A noise PSD estimation algorithm using derivative-based high-pass filter in non-stationary noise conditions

Roy and Paliwal *EURASIP Journal on Audio, Speech, and Music Processing* (2021) 2021:32
https://doi.org/10.1186/s13636-021-00220-9

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

A noise PSD estimation algorithm using derivative-based high-pass filter in non-stationary noise conditions

Sujan Kumar Roy[1*] and Kuldip K. Paliwal[1]

Abstract

The minimum mean-square error (MMSE)-based noise PSD estimators have been used widely for speech enhancement. However, the MMSE noise PSD estimators assume that the noise signal changes at a slower rate than the speech signal— which lacks the ability to track the highly non-stationary noise sources. Moreover, the performance of the MMSE-based noise PSD estimator largely depends upon the accuracy of the a priori SNR estimation in practice. In this paper, we introduce a noise PSD estimation algorithm using a derivative-based *high-pass* filter in non-stationary noise conditions. The proposed method processes the silent and speech frames of the noisy speech differently to estimate the noise PSD. It is due to the non-stationary noise that can be mixed with silent and speech-dominated frames non-uniformly. We first introduce a spectral-flatness-based adaptive thresholding technique to detect the speech activity of the noisy speech frames. Since the silent frame of the noisy speech is completely filled with noise, the noise periodogram is directly computed from it without applying any filtering. Conversely, a 4^{th} order derivative-based *high-pass* filter is applied during speech activity of the noisy speech frame to filter out the clean speech components while leaving behind mostly the noise. The noise periodogram is computed from the filtered signal—which counteracts the leaking of clean speech power. The noise PSD estimate is obtained by recursively averaging the previously estimated noise PSD and the current estimate of the noise periodogram. The proposed method is found to be effective in tracking the rapidly changing as well as the slowly varying noise PSD than the competing methods in non-stationary noise conditions for a wide range of signal-to-noise ratio (SNR) levels. Extensive objective and subjective scores on the NOIZEUS corpus demonstrate that the application of the proposed noise PSD with MMSE-based speech enhancement methods produce higher quality and intelligible enhanced speech than the competing methods.

Keywords: Noise PSD tracking, Spectral-flatness, Speech enhancement, High-pass filter, Derivative

1 Introduction

The speech processing systems have a close link to our daily life, such as mobile communication systems, hearing aid devices, and voiced operated autonomous systems. In practice, the clean speech signal is contaminated with the surrounding interfering noise sources and affects the performance of these systems. In most of the conditions, the interfering noise sources become non-stationary. In this circumstance, the speech enhancement algorithm (SEA) [1–9] aims to improve the quality and intelligibility of the noisy speech. It can be achieved by eliminating the embedded background noises from the noisy speech without distorting the clean speech. The noise PSD estimation is a crucial component in designing a SEA [10]. Since the noise PSD is unobserved in practice, it is quite difficult to accurately estimate noise PSD from noisy speech. In addition, an under-estimated noise PSD introduces *residual background* noise in the enhanced speech, while an over-estimation of it leads to speech distortion [10].

*Correspondence: sujankumar.roy@griffithuni.edu.au
[1]Signal Processing Laboratory, Griffith University, Nathan Campus, 170 Kessels Road, 4111 Brisbane, QLD, Australia

In stationary noise conditions, the noise PSD can be estimated during speech pauses of the noisy speech—that requires the speech activity detection (SAD) [11–13]. Since the non-stationary noise affects the speech spectrum non-uniformly, it is desired to estimate the noise PSD from both the silent/speech frames of the noisy speech. Therefore, it is challenging for a noise PSD tracking method to avoid the leaking of speech power into the estimated noise PSD during speech activity of the noisy speech.

Many approaches have been devoted to deal with the noise PSD estimation in non-stationary noise conditions in literature. Minimal tracking is the basis of early non-stationary noise PSD tracking methods. It operates with the principle that the spectral power level of the noisy speech in each frequency bin frequently decays to the spectral power level of the noise signal, even during speech presence ([14], Chapter 9). Following this strategy, Martin introduced a minimum statistics (MS)-based noise PSD estimation method [15]. In the MS method, the noise PSD estimate is given by tracking the minimum of the smoothed noisy speech power spectrum in each frequency bin within a fixed time window. However, the length of the time window has a significant impact on the accuracy of noise PSD estimates. Typically, a short time window may cause the noise PSD to be over-estimated due to the MS method might track the noisy speech spectral power instead of the noise spectral power. Conversely, a long time window results in a large delay in tracking the rapidly changing noise PSD. To address this, Doblinger introduced a continuous minima tracking-based noise PSD estimation method [16]. Unlike tracking the noisy speech spectral power within a fixed time window [15], the noise PSD is updated continuously by smoothing the noisy speech power spectra in each frequency bin using a non-linear smoothing rule. Thus, it reduces the delay of tracking the abrupt changing noise PSD. However, it performs continuous PSD smoothing without considering the speech presence/absence of noisy speech. As a result, the noise estimate increases whenever the noisy speech spectral power increases, which may be irrespective of the change in noise spectral power level. In [17], Martin introduced a further improvement of MS method [15]. It was observed that the computed noise PSD for a typical frequency band is lower than (or close to) its mean computed over the time window for that particular frequency band. That means the estimated noise PSD has a tendency of under-estimation as occurred in [15]. To address this, Martin introduced a bias compensation factor, which was multiplied with the estimated noise PSD to make it unbiased. In addition, an optimal time-frequency-dependent smoothing factor was computed for smoothing the noisy speech periodogram prior to minimum tracking. It balances the minimum tracking of the noise PSD when the noisy speech spectral power rises abruptly with respect to the change of noise power level.

The time-recursive averaging with speech presence uncertainty is known as another class of noise PSD estimation technique. It exploits the observation that the noise affects the speech spectrum non-uniformly. That means the speech spectrum at some frequency bins can be affected by noise more than others. Thus, we can update the noise PSD at a particular frequency bin containing a lower speech presence probability (SPP). It leads to the idea of noise PSD estimate, which is given by recursively averaging the past estimated noise PSD and the current noisy speech periodogram weighted by a frequency-dependent smoothing factor. In this method, an SPP is used to adjust the smoothing factor. On the basis of SPP computation, the minima controlled recursive averaging (MCRA) method [18], the improved MCRA (IMCRA) method [19], and MCRA2 method [20] have been introduced. Specifically, the SPP in MCRA method [18] is computed by the ratio of the smoothed noisy speech PSD to its local minimum and compare against a fixed threshold. It also uses the MS method [17] to search the minimum. In IMCRA method [19], the SPP computation involves two iterations of smoothing and minimum tracking. The first iteration gives a simple speech presence detector for each frequency bin. The second iteration of smoothing excludes the strong components of speech, thus allowing a short time window for minimum tracking. The SPP is estimated on the basis of a Gaussian statistical model and obtained from the ratio of the likelihood functions of speech presence and speech absence. The MCRA2 method [20] was proposed as a further improvement of the MCRA method [18]. Specifically, it employs a continuous spectral minimum tracking technique [16] rather than the fixed time window-based minimum tracking [17]. For the SPP estimate, the MCRA2 employs frequency-dependent thresholds instead of using a fixed threshold by MCRA method [18]. Since the noise PSD estimation methods [18–20] are proposed on the basis of MS principle [16, 17], the abrupt rising of noise power may increase the tracking delay as well as prone to an under-estimation of the noise PSD.

Unlike MS-based methods [17–20], Bayesian statistics-based noise PSD estimators are more prominent in rapid tracking of noise power with a shorter delay. In [21], Hendriks et al. introduced a Bayesian-motivated minimum-mean-squared-error (MMSE) noise PSD estimator with lower tracking delay (MMSE-LC). The MMSE estimator is derived by minimizing the mean-squared error (MSE) of the noisy speech spectral power to estimate the instantaneous noise power. The first-order recursive averaging using the past estimated noise PSD and the instantaneous estimate of noise power, giving the noise PSD. However, the instant noise power estimation using

the MMSE method requires the a priori and a posteriori SNR, which are unknown in practice. Typically, the noise power estimation is predominately affected by the accuracy of the a priori SNR estimate. The authors first employed a limited maximum-likelihood (ML) estimate of the a priori SNR to obtain an MMSE estimate of the noise power. However, the simple ML estimator leads to a bias in noise PSD estimates, which is minimized by multiplying a bias compensation factor (computed analytically) with it. The bias compensation factor used a second estimate of the a priori SNR, which is obtained by the *decision-directed* (DD) approach [4]. However, if the estimated noise PSD becomes too low as compared to the spectral noise power rising abruptly from one level to another, the noise PSD tracker gets stagnates. To address this, a *safety-net* is adopted, where the last 0.8 s of the noisy speech periodogram is stored. Specifically, the final noise PSD estimate is obtained by taking the maximum between the current estimated noise PSD and the minimum of the noisy speech periodogram (within the time span of 0.8 s). In [22], Gerkmann and Hendriks proposed an unbiased MMSE (U-MMSE)-based noise PSD estimator. The authors first showed that the noise PSD estimation process in [21] under the given ML a priori SNR estimator can be interpreted as a SAD (hard decision)-based estimator. Specifically, the noise PSD is updated only when the speech is absent. Thus, a bias compensation is necessary for the noise PSD estimator [21]. To cope with this problem, the SAD is replaced by a soft SPP-based estimator of the noise power. Specifically, the noise periodogram estimate is given by a sum of the past estimated noise PSD weighted by the conditional probability of speech presence and the noisy speech periodogram weighted by the conditional probability of speech absence. Then, the first-order recursive averaging using the past estimated noise PSD and the current estimate of the noise periodogram, giving the noise PSD. Therefore, unlike MMSE-LC method [21], bias compensation and the *safety-net* adaptation are unnecessary for the U-MMSE method [22]. In addition, U-MMSE also used a fixed non-adaptive a priori SNR as a parameter of the likelihood of speech presence, which avoids the necessity of the a priori SNR estimation. Therefore, the U-MMSE-based method [22] exhibits a faster noise PSD tracking capability than the MMSE-LC method [21]. In [23], Singh et al. proposed a Bayesian noise estimation in modulation-domain (MD). In this paper, the authors investigate the use of the modulation-domain to model the noise density function. Specifically, they showed that the modulation-domain-based Gamma density function better represents the noise density for all time-varying noise signals as compared to the non-modulation domain. The modulation-based Gamma density is then used to derive noise estimator via a Bayesian-motivated MMSE approach. It was claimed that the proposed noise estimator does not require bias compensation as like [21]. The proposed method yields better noise suppression as compared to the competing methods. In [24], Nielsen et al. proposed a model-based approach for noise PSD estimation. The authors claimed that the proposed method is effective in tracking the non-stationary noise PSD. However, this method requires to access the prior spectral information about the speech and noise sources, which are unobserved in practice.

In [25], Zhang et al. introduced an improvement of MMSE (IMMSE) method for noise PSD estimation. The authors incorporated a speech presence uncertainty (SPU) and a bias correction factor to compute the speech spectral power, which is used in the DD approach to improving the a priori SNR estimate. It was shown that the improvement of the a priori SNR estimate enrich the noise PSD tracking capability to some extent than that of the benchmark MMSE-based methods [21, 22]. Regardless of using the estimated speech spectral power, due to the use of past estimated noise power in the DD approach by the IMMSE method [25], it may still fail to track the abrupt changing noise PSD for the current noisy speech frame. Later on, Zhang et al. proposed a noise PSD tracking algorithm by incorporating a log-spectral power MMSE (MMSE-LSA) estimator. In this method, the smoothing parameter used in the recursive operation for noise PSD estimation is adjusted based on the SPP method. In addition, a spectral nonlinear weighting function was derived to estimate the noise spectral power, which depends on the a priori and the *a posteriori* signal-to-noise ratio (SNR). In general, the MMSE-based estimators [21, 22, 25, 26] suffer from the accurate estimates of a priori SNR in practice. In addition, the MMSE estimators commonly assume that the noise changes at a slower rate than the speech signal. Therefore, a delay is introduced during massive changes of instantaneous SNR. As a result, the MMSE-based noise PSD estimators are capable of tracking the moderately varying non-stationary noise sources; however, they do not adequately address the tracking of the highly non-stationary noise sources.

