IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 9, 2019, accepted June 6, 2019, date of publication June 19, 2019, date of current version July 3, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2923680

Fast Nonstationary Noise Tracking Based on Log-Spectral Power MMSE Estimator and Temporal Recursive Averaging

QIQUAN ZHANG, MINGJIANG WANG, YUN LU, MUHAMMAD IDREES, AND LU ZHANG
School of Electronic and Information Engineering, Harbin Institute of Technology at Shenzhen, Shenzhen 518055, China

Corresponding author: Mingjiang Wang (mjwang@hit.edu.cn)

This work was supported by the Basic Research Discipline Layout Project of Shenzhen under Grant JCYJ20170412151226061 and Grant JCYJ20180507182241622.

ABSTRACT Estimation of the noise power spectral density (PSD) plays a critical role in most existing single-channel speech enhancement algorithms. In this paper, we present a novel noise PSD tracking algorithm, which employs a log-spectral power minimum mean square error (MMSE) estimator. This method updates the noise PSD estimate by performing a temporal recursive averaging of log-spectral MMSE estimate of the current noise power to reduce the risk of speech leakage into noise estimate. A smoothing parameter used in the recursive operation is adjusted by speech presence probability (SPP). In this method, a spectral nonlinear weighting function is derived to estimate the noise spectral power which depends on the *a priori* and the a posteriori signal-to-noise ratio (SNR). An extensive performance comparison has been carried out with several state-of-the-art noise tracking algorithms, i.e., Minimum Statistics (MS), modified minima controlled recursive averaging algorithm (MCRA-2), MMSE-based method, and SPP-based method. It is clear from experimental results that the proposed algorithm exhibits more excellent noise tracking capability under various nonstationary noise environments and SNR levels. When employed in a speech enhancement framework, improved speech enhancement performance in terms of the segmental SNR (segSNR) improvements and three objective composite metrics is observed.

INDEX TERMS Acoustic noise, speech enhancement, noise PSD estimation, log-spectral, minimum mean-square error (MMSE) estimator, speech presence probability.

I. INTRODUCTION

Speech is one of the most important forms of human communication, which plays an important role in many applications such as mobile communications, digital hearing aids and human-computer interactions. However, in practical scenarios, clean speech signals will always, to some extent, be degraded by surrounding interference noises. In most situations, the interfering noise is usually nonstationary. The nonstationary interference noise will bring great challenges to speech signal processing applications. In human-computer interaction (e.g., automatic speech recognition), for instance, the degraded speech leads to a significant decrease of recognition accuracy. As a consequence, noise suppression technology [1]–[7] is of great importance, the aim of which is to suppress the disturbing noise component in noisy speech while preserving the original quality and intelligibility of clean speech. Single-channel noise suppression approaches [8]–[10] based on short-time Fourier transform (STFT, a sequence of Fourier transforms of a windowed signal) are often used to achieve this.

Noise power spectral density (PSD) is defined as the noise power per unit bandwidth. Noise PSD estimation is a crucial component in designing single-channel speech enhancement algorithms [11]–[17]. An underestimation of the noise PSD leads to an unnecessary amount of residual noise in the enhanced signal, while an overestimation introduces speech distortions, which may result in a loss of speech intelligibility. A conventional noise PSD method is to exploit a voice activity detector (VAD) [18]–[21] to identify speech pause periods, and then the noise PSD estimate is updated during speech absence. Although this is effective for highly

The associate editor coordinating the review of this manuscript and approving it for publication was Bora Onat.

stationary noise, it often fails at low SNR (where SNR is the ratio of signal power to the noise power) scenarios, especially when the noise is nonstationary. In past decades, a significant amount of work has been done to solve this problem. In general, most state-of-the-art methods for noise PSD estimation can be divided into four main groups [22], i.e., Minimum Statistics (MS) methods [23], [24], time-recursive averaging methods [25]–[27], subspace decomposition algorithms [28], [29], and other techniques based on Bayesian estimation principle [30]–[32].

In the first group of algorithms, the noise PSD is tracked via Minimum Statistics (MS) algorithms [23], [24], which rely on two assumptions that the noise and the speech are statistically independent, and that the power of the noisy speech signal frequently decays to the power level of the noise signal (e.g., in speech pauses). The noise PSD is estimated as the tracked minimum of the smoothed noisy spectrum within a finite time window. The expectation of the minima is smaller than the mean value of the spectral power, thus a bias compensation factor is derived to correct the bias [24]. Since MS method will result in speech leakage into noise PSD estimate when the time window is short, a sufficiently long time window is required to reduce the amount of speech leakage. Unfortunately, if the time window is chosen too long, fast noise level changes will be tracked with a rather large delay. Thus a trade-off is necessary, a typical size of window is in the order of 1 s. As the minimum value in a window is used, the noise PSD will always be underestimated or tracked with a large delay in case of increasing noise power level.

In the second category of algorithms, the noise PSD estimate is updated by recursively averaging the previous estimated noise PSD and the current noisy speech power spectrum, in which the smoothing factors are controlled by the speech presence probability (SPP). The representative methods of this class include minima controlled recursive averaging (MCRA) method [25] and its two modifications, i.e., improved MCRA (IMCRA) [26] and MCRA-2 [27]. The main distinction between MCRA, IMCRA and MCRA-2 is reflected in the way the SPP is calculated. In MCRA, the SPP is determined by the ratio of the smoothed noisy speech power spectrum to its local minimum obtained by minimum statistics technique [24], and for that reason this method is referred as the minima controlled recursive averaging (MCRA) algorithm. The presence of speech is detected when the ratio is above a certain fixed threshold. MCRA-2 employs the continuous spectral minimum-tracking algorithm [33] to obtain the minimum and is not constrained within a search window. Moreover, unlike the fixed threshold in MCRA, frequency-dependent thresholds are used in MCRA-2 to calculate the SPP. In IMCRA method, the SPP estimation is based on a Gaussian statistical model and obtained from the ratio of the likelihood functions of speech presence and speech absence. The derivation of IMCRA method involves two iterations of smoothing and minimum tracking. The first iteration provides a simple speech-presence detector for each frequency bin, while the second iteration of smoothing excludes high-energy speech components, thus allowing for smaller windows in minima tracking. However, since these approaches are proposed on the basis of the MS principle [24], they still show a considerable tracking delay in case of the increasing noise power level.

In the third family of methods, the decomposed noise-only subspace is used to update the noise PSD estimation. A famous subspace decomposition based approach, called subspace noise tracking (SNT) algorithm was proposed in [28]. The SNT is based on eigenvalue decompositions of correlation matrices that are constructed using time series of noisy discrete Fourier transform (DFT) coefficients. An improvement of this method, called minimum subspace noise tracking (MSNT) algorithm [29], exploits the limited-rank structure of the clean speech signal. MSNT combines the subspace structure and the minimum statistics tracking to estimate noise PSD. In comparison to, e.g., MS-based noise PSD trackers, the subspace decomposition based noise tracking algorithms allow for the faster noise tracking for many nonstationary noises [34]. However, the improved noise tracking performance of the subspace based noise trackers is accompanied by a significant increase in the computational complexity.

In the fourth group of methods, the derivation of the noise spectral power estimators is based on Bayesian estimation principle and assumed statistical model. In [30], [31], minimum mean square error (MMSE) estimator derived by minimizing the mean square error (MSE) of spectral power is used to estimate the instantaneous noise power and a first-order recursive smoothing technique is employed to update the noise PSD estimate. However, for noise power estimation, the simple bias compensation in [30] is motivated heuristically, whereas the bias compensation in [31] is derived rigorously based on assumed signal model. The SPP-based approach [32] is a further modification of the MMSE-based approach [31]. In the SPP-based method, the noise PSD estimate is obtained by the sum of the previous noise PSD estimate weighted by the conditional probability of speech presence and the periodogram of noisy speech weighted by the conditional probability of speech absence. These MMSE algorithms [31], [32] achieve fast noise spectral power tracking and are demonstrated to have a more robust noise estimation performance [34]. More recently, a model-based noise PSD estimation method was reported in [35] and [36], where different codebooks were trained for different noise and speech types. This model-based method [35] performs best for noise-types for which the algorithm is trained. However, since the number of models increases with the product of the codebook, this might lead to an intractable computational complexity.

Although the spectral mean-square error (MSE) distortion metric is mathematically tractable and also leads to good results in [30]–[32], it appears to be not perceptually meaningful. In fact, human ear has a logarithmic response to sound (whether speech or noise) intensity changes [37] and it is argued that a distortion metric based on the MSE of the log-

spectral is perceptually more relevant, and more appropriate for speech processing [38]. Based on such facts it was presented in [3] and [13] to estimate speech spectral amplitude by minimizing the log-spectral MSE. Recently, an algorithm was presented in [39] to track the speech and noise in the log-power spectral domain. Motivated by these facts, the noise is naturally regarded as "target" signal (not speech signal in [3], [13]), and we therefore exploit this distortion metric for noise estimation and develop a noise spectral power estimator that minimizes the MSE of log-spectral power. Moreover, speech estimators [3], [13] focus on reconstructing the instantaneous speech spectral amplitude, while noise tracking algorithms are interested in estimating the noise PSD (expectation of instantaneous noise spectral power). In this algorithm, the noise PSD estimation is obtained by recursively averaging the log-spectral MMSE estimate of the current noise power. The smoothing parameter is adjusted by the speech presence probability determined by the smoothed posteriori SNR. For the noise spectral power estimate, we derive a nonlinear spectral weighting function, which relies on the a priori and the a posteriori SNR. In this work, we consider the standard "decision-directed" (DD) estimator for the a priori SNR estimation. Experimental results show that for different nonstationary noises the proposed noise PSD tracker achieves a more accurate and rapid noise PSD estimate, and a better speech enhancement performance in terms of both the segmental SNR [22], [40] and three composite measures [41].

The remainder of this paper is organized as follows. Section II explains the used notation, and the signal model employed to derive the noise spectral power estimator. In Section III, we propose to employ Log-spectral MMSE estimate of noise power to recursively update the noise PSD estimate, which reduces the probability of speech leakage. Section IV gives a detailed derivation of the proposed Log-spectral MMSE noise power estimator. In Section V, we evaluate the performance of the proposed algorithm and make comparisons with four state-of-the-art methods, MS [24], MCRA-2 [27], MMSE-based algorithm [31], and SPP-based algorithm [32], in terms of tracking performance, and overall performance in a noise suppression framework. Conclusions are finally presented in Section VI.

