Applied Acoustics 185 (2022) 108355

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Robustness and sensitivity metrics-based tuning of the augmented Kalman filter for single-channel speech enhancement

Sujan Kumar Roy *, Kuldip K. Paliwal

Signal Processing Laboratory, Griffith University, Nathan Campus, Brisbane, QLD, 4111, Australia

ARTICLE INFO

Article history:
Received 4 March 2021
Received in revised form 28 June 2021
Accepted 9 August 2021

Keywords:
Speech enhancement
Kalman filter
Augmented Kalman filter
Robustness metric
Sensitivity metric
LPC

ABSTRACT

The inaccurate estimates of the speech and noise linear prediction coefficients (LPCs) introduce bias in augmented Kalman filter (AKF) gain, which impacts the quality and intelligibility of enhanced speech. Although current tuning methods *offset* the bias in AKF gain, particularly in colored noise conditions, they do not adequately address nonstationary noise conditions. This paper introduces a new tuning algorithm of the AKF gain for speech enhancement in real-life noise conditions. Due to this purpose, a speech presence probability (SPP) method first estimates the noise power spectral density (PSD) from each noisy speech frame to compute the noise LPC parameters. A whitening filter is constructed with the noise LPCs to pre-whiten each noisy speech frame prior to computing the speech LPC parameters. The AKF is then constructed with the estimated speech and noise LPC parameters. To achieve better noise reduction, the robustness metric is employed to dynamically *offset* the bias in AKF gain during speech absence of the noisy speech to that of the sensitivity metric during speech presence. The speech activity is obtained through adopting the speech and noise production model parameters. It is shown that the reduced-biased AKF gain achieved by the proposed tuning algorithm addresses speech enhancement in real-life noise conditions. Objective and subjective scores on the NOIZEUS corpus demonstrate that the proposed method produces enhanced speech with higher quality and intelligibility than the competing methods in real-life noise conditions for a wide range of signal-to-noise ratio (SNR) levels.

1. Introduction

The main objective of a speech enhancement algorithm (SEA) is to improve the quality and intelligibility of degraded speech. It can be achieved through eliminating the embedded noise from the degraded speech. SEA is useful in many applications, where the noise corrupted speech is unreliable. For example, mobile communication systems, hearing aid devices, and speech recognition systems typically rely upon the accuracy of speech enhancement for robustness. Various SEAs, such as spectral subtraction (SS) [1–4], the Wiener Filter (WF) [5–7], minimum mean square error (MMSE) [8–10], the Kalman filter (KF) [11], augmented KF (AKF) [12], computational auditory scene analysis (CASA) [13], and deep neural network (DNN) [14] have been introduced over the decades. This paper focuses on AKF-based single-channel speech enhancement in real-life noise conditions.

KF was first used for speech enhancement by Paliwal and Basu [11]. In KF, each clean speech frame is represented by an auto-regressive (AR) model, whose parameters comprise the linear prediction coefficients (LPCs) and prediction error variance. The LPC parameters and additive noise variances are used to construct the KF recursive equations. Given a frame of noisy speech samples, the KF gives a linear MMSE estimate of the clean speech samples using the recursive equations. Therefore, the KF performance for speech enhancement largely depends upon the accuracy of LPCs, prediction error variance, and additive noise variance estimation in practice.

The KF methods in [11] usually proposed for speech enhancement in stationary noise conditions. However, in practice, most of the additive noise is non-stationary in nature— containing time-varying amplitudes. To perform speech enhancement other than stationary noise condition, such as colored noise conditions, in [12], Gibson et al. introduced an augmented KF (AKF). In AKF, both the clean speech and additive noise are represented by two AR models. Unlike KF [11], the clean speech and noise LPC parameters are incorporated in an augmented matrix form to construct the recursive equations of AKF. Due to incorporating the dynamic model of additive noise in the recursive equations, it is more appropriate to apply AKF for speech enhancement in real-life noise

* Corresponding author.
E-mail addresses: sujankumar.roy@griffithuni.edu.au (S.K. Roy), k.paliwal@griffith.edu.au (K.K. Paliwal).

https://doi.org/10.1016/j.apacoust.2021.108355
0003-682X/

conditions. For example, in [12], the AKF processes the colored noise corrupted speech iteratively (usually three-four iterations) to eliminate the embedded noise, yielding the enhanced speech. During this, the LPC parameters for the current frame are computed from the corresponding filtered speech frame of the previous iteration by AKF. Although the AKF demonstrates an improvement in signal-to-noise ratio (SNR) of the noisy speech, however, it suffers from *musical* noise and speech *distortion*. Therefore, the AKF method in [12] does not adequately address the inaccurate speech and noise LPC parameter estimation issue in practice.

In [15], Roy et al. proposed a sub-band (SB) iterative KF (SBIT-KF)-based SEA. In this SEA, the noisy speech is first decomposed into 16 sub-bands (SBs). Then a partial reconstruction of noisy speech is made with the high-frequency SBs (HFSBs). An iterative KF (two iterations) is employed to the partially reconstructed noisy speech, yielding a partial enhanced speech. As in [12], the speech LPC parameters for the current frame are computed from the corresponding filtered speech frame of the previous iteration by KF. Also, the noise variance is estimated using a derivative-based *high-pass* filter from each frame of the partially reconstructed noisy speech. Conversely, the low-frequency SBs (LFSBs) keep unprocessed with the assumption that the impact of noise on LFSBs is negligible. The partial enhanced speech is then added with the LFSBs to reconstruct the final enhanced speech. However, the LFSBs can also be affected by noise typically when operating in conditions that have time-varying amplitudes. As demonstrated in [12], the iterative processing of the partially reconstructed noisy speech using KF [15] also produced distorted speech.

In [16], Saha et al. proposed a robustness metric and a sensitivity metric for tuning the bias in KF gain for instrument engineering applications. Later on, So et al. employed the tuning of KF gain in speech enhancement context [17]. Specifically, it is shown in [17] that 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. However, the inaccurate estimates of the LPC parameters introduce bias in KF gain, results in leaking a significant *residual* noise in the enhanced speech. In [17], a robustness metric is used to *offset* the bias in KF gain for speech enhancement. In [18], So et al. further showed that the robustness metric strongly suppresses the KF gain in speech regions, resulting in distorted speech. To cope with this problem, in [18], a sensitivity metric was used to *offset* the bias in KF gain. It was shown that the sensitivity tuning of the KF gain produced less distorted speech than that of [17]. However, both of the KF methods [17,18] address speech enhancement, particularly in stationary white noise condition. In [19], George et al. introduced a robustness metric-based tuning of the AKF (AKF-RMBT) for speech enhancement in colored noise conditions. Firstly, the noise LPC parameters are computed from the first noisy speech frame by assuming that there remains no speech. The computed noise LPC parameters remain constant during processing all noisy speech frames for a given noisy speech utterance. A whitening filter is also constructed with the noise LPCs to pre-whiten each noisy speech frame prior to computing the speech LPC parameters. Then construct the AKF with the estimated LPC parameters. As like [17], it is shown that the robustness metric-based tuning method *offsets* the bias in AKF gain for silent frames to some extent; however, it over-suppresses the components in speech regions, resulting in distorted speech. In addition, the speech and noise LPC parameters estimation process as well as the tuning method in [19] do not account for conditions that have time-varying amplitudes. In [20], Roy and Paliwal proposed an extension of the work [19] by employing a sensitivity metric-based tuning of the AKF (AKF-SMBT). In this SEA, the speech and noise LPC parameters are computed with a similar process as in [19]. It is demonstrated that the application of sensitivity metric in the proposed tuning method [20] minimizes the underestimation issue of AKF gain, particularly in speech regions [19]. It is also shown that the reduced-biased AKF gain in [20] minimizes the amount of *residual* noise as well as *distortion* in the enhanced speech as compared to [19]. However, this SEA [20] also does not account for conditions that have time-varying amplitudes.

Motivated by the shortcomings of previously proposed KF and AKF methods [17–20], in this paper, we introduce a new tuning algorithm to dynamically *offset* the bias in AKF gain— which addresses speech enhancement in conditions that have time-varying amplitudes. For this purpose, we first estimate the noise power spectral density (PSD) from each noisy speech frame using a speech presence probability (SPP) method to compute the noise LPC parameters. A whitening filter is also constructed with the noise LPCs to pre-whiten each noisy speech frame prior to computing the clean speech LPC parameters. The AKF is then constructed with the estimated clean speech and noise LPC parameters, where a robustness metric is employed to dynamically *offset* the bias in AKF gain when there is speech absent of the noisy speech to that of the sensitivity metric during speech presence to achieve better noise reduction. The proposed method aims to mitigate the weaknesses of previously proposed tuning methods by providing a reduced-biased AKF gain— even for noise conditions that have time-varying amplitudes. The motivation of this is to produce enhanced speech at a higher quality and intelligibility in real-life noise conditions.

The structure of this paper is as follows: background knowledge is presented in Section 2, including the signal model, AKF for speech enhancement, paradigm shift of the AKF recursive equations, and the impact of biased AKF gain on speech enhancement in colored as well as non-stationary noise conditions. Following this, Section 3 describes the proposed SEA, which includes speech and noise LPC parameter estimation and proposed AKF gain tuning method. Section 4 describes the experimental setup in terms of speech corpus, objective and subjective evaluation measures, and specifications of the competing SEAs. The experimental results are then presented in Section 5. Finally, Section 6 gives some concluding remarks.

2. Background

2.1. Signal model

The noisy speech $y(n)$, at discrete-time sample n, is assumed to be given by:

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

where $s(n)$ is the clean speech and $v(n)$ is uncorrelated additive noise. Since the AKF operates on a frame-by-frame basis for speech enhancement, firstly, a 20 ms rectangular window with 0% overlap is used to convert $y(n)$ into frames [19], denoted by $y(n,l)$:

$$y(n,l) = s(n,l) + v(n,l), \tag{2}$$

where $l_Q\{0,1,2,\ldots,L-1\}$ is the frame index, L is the total number of frames in an utterance, and N is the total number of samples within each frame, i.e., $n_Q\{0,1,\ldots,N-1\}$.

2.2. AKF for speech enhancement

For simplicity, the frame index is omitted in the AKF recursive equations. Each frame of the clean speech and noise signal in (2) can be represented with p^{th} and q^{th} order AR models, as in [21], Chapter 8:

$$s(n) = -\sum_{i=1}^{p} a_i s(n-i) + w(n), \tag{3}$$

$$v(n) = -\sum_{k=1}^{q} b_j v(n-j) + u(n), \tag{4}$$

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 variances ${\sigma_w}^2$ and ${\sigma_u}^2$, respectively.

The state-vector s(n) corresponding to clean speech samples $s(n)$ is represented as:

$$\boldsymbol{s}(n) = \begin{bmatrix} s(n) \\ s(n-1) \\ s(n-2) \\ \vdots \\ s(n-p+1) \end{bmatrix}. \tag{5}$$

The state transition matrix $\boldsymbol{\Phi}_s$ of s(n) is given by:

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

Equations (5)-(6) are used to form the state-space model (SSM) of clean speech as:

$$\boldsymbol{s}(n) \quad = \boldsymbol{\Phi}_s \boldsymbol{s}(n-1) + \boldsymbol{d}_s w(n), where\, \boldsymbol{d}_s = [1 \quad 0 \quad \ldots \quad 0]^\top. \tag{7}$$

where $d_s = [1 \quad 0 \quad \ldots 0]^\perp$.

The additive noise state-vector v(n) and the corresponding state transition matrix $\boldsymbol{\Phi}_v$ are given by:

$$\boldsymbol{v}(n) = \begin{bmatrix} v(n) \\ v(n-1) \\ v(n-2) \\ \vdots \\ v(n-q+1) \end{bmatrix}, \tag{8}$$

$$\boldsymbol{\Phi}_v = \begin{bmatrix} -b_1 & -b_2 & \ldots & -b_{q-1} & -b_q \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}. \tag{9}$$

Equations (8)-(9) are used to form the SSM of additive noise as:

$$\boldsymbol{v}(n) \quad = \boldsymbol{\Phi}_v \boldsymbol{v}(n-1) + \boldsymbol{d}_v u(n), \tag{10}$$

where $d_v = [1 \quad 0 \quad \ldots 0]^\perp$.

The SSMs of the speech and additive noise can be combined into augmented matrix form as:

$$\begin{bmatrix} \boldsymbol{s}(n) \\ \boldsymbol{v}(n) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}_s & 0 \\ 0 & \boldsymbol{\Phi}_v \end{bmatrix} \begin{bmatrix} \boldsymbol{s}(n-1) \\ \boldsymbol{v}(n-1) \end{bmatrix} + \begin{bmatrix} \boldsymbol{d}_s & 0 \\ 0 & \boldsymbol{d}_v \end{bmatrix} \begin{bmatrix} w(n) \\ u(n) \end{bmatrix}. \tag{11}$$

By replacing $X^{(n)} \begin{bmatrix} \boldsymbol{s}(n) \\ \boldsymbol{v}(n) \end{bmatrix}, \Phi = \begin{bmatrix} \boldsymbol{\Phi}_s & 0 \\ 0 & \boldsymbol{\Phi}_v \end{bmatrix} d = \begin{bmatrix} d_s & 0 \\ 0 & d_v \end{bmatrix}$, $and\, z^{(n)} = \begin{bmatrix} w(n) \\ u(n) \end{bmatrix}$ can be written as:

$$\boldsymbol{x}(n) \quad = \boldsymbol{\Phi} \boldsymbol{x}(n-1) + \boldsymbol{d} \boldsymbol{z}(n). \tag{12}$$

Whereas the noisy observation $y(n)$ in eq. (2) can be represented in augmented matrix form as:

$$y(n) = \begin{bmatrix} \boldsymbol{c}_s^\top & \boldsymbol{c}_v^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{s}(n) \\ \boldsymbol{v}(n) \end{bmatrix}, \tag{13}$$

where $C^s = [1 \quad 0 \quad \ldots 0]^T$ and $C^V = [1 \quad 0 \quad \ldots 0]^T$ are the $p \times 1$ and $q \times 1$ vectors, respectively.