Nowadays, deep neural network (DNN) has also been used for noise PSD estimation. In [27], Chinaev et al. proposed a DNN-based noise PSD estimation method. They used a single-channel DNN-based noise presence probability (NPP) estimation for noise PSD tracking, termed as NPP-DNN. It was claimed that the algorithm provides a causal noise PSD estimate—which addresses speech enhancement for communication purposes. In [28], Zhang et al. proposed a DeepMMSE framework, which utilizes a DNN technique to estimate the a priori SNR—a key parameter for MMSE noise PSD estimators [21, 22, 25, 26]. Specifically, a residual network and a temporal convolutional neural network (ResNet-TCN)

has been incorporated within the DeepMMSE framework, which learns to map the noisy speech magnitude spectrum to the a priori SNR. The estimated a priori SNR is then employed to the MMSE-STSA and MMSE-LSA noise PSD estimators in [25, 26]. The DeepMMSE shows better noise PSD tracking as well as speech enhancement performance in terms of objective scores than [25, 26]. However, the accurate estimates of the a priori SNR in real-life non-stationary noise conditions become degraded—which reduces noise PSD tracking capabilities.

In light of the shortcomings of existing methods in the literature, our key observations in proposing the noise PSD estimator are (i) a silent frame of noisy speech completely filled with noise and the noise PSD can be directly computed from it without applying any filtering and (ii) the contamination of clean speech with noise during speech activity of the noisy speech frame may lead to a risk of leaking speech power in the estimated noise PSD. In light of the observations, in this paper, we propose a derivative-based *high-pass* filter for noise PSD tracking in non-stationary noise conditions. Specifically, the proposed method estimates the noise PSD by differently processing the silent/speech frames of the noisy speech. For this purpose, the speech activity is first obtained using a *spectral-flatness* based adaptive thresholding technique. The noise periodogram is directly computed from the silent frames of the noisy speech due to completely filled with noise. Conversely, the application of a 4^{th} order derivative-based *high-pass* filter to the noisy speech frame during speech activity filtered out the clean speech components, what it remains mostly the noise. The noise periodogram is computed from the filtered signal—which mitigates the risk of leaking the speech power. Then the noise PSD estimate is obtained by recursively averaging the past estimated noise PSD and the current estimate of the noise periodogram. The motivation of this is to provide a better estimate of noise PSD with low tracking delays leading to a significant noise suppression performance when employed in the MMSE-based SEA in non-stationary noise conditions.

The rest of the paper is organized as follows: Section 2 describes the proposed noise PSD estimation system, including the proposed speech activity detection algorithm, proposed noise PSD estimation algorithm, and a summary of the proposed algorithm. Section 3 describes the experimental setup, including speech corpus, objective and subjective evaluation measures. The experimental results are then presented in Section 4. Finally, Section 5 gives some concluding remarks.

2 Proposed noise PSD tracking system

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

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

Figure 1 shows the block-diagram of the proposed noise PSD tracking system. Firstly, a 32 ms Hamming window ([29], Chapter 7) with 50% overlap at $f_s = 16$ kHz sampling frequency was considered for converting $y(n)$ into frames, $y(n,k)$.

The noisy speech in Eq. (1) can then be represented in terms of frames as:

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

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

The noisy speech, $y(n)$ (Eq. (1)), is also analysed framewise using the short-time Fourier transform (STFT) as:

$$Y(l,m) = S(l,m) + V(l,m), \tag{3}$$

where $m \epsilon \{0,1,2,\ldots,511\}$ is the discrete-frequency index and $Y(l,m)$, $S(l,m)$, and $V(l,m)$ represent the complex-valued STFT coefficients of the noisy speech, clean speech, and noise signal, respectively. A 32 ms Hamming window ([29], Chapter 7) with 50% overlap at $f_s = 16$ kHz sampling frequency was used for analysis and synthesis.

The next step of the proposed method is speech activity detection of the noisy speech frames followed by carrying out the tracking of noise PSD. These two steps are described in the following Sections.

2.1 Proposed speech activity detection algorithm

We introduce a *spectral-flatness* (denoted by ζ)-based adaptive thresholding technique for speech activity detection. For l^{th} frame, $\zeta(l)$ is computed by the ratio of geometric and arithmetic mean of the 257-point single-sided noisy speech magnitude spectrum, $|Y(l,m)|$, containing the DC and Nyquist frequency components as [30–32]:

$$\zeta(l) = \frac{\sqrt[M]{\prod_{m=0}^{M-1} |Y(l,m)|}}{\frac{1}{M}\sum_{m=0}^{M-1} |Y(l,m)|}, \tag{4}$$

where $M = 257$, i.e. $m \epsilon \{0,1,2,\ldots,M-1\}$.

The $\zeta(l)$ ranged between 0 and 1 in the sense that the arithmetic mean of $|Y(l,m)|$ is always greater than that of the geometric mean. To interpret $\zeta(l)$ as a speech activity detector on a framewise basis, we conduct an experiment, where an IEEE utterance sp05 ("Wipe the grease off his dirty face") from the NOIZEUS corpus ([14], Chapter 12) (sampled at 16 kHz) is corrupted with 5 dB white (computer generated) as well as *babble*, *traffic*, *passing car*, and *passing train* noise sources taken from the freesound database [33]. It can be seen that $\zeta(l)$ varies between 1 and 0 over the time-frames depending on the silent/speech activity of the noisy speech (Fig. 2b)[1].

[1] To interpret the responses of $\zeta(l)$ (Eq. (4)) as a speech activity detector, we simply map the time-frame index, l in $\zeta(l)$ to the actual time-index of the

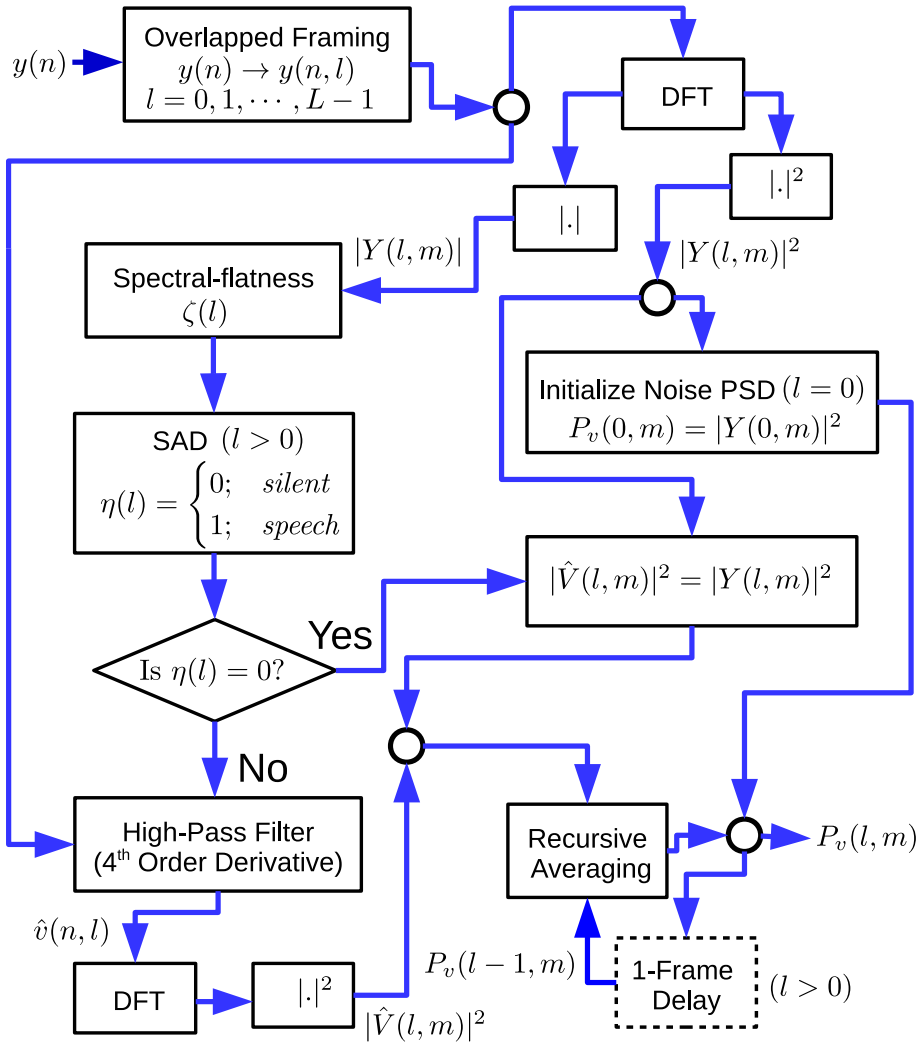

Fig. 1 Block diagram of the proposed noise PSD tracking system

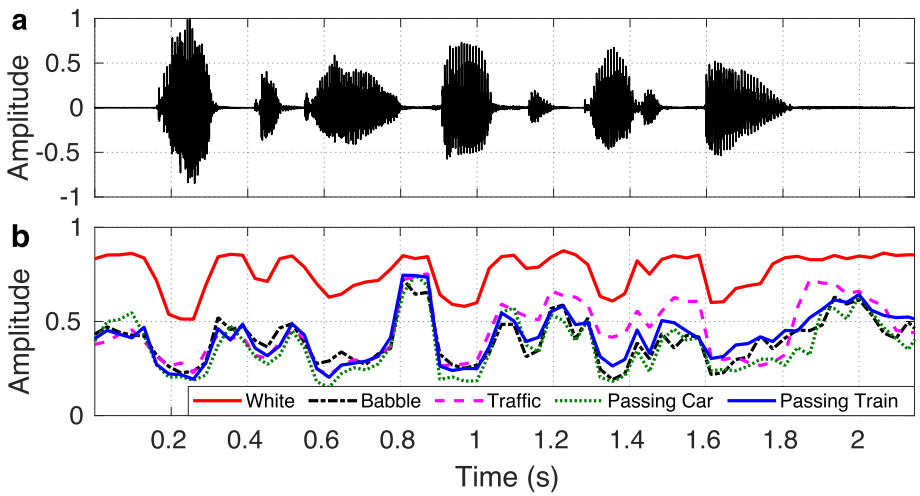

Fig. 2 a Clean speech and **b** computed $\zeta(k)$ over the time-frames of the utterance sp05 corrupted with 5 dB white, babble, traffic, passing car, and passing train noises