II. SIGNAL MODEL AND NOTATION

Let $y(n)$ denotes a noisy speech signal, which consists of a clean speech signal $x(n)$ contaminated with additive noise signal $d(n)$, i.e., $y(n) = x(n) + d(n)$, where n is the discrete time index. The noisy signal $y(n)$ is segmented into overlapping frames, followed by windowing with a square-root-Hann window. Subsequently, each frame is transformed by applying the short-time Fourier transform (STFT). The noisy speech signal in the time-frequency domain is expressed as

$$Y(l,k) = X(l,k) + D(l,k) \tag{1}$$

where $X(l,k)$ and $D(l,k)$ represent the complex STFT coefficients of the clean speech and additive noise term, respectively. Furthermore, l is the frame index and k is the frequency index. It is assumed that $X(l,k)$ and $D(l,k)$ are conditionally independent across time and frequency, and obey zero-mean complex Gaussian distributions with model parameters $E\left\{|X(l,k)|^2\right\} = \lambda_x(l,k)$ and $E\left\{|D(l,k)|^2\right\} = \lambda_d(l,k)$, respectively, where $E\{\cdot\}$ denotes the statistical expectation operator. $\lambda_x(l,k)$ and $\lambda_d(l,k)$ denote the PSDs (or variances) of the speech and the noise signals, respectively. In the sequel, the indexes l and k will be omitted for simplicity, whenever it is possible. The STFT coefficients can be represented in terms of their amplitude and phase, denoted as $Y = Re^{j\alpha}$, $X = Ae^{j\beta}$, and $D = Ne^{j\theta}$. We will call N^2 the (instantaneous) noise spectral power.

Further, we use the terms a priori SNR ξ and the a posteriori SNR γ, defined as

$$\xi = \frac{\lambda_x}{\lambda_d} \quad \text{and} \quad \gamma = \frac{R^2}{\lambda_d}, \tag{2}$$

respectively. A hat symbol is used to denote the estimated quantities of variables, e.g., $\hat{N}^2$ is an estimator of noise spectral power N^2.

III. TEMPORAL RECURSIVE SMOOTHING OF NOISE LOG-SPECTRAL POWER MMSE ESTIMATION

The temporal-recursive averaging algorithms, MCRA [25], IMCRA [26], obtain the noise PSD estimation by recursively smoothing the noisy speech spectral power R^2 [32], i.e.,

$$\hat{\lambda}_d(l,k) = \alpha_N(l,k)\hat{\lambda}_d(l-1,k) + (1-\alpha_N(l,k))R^2(l,k). \tag{3}$$

As the minimum values in a long time window are used to avoid speech leakage into noise PSD estimate, these methods show a slow response to fast increases in noise level [42]. In this paper, the noise PSD is estimated by recursively averaging a (instantaneous) noise spectral power estimator $\hat{N}^2$ instead of noisy spectral power R^2, given by

$$\hat{\lambda}_d(l,k) = \alpha_N(l,k)\hat{\lambda}_d(l-1,k) + (1-\alpha_N(l,k))\hat{N}^2(l,k). \tag{4}$$

Compared to recursive averaging technique with fixed smoothing factor, SPP-based recursive averaging technique is a more general and widely used method. Similar to MCRA and IMCRA, the time-varying smoothing parameter $\alpha_N(l,k)$ is also adjusted by an estimate $\hat{p}(l,k)$ of SPP

$$\alpha_N(l,k) = \alpha_n + (1-\alpha_n)\hat{p}(l,k) \tag{5}$$

where $\alpha_n (0 < \alpha_n < 1)$ is a smoothing parameter which usually has a value range of [0.8, 0.95] as suggested in [22] and is empirically set to 0.8 in this work. Utilizing noise spectral power estimate $\hat{N}^2$ instead of noisy spectral power R^2 has the benefit of reducing the amount of speech component leaking into noise PSD estimate. Therefore, an extremely accurate SPP estimator is not necessary. For $\hat{N}^2$ the Log-spectral MMSE estimator of the noise power N^2 is exploited. Different from IMCRA, this work uses a simpler estimation method for $\hat{p}(l,k)$ that allows for faster tracking.

A. SPEECH PRESENCE PROBABILITY ESTIMATION

Since the noise PSD estimate is updated with noise spectral power estimate $\hat{N}^2$, the risk of speech leakage is reduced. Accordingly, there is no need to design an extremely accurate SPP estimator. In this work, we employ a very simple SPP estimator, which depends on the smoothed posteriori SNR. Considering the correlation of speech presence in the neighboring time-frequency points [43], we calculate the smoothed posteriori SNR over a time-frequency region

$$\overline{\gamma}(l,k) = \frac{1}{M}\sum_{j=0}^{\Delta l}\sum_{i=-\Delta k}^{\Delta k}\gamma(l-j,k-i) \tag{6}$$

where $M = (2\Delta k+1)\cdot(\Delta l+1)$ is the number of neghboring time-frequency points which are averaged. Δk and Δl denote number of the adjacent frequency bins and successive time frames, respectively, set to 1 and 2. Then, the smoothed posteriori SNR is compared against a threshold to decide speech present regions as follows

$$\begin{aligned}&\text{if } \overline{\gamma}(l,k) > \Psi(k)\\ &\quad I(l,k) = 1 \quad \text{speech present}\\ &\text{else}\\ &\quad I(l,k) = 0 \quad \text{speech absent}\\ &\text{end}\end{aligned} \tag{7}$$

where $\Psi(k)$ is the threshold, which controls the trade-off between the update speed of noise PSD estimation and the amount of speech leakage. The higher the value, the faster the tracking speed, but the higher the risk of speech leakage. The speech presence probability $I(l,k)$ is smoothed over time using the following first-order recursion:

$$\hat{p}(l,k) = \alpha_p\hat{p}(l-1,k) + (1-\alpha_p)I(l,k) \tag{8}$$

where $\alpha_p(0<\alpha_p<1)$ is a smoothing parameter, set to 0.2 in our experiment as adopted in [25]. The smoothing parameter α_N is obtained by substituting (8) into (5). Here, using averaged priori SNR reduces random fluctuations in $\hat{p}(l,k)$, at the same time fast react to changing noise levels is achieved (minimum tracking is abandoned). Additionally, similar to MCRA-2, we exploit frequency-dependent thresholds $\Psi(k)$ instead of the fixed threshold in MCRA method, set to

$$\Psi(k) = \begin{cases} 5 & 1 \le k \le K/8 \\ 6.5 & K/8 < k \le 3K/8 \\ 8 & 3K/8 < k \le K/2+1. \end{cases}$$

where K is the window length as well as STFT length.

IV. LOG-SPECTRAL POWER MMSE ESTIMATOR

A. DERIVATION OF THE WEIGHTING FUNCTION

To estimate the noise PSD, in this section we derive an estimator of the noise spectral power $\hat{N}^2$, which minimizes the MSE of the log-spectral power, given by

$$\hat{N}^2 = \exp\left(E\left\{\log N^2|Y\right\}\right). \tag{9}$$

In [3] the MMSE estimator of the speech spectral magnitude in logarithmic domain was derived by exploiting moment generating function. Similar to [3], the moment generating function of $\log N^2$ given Y, i.e., $\log N^2|Y$, is exploited to derive the noise spectral power estimator according to (9). Let $P = \log N^2$, then the moment generating function of P given Y takes the form

$$\begin{aligned}M_{P|Y}(\mu) &= E\left\{\exp(\mu P)|Y\right\}\\ &= E\left\{N^{2\mu}|Y\right\}.\end{aligned} \tag{10}$$

By exploiting the first derivation of $M_{P|Y}(\mu)$ at $\mu = 0$, the estimator in (9) is obtained as

$$\hat{N}^2 = \exp\left\{E\left\{\log N^2|Y\right\}\right\} = \exp\left[M'_{P|Y}(\mu)|_{\mu=0}\right]. \tag{11}$$

Therefore, we need to evaluate the moment generating function $M_{P|Y}(\mu)$ and then to obtain the estimator $\hat{N}^2$ using (11). By applying Bayes' theorem, $M_{P|Y}(\mu)$ can be expressed as

$$M_{P|Y}(\mu) = \frac{\int_0^{+\infty}\int_0^{2\pi} n^{2\mu}f(Y|n,\theta)f(n,\theta)d\theta dn}{\int_0^{+\infty}\int_0^{2\pi} f(Y|n,\theta)f(n,\theta)d\theta dn}. \tag{12}$$

Under the assumed complex Gaussian distributions, $f(Y|n,\theta)$ and $f(n,\theta)$ are given by

$$f(Y|n,\theta) = \frac{1}{\pi\lambda_x}\exp\left\{-\frac{|Y-ne^{j\theta}|^2}{\lambda_x}\right\} \tag{13}$$

$$f(n,\theta) = \frac{n}{\pi\lambda_d}\exp\left\{-\frac{n^2}{\lambda_d}\right\}. \tag{14}$$

By substituting (13) and (14) into (12), followed by using [44, Eqs. 8.406.3, 6.631.1, and 9.212.1] we obtain

$$M_{P|Y}(\mu) = \lambda^{\mu}\Gamma(\mu+1)\,\Phi(-\mu,1;-\eta) \tag{15}$$

with $\lambda = \frac{\lambda_x\lambda_d}{\lambda_x+\lambda_d}$ and where $\Gamma(\cdot)$ is the gamma function, $\Phi(\cdot)$ is the confluent hypergeometric function [44, Eq. 9.210.1], and η satisfies the relation

$$\eta = \frac{\gamma}{\xi(1+\xi)}. \tag{16}$$

The first derivative of $M_{P|Y}(\mu)$ at $\mu = 0$ in (11) is then given by

$$\begin{aligned}\frac{d}{d\mu}M_{P|Y}(\mu)|_{\mu=0} = &\underbrace{\frac{d}{d\mu}\left\{\lambda^{\mu}\right\}|_{\mu=0}}_{\text{part 1}}\\ &+\underbrace{\frac{d}{d\mu}\left\{\Gamma(\mu+1)\right\}|_{\mu=0}}_{\text{part 2}}\\ &+\underbrace{\frac{d}{d\mu}\left\{\Phi(-\mu,1;-\eta)\right\}|_{\mu=0}}_{\text{part 3}}\end{aligned} \tag{17}$$

