Applied Acoustics 74 (2013) 770–781

Contents lists available at SciVerse ScienceDirect

Applied Acoustics

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

Noise estimation based on time–frequency correlation for speech enhancement

Wenhao Yuan, Jiajun Lin *, Wei An, Yu Wang, Ning Chen

School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

ARTICLE INFO

Article history:
Received 16 May 2012
Received in revised form 16 October 2012
Accepted 20 November 2012
Available online 8 January 2013

Keywords:
Noise estimation
Speech enhancement
Minimum search
Improved Minima Controlled Recursive Averaging

ABSTRACT

As a fundamental part of speech enhancement, noise estimation is particularly challenging in highly non-stationary noise environments. In this work, we propose an effective algorithm on the basis of the "Improved Minima Controlled Recursive Averaging (IMCRA)" with the objective to improve the performance of noise estimation. The main contributions of this work are: (i) in the algorithm, a rough decision about speech presence is proposed by calculating the autocorrelation and cross-channel correlation of the T–F (Time–Frequency) units; (ii) with this decision, we refine the smoothing parameters for the smoothing of noisy power spectrum and the recursive averaging in noise spectrum estimation as well as the weighting factor for the *a priori* SNR (Signal to Noise Ratio) estimation in the IMCRA; (iii) we improve the search of local minima during spectral bursts by adding a minimum search with a shorter window. Extensive experiments are carried out to evaluate the performance of our proposed algorithm. The experimental results illustrate that, compared with the IMCRA, the proposed approach significantly improves the accuracy of noise spectrum estimation and the quality of enhanced speech in the typical noise situations.

1. Introduction

Due to universal applicability and simplicity, single channel speech enhancement has been being a hot research spot of speech enhancement, for several years, that is an indispensable step in various fields, such as speech communication, speech coding and speech recognition in noisy environments [1–7]. As being a necessary step in most single channel speech enhancement algorithms, noise estimation significantly affects the performance of speech enhancement. Traditional noise estimation methods utilize voice activity detectors (VADs) for detecting the presence and absence of speech in noisy speech, and update the noise estimation during speech absence [8]. The VADs can achieve good results in stationary noise environments (e.g., white noise) because they mainly take advantage of the energy statistical properties and other characteristics of speech and noise signals. However, in real noise environments (e.g., factory noise), considering the rapid varying of energy statistical properties of the noise, VADs usually do not work well and result in losing track of the immediate changes of noise. Therefore, an urgent research work is to develop a more accurate and robust noise estimation algorithm such that it can update the noise power spectrum continuously without relying on the judgments of VADs.

Recently increasing efforts can be found in literature that focuses on noise estimation. Doblinger [9] updated the noise estimate continuously through tracking the spectral minima frame by frame. However, this method may tend to attenuate the speech spectrum, because it fails to differentiate between an increase in noise floor and an increase in the speech spectrum level. Hirsch and Ehrlicher [10] proposed a simple noise estimation method by using a first order recursive averaging, which updates the noise estimate by comparing the power spectrum of current frame to the noise estimate of past frames. However, this method fails to update the noise estimate in the case of abrupt increase of noise floor. Ris and Dupont [11] estimated the noise by combining the above techniques with narrow-band spectral analysis, while Stahl et al. [12] proposed a quantile-based noise estimation algorithm which filters out speech peaks and estimates the noise from the remaining spectrum with the use of a non-linear filter. However, the above two methods may fail to adapt fast to highly-varying noise. Martin [13] proposed a noise estimation algorithm, based on minimum statistics (MS), which tracks the minima values of a smoothed spectrum of the noisy speech over a finite window, and then multiplies the result by a bias factor to achieve the unbiased estimate of noise spectrum. The major drawback of this method is that the update of the noise spectrum spends more time than the duration of the minimum-search window when the noise floor increases abruptly. Cohen and Berdugo [14] proposed a minima controlled recursive averaging algorithm (MCRA). The MCRA searches the local minimum similarly to MS, and then compares the ratio of the noisy speech to the local minimum against a threshold to find the noise-only regions.

* Corresponding author. Tel./fax: +86 021 64253511.
E-mail addresses: ywhecust@126.com (W. Yuan), jjlin@ecust.edu.cn (J. Lin), anweiunl@gmail.com (W. An), rain198689@163.com (Y. Wang), nchen@ecust.edu.cn (N. Chen).

0003-682X/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.apacoust.2012.11.007

The noise estimate is updated by tracking the noise-only regions of the noisy speech spectrum. In [15], the improved MCRA (IMCRA) approach was proposed, which exploits a different method to track the noise-only regions based on the estimated speech presence probability. However, both the MCRA and IMCRA might take twice as much of the duration of the search window to adapt to a noise burst. Rangachari et al. [16] described a noise estimation algorithm based on voice activity detection, which updates the noise estimate in each frame. In [17], the method was further improved by the way of making the speech presence decision based on the ratio of noisy speech spectrum to its local minimum against a frequency-dependent threshold. This method reduces adaptation time to noise bursts, but it may occasionally attenuate some speech components.

In sum, the main drawbacks of these noise estimation algorithms are the inaccuracies for distinguishing speech from noise and the delayed responses to the abrupt increases of noise floor. In this work, we propose a novel noise estimation algorithm based on the IMCRA so that it can improve the tracking ability and the accuracy of noise estimate by introducing two feature functions that can effectively reflect the correlation of the signals.

(i) The T–F correlation is a significant feature to differentiate speech from noise. The IMCRA takes into account the correlation between consecutive frames and adjacent frequencies by carrying out the smoothing of noisy power spectrum, the recursive averaging in the noise spectrum estimation and the non-linear recursive procedure in the *a priori* SNR estimation. However, in the calculation of the speech presence probability based on the ratio of noisy speech spectrum to the local minima, the IMCRA mainly focuses on the comparison of the power spectrum in different frames, and does not entirely consider the relation between correlation features and speech presence.
In this paper, we first adopt the peripheral auditory processing in the CASA (Computational Auditory Scene Analysis) [18] to make an initial processing of the noisy speech signal, and convert the signal to T–F units distributing in 128 channels. Then, the autocorrelation and the cross-channel correlation of the T–F units are calculated, which can reflect the coherent relationship among different frames and different frequency bins. By comparing the calculation of two correlation functions to fixed thresholds, a rough decision about speech presence is made.

(ii) The smoothing of noisy power spectrum and the update of noise spectrum estimation has great impact on the resultant noise estimate. In the IMCRA, to implement the smoothing and the update, some smoothing parameters are introduced and defined as constants. However, the optimal values of the parameters should vary according to the speech presence probability rather than be certain constants. In addition, the estimate of the *a priori* SNR is one of the critical factors to affect the speech enhancement. In the IMCRA, the calculation of the speech presence probability used in noise estimation depends on the estimate of the *a priori* SNR, therefore the estimation of the *a priori* SNR is also one of the critical factors to affect the noise estimate. The weighting factor for the *a priori* SNR estimation in the IMCRA is a constant.
From the above discussion, it is worthy to be mentioned that that the choice of these parameters is closely related to the accuracy of the noise estimate. This paper refines above parameters to some values depending on the speech presence of each frame based on the rough decision we made previously.

(iii) To reduce the delay of minimum search in the case of spectral bursts, we add a parallel minimum search that is similar to the original one but using a shorter length window. And we combine it with the obtained rough decision about speech presence to improve the minimum search when spectral bursts are detected.

The rest of this paper is organized as follows. Section 2 briefly reviews the IMCRA method, and Section 3 presents the rough decision about speech presence based on correlation features. In Section 4, the decision is used to adjust the smoothing parameters for the smoothing of the noisy power spectrum and the recursive averaging in the noise spectrum estimation, and also the weighting factor for the *a priori* SNR estimation. Section 5 introduces a new minimum search method by combining the decision with two parallel minimum searches. Section 6 evaluates the performance of the proposed algorithm compared to the IMCRA. Finally, the conclusions are given in Section 7.

2. Review of improved minima controlled recursive averaging

Let y denote an observed noisy signal in the time domain, which is the sum of a clean speech x and an uncorrelated additive noise d. By applying the short-time Fourier transform (STFT), we have

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

in the time–frequency domain, where k represents the frequency bin index, and l is the frame index.

