Article

Robustness and Sensitivity Tuning of the Kalman Filter for Speech Enhancement

Sujan Kumar Roy*, Kuldip K. Paliwal

Signal Processing Laboratory, Griffith University, Nathan Campus, Brisbane, QLD, 4111, Australia; k.paliwal@griffith.edu.au
* Correspondence: sujankumar.roy@griffithuni.edu.au

Abstract: The inaccurate estimates of linear prediction coefficient (LPC) and noise variance introduce bias in Kalman filter (KF) gain and degrades speech enhancement performance. The existing methods proposed a tuning of the biased Kalman gain particularly in stationary noise condition. This paper introduces a tuning of the KF gain for speech enhancement in real-life noise conditions. First, we estimate noise from each noisy speech frame using a speech presence probability (SPP) method to compute the noise variance. Then construct a whitening filter (with its coefficients computed from the estimated noise) and employed to the noisy speech, yielding a pre-whitened speech, from where the speech LPC parameters are computed. Then construct KF with the estimated parameters, where the robustness metric offsets the bias in Kalman gain during speech absence to that of the sensitivity metric during speech presence to achieve better noise reduction. Where the noise variance and the speech model parameters are adopted as a speech activity detector. The reduced-biased Kalman gain enables the KF to minimize the noise effect significantly, yielding the enhanced speech. Objective and subjective scores on NOIZEUS corpus demonstrates that the enhanced speech produced by the proposed method exhibits higher quality and intelligibility than some benchmark methods.

Keywords: Speech enhancement; Kalman filter; Kalman gain; robustness metric; sensitivity metric; LPC, whitening filter; real-life noise.)

Citation: Roy, S. K; Paliwal, K. K Robustness and Sensitivity Tuning of the Kalman Filter for Speech Enhancement. *Preprints* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

The main objective of a speech enhancement algorithm (SEA) is to improve the quality and intelligibility of the noisy speech [1]. It can be achieved by eliminating the embedded noise from a noisy speech signal without distorting the speech. Many speech processing systems, such as speech communication systems, hearing aid devices, and speech recognition systems somehow relay upon the enhancement of noisy speech. Various SEAs, namely spectral subtraction (SS) [2–5], Wiener Filter (WF) [6–8], minimum mean square error (MMSE) [9–11], Kalman filter (KF) [12], augmented KF (AKF) [13], and deep neural network (DNN) [14–16] have been introduced over the decades. This paper focuses on KF-based speech enhancement in real-life noise conditions.

Kalman filter (KF) was first used for speech enhancement by Paliwal and Basu [12]. In KF, a speech signal is represented by an auto-regressive (AR) process, whose parameters comprise the linear prediction coefficients (LPCs) and prediction error variance. The LPC parameters and noise variance are used to construct the KF recursion equations. KF gives a linear MMSE estimate of the current state of the clean speech given the observed noisy speech for each sample within a frame. Therefore, the performance of KF-based SEA largely depends on how accurately the LPC parameters and noise variance are estimated. Experiments demonstrated that the KF shows excellent performance in stationary white Gaussian noise (WGN) condition when the LPC parameters are estimated from the clean speech [12]. On the contrary, the LPC parameters and the noise variance directly computing from the noisy speech would be inaccurate and unreliable that leads to performance degradation. In [13], Gibson et al. introduced an augmented KF (AKF) to enhance colored noise corrupted speech. In this SEA, both the clean speech and noise signal are represented by two AR processes. The speech and noise LPC parameters are incorporated in an augmented matrix form to construct the recursive equations of AKF. In [13], the AKF processes the colored noise corrupted speech iteratively (usually 3-4 iterations) to eliminate the noise, yielding the enhanced speech. Specifically, the LPC parameters for the current frame are computed from the corresponding filtered speech frame of the previous iteration by AKF. Although the enhanced speech of the AKF demonstrates an improvement in signal-to-noise ratio (SNR), it suffers from *musical* noise and speech *distortion*. Therefore, this method [13] does not adequately address the inaccurate LPC parameter estimation issue in practice. Roy et al. introduced a sub-band (SB) iterative KF (SBIT-KF)-based SEA [17]. This method enhances only the high-frequency sub-bands (SBs) using iterative KF among the 16 decomposed SBs of noisy speech for a given utterance, with the assumption that the impact of noise in low-frequency SBs are negligible. However, the low-frequency SBs can also be affected by noise typically when operating in real-life noise conditions. As demonstrated in [13], the SBIT-KF [17] also suffers from speech *distortion* due to the iterative processing of noisy speech by KF.

In [18], Saha et al. proposed a robustness metric and a sensitivity metric for tuning of the biased KF gain for instrument engineering applications. Later on, So et all. applied the tuning of KF gain for speech enhancement in WGN condition [19,20]. Specifically, the enhanced speech (for each sample within a noisy speech frame) is given by recursively averaging the observed noisy speech and the predicted speech weighted by a scalar KF gain [19]. However, the inaccurate estimates of LPC parameters introduce bias in the KF gain, results in leaking a significant residual noise in the enhanced speech. In [19], a robustness metric is used to *offset* the bias in KF gain for speech enhancement. However, So et al. further showed that the robustness metric strongly suppresses the KF gain in speech regions, resulting distorted speech [20]. In [20], a sensitivity metric was used to offset the bias in KF gain, which produced less distorted speech. In [21], George et al. proposed a robustness metric-based tuning of the AKF (AKF-RMBT) for enhancing colored noise corrupted speech. As in [19], the adjusted AKF gain is under-estimated in speech regions, resulting in distorted speech.

The existing KF methods [19,20] address tuning of biased Kalman gain in WGN condition with the prior assumption that the impact of WGN on LPCs is negligible. Though the AKF method [21] performs tuning of biased gain in colored noise conditions, however, it still produced distorted speech. In this paper, we address tuning of KF gain for speech enhancement in real-life noise conditions. For this purpose, we estimate noise from each noisy speech frame using a SPP-based method to compute the noise variance. To minimize bias in the LPC parameters, we compute them from a pre-whitened speech. Then KF is constructed with the estimated parameters. To achieve better noise reduction, the robustness metric is employed to offset the bias in Kalman gain when there is speech absent to that of the sensitivity metric during speech presence of the noisy speech. We also adopt the noise variance and the AR model parameters as a speech activity detector. The reduced-biased KF gain exhibits better suppression of noise in the enhanced speech. The performance of the proposed SEA is compared against some benchmark methods using objective and subjective testing.

The structure of this paper is as follows: Section 2 describes the KF for speech enhancement, including the paradigm shift of the KF recursive equations, impact of biased KF gain on KF-based speech enhancement in WGN and real-life noise conditions. In section 3, we describe the proposed SEA, which includes the proposed parameter estimation as well as proposed Kalman gain tuning algorithm. Following this, section 4 describes the experimental setup in terms of speech corpus, objective and subjective evaluation metrics, and specifications of competitive SEAs. The experimental results are then presented in section 5. Finally, section 6 gives some concluding remarks.

2. Kalman Filter for Speech Enhancement

Assuming that the noise, $v(n)$, to be additive and uncorrelated with the clean speech, $s(n)$, at sample n, the noisy speech, $y(n)$, can be represented as:

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

The clean speech, $s(n)$, can be represented by a p^{th} order autoregressive (AR) model as [22, Chapter 8]:

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

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

Equations (1)-(2) can be used to form the following state-space model (SSM) of the KF (where the **bold** variables denote vector/matrix quantities, as opposed to unbolded variables for scalar quantities):

$$\boldsymbol{x}(n) = \boldsymbol{\Phi}\boldsymbol{x}(n-1) + \boldsymbol{d}w(n), \tag{3}$$

$$y(n) = \boldsymbol{c}^\top \boldsymbol{x}(n) + v(n). \tag{4}$$

In the above SSM,

1. $\boldsymbol{x}(n)$ is a $p \times 1$ state vector at sample n, given by:

$$\boldsymbol{x}(n) = [s(n) \quad s(n-1) \quad \ldots \quad s(n-p+1)]^\top, \tag{5}$$

2. $\boldsymbol{\Phi}$ is a $p \times p$ state transition matrix, represented as:

$$\boldsymbol{\Phi} = \begin{bmatrix} -a_1 & -a_2 & \ldots & -a_{p-1} & -a_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}, \tag{6}$$

3. $\boldsymbol{d}$ and $\boldsymbol{c}$ are the $p \times 1$ measurement vectors for the excitation noise and observation, written as:

$$\boldsymbol{d} = \boldsymbol{c} = \begin{bmatrix}1 & 0 & \ldots & 0\end{bmatrix}^T,$$

4. $y(n)$ is the observed noisy speech at sample n.

During the operation of KF, the noisy speech, $y(n)$, is windowed into non-overlapped and short (e.g., 20 ms) frames. For a particular frame, the KF recursively computes an unbiased linear MMSE estimate, $\hat{\boldsymbol{x}}(n|n)$, of the state-vector, $\boldsymbol{x}(n)$, given the observed noisy speech up to sample n, i.e., $y(1), y(2), ..., y(n)$, using the following equations [12]:

$$\hat{\boldsymbol{x}}(n|n-1) = \boldsymbol{\Phi}\hat{\boldsymbol{x}}(n-1|n-1), \tag{7}$$

$$\boldsymbol{\Psi}(n|n-1) = \boldsymbol{\Phi}\boldsymbol{\Psi}(n-1|n-1)\boldsymbol{\Phi}^\top + \sigma_w^2 \boldsymbol{d}\boldsymbol{d}^\top, \tag{8}$$

$$\boldsymbol{K}(n) = \boldsymbol{\Psi}(n|n-1)\boldsymbol{c}[\boldsymbol{c}^\top \boldsymbol{\Psi}(n|n-1)\boldsymbol{c} + \sigma_v^2]^{-1}, \tag{9}$$

$$\hat{\boldsymbol{x}}(n|n) = \hat{\boldsymbol{x}}(n|n-1) + \boldsymbol{K}(n)[y(n) - \boldsymbol{c}^\top \hat{\boldsymbol{x}}(n|n-1)], \tag{10}$$

$$\boldsymbol{\Psi}(n|n) = [\boldsymbol{I} - \boldsymbol{K}(n)\boldsymbol{c}^\top]\boldsymbol{\Psi}(n|n-1). \tag{11}$$

In the above eqs. (7)-(11), $\boldsymbol{\Psi}(n|n-1)$ and $\boldsymbol{\Psi}(n|n)$ are the error covariance matrices of the *a priori* and *a posteriori* state estimates, $\hat{\boldsymbol{x}}(n|n-1)$ and $\hat{\boldsymbol{x}}(n|n)$; $\boldsymbol{K}(n)$ is the Kalman gain; σ_v^2 is the variance of the additive noise, $v(n)$; and $\boldsymbol{I}$ is the identity matrix. During processing each frame, the estimated LPC parameters, $(\{a_i\}, \sigma_w^2)$, and noise variance, σ_v^2, remain unchanged for that frame, while $\boldsymbol{K}(n)$, $\boldsymbol{\Psi}(n|n)$, and $\hat{\boldsymbol{x}}(n|n)$ are continually updated on a samplewise basis. The estimated speech at sample n is given by: $\hat{s}(n|n) = \boldsymbol{c}^\top\hat{\boldsymbol{x}}(n|n)$. Once all noisy speech frames being processed, synthesis over the enhanced frames yielding the enhanced speech, $\hat{s}(n)$.