According to the basic derivative rules, the part 1 in (17) is given by

$$\frac{d}{d\mu}\left\{\lambda^{\mu}\right\}|_{\mu=0} = \log\lambda. \tag{18}$$

For the derivation of the part 2, we can obtain the derivative of $\Gamma(\mu+1)$ through the derivative of $\log\Gamma(\mu+1)$. Exploiting the series expansion of $\log\Gamma(\mu+1)$ [44, Eq. 8.342.1], we obtain

$$\frac{d}{d\mu}\Gamma(\mu+1) = \Gamma(\mu+1)\frac{d}{d\mu}\log\Gamma(\mu+1) = -c \tag{19}$$

where c is the Eulers constant. For the computation of part 3, utilizing [44, Eq. 9.210.1] and derivative rules, it can be written as

$$\frac{d}{d\mu}\Phi(-\mu, 1; -\eta)\,|_{\mu=0} = -\sum_{k=1}^{\infty}\frac{(-\eta)^k}{k!}\frac{1}{k}. \tag{20}$$

Now, summing the results of (18), (19) and (20), and followed by utilizing (11), (16) and [44, Eqs. 8.211.1 8.214.1], the Log-spectral MMSE estimator of noise power is obtained as

$$\hat{N}^2 = \left(\frac{1}{1+\xi}\right)^2 \exp\left\{\int_{\frac{\gamma}{\xi(1+\xi)}}^{\infty}\frac{e^{-t}}{t}dt\right\} R^2. \tag{21}$$

The noise spectral power estimation is obtained from the noisy speech through a multiplicative nonlinear weighting function which depends only on the a priori and the a posteriori SNR. The weighting function is defined as

$$G_{N^2} = \frac{\hat{N}^2}{R^2} = \left(\frac{1}{1+\xi}\right)^2 \exp\left\{\int_{\frac{\gamma}{\xi(1+\xi)}}^{\infty}\frac{e^{-t}}{t}dt\right\} \tag{22}$$

After estimating noise spectral power $\hat{N}^2$ with (21), the noise PSD estimation is updated via (4) and (5) as

$$\begin{aligned}\hat{\lambda}_d(l,k) &= \hat{p}(l,k)\hat{\lambda}_d(l-1,k) \\ &\quad + \left(1-\hat{p}(l,k)\right)\left\{\alpha_n\hat{\lambda}_d(l-1,k) + (1-\alpha_n)G_{N^2}R^2\right\}\end{aligned} \tag{23}$$

B. PRIORI SNR ESTIMATION FOR NOISE PSD ESTIMATION

It is observed from (22) that the weighting function takes the priori and posteriori SNRs as parameters. As these parameters are unknown in practice, it is necessary to make an estimation. We have known that noise PSD tracking performance depends on the particular priori SNR estimator used. For the a priori SNR estimate, the DD approach and the ML approach are proposed in [2]. The DD approach is based on a heuristic knowledge and is widely accepted in literature. In this work, the standard DD priori SNR estimator is exploited to estimate the a priori SNR used for noise PSD estimate:

$$\begin{aligned}\hat{\xi}(l,k) = \max\Bigg[&\alpha_{NS}\frac{\hat{A}^2(l-1,k)}{\hat{\lambda}_d(l-1,k)} \\ &+(1-\alpha_{NS})\left[\frac{R^2(l,k)}{\hat{\lambda}_d(l-1,k)} - 1\right], \xi_{\min}\Bigg]\end{aligned} \tag{24}$$

where $\xi_{\min} = -15$ dB is the minimum value allowed for the priori SNR ξ, α_{NS} is the smoothing factor, $\hat{\lambda}_d$ is the estimated noise PSD, and $\hat{A}^2(l-1,k)$ is the speech spectral power estimate obtained in the previous frame. The smoothing factor α_{NS} typically lies in the range [0.9, 0.99] [45] and is set to 0.98 in this work.

C. SAFETY NET

Moreover, as in MMSE-based algorithm [31], in order to ensure that the noise PSD estimator continues to work properly in the extreme situation where the noise power level abruptly changes from one level to another, an effective and simple safety-net presented in [42] is adopted. In the safety-net, some memory resources are required to store the previous 0.8 seconds of the smoothed periodogram $S(l,k)$ of noisy speech $|Y(l,k)|^2$, where $S(l,k)$ is given by $S(l,k) = 0.1\,S(l-1,k) + 0.9\,|Y(l,k)|^2$. The minima $S_{\min}(l,k)$ of $S(l,k)$ is used as a reference value. Then, the noise PSD estimation $\hat{\lambda}_d(l,k)$ obtained with (23) is checked whether it fulfills the condition: $\hat{\lambda}_d(l,k)/S_{\min}(l,k) < 1.5$. If that happens, the final noise PSD estimation is updated by $\hat{\lambda}_d(l,k) = \max\left[1.5\cdot S_{\min}(l,k), \hat{N}^2(l,k)\right]$.

V. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, several comparisons and experiments are carried out to evaluate the performance of noise PSD trackers and demonstrate the superiority of proposed algorithm over other four state-of-the-art methods. Performance evaluations are conducted on the NOIZEUS database, which contains 30 IEEE sentences produced by three female and three male speakers [22], [46]. Clean speech signals are corrupted by five distinct types of noise sources at five input SNR levels, namely -5, 0, 5, 10, and 15 dB. The noise sources are modulated white Gaussian noise, babble noise from NOISEX-92 database [47], passing car noise, passing train noise, and traffic noise. The modulated white Gaussian noise is obtained through modulating white Gaussian noise by the following function

$$f(n) = 0.1 + 0.5\sin\left(2\pi n\frac{f_{mod}}{f_s} - \pi\right) \tag{25}$$

where n is the discrete-time index, f_s the sampling frequency, and $f_{mod} = 0.2$ Hz denotes the modulation frequency. The passing car noise, passing train noise, and traffic noise are taken from Freesound database [48]. Speech and noise signals used in our experiments are sampled at a frequency of $f_s = 8$ kHz. All noise PSD trackers employ a overlapping square-root-Hann window for spectral analysis and synthesis. The window length as well as the DFT length is K = 256 samples (32 ms), and the amount of the overlap between successive frames is 50%.

In section V-A, we first compare the noise estimation accuracy of all noise trackers in five different noise environments. Subsequently, in section V-B the noise PSD estimators are integrated into a noise suppression framework and the speech enhancement performance is compared. Finally, the computational complexity is analyzed in section V-C.

A. NOISE ESTIMATION ACCURACY

The noise estimation accuracy is measured using the averaged logarithmic spectral error distance between the estimated noise PSD $\hat{\lambda}_d(l,k)$ and the ideal reference noise PSD $\lambda_d(l,k)$. The ideal reference noise PSD $\lambda_d(l,k)$ is calculated

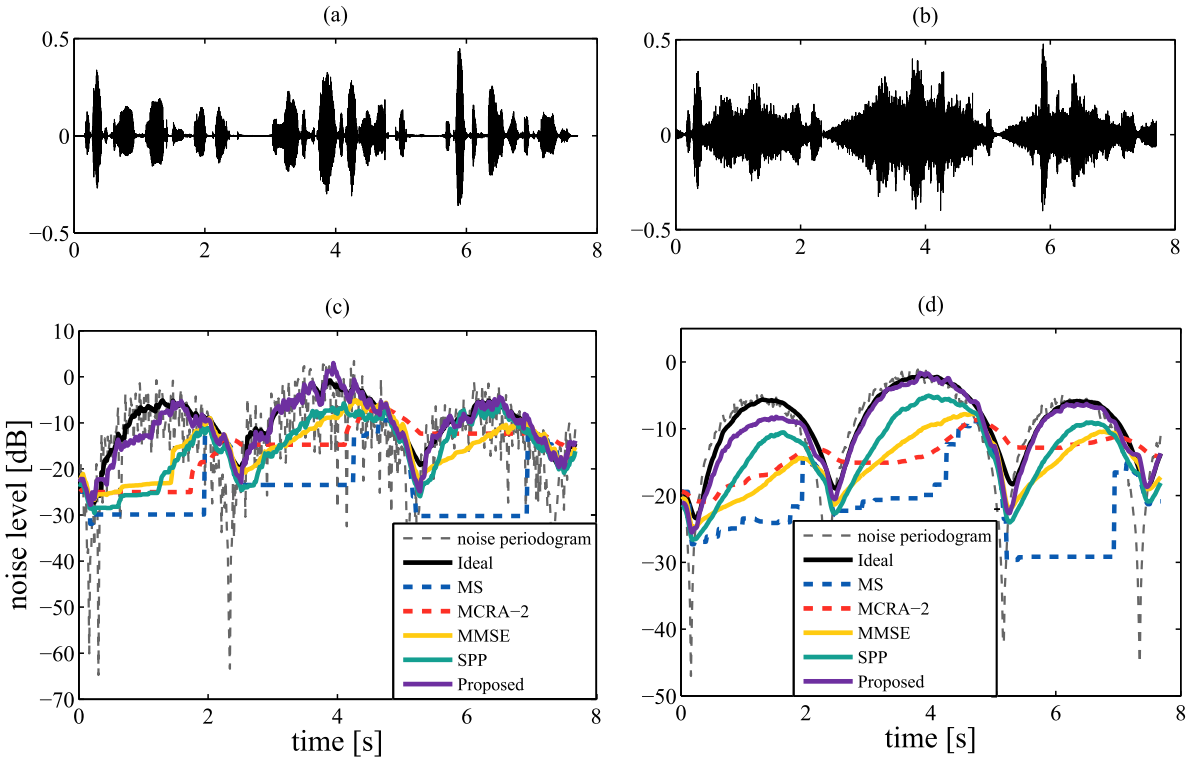

FIGURE 1. (a) Clean speech signal. (b) Speech signal contaminated with modulated Gaussian white noise at an overall input SNR of 0 dB. (c) Comparison between proposed approach and the four state-of-the-art noise estimators for a single frequency bin $k = 36$. (d) The estimated noise PSDs averaged over all frequency bins.