In the IMCRA, the noise is estimated by recursively averaging past spectral power values of the noisy measurement during periods of speech absence and holding the estimate during speech presence [15]. Under speech presence uncertainty, the conditional speech presence probability is employed, and the recursive averaging can be obtained by

$$\bar{\lambda}_d(k,l+1) = \tilde{\alpha}_d(k,l)\bar{\lambda}_d(k,l) + [1-\tilde{\alpha}_d(k,l)]|Y(k,l)|^2 \tag{2}$$

where

$$\tilde{\alpha}_d(k,l) \triangleq \alpha_d + (1-\alpha_d)p(k,l) \tag{3}$$

is a time-varying frequency-dependent smoothing parameter. $\alpha_d (0 < \alpha_d < 1)$ denotes a smoothing parameter, and $p(k,l)$ is the conditional speech presence probability. Through introducing a bias compensation factor β, the noise estimate is given by

$$\hat{\lambda}_d(k,l+1) = \beta \cdot \bar{\lambda}(k,l+1) \tag{4}$$

In order to calculate the speech presence probability, two iterations of smoothing and minimum tracking are carried out. The time smoothing in the first iteration is performed by a first-order recursive averaging as

$$S(k,l) = \alpha_s S(k,l-1) + (1-\alpha_s)S_f(k,l) \tag{5}$$

where α_s $(0 < \alpha_s < 1)$ is a smoothing parameter, and $S_f(k,l)$ is obtained by the frequency smoothing of the noisy power spectrum

$$S_f(k,l) = \sum_{i=-\omega}^{\omega} b(i)|Y(k-i,l)|^2 \tag{6}$$

where b denotes a normalized window function. The time smoothing in the second iteration is similar to that in the first iteration, and utilizes the same smoothing parameter.

The method of minimum search in the IMCRA is in accordance with that used in MS [13], the local minima of $S(k,l)$ is searched within a finite window of length D, for each frequency bin

$$S_{\min}(k,l) \triangleq \min\{S(k,l')|l-D+1 \leqslant l' \leqslant l\}. \tag{7}$$

To reduce the computational complexity, the window of D samples is generally divided into U sub-windows of V samples ($D = UV$).

The computation of the speech presence probability also requires an estimate of the *a priori* SNR. In the IMCRA, the *a priori* SNR is commonly estimated by

$$\hat{\xi}(k,l) = \alpha G_{H_1}^2(k,l-1)\gamma(k,l-1) + (1-\alpha)\max\{\gamma(k,l)-1,0\} \quad (8)$$

where α is a weighting factor, $G_{H_1}(k,l)$ is the spectral gain function, and $\gamma(k,l) \triangleq |Y(k,l)|^2/\lambda_d(k,l)$ represents the *a posteriori* SNR.

3. Correlation-based rough decision about speech presence

The T–F correlation is a significant feature for the judgment of speech presence. The IMCRA takes into account the correlation between consecutive frames and adjacent frequencies by carrying out the smoothing of noisy power spectrum, the recursive averaging in the noise spectrum estimation and the non-linear recursive procedure in the *a priori* SNR estimation. However, it does not consider the direct relation between correlation features and speech presence. In this section, by introducing two correlation functions, we propose a rough decision about speech presence based on the correlation features.

3.1. Peripheral auditory processing

To extract correlation features, the preprocessing measure is required for the noisy speech signal. Similar to the peripheral auditory processing in the CASA [18], we decompose the noisy speech into a series of T–F units by using a 128-channel gammatone filterbank, where the impulse response of a gammatone filter is described in [18] as

$$g(t) = \begin{cases} t^{r-1}\exp(-2\pi bt)\cos(2\pi ft), & \text{if } t \geqslant 0 \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

where $r = 4$ is the order of the filter, b is the equivalent rectangular bandwidth, and f is the center frequency of the filter which is quasi-logarithmically spaced from 80 to 5000 Hz. Further, the response of each gammatone filter is transduced into the hair cell output $h(c,n)$ with the Meddis model of inner hair cells [19], where c represents the channel index, and n the time step.

3.2. Correlation function

To calculate the correlation of T–F units, we introduce two functions: the autocorrelation function and the cross-channel correlation function.

3.2.1. Autocorrelation function

Due to the capability of effectively detecting the periodicity of signals, autocorrelation function has been widely exploited in designing pitch detection algorithms. By calculating the autocorrelation of each filter response, the correlation between two adjacent frames in the same channel can be derived. The autocorrelation function of the hair cell output $h(c,n)$ is given in [18] by

$$A_H(c,l,\tau) = \frac{1}{N_c}\sum_{n=0}^{N_c-1} h(c,lT-n)h(c,lT-n-\tau) \quad (10)$$

where the output $A_H(c,l,\tau)$ is three-dimensional, l represents the time frame, and $\tau \in [0, 12.5\text{ ms}]$ represents the time delay, in which the maximum delay 12.5 ms corresponds to the minimum frequency of the speech at 80 Hz. N_c is the corresponding number of samples.

3.2.2. Cross-channel correlation function

The cross-channel correlation reflects the relations between any two adjacent channels. Literature [18] indicates that the high degree of the cross-channel correlation between adjacent filter channels implies that they are likely to respond to the same source. Therefore, the cross-channel correlation can be used as a significant feature to the extraction of speech from noisy speech. The cross-channel correlation function is calculated in [18] as

$$C_H(c,l) = \sum_{\tau=0}^{L-1}\widehat{A}_H(c,l,\tau)\widehat{A}_H(c+1,l,\tau) \quad (11)$$

where $\widehat{A}_H(c,l,\tau)$ is the normalized result of the autocorrelation function $A_H(c,l,\tau)$, and $L = 201$ corresponds to the maximum delay 12.5 ms.

3.3. Rough decision about speech presence

By using the autocorrelation at $\tau = 0$ and the cross-channel correlation, a binary value is calculated as

$$M_{sp}(c,l) = \begin{cases} 1, & \text{if } A_H(c,l,0) > \theta_H^2 \text{ and } C_H(c,l) > \theta_C \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

where the thresholds θ_H and θ_C are chosen to be the same as in [18]. The binary values of different channels in the same frame are accumulated, which can be written as

$$SM_{sp}(l) = \sum_c M_{sp}(c,l) \quad (13)$$

By comparing the sum against a threshold θ_P, the following decision can be made

$$P_{SM}(l) = \begin{cases} 1, & \text{if } SM_{sp}(l) > \theta_P \\ \left(\frac{SM_{sp}(l)}{\theta_P}\right)^2, & \text{otherwise} \end{cases} \quad (14)$$

P_{SM} is related to the energy of each T–F unit and the correlation among the T–F units at different frames or different channels, and also can make a rough decision about speech presence. Fig. 1 shows an example of P_{SM} for a segment of speech corrupted by Gaussian white noise at 0 dB SNR. The waveforms of clean speech and noisy speech are also provided in Fig. 1 as references. It is observable that P_{SM} achieves bigger values during periods of speech presence; while the smaller values are achieved during speech absence.

It is known that the judgment of speech presence is a critical factor in the IMCRA, therefore the rough decision about speech presence calculated from the correlation will contribute to the improvement of the noise estimation. In this paper, we utilize P_{SM} to refine the smoothing parameters for the smoothing of noisy power spectrum and the recursive averaging in the noise spectrum estimation, and also refine the weighting factor that controls the tradeoff between noise reduction and speech distortion in the *a priori* SNR estimation. We also try to solve the delay problem of minimum search in the case of spectral bursts by using the rough decision and adding a searching window with shorter length.

4. Parameter refinement

In this section, three key parameters related to the noise estimation in the IMCRA are refined, including: the smoothing parameter for the smoothing of noisy power spectrum, the smoothing parameter for the recursive averaging in the noise spectrum estimation, and the weighting factor in the *a priori* SNR estimation.