By replacing $C^\top = \begin{bmatrix} C_s^\top & C_v^\top \end{bmatrix}^T$, eq. (13) becomes:

$$y(n) = \boldsymbol{c}^\top \mathbf{x}(n). \tag{14}$$

Equations (12) and (14) together form the augmented SSM (ASSM) of AKF. For each noisy speech frame, the AKF computes an unbiased linear MMSE estimate, $\hat{\mathrm{x}}(n|n)$ at sample n, given the observed noisy speech, $y(n)$ by using the following recursive equations [12]:

$$\hat{\boldsymbol{x}}(n|n-1) = \boldsymbol{\Phi} \hat{\boldsymbol{x}}(n-1|n-1), \tag{15}$$

$$\boldsymbol{\Psi}(n|n-1) = \boldsymbol{\Phi} \boldsymbol{\Psi}(n-1|n-1) \boldsymbol{\Phi}^\top + \mathbf{Q} \mathbf{d} \mathbf{d}^\top, \tag{16}$$

$$\boldsymbol{K}(n) = \boldsymbol{\Psi}(n|n-1) \boldsymbol{c} (\boldsymbol{c}^\top \boldsymbol{\Psi}(\mathrm{n}|\mathrm{n}-1) \mathbf{c})^{-1}, \tag{17}$$

$$\hat{\boldsymbol{x}}(n|n) = \hat{\boldsymbol{x}}(n|n-1) + \boldsymbol{K}(n)[y(n) - \boldsymbol{c}^\top \hat{\mathbf{x}}(\mathrm{n}|\mathrm{n}-1)], \tag{18}$$

$$\boldsymbol{\Psi}(n|n) = [\boldsymbol{I} - \boldsymbol{K}(n) \boldsymbol{c}^\top] \boldsymbol{\Psi}(\mathrm{n}|\mathrm{n}-1), \tag{19}$$

where the process noise covariance matrix Q is given by:

$$\boldsymbol{Q} = \begin{bmatrix} \sigma_w^2 & 0 \\ 0 & \sigma_u^2 \end{bmatrix}. \tag{20}$$

For a noisy speech frame, the error covariances ($\boldsymbol{\Psi}(n|n-1)$ and $\boldsymbol{\Psi}(n|n)$ corresponding to $\hat{\mathrm{x}}(n|n-1)$ and $\hat{\mathrm{x}}(n|n)$) and the Kalman gain, K(n) are continually updated on a sample-by-sample basis, while ($\{a_i\}$, ${\sigma_w}^2$) and ($\{b_k\}$, ${\sigma_u}^2$) remain unchanged. Once all noisy speech frames for a given utterance being processed, synthesis over the enhanced frames gives the enhanced speech, $\hat{s}(n)$.

2.3. Paradigm shift of AKF recursive equations

The paradigm shift of recursive equations (15)-(19) transform them in scalar form. It exploits the understanding of the AKF operation in speech enhancement context. For this purpose, at sample n, we first extract the estimated speech, $\hat{s}(n|n)$ (the output of the AKF) as:

$g^\top \hat{\mathrm{x}}(n|n)$, where $g = [1 \quad 0 \quad 0 \quad \ldots 0]^\perp$ column vector. $g^\top \hat{\mathrm{x}}(n|n)$ is simplified as [19]:

$$\boldsymbol{g}^\top \hat{\mathbf{x}}(\mathrm{n}|\mathrm{n}) = [1 \quad 0 \quad 0 \quad \ldots \quad 0] \begin{bmatrix} \hat{\mathrm{s}}(\mathrm{n}|\mathrm{n}) \\ \hat{\mathrm{s}}(\mathrm{n}|\mathrm{n}-1) \\ \vdots \\ \hat{\mathrm{s}}(\mathrm{n}|\mathrm{n}-\mathrm{p}+1) \\ \hat{\mathrm{v}}(\mathrm{n}|\mathrm{n}) \\ \hat{\mathrm{v}}(\mathrm{n}|\mathrm{n}-1) \\ \vdots \\ \hat{\mathrm{v}}(\mathrm{n}|\mathrm{n}-\mathrm{q}+1) \end{bmatrix}. \tag{21}$$

The matrix multiplication in eq.(21) gives:

$$\boldsymbol{g}^\top \hat{\mathbf{x}}(\mathrm{n}|\mathrm{n}) = \hat{\mathrm{s}}(\mathrm{n}|\mathrm{n}). \tag{22}$$

By multiplying $g^\top$ on both side of eq. (18) gives:

$$\boldsymbol{g}^\top \hat{\mathbf{x}}(\mathrm{n}|\mathrm{n}) \quad = \boldsymbol{g}^\top \hat{\mathbf{x}}(\mathrm{n}|\mathrm{n}-1) + \boldsymbol{g}^\top \mathbf{K}(\mathrm{n})[\mathrm{y}(\mathrm{n}) - \boldsymbol{c}^\top \hat{\mathbf{x}}(\mathrm{n}|\mathrm{n}-1)]. \tag{23}$$

According to eq. (22), $g^{\top}\hat{x}(n|n-1)$ is given by:

$$\boldsymbol{g}^{\top}\hat{\mathbf{x}}(\mathrm{n}|\mathrm{n}-1)=\hat{\mathrm{s}}(\mathrm{n}|\mathrm{n}-1). \tag{24}$$

Also $c^{\top}\hat{x}(n|n-1)$ is re-written as:

$$\begin{aligned}\boldsymbol{c}^{\top}\hat{\mathbf{x}}(\mathrm{n}|\mathrm{n}-1)&=\begin{bmatrix}\mathbf{c}_{\mathrm{s}}^{\top} & \mathbf{c}_{\mathrm{v}}^{\top}\end{bmatrix}\begin{bmatrix}\hat{\mathrm{s}}(\mathrm{n}|\mathrm{n}-1)\\ \hat{\mathrm{s}}(\mathrm{n}|\mathrm{n}-2)\\ \vdots\\ \hat{\mathrm{s}}(\mathrm{n}|\mathrm{n}-\mathrm{p}+1)\\ \hat{\mathrm{v}}(\mathrm{n}|\mathrm{n}-1)\\ \hat{\mathrm{v}}(\mathrm{n}|\mathrm{n}-2)\\ \vdots\\ \hat{\mathrm{v}}(\mathrm{n}|\mathrm{n}-\mathrm{q}+1)\end{bmatrix},\\ &=\begin{bmatrix}1 & 0 & \dots & 0 & 1 & 0 & \dots & 0\end{bmatrix}\begin{bmatrix}\hat{s}(n|n-1)\\ \hat{s}(n|n-2)\\ \vdots\\ \hat{s}(n|n-p+1)\\ \hat{v}(n|n-1)\\ \hat{v}(n|n-2)\\ \vdots\\ \hat{v}(n|n-q+1)\end{bmatrix}.\end{aligned} \tag{25}$$

The matrix multiplication in eq. (25) gives:

$$\boldsymbol{c}^{\top}\hat{\mathbf{x}}(\mathrm{n}|\mathrm{n}-1)=\hat{\mathrm{s}}(\mathrm{n}|\mathrm{n}-1)+\hat{\mathrm{v}}(\mathrm{n}|\mathrm{n}-1). \tag{26}$$

In eq. (23), $g^{\top}K(n)$ gives the first component, $K_0(n)$ of Kalman gain vector, $K(n)$, which is written as:

$$K_0(n)\quad=\boldsymbol{g}^{\top}\mathbf{K}(\mathrm{n}). \tag{27}$$

Substituting eq. (17) into eq. (27) gives:

$$K_0(n)\quad=\frac{\boldsymbol{g}^{\top}\boldsymbol{\Psi}(\mathrm{n}|\mathrm{n}-1)\mathbf{c}}{\mathbf{c}^{\top}\boldsymbol{\Psi}(\mathrm{n}|\mathrm{n}-1)\mathbf{c}}. \tag{28}$$

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

$$\begin{aligned}\boldsymbol{c}^{\top}\boldsymbol{\Psi}(\mathrm{n}|\mathrm{n}-1)\mathbf{c}&=\mathbf{c}^{\top}\boldsymbol{\Phi}\boldsymbol{\Psi}(\mathrm{n}-1|\mathrm{n}-1)\boldsymbol{\Phi}^{\top}\mathbf{c}\\ &\quad+\boldsymbol{c}^{\top}\mathbf{Q}\mathbf{d}\mathbf{d}^{\top}\mathbf{c}.\end{aligned} \tag{29}$$

$Qdd^{\top}$ in the second term of eq. (29) is written as:

$$\begin{aligned}\boldsymbol{Qdd}^{\top}&=\begin{bmatrix}\mathbf{d}_{\mathrm{s}} & 0\\ 0 & \mathbf{d}_{\mathrm{v}}\end{bmatrix}\begin{bmatrix}\sigma_{\mathrm{w}}^2 & 0\\ 0 & \sigma_{\mathrm{u}}^2\end{bmatrix}\begin{bmatrix}\mathbf{d}_{\mathrm{s}}^{\top} & 0\\ 0 & \mathbf{d}_{\mathrm{v}}^{\top}\end{bmatrix},\\ &=\begin{bmatrix}\boldsymbol{d}_s\sigma_w^2 & 0\\ 0 & \boldsymbol{d}_v\sigma_u^2\end{bmatrix}\begin{bmatrix}\boldsymbol{d}_s^{\top} & 0\\ 0 & \boldsymbol{d}_v^{\top}\end{bmatrix},\\ &=\begin{bmatrix}\sigma_w^2\boldsymbol{d}_s\boldsymbol{d}_s^{\top} & 0\\ 0 & \sigma_u^2\boldsymbol{d}_v\boldsymbol{d}_v^{\top}\end{bmatrix},\\ &=\begin{bmatrix}\sigma_w^2 & 0\\ 0 & \sigma_u^2\end{bmatrix}.\end{aligned} \tag{30}$$

Now, $c^{\top}Qdd^{\top}c$ is simplified as:

$$\begin{aligned}\boldsymbol{c}^{\top}\mathbf{Q}\mathbf{d}\mathbf{d}^{\top}\mathbf{c}&=\begin{bmatrix}\mathbf{c}_{\mathrm{s}}^{\top} & \mathbf{c}_{\mathrm{v}}^{\top}\end{bmatrix}\begin{bmatrix}\sigma_{\mathrm{w}}^2 & 0\\ 0 & \sigma_{\mathrm{u}}^2\end{bmatrix}\begin{bmatrix}\mathbf{c}_{\mathrm{s}}\\ \mathbf{c}_{\mathrm{v}}\end{bmatrix},\\ &=\begin{bmatrix}\mathbf{c}_s^{\top} & \boldsymbol{c}_v^{\top}\end{bmatrix}\begin{bmatrix}\sigma_w^2\boldsymbol{c}_s\\ \sigma_u^2\boldsymbol{c}_v\end{bmatrix},\\ &=\sigma_w^2\boldsymbol{c}_s^{\top}\mathbf{c}_{\mathrm{s}}+\sigma_{\mathrm{u}}^2\mathbf{c}_{\mathrm{v}}^{\top}\mathbf{c}_{\mathrm{v}},\\ &=\sigma_w^2+\sigma_u^2.\end{aligned} \tag{31}$$

In eq. (29), $c^{\top}\boldsymbol{\Phi}\boldsymbol{\Psi}(n-1|n-1)\boldsymbol{\Phi}^{\top}c$ is written as:

$$\begin{aligned}\boldsymbol{c}^{\top}\boldsymbol{\Phi}\boldsymbol{\Psi}(\mathrm{n}-1|\mathrm{n}-1)\boldsymbol{\Phi}^{\top}\mathbf{c}&=\begin{bmatrix}\mathbf{c}_{\mathrm{s}}^{\top} & \mathbf{c}_{\mathrm{v}}^{\top}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Phi}_{\mathrm{s}} & 0\\ 0 & \boldsymbol{\Phi}_{\mathrm{v}}\end{bmatrix}\\ &\begin{bmatrix}\boldsymbol{\Psi}_s(n-1|n-1) & 0\\ 0 & \boldsymbol{\Psi}_v(n-1|n-1)\end{bmatrix}\begin{bmatrix}\boldsymbol{\Phi}_s^{\top} & 0\\ 0 & \boldsymbol{\Phi}_v^{\top}\end{bmatrix}\begin{bmatrix}\boldsymbol{c}_s\\ \boldsymbol{c}_v\end{bmatrix},\\ &=\begin{bmatrix}\boldsymbol{c}_s^{\top}\boldsymbol{\Phi}_{\mathrm{s}} & \boldsymbol{c}_v^{\top}\boldsymbol{\Phi}_{\mathrm{v}}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Psi}_s(n-1|n-1) & 0\\ 0 & \boldsymbol{\Psi}_v(n-1|n-1)\end{bmatrix}\\ &\quad\begin{bmatrix}\boldsymbol{\Phi}_s^{\top}\mathbf{c}_{\mathrm{s}}\\ \boldsymbol{\Phi}_v^{\top}\mathbf{c}_{\mathrm{v}}\end{bmatrix},a\\ &=\begin{bmatrix}\boldsymbol{c}_s^{\top}\boldsymbol{\Phi}_{\mathrm{s}}\boldsymbol{\Psi}_{\mathrm{s}}(\mathrm{n}-1|\mathrm{n}-1) & \boldsymbol{c}_v^{\top}\boldsymbol{\Phi}_{\mathrm{v}}\boldsymbol{\Psi}_{\mathrm{v}}(\mathrm{n}-1|\mathrm{n}-1)\end{bmatrix}\\ &\quad\begin{bmatrix}\boldsymbol{\Phi}_s^{\top}\mathbf{c}_{\mathrm{s}}\\ \boldsymbol{\Phi}_v^{\top}\mathbf{c}_{\mathrm{v}}\end{bmatrix},\\ &=\mathbf{c}_s^{\top}\boldsymbol{\Phi}_{\mathrm{s}}\boldsymbol{\Psi}_{\mathrm{s}}(\mathrm{n}-1|\mathrm{n}-1)\boldsymbol{\Phi}_{\mathrm{s}}^{\top}\mathbf{c}_{\mathrm{s}}+\\ &\quad\boldsymbol{c}_v^{\top}\boldsymbol{\Phi}_{\mathrm{v}}\boldsymbol{\Psi}_{\mathrm{v}}(\mathrm{n}-1|\mathrm{n}-1)\boldsymbol{\Phi}_{\mathrm{v}}^{\top}\mathbf{c}_{\mathrm{v}},\\ &=\alpha^2(n)+\beta^2(n).\end{aligned} \tag{32}$$