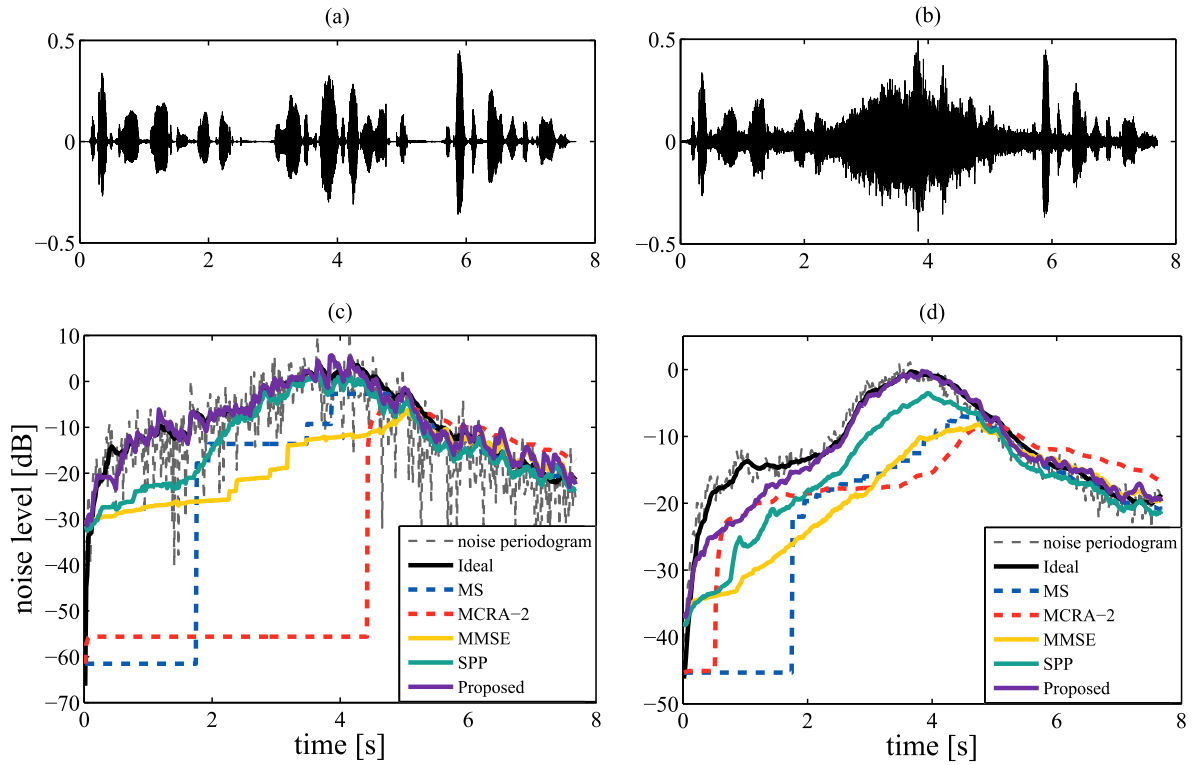

FIGURE 2. (a) Clean speech signal. (b) Speech signal degraded by passing train noise at an overall input SNR of 0 dB. (c) Comparison between proposed method and the four state-of-the-art noise estimators for a single frequency bin $k = 36$. (d) The estimated noise PSDs averaged over all frequency bins.

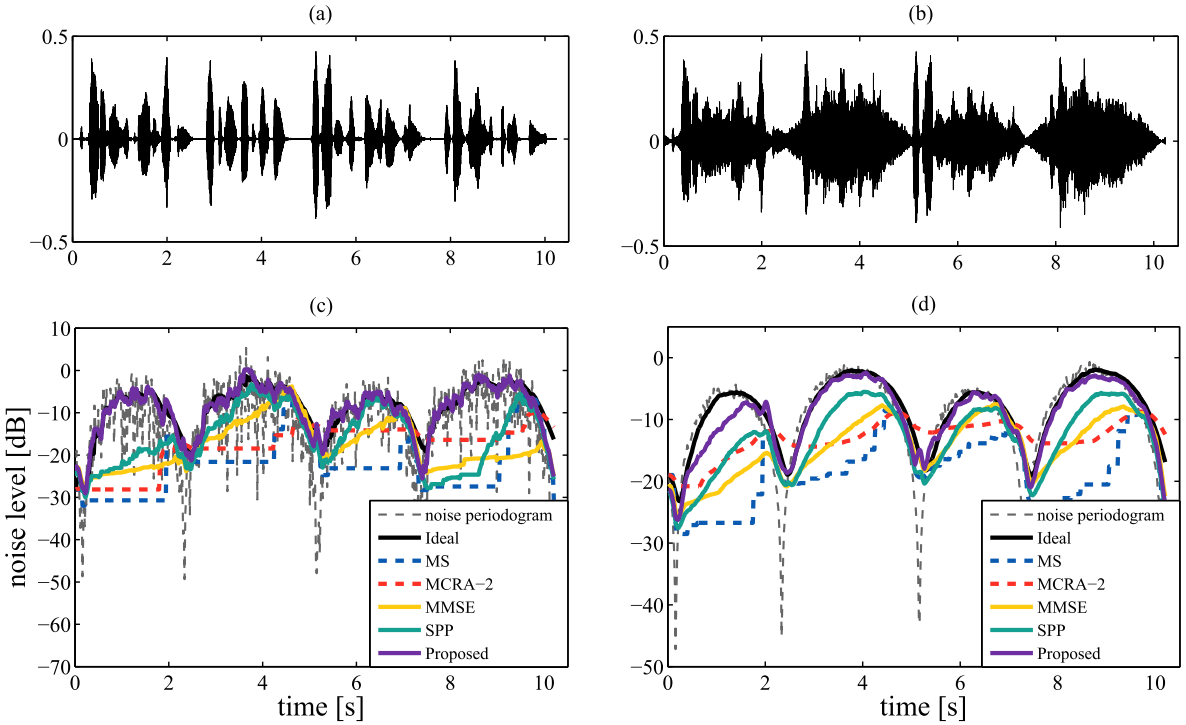

FIGURE 3. (a) Clean speech signal. (b) Speech signal degraded by modulated Gaussian white noise at an overall input SNR of 0 dB. (c) Comparison between proposed approach and the four state-of-the-art noise trackers for a single frequency bin $k = 36$. (d) The estimated noise PSDs averaged over all frequency bins.

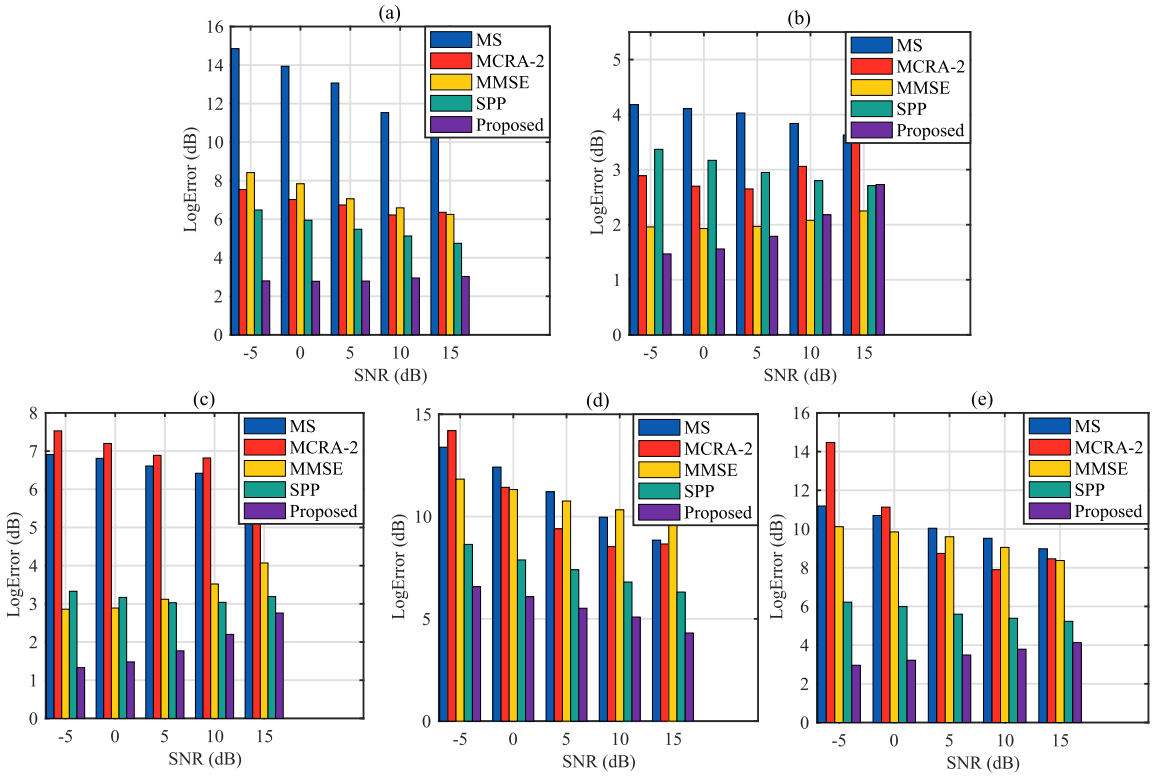

FIGURE 4. Performance comparison in terms of LogError (dB). (a) modulated Gaussian white noise with $f_{mod} = 0.2$ Hz, (b) babble noise, (c) passing car noise, (d) passing train noise, (e) traffic noise.