2.1. *Paradigm shift of recursive equations*

The paradigm shift of the recursive eqs. (7)-(11) transforms them in scalar form. It exploits the understanding as well as analysis of the KF operation in speech enhancement context. The simplification starts with the output of the KF, $\hat{s}(n|n) = \boldsymbol{c}^\top\hat{\boldsymbol{x}}(n|n)$, which is re-written as:

$$\begin{aligned} \boldsymbol{c}^\top\hat{\boldsymbol{x}}(n|n) &= \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \hat{s}(n|n) \\ \hat{s}(n|n-1) \\ \vdots \\ \hat{s}(n|n-p+1) \end{bmatrix}, \\ &= \hat{s}(n|n). \end{aligned} \tag{12}$$

Multiply $\boldsymbol{c}^\top$ on both sides of eq. (10) gives:

$$\boldsymbol{c}^\top\hat{\boldsymbol{x}}(n|n) = \boldsymbol{c}^\top\hat{\boldsymbol{x}}(n|n-1) + \boldsymbol{c}^\top\boldsymbol{K}(n)[y(n) - \boldsymbol{c}^\top\hat{\boldsymbol{x}}(n|n-1)]. \tag{13}$$

According to eq. (12), $\boldsymbol{c}^\top\hat{\boldsymbol{x}}(n|n-1)$ is also given by:

$$\boldsymbol{c}^\top\hat{\boldsymbol{x}}(n|n-1) = \hat{s}(n|n-1). \tag{14}$$

In eq. (13), $\boldsymbol{c}^\top\boldsymbol{K}(n)$ represents the first component, $K_0(n)$, of the Kalman gain vector, $\boldsymbol{K}(n)$, i.e.,:

$$K_0(n) = \boldsymbol{c}^\top\boldsymbol{K}(n). \tag{15}$$

Substituting eq. (9) into eq. (15) gives:

$$K_0(n) = \frac{\boldsymbol{c}^\top\boldsymbol{\Psi}(n|n-1)\boldsymbol{c}}{\boldsymbol{c}^\top\boldsymbol{\Psi}(n|n-1)\boldsymbol{c} + \sigma_v^2}. \tag{16}$$

With eq. (8), $\boldsymbol{c}^\top\boldsymbol{\Psi}(n|n-1)\boldsymbol{c}$ of eq. (16) is simplified as:

$$\boldsymbol{c}^\top\boldsymbol{\Psi}(n|n-1)\boldsymbol{c} = \boldsymbol{c}^\top\boldsymbol{\Phi}\boldsymbol{\Psi}(n-1|n-1)\boldsymbol{\Phi}^\top\boldsymbol{c} + \boldsymbol{c}^\top\sigma_w^2\boldsymbol{d}\boldsymbol{d}^\top\boldsymbol{c}. \tag{17}$$

The linear algebra operation on $\boldsymbol{c}^\top\sigma_w^2\boldsymbol{d}\boldsymbol{d}^\top\boldsymbol{c}$, gives:

$$\boldsymbol{c}^\top\sigma_w^2\boldsymbol{d}\boldsymbol{d}^\top\boldsymbol{c} = \sigma_w^2, \tag{18}$$

and $\boldsymbol{c}^\top\boldsymbol{\Phi}\boldsymbol{\Psi}(n-1|n-1)\boldsymbol{\Phi}^\top\boldsymbol{c}$ represents the transmission of *a posteriori* error variance by the speech model from the previous time sample, $n-1$, denoted as [20]:

$$\boldsymbol{c}^\top\boldsymbol{\Phi}\boldsymbol{\Psi}(n-1|n-1)\boldsymbol{\Phi}^\top\boldsymbol{c} = \alpha^2(n). \tag{19}$$

Substituting eqs. (18)-(19) into eq. (17) gives:

$$\boldsymbol{c}^\top\boldsymbol{\Psi}(n|n-1)\boldsymbol{c} = \alpha^2(n) + \sigma_w^2. \tag{20}$$

From eq. (20) and (16), $K_0(n)$ in given by:

$$K_0(n) = \frac{\alpha^2(n) + \sigma_w^2}{\alpha^2(n) + \sigma_w^2 + \sigma_v^2}. \tag{21}$$

Substituting eqs. (12), (14), and (15) into eq. (13) gives:

$$\hat{s}(n|n) = \hat{s}(n|n-1) + K_0[y(n) - \hat{s}(n|n-1)]. \tag{22}$$

Re-arranging eq. (22) yields:

$$\hat{s}(n|n) = [1 - K_0(n)]\hat{s}(n|n-1) + K_0(n)y(n). \tag{23}$$

Eq. (23) implies that the accurate estimates of $\hat{s}(n|n)$ (output of the KF) will be achieved if $K_0(n)$ becomes unbiased. However, in practice, the inaccurate estimates of $(\{a_i\}, \sigma_w^2)$ and σ_v^2 introduce bias in $K_0(n)$, resulting degraded $\hat{s}(n|n)$. In [18], Saha et al. introduced a robustness and a sensitivity metrics to quantify the level of robustness and sensitivity of the KF, which can be used to *offset* the bias in $K_0(n)$. In speech enhancement context, these metrics can be computed by simplifying the mean squared error, $\boldsymbol{c}^\top \boldsymbol{\Psi}(n|n)\boldsymbol{c}$ of the KF output, $\hat{s}(n|n)$ as:

$$\begin{aligned} \boldsymbol{c}^\top \boldsymbol{\Psi}(n|n)\boldsymbol{c} &= \boldsymbol{c}^\top[\boldsymbol{I} - \boldsymbol{K}(n)\boldsymbol{c}^\top]\boldsymbol{\Psi}(n|n-1)\boldsymbol{c}, \quad [from\ (11)] \\ &= \boldsymbol{c}^\top \boldsymbol{\Psi}(n|n-1)\boldsymbol{c} - \boldsymbol{c}^\top \boldsymbol{K}(n)\boldsymbol{c}^\top \boldsymbol{\Psi}(n|n-1)\boldsymbol{c}. \end{aligned} \tag{24}$$

Substituting eqs. (15) and (20) into (24) gives:

$$\begin{aligned} &\Psi_{0,0}(n|n) = \alpha^2(n) + \sigma_w^2 - K_0(n)[\alpha^2(n) + \sigma_w^2], \\ &\Psi_{0,0}(n|n) - \alpha^2(n) = \sigma_w^2 - \frac{[\alpha^2(n) + \sigma_w^2]^2}{\alpha^2(n) + \sigma_w^2 + \sigma_v^2}, \\ &\frac{\Psi_{0,0}(n|n) - \alpha^2(n)}{\alpha^2(n) + \sigma_w^2} = \frac{\sigma_w^2}{\alpha^2(n) + \sigma_w^2} - \frac{\alpha^2(n) + \sigma_w^2}{\alpha^2(n) + \sigma_w^2 + \sigma_v^2}, \\ &\frac{\Psi_{0,0}(n|n) - \alpha^2(n)}{\alpha^2(n) + \sigma_w^2} = \frac{\sigma_w^2}{\alpha^2(n) + \sigma_w^2} + \frac{\sigma_v^2}{\alpha^2(n) + \sigma_w^2 + \sigma_v^2} - 1, \\ &\frac{\Psi_{0,0}(n|n) - \alpha^2(n)}{\alpha^2(n) + \sigma_w^2} + 1 = \frac{\sigma_w^2}{\alpha^2(n) + \sigma_w^2} + \frac{\sigma_v^2}{\alpha^2(n) + \sigma_w^2 + \sigma_v^2}, \\ &\Delta\Psi(n|n) + 1 = J_2(n) + J_1(n), \end{aligned} \tag{25}$$

where $J_2(n)$ and $J_1(n)$ are the robustness and sensitivity metrics of the KF, given as [19,20]:

$$J_2(n) = \frac{\sigma_w^2}{\alpha^2(n) + \sigma_w^2}, \tag{26}$$

$$J_1(n) = \frac{\sigma_v^2}{\alpha^2(n) + \sigma_w^2 + \sigma_v^2}. \tag{27}$$

The KF-based SEAs in [19,20] address tuning of $K_0(n)$ using $J_2(n)$ and $J_1(n)$ metrics for speech enhancement in WGN condition as described next.

2.2. Impact of Biased $K_0(n)$ on KF-based Speech Enhancement in WGN Condition

We analyze the shortcomings of existing KF-based SEAs [19,20] in terms of biased interpretation of $K_0(n)$. For this purpose, we conduct an experiment with the utterance sp05 (*Wipe the grease off his dirty face*) of NOIZEUS corpus [1, Chapter 12] (sampled at 8 kHz) corrupted with 5 dB WGN noise [23]. In [19,20], a 20 ms non-overlapped rectangular window was considered for converting $y(n)$ into frames as:

$$y(n,k) = s(n,k) + v(n,k), \tag{28}$$

where $k\epsilon\{0,1,2,\ldots,N-1\}$ is the frame index, N is the total number of frames in an utterance, and M is the total number of samples in each frame, i.e., $n\epsilon\{0,1,2,\ldots,M-1\}$.

In [19], So et al. first analyze $K_0(n)$ in oracle case, where $(\{a_i\},\sigma_w^2)$ $(p = 10)$ and σ_v^2 are computed from each frame of the clean speech and the noise signal, $s(n,k)$ and $v(n,k)$. It can be seen that $K_0(n)$ approaching 1 when there is speech presence of the noisy speech, which passes almost clean speech to the output (e.g., 0.16-0.33 s or 0.9-1.06 s in Figure 1 (d)-(e)). Conversely, $K_0(n)$ remains approximately 0 during speech absent of the noisy speech, which does not pass any corrupting noise (e.g., 0-0.15 s or 1.8-2.19 s in Figure 1 (d)-(e)). Thus, KF-Oracle method produced almost identical speech (Figure 1 (e)) to the clean speech (Figure 1 (a)).

In non-oracle case, $(\{a_i\},\sigma_w^2)$ are computed from noisy speech, resulting biased $(\{\tilde{a}_i\},\tilde{\sigma}_w^2)$. Then $K_0(n)$ in (21) using biased $\tilde{\sigma}_w^2$ is given by:

$$K_0(n) = \frac{\alpha^2(n) + \tilde{\sigma}_w^2}{\alpha^2(n) + \tilde{\sigma}_w^2 + \sigma_v^2}. \tag{29}$$

In [19,20], So et all assumed that the impact of WGN in $\{\tilde{a}_i\}$ is negligible. Thus, $\tilde{\sigma}_w^2$ could be approximately estimated as: $\tilde{\sigma}_w^2 \approx \sigma_w^2 + \sigma_v^2$ [19,20]. Substituting $\tilde{\sigma}_w^2 \approx \sigma_w^2 + \sigma_v^2$ in eq. (29) and re-arranging yields:

$$K_0(n) = \frac{\alpha^2(n) + \sigma_w^2 + \sigma_v^2}{\alpha^2(n) + \sigma_w^2 + 2\sigma_v^2}. \tag{30}$$

During speech pauses of $y(n,k)$, $s(n,k) = 0$ gives $\alpha^2(n) = 0$ and $\sigma_w^2 = 0$. According to eq. (30), $K_0(n)$ becomes biased around 0.5 (e.g., 0-0.15 s or 1.8-2.19 s in Figure 1 (d)). The biased $K_0(n)$ leaking a significant residual noise in the enhanced speech as shown in Figure 1 (f).

In non-oracle case, it is also observed that $J_2(n) \approx 1$ typically during speech pauses of $y(n,k)$ (e.g., 0-0.15 s or 1.8-2.19 s in Figure 1 (c)). Therefore, $J_2(n)$ metric is found to be useful in tuning biased $K_0(n)$ as [19]:

$$K_0'(n) = K_0(n)[1 - J_2(n)]. \tag{31}$$

Figure 1 (d) reveals that $K_0'(n) \approx 0$ during speech pauses. However, $K_0'(n)$ is over-suppressed during speech presence of $y(n,k)$, resulting distorted speech as shown in Figure 1 (g).

To address this, So et al. proposed a $J_1(n)$ metric-based tuning of $K_0(n)$ [20]. It can be seen from Figure 1 (c) that $J_1(n)$ lies around 0.5 during speech pauses (e.g., 0-0.15 s or 1.8-2.19 s), whereas approaching 0 at speech regions (e.g., 0.16-0.33 s or 0.9-1.06 s). Therefore, the tuning of biased $K_0(n)$ using $J_1(n)$ metric is performed as [20]:

$$K_0''(n) = K_0(n) - J_1(n). \tag{32}$$

It can be seen from Figure 1 (d) that $K_0''(n)$ is closely similar to the oracle $K_0(n)$, which minimizes distortion in the enhanced speech (Figure 1 (h)) as compared to (Figure 1 (g)).