where $\alpha^2(n)$ and $\beta^2(n)$ represents the transmission of *a posteriori* error variance of the speech and noise from the previous time sample, given by [19]:

$$\alpha^2(n)=\boldsymbol{c}_s^{\top}\boldsymbol{\Phi}_{\mathrm{s}}\boldsymbol{\Psi}_{\mathrm{s}}(\mathrm{n}-1|\mathrm{n}-1)\boldsymbol{\Phi}_{\mathrm{s}}^{\top}\mathbf{c}_{\mathrm{s}}, \tag{33}$$

$$\beta^2(n)=\boldsymbol{c}_v^{\top}\boldsymbol{\Phi}_{\mathrm{v}}\boldsymbol{\Psi}_{\mathrm{v}}(\mathrm{n}-1|\mathrm{n}-1)\boldsymbol{\Phi}_{\mathrm{v}}^{\top}\mathbf{c}_{\mathrm{v}} \tag{34}$$

In equations (33)-(34), $\boldsymbol{\Psi}_s(n-1|n-1)$ and $\boldsymbol{\Psi}_v(n-1|n-1)$ represent the error covariance matrices of the *a priori* state estimates, $\hat{x}(n|n-1)$ and $\hat{v}(n|n-1)$, form $\boldsymbol{\Psi}$ as:

$$\boldsymbol{\Psi}=\begin{bmatrix}\boldsymbol{\Psi}_s & 0\\ 0 & \boldsymbol{\Psi}_v\end{bmatrix}. \tag{35}$$

By substituting equations (31)-(32) into eq. (29) gives:

$$\boldsymbol{c}^{\top}\boldsymbol{\Psi}(\mathrm{n}|\mathrm{n}-1)\mathbf{c}=\alpha^2(\mathrm{n})+\beta^2(\mathrm{n})+\sigma_{\mathrm{w}}^2+\sigma_{\mathrm{u}}^2. \tag{36}$$

Now, $g^{\top}\boldsymbol{\Psi}(n|n-1)c$ in eq. (28) can be expressed as:

$$\begin{aligned}\boldsymbol{g}^{\top}\boldsymbol{\Psi}(\mathrm{n}|\mathrm{n}-1)\mathbf{c}&=\mathbf{g}^{\top}\boldsymbol{\Phi}\boldsymbol{\Psi}(\mathrm{n}-1|\mathrm{n}-1)\boldsymbol{\Phi}^{\top}\mathbf{c}\\ &\quad+\boldsymbol{g}^{\top}\mathbf{Q}\mathbf{d}\mathbf{d}^{\top}\mathbf{c}.\end{aligned} \tag{37}$$

Using the expression similar to the derivation of eq. (32), it can be shown that:

$$\begin{aligned}\boldsymbol{g}^{\top}\boldsymbol{\Phi}\boldsymbol{\Psi}(\mathrm{n}-1|\mathrm{n}-1)\boldsymbol{\Phi}^{\top}\mathbf{c}&=\mathbf{g}^{\top}\boldsymbol{\Phi}\boldsymbol{\Psi}(\mathrm{n}-1|\mathrm{n}-1)\boldsymbol{\Phi}^{\top}\mathbf{g},\\ &=\boldsymbol{c}_s^{\top}\boldsymbol{\Phi}_{\mathrm{s}}\boldsymbol{\Psi}_{\mathrm{s}}(\mathrm{n}-1|\mathrm{n}-1)\boldsymbol{\Phi}_{\mathrm{s}}^{\top}\mathbf{c}_{\mathrm{s}}=\alpha^2(\mathrm{n}).\end{aligned} \tag{38}$$

Also, using the similar derivation in eq. (31) gives:

$$\boldsymbol{g}^{\top}\mathbf{Q}\mathbf{d}\mathbf{d}^{\top}\mathbf{c}=\mathbf{g}^{\top}\mathbf{Q}\mathbf{d}\mathbf{d}^{\top}\mathbf{g}=\sigma_{\mathrm{w}}^2. \tag{39}$$

Substituting equations (38)-(39) into eq. (37) yields:

$$\begin{aligned}\boldsymbol{g}^{\top}\boldsymbol{\Psi}(\mathrm{n}|\mathrm{n}-1)\mathbf{c}&=\mathbf{g}^{\top}\boldsymbol{\Psi}(\mathrm{n}|\mathrm{n}-1)\mathbf{g}=\mathbf{c}^{\top}\boldsymbol{\Psi}(\mathrm{n}|\mathrm{n}-1)\mathbf{g}\\ &=\alpha^2(n)+\sigma_w^2.\end{aligned} \tag{40}$$

Substituting equations (40) and (36) into eq. (28) gives:

$$K_0(n)=\frac{\alpha^2(n)+\sigma_w^2}{\alpha^2(n)+\beta^2(n)+\sigma_w^2+\sigma_u^2}. \tag{41}$$

Substituting equations (22), (24), (26)-(27) into eq. (23) yields:

$$\begin{aligned}\hat{s}(n|n)\quad&=\hat{s}(n|n-1)+K_0(n)y(n)-K_0(n)[\hat{s}(n|n-1)\\ &\quad+\hat{v}(n|n-1)],\\ &=[1-K_0(n)]\hat{s}(n|n-1)+K_0(n)[y(n)-\\ &\quad\hat{v}(n|n-1)].\end{aligned} \tag{42}$$

Equation (42) implies that the estimated speech at sample n, $\tilde{s}(n|n)$ is given by a sum of the predicted speech, $\hat{s}(n|n-1)$ and the measurement innovation, $[y(n) - \hat{v}(n|n-1)]$ weighted by the scalar Kalman gain, $K_0(n)$. Therefore, the temporal trajectory of $K_0(n)$ is a useful indicator of $\hat{s}(n|n)$ estimate. In practice, the inaccurate estimates of $(\{a_i\}, {\sigma_w}^2)$ and $(\{b_j\}, {\sigma_u}^2)$ introduce bias in $K_0(n)$, resulting in degraded $\hat{s}(n|n)$. Therefore, there should have a performance metric/index that can quantify the level of biasness in $K_0(n)$. George et al. introduced a robustness metric and a sensitivity metric, which can be used to *offset* the bias in $K_0(n)$ [19]. In AKF-based SEA [19, Section 3.2], the robustness and sensitivity metrics are defined by simplifying the mean squared error, $\mathbf{g}^\top \mathbf{\Psi}(n|n)\mathbf{g}$ of the AKF output, $\tilde{s}(n|n)$ as:

$$\begin{aligned}\mathbf{g}^\top\mathbf{\Psi}(\mathrm{n|n})\mathbf{g} &= \mathbf{g}^\top[\mathbf{I}-\mathbf{K}(\mathrm{n})\mathbf{c}^\top]\mathbf{\Psi}(\mathrm{n|n-1})\mathbf{g}, \text{[from [eq : 19]}\\ &= \mathbf{g}^\top\mathbf{\Psi}(\mathrm{n|n-1})\mathbf{g} - \mathbf{g}^\top\mathbf{K}(\mathrm{n})\mathbf{c}^\top\mathbf{\Psi}(\mathrm{n|n-1})\mathbf{g}, \\ &= \mathbf{g}^\top\mathbf{\Psi}(\mathrm{n|n-1})\mathbf{g} - \mathrm{K_0(n)}\mathbf{c}^\top\mathbf{\Psi}(\mathrm{n|n-1})\mathbf{g}.\end{aligned} \tag{43}$$

Substituting equations (40)-(41) into (43) gives:

$$\Psi_{0,0}(n|n) = \alpha^2(n) + \sigma_w^2 - \frac{[\alpha^2(n)+\sigma_w^2]^2}{\alpha^2(n)+\beta^2(n)+\sigma_w^2+\sigma_u^2},$$

$$\Psi_{0,0}(n|n) - \alpha^2(n) = \sigma_w^2 - \frac{[\alpha^2(n)+\sigma_w^2]^2}{\alpha^2(n)+\beta^2(n)+\sigma_w^2+\sigma_u^2},$$

$$\frac{\Psi_{0,0}(n|n)-\alpha^2(n)}{\alpha^2(n)+\sigma_w^2} = \frac{\sigma_w^2}{\alpha^2(n)+\sigma_w^2} - \frac{\alpha^2(n)+\sigma_w^2}{\alpha^2(n)+\beta^2(n)+\sigma_w^2+\sigma_u^2},$$

$$\frac{\Psi_{0,0}(n|n)-\alpha^2(n)}{\alpha^2(n)+\sigma_w^2} = \frac{\sigma_w^2}{\alpha^2(n)+\sigma_w^2} + \frac{\beta^2(n)+\sigma_u^2}{\alpha^2(n)+\beta^2(n)+\sigma_w^2+\sigma_u^2} - 1,$$

$$\begin{aligned}\frac{\Psi_{0,0}(n|n)-\alpha^2(n)}{\alpha^2(n)+\sigma_w^2} + 1 &= \frac{\sigma_w^2}{\alpha^2(n)+\sigma_w^2} + \frac{\beta^2(n)+\sigma_u^2}{\alpha^2(n)+\beta^2(n)+\sigma_w^2+\sigma_u^2}, \\ \Delta\Psi(n|n) + 1 &= J_2(n) + J_1(n),\end{aligned} \tag{44}$$

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

$$J_2(n) = \frac{\sigma_w^2}{\alpha^2(n)+\sigma_w^2}, \tag{45}$$

$$J_1(n) = \frac{\beta^2(n)+\sigma_u^2}{\alpha^2(n)+\beta^2(n)+\sigma_w^2+\sigma_u^2}. \tag{46}$$

In AKF-RMBT [19], a $J_2(n)$ metric-based tuning of $K_0(n)$ has been proposed for speech enhancement in colored noise conditions. In AKF-SMBT [20], an extension of AKF-RMBT [19] using a $J_1(n)$ metric-based tuning of $K_0(n)$ has been proposed. Section 2.4 demonstrates the shortcomings of AKF-RMBT and AKF-SMBT [19,20] in terms of biased interpretation of $K_0(n)$.

2.4. Impact of biased $K_0(n)$ on AKF-based speech enhancement in colored noise conditions

To analyze the shortcomings of AKF-RMBT and AKFSMBT [19,20], we conduct an experiment with the utterance sp27 ("*Bring your best compass to the third class*") of the NOIZEUS corpus [22], Chapter 12 (sampled at 8 kHz) corrupted with colored (*factory*) noise (taken from RSG-10 database [23]) at 5 dB SNR level. As like [19,20], $p = 10$ and $q = 40$ have been used in this analysis.

In oracle case, $(\{a_i\}, {\sigma_w}^2)$ and $(\{b_j\}, {\sigma_u}^2)$ are computed from the clean speech and the additive noise, respectively. During speech pauses of the observed noisy speech; since $s(n,l) = 0$, it gives $\{a_i\} = 0$, $\tilde{s}(n|n-1) = 0$, and $[\alpha^2(n) + {\sigma_w}^2] = 0$, which turns $K_0(n) = 0$ (according to eq. (41)). For example, it is shown that $K_0(n) = 0$ between 0 and 0.2 s or 2.2–2.52 s of Fig. 1 (d)). With $K_0(n) = 0$ and $\hat{s}(n|n-1) = 0$, eq. (42) implies that nothing is passed to the enhanced speech (i.e., $\tilde{s}(n|n) = 0$) (e.g., 0–0.2 s or 2.22.52 s in Fig. 1 (e)). Conversely, during speech presence of the noisy speech, it is observed that $K_0(n)$ approaching 1 (e.g., 0.2–0.6 s of Fig. 1 (d)). With $K_0(n) \approx 1$, the first part in eq. (42) approaching 0, while the predicted noise, $\hat{v}(n|n-1)$ subtracted from the observed noisy speech, $y(n)$ in the second part, termed as measurement innovation, $[y(n) - \hat{v}(n|n-1)]$ scaled by $K_0(n)$ almost retains the clean speech. As a result, the enhanced speech produced by AKF-Oracle (Fig. 1(e)) is almost identical to the clean speech (Fig. 1(a)).