TABLE 1. Performance comparison in terms of segSNR [dB].

noise source	input SNR [dB]	noisy	MS [24]	MCRA-2 [27]	MMSE-based [31]	SPP-based [32]	Prop.
modulated white noise	-5	-5.72	-5.05	-4.83	-4.65	-3.76	**-2.67**
	0	-3.12	-2.68	-2.32	-2.19	-1.11	**-0.50**
	5	-1.72	0.08	0.50	0.85	1.75	**2.39**
	10	3.01	3.22	3.48	3.86	4.69	**4.88**
	15	6.57	6.79	6.36	7.34	7.96	**8.07**
babble noise	-5	-6.92	-5.02	-4.67	-4.06	-4.27	**-3.86**
	0	-4.58	-2.68	-2.23	-1.89	-2.05	**-1.88**
	5	-1.87	0.07	0.59	**0.85**	0.60	0.77
	10	1.26	3.22	3.56	**3.89**	3.73	3.64
	15	4.8	6.58	6.27	**7.15**	7.00	6.72
passing car noise	-5	-5.41	-3.39	-3.68	-1.41	-1.37	**-0.38**
	0	-2.89	-0.91	-1.16	0.80	0.94	**1.76**
	5	0.07	1.84	1.59	3.14	3.46	**3.85**
	10	3.31	4.87	4.26	5.80	6.20	**6.22**
	15	6.83	7.99	6.90	8.91	**9.30**	9.10
passing train noise	-5	-3.72	-2.27	-2.56	-2.27	-1.49	**-0.7**4
	0	-0.99	0.46	0.08	0.41	1.27	**1.92**
	5	1.94	3.36	2.70	3.16	4.09	**4.60**
	10	5.15	6.47	5.27	6.13	7.22	**7.23**
	15	8.32	9.52	7.55	9.22	10.02	**10.06**
traffic noise	-5	-5.08	-3.20	-3.66	-3.38	-2.23	**-1.40**
	0	-2.49	-0.55	-1.09	-0.65	0.46	**1.13**
	5	0.53	2.42	1.66	2.25	3.48	**3.80**
	10	3.87	5.63	4.58	5.38	6.74	**6.77**
	15	7.59	9.08	7.27	8.97	9.72	**9.88**

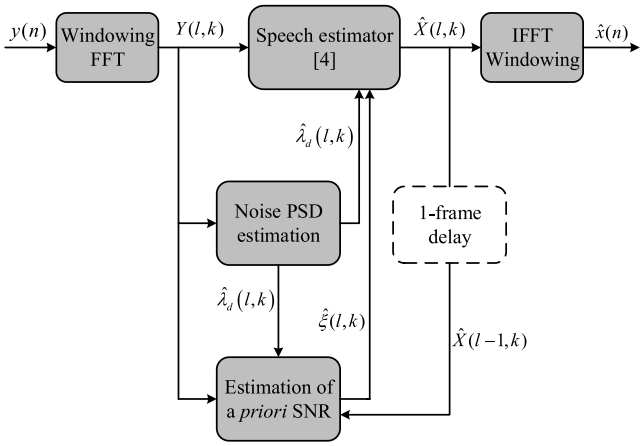

FIGURE 5. Block diagram of the standard DFT-based single channel speech enhancement scheme.

by employing a recursive temporal smoothing of noise periodograms [28], [34] i.e.,

$$\lambda_d(l,k) = \alpha_d \lambda_d(l-1,k) + (1-\alpha_d)\,|D(l,k)|^2 \quad (26)$$

with a smoothing parameter $\alpha_d = 0.9$ [28], [34]. The averaged logarithmic spectral error distance (LogErr) is defined as follows [28], [34]

$$\text{LogErr} = \frac{10}{LK}\sum_{k=1}^{K}\sum_{l=1}^{L}\left|\log_{10}\left[\frac{\lambda_d(l,k)}{\hat{\lambda}_d(l,k)}\right]\right| \text{ (dB)} \quad (27)$$

where L and K indicate the number of signal-frames and frequency bins respectively. The lower LogErr value, the better the tracking capability.

To illustrate the noise tracking performance of the proposed method in comparison to four competing trackers, we consider an example where three speech signals obtained from one female and two male speakers are concatenated and is degraded by modulated white Gaussian noise at an overall SNR of 0 dB. In Fig. 1, the estimated noise PSDs are shown for proposed method and four competing noise estimators together with ideal reference noise PSD. The clean and noisy speech signals are shown in Fig. 1(a) and Fig. 1(b), respectively. Fig. 1(c) exhibits the results of noise PSD estimation at frequency bin $k = 36$. This frequency bin index corresponds to the DFT band centered around 1125 Hz. Fig. 1(d) displays the estimated noise PSDs averaged over all frequency bins. It is observed that the proposed noise estimator tracks the increases and decreases in noise level much better than other four approaches. As expected, MS is not capable of tracking

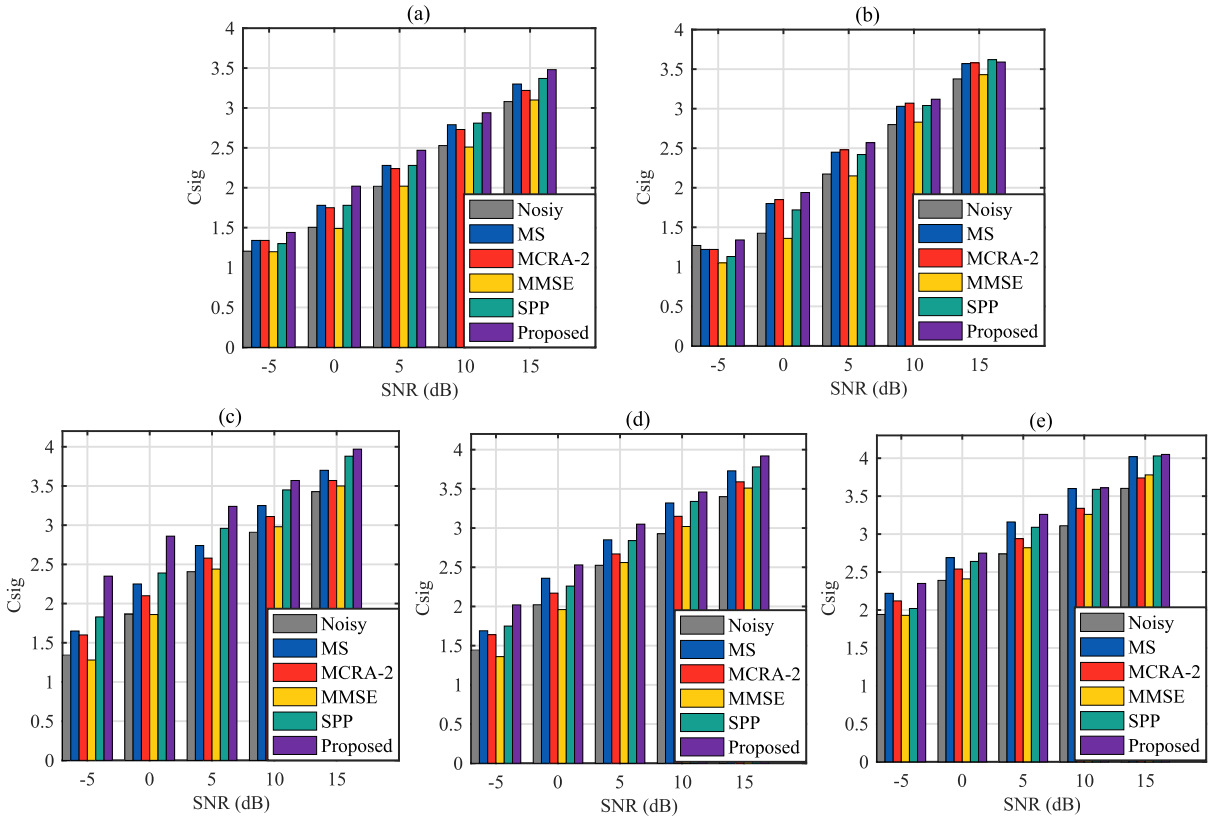

FIGURE 6. Performance comparison in terms of C_{sig}. (a) modulated Gaussian white noise with $f_{mod} = 0.2$ Hz, (b) babble noise, (c) passing car noise, (d) passing train noise, (e) traffic noise.

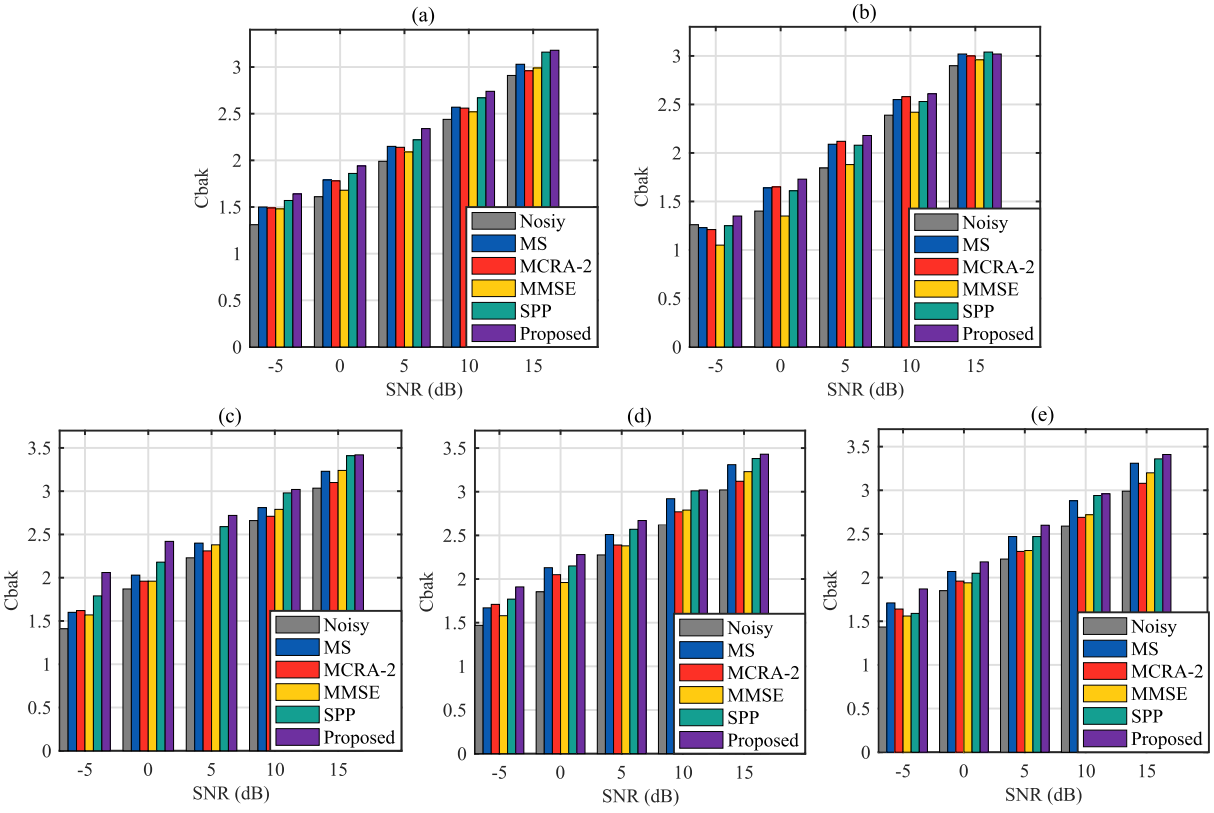

FIGURE 7. Performance comparison in terms of C_{bak}. (a) modulated Gaussian white noise with $f_{mod} = 0.2$ Hz, (b) babble noise, (c) passing car noise, (d) passing train noise, (e) traffic noise.