4.1. Refining parameter for the smoothing of noisy power spectrum

In the smoothing of noisy power spectrum (Eq. (5)), the IMCRA utilizes the smoothing parameter α_s as a constant to control the rate of the noisy power spectrum of the current frame to the

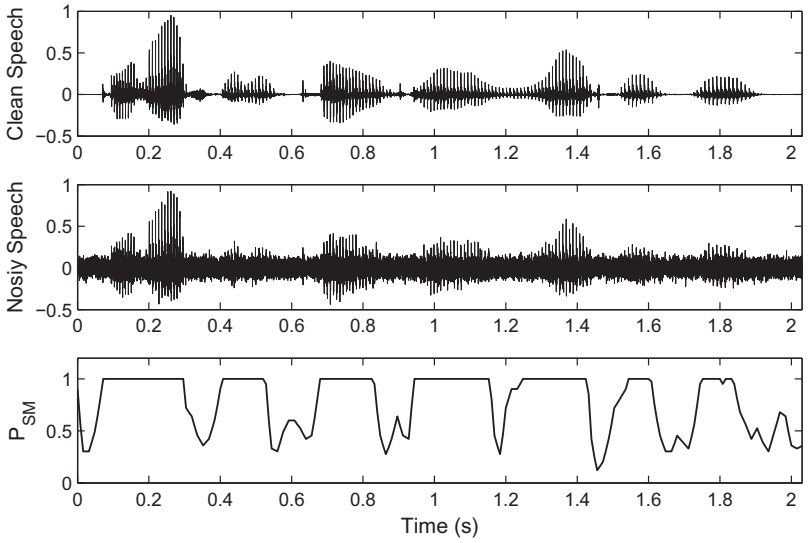

Fig. 1. Example of P_{SM} for a segment of speech corrupted by Gaussian white noise at 0 dB SNR. (From top to bottom) Waveform of clean speech, waveform of noisy speech and the rough decision about speech presence.

smoothing results. However, when the current frame contains speech component, the pace of changes of noisy power spectrum is accelerated, to retain more speech spectral component of the current frame in the smoothing results, this rate should be increased [20,21]. Let

$$\alpha'_s(l) = \alpha_{sp} + (1 - P_{SM}(l))\beta_s \tag{15}$$

where $0 < \alpha_{sp} < 1$ is a constant, and $\beta_s = 0.05$. Replace α_s with $\alpha'_s(l)$, we have

$$S(k,l) = \alpha'_s(l)S(k,l-1) + (1 - \alpha'_s(l))S_f(k,l) \tag{16}$$

In Eq. (15), the bigger $P_{SM}(l)$ is, the smaller $\alpha'_s(l)$ is. Thereby the noisy power spectrum of the lth frame $S_f(k,l)$ accounts for bigger proportion of $S(k,l)$, and the speech spectral component of the lth frame can be retained better. $\alpha'_s(l)$ for a segment of speech corrupted by Gaussian white noise at 0 dB SNR is illustrated in Fig. 2.

Similarly, when carrying out the second iteration of smoothing, $\alpha'_s(l)$ is also used to replace α_s as

$$\widetilde{S}(k,l) = \alpha'_s(l)\widetilde{S}(k,l-1) + (1 - \alpha'_s(l))\widetilde{S}_f(k,l) \tag{17}$$

4.2. Refining smoothing parameter in noise spectrum estimation

The noise spectrum estimation in the IMCRA is controlled by the smoothing parameter $\tilde{\alpha}_d(k,l)$ in the recursive averaging. According to Eq. (2), when speech is present in the lth frame, $\tilde{\alpha}_d(k,l)$ should choose a big value so as to reduce the speech component of $Y(k,l)$ in the estimated noise spectrum; while when speech is absent in the l th frame, $\tilde{\alpha}_d(k,l)$ should choose a small value such that the noise spectrum estimate can be updated timely.

As shown in Eq. (3), IMCRA exploits a constant α_d combined with the speech presence probability $p(k,l)$ to control the magnitude of $\tilde{\alpha}_d(k,l)$. To make $\tilde{\alpha}_d(k,l)$ more effectively adaptive to the states (present and absent) of the speech, based on the rough decision in Eq. (14), we define a time-varying parameter as

$$\alpha'_d(l) = \min\{\alpha_d + (1 - \alpha_d)P'_{SM}(l), \alpha'_{d\max}\} \tag{18}$$

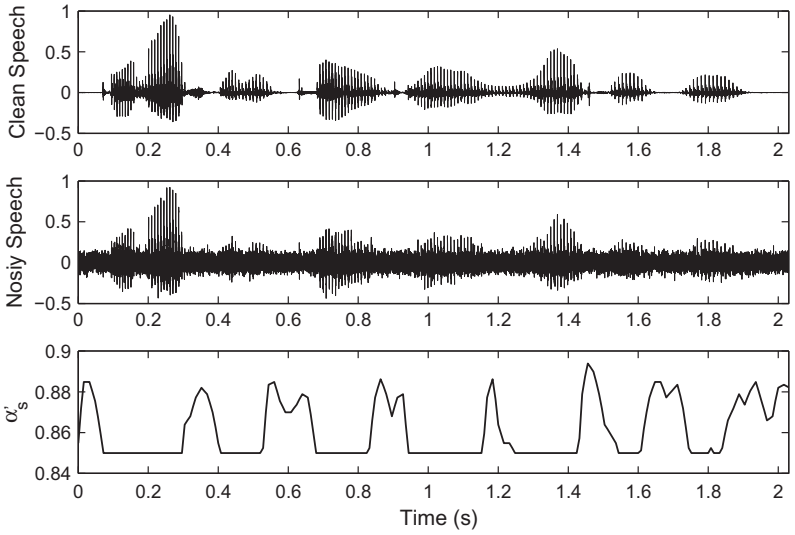

Fig. 2. Example of $\alpha'_s(l)$ for a segment of speech corrupted by Gaussian white noise at 0 dB SNR. (From top to bottom) Waveform of clean speech, waveform of noisy speech and the refined smoothing parameter of noisy power spectrum.

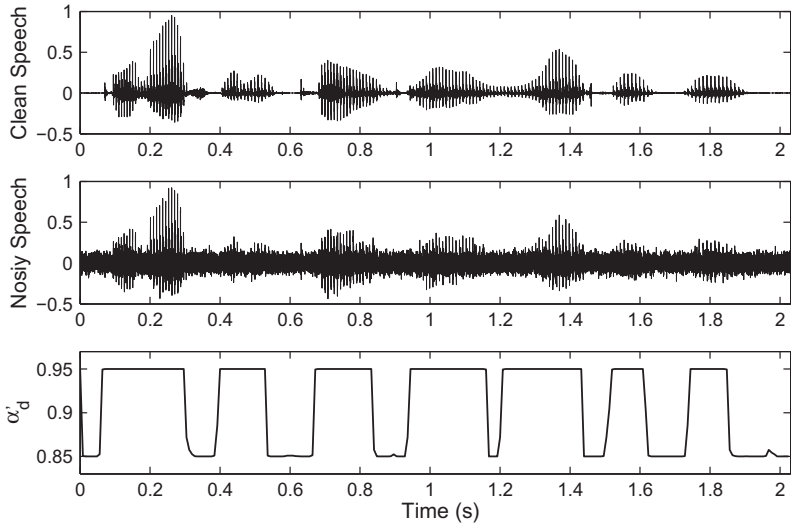

Fig. 3. Example of $\alpha'_d(l)$ for a segment of speech corrupted by Gaussian white noise at 0 dB SNR. (From top to bottom) Waveform of clean speech, waveform of noisy speech and the refined smoothing parameter in noise spectrum estimation.

to replace the constant α_d by $\alpha'_d(l)$ in Eq. (3), where

$$P'_{SM}(l) = \frac{1}{1+\exp(-K_P(A_P P_{SM}(l) - B_P))} \tag{19}$$

is the sigmoid type transformation of $P_{SM}(l)$, and the parameters $K_P = -2$, $A_P = 14$, $B_P = 11$ are determined through a large number of experiments. Obviously, the value of $\alpha'_d(l)$ increases as the value of $P_{SM}(l)$ increases, and vice verse. Fig. 3 shows the waveform of a speech segment, the waveform of its corresponding noisy speech corrupted by Gaussian white noise at 0 dB SNR and $\alpha'_d(l)$ for the noisy speech segment. It is observable that $\alpha'_d(l)$ reaches bigger value when the speech is present, while smaller value of $\alpha'_d(l)$ is achieved when the speech is absent.