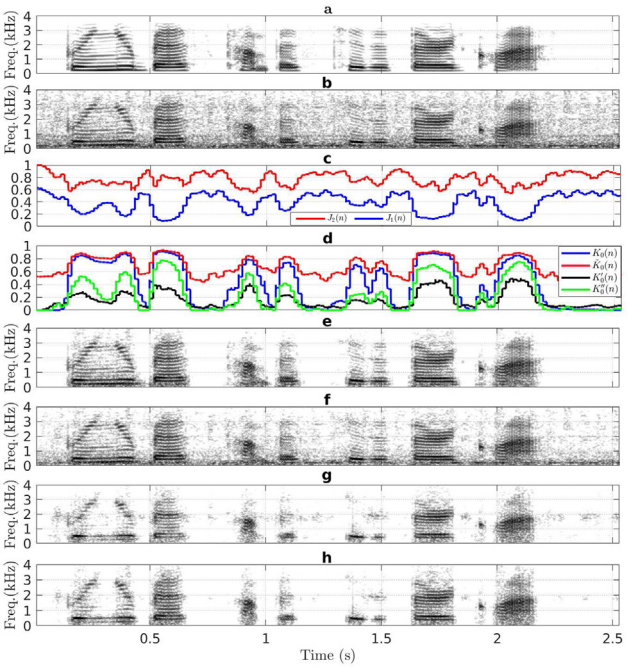

Fig. 1. Spectrograms of the: (a) clean speech (utterance sp27), (b) noisy speech (corrupt (a) with 5 dB *factory* noise), (c) $J_2(n)$ and $J_1(n)$ metrics, (d) oracle and non-oracle $K_0(n)$ with adjusted) and), spectrogram of enhanced speech produced by: (e) AKF-Oracle, (f) AKF-Non-oracle, (g) AKF-RMBT [19], and (h) AKF-SMBT [20] methods, respectively.

In non-oracle case, $(\{b_j\}, {\sigma_u}^2)$ are computed from the first noisy speech frame by assuming that there remains no speech. The computed $(\{b_j\}, {\sigma_u}^2)$ remains constant during processing all frames for a given utterance [19]. That means, the total *a priori* prediction error of the noise model, $[\beta^2(n) + {\sigma_u}^2]$ also remains constant for all noisy speech frames. Conversely, $(\{a_i\}, {\sigma_w}^2)$ computed from the noisy speech frames becomes biased, i.e., $(\{\tilde{a}_i\}, {\tilde{\sigma}_w}^2)$, which results in biased total *a priori* prediction error of the speech model, $[\tilde{\alpha}^2(n) + {\tilde{\sigma}_w}^2]$. Since the silent frames of the noisy speech are completely filled with noise, it gives $[\tilde{\alpha}^2(n) + {\tilde{\sigma}_w}^2] \approx [\beta^2(n) + {\sigma_u}^2]$. According to eq. (41), this condition introduces 0.5 bias in $\tilde{K}_0(n)$ (e.g., 0–0.2 s or 2.2–2.52 s of Fig. 1 (d)). With 0.5 biased $\tilde{K}_0(n)$, eq. (42) implies that 50% of the measurement innovation, i.e., $[y(n) - \hat{v}(n|n-1)]$ leaking into the enhanced speech, $\hat{s}(n|n)$ (e.g., 0.2–0.6 s of Fig. 1 (f)). $(\{\tilde{a}_i\}, {\tilde{\sigma}_w}^2)$ also produces biased $\tilde{K}_0(n)$ in speech regions. The biased $\tilde{K}_0(n)$ passes a significant *residual* noise to the enhanced speech, $\hat{s}(n|n)$ as shown in Fig. 1 (f)).

In AKF-RMBT [19], a $J_2(n)$ metric is used to *offset* the bias in $\tilde{K}_0(n)$ as:

$$K_0'(n) = \tilde{K}_0(n)[1 - J_2(n)]. \tag{47}$$

To perform the tuning of $K_{\sim 0}(n)$ using eq. (47), it requires $J_2(n) \approx 1$. However, it is shown in [19, Fig. 4 (d)] that the colored noise effect in ($\{a_{\sim i}\}$, $\sim\sigma_w^2$) changes the behaviour of $J_2(n)$ apart from approaching 1. To cope with this problem, George et al. employed a whitening filter, $H_w(z)$ to each noisy speech frame, yielding a pre-whitened speech, $y_w(n,l)$. With $\{\hat{b}_j\}$, $H_w(z)$ is constructed as [19]:

$$H_w(z) = 1 + \sum_{j=1}^{q} \hat{b}_j z^{-j}. \tag{48}$$

Now, ($\{a_i\}$,σ_w^2) are computed from $y_w(n,l)$ using the autocorrelation method [21], Chapter 8. It can be seen that the improved ($\{a_i\}$,σ_w^2) enables $J_2(n) \approx 1$ during speech pauses, resulting 0 as shown in Fig. 1 (c)-(d). However, the tuning process in eq. (47) results in a significantly reduced) in speech regions as compared to the oracle $K_0(n)$ (Fig. 1 (d)). Therefore,) causes over-suppression of the speech components, resulting in distorted speech as shown in Fig. 1 (g).

To address the problem in AKF-RMBT [19], a $J_1(n)$ metric-based tuning of $K_{\sim 0}(n)$ has been proposed in AKFSMBT [20] as:

$$K_0''(n) \quad = \tilde{K}_0(n) - J_1(n). \tag{49}$$

It can be seen that $\tilde{K}_0(n)$ and $J_1(n)$ approaching 0.5 during speech pauses (e.g., 0–0.2 s or 2.2–2.52 s of Fig. 1 (c)(d)). While $J_1(n) \approx 0$ in speech regions (e.g., 0.2–0.6 s of Fig. 1 (c)). Therefore, the subtraction of $J_1(n)$ from $\tilde{K}_0(n)$ (eq. (49)) results in $K_0''(n) \approx 0$ during speech pauses, while $K_0''(n) \approx 1$ in speech regions. It is shown in Fig. 1 (d) that under-estimation issue in the speech region is minimized in $K_0''(n)$ as compared to $K_0''(n)$. As a result, AKF-SMBT [20] produced less distorted speech (Fig. 1 (h)) than that of [19] (Fig. 1 (g)).

Technically, ($\{b_k\}$,σ_u^2) must be computed from each noisy speech frame in non-stationary noise conditions. Thus, ($\{b_k\}$,σ_u^2) computed from the first noisy speech frame in AKF-RMBT and AKF-SMBT [19,20] does not adequately address the non-stationary noise conditions. In addition, the whitening filter, $H_w(z)$ (eq. (48)) constructed with the constant $\{b_k\}$ in AKF-RMBT and AKF-SMBT [19] failed to reduce bias in the estimated ($\{a_i\}$, σ_w^2). In light of the observations, AKF-RMBT and AKF-SMBT [19,20] do not adequately address speech enhancement in non-stationary noise conditions. In Section 2.5, we further demonstrate the biasing impact of $K_0(n)$ on AKF-based speech enhancement in non-stationary noise conditions.

2.5. Impact of biased $K_0(n)$ on AKF-based speech enhancement in non-stationary noise conditions

To analyze the impact of biased $K_0(n)$ on AKF-based speech enhancement in non-stationary noise condition, we repeat the experiment in Fig. 1 except the utterance sp27 is corrupted with 5 dB *babble* noise (taken from AURORA database [24]). In this study, a 32 ms rectangular window with 50% overlap [25], Sec 7.2.1 was considered for converting $y(n)$ into frames, $y(n,k)$ (as in eq. (2)). We have also used $p = 16$ and $q = 40$.

As demonstrated in Section 2.4, in oracle case, the silent frames of $y(n,l)$ gives $s(n,l) = 0$ such that $a_i = 0$ for $i = 1,2,\ldots,p$, which turns $\hat{s}(n|n-1) = 0$ as well as $[\alpha^2(n) + \sigma_w^2] = 0$ (e.g., 0–0.2 s or 2.2–2.52 s of Fig. 2 (d)). Substituting $[\alpha^2(n) + \sigma_w^2] = 0$ in eq. (41) gives $K_0(n) = 0$, which in turn $\hat{s}(n|n) = 0$ (eq. (42)), i.e., nothing is passed to the enhanced speech (e.g., 0–0.2 s or 2.2–2.52 s of Fig. 2 (c) and (g)). Conversely, it gives $[\alpha^2(n) + \sigma_w^2] > [\beta^2(n) + \sigma_u^2]$ for speech dominated frames, resulting in $K_0(n) \approx 1$ (e.g., 0.2–0.6 s of Fig. 2 (c)). As demonstrated in Section 2.4, $K_0(n) \approx 1$ almost passes the clean speech to the output. Therefore, the enhanced speech produced by AKF-Oracle (Fig. 2(g)) is almost identical to the clean speech (Fig. 2(a)).

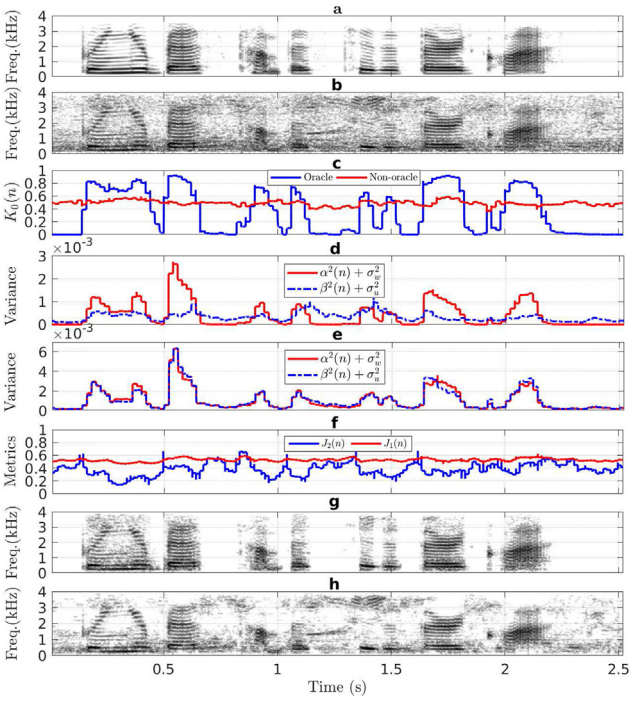

Fig. 2. Biasing effect demonstration of $K_0(n)$, spectrogram of: (a) clean speech (utterance sp27), (b) noisy speech (corrupt sp27 with 5 dB *babble* noise), (c) $K_0(n)$ computed in oracle and non-oracle cases, (d)-(e) $[\alpha^2(n) + \sigma_w^2]$ and $[\beta^2(n) + \sigma_u^2]$ computed in oracle and non-oracle cases, (f) $J_2(n)$ and $J_1(n)$ metrics computed from the noisy speech in (b), spectrogram of enhanced speech produced by: (g) AKF-Oracle method, and (h) AKF-Non-oracle method.

In non-oracle case, ($\{a_i\}$, σ_w^2) and ($\{b_j\}$, σ_u^2) are computed from each noisy speech frame, resulting in biased ($\{a_{\sim i}\}$, $\sim\sigma_w^2$) and ($\{_{\sim} b_j\}$, $\sim\sigma_u^2$), which in turn $[\alpha^2(n) + \sigma_w^2] \approx [\beta^2(n) + \sigma_u^2]$ (e.g., 0–0.2 s or 2.2–2.52 s of Fig. 2 (e)). According to eq. (41), this condition introduces around 0.5 bias in $K_0(n)$ (e.g., 0–0.2 s or 2.2–2.52 s of Fig. 2 (c)). During speech activity of $y(n,l)$, it is observed that $[\alpha^2(n) + \sigma_w^2] \geq [\beta^2(n) + \sigma_u^2]$, resulting in an under-estimated $K_0(n)$ as compared to the oracle case (e.g., 0.2–0.6 s of Fig. 2 (c)). 0.5 biased $K_0(n)$ in silent regions leaking 50% of $[y(n) - \hat{v}(n|n-1)]$ to the enhanced speech (Fig. 2 (h)). In addition, the under-estimated $K_0(n)$ in speech regions produced distorted speech (Fig. 2 (h)). Also, $J_2(n)$ and $J_1(n)$ metrics (Fig. 2 (f)) do not achieve the similar characteristics as found in the colored noise condition (Fig. 1 (c)), which leaves them inappropriate in tuning the biased $K_0(n)$ (Fig. 2 (c)) using equations. (47) and (49).

In light of the observations in this section, the objective of proposed SEA falls in twofold: firstly, to improve the estimates of ($\{a_i\}$, σ_w^2) and ($\{b_j\}$,σ_u^2) in real-life noise conditions so that $J_2(n)$ and $J_1(n)$ achieve the similar characteristics as found in colored noise conditions (Fig. 1 (c)). Secondly, incorporate both the improved $J_2(n)$ and $J_1(n)$ metrics for tuning the biased $K_0(n)$ to achieve better noise reduction by AKF–even for conditions that have time-varying amplitudes.

3. Proposed speech enhancement algorithm

Fig. 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 2.5. The next step of the proposed

SEA is ($\{a_i\}$,σ_w^2) and ($\{b_j\}$,σ_u^2) estimation as described in Section 3.1.

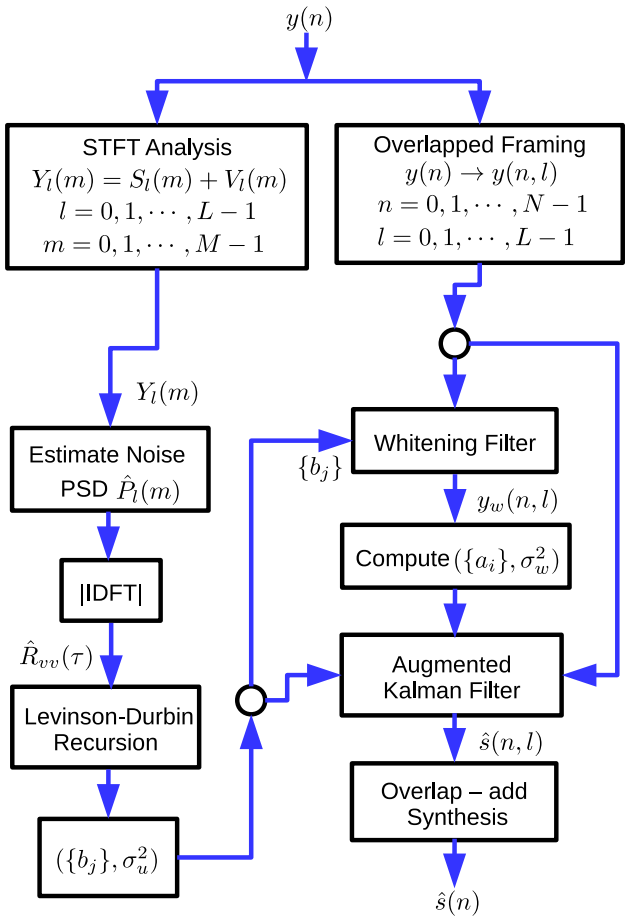

Fig. 3. Block diagram of the proposed AKF-based SEA.