Technically, the real-life noise (colored/non-stationary) may contain time varying amplitudes, which impact $(\{a_i\},\sigma_w^2)$ significantly as opposed to negligible impact of WGN in these parameters [19,20]. Therefore, the assumption of $\tilde{\sigma}_w^2 \neq \sigma_w^2 + \sigma_v^2$ made in [19,20] is invalid for real-life noise conditions. Moreover, the existing methods [19,20] do not analyze the impact of noise variance, σ_v^2 on $K_0(n)$. According to eq. (21), besides $\alpha^2(n)$ and σ_w^2, σ_v^2 is also an important parameter to compute $K_0(n)$ accurately. In light of

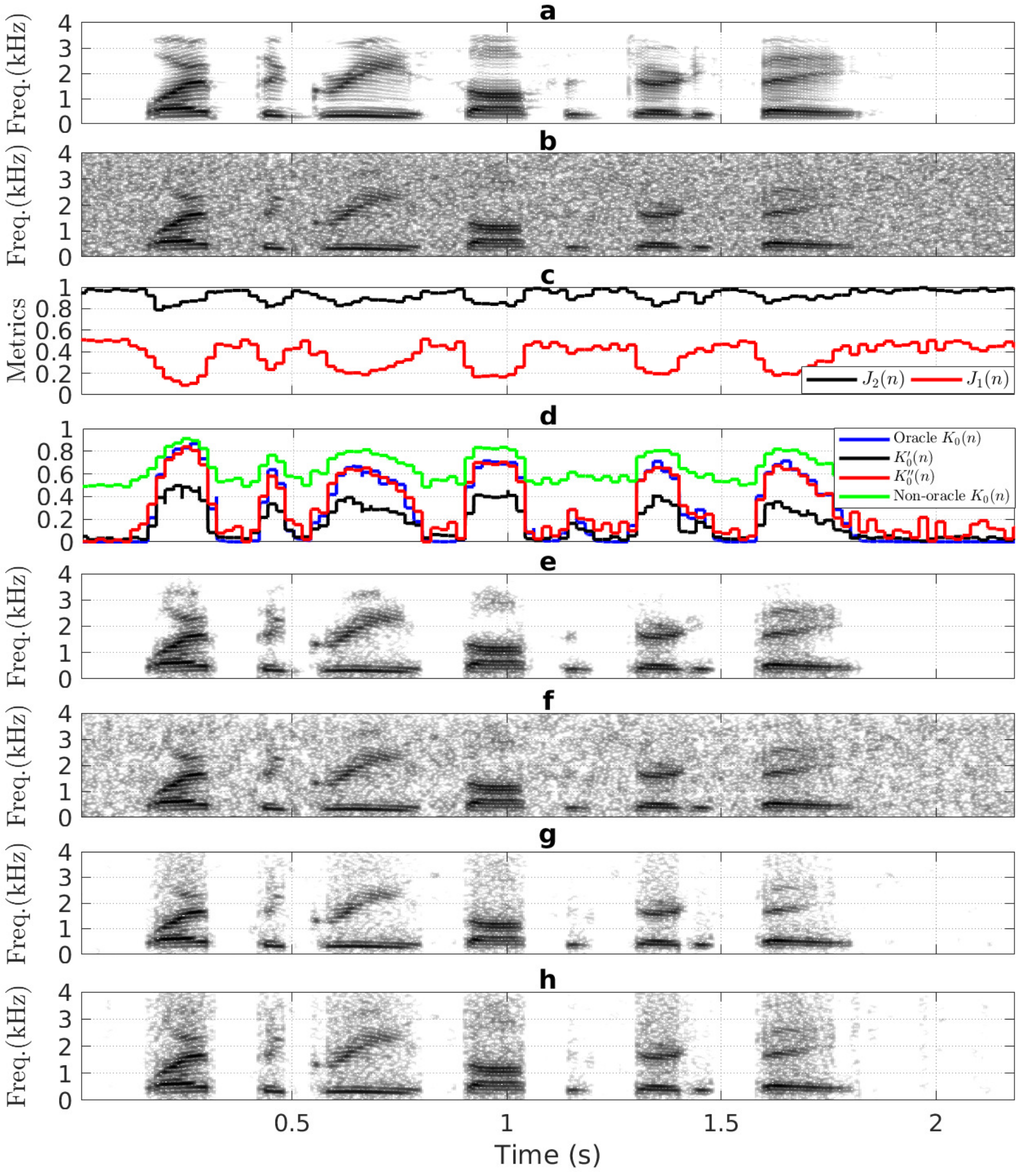

Figure 1. Review of existing KF-based SEA: (a)-(b) spectrograms of the clean speech (utterance sp05) and the noisy speech (corrupt (a) with 5 dB WGN), (c) $J_2(n)$ and $J_1(n)$ metrics, (d) oracle and non-oracle $K_0(n)$ with adjusted $K'_0(n)$ and $K''_0(n)$, spectrogram of enhanced speech produced by: (e) KF-Oracle method, (f) KF-Non-oracle method, (g)-(h) methods in [19,20].

these observations, the methods in [19,20] are not applicable for speech enhancement in real-life noise conditions. Therefore, we perform detail analysis of biasing effect of $K_0(n)$ on KF-based speech enhancement in real-life noise conditions.

2.3. Impact of Biased $K_0(n)$ on KF-based Speech Enhancement in Real-life Noise Condition

To analyze $K_0(n)$ and its impact on KF-based speech enhancement, we repeat the experiment in Figure 1 except the utterance sp05 is corrupted with a typical real-life non-stationary noise, *babble* [23] at 5 dB SNR. A 32 ms rectangular window with 50% overlap [24, Sec 7.2.1] was considered for converting $y(n)$ into frames, $y(n,k)$ (as in eq. (28)).

As shown in Section 2.2, in oracle case, $K_0(n)$ also shows a smooth transition between 0 and 1 depending on the speech absent and speech presence of noisy speech (Figure 2 (c)). Technically, during speech pauses of $y(n,k)$, the total *a priori* prediction error of the AR model, $[\alpha^2(n) + \sigma_w^2] = 0$ (e.g., 0-0.15 s or 1.8-2.19 s in Figure 2 (d)).

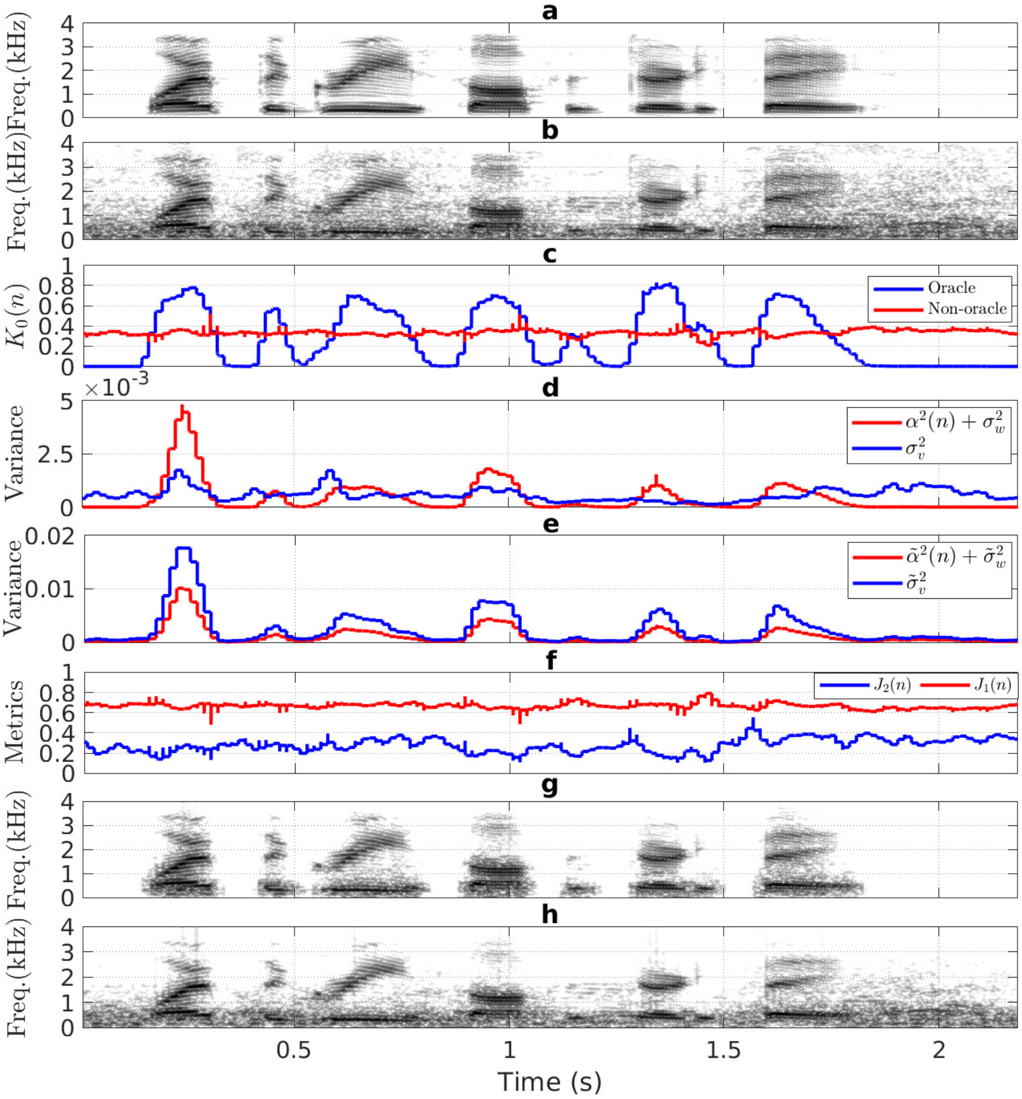

Figure 2. Biasing effect of $K_0(n)$: (a)-(b) spectrograms of the clean speech and the noisy speech (corrupt sp05 with 5 dB *babble* noise), (c) $K_0(n)$ computed in oracle and non-oracle cases, (d)-(e) $[\alpha^2(n)+\sigma_w^2]$ and σ_v^2 computed in oracle and non-oracle cases, (f) $J_2(n)$ and $J_1(n)$ computed from the noisy speech in (b), spectrogram of enhanced speech produced by: (g) KF-Oracle method, and (h) KF-Non-oracle method.

Substituting $[\alpha^2(n)+\sigma_w^2]=0$ in eq. (21)), gives $K_0(n)=0$, which in turn $\hat{s}(n|n)=0$ (eq. (23)), i.e., nothing is passed to the output (e.g., 0-0.15 s or 1.8-2.19 s of Figure 2 (c), (g)). Conversely, it is observed that $[\alpha^2(n)+\sigma_w^2] >> \sigma_v^2$ in speech regions of $y(n,k)$, for which $K_0(n)$ is approaching 1 (e.g., 0.16-0.33 s or 0.9-1.06 s in Figure 2 (c)). As discussed in section 2.2, the higher $K_0(n)$ almost passes the clean speech to the output. Therefore, the enhanced speech in oracle case (Figure 2(g)) is closely similar to the clean speech (Figure 2(a)).