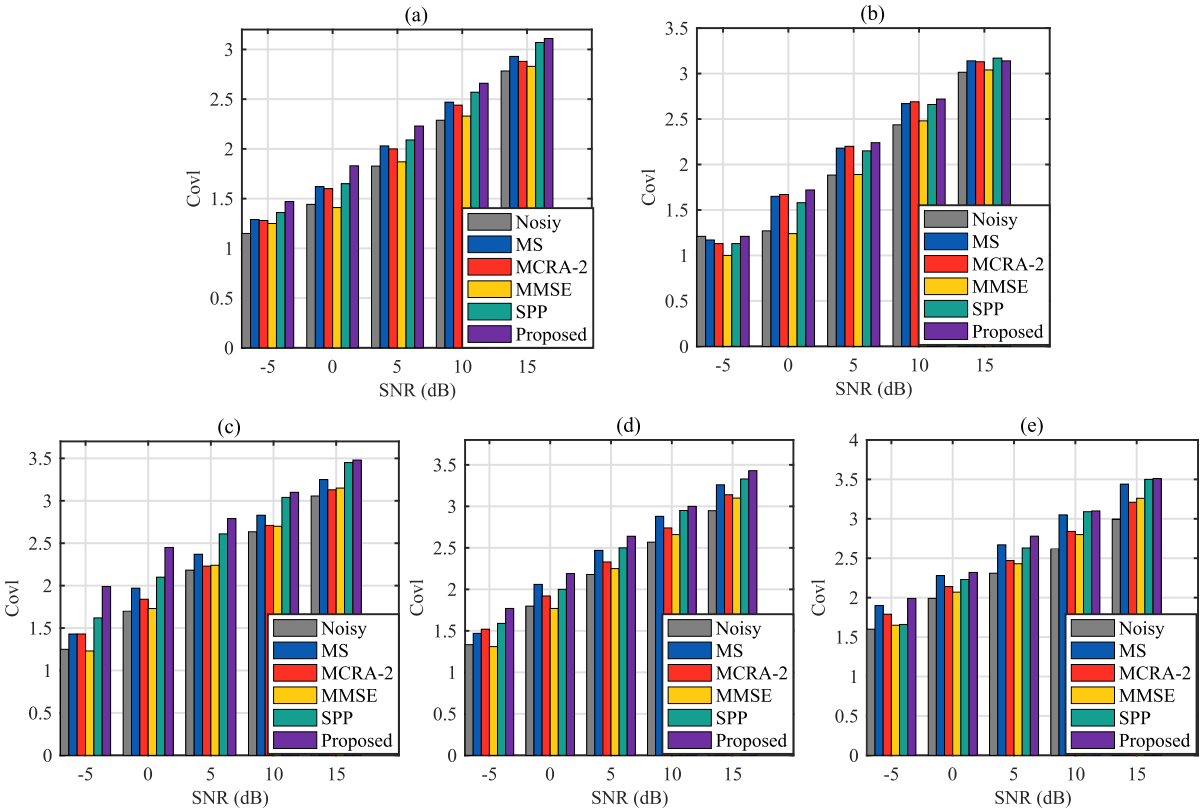

FIGURE 8. Performance comparison in terms of C_{ovl}. (a) modulated Gaussian white noise with $f_{mod} = 0.2$ Hz, (b) babble noise, (c) passing car noise, (d) passing train noise, (e) traffic noise.

the changes when noise PSD increases. MCRA-2 is based on the minimum-tracking principle and therefore also shows a relatively large delay in tracking the increasing noise PSD. Compared to MS and MCRA-2, the MMSE and SPP algorithms perform better, but still have a tracking delay when the noise PSD rises.

In Fig. 2, we show a second example where the same speech signal is corrupted by noise originating from passing train at an overall SNR of 0 dB. It is observed again that the proposed method exhibits better performance of handling both fast increases and decreases in noise level than other four reference approaches. For a rapidly increasing noise PSD, i.e., in the time-interval from 0-4 seconds, the proposed algorithm has a shortest tracking delay. When the noise is decreasing, e.g., in the time-span from 5 till 8 seconds, the proposed method, MS, MMSE and SPP exhibit similar performance. Compared to MS, the MCRA-2 algorithm is slightly better in tracking the increasing noise level, but it has the tendency to overestimate the noise PSD when the noise is decreasing. Fig. 3 shows another example where four different speech signals spoken by two male and two female speakers are concatenated and is corrupted by modulated white noise at an SNR of 0 dB. It is evident from Fig. 3 that the proposed noise tracker shows a better tracking performance than other competing methods.

The quantitative evaluation results of noise tracking performance of all noise PSD estimators are given in Fig. 4 in terms of LogErr measure. It can be observed from the results in Fig. 4 that the proposed algorithm clearly outperforms other four competing methods in terms of LogErr for almost all noise sources and SNR levels, except for babble noise at 10 and 15 dB input SNR, where MMSE performs slightly better. As the proposed method can quickly update the noise PSD estimate, the superiority in terms of tracking performance is obvious especially at low SNR conditions. However, with the increase of SNR, the proposed tracker updates noise estimate quickly which may lead to overestimation of noise, and shows an increase in terms of LogErr.

TABLE 2. Comparison of the computational complexity in terms of normalized processing time.

Approach	MS	MCRA-2	MMSE	SPP	Prop.
Proc. Time	0.67	0.51	0.38	0.21	1

B. NOISE SUPPRESSION PERFORMANCE

In order to investigate the impact of noise PSD trackers on noise suppression performance, the estimated noise PSDs are then incorporated into a DFT domain-based single channel speech enhancement system. The block diagram of the standard DFT-based single channel speech enhancement framework is depicted in Fig. 5. For the speech estimator, this work employs an MMSE amplitude estimator, which is derived under the assumption that the speech DFT coefficients follow a generalized-Gamma distribution with distribution

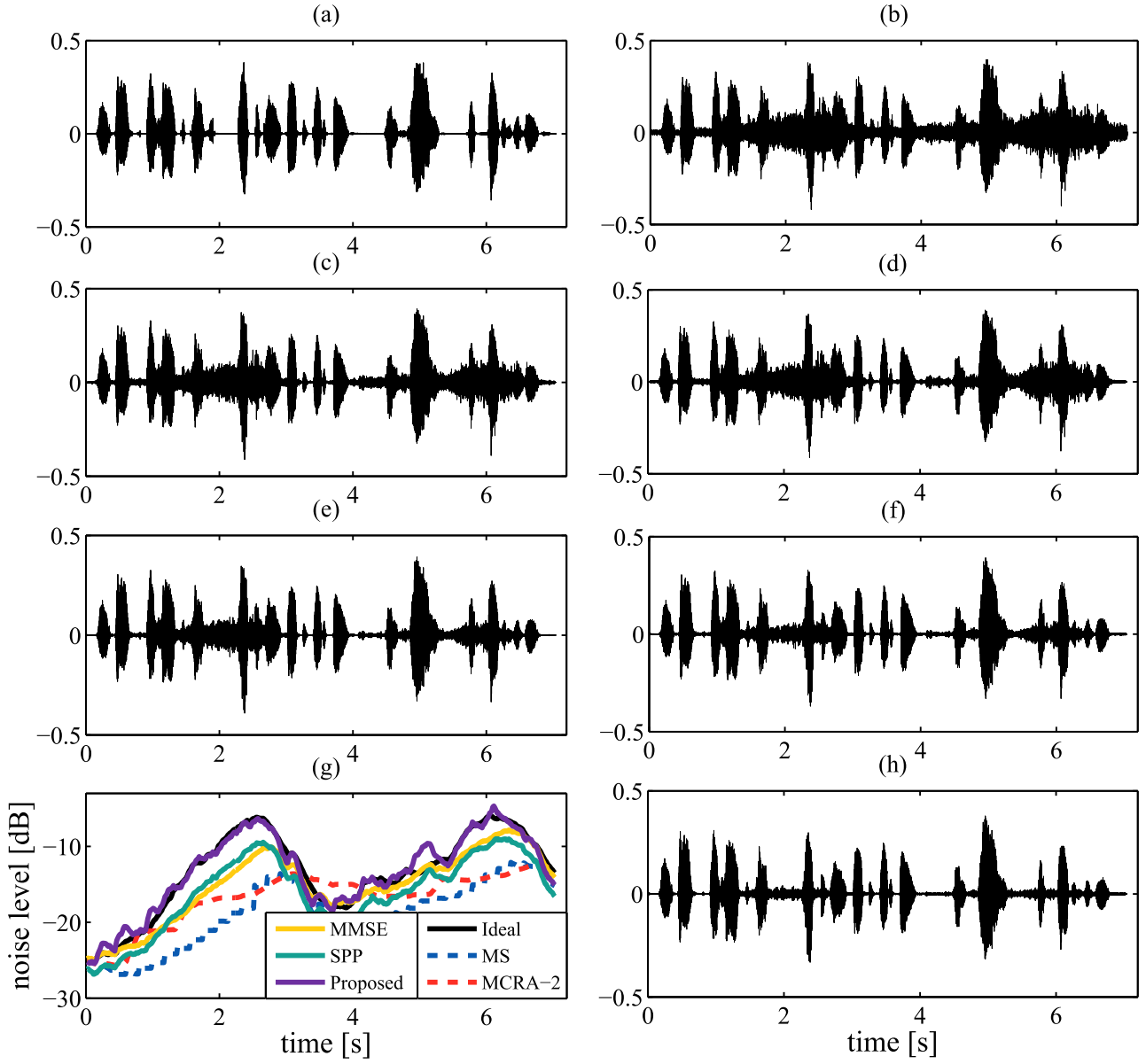

FIGURE 9. Waveforms of enhanced speech signal obtained using different noise estimators. (a) clean speech signal. (b) noisy speech signal corrupted by traffic noise at an overall input SNR of 5 dB. The enhanced speech signals with: (c) MS, (d) MCRA-2 method, (e) MMSE-based method, (f) SPP-based algorithm, (h) Proposed algorithm. (g) shows the ideal noise PSD and estimated noise PSDs obtained with different noise trackers.

parameters $\gamma = 1$ and $\nu = 0.6$ [4]. In this speech enhancement system, we estimate the priori SNR using the decision-directed approach with a smoothing parameter $\alpha_{dd} = 0.98$.