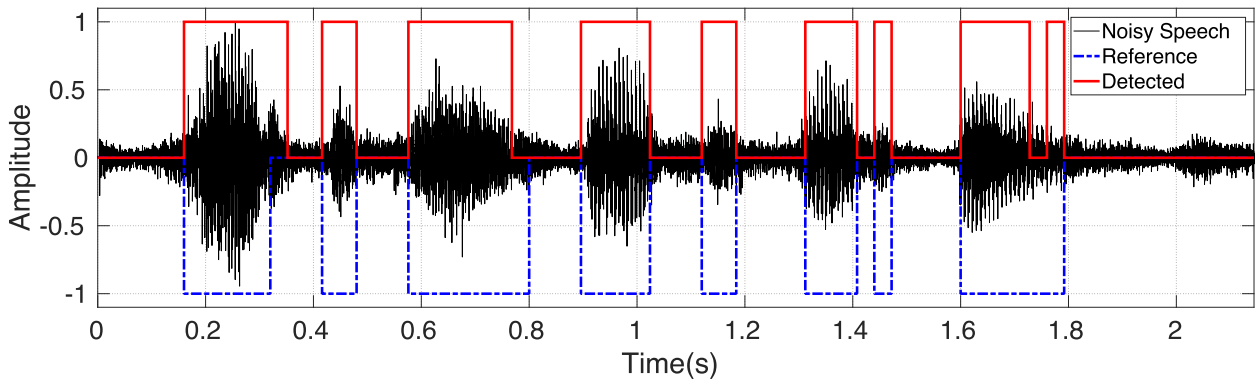

Fig. 3 Comparing the detected silent/speech flags from the utterance sp05 corrupted with 5 dB *babble* noise to that of the reference

Typically, $|Y(l,m)|$ is dominated by noise during silent activity. In white noise condition, $|Y(l,m)|$ approximately contains similar power at all frequency bins, i.e. remains flat during speech pauses, resulting in $\zeta(l) \approx 1$ (e.g. 0–0.15 s or 1.8-2.19 s of Fig. 2b). Conversely, $|Y(l,m)|$ remains non-uniform in active speech regions, yielding lower (approaching 0) $\zeta(l)$ (e.g. 0.16–0.33 s or 0.9–1.06 s of Fig. 2b). This grasps the main idea of $\zeta(l)$ being used as a speech activity detector of the noisy speech on a frame-wise basis. However, the non-stationary noise sources, such as *babble*, *traffic*, *passing car*, and *passing train*, may affect the spectrum of $|Y(l,m)|$ non-uniformly. As a result, $|Y(l,m)|$ does not remain flat during speech pauses, resulting in $\zeta(l)$ not necessarily approaching 1, but still remains higher than that of it in speech regions (Fig. 2b). To adopt $\zeta(l)$ as speech activity detector in such conditions, we propose an adaptive thresholding technique. We have found that the adaptive threshold (t_ζ) (the average of the previous $\zeta(l)$'s) can be used to detect the speech activity of the noisy speech frames. Specifically, by assuming the first $y(n,0)$ ($l = 0$) as silent, $\mathbb{S}_\zeta$ (sum of $\zeta(l)$'s in previous frames) is initialized as: $\mathbb{S}_\zeta = \zeta(0)$. For l^{th} frame ($l \geq 1$), t_ζ is computed as: $t_\zeta = \mathbb{S}_\zeta / l$, where $\mathbb{S}_\zeta = \mathbb{S}_\zeta + \zeta(l)$. If $\zeta(l) > t_\zeta$ ($l \geq 1$), $y(n,l)$ is detected as silent; otherwise speech activity is present. Since the updated $\mathbb{S}_\zeta$ at l^{th} frame is used to compute t_ζ at $(l+1)^{th}$ frame, it does not require infinite memory. The computed t_ζ is also able to capture the long-term variability of the noisy speech during speech activity detection in the sense that it takes the average of all previously computed $\zeta(l)$'s to that of the current estimate. Therefore, it minimizes the impact of the abrupt change of the noise amplitude between two successive frames during speech activity detection. The whole process is summarized in Section 2.3.

Figure 3 compares the detected flags (0/1: silent/speech) from the utterance sp05 corrupted with 5 dB *babble* noise with the reference flags (0/-1: silent/speech). It can be seen that the detected flags are closely similar to that of the reference. In this experiment, the reference flags are generated by visually inspecting the frames of sp05 (Fig. 2a). More details about the performance evaluation of the proposed SAD with existing SAD methods are given in Section 4.1.

2.2 Proposed noise PSD tracking algorithm

During silent activity of $y(n,l)$, $s(n,l) \approx 0$ (Eq. (2)), meaning that the $y(n,l)$ is completely filled with the additive noise, $v(n,l)$. Thus, unlike the benchmark methods [16–22, 25], the proposed method keeps the detected silent frames of $y(n,l)$ unprocessed. To start the algorithm, the first noisy speech frame, $y(n,0)$ ($l = 0$), is assumed to be silent, which gives an estimate of noise. Therefore, $|Y(l,m)|^2$ corresponding to $y(n,l)$ ($l = 0$) is used to initialize the noise periodogram, $|\hat{V}(0,m)|^2 = |Y(0,m)|^2$ and the noise PSD, $P_v(0,m) = |Y(0,m)|^2$. The noisy speech periodogram, $|Y(l,m)|^2$ is computed as:

$$|Y(l,m)|^2 = \frac{1}{N}\left|\sum_{n=0}^{N-1} y(n,l)e^{-j\frac{2\pi}{N}nm}\right|^2. \quad (5)$$

Specifically, during silent activity of $y(n,l)$ ($1 \leq l \leq L-1$), $|Y(l,m)|^2$ gives an estimate of the noise periodogram, i.e. $|\hat{V}(l,m)|^2 = |Y(l,m)|^2$. On the other hand, during speech activity of $y(n,l)$ ($1 \leq l \leq L-1$), $s(n,l)$ remains embedded with $v(n,l)$—which leads to a risk of leaking speech power to the estimated noise power, $|\hat{V}(l,m)|^2$. To cope with this problem, we have found that the application of a derivative based *high-pass* filter to $y(n,l)$ during speech activity filtered out the components of $s(n,l)$ before estimating $|\hat{V}(l,m)|^2$. Specifically, the clean speech, $s(n,l)$ (Eq. (2)) is smooth enough to be locally approximated with a *lower-order* polynomial terms, which can be thought of as a truncated Taylor series, whilst the noise signal, $v(n,l)$ contains a *higher-order* polynomial terms to the series of noisy speech, $y(n,l)$ [34]. It is demonstrated in Ogrodzki ([34],

noisy speech in Fig. 2. Specifically, each time-frame index, l is mapped to the time-index as: $l \times \left(\frac{M}{f_s}\right)$ (in sec).

Eq. (5.80)) that a smooth signal can be approximated by a 3^{rd} order polynomial terms, which is interpreted as a 3^{rd} order truncated Taylor series. Motivated by this observation, the application of a 4^{th} order derivative to $y(n,l)$ (Eq. (2)) acts as a *high-pass* filter, which filters out the components of $s(n,l)$, what it remains mostly the components of $v(n,l)$. Therefore, the filtered-signal gives an estimate of the additive noise, $\hat{v}(n,l)$. The filtering operation is represented as a convolution of $y(n,l)$ with a 4^{th} order derivative template, $w(n) = [1 \quad -4 \quad 6 \quad -4 \quad 1]$ as [35]:

$$\hat{v}(n,l) = \sum_{i=0}^{4} w(i) y(n-i,l). \tag{6}$$

Using the estimated $\hat{v}(n,l)$, the corresponding noise periodogram, $|\hat{V}(l,m)|^2$, is computed as:

$$|\hat{V}(l,m)|^2 = \frac{1}{N}\left|\sum_{n=0}^{N-1} \hat{v}(n,l) e^{-j\frac{2\pi}{N}nm}\right|^2. \tag{7}$$

Note that the proposed 4^{th} order derivative-based high-pass filter is used to filter out the clean speech components, not for filtering additive noise. Therefore, The application of a 4^{th} order derivative-based *high-pass* filter to $y(n,l)$ reduces most of the clean speech components, $s(n,l)$, resulting a noise dominated signal, $\hat{v}(n,l)$. As a result, although the *high-pass* filter is designed with the fixed parameter (4^{th} order derivative), it does not impact the noise having different frequency distribution. Since the components of $s(n,l)$ are filtered out prior to estimating $\hat{v}(n,l)$, it mitigates the risk of leaking speech power, $|S(l,m)|^2$ to the computed noise periodogram, $|\hat{V}(l,m)|^2$. However, the *high-pass* filter may reduce some smoothed noise components—which remain closely coincide with the clean speech, such as *babble* noise in the filtered signal, $\hat{v}(n,l)$. Therefore, for preserving the closely coincide noise components in the filtered signal, we perform a recursive averaging with the estimated noise power, $|\hat{V}(l,m)|^2$, and the noise PSD, $P_v(l,m)$ ($l > 0$) as:

$$P_v(l,m) = \beta P_v(l-1,m) + (1-\beta)|\hat{V}(l,m)|^2, \tag{8}$$

where β (ranged between 0 and 1) is a smoothing factor. The choice of β impacts the estimate of $P_v(l,m)$ to some extent. It is observed that $\beta \approx 1$ for speech-dominated frames of noise corrupted speech than that of silent frames relatively containing a bit lower value ([14], Section 9.4.1). Motivated by this observation, we empirically set $\beta = 0.98$ for $\eta(l) = 1$, and $\beta = 0.9$ for $\eta(l) = 0$, which gives a better estimate of $P_v(l,m)$.

2.3 Summary of the proposed noise PSD estimator

By integrating the discussions in Sections 2.1 and 2.2, the proposed noise PSD estimator can be summarized as:

1. **Initialization:** ($l = 0$)
   a) Compute $\zeta(0)$ using Eq. (4)
   b) Assume $\mathbb{S}_\zeta = \zeta(0)$
   c) Compute $|Y(0,m)|^2$ using Eq. (5)
   d) Assume $|\hat{V}(0,m)|^2 = |Y(0,m)|^2$
   e) Assume $P_v(0,m) = |Y(0,m)|^2$
2. **for** $l = 1$ *to* $L-1$ **do** [*framewise processing loop*]
   a) Compute $\zeta(l)$ using Eq. (4)
   b) Compute $|Y(l,m)|^2$ using Eq. (5)
   c) $\mathbb{S}_\zeta = \mathbb{S}_\zeta + \zeta(l)$
   d) $t_\zeta = \mathbb{S}_\zeta / l$
   e) **if** $\zeta(l) > t_\zeta$ **then** [*silent activity*]
      i. $|\hat{V}(l,m)|^2 = |Y(l,m)|^2$
      ii. $\beta = 0.9$
   else [*speech activity*]
      i. Estimate $\hat{v}(n,l)$ using Eq. (6)
      ii. Compute $|\hat{V}(l,m)|^2$ using Eq. (7)
      iii. $\beta = 0.98$
   end if
   f) Update $P_v(l,m)$ using Eq. (8)
   end for

2.4 Speech enhancement using estimated noise PSD

To evaluate the performance of the proposed noise PSD estimator against the benchmark methods, it is employed to the traditional MMSE-based speech enhancement system. Typically, the estimated noise PSD is used in the DD approach to computing the a priori SNR—a key parameter of the MMSE gain function used for speech enhancement [4]. Specifically, given the noisy speech magnitude spectrum, $|Y(l,m)|$, an estimate of the clean speech magnitude spectrum, $|\hat{S}(l,m)|$, is obtained as [4]:

$$|\hat{S}(l,m)| = G(l,m)|Y(l,m)| \tag{9}$$

where $G(l,m)$ is a gain function.