In non-oracle case, the biased estimates of $(\{\tilde{a}_i\},\tilde{\sigma}_w^2)$ and $\tilde{\sigma}_v^2$, resulting $[\tilde{\alpha}^2(n)+\tilde{\sigma}_w^2]\approx\tilde{\sigma}_v^2$ (e.g., 0-0.15 s or 1.8-2.19 s in Figure 2 (e)). According to eq. (21), this condition introduces around 0.5 bias in $K_0(n)$ (e.g., 0-0.15 s or 1.8-2.19 s in Figure 2 (c)). During speech presence of $y(n,k)$, it is observed that $\tilde{\sigma}_v^2 >> [\tilde{\alpha}^2(n)+\tilde{\sigma}_w^2]$ (e.g., 0.16-0.33 s or 0.9-1.06 s of Figure 2 (e)), resulting under-estimated $K_0(n)$ as compared to oracle case (Figure 2 (c)). The 0.5 biased $K_0(n)$ leaking 50% *residual* noise to $\hat{s}(n|n)$ particularly in silent regions (Figure 2 (h)). While the under-estimated $K_0(n)$ in speech regions introduce a significant *distortion* in the enhanced (Figure 2 (h)). In addition, $J_2(n)$ and

$J_1(n)$ metrics (Figure 2 (f)) do not comply with the desired characteristics as found in WGN condition (Figure 1 (c)). Therefore, it is inappropriate to apply the metrics (Figure 2 (f)) for tuning of the biased $K_0(n)$ (Figure 2 (c)) using eqs. (31)-(32).

In AKF-RMBT method, the speech LPC parameters were computed from the pre-whitened speech to utilize $J_2(n)$ metric for the tuning of biased $K_0(n)$ in colored noise conditions [21, Figure 5 (d)]. As in [19], $J_2(n)$ metric-based tuning of $K_0(n)$ still produces distorted speech. In addition, the noise LPC parameters computed from initial speech pauses keep constant during processing all noisy speech frames for an utterance. The whitening filter is also constructed with the constant noise LPCs to pre-whiten each noisy speech frame prior to compute speech LPC parameters. As a result, the tuning of $K_0(n)$ [21] becomes irrespective in conditions having time-varying amplitudes, such as *babble* noise.

Motivated by the shortcomings of [19–21], we propose $J_2(n)$ and $J_1(n)$ metrics-based tuning of KF gain, $K_0(n)$ for speech enhancement in real-life noise conditions.

3. Proposed Speech Enhancement Algorithm

Figure 3 shows the block diagram of the proposed SEA. Firstly, $y(n)$ is converted into frames, $y(n,k)$ with the same setup as used in section 2.3. To carried out the tuning of $K_0(n)$ in real-life noise conditions, unlike biased $J_2(n)$ and $J_1(n)$ metrics (Figure 2 (f)), they should achieve similar characteristics that occur in WGN condition (Figure 1 (c)). It can be achieved through improving the estimates of $(\{\hat{a}_i\},\hat{\sigma}^2_w)$ and $\hat{\sigma}^2_v$ as described in the next section.

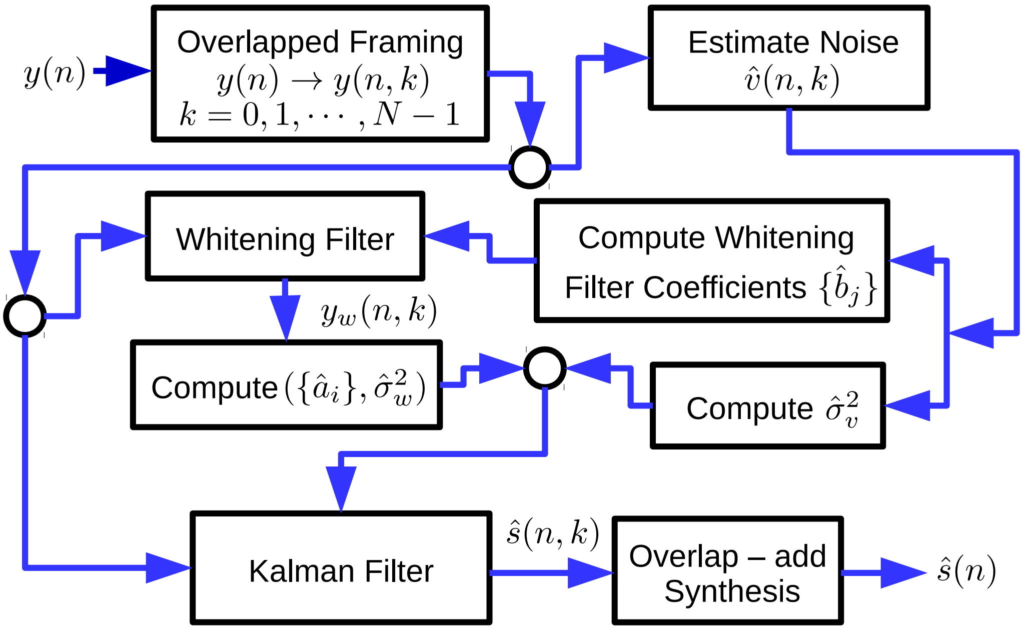

Figure 3. Block diagram of the proposed KF-based SEA.

3.1. Parameter Estimation

In is known that $(\{a_i\},\sigma^2_w)$ are very sensitive to real-life noises. Since the clean speech, $s(n,k)$ is unavailable in practice, it is difficult to accurately estimate these parameters. Therefore, we first focus on noise estimation, $\hat{v}(n,k)$ for each noisy speech frame using speech presence probability (SPP) method (described in section 3.2) [25], to compute $\hat{\sigma}^2_v$. Given $\hat{v}(n,k)$, $\hat{\sigma}^2_v$ is computed as:

$$\hat{\sigma}^2_v = \frac{1}{M}\sum_{n=0}^{M-1}\hat{v}^2(n,k). \tag{33}$$

To reduce bias in the estimated $(\{\hat{a}_i\}, \hat{\sigma}_w^2)$ for each noisy speech frame, we compute them from the corresponding pre-whitened speech, $y_w(n,k)$ using the autocorrelation method [22]. The framewise $y_w(n,k)$ is obtained by employing a whitening filter, $H_w(z)$ to $y(n,k)$. $H_w(z)$ is given by [22]:

$$H_w(z) = 1 + \sum_{j=1}^{q} \hat{b}_j z^{-j}, \tag{34}$$

where the coefficients, $\{\hat{b}_j\}$ $(q = 20)$ are computed from $\hat{v}(n,k)$ using the autocorrelation method [22].

3.2. Proposed $\hat{v}(n,k)$ Estimation Method

The proposed noise estimation is performed in acoustic-domain using the SPP method [25]. For this purpose, the noisy speech, $y(n)$ (eq. (1)) is analyzed frame-wise using the short-time Fourier transform (STFT):

$$Y_k(m) = S_k(m) + V_k(m), \tag{35}$$

where $Y_k(m)$, $S_k(m)$, and $V_k(m)$ denote the complex-valued STFT coefficients of the noisy speech, the clean speech, and the noise signal, respectively, for time-frame index k and frequency bin index $m\epsilon\{0, 1, \ldots, 255\}$.

A Hamming window with 50% overlap is used in STFT analysis [24, Sec 7.2.1]. In polar form, $Y_k(m)$, $S_k(m)$, and $V_k(m)$ can be expressed as: $Y_k(m) = R_k(m)e^{j\phi_k(m)}$, $S_k(m) = A_k(m)e^{j\varphi_k(m)}$, and $V_k(m) = D_k(m)e^{j\theta_k(m)}$, where $R_k(m)$, $A_k(m)$, and $D_k(m)$ are the magnitude spectrums of the noisy speech, the clean speech, and the noise signal, respectively, and $\phi_k(m)$, $\varphi_k(m)$, and $\theta_k(m)$ are the corresponding phase spectrums. We process each frequency bin of the single-sided noisy speech power spectrum, $R_k^2(m)$ to estimate the noise power spectrum, $\hat{D}_k^2(m)$, where $m\epsilon\{0, 1, \ldots, 128\}$ containing the DC and Nyquist frequency components. To initialize the algorithm, we consider the first frame $(k = 0)$ of $R_0^2(m)$ as silent, giving an estimate of noise power, $\hat{D}_0^2(m) = R_0^2(m)$. The noise PSD, $\hat{\lambda}_0(m)$ is also initialized as: $\hat{\lambda}_0(m) = \hat{D}_0^2(m)$. For $k \geq 1$, using the speech presence uncertainty principle [25], an MMSE estimate of $\hat{D}_k^2(m)$ at m^{th} frequency bin is given by:

$$\hat{D}_k^2(m) = P(H_0^m | R_k(m)) R_k^2(m) + P(H_1^m | R_k(m)) \hat{\lambda}_{k-1}(m), \tag{36}$$

where $P(H_0^m | R_k(m))$ and $P(H_1^m | R_k(m))$ are the conditional probability of the speech absence and the speech presence, given $R_k(m)$ at m^{th} frequency bin.

The simplified $P(H_1^m | R_k(m))$ estimate is given by[1] [25]:

$$P(H_1^m | R_k(m)) = \left[1 + (1 + \xi_{opt}) \exp\left\{ \left(-\frac{R_k^2(m)}{\hat{\lambda}_{k-1}(m)} \right) \left(\frac{\xi_{opt}}{1 + \xi_{opt}} \right) \right\} \right]^{-1}, \tag{37}$$

where ξ_{opt} is the optimal *a priori* SNR.

In [25], the optimal choice for ξ_{opt} is found to be $10 \log_{10}(\xi_{opt}) = 15$ dB, and $P(H_0^m | R_k(m))$ is given by $P(H_0^m | R_k(m)) = 1 - P(H_1^m | R_k(m))$. If $P(H_1^m | R_k(m)) = 1$ occurs at m^{th} frequency bin, it causes stagnation, which stops updating $\hat{D}_k^2(m)$ (eq. (36)). Unlike monitoring the status of $P(H_1^m | R_k(m)) = 1$ for a long time as reported in [25], we simply resolve this issue by setting $P(H_1^m | R_k(m)) = 0.99$ once this condition occurs prior to update $\hat{D}_k^2(m)$.

It is observed that $R_k^2(m)$ is completely filled with additive noise during silent activity, thus giving an estimate of noise power. Therefore, unlike updating $\hat{D}_k^2(m)$ using

[1] The simplification is a result of assuming the *a priori* probability of the speech absence and presence, $P(H_0)$ and $P(H_1)$ as: $P(H_0) = P(H_1)$.

eq. (36) by existing method [25], we do it differently depending on the silent/speech activity of $R_k^2(m)$ (for each frequency bin m). Specifically, at m^{th} frequency bin ($k \geq 1$), if $P(H_1^m|R_k(m)) < 0.5$, $R_k^2(m)$ yields silent activity, resulting $\hat{D}_k^2(m) = R_k^2(m)$, otherwise, $\hat{D}_k^2(m)$ is estimated using eq. (36). With estimated $\hat{D}_k^2(m)$, $\hat{\lambda}_k(m)$ is updated as:

$$\hat{\lambda}_k(m) = \eta\hat{\lambda}_{k-1}(m) + (1-\eta)\hat{D}_k^2(m), \tag{38}$$

where the smoothing constant, η is set to 0.9.

The |IDFT| of $P_v(m)e^{j\phi_k(m)}$ yields the estimated noise, $\hat{v}(n,k)$, where $P_v(m) = \sqrt{\hat{\lambda}_k(m)}$. To ensure the conjugate symmetry, the components of $P_v(m)$ at $m\epsilon\{1,2,\ldots,127\}$ are flipped to that of the $m\epsilon\{129,130,\ldots,255\}$ of $P_v(m)$ before taking the |IDFT|.

3.3. Proposed $K_0(n)$ Tuning Method

Firstly, we construct KF with $\{\hat{a}_i\}, \hat{\sigma}_w^2)$ and $\hat{\sigma}_v^2$ and extract the tuning parameters as shown in Figure 4. It can be seen from Figure 4 (a) that $[\hat{\alpha}^2(n) + \hat{\sigma}_w^2]$ and $\hat{\sigma}_v^2$ achieves very similar characteristics as like KF-Oracle method (Figure 2 (e)). The improvement of these parameters also enables $J_2(n)$ and $J_1(n)$ metrics (Figure 4 (b)) to achieve quite similar characteristics as appear in WGN condition (Figure 1 (c)). Therefore, $J_2(n)$ and $J_1(n)$ metrics (Figure 4 (b)) are now eligible to dynamically tune $K_0(n)$ in real-life noise conditions. However, our investigation reveals that $J_2(n)$ metric is useful in tuning $K_0(n)$ during speech pauses, since it results under-estimated $K_0(n)$ during speech presence of noisy speech [20]. On the contrary, since $J_1(n)$ metric approaches 0 in speech regions of the noisy speech, according to eq. (32), it minimizes the under-estimation of $K_0(n)$. In light of these observations, for each sample of $y(n,k)$, we incorporate $J_2(n)$ metric during speech pauses and $J_1(n)$ metric during speech presence to dynamically *offset* the bias in $K_0(n)$.