By replacing α_d in Eq. (3) with $\alpha'_d(l)$, we have

$$\tilde{\alpha}'_d(k,l) \triangleq \alpha'_d + (1-\alpha'_d)p(k,l) \tag{20}$$

Finally, we derive a new equation of recursive averaging for the noise spectrum estimation by using $\tilde{\alpha}'_d(k,l)$ instead of $\tilde{\alpha}_d(k,l)$ in Eq. (2),

$$\bar{\lambda}_d(k,l+1) = \tilde{\alpha}'_d(k,l)\bar{\lambda}_d(k,l) + [1-\tilde{\alpha}'_d(k,l)]|Y(k,l)|^2 \tag{21}$$

4.3. Refining weighting factor in the a priori SNR estimation

In the *a priori* SNR estimation, the weighting factor α controls the tradeoff between noise reduction and speech distortion [15] and is set as a constant in the IMCRA. In fact, in the cases of speech presence and transition from non-speech frames to speech frames, the rapid changes in speech power spectrum cause significant changes in SNR. At these points, if α takes a large value, the estimate of the *a priori* SNR will contain too much of the previous frame and cannot response to the changes immediately. Hence, α is expected to obtain a small value in these cases, so that the estimation of the SNR can track the changes of speech power spectrum to reduce the speech distortion. And conversely, when speech is absent, α is expected to obtain a large value to reduce musical noise [22]. We define

$$\alpha'(l) = (1-P_{SM}(l))\alpha_2 + P_{SM}(l)\alpha_1 \tag{22}$$

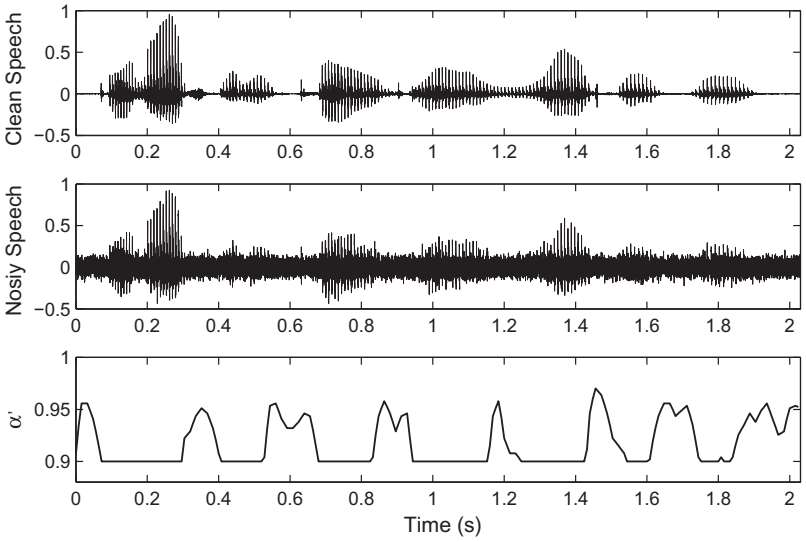

Fig. 4. Example of $\alpha'(l)$ for a segment of speech corrupted by Gaussian white noise at 0 dB SNR. (From top to bottom) Waveform of clean speech, waveform of noisy speech and the refined weighting factor in the *a priori* SNR estimation.

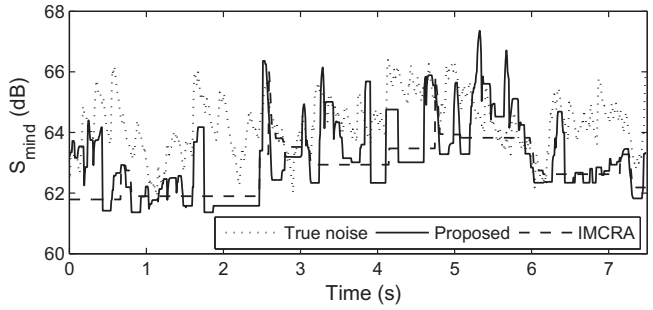

Fig. 5. Example of local minima using the proposed algorithm and IMCRA for a segment of speech corrupted by Gaussian white noise at 0 dB SNR at frequency bin $k = 45$. (From top to bottom) Waveform of noisy speech, comparison of local minima in a subband.

Table 1
Parameter values used in our proposed algorithm.

$\theta_H = 50$	$\theta_C = 0.985$	$\theta_P = 20$	$\theta_I = 12$
$\alpha'_{d\max} = 0.95$	$\alpha_1 = 0.9$	$\alpha_2 = 0.98$	

where $0 < \alpha_1 < \alpha_2 < 1$. In Eq. (22), the larger $P_{SM}(l)$ leads to the smaller $\alpha'(l)$, while $\alpha'(l)$ would take the larger value for smaller $P_{SM}(l)$. $\alpha'(l)$ for a segment of speech corrupted by Gaussian white noise at 0 dB SNR is given as an example in Fig. 4, from which we can notice that the variation of $\alpha'(l)$ follows the presence of speech in noisy speech. Using $\alpha'(l)$ instead of α in Eq. (8), we derive a new estimation for the *a priori* SNR as

$$\hat{\xi}(k,l) = \alpha'(l)G^2_{H_1}(k,l-1)\gamma(k,l-1) + (1-\alpha'(l))\max\{\gamma(k,l)-1,0\} \tag{23}$$

5. Improved minimum search

According to the theory of the minimum search in the IMCRA, it may lead $D + V$ samples delay into the update of local minima during spectral bursts. To reduce the update delay, we append a minimum search which is similar to that in the IMCRA but using a window with shorter length D', and $D' \approx 0.1D$, $D' = U'V'$. The minimum search is given by

$$S'_{\min}(k,l) \triangleq \min\{S(k,l')|l - D' + 1 \leqslant l' \leqslant l\}. \tag{24}$$

To detect the spectral bursts, we propose the following rough decision

$$I'(k,l) = \begin{cases} 0, & \text{if } \gamma_{\min}(k,l) < \gamma_0 \text{ and } \zeta(k,l) < \zeta_0 \\ 1, & \text{otherwise} \end{cases} \tag{25}$$

where $\gamma_{\min}(k,l) \triangleq |Y(k,l)|^2/(B_{\min}S_{\min}(k,l))$, and $\zeta(k,l) \triangleq S(k,l)/(B_{\min}S_{\min}(k,l))$. When $I'(k,l) = 1$, we can determine a spectral burst in the kth frequency bin of the lth frame. When $\sum_k I'(k,l) > \theta_I$, that is that the number of the frequency bins reporting spectral burst is greater than the threshold in the lth frame, we can determine that the lth frame reports a spectral burst. The reason of a spectral burst may be a sudden increase of the speech power spectrum or a sudden increase of the noise power spectrum. If a spectral burst is due to a sudden increase of the noise power spectrum, $S_{\min}$ should be updated immediately to track the change of noise floor. To solve the delay problem in minimum search during spectral bursts, this paper proposes a new minimum search method as

$$S_{\min}(k,l) = \begin{cases} P_{SM}(l)S_{\min}(k,l) + (1-P_{SM}(l))S'_{\min}(k,l), & \text{if } \sum_k I'(k,l) > \theta_I \\ S_{\min}(k,l), & \text{otherwise} \end{cases} \tag{26}$$

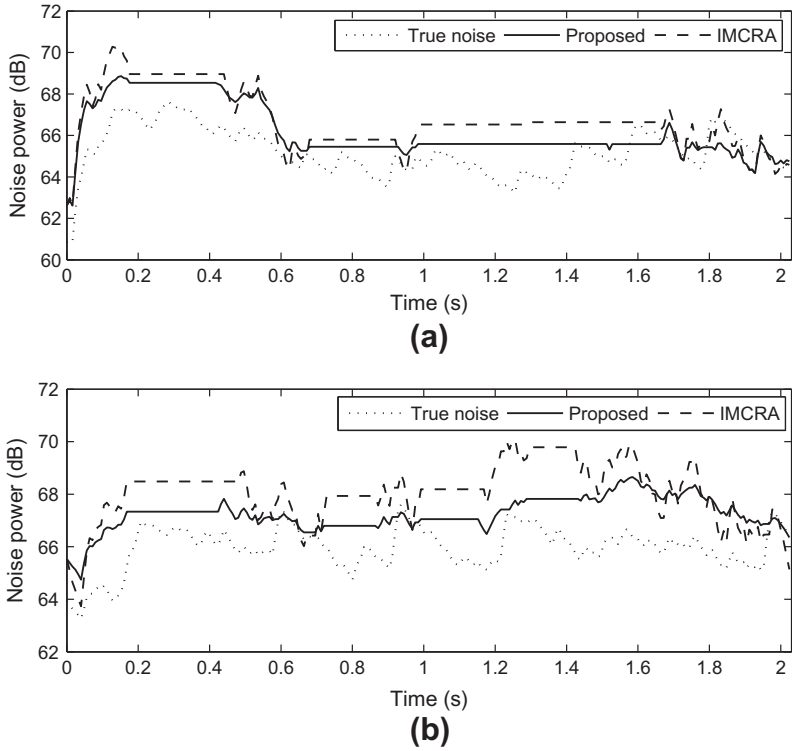

Fig. 6. Example of noise estimation in a subband. (a) Comparison of the noise spectrum (for frequency bin $k = 45$) estimated by the proposed algorithm and IMCRA for a segment of speech corrupted by Gaussian white noise at 0 dB SNR. (b) Comparison of the noise spectrum (for frequency bin $k = 45$) estimated by the proposed algorithm and IMCRA for a segment of speech corrupted by F-16 cockpit noise at 0 dB SNR.