The MMSE-STSA gain function is given by [4]:

$$G_{\text{MMSE-STSA}}(l,m) = \frac{\sqrt{\pi}}{2}\frac{\sqrt{\nu(l,m)}}{\gamma(l,m)} \exp\left(\frac{-\nu(l,m)}{2}\right) \left[(1+\nu(l,m))I_0\left(\frac{\nu(l,m)}{2}\right) + \nu(l,m) I_1\left(\frac{\nu(l,m)}{2}\right)\right], \tag{10}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of zero and first order, and $\nu(l,m)$ is given by:

$$\nu(l,m) = \frac{\xi(l,m)}{1+\xi(l,m)}\gamma(l,m), \tag{11}$$

where $\xi(l,m)$ and $\gamma(l,m)$ are the a priori and *a posteriori* SNR, respectively, defined as [4]:

$$\xi(l,m) = \frac{\lambda_s(l,m)}{\lambda_v(l,m)}, \tag{12}$$

$$\gamma(n,k) = \frac{|Y(l,m)|^2}{\lambda_v(l,m)}, \tag{13}$$

where $\lambda_s(l,m) = \mathrm{E}\{|S(l,m)|^2\}$ is the variance of the clean speech spectral component and $\lambda_v(l,m) = \mathrm{E}\{|V(l,m)|^2\}$ is the variance of the noise spectral component. In practice, we do not have access to $|S(l,m)|^2$ and $|V(l,m)|^2$ for computing $\lambda_s(l,m)$ and $\lambda_v(l,m)$. Thus, we need to estimate $\lambda_s(l,m)$ and $\lambda_v(l,m)$ from noisy speech for computing $\xi(l,m)$ and $\gamma(l,m)$. In this paper, we have used the noise PSD, $P_v(l,m)$ estimated by the proposed and benchmark methods, which replace $\lambda_v(l,m)$ to compute $\gamma(l,m)$. With the computed $\gamma(l,m)$, the traditional DD approach gives an estimate of $\hat{\xi}(l,m)$ as [4]:

$$\hat{\xi}(l,m) = \eta \frac{|\hat{S}(l-1,m)|^2}{P_v(l-1,m)} + (1-\eta)\max(\hat{\gamma}(l,m) - 1, 0), \tag{14}$$

where max(.) is the maximum function, η is the smoothing factor usually set to 0.98, and $|\hat{S}(l-1,m)|^2$ and $P_v(l-1,m)$ represent the estimated clean speech power spectrum and noise PSD at $(l-1)^{th}$ $(l > 0)$ frame, respectively.

Using the estimated $\hat{\xi}(l,m)$, the MMSE-LSA gain function is given by [5]:

$$G_{\text{MMSE-LSA}}(l,m) = \frac{\hat{\xi}(l,m)}{1+\hat{\xi}(l,m)} \exp\left\{\frac{1}{2}\int_{v(l,m)}^{\infty} \frac{e^{-t}}{t} dt\right\}, \tag{15}$$

where the integral part belongs to the exponential integral.

We have also used the square-root Wiener filter (SRWF) gain function, which can be represented in terms of $\hat{\xi}(l,m)$ as Loizou ([14], Section 6.5.1 of Chapter 6):

$$G_{\text{SRWF}}(l,m) = \sqrt{\frac{\hat{\xi}(l,m)}{1+\hat{\xi}(l,m)}}. \tag{16}$$

3 Experimental setup

3.1 Speech corpus

The performance evaluations are conducted on the NOIZEUS speech corpus, which includes 30 phonetically balanced IEEE utterances belonging to six speakers (three male and three female) ([14], Chapter 12). For objective experiments, the clean speech utterances are corrupted with five different non-stationary noise sources at multiple SNR levels (from −10 to +10 dB, in 5 dB increments). This provides 30 examples per condition with 25 total conditions. The first non-stationary noise source is a computer-generated modulated white Gaussian noise (Mod. WGN). It can be generated by modulating the white Gaussian noise as follows [25]:

$$v(n) = 0.1 + 0.5\sin\left(2\pi n \frac{f_{mod}}{f_s} - \pi\right), \tag{17}$$

where f_{mod} is the modulating frequency. We have chosen $f_{mod} = 0.1$ Hz in the simulation. The other non-stationary noise sources, such as *babble*, *traffic*, *passing car*, and *passing train,* noise sources are taken from the freesound database [33]. Note that the clean speech recordings in the NOIZEUS corpus actually taken from IEEE dataset [14] having a sampling frequency of 25 kHz. On the other hand, all the noise recordings taken from the freesound database [33] having a sampling frequency ranged between 20 and 41 kHz. For objective experiments, all clean speech and noise recordings were down-sampled to 16 kHz prior to generating the noisy speech dataset.

3.2 Evaluation measures

Three levels of objective evaluation are carried out in this paper: firstly, the performance comparison of the proposed speech activity detection (SAD) method; secondly, the performance of the noise PSD tracking among the competing methods; and finally, we compare the quality and intelligibility of the enhanced speech produced by the MMSE-based SEAs in Section 2.4—that used the noise PSD estimated by the competing methods.

3.2.1 Objective measure for SAD

The objective measure for SAD defined in [36] also used in this paper. In this measure, the speech-dominated frames are regarded as voiced frame and silent or other non-speech frames are regarded as unvoiced. Specifically, voiced-to-unvoiced (V-Uv) and unvoiced-to-voiced (Uv-V) error rates denote the accuracy in correctly classifying voiced/unvoiced speech frames given the noisy speech. A Uv-V error occurs when an unvoiced frame is classified erroneously as voiced, and a V-Uv error occurs once a voiced frame is detected as unvoiced. The overall error rate is obtained by summing up the V-Uv and Uv-V errors. For performance evaluation, the reference database is created using 30 IEEE clean speech recordings from the NOIZEUS corpus in Section 3.1 by labelling 1 for a voiced frame and 0 for an unvoiced frame. For the objective experiment, the competing SAD methods (Section 3.4) are applied to the noisy speech dataset (Section 3.1) and similarly labelled the detected voiced frame as 1 and unvoiced frame as 0. Then compute the V-Uv and Uv-V errors with respect to the reference dataset.

3.2.2 Objective measure for noise PSD tracking

The efficiency of the estimated noise PSD is measured in terms of the logarithmic error distance (LogErr) (dB) measure, given by [22, 25]:

$$\text{LogErr} = \frac{1}{LM}\sum_{l=0}^{L-1}\sum_{m=0}^{M-1}\left|10\log_{10}\left[\frac{P_r(l,m)}{P_v(l,m)}\right]\right|, \tag{18}$$

where $P_r(l,m)$ is the reference noise PSD.

The additive noise $v(n)$ that is known first convert into frames, $v(n,k)$, with the same specifications used in generating, $y(n,k)$. Then, the computed noise periodogram, $|V(l,m)|^2$ from $v(n,k)$ is used to recursively update the reference noise PSD, $P_r(l,m)$ $(l>0)$, as [22]:

$$P_r(l,m) = \alpha P_r(l-1,m) + (1-\alpha)|V(l,m)|^2, \quad (19)$$

where $P_r(0,m) = |V(0,m)|^2$ (for $l=0$) and $\alpha = 0.9$ is chosen to smooth the power fluctuations in $P_r(l,m)$ [21].

A lower LogErr (dB) indicates a better noise PSD tracking capability. However, the error measure (LogErr) can be separated into overestimation (denoted as $\text{LogErr}_{\text{ov}}$) and underestimation (denoted as $\text{LogErr}_{\text{un}}$) as [22, 25]:

$$\text{LogErr} = \text{LogErr}_{\text{ov}} + \text{LogErr}_{\text{un}}, \quad (20)$$

where $\text{LogErr}_{\text{ov}}$ and $\text{LogErr}_{\text{un}}$ are defined as [22, 25]:

$$\text{LogErr}_{\text{ov}} = \frac{1}{LM}\sum_{l=0}^{L-1}\sum_{m=0}^{M-1}\left|\min\left(0, 10\log_{10}\left[\frac{P_r(l,m)}{P_v(l,m)}\right]\right)\right|, \quad (21)$$

$$\text{LogErr}_{\text{un}} = \frac{1}{LM}\sum_{l=0}^{L-1}\sum_{m=0}^{M-1}\left|\max\left(0, 10\log_{10}\left[\frac{P_r(l,m)}{P_v(l,m)}\right]\right)\right|. \quad (22)$$

The overestimation of noise PSD (measured by $\text{LogErr}_{\text{ov}}$) leads to speech distortion in speech enhancement contexts. Conversely, the underestimation of noise PSD (measured by $\text{LogErr}_{\text{un}}$) results in an increasing amount of residual background noise in the enhanced speech. More details about the impact of the estimated noise PSD by the proposed and competing methods are given in Sections 4.4–4.7.

3.2.3 Objective measure for speech enhancement

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 measures have been used in this paper:

- Perceptual Evaluation of Speech Quality (PESQ) for objective quality evaluation [37]. 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 [38]. It ranges between 0 and 100%. A higher STOI score indicates better speech intelligibility.

We also analysed the spectrograms of enhanced speech produced by the competitive methods to visually quantify the level of *residual background* noise and *speech distortion*. For this purpose, we generate a stimuli set by concatenating the utterances sp05 and sp12 corrupted with 5 dB *passing train* noise.

3.3 Subjective evaluation measure for speech enhancement

The subjective evaluation was carried out through a series of blind AB listening tests ([9], Section 3.3.4). To perform the tests, we generated a set of stimuli by corrupting the IEEE clean speech utterances *sp05*, *sp10*, *sp21*, *sp26*, and *sp27* from the NOIZEUS corpus ([14], Chapter 12). The reference transcript for utterance *sp05* is: "Wipe the grease off his dirty face", and is corrupted with *mod. WGN* at 0 dB. The reference transcript for utterance *sp10* is: "The sky that morning was clear and bright blue", and is corrupted with *babble* at 5 dB. The reference transcript for utterance *sp21* is: "Clams are small, round, soft and tasty", and is corrupted with *traffic* at 10 dB. The reference transcript for utterance *sp26* is: "She has a smart way of wearing clothes", and is corrupted with *passing car* at 0 dB. The reference transcript for utterance *sp27* is: "Bring your best compass to the third class", and is corrupted with *passing train* at 5 dB. Utterances *sp05*, *sp10*, and *sp21* were uttered by male and utterances *sp26* and *sp27* were uttered by female, respectively.

In this test, the enhanced speech produced by seven SEAs as well as the corresponding clean speech and noisy speech signals were played as stimuli pairs to the listeners. Specifically, the test is performed on a total of 360 stimuli pairs (72 for each utterance) played in a random order to each listener, excluding the comparisons between the same method. The listener 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 [2]. The average of the preference scores given by the listeners, termed as mean subjective preference score (%), is used to compare the efficiency among the SEAs.

3.4 Specifications of competing methods

The performance evaluation of the proposed SAD method (Section 2.1) is carried out by comparing it with the following competing methods: spectrum energy-based SAD (SE-SAD) [41], formant-based SAD (FrSAD) [40], linear model of empirical mode decomposition (LmEMD) [36], multi-feature-based SAD (MF-SAD) [30], empirical mode decomposition (EMD)-based noise filtering (NfEMD), and without noise filtering (WnF) [39].

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

For noise PSD tracking and speech enhancement, the performance of the proposed noise PSD tracking method (Section 2.2) is carried out by comparing it with the following competing methods: noise-presence-probability and DNN-based noise PSD tracking method (NPP-DNN) [27], improved MMSE (IMMSE) method [25], unbiased MMSE (U-MMSE) method [22], MMSE with low complexity (MMSE-LC) method [21], MCRA method [18], and MS method [17].