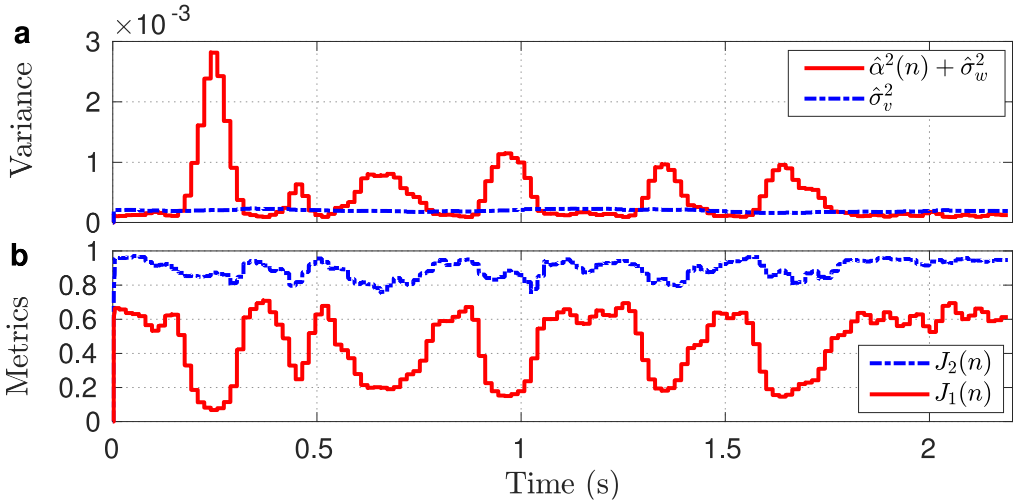

Figure 4. Comparing the estimated: (a) $[\hat{\alpha}^2(n) + \hat{\sigma}_w^2]$, $\hat{\sigma}_v^2$ and (b) $J_2(n)$, $J_1(n)$ metrics from the noisy speech in Figure 2 (b).

We studied and found that $[\hat{\alpha}^2(n) + \hat{\sigma}_w^2]$ and $\hat{\sigma}_v^2$ can be adopted as a speech activity detector for each sample of $y(n,k)$. For example, during speech pauses, the condition $\hat{\sigma}_v^2 \geq [\hat{\alpha}^2(n) + \hat{\sigma}_w^2]$ holds (e.g., 0-0.15 s or 1.8-2.19 s of Figure 4 (a)). Conversely, $[\hat{\alpha}^2(n) + \hat{\sigma}_w^2] >> \hat{\sigma}_v^2$ is found in speech regions (e.g., 0.16-0.33 s or 0.9-1.06 s of Figure 4 (a)). Therefore, at sample n, if $\hat{\sigma}_v^2 \geq [\hat{\alpha}^2(n) + \hat{\sigma}_w^2]$, $y(n,k)$ is termed as silent and set the decision parameter (denoted by ζ) as $\zeta(n) = 0$; otherwise speech activity occurs and $\zeta(n) = 1$. Figure 5 reveals that the detected flags (0/1: silent/speech) by the proposed method is closely similar to that of the reference (0/-1: silent/speech; generated by visually inspecting the utterance sp05).

doi:10.20944/preprints202103.0221.v1

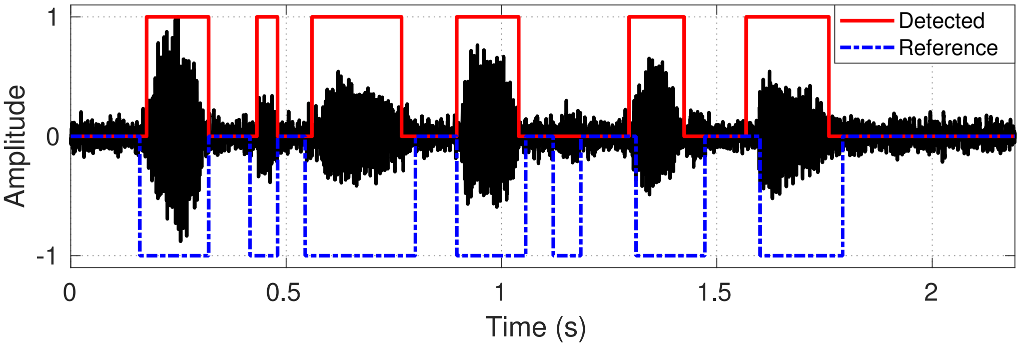

Figure 5. Comparing the detected flags of Figure 2 (b) to that of the reference corresponding to Figure 2 (a).

At sample n, if $\zeta(n) = 0$, the adjusted $K'_0(n)$ in the proposed SEA is given by:

$$\begin{aligned} K'_0(n) &= K_0(n)[1 - J_2(n)], \\ &= \left[\frac{\hat{\alpha}^2(n) + \hat{\sigma}_w^2}{\hat{\alpha}^2(n) + \hat{\sigma}_w^2 + \hat{\sigma}_v^2}\right]\left[\frac{\hat{\alpha}^2(n)}{\hat{\alpha}^2(n) + \hat{\sigma}_w^2}\right], \\ &= \frac{\hat{\alpha}^2(n)}{\hat{\alpha}^2(n) + \hat{\sigma}_w^2 + \hat{\sigma}_v^2}. \end{aligned} \tag{39}$$

To justify the validity of $K'_0(n)$, Figure 6 (a) shows the numerator and the denominator of eq. (39) computed from the noisy speech in Figure 2 (b). It can be seen that $\hat{\alpha}^2(n) \approx 0$ during speech pauses (e.g., 0-0.15 s or 1.8-2.19 s of Figure 6 (a)). According to eq. (39), it results $K'_0(n) \approx 0$. Since $[\hat{\alpha}^2(n) + \hat{\sigma}_w^2 + \hat{\sigma}_v^2] >> \hat{\alpha}^2(n)$ occurs during speech presence (e.g., 0.16-0.33 s or 0.9-1.06 s of Figure 6 (a)), it may result under-estimated $K'_0(n)$ as like WGN experiment (Figure 1 (d)). Thus, $J_2(n)$ metric-based tuning of $K'_0(n)$ in speech activity of $y(n,k)$ is inappropriate.

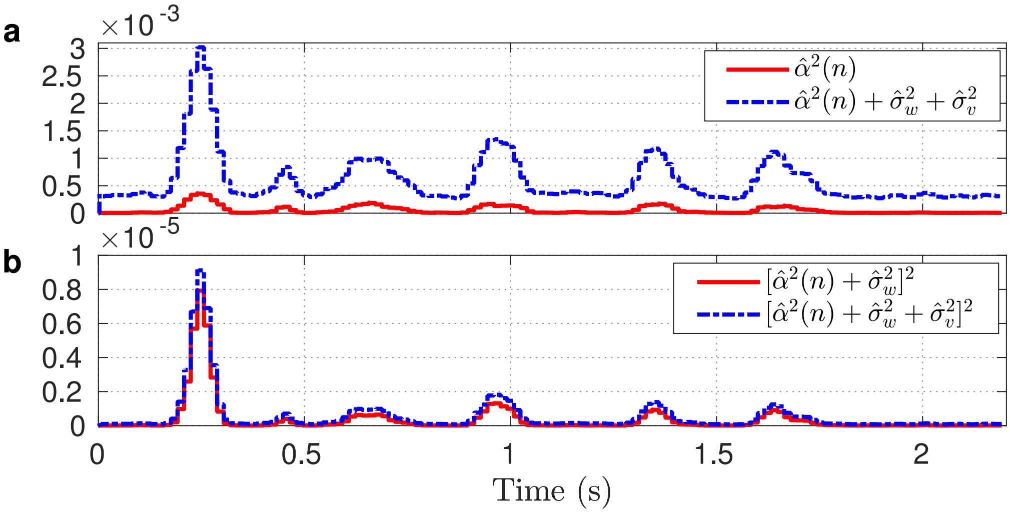

Figure 6. $K'_0(n)$ responses in terms of: (a) $\hat{\alpha}^2(n)$ and $\hat{\alpha}^2(n) + \hat{\sigma}_w^2 + \hat{\sigma}_v^2$, and (b) $[\hat{\alpha}^2(n) + \hat{\sigma}_w^2]^2$ and $[\hat{\alpha}^2(n) + \hat{\sigma}_w^2 + \hat{\sigma}_v^2]^2$, where the same experimental setup of Figure 2 (b) is used.

As discussed earlier, we do tuning biased $K_0(n)$ using $J_1(n)$ metric during speech activity of $y(n,k)$. However, our further investigation on $J_1(n)$ metric-based tuning in eq. (32) reveals that the subtraction of $J_1(n)$ from biased $K_0(n)$ may still produce under-estimated $K_0'(n)$. To cope with this problem, at sample n, if $\zeta(n) = 1$, we find a more effective solution for tuning of biased $K_0(n)$ using $J_1(n)$ metric as:

$$\begin{aligned} K_0'(n) &= K_0(n)[1 - J_1(n)], \\ &= \left[\frac{\hat{\alpha}^2(n) + \hat{\sigma}_w^2}{\hat{\alpha}^2(n) + \hat{\sigma}_w^2 + \hat{\sigma}_v^2}\right]\left[\frac{\hat{\alpha}^2(n) + \hat{\sigma}_w^2}{\hat{\alpha}^2(n) + \hat{\sigma}_w^2 + \hat{\sigma}_v^2}\right], \\ &= \frac{[\hat{\alpha}^2(n) + \hat{\sigma}_w^2]^2}{[\hat{\alpha}^2(n) + \hat{\sigma}_w^2 + \hat{\sigma}_v^2]^2}. \end{aligned} \tag{40}$$

To justify the validity of $K_0'(n)$, the numerator and the denominator of eq. (40) are shown in Figure 6 (b). It can be seen that $[\hat{\alpha}^2(n) + \hat{\sigma}_w^2 + \hat{\sigma}_v^2]^2 \geq [\hat{\alpha}^2(n) + \hat{\sigma}_w^2]^2$ during speech presence of $y(n,k)$ (e.g., 0.16-0.33 s or 0.9-1.06 s), which causes $K_0'(n)$ approaching 1.

To examine the performance of the proposed tuning algorithm in real-life non-stationary noise conditions, we repeat the experiment in Figure 2. It can be seen from

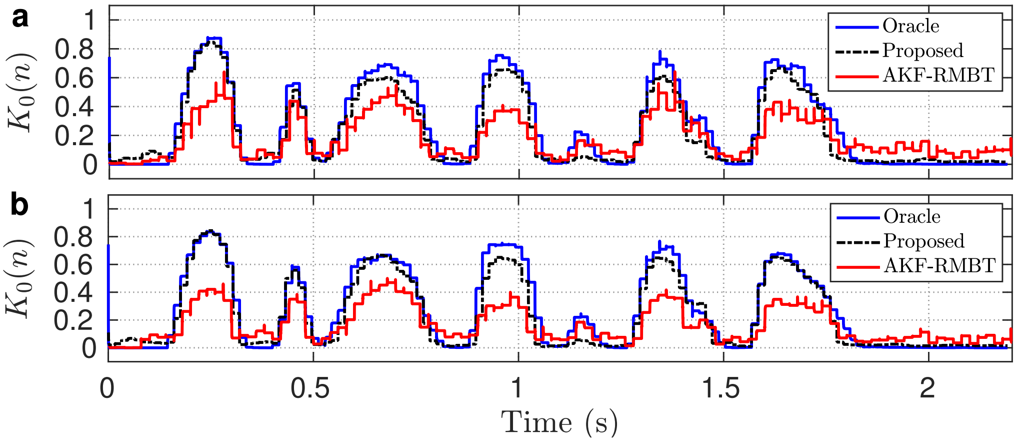

Figure 7. Comparing $K_0(n)$ obtained using KF-Oracle, Proposed, and AKF-RMBT [21] methods from the utterance sp05 corrupted with 5 dB: (a) non-stationary (*babble*) and (b) colored (*f16*) noises.