The speech enhancement performance is evaluated in terms of the segSNR metric and three composite objective metrics. The segmental SNR (segSNR) is defined as follows [22], [40]

$$\text{segSNR} = \frac{1}{L}\sum_{l=0}^{L-1}\Phi\left\{10\log_{10}\frac{\sum_{n=0}^{N}x^2(lN+n)}{\sum_{n=0}^{N}\left(x(lN+n)-\hat{x}(lN+n)\right)^2}\right\}. \tag{28}$$

where L and N denote the number of frames in the signal and the frame length, respectively, and $\Phi(x) = \min\{\max(x,-10),35\}$. For the segSNR computation, only the signal segments containing speech are taken into account. The segSNR values are limited in the range of $[-10\text{dB}, 35\text{dB}]$ thereby avoiding the need for a speech/silence detector. The segSNR measure results obtained with different noise PSD tracking algorithms are given in Table 1. It is found that the proposed noise PSD tracker yields larger segSNR improvements than all other algorithms for almost each noise source except for babble noise. For babble noise, MMSE obtains slightly higher segSNR values in case of input SNR more than 5 dB. However, the segSNR measure, which is widely

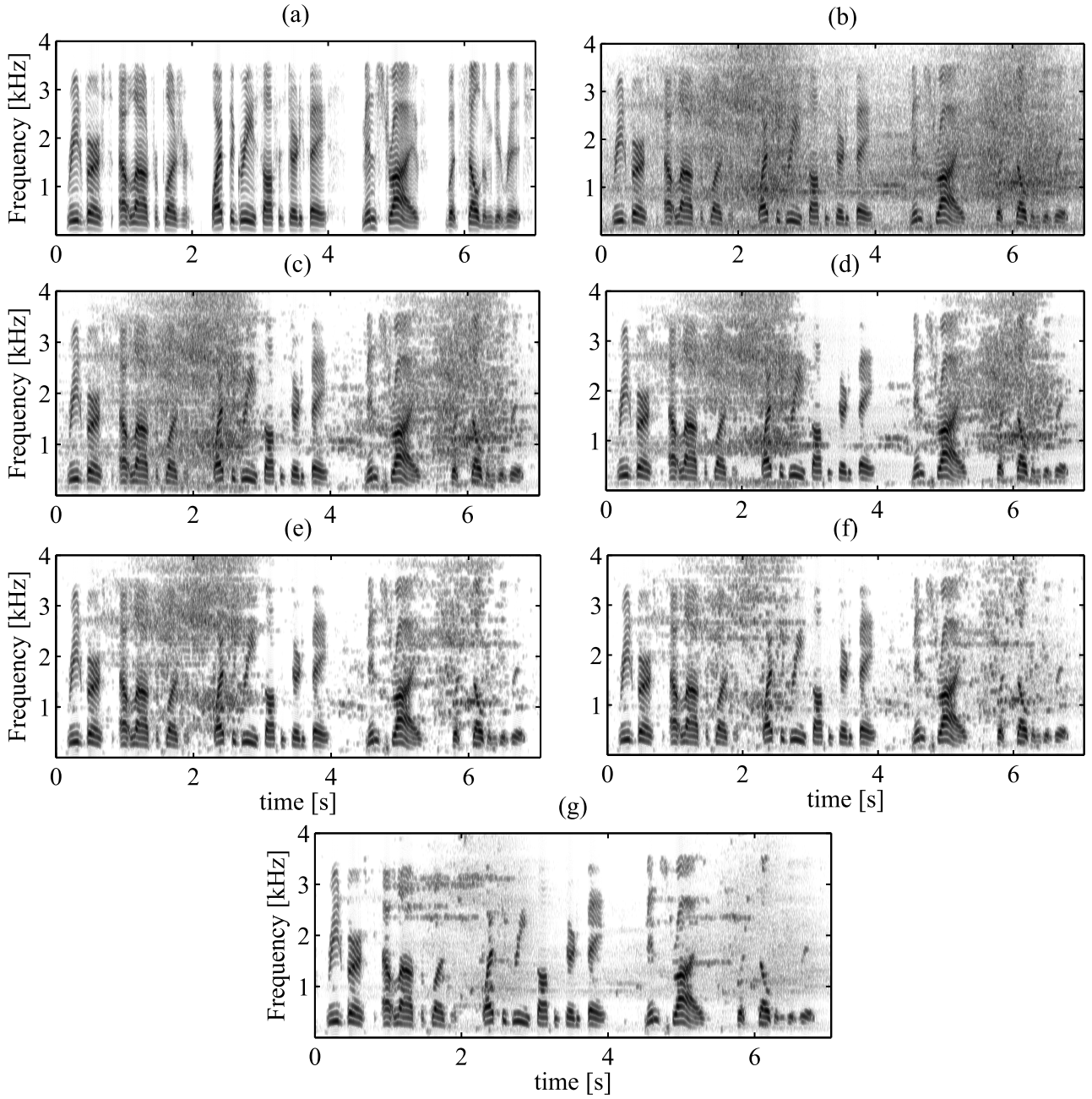

FIGURE 10. Spectrograms of (a) clean speech signal and (b) noisy speech signal corrupted by traffic noise at an overall input SNR of 5 dB. The spectrograms of enhanced speech signals using: (c) MS, (d) MCRA-2 method, (e) MMSE-based method, (f) SPP-based algorithm (g) Proposed algorithm.

used to evaluate noise reduction performance of speech enhancement algorithms, yields a poor correlation coefficients with subjective measure. For this, three composite objective metrics are employed to evaluate the enhancement performance.

The three composite objective metrics are C_{sig}, C_{bak}, and C_{vol}, which are obtained by linearly combining existing widely used measures, segSNR, weighted-slope spectral (WSS) distance [49], perceptual evaluation of speech quality (PESQ) [50], log likelihood ration (LLR) and Itakura-Saito (IS) distance measure [51]. The three composite metrics are given below [41]:

$$\begin{cases} C_{sig} = 3.093 - 1.029\text{LLR} + 0.603\text{PESQ} - 0.009\text{WSS} \\ C_{bak} = 1.634 + 0.478\text{PESQ} - 0.007\text{WSS} + 0.063\text{segSNR} \\ C_{ovl} = 1.594 + 0.805\text{PESQ} - 0.512\text{LLR} - 0.007\text{WSS} \end{cases} \tag{29}$$

C_{sig}, C_{bak} and C_{vol} are designed to provide the high correlations with three subjective measures, i.e., Mean opinion score (MOS) predictor of speech distortion (SIG), MOS predictor of background intrusiveness (BAK), and MOS predictor of overall speech quality (OVRL).

The scores of three composite objective metrics obtained with all noise PSD estimation methods are shown in Figs. 6-8. Since the three composite measures provide very high correlation coefficients with subjective measures, especially C_{ovl} measure has the highest correlation with the real subjective test, the evaluation results of composite measures are more important than segSNR measure. From the scores in Figs. 6-8, we observe that the proposed noise estimator is clearly superior to other noise tracking methods for all noise types and SNR conditions, except for babble noise at 15 dB.

Figs. 9 and 10 present the enhanced waveforms and spectrograms obtained with different noise estimators for a speech example which is degraded by the traffic noise at 5 dB input SNR. In this way, the enhancement performance of speech enhancement algorithm combined with different noise estimators can be seen more directly. Fig. 9(a)-(b) and Fig. 10(a)-(b) show waveforms and spectrograms of clean speech and noisy speech, respectively. Fig. 9(c)-(f) display the enhanced speech waveforms obtained using four competing noise trackers, and the respective spectrograms are shown in Fig. 10(c)-(f). Fig. 9(h) and Fig. 10(g) show the enhanced waveform and spectrogram using proposed algorithm, respectively. Additionally, Fig. 9(g) also shows the estimated noise PSDs together with ideal reference noise PSD. Clearly, the proposed method performs better than other four competing algorithms. In general, the proposed approach shows a good tradeoff between noise suppression and speech distortion as it obtains higher segSNR and higher three composite measures.

C. COMPUTATIONAL COMPLEXITY ANALYSIS

To investigate the computational complexity of proposed algorithm and other four competing algorithms, we compare the execution time of Matlab implementations of these algorithms in this section [32]. The Matlab implementations of these methods run on a PC with a Intel Core i7-7700 processor. Table 2 shows the execution times of all five methods, normalized by the execution time of the proposed method. It is observed that the proposed algorithm exhibits a higher computational complexity than other methods. The computational complexity of the proposed method is mainly determined by the computation of the nonlinear weighting function (22), as exponential operation of the special exponential integral function needs to be computed.

However, in a practical system, all nonlinear weighting functions can be computed offline for the relevant range of the parameters and stored in a lookup table. In this way, the noise PSD tracker can be implemented with significantly reduced execution time (normalized execution time: 0.52). The computation complexity is not an issue then. In addition, since more and more computational power will be available with improved technology, this problem will be easily solved. Notice, that the numbers as given in Table 2 are rough estimates since there will be some changes depending on the implementation details. The number in Table 2 reflects all processing steps of the proposed algorithm.