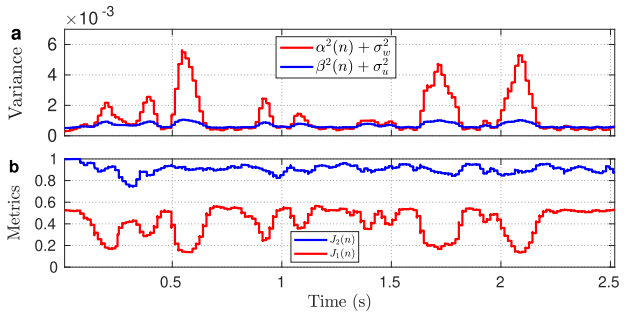

Fig. 4. Comparing the estimated: (a) $[\alpha^2(n) + \sigma_w{}^2]$ and $[\beta^2(n) + \sigma_u{}^2]$ and (b) $J_2(n)$, $J_1(n)$ metrics from the noisy speech in Fig. 2 (b).

3.1. Proposed ($\{a_i\}$,σ_w^2) and ($\{b_j\}$,σ_u^2) estimation method

The speech and noise LPC parameters, ($\{a_i\}$, $\sigma_w{}^2$) and ($\{b_j\}$, $\sigma_u{}^2$) are sensitive to noise. Since the clean speech, $s(n,l)$ and the noise, $v(n,l)$ are unobserved in practice, it is difficult to accurately estimate ($\{a_i\}$, $\sigma_w{}^2$) and ($\{b_j\}$, $\sigma_u{}^2$) from noisy speech, $y(n,l)$. It is already demonstrated that ($\{a_i\}$, $\sigma_w{}^2$) and ($\{b_j\}$, $\sigma_u{}^2$) estimates in AKF-RMBT and AKF-SMBT [19,20] do not address the conditions that have time varying amplitudes.

To cope with this problem, in this paper, we first estimate noise PSD, $\hat{P}_v(l,m)$ from each noisy speech frame using an SPP method [26] (described in section 3.2). Then employ inverse Fourier transform to $\hat{P}_v(l,m)$, yields an esτ is the autocorrelation lag. By solving $Rvvb(\tau)$ using the timate of the noise autocorrelation matrix, $R_{vv}(\tau)$, where Levinson-Durbin recursion [21], Chapter 8, gives (b $\{b_j\}$, $\sigma_u{}^2$) (q = 40). As in [19], to reduce bias in the estimated ($\{a_i\}$, $\sigma_w{}^2$) for each noisy speech frame, we compute them from the corresponding pre-whitened speech, $y_w(n,l)$ using the autocorrelation method [21], Chapter 8. The framewise $y_w(n,l)$ is obtained by employing a whitening filter, $H_w(z)$ to $y(n,l)$. With estimated $\{b_j\}$; $H_w(z)$ is constructed as in eq. (48). Unlike AKF-RMBT and AKF-SMBT [19,20], since $H_w(z)$ is constructed with $\{b_j\}$ for each noisy speech frame, the estimates of ($\{a_i\}$, $\sigma_w{}^2$) address conditions that have time-varying amplitudes.

3.2. Noise PSD estimation

In this paper, we incorporate an SPP method [26] to estimate noise PSD from each noisy speech frame. For this purpose, the noisy speech, $y(n)$ (eq. (1)) is next analyzed frame-wise using the short-time Fourier transform (STFT):

$$Y_l(m) = S_l(m) + V_l(m), \tag{50}$$

where $Y_l(m)$, $S_l(m)$, and $V_l(m)$ denote the complex-valued STFT coefficients of the noisy speech, clean speech, and noise signal, respectively, for time-frame index l and frequency bin $m_\varrho\{0,1,2,\ldots,M-1\}$ with M being the total number of frequency-bins within each frame.

A Hamming window with 50% overlap is used in STFT analysis [25], Section 7.2.1. In polar form, $Y_l(m)$, $S_l(m)$, and $V_l(m)$ can be expressed as: $Y_l(m) = R_l(m)e^{j\varphi l(m)}$, $Sl(m) = Al(m)ej\phi l(m)$, and $Vl(m) = Dl(m)ej\theta l(m)$, where $R_l(m)$, $A_l(m)$, and $D_l(m)$ are the magnitude spectrums of the noisy speech, the clean speech, and the noise signal, respectively, and $\varphi_l(m)$, $\phi_l(m)$, and $\theta_l(m)$ are the corresponding phase spectrums. We process each frequency bin of the single-sided noisy speech power spectrum, $R_l{}^2(m)$ (where m_ϱ $\{0,1,\ldots,128\}$ containing the DC and Nyquist frequency components) to estimate the noise power spectrum, $\hat{D}_l{}^2(m)$. To initialize the algorithm, we assume the first frame ($l = 0$) of $R_0^2(m)$ as silent, which gives an estimate of noise power as: $\hat{D}_0^2(m) = R_0^2(m)$. The noise PSD, $\lambda_0(m)$ is also initialized as: $\lambda_0(m) = \hat{D}_0^2(m)$. For $l \geq 1$, using the speech presence uncertainty principle, an MMSE estimate of $\hat{D}_l{}^2(m)$ at m^{th} frequency bin is given by [26]:

$$\hat{D}_l^2(m) \quad = P(H_0^m|R_l(m))R_l^2(m) + P(H_1^m|R_l(m))\hat{\lambda}_{l-1}(m), \tag{51}$$

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

The simplified)) estimate is given by[1] [26]:

$$P(H_1^m|R_l(m)) = [1 + (1 + \xi_{opt})\exp\{(-\frac{R_l^2(m)}{\hat{\lambda}_{l-1}(m)})(\frac{\xi_{opt}}{1+\xi_{opt}})\}]^{-1}, \tag{52}$$

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

The optimal choice for ξ_{opt} is found as $10\log_{10}(\xi_{opt}) = 15$ dB [26], and $P(H_0^m|R_l(m))$ is given by$P(H_0^m|R_l(m)) = 1 - P(H_1^m\}R_l(m))$. However, if $P(H_0^m|R_l(m)) = 1$ occurs at m^{th} frequency bin, it causes stagnation, which stops updating $\hat{D}_l{}^2(m)$ (eq. (51)). Unlike monitoring the status of $P(H_1^m|R_l(m))) = 1$ for a long time as reported in [26], we simply resolve this issue by setting $P(H_1^m|R_l(m)) = 0.99$once this condition occurs prior to update $\hat{D}_l{}^2(m)$.

It is observed that $R_l{}^2(m)$ is completely filled with additive noise during silent activity, thus giving an estimate of noise power. Therefore, unlike updating $\hat{D}_l{}^2(m)$ using eq. (51) by existing method [26], we do it differently depending on the silent/speech activity of $R_l{}^2(m)$ (for each frequency bin m). Specifically, at m^{th} frequency bin ($l \geq 1$), if$P(H_1^m|R_l(m)) < 0.5, R_l^2(m)$ yields silent activity, resulting in $\hat{D}_l{}^2(m) = R_l{}^2(m)$, otherwise, $\hat{D}_l{}^2(m)$ is estimated using eq. (51). With estimated $\hat{D}_l{}^2(m)$, $\lambda_l(m)$ is updated as:

[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: $P(H_0) = P(H_1)$.

$$\hat{\lambda}_l(m) = \eta\hat{\lambda}_{l-1}(m) + (1-\eta)\hat{D}_l^2(m), \tag{53}$$

where the smoothing constant, η is set to 0.9.

The 256-point noise PSD is given as: $\hat{P}_v(l,m) = \lambda_l(m)$, where the components of $\hat{P}_v(l,m)$ at $m_\varrho\{1,2,\ldots,127\}$ are flipped to that of the $m_\varrho\{129,130,\ldots,255\}$ of $\hat{P}_v(l,m)$.

3.3. Proposed $K_0(n)$ tuning method

Firstly, the AKF is constructed with the estimated $(\{a_i\},\sigma_w^2)$ and $(\{b_j\},\sigma_u^2)$. Then we extract the tuning parameters as shown in Fig. 4. It can be seen from Fig. 4 (a) that $[\alpha^2(n)+\sigma_w^2]$ and $[\beta^2(n)+\sigma_u^2]$ achieves very similar characteristics as like AKF-Oracle method (Fig. 2 (d)). The improvement of these parameters also causes $J_2(n)$ and $J_1(n)$ metrics (Fig. 4 (b)) to achieve the similar characteristics as appear in the colored noise condition (Fig. 1 (c)). Therefore, $J_2(n)$ and $J_1(n)$ metrics (Fig. 4 (b)) are now eligible to dynamically *offset* the bias in $K_0(n)$—even for non-stationary noise conditions. However, as demonstrated in Section 2.4, $J_2(n)$ metric is useful in tuning $K_0(n)$ during speech pauses of the noisy speech, since it results in under-estimated $K_0(n)$ during speech presence. On the contrary, since $J_1(n)$ metric approaches 0 in speech regions of the noisy speech, according to eq. (49), it minimizes the under-estimation of $K_0(n)$. In light of the observations, for each sample of $y(n,l)$, we incorporate $J_2(n)$ metric during speech pauses and $J_1(n)$ metric during speech presence to dynamically *offset* the bias in $K_0(n)$.

We observed that the total *a priori* prediction errors of the speech and noise AR models; $[\alpha^2(n)+\sigma_w^2]$ and $[\beta^2(n)+\sigma_u^2]$ can be adopted as a speech activity detector for each sample of $y(n,l)$. For example, during speech pauses, the condition $[\beta^2(n)+\sigma_u^2] \geq [\alpha^2(n)+\sigma_w^2]$ holds (e.g., 0–0.2 s or 2.2–2.52 s of Fig. 4 (a)). Conversely, $[\alpha^2(n)+\sigma_w^2] \gg [\beta^2(n)+\sigma_u^2]$ is found in speech regions (e.g., 0.2–0.6 s of Fig. 4 (a)). Therefore, at sample n, if $[\beta^2(n)+\sigma_u^2] \geq [\alpha^2(n)+\sigma_w^2]$; $y(n,k)$ is termed as silent and set the decision parameter (denoted by ζ) as $\zeta(n)=0$; otherwise speech activity occurs and $\zeta(n)=1$. It can be seen from Fig. 5 that the detected flags (0/1: silent/speech) by proposed method is closely similar to that of the reference (0/-1: silent/speech). The reference flags are generated by visually inspecting the corresponding clean speech (Fig. 2 (a)) frames.

At sample n, if $\zeta(n)=0$, the adjusted $K_0'(n)$ in the proposed SEA is given by:

$$\begin{aligned} K_0'(n) &= K_0(n)[1-J_2(n)], \\ &= \left[\frac{\alpha^2(n)+\sigma_w^2}{\alpha^2(n)+\beta^2(n)+\sigma_w^2+\sigma_u^2}\right]\left[1-\frac{\sigma_w^2}{\alpha^2(n)+\sigma_w^2}\right], \\ &= \frac{\alpha^2(n)}{\alpha^2(n)+\beta^2(n)+\sigma_w^2+\sigma_u^2}. \end{aligned} \tag{54}$$

To justify the validity of $K_0'(n)$, Fig. 6 (a) shows the numerator and the denominator of eq. (54) computed from the noisy speech in Fig. 2 (b). It can be seen that $\alpha^2(n) \approx 0$ during speech pauses (e.g., 0–0.2 s or 2.2–2.52 s of Fig. 6 (a)). According to eq. (54), it results $K_0'(n) \approx 0$. Since $[\alpha^2(n)+\beta^2(n)+\sigma_w^2+\sigma_u^2] \gg \alpha^2(n)$ occurs during speech presence (e.g., 0.2–0.6 s of Fig. 6 (a)), it may result in under-estimated $K_0'(n)$ as like colored noise experiment (Fig. 1 (d)). Thus, $J_2(n)$ metric-based tuning of $K_0'(n)$ in speech activity of $y(n,l)$ is inappropriate.

As discussed earlier, we employ $J_1(n)$ metric to *offset* the bias in $K_0(n)$ during speech activity of $y(n,l)$.

However, our further investigation on $J_1(n)$ metric-based tuning in eq. (49) reveals that the subtraction of $J_1(n)$ from biased $K_0(n)$ still produced under-estimated as shown in Fig. 1 (d). To cope with this problem, at sample n, if $\zeta(n)=1$, we propose the tuning of biased $K_0(n)$ using $J_1(n)$ metric as:

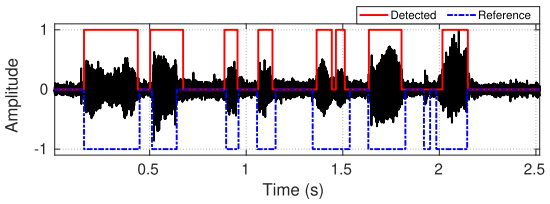

Fig. 5. Comparing the detected flags from noisy speech in Fig. 2 (b) to that of the reference corresponding to Fig. 2 (a).