4 Results and discussions

4.1 Evaluation of speech activity detection

Table 1 presents the average V-Uv (%), Uv-V (%), and Overall (%) error rate comparison for each method. Note that the results were computed by taking the average of V-Uv (%), Uv-V (%), and overall (%) error rates for all frames of the noisy speech (Section 3.1) at each SNR levels with respect to the reference SAD dataset as specified in Section 3.2.1. We have also included the average V-Uv (%), Uv-V (%), and overall (%) error rates for each method in oracle case (apply each SAD method to the 30 clean speech in Section 3.1). It can be seen that the proposed SAD method consistently demonstrates lower V-Uv (%), Uv-V (%), and overall (%) error rate as compared to the competing methods at all SNR levels and oracle case. Amongst the competing methods, LmEMD [36] relatively shows better V-Uv (%), Uv-V (%), and Overall (%) followed by NfEMD [39], FrSAD [40], MF-SAD [30], and SE-SAD [41]. The WnF method [39] shows the lowest V-Uv (%), Uv-V (%), and overall (%). In light of the comparative study, it is evident to say that the proposed SAD method outperforms the competing methods in various non-stationary noise conditions for a wide range of SNR levels. The accuracy of the proposed SAD method will improve the performance of the proposed noise PSD estimator (Section 2.2).

Table 1 Average V-Uv (%), Uv-V (%), and overall (%) error rate comparison for each SAD method over the noisy speech dataset in Section 3.1

		Errors		
SNR (dB)	Methods	V-Uv (%)	Uv-V (%)	Overall (%)
Oracle	Proposed	**0.09**	**0.25**	**0.34**
	LmEMD [36]	0.21	0.42	0.63
	NfEMD [39]	0.29	0.49	0.78
	FrSAD [40]	0.35	0.56	0.91
	MF-SAD [30]	0.41	0.71	1.12
	SE-SAD [41]	0.48	0.77	1.25
	WnF [39]	0.63	0.85	1.48
15	Proposed	**0.13**	**0.36**	**0.49**
	LmEMD [36]	0.33	0.64	0.97
	NfEMD [39]	0.46	0.77	1.23
	FrSAD [40]	0.56	0.89	1.45
	MF-SAD [30]	0.67	1.14	1.81
	SE-SAD [41]	0.78	1.24	2.02
	WnF [39]	0.97	1.38	2.35
10	Proposed	**0.25**	**0.63**	**0.88**
	LmEMD [36]	0.48	0.93	1.41
	NfEMD [39]	0.69	1.13	1.82
	FrSAD [40]	0.82	1.32	2.14
	MF-SAD [30]	0.97	1.68	2.65
	SE-SAD [41]	1.45	1.84	2.99
	WnF [39]	1.78	2.13	3.91
5	Proposed	**0.32**	**0.87**	**1.19**
	LmEMD [36]	0.75	1.47	2.22
	NfEMD [39]	1.03	1.75	2.78
	FrSAD [40]	1.26	2.02	3.28
	MF-SAD [30]	1.48	2.57	4.05
	SE-SAD [41]	1.74	2.81	4.55
	WnF [39]	2.83	3.19	6.02
0	Proposed	**0.41**	**1.12**	**1.53**
	LmEMD [36]	0.96	1.87	2.83
	NfEMD [39]	1.33	2.11	3.44
	FrSAD [40]	1.66	2.67	4.33
	MF-SAD [30]	1.89	2.82	4.71
	SE-SAD [41]	2.15	3.26	5.41
	WnF [39]	3.11	5.23	8.34
-5	Proposed	**0.88**	**1.94**	**2.82**
	LmEMD [36]	1.91	2.82	2.83
	NfEMD [39]	2.63	3.71	4.73
	FrSAD [40]	3.11	4.17	7.28
	MF-SAD [30]	3.83	5.29	9.12
	SE-SAD [41]	4.05	5.96	10.01
	WnF [39]	5.83	7.11	12.94

Boldface indicates best objective result for a competing method

4.2 Objective evaluation of estimated noise PSD

To illustrate the performance of noise PSD tracking among the noise PSD estimators, eight IEEE utterances from the NOIZEUS corpus ([14], Chapter 12) are concatenated and corrupted with 0 dB computer-generated non-stationary, i.e. Mod. WGN noise. Figure 4 shows the noise PSD estimates obtained by each method for a particular frequency bin, $m = 28$, corresponding to the DFT band around 875 Hz. It can be seen that the proposed method effectively track the slowly varying noise PSD, i.e. between 6.5 s and 9 s or 11.5 s and 14 s (Fig. 4a). Most of the competing methods failed to track this slowly varying noise PSD properly, except NPP-DNN [27] and IMMSE [25]. On the other hand, the proposed method is also found to be effective in tracking the abrupt rising of noise PSD, e.g. between 15 and 17 s (changing at a rate of 12 dB/s), while the competing methods are found to be inappropriate in tracking this noise PSD, except NPP-DNN [27]. Figure 5 also compares the noise PSD estimates obtained by each method, where the same concatenated

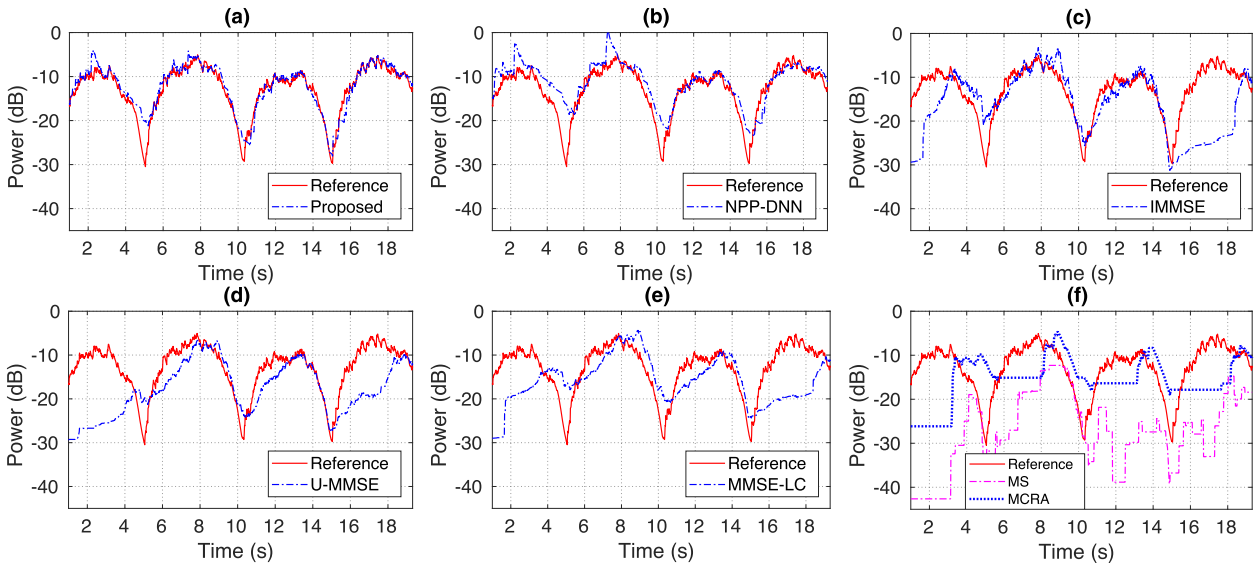

Fig. 4 Comparing the estimated noise PSD at frequency bin $m = 28$ on 8 concatenated utterances of NOIZEUS corpus ([14], Chapter 12) corrupted with 0 dB Mod. WGN noise using: **a** proposed (*LogErr* (dB)=1.22), **b**–**f** competing (*LogErr* (dB)= 2.01, 5.38, 5.77, 5.52, 6.65, 14.43) methods

sentence is corrupted with 0 dB real-world non-stationary, i.e. *passing train* noise. It can be seen that the proposed method still tracks the slowly varying noise PSD (e.g. between 7 and 14 s), which is closely similar to that of the reference (Fig. 5a). Amongst the benchmark methods, NPP-DNN [27], IMMSE [25], and U-MMSE [22] track this slowly varying noise PSD relatively well than MMSE-LC [21] and MCRA [18], while MS [17] completely failed to track this noise PSD. During the abrupt rising of noise, e.g. between 1 and 7 s at a rate of 5 dB/s, the proposed method is also found to be effective in tracking the noise PSD. Amongst the competing methods, NPP-DNN [27] relatively tracks a part of this abrupt changing noise PSD as compared to IMMSE [25] and U-MMSE [22]; however, MMSE-LC [21], MCRA [18], and MS [17] completely failed to track this noise PSD. Usually, MCRA [18] and MS [17] methods prone to an under-estimation of noise PSD due to using minimum statistics principle, whilst

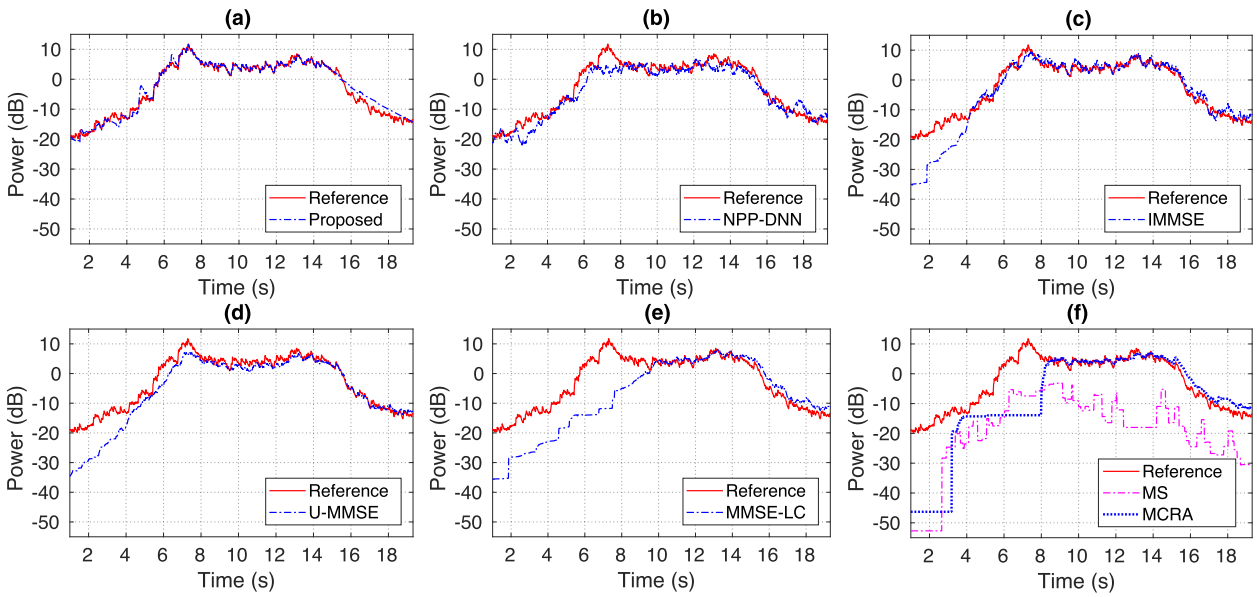

Fig. 5 Comparing the estimated noise PSD at frequency bin $m = 28$ on 8 concatenated utterances of NOIZEUS corpus ([14], Chapter 12) corrupted with 0 dB passing train noise using **a** proposed (*LogErr* (dB)=1.30), **b**–**f** competing (*LogErr* (dB)=2.13, 3.35, 3.76, 7.15, 8.57, 17.17) methods

the MMSE-based methods [21, 22, 25] give an underestimated noise PSD when the reference noise PSD rises abruptly than that of its slowly variations. Overall, the proposed method is found to be consistent in tracking both the fast-changing as well as the slowly varying reference noise PSD across the tested conditions.