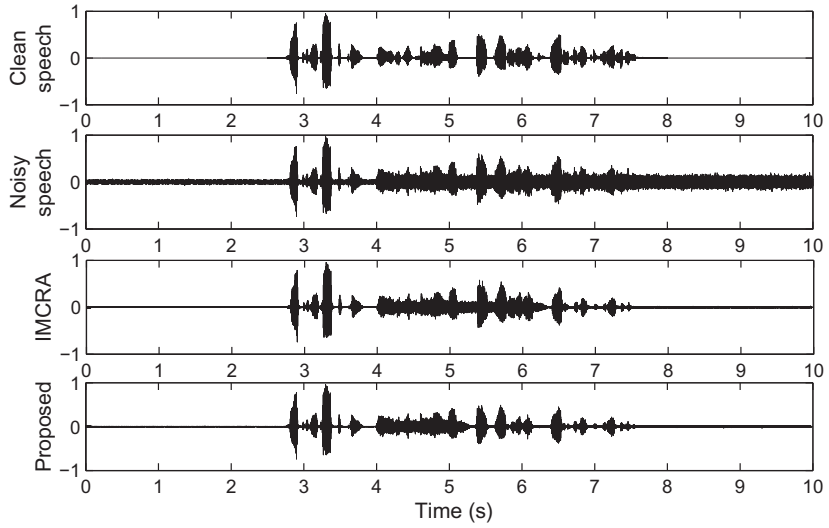

Fig. 7. Example of noise estimation for a segment of speech corrupted by Gaussian white noise with 10 dB SNR before $T = 4$ s and 0 dB SNR after $T = 4$ s. (From top to bottom) Waveform of clean speech, waveform of noisy speech, waveform of speech enhanced by IMCRA and waveform of speech enhanced by the proposed algorithm.

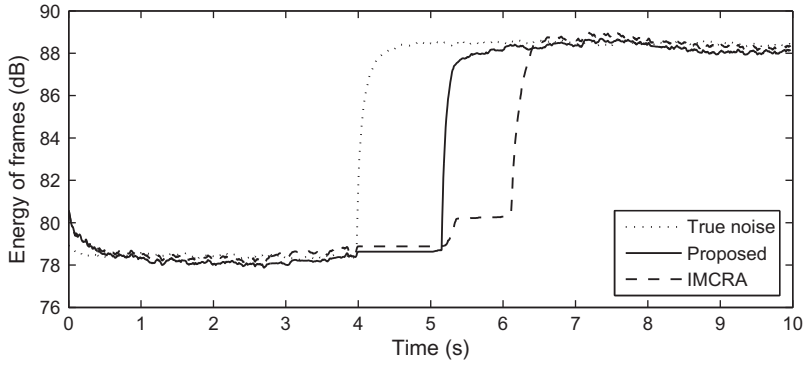

Fig. 8. Energy of frames of true noise and estimated noise using the proposed algorithm and the IMCRA.

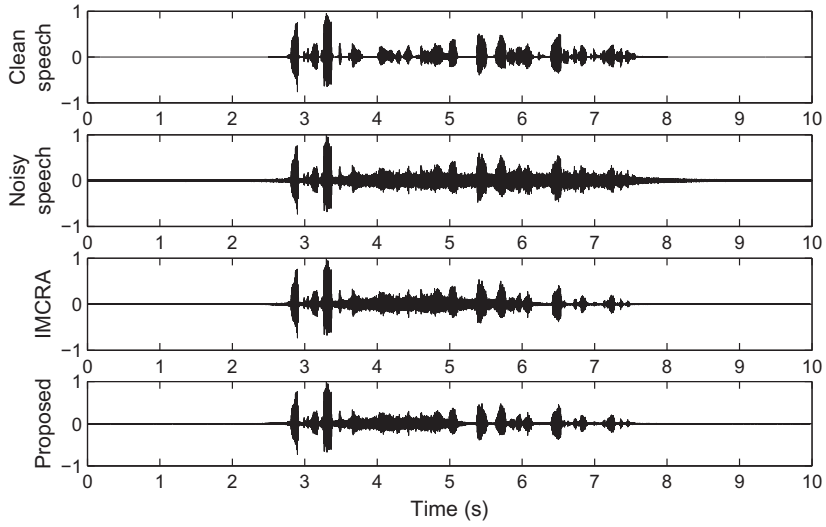

Fig. 9. Example of noise estimation for a segment of speech corrupted by Gaussian white noise with SNRs in Eq. (27). (From top to bottom) Waveform of clean speech, waveform of noisy speech, waveform of speech enhanced by the IMCRA and waveform of speech enhanced by the proposed algorithm.

where $S_{\min}$ and $S'_{\min}$ are calculated by Eqs. (7) and (24) respectively. In Eq. (26), when the rough decision about speech presence $P_{SM}(l)$ is small, which implies the spectral bursts are more likely caused by sudden increases of noise spectrum, S_{mind} will be decided mainly

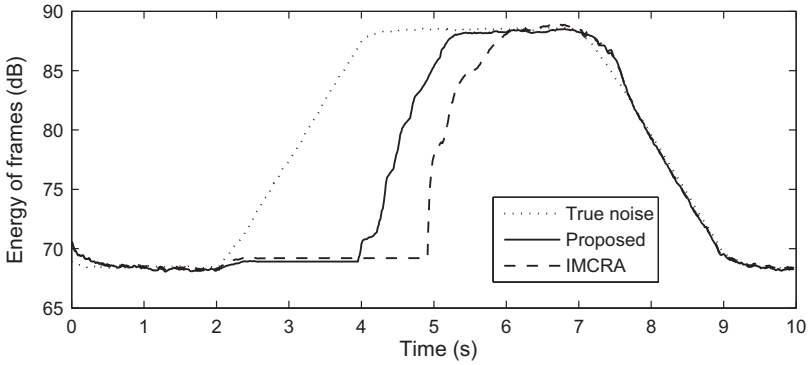

Fig. 10. Energy of frames of true noise and estimated noise using the proposed algorithm and the IMCRA.

by $S'_{\min}$, which leads to a more immediate search of local minima. When $P_{SM}(l)$ is large, which implies the spectral bursts are more likely caused by sudden increases of speech power spectrum, $S_{\min d}$ will be decided mainly by $S_{\min}$, and the over-estimation of noise can be avoided. With this, we can achieve the timely and accurate tracking of the noise floor.

Local minima at frequency bin $k = 45$ for a speech segment degraded by Gaussian white noise at 0 dB SNR are plotted in Fig. 5, where the solid line and the dashed line represent the local minima computed by the proposed method and the IMCRA respectively. The true noise spectrum is also given as comparison with dotted line. From the figure we can see that our proposed method is more sensitive to the changes of noise floor, compared with the IMCRA, our proposed method can detect much more noise bursts and adapt to them rapidly.

6. Performance evaluation

To evaluate the performance, we compared our proposed algorithm with the IMCRA in noise estimation and speech enhancement. The typical values of the parameters used in our proposed algorithm are given in Table 1. The choices of other parameters are the same as those in the IMCRA.

The clean speech signals used in our experiments are from six different utterances, which, half from male speakers and half from female speakers, are taken from TIMIT database [23]. The noise signals used in our experiments include six different types of noise taken from the Noisex92 database [24]: Gaussian white noise, pink noise, HF channel noise, factory floor noise 2, F-16 cockpit noise and Volvo noise. Both the speech and the noise signals are sampled at 16 kHz.