Figure 7 (a) that $K_0'(n)$ is closely similar to the oracle $K_0(n)$. Specifically, it maintains a smooth transition at the edges and the temporal changes in speech regions are closely matched to the oracle $K_0(n)$. Whereas the AKF-RMBT method [21] produces a significant under-estimated $K_0(n)$ in speech regions. Therefore, the reduced-biased $K_0'(n)$ in proposed method is more appropriate to mitigate the risks of *distortion* in the enhanced speech than that of AKF-RMBT method [21]. We also repeat the experiment in Figure 2 except the utterance sp05 is corrupted by 5 dB colored (*f16*) noise. Figure 7 (b) reveals that the biasing effect is reduced significantly in $K_0'(n)$ and closely similar to the oracle $K_0(n)$. However, the AKF-RMBT method [21] still produced under-estimated $K_0(n)$ in speech regions. In light of the comparative study, it is evident to say that the proposed method adequately addresses the tuning of biased $K_0(n)$ both in real-life non-stationary and colored noise conditions.

4. Speech Enhancement Experiment

4.1. Corpus

To evaluate the efficiency of the competitive methods, 30 phonetically balanced utterances belonging to six speakers (3 male and 3 female) are taken from the NOIZEUS corpus sampled at 8 kHz [1, Chapter 12]. For objective experiments, we generate a noisy speech data set that has been corrupted by real-world non-stationary (*babble, street*) and colored (*factory2* and *f16*) noises for a wide range of SNR levels (from -5 dB to 15 dB). The *street* noise is taken from [26] and the rest of the noises are from [23].

4.2. Objective Evaluation

The objective measures are used to evaluate the quality and intelligibility of the enhanced speech with respect to the corresponding clean speech. The following objective evaluation metrics have been used in this paper:

- Perceptual Evaluation of Speech Quality (PESQ) for objective quality evaluation [27]. The PESQ score ranged between -0.5 to 4.5. A higher PESQ score indicating better speech quality.
- Short-time objective intelligibility (STOI) measure for objective intelligibility evaluation [28]. It ranged between 0 to 1 (or 0 to 100%). The higher STOI score indicates better speech intelligibility.

4.3. Spectrogram Evaluation

We also analyzed the spectrograms of enhanced speech produced by the proposed and the competitive methods to visually quantify the level of *residual* noise as well as *distortion*. For this purpose, we generate a noisy speech data set by corrupting the utterance sp05 with 5 dB *babble* (non-stationary) and 5 dB *f16* (colored) noises.

4.4. Subjective Evaluation

The subjective evaluation was carried out through a series of blind AB listening tests [5, Section 3.3.4]. To perform these tests, we have used the same noisy speech dataset (4.3). In this test, the enhanced speech produced by 6 SEAs as well as the corresponding clean speech and noise corrupted speech signals were played as stimuli pairs to the listeners. Specifically, the test is performed on a total of 112 stimuli pairs (56 for each utterance) played in a random order to each listener, excluding the comparisons between the same method.

The listener gives the following ratings for each stimuli pair: prefers the first or second stimuli which is perceptually better, or a third response indicating no difference is found between them. For a pairwise scoring, 100% award is given to the preferred method, 0% to the other, and 50% for the similar preference response. The participants could re-listen to stimuli if required. Ten English speaking listeners participate in the blind AB listening tests. The average of the preference scores given by the listeners, termed as mean preference score (%), which is used to compare the efficiency among the SEAs.

4.5. Specifications of the Competitive SEAs

The performance of the proposed SEA is carried out by comparing it with the following benchmark SEAs (p : order of $\{a_i\}$, σ_w^2: the excitation variance of AR model, w : analysis frame duration (ms), s : analysis frame shift (ms)).

1. **Noisy**: No-enhancement (speech corrupted with noise).
2. **KF-Oracle:** KF, where ($\{a_i\}$, σ_w^2) and σ_v^2 are computed from the clean speech and the noise signal, $p = 10$, $w = 32$ ms, $s = 16$ ms, and rectangular window is used for framing.
3. **KF-Non-oracle:** KF, where ($\{a_i\}$, σ_w^2) and σ_v^2 are computed from the noisy speech, $p = 10$, $w = 32$ ms, $s = 16$ ms, and rectangular window is used for framing.
4. **MMSE-STSA** [9]: It used $w = 25$ ms, $s = 10$ ms, and Hamming window for framing.
5. **AKF-IT** [13]: AKF operates with two iterations, where initial ($\{a_i\}$, σ_w^2) and ($\{b_j\}$, σ_u^2) are computed from the noisy speech followed by re-estimation of them from the processed speech after first iteration, $p = 10$, noise LPC order $q = 10$, $w = 20$ ms, $s = 0$ ms, and rectangular window is used for framing.
6. **AKF-RMBT** [21]: Robustness metric-based tuning of the AKF, where ($\{a_i\}$, σ_w^2) and ($\{b_j\}$, σ_u^2) are computed from the pre-whitened speech and initial silent frames, $p = 10$, $q = 40$, $w = 20$ ms, $s = 0$ ms, and rectangular window is used for framing.

7. **Proposed:** Robustness and sensitivity tuning of the KF, where $(\{\hat{a}_i\}, \hat{\sigma}_w^2)$ and $\hat{\sigma}_v^2$ are computed from the pre-whitened speech and estimated noise, $p = 20$, $q = 20$, $w = 32$ ms, $s = 16$ ms, rectangular window is used for time-domain frames, and Hamming window is used for acoustic frames.

5. Results and Discussion

5.1. Objective Quality Evaluation

Figure 8 shows the average PESQ score (found over all frames for each test condition in Section 4.1) for each SEA. It can be seen that the KF-Oracle method exhibits the highest PESQ score for all test conditions. It is due to $(\{a_i\}, \sigma_w^2)$ and σ_v^2 are computed from the clean speech and the noise signal. Thus, informally, it can be considered as an upper-bound of PESQ score improvement. The improvement of average PESQ score for KF-Non-oracle method is marginal as compared to the no enhancement (Noisy). The proposed SEA shows a considerable PESQ score improvement than the benchmark methods across the test conditions. The average PESQ score for proposed method is also very similar to that of the KF-Oracle method. It is due to the reduced-biased Kalman gain obtained by proposed tuning algorithm is closely similar to that of the KF-Oracle method (Figure 7). Among the benchmark methods, the AKF-RMBT method [21] and the MMSE-STSA method [9] relatively show competitive PESQ scores for all noise conditions (Figure 9 (a)-(d)). On the other hand, since the AKF-IT method [13] suffers from *distortion* and *musical* noise in the enhanced speech, it gives reduced PESQ scores than other benchmark methods across the test conditions. In light of this comparative study, it is evident to say that the proposed SEA exhibits better quality in the enhanced speech than that of the benchmark methods for all noise conditions.

5.2. Objective Intelligibility Evaluation

Figure 9 shows the average STOI score (found over all frames for each test condition in Section 4.1). The high STOI score of enhanced speech indicates intelligibility improvements. As like PESQ score comparison (section 5.1), the KF-Oracle method also achieves the highest STOI score for all noise conditions. When comparing the average STOI score, the proposed method consistently outperforming all other methods across the noise conditions. The STOI score improvement by proposed method is also very similar to that of the KF-Oracle method. When comparing the average STOI score among the benchmark methods, the AKF-RMBT method [21] is found to be competitive with the proposed method in *street* and other colored noise conditions typically at high SNR levels (Figure 9 (b)-(d)). While the AKF-IT method [13] shows improved STOI score at low SNR levels than the MMSE-STSA method [9] and the KF-Non-oracle method across the noise conditions. On the other hand, the STOI score for no enhancement case (Noisy) is competitive, even shows slightly better at high SNR levels than the KF-Non-oracle method. Since the noisy speech signal remains unprocessed, it may produce competitive STOI score regardless of the background noise as observed in this study.

5.3. Spectrogram Analysis of the SEAs

Figs. 10-11 compare the spectrograms of enhanced speech produced by each SEA for noisy speech data set (section 4.2). Typically, the noise reduction is visibly improved when going from the KF-Non-oracle method to the KF-Oracle method. Specifically, the biased gain of the KF-Non-oracle method passes a significant residual noise in the enhanced speech (Figs. 10 (c)-11 (c)). Also, the poor estimates of the *a priori* SNR introduces a high degree of residual noise in the enhanced speech produced by the MMSE-STSA method [9] (Figs. 10 (d)-11 (d)). Whereas the degree of residual noise decreases in the enhanced speech produced by the AKF-IT method [13] (Figs. 10 (e)-11 (e)). However, the residual noise appears as musical noise. The enhanced speech also gets distorted due to processing the noisy speech iteratively by AKF. The AKF-RMBT method [21] exhibits less residual noise in the enhanced speech, however, suffering

doi:10.20944/preprints202103.0221.v1

from distortion due to the under-estimated Kalman gain (Figs. 10 (f)-11 (f)). When comparing the proposed method, the enhanced speech is almost free of residual noise and distortion, which are closely similar to the KF-Oracle method. It is due to the reduced-biased Kalman gain of the proposed method, which is very similar to that of the KF-Oracle method. Among the benchmark methods, the informal listening tests also confirm that the MMSE-STSA [9] and AKF-IT [13] methods produce annoying sounds as compared to the AKF-RMBT method [21] in the enhanced speech. While the enhanced speech produced by the proposed method has a negligible audio artifact as in KF-Oracle method across the noise conditions.

5.4. Subjective Evaluation by AB listening test

The mean preference score (%) comparison for all methods are shown in Figs. 12-13. The non-stationary (*babble*) noise experiment in Figure 12 reveals that the proposed method is widely preferred (73%) by the listeners to that of the benchmark methods, apart from the clean speech and the KF-Oracle method (81%). Whereas the AKF-RMBT [21] method is found to be the best preferred method (60%) among the benchmark methods by the listeners. Although the AKF-IT method [21] produced distorted speech as confirmed by objective PESQ and STOI score comparison as well as spectrogram analysis, however, the listeners prefer it (47%) over the MMSE-STSA method (31%) [9]. The subjective testing implies that it was considered as an improvement of noise reduction in speech region than a distortion. The colored (*f16*) noise experiment (Figure 13) also confirms that the proposed method achieves a significant preference score (75%) than the benchmark methods, except the clean speech and the KF-Oracle method (82%). Among the benchmark methods, the AKF-RMBT [21] method is found to be the best preferred method (64%) followed by the AKF-IT method [13] (around 48%) by the listeners. In light of the blind AB listening tests, the proposed method is found to be quite successful in reducing noise for both male and female utterances corrupted by real-life non-stationary and colored noises.

6. Conclusion

Robustness and sensitivity metrics-based tuning of the Kalman filter gain for single-channel speech enhancement has been investigated in this paper. At first, the noise variance is computed from the estimated noise for each noisy speech frame using a speech presence probability method. A whitening filter is also constructed to pre-whiten each noisy speech frame prior to estimate LPC parameters. Then, the robustness and the sensitivity metrics are incorporated differently depending on the speech activity of the noisy speech to dynamically *offset* the bias in Kalman gain. Where the noise variance and the AR model parameters are adopted as a speech activity detector. It is shown that the proposed tuning algorithm yields a significant reduced-biased Kalman gain, which enables the KF to minimize the residual noise and distortion in the enhanced speech. Extensive objective and subjective testing on NOIZEUS corpus demonstrates that the proposed method outperforms the benchmark methods in real-life noise conditions for a wide range of SNR levels.