The performance of noise PSD tracking is also evaluated in terms of the average LogErr (dB) distortion measure. Due to this purpose, Fig. 6 shows the average LogErr (dB) for each noise PSD estimator over the noisy speech dataset in Section 3.1. The lower shaded area of the bar in Fig. 6 indicates the measure $\text{LogErr}_{\text{ov}}$, caused due to the noise PSD overestimation, and the upper part corresponds to the measure of $\text{LogErr}_{\text{un}}$, caused due to noise PSD underestimation. The lower LogErr indicates a better tracking of the noise PSD. It can be seen from Fig. 6 that the proposed method demonstrates a lower average LogErr (dB) for most of the tested conditions. Amongst the competing methods, NPP-DNN [27] shows competitive LogErr (dB) with the proposed method and even exhibits bit higher LogErr (dB) for some conditions, such as 10 dB SNR levels in Fig. 6a, b. Amongst the MMSE-based methods, IMMSE [25] shows lower LogErr (dB) than U-MMSE [22] and MMSE-LC [21], whilst MS-based methods, such as MCRA and MS [17, 18], exhibit worse LogErr (dB) for all tested conditions. As demonstrated in earlier experiments (Figs. 4 and 5), MCRA and MS methods [17, 18] were developed with the principle of minimum statistics—which usually prone to an underestimation of noise PSD. As a result, the average LogErr for MCRA and MS methods [17, 18] become significantly higher than that of MMSE-based methods [21, 22, 25], apart from NPP-DNN [27] and proposed methods. Overall, the consistent lower LogErr across the wide range of

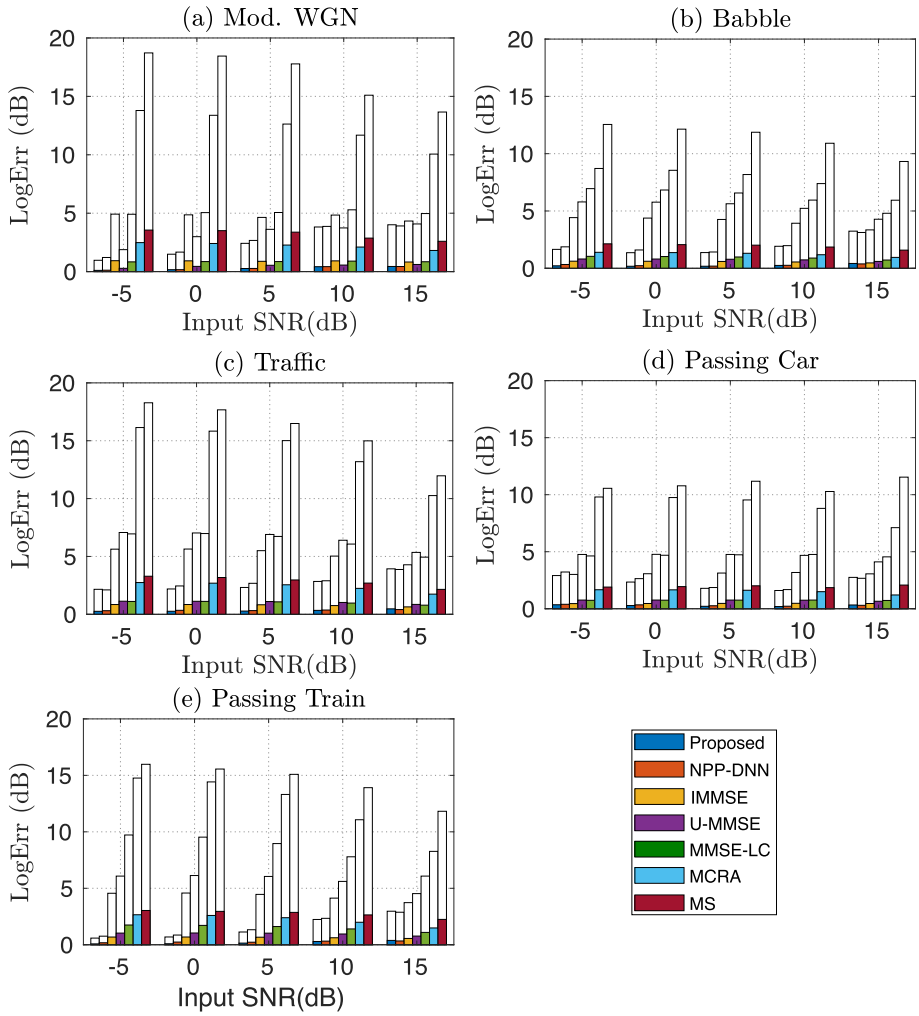

Fig. 6 Average LogErr (dB) comparison for each noise PSD tracking method over the noisy speech dataset in Section 3.1. The lower part of the bars represents $LogErr_{ov}$ (dB) and the upper part in white represents $LogErr_{un}$ (dB)

non-stationary noise conditions achieved by the proposed method indicates the superiority of noise PSD tracking over the competing methods.

4.3 Computational complexity evaluation of noise PSD estimators

Computation cost is also an important measure to justify the efficiency of a noise PSD tracking method, particularly in non-stationary noise conditions. The computational complexity in terms of normalized processing time of Matlab implementation [21, 22, 25] for all methods is given in Table 2. It can be seen that the proposed method takes the lowest computational time as compared to the competing methods. Amongst the competing methods, NPP-DNN [27] takes the next lowest computational time (1.08—excluding the training time of DNN prior to noise PSD estimation) with U-MMSE [22] (1.12). It is also found that the MCRA method [18] becomes computationally worse than any other methods. It is due to the computation of the smoothed noisy speech periodogram across the neighbouring frequency bin for all frames, as well as the minimum tracking during SPP estimation increase the computation cost significantly. In comparison, IMMSE [25] is found to be next to worse in computational complexity amongst the competing methods.

The lowest computational cost of the proposed method indicates the minimum tracking delay of the abrupt rising noise PSD. The lower LogErr (dB) of the proposed method also proves the significance of noise PSD tracking in non-stationary noise conditions. However, it is also important to compare the noise PSD tracking efficiency in speech enhancement context, since the lower LogErr (dB) does not guarantee better speech enhancement performance. Therefore, the following Sections describe the objective and subjective evaluation of enhanced speech, where the noise PSD estimated by the competing methods are incorporated in the DD approach to utilize the a priori SNR estimation of the MMSE-based speech enhancement systems in Section 2.4.

Table 2 Comparing the normalized processing time between the proposed and competing noise PSD tracking methods

Methods	Normalized Processing Time
Proposed	1.00
NPP-DNN [27]	1.08
IMMSE [25]	7.21
U-MMSE [22]	1.12
MMSE-LC [21]	4.87
MCRA [18]	44.52
MS [17]	3.73

4.4 Objective quality evaluation of enhanced speech

Figure 7 shows the average PESQ score for MMSE-based speech enhancement systems (Section 2.4). The experiments were conducted on the noisy speech data set in Section 3.1. It can be seen from Fig. 7a that the SRWF-based SEA with the proposed noise PSD estimator demonstrates consistent PESQ score improvement than the competing methods for most of the tested conditions. Amongst the competing methods, NPP-DNN [27] produces a very competitive PESQ score with the proposed method, particularly at high SNR levels, such as 10 dB SNR level of *mod. WGN* and *passing car* noise sources (Fig. 7a). The average PESQ scores of the MMSE-STSA [4] and MMSE-LSA [5] methods using the proposed noise PSD estimator (Fig. 7b, c) also show improvement to that of the competing noise PSD estimators. Amongst the competing methods, NPP-DNN [27] produces a competitive PESQ score with the proposed methods at a high SNR level only. Except NPP-DNN [27], the other competing methods produce significantly lower PESQ scores for all tested conditions (Fig. 7a–c). This is due to the lacking of the noise PSD tracking capability by the classical methods, which impact the a priori SNR estimation of the MMSE-speech enhancement systems in Section 2.4. Conversely, due to showing better noise PSD tracking capability by proposed and NPP-DNN methods, both of the methods produce better PESQ scores when applied in speech enhancement contexts. Overall, the comparative study reveals that the proposed method produces better quality enhanced than the NPP-DNN method [27].

We also compare the improvement of average PESQ scores (taking the average of PESQ scores for all tested conditions) for the proposed noise PSD estimator with the competing methods. It can be seen from Table 3 that the proposed noise PSD estimator also produced higher PESQ scores than any of the competing methods when incorporated in MMSE-based speech enhancement system in Section 2.4 (with an improvement of 0.06 for SRWF, 0.04 for MMSE-STSA, and 0.09 for MMSE-LSA over the next best noise PSD estimator, NPP-DNN [27]). It is also observed that the average PESQ improvement of the proposed noise PSD estimator is consistently increasing as compared to the IMMSE [25], U-MMSE[22], MMSE-LC [21], MCRA [18], and MS [17] noise PSD estimators.

4.5 Objective intelligibility evaluation of enhanced speech

Figure 8 shows the average STOI score for MMSE-based speech enhancement systems (Section 2.4). The experiments were conducted on the same noisy speech data set as in Fig. 7. It can be seen from Fig. 8a that the SRWF-based SEA with the proposed noise PSD estimator exhibits consistent average STOI score improvement than the competing noise PSD estimators across the tested

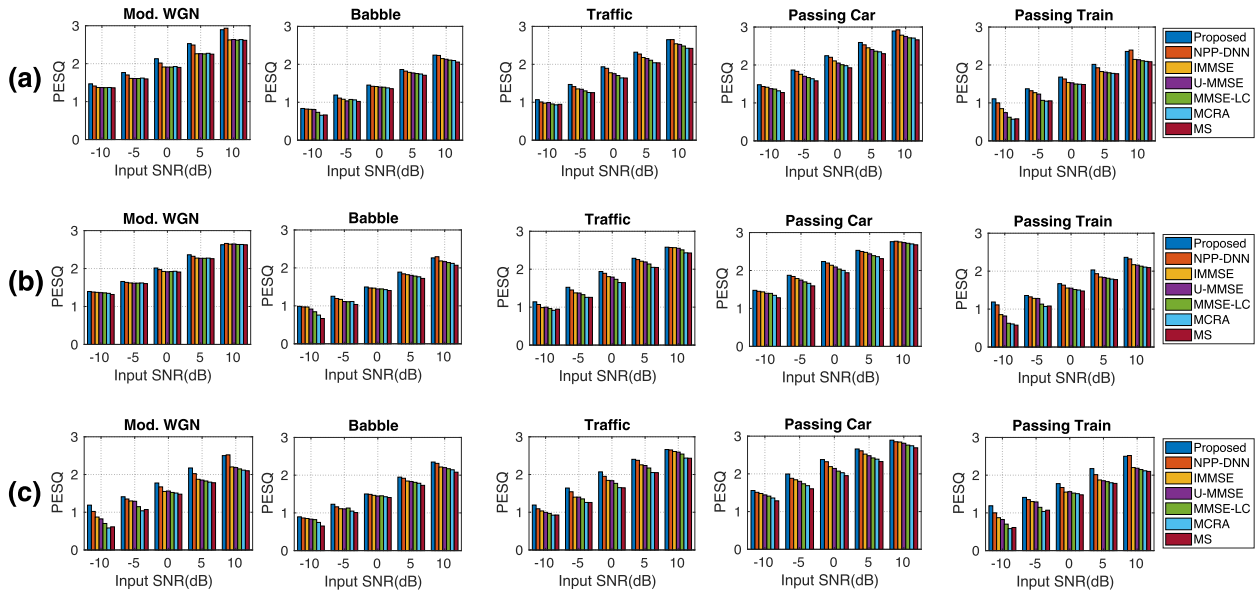

Fig. 7 Average PESQ score comparison on the NOIZEUS corpus corrupted with Mod. WGN, babble, traffic, passing car, and passing train noises at multiple SNR levels (from −10 to +10dB, in 5 dB increments), where the estimated noise PSD by the competing methods are used in the DD approach to compute the a priori SNR for utilizing the **a** SRWF ([14], Section 6.5.1 of Chapter 6), **b** MMSE-STSA [4], and **c** MMSE-LSA [5] gain functions in speech enhancement context

conditions. Amongst the competing methods, NPP-DNN [27] produces competitive STOI scores with the proposed method, particularly at high SNR levels. The average STOI score for the MMSE-STSA [4] and MMSE-LSA [5] based SEAs with the proposed noise PSD estimator (Fig. 8b, c) also demonstrates consistence improvement to that of the competing noise PSD estimators for most of the tested conditions. Amongst the competing methods, NPP-DNN [27] produces the competing STOI scores with the proposed method (Fig. 8b, c).

