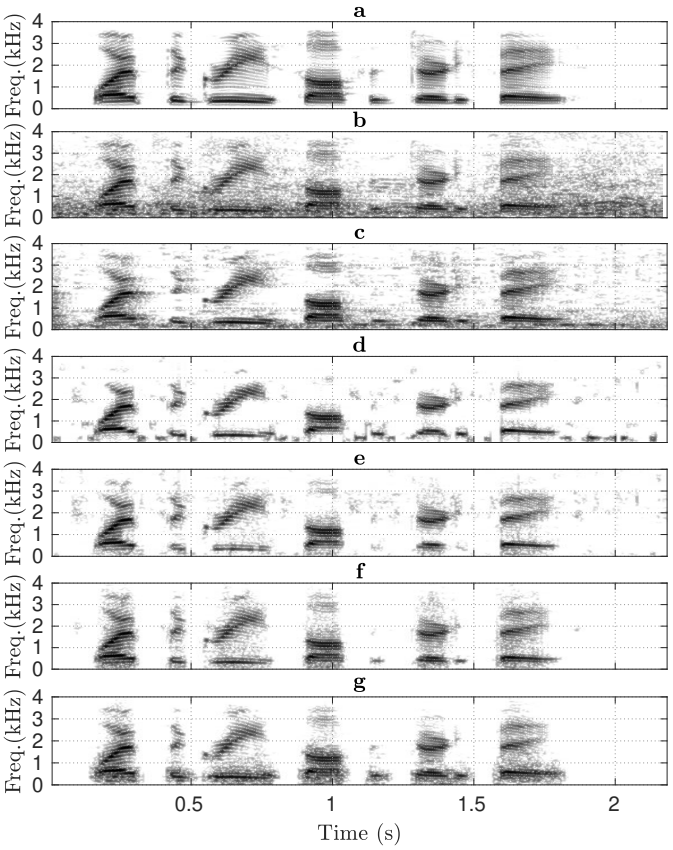

Fig. 6. Comparing the spectrograms of: (a) clean speech (utterance sp05); (b) noisy speech (corrupt sp05 with 5 dB babble noise) (PESQ=2.23); to that of the enhanced speech produced by: (c) MMSE-STSA [3] (PESQ=2.47), (d) AKF-IT [10] (PESQ=2.41), (e) AKF-RMBT [14] (PESQ=2.58), (f) Proposed (PESQ=2.73), and (g) AKF-Ideal (PESQ=2.81) methods.

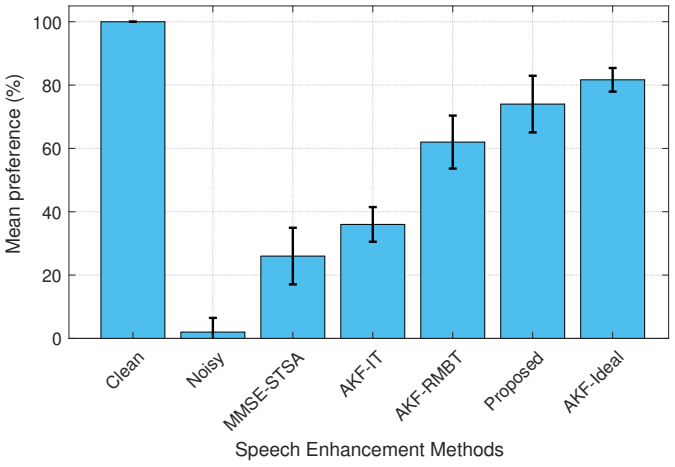

Fig. 7. The mean preference score (%) comparison among the SEAs on sp05 corrupted with 5 dB *factory2* noise.

It was shown that the under-estimated Kalman gain achieved through the existing robustness metric-based tuning of the AKF introduced *distortion* in the enhanced speech. While the sensitivity tuning of Kalman gain in the proposed SEA was found to be closely matched to that of the ideal gain. It enables the AKF to minimize *distortion* as well as *residual* noise in the enhanced speech. Objective and subjective testing reveal that the proposed method outperforms the benchmark methods in real-world colored and non-stationary noise conditions for a wide range SNR levels.

REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, April 1979.

[2] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 208–211, April 1979.

[3] 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, December 1984.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, April 1985.

[5] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 629–632, May 1996.

[6] C. Plapous, C. Marro, L. Mauuary, and P. Scalart, "A two-step noise reduction technique," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 289–292, May 2004.

[7] K. Paliwal and A. Basu, "A speech enhancement method based on kalman filtering," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 12, pp. 177–180, April 1987.

[8] 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, 2015.

[9] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, FL, USA: CRC Press, Inc., 2013.

[10] 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, August 1991.

[11] S. K. Roy, W. P. Zhu, and B. Champagne, "Single channel speech enhancement using subband iterative kalman filter," *IEEE International Symposium on Circuits and Systems*, pp. 762–765, May 2016.

[12] S. So, A. E. W. George, R. Ghosh, and K. K. Paliwal, "A non-iterative kalman filtering algorithm with dynamic gain adjustment for single-channel speech enhancement," *International Journal of Signal Processing Systems*, vol. 4, pp. 263–268, August 2016.

[13] S. So, A. E. W. George, R. Ghosh, and K. K. Paliwal, "Kalman filter with sensitivity tuning for improved noise reduction in speech," *Circuits, Systems, and Signal Processing*, vol. 36, pp. 1476–1492, April 2017.

[14] A. E. W. George, S. So, R. Ghosh, and K. K. Paliwal, "Robustness metric-based tuning of the augmented kalman filter for the enhancement of speech corrupted with coloured noise," *Speech Communication*, vol. 105, pp. 62 – 76, December 2018.

[15] S. V. Vaseghi, "Linear prediction models," in *Advanced Digital Signal Processing and Noise Reduction*. John Wiley & Sons, 2009, ch. 8, pp. 227–262.

[16] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, July 1993.

[17] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 749–752, May 2001.

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

[19] K. K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication*, vol. 52, no. 5, pp. 450–475, May 2010.