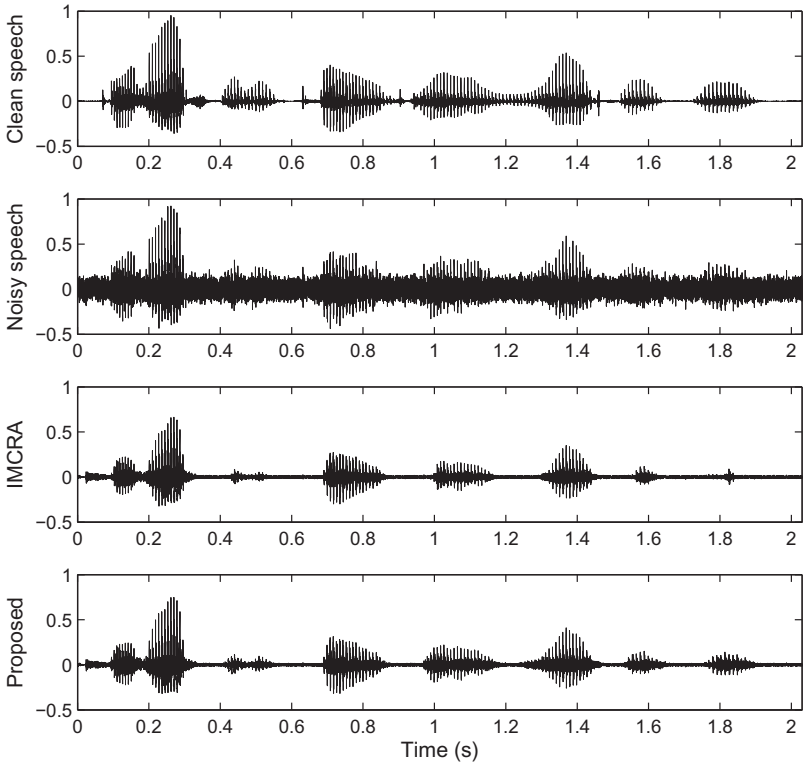

Fig. 11. Example of speech enhancement for a segment of speech corrupted by 0 dB Gaussian white noise. (From top to bottom) Waveform of clean speech, waveform of noisy speech, waveform of speech enhanced by IMCRA and waveform of speech enhanced by the proposed algorithm.

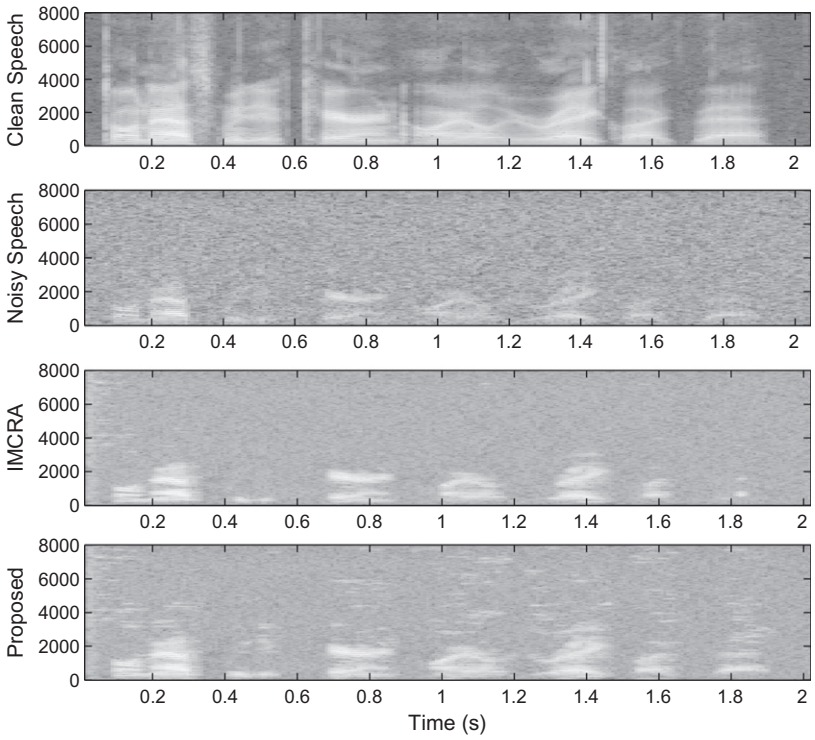

Fig. 12. Example of speech enhancement for a segment of speech corrupted by 0 dB Gaussian white noise. (From top to bottom) Spectrogram of clean speech, spectrogram of noisy speech, spectrogram of speech enhanced by IMCRA and speech enhanced by the proposed algorithm.

Table 2
Average segmental SNR for various noise type and levels, obtained by the proposed algorithm and IMCRA. The values in bold represent the best results.

Noise type	Global SNR (dB)	Proposed (dB)	IMCRA (dB)
White	0	**2.053**	1.364
	5	**4.503**	3.880
	10	**7.215**	6.494
Pink	0	**1.537**	0.770
	5	**4.142**	3.493
	10	**7.004**	6.226
Hfchannel	0	**1.606**	1.157
	5	**3.925**	3.657
	10	**6.744**	6.201
Volvo	0	**11.150**	10.837
	5	**13.265**	12.683
	10	**15.749**	15.107
Factory2	0	**2.823**	2.491
	5	**6.056**	5.649
	10	**8.413**	8.032
F-16	0	**1.812**	1.436
	5	**4.400**	3.811
	10	**7.206**	6.667

Table 3
Average PESQ score for various noise type and levels, obtained by the proposed algorithm and IMCRA. The values in bold represent the best results.

Noise type	Global SNR (dB)	Proposed	IMCRA
White	0	**2.119**	1.838
	5	**2.489**	2.307
	10	**2.842**	2.756
Pink	0	**2.125**	1.894
	5	**2.564**	2.378
	10	**2.784**	2.662
Hfchannel	0	**1.935**	1.682
	5	**2.260**	2.184
	10	**2.695**	2.622
Volvo	0	**3.437**	3.389
	5	**3.613**	3.571
	10	**3.696**	3.647
Factory2	0	**2.331**	2.220
	5	**2.784**	2.710
	10	**3.005**	2.984
F16	0	**2.165**	2.007
	5	**2.574**	2.394
	10	**2.842**	2.765

6.1. Performance evaluation on noise estimation

We evaluate the noise estimation performance of our proposed algorithm by carrying out two sets of experiments. In the first set of experiments, we compare the performance of noise estimation in subbands. And in the second set of experiments, we investigate the tracking capability of the proposed algorithm in the cases of noise bursts and increases of noise floor.

6.1.1. Performance of noise estimation in subbands

Fig. 6a and b gives the noise estimate at frequency bin $k = 45$ of two segments of noisy speech, respectively, which are corrupted by 0 dB Gaussian white noise and F-16 cockpit noise. The true noise taken as the recursively smoothed periodogram of the noise $|D(k,l)|^2$ is also given as reference. In both of the two subfigures, it is seen that the solid line from the noise estimated by the proposed algorithm is much closer to the dotted line from the true noise than the dashed line from the noise estimated by the IMCRA, which indicates that the proposed algorithm obtains more accurate noise estimate than the IMCRA. It also can be noticed that the proposed algorithm is more sensitive to the changes of noise power and tracks them more accurately. The same conclusions can be drawn in other subbands from the experiments.

Table 4
Average WSS distance for various noise type and levels, obtained by the proposed algorithm and IMCRA. The values in bold represent the best results.

Noise type	Global SNR (dB)	Proposed	IMCRA
White	0	**44.544**	45.727
	5	**35.944**	38.172
	10	**29.599**	33.011
Pink	0	**54.766**	56.912
	5	**42.811**	44.584
	10	**33.126**	35.744
Hfchannel	0	**53.997**	54.895
	5	47.283	**47.209**
	10	**36.034**	37.293
Volvo	0	**25.315**	27.241
	5	**22.113**	24.095
	10	**17.752**	19.347
Factory2	0	**52.286**	52.616
	5	**37.059**	38.580
	10	**32.076**	32.438
F16	0	59.808	**58.940**
	5	47.523	**47.358**
	10	**36.222**	37.767

6.1.2. Tracking capability

In this part, in order to examine the tracking capability, we assume a 10-s period of noisy speech, in which the speech is corrupted by Gaussian white noise with 10 dB SNR before $T = 4$ s and 0 dB SNR after $T = 4$ s, as shown in Fig. 7 (noisy speech). Fig. 8 illustrates the energy of frames of true noise (dotted line) and noise estimated by the proposed algorithm (solid line) and the IMCRA (dashed line). It is obvious that our proposed algorithm can response to the noise bursts almost 1 s earlier than the IMCRA. Comparing the waveforms of the enhanced speech using the proposed algorithm and the IMCRA in Fig. 7, we can also notice the distinction about the delay to noise burst.

In the other case, we assume a non-stationary Gaussian white noise similar to that in [15], which is simulated by increasing or decreasing the level of stationary Gaussian white noise at the same rate for a period of several seconds. Fig. 9 illustrates the waveform of a 10-s segment of speech that is corrupted by the non-stationary Gaussian white noise with the SNRs in the following equation:

$$SNR = \begin{cases} 20 & 0 \leqslant T < 2 \\ 20 - 10(T-2) & 2 \leqslant T < 4 \\ 0 & 4 \leqslant T < 7 \\ 10(T-7) & 7 \leqslant T < 9 \\ 20 & T \geqslant 9 \end{cases} \quad (27)$$

Fig. 9 also illustrates the waveforms of the original clean speech and the enhanced speech using the IMCRA and the proposed algorithm.