As in Section 4.4, we also compare the improvement of average STOI scores (taking the average of STOI scores for all tested conditions) for the proposed noise PSD estimator with the competing methods. It can be seen from Table 4 that the proposed noise PSD estimator produced a higher STOI score than any of the competing methods (with an improvement of 2.11% for SRWF, 0.78% for MMSE-STSA, and 1.12% for MMSE-LSA over the next best method, NPP-DNN [27]). It is also observed that the average STOI improvement of the proposed noise PSD estimator is consistently increasing as compared to the IMMSE [25], U-MMSE[22], MMSE-LC [21], MCRA [18], and MS [17] noise PSD estimators.

In light of the experiment in Figs. 7 and 8, it is evident to say that the proposed noise PSD estimator produces a higher quality and intelligibility in the enhanced speech when incorporated in MMSE-based SEAs than the competing noise PSD estimators. We also analyse the average PESQ and STOI (%) score improvements of the SRWF-based SEA against MMSE-STSA and MMSE-LSA based SEAs, where the proposed noise PSD estimator is incorporated. It can be seen from Table 5 that the SRWF method produced higher PESQ and STOI (%) scores with an improvement of 2.42 and 5.19% for MMSE-STSA and 0.02 and 3.57% for MMSE-LSA. In light of the study, it is evident to say that the proposed noise PSD estimator with SRWF-based SEA shows better speech enhancement performance than that of the MMSE-LSA and MMSE-STSA based SEAs. Therefore, the following two sections perform the spectrogram analysis and the subjective evaluation of the enhanced speech produced by the SRWF-based SEA by incorporating the proposed and competing noise PSD estimators.

Table 3 The average PESQ score improvement of the proposed noise PSD estimator against the competing methods, when incorporated in the MMSE-based speech enhancement systems

	Speech enhancement methods		
Noise PSD estimators	**SRWF**	**MMSE-STSA**	**MMSE-LSA**
Proposed Vs.			
NPP-DNN [27]	0.06	0.04	0.09
IMMSE [25]	0.13	0.07	0.16
U-MMSE [22]	0.15	0.10	0.19
MMSE-LC [21]	0.19	0.14	0.24
MCRA [18]	0.21	0.17	0.30
MS [17]	0.23	0.20	0.33

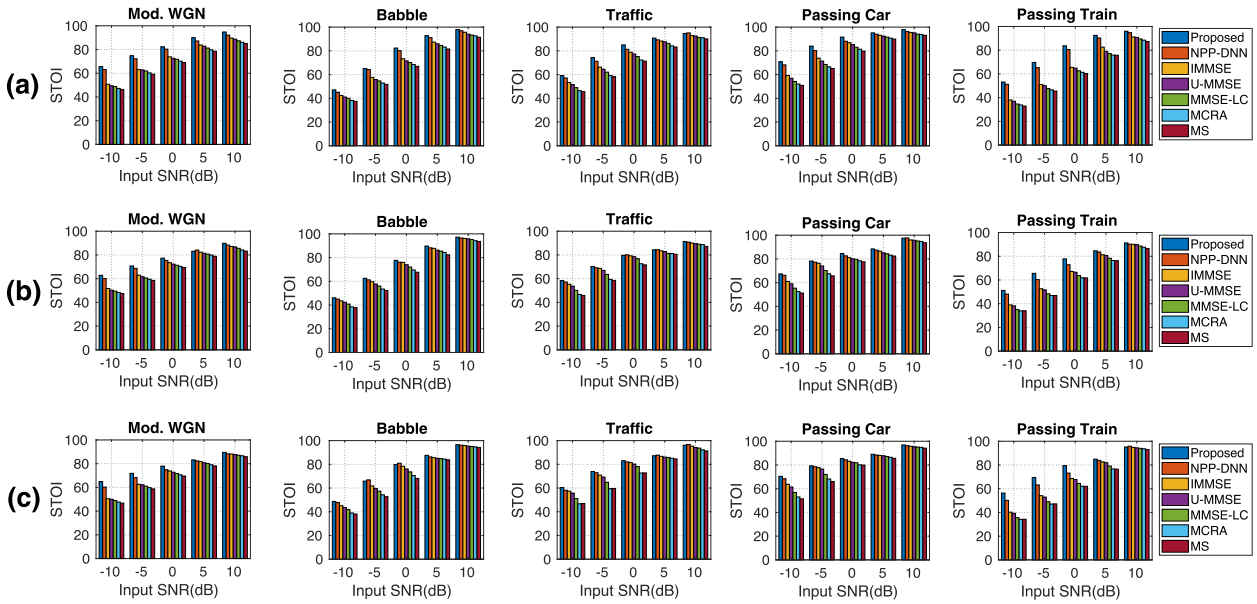

Fig. 8 Average STOI score comparison on NOIZEUS corpus corrupted with Mod. WGN, babble, traffic, passing car, and passing train noises at multiple SNR levels (from −10 to +10dB, in 5 dB increments), where the estimated noise PSD by the competitive methods are used in the DD approach to compute the a priori SNR for utilizing the **a** SRWF ([14], Section 6.5.1 of Chapter 6), **b** MMSE-STSA [4], and **c** MMSE-LSA [5] gain functions in speech enhancement context

4.6 Spectrogram analysis of enhanced speech

Figure 9 compares the spectrogram of the enhanced speech produced by SRWF-based SEA, which incorporates the competing noise PSD estimators. The experiment is conducted on a concatenation of sp05 (male) and sp12 (female) utterances corrupted with 5 dB *passing train* noise. It can be seen that the enhanced speech produced by SRWF-based SEA with the proposed noise PSD estimator (Fig. 9i) shows a significant reduction of *residual background* noise as compared to other noise PSD estimators, apart from the clean speech (Fig. 9a). Specifically, the spectrogram produced by SRWF-based SEA with the proposed noise PSD estimator contains less noise floor to that of the significant *residual backround* noise and *speech distortion* by other noise PSD estimators. Amongst the competing methods, the notable noise floor is found in the enhanced speech produced by NPP-DNN-based noise PSD estimator [27] (Fig. 9h). The rest of the methods suffer from significant *residual background* noise (Fig. 9c-g).

Table 4 The average STOI score (%) improvement of the proposed noise PSD estimator against the competing methods, when incorporated in the MMSE-based speech enhancement systems

	Speech enhancement methods		
Noise PSD estimators	**SRWF**	**MMSE-STSA**	**MMSE-LSA**
Proposed Vs.			
NPP-DNN [27]	2.11	0.78	1.12
IMMSE [25]	8.12	1.60	2.01
U-MMSE [22]	9.11	2.50	2.94
MMSE-LC [21]	10.60	4.49	5.19
MCRA [18]	11.93	6.32	7.61
MS [17]	12.41	6.82	8.22

4.7 Subjective evaluation by AB listening test

Figure 10 shows the mean subjective preference score (%) comparison for each method over the stimuli set in Section 3.3. It can be seen that the enhanced speech produced by the proposed method is widely preferred by the listeners (82.86%) to that of the competing methods, apart from the clean speech (100%). Amongst the competing methods, NPP-DNN [27] shows very competitive score (81.73%) with the proposed method followed by IMMSE [25] (70.22%), U-MMSE [22] (56.71%), MMSE-LC [21] (39.31%), MCRA [18] (29%), and MS [17] (17.5%). In light of the series of blind AB listening tests, it is evident to say that the enhanced speech produced by the proposed

Table 5 The average PESQ and STOI (%) scores improvement of the SRWF based SEA with proposed noise PSD estimator to that of the MMSE-STSA and MMSE-LSA based SEAs

Methods	**PESQ**	**STOI**
SRWF Vs. MMSE-STSA	2.42	5.19%
SRWF Vs. MMSE-LSA	0.02	3.57%

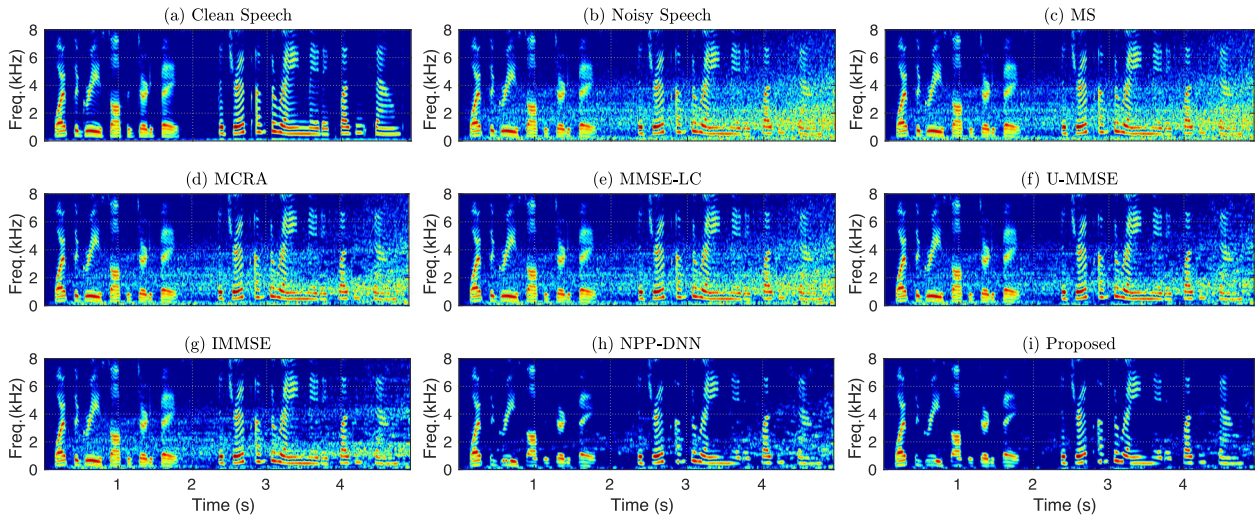

Fig. 9 Comparing the spectrograms of **a** clean speech (concatenate sp05 and sp12) and **b** noisy speech (corrupt (**a** with 5 dB *passing train* noise (PESQ=1.80), to that of the enhanced speech produced by the SRWF-based SEA ([14], Section 6.5.1 of Chapter 6), where the noise PSD estimates are given by **c** MS [17] (PESQ=1.82), **d** MCRA [18] (PESQ=1.83), **e** MMSE-LC [21] (PESQ=1.85), **f** U-MMSE [22] (PESQ=1.93), **g** IMMSE [25] (PESQ=1.92), **h** NPP-DNN [27] (PESQ=2.01), and **i** Proposed (PESQ=2.13) methods

method exhibits a significant improvement in perceived quality as compared to the benchmark methods, except the NPP-DNN [27] for the tested condition specified in Section 3.3.