Author Contributions: The contribution of Sujan Kumar Roy includes: preliminary experiments, experiment design, conducted the experiments, code writing, design of models, analysis of results, literature review, and manuscript writing. Kuldip K. Paliwal provided supervision and aided with editing the final manuscript

Institutional Review Board Statement: The subjective AB listening tests in Section were conducted on the approval of Griffith University Human Research Ethics: database protocol number 2018/671.

Conflicts of Interest: The authors declare no conflict of interest.

References

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

2. Boll, S. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **1979**, *27*, 113–120. doi:10.1109/TASSP.1979.1163209.
3. Berouti, M.; Schwartz, R.; Makhoul, J. Enhancement of speech corrupted by acoustic noise. *IEEE International Conference on Acoustics, Speech, and Signal Processing* **1979**, *4*, 208–211. doi:10.1109/TASSP.1979.1163209.
4. Kamath, S.; Loizou, P. A Multi-Band Spectral Subtraction Method for Enhancing Speech Corrupted by Colored Noise. *IEEE International Conference on Acoustics, Speech, and Signal Processing* **2002**, *4*, 4160–4164. doi:10.1109/ICASSP.2002.5745591.
5. Paliwal, K.; Wójcicki, K.; Schwerin, B. Single-channel Speech Enhancement Using Spectral Subtraction in the Short-time Modulation Domain. *Speech Communication* **2010**, *52*, 450–475. doi:https://doi.org/10.1016/j.specom.2010.02.004.
6. Lim, J.S.; Oppenheim, A.V. Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE* **1979**, *67*, 1586–1604. doi:10.1109/PROC.1979.11540.
7. Scalart, P.; Filho, J.V. Speech enhancement based on a priori signal to noise estimation. *IEEE International Conference on Acoustics, Speech, and Signal Processing* **1996**, *2*, 629–632.
8. Plapous, C.; Marro, C.; Mauuary, L.; Scalart, P. A two-step noise reduction technique. *IEEE International Conference on Acoustics, Speech, and Signal Processing* **2004**, *1*, 289–292.
9. Ephraim, Y.; Malah, D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **1984**, *32*, 1109–1121. doi:10.1109/TASSP.1984.1164453.
10. Ephraim, Y.; Malah, D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **1985**, *33*, 443–445. doi:10.1109/TASSP.1985.1164550.
11. Paliwal, K.; Schwerin, B.; Wójcicki, K. Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator. *Speech Communication* **2012**, *54*, 282–305. doi:https://doi.org/10.1016/j.specom.2011.09.003.
12. Paliwal, K.; Basu, A. A speech enhancement method based on Kalman filtering. *IEEE International Conference on Acoustics, Speech, and Signal Processing* **1987**, *12*, 177–180. doi:10.1109/ICASSP.1987.1169756.
13. Gibson, J.D.; Koo, B.; Gray, S.D. Filtering of colored noise for speech enhancement and coding. *IEEE Transactions on Signal Processing* **1991**, *39*, 1732–1742. doi:10.1109/78.91144.
14. Wang, Y.; Wang, D. Towards Scaling Up Classification-Based Speech Separation. *IEEE Transactions on Audio, Speech, and Language Processing* **2013**, *21*, 1381–1390.
15. Xu, Y.; Du, J.; Dai, L.; Lee, C. An Experimental Study on Speech Enhancement Based on Deep Neural Networks. *IEEE Signal Processing Letters* **2014**, *21*, 65–68.
16. Williamson, D.S.; Wang, Y.; Wang, D. Complex Ratio Masking for Monaural Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2016**, *24*, 483–492.
17. Roy, S.K.; Zhu, W.P.; Champagne, B. Single channel speech enhancement using subband iterative Kalman filter. *IEEE International Symposium on Circuits and Systems* **2016**, pp. 762–765. doi:10.1109/ISCAS.2016.7527352.
18. Saha, M.; Ghosh, R.; Goswami, B. Robustness and Sensitivity Metrics for Tuning the Extended Kalman Filter. *IEEE Transactions on Instrumentation and Measurement* **2014**, *63*, 964–971. doi:10.1109/TIM.2013.2283151.
19. So, S.; George, A.E.W.; Ghosh, R.; Paliwal, K.K. A non-iterative Kalman filtering algorithm with dynamic gain adjustment for single-channel speech enhancement. *International Journal of Signal Processing Systems* **2016**, *4*, 263–268. doi:10.18178/ijsps.4.4.263-268.
20. So, S.; George, A.E.W.; Ghosh, R.; Paliwal, K.K. Kalman Filter with Sensitivity Tuning for Improved Noise Reduction in Speech. *Circuits, Systems, and Signal Processing* **2017**, *36*, 1476–1492. doi:10.1007/s00034-016-0363-y.
21. George, A.E.; So, S.; Ghosh, R.; Paliwal, K.K. Robustness metric-based tuning of the augmented Kalman filter for the enhancement of speech corrupted with coloured noise. *Speech Communication* **2018**, *105*, 62–76. doi:https://doi.org/10.1016/j.specom.2018.10.002.
22. V. Vaseghi, S. Linear Prediction Models. In *Advanced Digital Signal Processing and Noise Reduction*; John Wiley & Sons, 2009; chapter 8, pp. 227–262.
23. Varga, A.; Steeneken, H.J. 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* **1993**, *12*, 247–251. doi:https://doi.org/10.1016/0167-6393(93)90095-3.
24. Oppenheim, A.V.; Schafer, R.W. *Discrete-Time Signal Processing*, 3rd ed.; Prentice Hall Press: Upper Saddle River, NJ, USA, 2009.
25. Gerkmann, T.; Hendriks, R.C. Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay. *IEEE Transactions on Audio, Speech, and Language Processing* **2012**, *20*, 1383–1393. doi:10.1109/TASL.2011.2180896.
26. Pearce, D.; Hirsch, H.G. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. INTERSPEECH. ISCA, 2000, pp. 29–32.
27. Rix, A.W.; Beerends, J.G.; Hollier, M.P.; Hekstra, A.P. 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* **2001**, *2*, 749–752. doi:10.1109/ICASSP.2001.941023.
28. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. *IEEE Transactions on Audio, Speech, and Language Processing* **2011**, *19*, 2125–2136. doi:10.1109/TASL.2011.2114881.

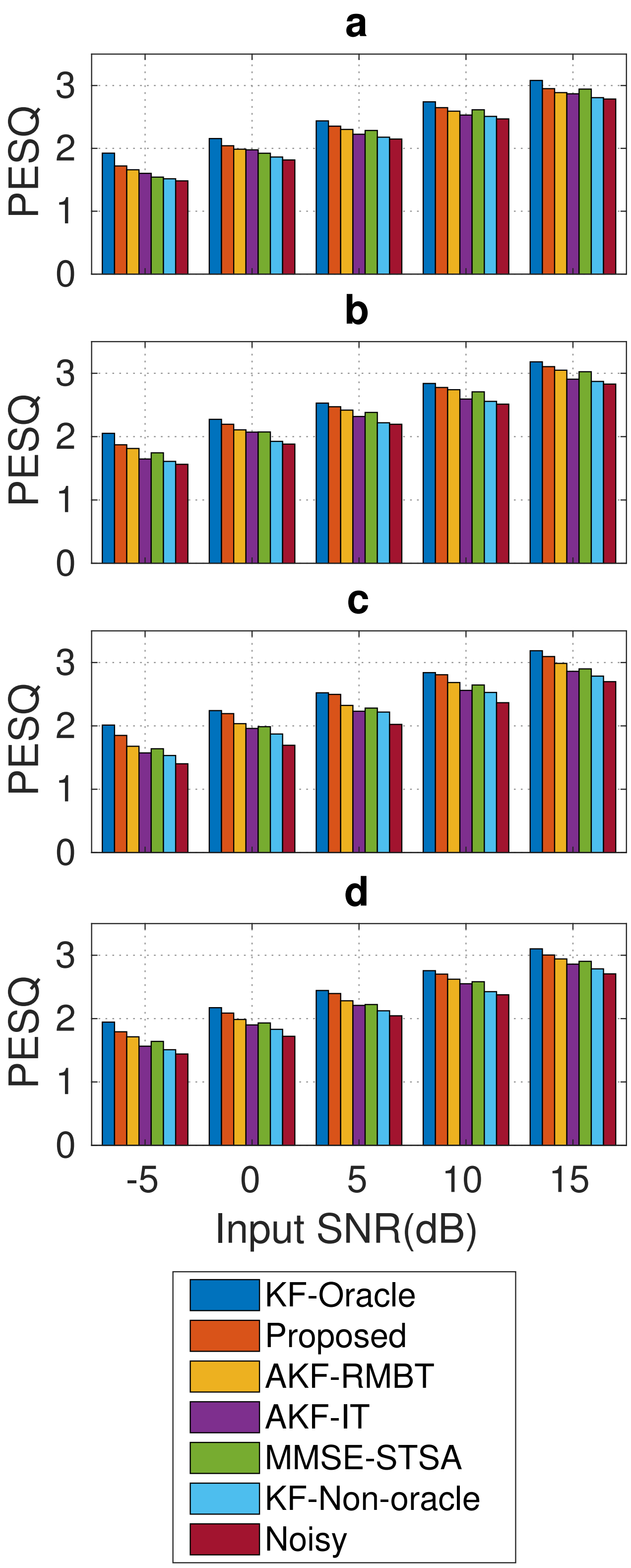

Figure 8. Average PESQ score comparison between the proposed and benchmark SEAs on NOIZEUS corpus corrupted with: (a) *babble*, (b) *street*, (c) *factory2*, and (d) *f16* noises for a wide range of SNR levels (from -5 dB to 15 dB).

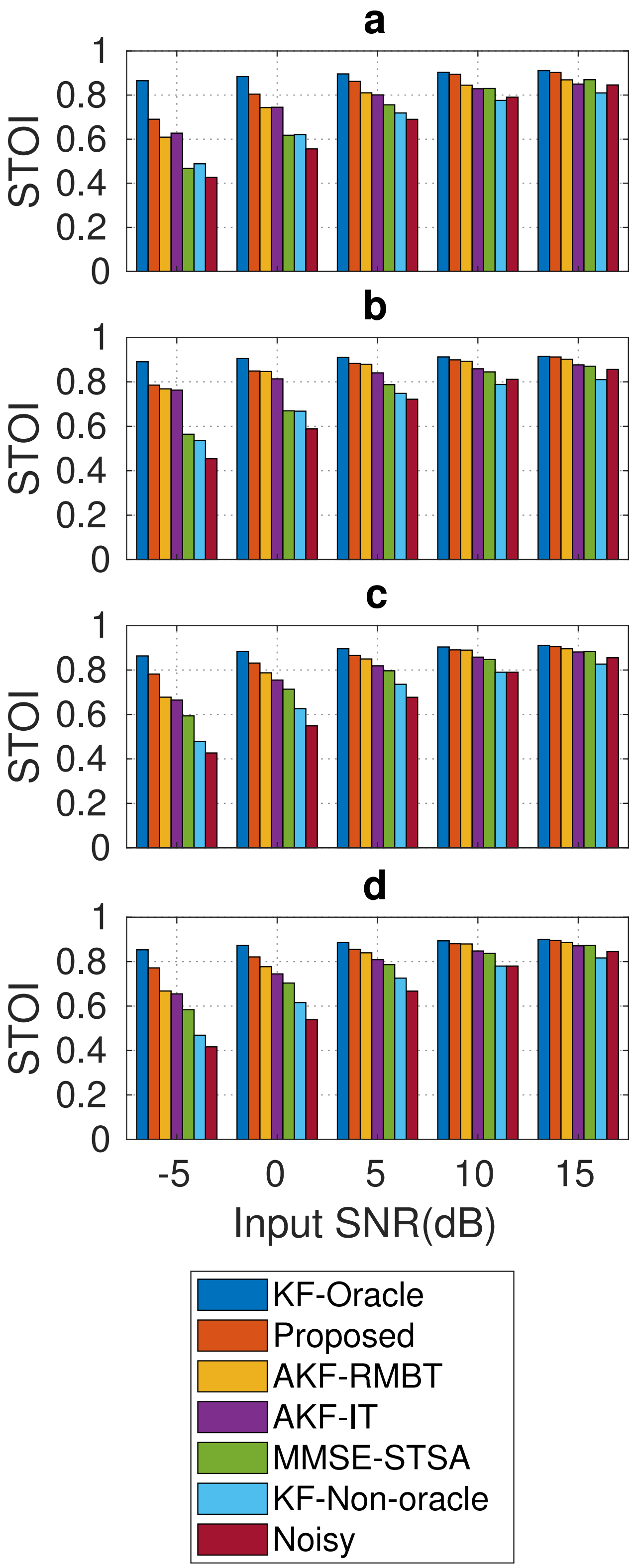

Figure 9. Average STOI score comparison between the proposed and benchmark SEAs on NOIZEUS corpus corrupted with: (a) *babble*, (b) *street*, (c) *factory2*, and (d) *f16* noises for a wide range of SNR levels (from -5 dB to 15 dB).