The energy of frames of true noise and estimated noise are depicted in Fig. 10. The solid line represents the estimated noise energy using the proposed algorithm, the dashed line represents the estimated noise energy using the IMCRA, and the dotted line represents the true noise energy. Due to the varying parameters adaptive to speech presence and the new minimum search method, the proposed algorithm can response to the increasing noise floor much faster than the IMCRA. For the decreasing noise floor, both the proposed algorithm and the IMCRA can response to it immediately. From the waveforms of enhanced speech in Fig. 9, we can reach the same conclusion more visually.

6.2. Performance of Speech Enhancement

To obtain the noisy speech for speech enhancement experiments, the speech signals are degraded by the six kinds of noise

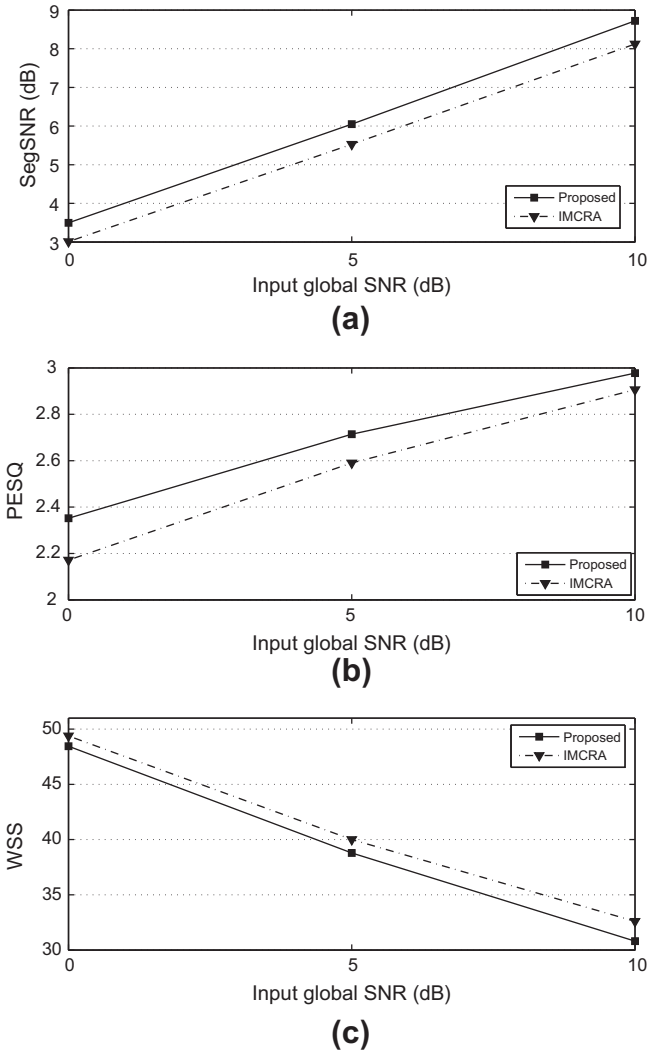

Fig. 13. Comparison of three measures for different input SNRs (global SNR). (a) Average segmental SNR of enhanced speech using the proposed algorithm and IMCRA for different input SNRs. (b) Average PESQ score of enhanced speech using the proposed algorithm and IMCRA for different input SNRs. (c) Average WSS distance of enhanced speech using the proposed algorithm and IMCRA for different input SNRs.

with 0 dB, 5 dB and 10 dB global SNRs respectively, where the global SNR is defined as

$$SNR_{Global} = 10 \log \frac{\sum_n x(n)^2}{\sum_n (y(n) - x(n))^2} \quad (28)$$

The noise spectrum estimation is executed by the proposed algorithm and the IMCRA, and the spectral gain function of the OM-LSA [14] estimator (Eqs. (29) and (30)) is used to implement speech enhancement to the $3 \times 6 \times 6$ segments of noisy speech.

$$G_{H_1}(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} \exp\left(\frac{1}{2}\int_{\upsilon(k,l)}^{\infty} \frac{e^{-t}}{t}\, dt\right) \quad (29)$$

$$G(k,l) = \{G_{H_1}(k,l)\}^{p(k,l)} G_{\min}^{1-p(k,l)} \quad (30)$$

A segment of noisy speech corrupted by 0 dB Gaussian white noise and the speech enhanced by the proposed algorithm and the IMCRA are given in Figs. 11 and 12. From both the waveforms (Fig. 11) and the spectrograms (Fig. 12), it can be seen clearly that the speech enhanced by the proposed algorithm is closer to the original clean speech.

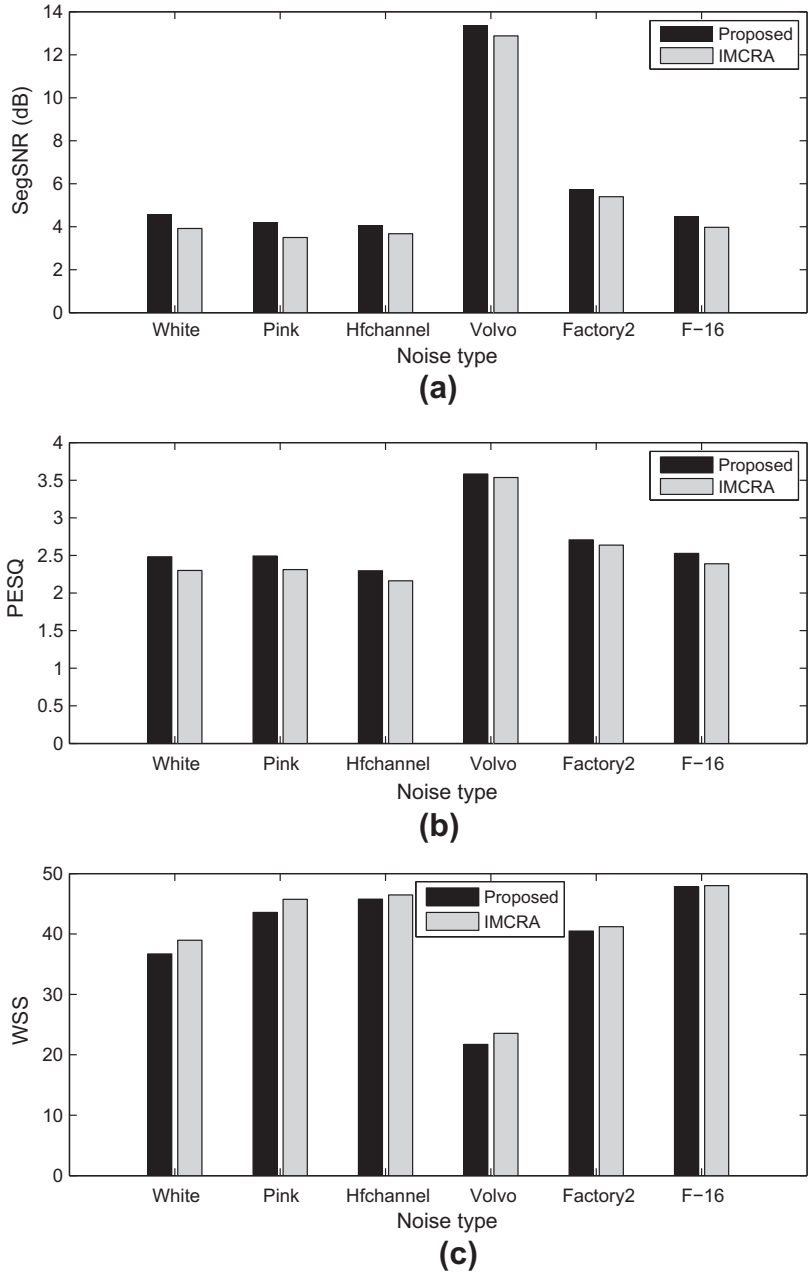

Fig. 14. Comparison of three measures for different noise types. (a) Average segmental SNR of enhanced speech using the proposed algorithm and IMCRA for different noise types. (b) Average PESQ score of enhanced speech using the proposed algorithm and IMCRA for different noise types. (c) Average WSS distance of enhanced speech using the proposed algorithm and IMCRA for different noise types.