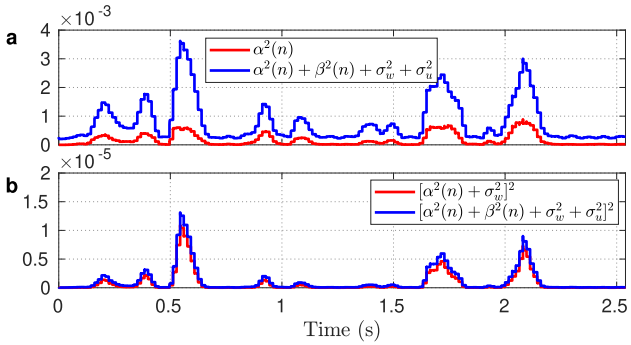

Fig. 6. $K_0'(n)$ responses in terms of: (a) $\alpha^2(n)$ and $\alpha^2(n)+\beta^2(n)+\sigma_w^2+\sigma_u^2$, and (b) $[\alpha^2(n)+\sigma_w^2]^2$ and $[\alpha^2(n)+\beta^2(n)+\sigma_w^2+\sigma_u^2]^2$, where the same experimental setup of Fig. 2 is used.

$$\begin{aligned} K_0'(n) &= K_0(n)[1-J_1(n)], \\ &= \left[\frac{\alpha^2(n)+\sigma_w^2}{\alpha^2(n)+\beta^2(n)+\sigma_w^2+\sigma_u^2}\right] \\ &\quad \left[\frac{\alpha^2(n)+\sigma_w^2}{\alpha^2(n)+\beta^2(n)+\sigma_w^2+\sigma_u^2}\right], \\ &= \frac{[\alpha^2(n)+\sigma_w^2]^2}{[\alpha^2(n)+\beta^2(n)+\sigma_w^2+\sigma_u^2]^2}. \end{aligned} \tag{55}$$

To justify the validity of), the numerator and the denominator of eq. (55) are shown in Fig. 6 (b). It can be seen that $[\alpha^2(n)+\beta^2(n)+\sigma_w^2+\sigma_u^2]^2 \geq [\alpha^2(n)+\sigma_w^2]^2$ during the speech presence of $y(n,l)$ (e.g., 0.2–0.6 s of Fig. 6 (b)), which results in) approaching 1.

To evaluate the performance of the proposed tuning method in non-stationary noise conditions, we conduct an experiment with the same setup as in Fig. 2. It can be seen from Fig. 7 (a) that the adjusted $K_0(n)$ by the proposed method shows significantly less bias and closely similar to that of 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 that of the oracle $K_0(n)$. Amongst the benchmark methods, the adjusted $K_0(n)$ by AKF-SMBT [20] shows less bias than that of the AKF-RMBT [19]. However, AKF-SMBT [20] still produces under-estimated $K_0(n)$ in speech regions. We also repeat the Fig. 7 (a) experiment except the utterance sp27 is corrupted with 5 dB *factory* noise to evaluate the performance of the proposed tuning method in colored noise conditions. Fig. 7 (b) reveals that the biasing effect is reduced significantly in the adjusted $K_0(n)$ by the proposed method, which is closely similar to that of the oracle $K_0(n)$. As in the previous experiment, AKF-SMBT [20] also produce under-estimated $K_0(n)$ in speech regions. The AKF-RMBT method [20] produced the most underestimated $K_0(n)$ amongst the competing methods. In light of the comparative study, the reduced-biased $K_0(n)$ achieved by

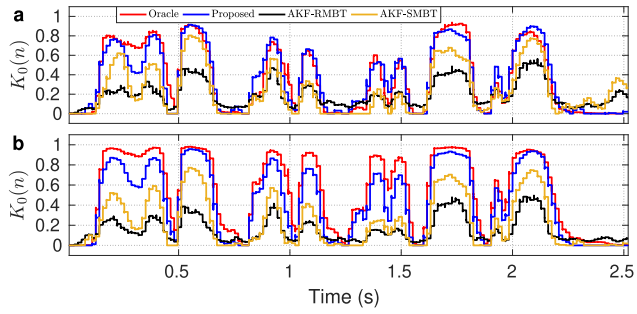

Fig. 7. Comparing $K_0(n)$ trajectories corresponding to the AKF-Oracle method, proposed method , AKF-RMBT method [19], and AKF-SMBT method [20], where the utterance sp27 is corrupted with 5 dB: (a) *babble* and (b) *factory* noises.

the proposed tuning algorithm will be of benefit to the AKF for speech enhancement in various noise conditions.

4. Experimental setup

4.1. Speech corpus

For the objective experiments, 30 phonetically balanced utterances belonging to six speakers (three male and three female) are taken from the NOIZEUS corpus [22], Chapter 12. The noisy speech for the test set is generated by mixing the clean speech with real-world non-stationary (*babble* and *street*) and colored (*factory* and *f16*) noises at multiple SNR levels (from −5dB to + 15 dB, in 5 dB increments). The *babble* noise is taken from AURORA database [24], the *street* noise is taken from Nonspeech database [27], and the *factory* and *f16* noises are taken from RSG-10 database [23]. All the clean speech and noise recordings are single-channel with a sampling frequency of 8 kHz. The noisy speech dataset provides 30 examples per condition with 20 total conditions.

4.2. Objective evaluation

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 [28]. It ranges between −0.5 and 4.5. A higher PESQ score indicates better speech quality.
- The short-time objective intelligibility (STOI) measure for objective intelligibility evaluation [29]. It ranges between 0 and 1 (or 0 to 100%). A higher STOI score indicates better speech intelligibility.

We also analyzed the spectrograms of enhanced speech produced by the competing SEAs to visually quantify the level of *residual* noise as well as *distortion*.

4.3. Subjective evaluation

The subjective evaluation was carried out through a series of blind AB listening tests [4, Section 3.3.4]. To perform the tests, we generate a set of stimuli by corrupting the utterances sp05 and sp27 from the NOIZEUS corpus [22], Chapter 12. The reference transcript for utterance sp05 is: "*Wipe the grease off his dirty face*", and is corrupted with 5 dB *factory* noise. The reference transcript for utterance sp27 is: "*Bring your best compass to the third class*", and is corrupted with 5 dB *babble* noise. Utterances sp05 and sp27 were uttered by a male and a female, respectively. In the tests, the enhanced speech produced by eight SEAs as well as the corresponding clean speech and noisy speech signals were played as stimuli pairs to the listeners. Specifically, the tests were performed on a total of 180 stimuli pairs (90 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 pairwise scoring, 100% award is given to the preferred method, and 0% to the other. For a similar preference response, each method is awarded a score of 50%. Participants could re-listen to stimuli if required. Ten English speaking listeners participate in the blind AB listening tests[2]. The average of the preference scores given by the listeners termed as mean subjective preference score (%), which is used to compare the performance among the SEAs.

4.4. Specifications of the competing SEAs

The performance of the proposed SEA is compared to the following SEAs (the following notation is used for convenience: (p, q) : is the order of $\{a_i\}$ and $\{b_k\}$, $(\sigma_w{}^2, \sigma_u{}^2)$ are the prediction error variances of the speech and noise AR models, w_f is the analysis frame duration (ms), and s_f is the analysis frame shift (ms)).

i. Noisy: speech corrupted with additive noise.
ii. AKF-Oracle: AKF, where $(\{a_i\}, \sigma_w{}^2)$ and $(\{b_j\}, \sigma_u{}^2)$ are computed from each frame of the clean speech and the additive noise, $p = 16$, $q = 40$, $w_f = 20$ ms, $s_f = 0$ ms, and rectangular window is used for framing.
iii. AKF-Non-oracle: AKF, where $(\{a_i\}, \sigma_w{}^2)$ and $(\{b_j\}, \sigma_u{}^2)$ are computed from each frame of the noisy speech, $p = 16$, $q = 40$, $w_f = 20$ ms, $s_f = 0$ ms, and rectangular window is used for framing.
iv. MMSE-STSA [8]: It used $w_f = 25$ ms, $s_f = 10$ ms, and Hamming window is used for analysis and synthesis.
v. AKF-IT [12]: AKF operates with two iterations, where the initial $(\{a_i\}, \sigma_w{}^2)$ and $(\{b_j\}, \sigma_u{}^2)$ are computed from each frame of the noisy speech followed by re-estimation of them from the processed speech frame after first iteration, $p = 10$, $q = 10$, $w_f = 20$ ms, $s_f = 0$ ms, and rectangular window is used for framing.
vi. SBIT-KF [15]: Subband iterative KF with two iterations, where the initial $(\{a_i\}, \sigma_w{}^2)$ are computed from each frame of the noisy speech followed by reestimation of them from the processed speech frame after first iteration, $\sigma_v{}^2$ is estimated from each noisy speech frame at 1*st* iteration, $p = 8$, $w_f = 32$ ms, $s_f = 0$ ms, and rectangular window is used for framing.
vii. AKF-RMBT [19]: Robustness metric-based tuning of AKF, where $(\{b_j\}, \sigma_u{}^2)$ are computed from the first noisy speech frame being considered as silent, $(\{a_i\}, \sigma_w{}^2)$ are computed from pre-whitened speech frame, $p = 10$, $q = 40$, $w_f = 20$ ms, $s_f = 0$ ms, and rectangular window is used for framing.
viii. AKF-SMBT [20]: Sensitivity metric-based tuning of AKF, where $(\{b_j\}, \sigma_u{}^2)$ are computed from the first noisy speech frame being considered as silent, $(\{a_i\}, \sigma_w{}^2)$ are computed from pre-whitened speech frame, $p = 10$, $q = 40$, $w_f = 32$ ms, $s_f = 16$ ms, and rectangular window is used for framing.
ix. Proposed: Robustness and sensitivity metrics-based tuning of the AKF, where $(\{b_j\}, \sigma_u{}^2)$ are computed from each frame of the estimated noise, $(\{a_i\}, \sigma_w{}^2)$ are computed from each

[2] The AB listening tests were conducted with approval from the Griffith University's Human Research Ethics Committee: database protocol number 2018/671.

frame of the pre-whitened speech, $p = 16$, $q = 40$, $w_f = 32$ ms, $s_f = 16$ ms, rectangular window is used for generating time domain frames, and Hamming window is used for acoustic domain analysis and synthesis.

5. Results and discussions

5.1. Objective quality evaluation

Fig. 8 shows the average PESQ score for each SEA. It can be seen that the AKF-Oracle method attained the highest average PESQ score for all of the tested conditions. This is due to ($\{a_i\}$, σ_w^2) and ($\{b_k\}$, σ_u^2) being computed from the clean speech and the noise signal, which is unobserved in practice. Thus, AKF-Oracle provides an indication of the upper-bound for the AKF in terms of average PESQ score. Conversely, the average PESQ score for Noisy indicates the lower bound of the average PESQ score for each of the tested conditions. The proposed SEA consistently produces a higher average PESQ score than the competing SEAs across the tested conditions. The average PESQ score for the proposed method is also very similar to that of the KF-Oracle method. This is likely due to the reduced-biased AKF gain achieved by the proposed tuning algorithm (Fig. 7). Amongst the competing methods, AKF-SMBT [20]

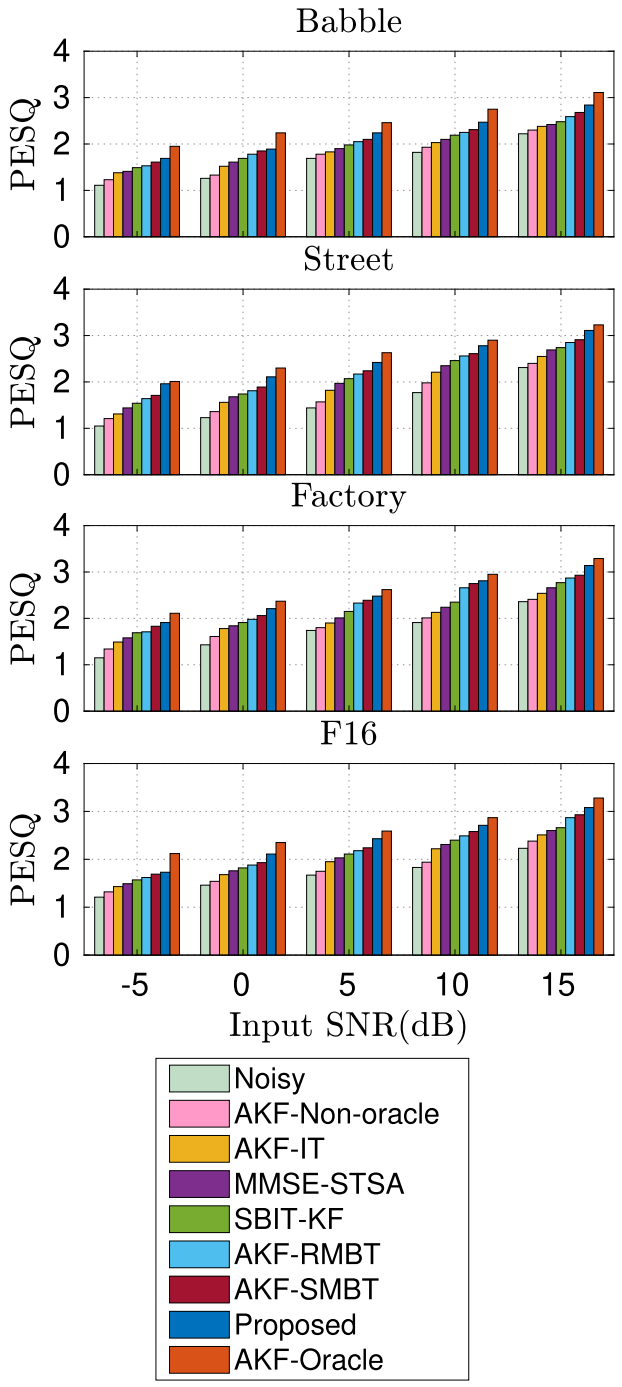

Fig. 8. Average PESQ score for each SEA found over all frames for each condition described in Section 4.1.