doi:10.20944/preprints202103.0221.v1

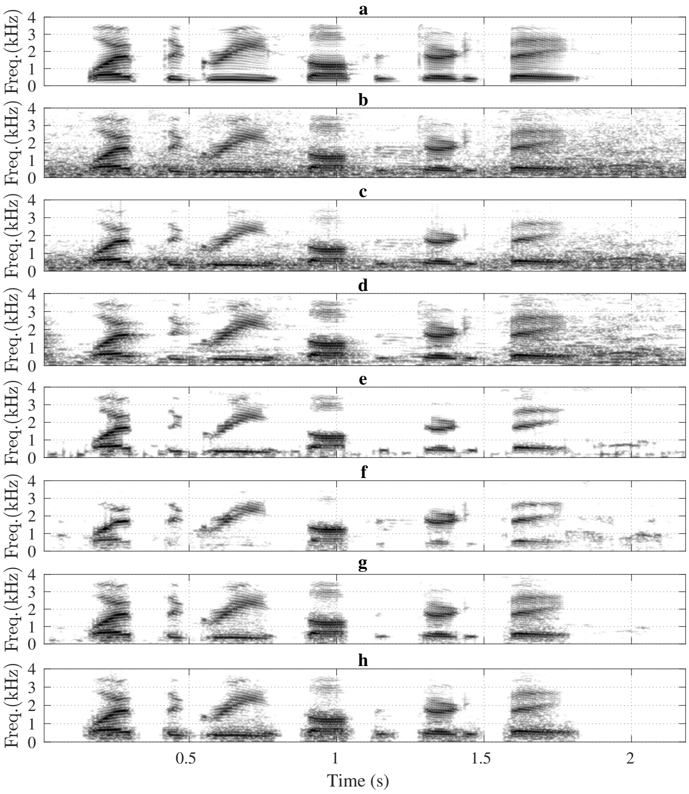

Figure 10. Comparing the spectrograms of: (a) clean speech (utterance sp05), (b) noisy speech (corrupt sp05 with 5 dB *babble* noise) (PESQ=2.10), enhanced speech produced by the: (c) KF-Non-oracle (PESQ=2.18), (d) MMSE-STSA [9] (PESQ=2.32), (e) AKF-IT [13] (PESQ=2.26), (f) AKF-RMBT [21] (PESQ=2.42), (g) Proposed (PESQ=2.55), and (h) KF-Oracle (PESQ=2.61) methods.

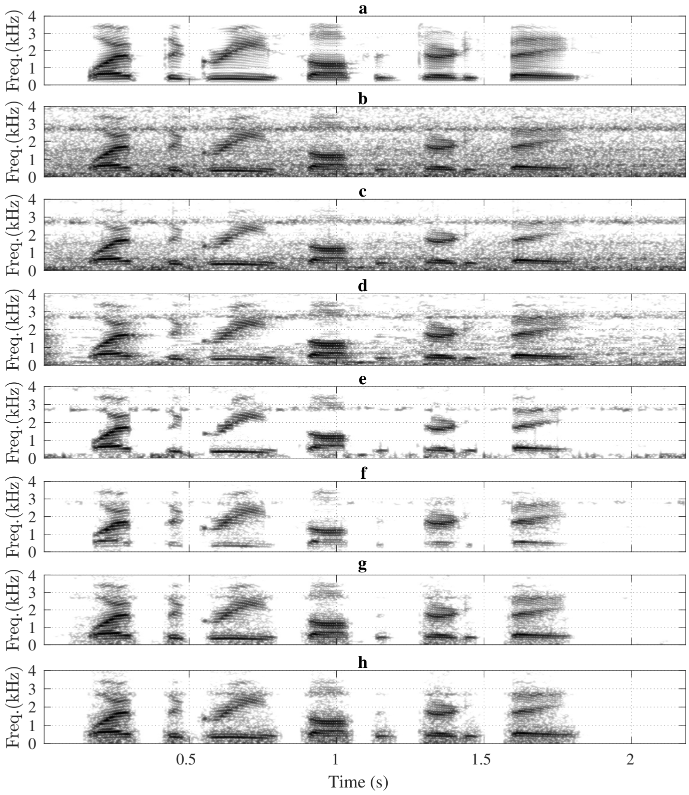

Figure 11. Comparing the spectrograms of: (a) clean speech (utterance sp05), (b) noisy speech (corrupt sp05 with 5 dB *f16* noise) (PESQ=2.14), enhanced speech produced by the: (c) KF-Non-oracle (PESQ=2.26), (d) MMSE-STSA [9] (PESQ=2.39), (e) AKF-IT [13] (PESQ=2.31), (f) AKF-RMBT [21] (PESQ=2.53), (g) Proposed (PESQ=2.65), and (h) KF-Oracle (PESQ=2.70) methods.

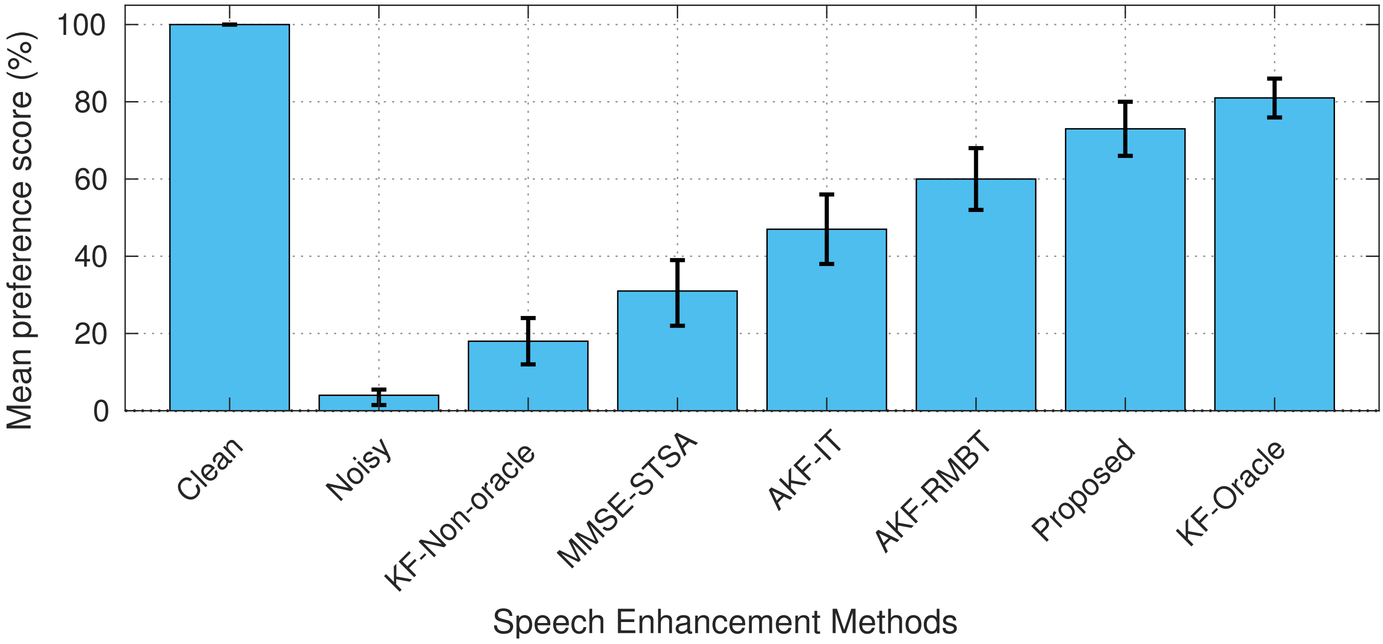

Figure 12. The mean preference score (%) comparison between the proposed and benchmark SEAs for the utterance sp05 corrupted with 5 dB non-stationary *babble* noise.

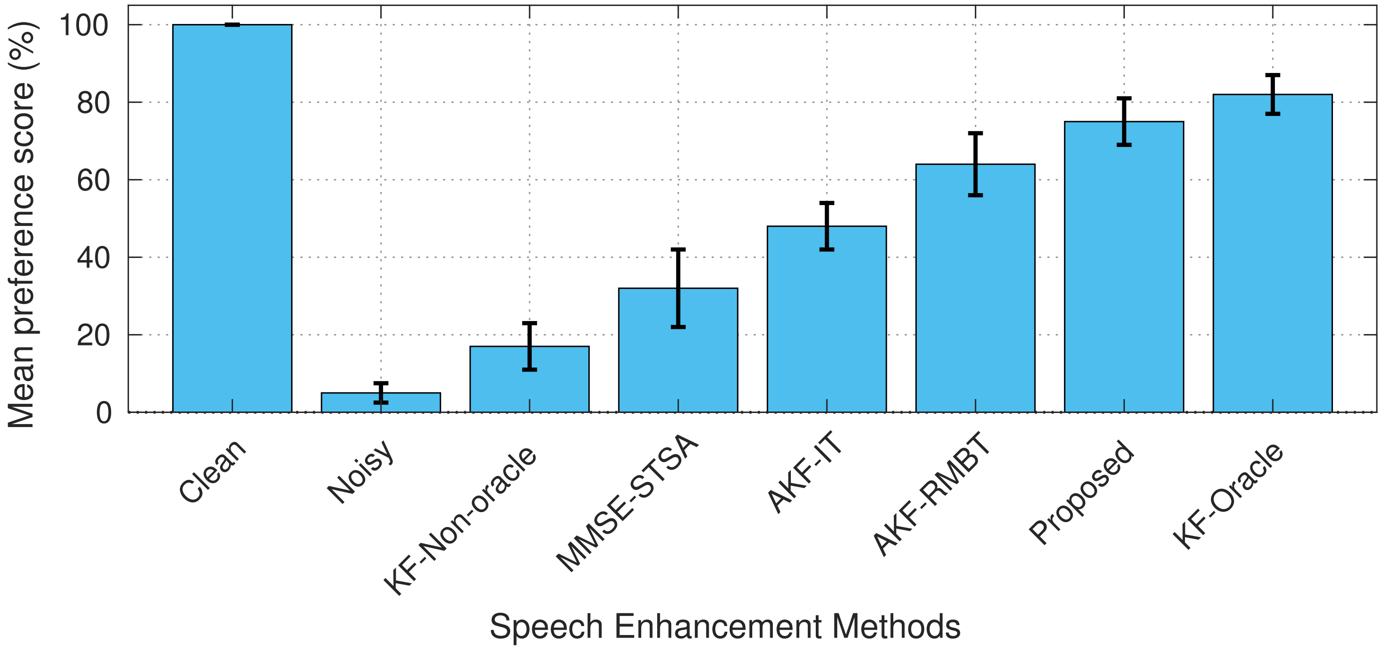

Figure 13. The mean preference score (%) comparison between the proposed and benchmark SEAs for the utterance sp27 corrupted with 5 dB colored *f16* noise.