The quantitative comparisons of the performance of speech enhancement between the proposed algorithm and the IMCRA are obtained by evaluating the segmental SNR (SegSNR) [25], the Perceptual Evaluation of Speech Quality (PESQ) measure [26] and the Weighted Spectral Slope (WSS) distance measure [27] under the given environmental conditions.

Segmental SNR is a significant measure for speech quality, which is much closer to the actual speech quality than the global SNR. It asserts that the larger value of segmental SNR corresponds to the higher speech quality. The segmental SNR is defined in [25] as

$$\mathrm{SegSNR} = \frac{10}{M}\sum_{m=0}^{M-1}\log\frac{\sum_{n=Nm}^{Nm+N-1}x^2(n)}{\sum_{n=Nm}^{Nm+N-1}(x(n)-\hat{x}(n))^2} \tag{31}$$

where $\hat{x}(n)$ is the enhanced speech signal. M and N are the number of frames in the signal and the window length in samples, respectively.

PESQ is the speech quality evaluation criteria described by ITU-T P.862, which is an objective speech quality evaluation method. The value of PESQ ranges from -0.5 to 4.5, and the larger value indicates the higher quality of speech.

The WSS distance measure calculates the weighted difference between the spectral slopes in each frequency bin, where the spectral slope describes the difference between adjacent spectral magnitudes in decibels. The smaller WSS distance implies that the enhanced speech is closer to the original clean speech. The WSS measure evaluated in this paper is given in [27] by

$$d_{WSS} = \frac{1}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K}W(j,m)(S_c(j,m)-S_p(j,m))^2}{\sum_{j=1}^{K}W(j,m)} \tag{32}$$

where $W(j,m)$ represent the weights computed as per literature [27], M is the frame number, $K = 25$ is the frequency bin number,

and $S_c(j,m)$, $S_p(j,m)$ are the spectral slopes for the jth frequency bin of the mth frame of the clean and enhanced speech, respectively.

According to noise types and levels, the noisy speech signals are divided into 18 groups, and each group is composed of six segments of noisy speech which are from different utterances and degraded by the same type of noise with the same global SNR. The noisy speech signals are processed by the proposed algorithm and the IMCRA respectively. Computing the average segmental SNR, average PESQ score and average WSS distance of the enhanced speech within each group, the comparisons of the segmental SNR, PESQ and WSS between the proposed algorithm and the IMCRA are shown in Tables 2–4. In Table 2, it is seen that the proposed algorithm achieves larger segmental SNR values in all of the 18 groups of data, and the average improvement is 0.536 dB. In Table 3, it is seen that the proposed algorithm achieves a larger PESQ scores in all of the 18 groups of data, and the average improvement is 0.125. In Table 3, it is seen that the proposed algorithm achieves smaller WSS distances in 15 out of 18 groups of data, and the average improvement is 1.315.

The improvement of three measures to the IMCRA for different input global SNRs is shown in Fig. 13, and the bar charts in Fig. 14 illustrates the improvement for different noise types. From these two figures, it is obvious that the proposed algorithm is superior to the IMCRA algorithm for various noise types and levels, which agrees with the results in Tables 2–4.

7. Conclusions

The IMCRA is a commonly used algorithm for estimating the noise power spectrum. This paper improves the IMCRA algorithm through the introduction of two feature functions reflecting the correlation between T–F units. After the peripheral auditory processing of noisy speech signal, we calculate the autocorrelation and cross-channel correlation between T–F units and further make a rough decision about speech presence. The smoothing parameters for the smoothing of noisy power spectrum and the recursive averaging in the noise spectrum estimation are refined with the use of the rough decision, and the weighting factor for the *a priori* SNR estimation is also improved. Moreover, by appending a minimum search with a shorter length of window and combing it with the rough decision about speech presence, we improve the minimum search during spectral bursts. The proposed algorithm has been tested and compared to the conventional IMCRA in various noise types and levels. Three objective measures are employed in the evaluation, and the results demonstrate that the proposed algorithm achieves better performance under most of tested environmental conditions.

Although our proposed algorithm shows superiority over the IMCRA, there are still many challenges to face. One of them is that we can only make a rough decision about speech presence, which prevents us to distinguish speech and non-speech frames accurately. A more precise decision about speech presence based on the features is needed. We are working to exploit more features to distinguish human speech from noise.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Nos. 60903186, 61271349).

References

[1] Ming J, Srinivasan R, Crookes D. A corpus-based approach to speech enhancement from nonstationary noise. IEEE Trans Audio Speech Lang Process 2011;19(4):822–36.
[2] Gunawan TS, Ambikairajah E, Epps J. Perceptual speech enhancement exploiting temporal masking properties of human auditory system. Speech Commun 2010;52(5):381–93.
[3] Ding H, Lu J, Qiu X, Xu B. An adaptive speech enhancement method for siren noise cancellation. Appl Acoust 2004;65(4):385–99.
[4] Paliwal K, Wojcicki K, Schwerin B. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. Speech Commun 2010;52(5):450–75.
[5] Lu C-T, Tseng K-F. A gain factor adapted by masking property and SNR variation for speech enhancement in colored-noise corruptions. Comput Speech Lang 2010;24(4):632–47.
[6] Lee W, Song J-H, Chang J-H. Minima-controlled speech presence uncertainty tracking method for speech enhancement. Signal Process 2011;91(1):155–61.
[7] Lu C-T. Enhancement of single channel speech using perceptual-decision-directed approach. Speech Commun 2011;53(4):495–507.
[8] Sohn J, Kim NS, Sung W. A statistical model-based voice activity detection. IEEE Signal Process Lett 1999;6:1–3.
[9] G. Doblinger, Computationally efficient speech enhancement by spectral minima tracking in subbands. In: Proc Eurospeech, 1995. p. 1513–6.
[10] Hirsch H, Ehrlicher C. Noise estimation techniques for robust speech recognition. In: Proc IEEE ICASSP, vol. 1, 1995. p. 153–6.
[11] Ris C, Dupont S. Assessing local noise level estimation methods: application to noise robust ASR. Speech Commun 2001;34(1-2):141–58.
[12] Stahl V, Fischer A, Bippus R. Quantile based noise estimation for spectral subtraction and wiener filtering. In: Proc IEEE ICASSP, vol. 3, 2000. p. 1875–8.
[13] Martin R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans Speech Audio Process 2001;9(5):504–12.
[14] Cohen I, Berdugo B. Noise estimation by minima controlled recursive averaging for robust speech enhancement. IEEE Signal Process Lett 2002;9(1):12–5.
[15] Cohen I. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans Speech Audio Process 2003;11(5):466–75.
[16] Rangachari S, Loizou P, Hu Y. A noise estimation algorithm with rapid adaptation for highly nonstationary environments. In: Proc IEEE ICASSP, vol. 1. 2004. p. 305–8.
[17] Rangachari S, Loizou PC. A noise-estimation algorithm for highly non-stationary environments. Speech Commun 2006;48(2):220–31.
[18] Wang D, Brown GJ. Computational auditory scene analysis. USA: IEEE Press; 2006.
[19] Meddis R. Simulation of auditory–neural transduction: further studies. J Acoust Soc Am 1988;83:1056–63.
[20] Talmon R, Cohen I, Gannot S. Single-channel transient interference suppression with diffusion maps. IEEE Trans Audio Speech Lang Process 2012;99:1.
[21] Hirszhorn A, Dov D, Talmon R, Cohen I. Transient interference suppression in speech signals based on the OM-LSA algorithm. In: Proc 13th international workshop on acoustic echo and noise control, 2012.
[22] Cappe O. Elimination of the musical noise phenomenon with the ephraim and Malah noise suppressor. IEEE Trans Speech Audio Process 1994;2(2):345–9.
[23] Fisher W, Doddington G, Marshall GK. The DARPA speech recognition research database: specification and status. In: Proc DARPA Speech Recognition Workshop, 1986. p. 93–100.
[24] Varga A, Steeneken HJ. 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 1993;12(3):247–51.
[25] Hansen JHL, Pellom BL. An effective quality evaluation protocol for speech enhancement algorithms. In: Proc the international conference on spoken language processing, 1998. p. 2819–22.
[26] Hu Y, Loizou P. Evaluation of objective quality measures for speech enhancement. IEEE Trans Audio Speech Lang Process 2008;16(1):229–38.
[27] Klatt D. Prediction of perceived phonetic distance from critical-band spectra: a first step. In: Proc IEEE ICASSP, vol. 7. 1982. p. 1278–81.