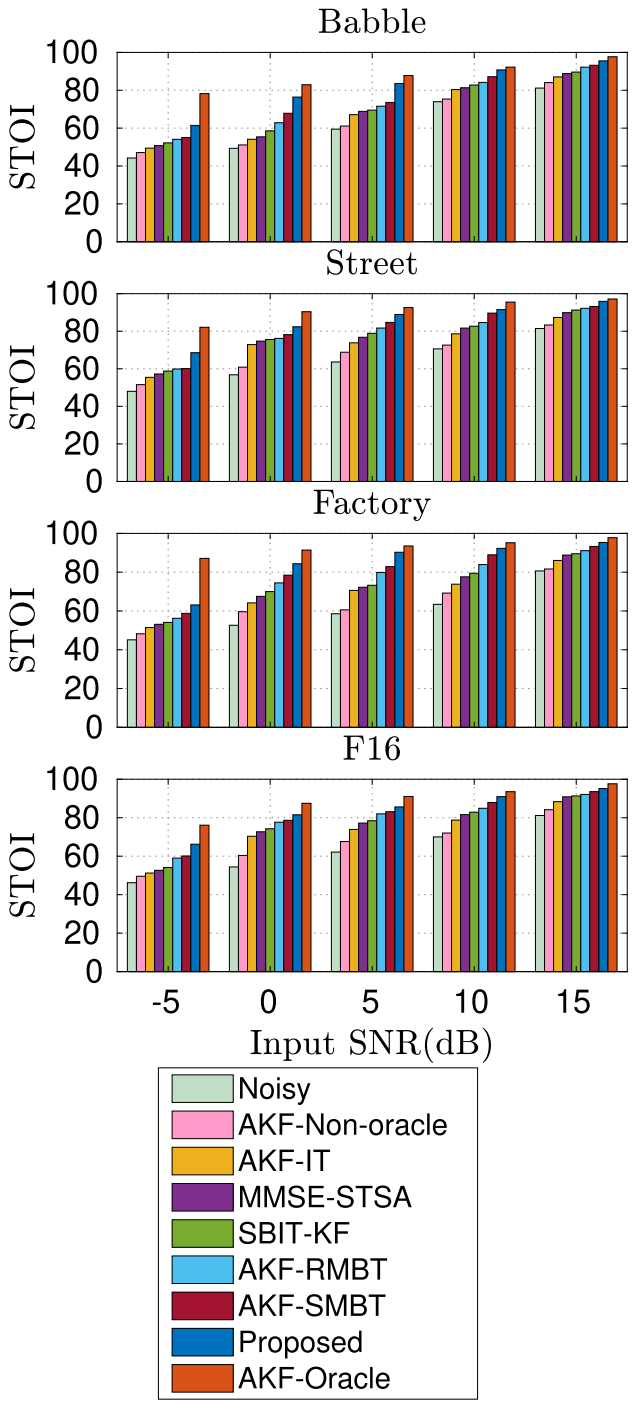

Fig. 9. Average STOI score for each SEA found over all frames for each condition described in Section 4.1.

relatively produced higher average PESQ scores for each of the tested conditions (Fig. 8). In light of this study, it is evident to say that the proposed SEA produces higher quality enhanced speech than that of the competing SEAs across the tested conditions.

5.2. Objective intelligibility evaluation

Fig. 9 shows the average STOI score for each SEA. As in Section 5.1, the AKF-Oracle method achieves the highest average STOI score for each tested condition. On the other hand, Noisy shows the

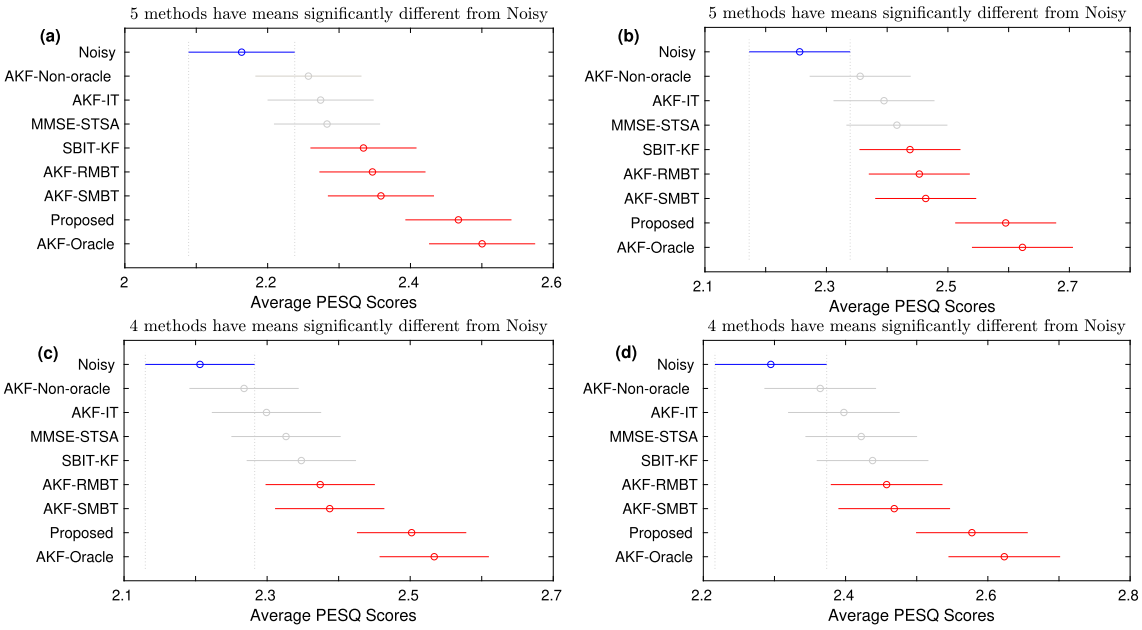

Fig. 10. Results obtained from comparative statistical analysis of PESQ for each method on the NOIZEUS corpus corrupted with: (a) *babble*, (b) *street*, (c) *factory*, and (d) *f16* noise sources at multiple SNR levels. The circle markers represent the PESQ score means along with their uncertainty intervals, which are represented by the horizontal solid lines. The red color refers to the methods that have PESQ score means significantly different from those of Noisy, which are highlighted in blue. The black color refers to the methods that show that their mean PESQ scores are not significantly different from the scores of Noisy. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

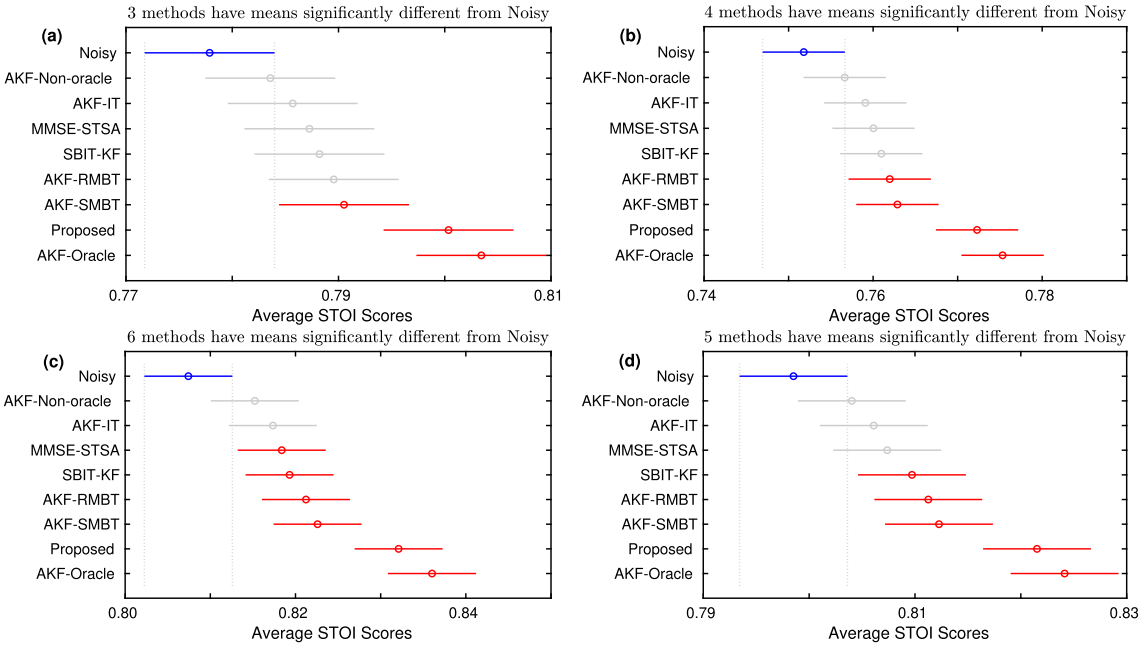

Fig. 11. Results obtained from comparative statistical analysis of STOI for each method on the NOIZEUS corpus corrupted with: (a) *babble*, (b) *street*, (c) *factory*, and (d) *f16* noise sources at multiple SNR levels. The circle markers represent the STOI score means along with their uncertainty intervals, which are represented by the horizontal solid lines. The red color refers to the methods that have STOI score means significantly different from those of Noisy, which are highlighted in blue. The black color refers to the methods that show that their mean STOI scores are not significantly different from the scores of Noisy. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

lowest average STOI score in any tested condition. The proposed method attained the highest average STOI score for each tested condition, apart from the AKF-Oracle method. Amongst the competing methods, AKF-SMBT [20] attained the highest average STOI scores. In light of this study, it is evident to say that the proposed SEA produces more intelligible enhanced speech than the competing SEAs across the tested conditions.

5.3. Statistical ANOVA test on PESQ and STOI score

To assess the significant differences between the quality and intelligibility scores for each method, we performed an analysis of variance (ANOVA) testing on the PESQ and STOI scores on the noisy speech dataset in Section 4.1 [30]. Specifically, a multiple comparison statistical test based on Tukey's honestly significant difference procedure with a significance level of 0.05 was conducted on the ANOVA test results to compute the significant differences among the SEAs [30]. The multiple comparison experimental results on the PESQ and STOI are presented in Figs. 10-11, where the circle markers indicate the objective score (PESQ or STOI) means along with their uncertainty intervals— which are represented by the horizontal solid lines [30]. The evaluation is performed on the basis of: any two objective scores (PESQ or STOI) means are significantly different from each other, if and only if their uncertainty intervals do not overlap [31]. In Figs. 10,11, each sub-figure highlights the methods where their enhanced speech signals have a mean objective score (red circle markers) significantly different from that of the Noisy (blue circle marker). In Figs. 10-11, each sub-figure also identifies the methods that performed poorly (black circle markers), with means that are not significantly different from the of Noisy.

Table 1
Comparing the normalized processing time between the proposed and competing SEAs.

Methods	Normalized Processing Time
AKF-Oracle	1.00
Proposed	1.09
AKF-SMBT [20]	1.16
AKF-RMBT [19]	1.16
SBIT-KF [15]	3.78
MMSE-STSA [8]	1.12
AKF-IT [12]	3.89
AKF-Non-oracle	1.19

It can be seen from Fig. 10 (a)-(b) that the five competing methods have means significantly different from Noisy for *babble* and *street* noise experiments. On the other hand, four competing methods have means significantly different from the Noisy for *factory* and *f16* noise sources (Fig. 10 (c)-(d)). For all noise conditions (Fig. 10 (a)-(d)), the proposed method has mean significantly different from the competing methods showing improvement against Noisy, apart from the AKF-Oracle.

Fig. 11 shows the results obtained from the comparative statistical analysis of STOI for each method as in Fig. 10. It can be seen from Fig. 11 (a)-(d) that the three methods have means significantly different from Noisy for *babble* to that of the four methods for *street*, six meth-ods for *factory* and five methods for *f16* noise sources. For all noise conditions (Fig. 11 (a)-(d)), the proposed method has a mean significantly different from the competing methods showing improvement against Noisy, apart from the AKF-Oracle.

In light of the comparative study on Figs. 10-11, it is evident to say that the proposed method has shown a significant PESQ and STOI score improvement, except the AKF-Oracle method for all noise conditions. Amongst the competing method, the AKF-SMBT [20] exhibits most PESQ and STOI score improvements followed by AKFRMBT [19].

5.4. Computational complexity evaluation of each SEA

Computation cost is also an important measure to justify the efficiency of a SEA. The computational complexity in terms of normalized processing time of Matlab implementation [26] for all methods is given in Table 1. It can be seen that the proposed method takes a lower computational time as compared to the competing methods, except the AKF-Oracle method. Amongst the competing methods, MMSE-STSA [8] takes almost similar computation time with the proposed method (1.12) followed by AKF-SMBT [20] and AKF-RMBT [19] (1.16), and AKFNon-oracle (1.19). It is also found that the SBIT-KF [15] (3.78), and AKF-IT [12] (3.89) become computationally worse than any other methods. It is due to the iterative processing of noisy speech by SBIT-KF [15] and AKF-IT [12] methods for speech enhancement.