VI. CONCLUSION

A crucial component of single-channel speech enhancement algorithms is the estimation of noise PSD. This paper develops a novel algorithm for noise PSD estimation. In this method, a nonlinear weighting function of the log-spectral power MMSE estimator is derived to estimate instantaneous noise spectral power, which depends on the a priori and the a posteriori SNR. Then, the noise PSD estimation is updated by performing a temporal recursive averaging of log-spectral MMSE estimation of the current noise power. The smoothing parameter in the temporal recursive smoothing operation is adjusted by a simple estimate of speech presence probability.

Experimental results of LogErr measure demonstrate that the proposed algorithm achieves faster and more accurate noise PSD tracking. Additionally, evaluation results of segSNR and three composite measures (C_{sig}, C_{bak}, C_{ovl}) show that the enhancement performance of proposed method is clearly superior to other competing methods in the presence of various noise sources and levels. The overall performance improvements of the proposed noise tracker come with an increase of computational complexity, which is mainly determined by the nonlinear weighting function computation. However, in a practical system, all weighting function can be evaluated offline and stored in a lookup table, thus the proposed method can be implemented with a significant decrease in computational complexity. As a result, the proposed method leads to a better tradeoff between the computational complexity and the overall performance. The techniques developed in this paper are of importance for many applications, such as hearing aids, speaker identification, human-computer interactions and many others.

ACKNOWLEDGMENT

Authors thank a lot the Circuit and Systems (signal processing) Group at Delft University of Technology for providing the matlab code of the MMSE speech estimator with generalized Gamma priors.

REFERENCES

[1] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Appl. Signal Process.*, vol. 2005, May 2005, Art. no. 354850.

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

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[4] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[5] M. S. Kavalekalam, J. K. Nielsen, J. B. Boldt, and M. G. Christensen, "Model-based speech enhancement for intelligibility improvement in binaural hearing aids," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 99–113, Jan. 2019.

[6] M. S. Kavalekalam, M. G. Christensen, and J. B. Boldt, "Model based binaural enhancement of voiced and unvoiced speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, LA, USA, Mar. 2017, pp. 666–670.

[7] T. Hussain, S. M. Siniscalchi, C.-C. Lee, S.-S. Wang, Y. Tsao, and W.-H. Liao, "Experimental study on extreme learning machine applications for speech enhancement," *IEEE Access*, vol. 5, pp. 25542–25554, Oct. 2017.
[8] R. C. Hendriks, T. Gerkmann, and J. Jensen, "DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art," *Synth. Lectures Speech Audio Process.*, vol. 9, no. 1, pp. 1–80, Jan. 2013.
[9] J. Benesty, J. Chen, and E. A. Habets, *Speech Enhancement in the STFT Domain* (SpringerBriefs in Electrical and Computer Engineering). Berlin, Germany: Springer-Verlag, 2011.
[10] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Hoboken, NJ, USA: Wiley, 2006.
[11] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, USA, Apr. 2008, pp. 4037–4040.
[12] B. M. Mahmmod, A. R. Ramli, S. H. Abdulhussain, S. Al-Haddad, and W. A. Jassim, "Low-distortion MMSE speech enhancement estimator based on laplacian prior," *IEEE Access*, vol. 5, pp. 9866–9881, 2017.
[13] A. Chinaev and R. Haeb-Umbach, "A generalized log-spectral amplitude estimator for single-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, LA, USA, Mar. 2017, pp. 4980–4984.
[14] N. Dionelis and M. Brookes, "Phase-aware single-channel speech enhancement with modulation-domain Kalman filtering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 5, pp. 937–950, May 2018.
[15] Y. Wang and M. Brookes, "Speech enhancement using an MMSE spectral amplitude estimator based on a modulation domain Kalman filter with a Gamma prior," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5225–5229.
[16] G. Enzner and P. Thüne, "Robust MMSE filtering for single-microphone speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, LA, USA, Mar. 2017, pp. 4009–4013.
[17] B. Fodor and T. Fingscheidt, "MMSE speech enhancement under speech presence uncertainty assuming (generalized) gamma speech priors throughout," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 4033–4036.
[18] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. Audio, Speech Language Process.*, vol. 21, no. 4, pp. 697–710, Apr. 2013.
[19] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.
[20] J.-H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1965–1976, Jun. 2006.
[21] K. W. Jang, D. K. Kim, and J.-H. Chang, "A uniformly most powerful test for statistical model-based voice activity detection," in *Proc. ISCA Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 2917–2920.
[22] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton FL, USA: CRC Press, 2013.
[23] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 1994, pp. 1182–1185.
[24] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.
[25] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.
[26] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.
[27] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Commun.*, vol. 48, no. 2, pp. 220–230, Feb. 2006.
[28] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Trans. Speech, Lang. Process.*, vol. 16, no. 3, pp. 541–553, Mar. 2008.
[29] M. Triki and K. Janse, "Minimum subspace noise tracking for noise power spectral density estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 29–32.
[30] R. Yu, "A low-complexity noise estimation algorithm based on smoothing of noise power estimation and estimation bias correction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 4421–4424.
[31] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, USA, Mar. 2010, pp. 4266–4269.
[32] T. Gerkmann and R. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.
[33] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," in *Proc. Eurospeech*, Sep. 1995, pp. 1513–1516.
[34] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, USA, May 2011, pp. 4640–4643.
[35] J. K. Nielsen, M. S. Kavalekalam, M. G. Christensen, and J. B. Boldt, "Model-based noise PSD estimation from speech in non-stationary noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 5424–5428.
[36] M. S. Kavalekalam, J. K. Nielsen, M. G. Christensen, and J. B. Boldt, "A study of noise PSD estimators for single channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 5464–5468.
[37] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Berlin, Germany: Springer-Verlag, Mar. 2013.
[38] R. Gray, A. Buzo, A. Gray, and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 367–376, Aug. 1980.
[39] N. Dionelis and M. Brookes, "Modulation-domain speech enhancement using a Kalman filter with a Bayesian update of speech and noise in the log-spectral domain," in *Proc. IEEE Int. Workshop Hands-Free Speech Commun. Microphone Arrays*, San Francisco, CA, USA, Mar. 2017, pp. 111–115.
[40] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. Int. Conf. Spoken Lang. Process.*, Nov. 1998, pp. 2819–2822.
[41] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in *Proc. Interspeech*, Sep. 2006, pp. 1447–1450.
[42] J. S. Erkelens and R. Heusdens, "Tracking of nonstationary noise based on data-driven recursive noise power estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 6, pp. 1112–1123, Aug. 2008.
[43] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, Nov. 2001.
[44] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, D. Zwillinger and V. Moll, 7th ed. New York, NY, USA: Academic, 2007.
[45] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 1994.
[46] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech commun.*, vol. 49, nos. 7–8, pp. 588–601, Jul. 2007.
[47] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.
[48] *Freesound.org*. [Online]. Available: http://www.freesound.org/
[49] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 1982, pp. 1278–1281.
[50] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2001, pp. 749–752.
[51] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.

QIQUAN ZHANG was born in Yunnan, China, in 1993. He received the B.S. degree in electronic science and technology from the Harbin Institute of Technology, Weihai Campus, Shandong, China, in 2015. He is currently pursuing the Ph.D. degree in electrical science and technology with the Shenzhen Key Laboratory of Internet of Things Terminal Technology, Harbin Institute of Technology at Shenzhen, Shenzhen. His research interests include digital speech and audio signal processing, speech enhancement algorithms, and microphone array signal processing.

MINGJIANG WANG was born in Heilongjiang, China, in 1968. He received the B.S. and M.S. degrees in semiconductor physics and devices from the Harbin Institute of Technology, Heilongjiang, in 1990 and 1993, respectively, and the Ph.D. degree in electronic engineering from Fudan University, Shanghai, China, in 1998. From 1993 to 1995, he was an Associate Professor with Southeast University, Jiangsu, China. He was also a Senior Engineer with Huawei Technologies Company, Ltd., from 1998 to 2000. Since 2009, he has been a Professor with the Electronic and Information Engineering Department, Harbin Institute of Technology at Shenzhen, Shenzhen Campus. He is currently the Director of the Shenzhen Key Laboratory of Internet of Things Terminal Technology. His research interests include the low-power loss chip design, speech signal processing, speech coding, speech enhancement, and audio/image deep learning algorithm with application to the AI processing chip design.

YUN LU received the B.S. degree in microelectronics from Xiangtan University, Hunan, China, in 2009, and the M.S. degree in optical engineering from Sun Yat-sen University, Guangdong, China, in 2011. He is currently pursuing the Ph.D. degree in electrical science and technology with the Shenzhen Key Laboratory of Internet of Things Terminal Technology, Harbin Institute of Technology at Shenzhen, Shenzhen. From 2011 to 2014, he was with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, working on the design of mixed-signal front-end circuits for biomedical applications. His current research interests include information processing of acoustic signals at the interface of engineering and neuroscience, innovative methods for brain-machine interface, and neural prosthesis.

MUHAMMAD IDREES received the B.S. degree in electrical engineering from the COMSATS Institute of Information Technology, Wah, Pakistan, in 2013, and the master's degree in telecommunication engineering from the University of Engineering and Technology (UET), Taxila, Pakistan, in 2015. He is currently pursuing the Ph.D. degree in electronic science and technology with the Shenzhen Key Laboratory of Internet of Things Terminal Technology, Harbin Institute of Technology at Shenzhen, Shenzhen China. He was a full-time Research Scholar with Microwaves, Antennas and Propagation (MAP) Research Group, UET Taxila, from 2013 to 2015, and he was mainly engaged in research on frequency selective surfaces, band filtering and MIMO antenna arrays. In 2015, he joined National Radio and Telecommunication Corporation (NRTC), Pakistan, where he was with the RF & Microwave Section as an Antenna Design Engineer. His research interests include deep learning, audio signal processing, speaker recognition, and speech enhancement.

LU ZHANG was born in Shandong, China, in 1993. He received the B.S. degree in electronic science and technology from the Harbin Institute of Technology, Weihai Campus, Shandong, in 2016. He is currently pursuing the Ph.D. degree in electrical science and technology with the Shenzhen Key Laboratory of Internet of Things Terminal Technology, Harbin Institute of Technology at Shenzhen, Shenzhen. His research interests include digital speech and audio signal processing, speech enhancement, and microphone array signal processing.