5 Conclusion

This paper presents a noise PSD estimation algorithm using derivative-based *high-pass* filter in non-stationary noise conditions. Firstly, a spectral-flatness based adaptive thresholding technique detects the speech activity for each noisy speech frame. Since the noisy speech in the silent frame gives an estimate of noise, the noise periodogram is directly computed from it. Conversely, the application of a 4^{th} order derivative-based *high-pass* filter to the noisy speech frame during speech presence filtered out the clean speech components while leaving behind

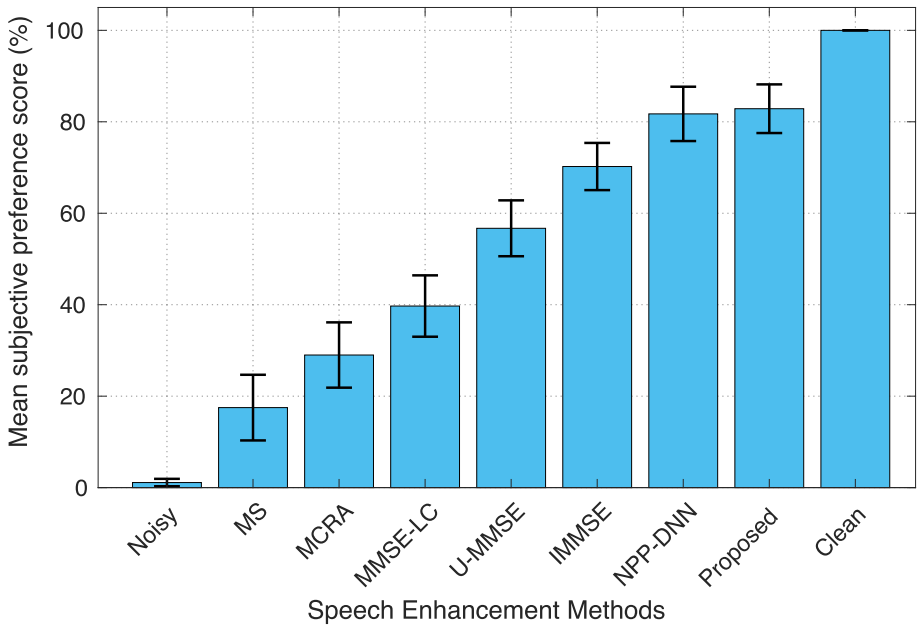

Fig. 10 The mean subjective preference score (%) comparison of the enhanced speech produced by SRWF-based SEA, where the competing noise PSD estimators have been incorporated. The experiments were conducted over the stimuli set in Section 3.3. The error bars indicate the standard deviation of the scores

mostly the noise. The noise periodogram is computed from the filtered signal—which mitigates the risk of leaking speech power. The noise PSD estimate is obtained by recursively averaging the past estimated noise PSD and the current estimate of noise periodogram weighted by a smoothing constant. Experimental results demonstrate that the proposed noise PSD estimator outperforms in tracking the rapidly changing as well as the slowly varying noise PSD than the competing methods in non-stationary noise conditions for a wide range of SNR levels. Extensive objective and subjective scores also reveal that the MMSE-based SEAs with the proposed noise PSD estimator produced higher quality and intelligible enhanced speech than the competing noise PSD estimators.

Abbreviations

PSD: Power spectral density; SNR: Signal-to-noise ratio; MMSE: Minimum mean square error; SEA: Speech enhancement algorithm; VAD: Voice activity detector; MS: Minimum statistics; SPP: Speech presence probability; MCRA: Minima controlled recursive averaging; IMCRA: Improved minima controlled recursive averaging; ML: Maximum-likelihood; DD: Decision-directed; STFT: Short-time Fourier transform; PESQ: Perceptual evaluation of speech quality

Acknowledgements

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

Authors' information

Not applicable

Authors' contributions

SK performed the entire research and its writing and KK supervised the research. Both authors read and approved the final manuscript.

Funding

Not applicable

Availability of data and materials

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 21 April 2021 Accepted: 1 August 2021
Published online: 14 August 2021

References

1. S. Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. **27**, 113–120 (1979). https://doi.org/10.1109/TASSP.1979.1163209
2. M. Berouti, R. Schwartz, J. Makhoul, Enhancement of speech corrupted by acoustic noise. IEEE Int. Conf. Acoust. Speech Signal Process. **4**, 208–211 (1979). https://doi.org/10.1109/TASSP.1979.1163209
3. S. Kamath, P. Loizou, A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. IEEE Int. Conf. Acoust. Speech Signal Process. **4**, 4160–4164 (2002). https://doi.org/10.1109/ICASSP.2002.5745591
4. Y. Ephraim, D. Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **32**(6), 1109–1121 (1984). https://doi.org/10.1109/TASSP.1984.1164453
5. Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **33**(2), 443–445 (1985). https://doi.org/10.1109/TASSP.1985.1164550
6. P. Scalart, J. V. Filho, Speech enhancement based on a priori signal to noise estimation. IEEE Int. Conf. Acoust. Speech Signal Process. **2**, 629–632 (1996). https://doi.org/10.1109/ICASSP.1996.543199
7. C. Plapous, C. Marro, L. Mauuary, P. Scalart, A two-step noise reduction technique. IEEE Int. Conf. Acoust. Speech Signal Process. **1**, 289–292 (2004). https://doi.org/10.1109/ICASSP.2004.1325979
8. C. Plapous, C. Marro, P. Scalart, Speech enhancement using harmonic regeneration. IEEE Int. Conf. Acoust. Speech. Signal Process. **1**, 157–160 (2005). https://doi.org/10.1109/ICASSP.2005.1415074
9. K. Paliwal, K. Wójcicki, B. Schwerin, Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. Speech Commun. **52**(5), 450–475 (2010). https://doi.org/10.1016/j.specom.2010.02.004
10. M. S. Kavalekalam, J. K. Nielsen, M. G. Christensen, J. B. Boldt, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A study of noise psd estimators for single channel speech enhancement, (2018), pp. 5464–5468. https://doi.org/10.1109/ICASSP.2018.8461703
11. J. Sohn, W. Sung, A voice activity detector employing soft decision based noise spectrum adaptation. IEEE Int. Conf. Acoust. Speech. Signal Process. **1**, 365–368 (1998). https://doi.org/10.1109/ICASSP.1998.674443
12. J. Sohn, N. S. Kim, W. Sung, A statistical model-based voice activity detection. IEEE Signal Process. Lett. **6**(1), 1–3 (1999). https://doi.org/10.1109/97.736233
13. J. H. Chang, N. S. Kim, S. K. Mitra, Voice activity detection based on multiple statistical models. IEEE Trans. Signal Process. **54**(6), 1965–1976 (2006). https://doi.org/10.1109/TSP.2006.874403
14. P. C. Loizou, *Speech Enhancement: Theory and Practice, 2nd ed.* (CRC Press, Inc., Boca Raton, 2013)
15. R. Martin, Spectral subtraction based on minimum statistics. European Signal Processing Conference (EUSIPCO), 1182–1185 (1994)
16. G. Doblinger, Computationally efficient speech enhancement by spectral minima tracking in subbands. Eur. Conf. Speech Commun. Technol. (EUROSPEECH), 1513–1516 (1995). https://doi.org/10.1016/j.specom.2005.08.005
17. R. Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. **9**(5), 504–512 (2001). https://doi.org/10.1109/89.928915
18. I. Cohen, B. Berdugo, Noise estimation by minima controlled recursive averaging for robust speech enhancement. IEEE Signal Process. Lett. **9**(1), 12–15 (2002). https://doi.org/10.1109/97.988717
19. I. Cohen, Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process. **11**(5), 466–475 (2003). https://doi.org/10.1109/TSA.2003.811544
20. S. Rangachari, P. C. Loizou, A noise-estimation algorithm for highly non-stationary environments. Speech Commun. **48**(2), 220–231 (2006). https://doi.org/10.1016/j.specom.2005.08.005
21. R. C. Hendriks, R. Heusdens, J. Jensen, MMSE based noise PSD tracking with low complexity. IEEE Int. Conf. Acoust. Speech Signal Process., 4266–4269 (2010). https://doi.org/10.1109/ICASSP.2010.5495680
22. T. Gerkmann, R. C. Hendriks, Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans. Audio Speech Lang. Process. **20**(4), 1383–1393 (2012). https://doi.org/10.1109/TASL.2011.2180896
23. T. Gerkmann, R. C. Hendriks, Bayesian noise estimation in the modulation domain. Speech Commun. **96**, 81–92 (2018). https://doi.org/10.1016/j.specom.2017.11.008
24. J. K. Nielsen, M. S. Kavalekalam, M. G. Christensen, J. Boldt, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Model-based noise PSD estimation from speech in non-stationary noise, (2018), pp. 5424–5428. https://doi.org/10.1109/ICASSP.2018.8461683
25. Q. Zhang, M. Wang, Y. Lu, L. Zhang, M. Idrees, A novel fast nonstationary noise tracking approach based on MMSE spectral power estimator. Digit. Signal Process. **88**, 41–52 (2019). https://doi.org/10.1016/j.dsp.2019.01.019
26. Q. Zhang, M. Wang, Y. Lu, M. Idrees, L. Zhang, Fast nonstationary noise tracking based on log-spectral power mmse estimator and temporal recursive averaging. IEEE Access. **7**, 80985–80999 (2019). https://doi.org/10.1109/ACCESS.2019.2923680
27. A. Chinaev, J. Heymann, L. Drude, R. Haeb-Umbach, in *Speech Communication; 12. ITG Symposium*. Noise-presence-probability-based noise PSD estimation by using DNNs, (2016), pp. 1–5

28. Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, C. Wang, DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 1404–1415 (2020). https://doi.org/10.1109/TASLP.2020.2987441
29. A. V. Oppenheim, R. W. Schafer, *Discrete-Time Signal Processing, 3rd ed.* (Prentice Hall Press, Upper Saddle River, 2009)
30. M. H. Moattar, M. M. Homayounpour, in *17th European Signal Processing Conference*. A simple but efficient real-time voice activity detection algorithm, (2009), pp. 2549–2553
31. N. Madhu, Note on measures for spectral flatness. Electron. Lett. **45**(23), 1195–1196 (2009). https://doi.org/10.1049/el.2009.1977
32. S. Graf, T. Herbig, M. Buck, G. Schmidt, Features for voice activity detection: a comparative analysis. EURASIP J. Adv. Signal Process. **2015**(1), 91 (2015). https://doi.org/10.1186/s13634-015-0277-z
33. Free Sound Database. https://freesound.org/. Accessed October 2019
34. J. Ogrodzki, *Circuit Simulation Methods and Algorithms, 1st ed.* (CRC Press, Inc., Boca Raton, 1994)
35. T. ÓHaver, Fourier Convolution. https://terpconnect.umd.edu/~toh/spectrum/Convolution.html. Accessed December 2019
36. M. K. I. Molla, K. Hirose, S. K. Roy, S. Ahmad, in *2011 IEEE International Symposium of Circuits and Systems (ISCAS)*. Adaptive thresholding approach for robust voiced/unvoiced classification, (2011), pp. 2409–2412. https://doi.org/10.1109/ISCAS.2011.5938089
37. A. W. Rix, J. G. Beerends, M. P. Hollier, A. P. Hekstra, Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. IEEE Int. Conf. Acoust. Speech. Signal Process. **2**, 749–752 (2001). https://doi.org/10.1109/ICASSP.2001.941023
38. C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2125–2136 (2011). https://doi.org/10.1109/TASL.2011.2114881
39. M. K. I. Molla, K. Hirose, N. Minematsu, in *Inter Speech*. Robust voiced/unvoiced speech classification using empirical mode decomposition and periodic correlation model, (2008), pp. 2530–2533
40. I.-C. Yoo, H. Lim, D. Yook, Formant-based robust voice activity detection. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(12), 2238–2245 (2015). https://doi.org/10.1109/TASLP.2015.2476762
41. J. Pang, in *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)*. Spectrum energy based voice activity detection, (2017), pp. 1–5. https://doi.org/10.1109/CCWC.2017.7868454

Publisher's Note