5.5. Spectrogram analysis of each SEA

Fig. 12 (a) shows the spectrogram of clean speech (female utterance sp27). The clean speech is corrupted by *babble* noise at 5 dB

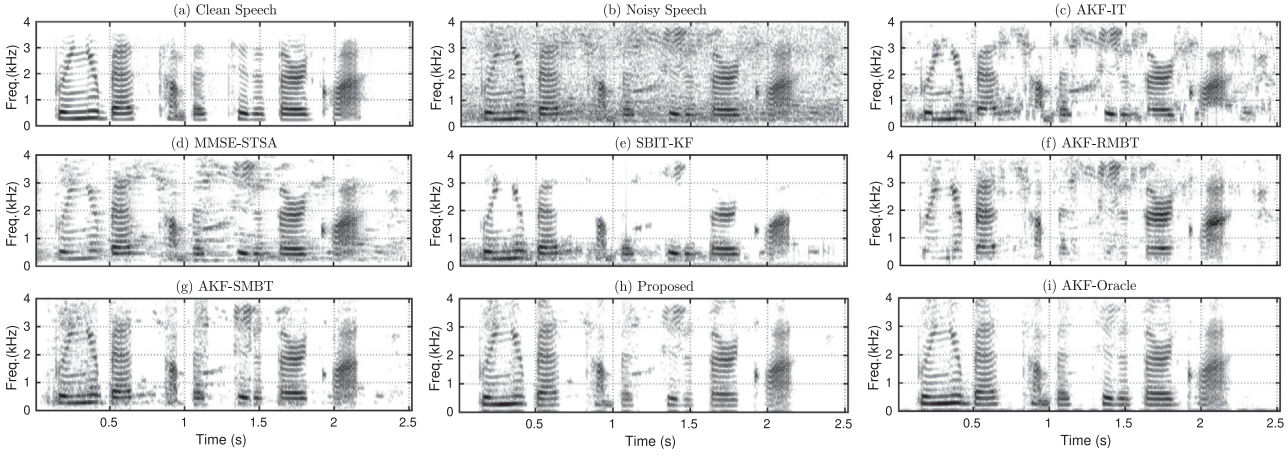

Fig. 12. Spectrograms of: (a) clean speech (utterance sp27), (b) noisy speech ((a) corrupted with 5 dB *babble* noise) (PESQ = 1.72), enhanced speech produced by: (c) AKF-IT [12] (PESQ = 1.91), (d) MMSE-STSA [8] (PESQ = 2.06), (e) SBIT-KF [15] (PESQ = 2.17), (f) AKF-RMBT [19] (PESQ = 2.22), (g) AKF-SMBT [20] (PESQ = 2.30), (h) Proposed (PESQ = 2.49), and (i) AKF-Oracle (PESQ = 2.69).

SNR level to create the noisy speech (Fig. 12 (b)). This is a particularly tough condition for speech enhancement since the background noise exhibits characteristics similar to the speech produced by the target speaker. The enhanced speech produced by AKFIT [12] is shown in Fig. 12 (c). The enhanced speech suffers from significant speech distortion. A significant residual background noise is also present in the enhanced speech. Fig. 12 (d) shows the enhanced speech produced by MMSE-STSA [8]. This method produced less distorted speech than AKF-IT (Fig. 12 (c)); however, residual background noise still remains. Less residual background noise is present in the enhanced speech produced by SBIT-KF [15] (Fig. 12 (e)) than MMSE-STSA (Fig. 12 (d)), however, the speech is more distorted. The AKF-RMBT [19] method produced less distorted speech (Fig. 12 (f)) than that of SBIT-KF (Fig. 12 (e)). The enhanced speech produced by AKF-SMBT [20] (Fig. 12 (g)) relatively shows less distortion as well as less residual background noise than that of AKF-RMBT (Fig. 12 (f)). The enhanced speech produced by the proposed method is shown in Fig. 12 (h). It can be seen that there is less residual background noise in the enhanced speech than AKF-SMBT (Fig. 12

(g)). Finally, the enhanced speech produced by the AKFOracle method is shown in Fig. 12 (i), which is most similar to the clean speech in Fig. 12 (a). This is due to AKF-Oracle uses clean speech and noise (unobserved in practice) for LPC parameter estimation.

5.6. Subjective evaluation by AB listening test

The mean subjective preference score (%) for each SEA is shown in Figs. 13-14. The colored (*factory*) noise experiment in Fig. 13 reveals that the enhanced speech produced by the proposed method is widely preferred by the listeners (74%) over the competing methods, apart from the clean speech (100%) and the AKF-Oracle method (84%). AKF-SMBT [20] is found to be the most preferred method (around 67%) amongst the benchmark methods by the listeners. For the non-stationary (*babble*) noise experiment (Fig. 14), the listeners again preferred the proposed method (72%) over the competing methods, with only clean speech (100%) and AKF-Oracle (82%) being more preferred. As in the previous experiment (Fig. 13), AKFSMBT [20] was the most preferred (65%) amongst the competing methods, with AKF-RMBT [19] (around 58%) being the next most preferred. In light of the blind AB listening tests, it is evident to say that the enhanced speech of the proposed method exhibits the best-perceived quality amongst all tested methods for both male and female utterances corrupted by real-life colored as well as nonstationary noise sources.

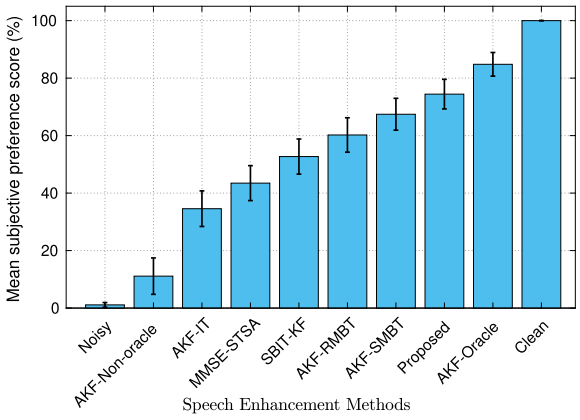

Fig. 13. The mean preference score (%) comparison between the proposed and benchmark SEAs for the utterance sp05 corrupted with 5 dB colored *factory* noise.

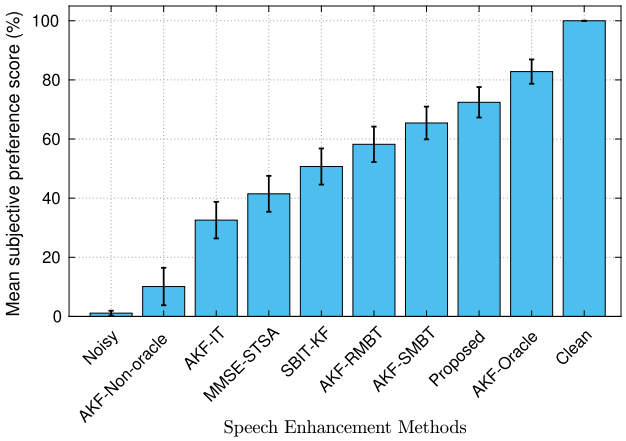

Fig. 14. The mean preference score (%) comparison between the proposed and benchmark SEAs for the utterance sp27 corrupted with 5 dB non-stationary *babble* noise.

6. Conclusion

This paper investigates robustness and sensitivity metricsbased tuning of the AKF gain for single-channel speech enhancement in real-life noise conditions. At first, an SPP method estimates the noise PSD from each noisy speech frame to compute the noise LPC parameters. A whitening filter is also constructed with the estimated noise LPCs to pre-whiten each noisy speech frame prior to computing the speech LPC parameters. Then construct the AKF with the estimated speech and noise LPC parameters. To achieve better noise reduction, the robustness metric is employed to *offset* the bias in AKF gain during speech pauses of the noisy speech to that of the sensitivity metric during speech presence. The speech and noise model parameters are adopted as a speech activity detector. It is shown that the reduced-biased AKF gain achieved by the proposed tuning algorithm addresses speech enhancement in real-life noise conditions. Objective and subjective scores on the NOIZEUS corpus demonstrate that the proposed method outperforms the competing methods in various noise conditions for a wide range of SNR levels.

The proposed method performs single-channel speech enhancement in the presence of additive background noise. However, in practice, clean speech can be corrupted by room impulse responses in the form of surface reflections (or noisy-reverberant speech) and background noise. Therefore, our future research direction is on augmented Kalman filtering for speech enhancement in the presence of noisyreverberant speech.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We gratefully thank the reviewers and editors for their effort in the improvement of this work.

References

[1] Boll S. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans Acoust Speech Signal Process 1979;27:113–20. https://doi.org/10.1109/TASSP.1979.1163209.

[2] Berouti M, Schwartz R, Makhoul J. Enhancement of speech corrupted by acoustic noise. IEEE Internatl Conf Acoust Speech Signal Process 1979;4:208–11. https://doi.org/10.1109/TASSP.1979.1163209.
[3] Kamath S, Loizou P. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. IEEE Internatl Conf Acoust Speech Signal Process 2002;4:4160–4. https://doi.org/10.1109/ICASSP.2002. 5745591.
[4] Paliwal K, Wójcicki K, Schwerin B. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. Speech Commun 2010;52(5):450–75. https://doi.org/10.1016/j.specom.2010.02.004.
[5] Lim JS, Oppenheim AV. Enhancement and bandwidth compression of noisy speech. Proc IEEE 1979;67(12):1586–604. https://doi.org/10.1109/PROC.1979.11540.
[6] Scalart P, Filho JV. Speech enhancement based on a priori signal to noise estimation. IEEE Internatl Conf Acoust Speech Signal Process 1996;2:629–32.
[7] Plapous C, Marro C, Mauuary L, Scalart P. A two-step noise reduction technique. IEEE Internatl Conf Acoust Speech Signal Process 2004;1:289–92.
[8] Ephraim Y, Malah D. Speech enhancement using a minimummean square error short-time spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process 1984;32(6):1109–21. https://doi.org/10.1109/TASSP.1984.1164453.
[9] Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process 1985;33(2):443–5. https://doi.org/10.1109/TASSP.1985.1164550.
[10] Paliwal K, Schwerin B, Wójcicki K. Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator. Speech Commun 2012;54(2):282–305. https://doi.org/10.1016/j.specom.2011.09. 003.
[11] Paliwal K, Basu A. A speech enhancement method based on kalman filtering. IEEE Internatl Conf Acoust Speech Signal Process 1987;12:177–80. https://doi.org/10.1109/ICASSP.1987.1169756.
[12] Gibson JD, Koo B, Gray SD. Filtering of colored noise for speech enhancement and coding. IEEE Trans Signal Process 1991;39(8):1732–42. https://doi.org/10.1109/78.91144.
[13] Brown GJ, Wang D. Separation of Speech By Computational Auditory Scene Analysis. Berlin Heidelberg: Springer Berlin Heidelberg; 2005. 10.1007/3-540-27489-8_16.
[14] Xu Y, Du J, Dai L, Lee C. An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process Lett 2014;21(1):65–8. https://doi.org/10.1109/LSP.2013. 2291240.
[15] Roy SK, Zhu WP, Champagne B. Single channel speech enhancement using subband iterative kalman filter. IEEE Internat Symp Circuits Systems 2016:762–5. https://doi.org/10.1109/ISCAS.2016.7527352.
[16] Saha M, Ghosh R, Goswami B. Robustness and sensitivity metrics for tuning the extended kalman filter. IEEE Trans Instrum Meas 2014;63(4):964–71. https://doi.org/10.1109/TIM.2013.2283151.
[17] So S, George AEW, Ghosh R, Paliwal KK. A noniterative kalman filtering algorithm with dynamic gain adjustment for single-channel speech enhancement. Internat J Signal Process Syst 2016:263–8. https://doi.org/10.18178/ijsps.4.4.263-268.
[18] So S, George AEW, Ghosh R, Paliwal KK. Kalman filter with sensitivity tuning for improved noise reduction in speech. Circuits Syst Signal Process 2017;36(4):1476–92. https://doi.org/10.1007/s00034-016-0363-y.
[19] George AE, So S, Ghosh R, Paliwal KK. Robustness metric-based tuning of the augmented kalman filter for the enhancement of speech corrupted with coloured noise. Speech Commun 2018;105:62–76. https://doi.org/10.1016/j.specom.2018.10.002.
[20] Roy SK, Paliwal KK. Sensitivity metric-based tuning of the augmented Kalman filter for speech enhancement. In: 14th International Conference on Signal Processing and Communication Systems (ICSPCS). p. 9310005. 10.1109/ICSPCS50536.2020.
[21] S. V. Vaseghi, Linear prediction models, in: Advanced Digital Signal Processing and Noise Reduction, John Wiley & Sons, 2009, Ch. 8, pp. 227–262.
[22] Loizou PC. Speech Enhancement: Theory and Practice. 2nd ed. Boca Raton, FL, USA: CRC Press Inc; 2013.
[23] H. J. Steeneken, F. W. Geurtsen, Description of the RSG-10 noise database, Report IZF 1988-3, TNO Institute for Perception, Soesterberg, The Netherlands.
[24] Pearce D, Hirsch H-G. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. INTERSPEECH 2000:29–32.
[25] Oppenheim AV, Schafer RW. Discrete-Time Signal Processing. 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall Press; 2009.
[26] Gerkmann T, Hendriks RC. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans Audio Speech Lang Process 2012;20(4):1383–93. https://doi.org/10.1109/TASL.2011.2180896.
[27] G. Hu, 100 nonspeech environmental sounds, The Ohio State University, Department of Computer Science and Engineering.
[28] Rix AW, Beerends JG, Hollier MP, Hekstra AP. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. IEEE Internatl Conf Acoust Speech Signal Process 2001;2:749–52. https://doi.org/10.1109/ICASSP.2001. 941023.
[29] Taal CH, Hendriks RC, Heusdens R, Jensen J. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Trans Audio Speech Lang Process 2011;19(7):2125–36. https://doi.org/10.1109/TASL.2011.2114881.
[30] Jassim WA, Harte N. Estimation of a priori signal-to-noise ratio using neurograms for speech enhancement. J Acoust Soc Am 2020;147(6):3830–48. https://doi.org/10.1121/10.0001324.
[31] Hochberg Y, Tamhane AC. Multiple Comparison Procedures. Hoboken, NJ: John Wiley & Sons; 1987.