EURASIP Journal on Audio,
Speech, and Music Processing

RESEARCH Open Access

An adaptive a priori SNR estimator for perceptual speech enhancement

Lara Nahma[1*], Pei Chee Yong[2], Hai Huyen Dam[1] and Sven Nordholm[1]

Abstract

In this paper, an adaptive averaging a priori SNR estimation employing critical band processing is proposed. The proposed method modifies the current decision-directed a priori SNR estimation to achieve faster tracking when SNR changes. The decision-directed estimator (DD) employs a fixed weighting with the value close to one, which makes it slow in following the onsets of speech utterances. The proposed SNR estimator provides a means to solve this issue by employing an adaptive weighting factor. This allows an improved tracking of onset changes in the speech signal. As a consequence, it results in better preservation of speech components. This adaptive technique ensures that the weighting between the modified decision-directed a priori estimate and the maximum likelihood a priori estimate is a function of the speech absence probability. The estimate of the speech absence probability is modeled by a sigmoid function. Furthermore, a critical band mapping for the short-time Fourier transform analysis-synthesis system is utilized in the speech enhancement to achieve less musical noise. In addition, to evaluate the ability of the a priori SNR estimation method in preserving speech components, we proposed a modified objective measurement known as modified hamming distance. Evaluations are performed by utilizing both objective and subjective measurements. The experimental results show that the proposed method improves the speech quality under different noise conditions. Moreover, it maintains the advantage of the DD approach in eliminating the musical noise under different SNR conditions. The objective results are supported by subjective listening tests using 10 subjects (5 males and 5 females).

Keywords: Single-channel speech enhancement, A priori SNR estimation, Decision-directed approach, Adaptive smoothing factor, Auditory system

1 Introduction

Noise suppression and speech enhancement are essential techniques employed in many products, for instance, mobile phones, hearing aids, and assistive listening devices. Particularly, hearable devices have been poised to assist people with difficulties in hearing in social environments [1]. For noise suppression and speech enhancement to work in the environments where acoustic noise becomes more intrusive, it is vital to maintaining weak speech components while still balancing the amount of noise reduction. Accordingly, techniques that can enhance speech signals while preserving weak speech components under a large variety of acoustic scenarios are key to successful products [2–4]. In this context, it is important to consider not only the speech but also the quality of noise after suppression. Unnatural sounding background noise is bothersome for users of hearable devices or hearing aids.

Traditionally, speech enhancement techniques have been utilizing the frequency domain for processing where the short-time Fourier transform (STFT) has been used as a tool to process the input data using frame-based oversampling techniques [3, 5–7]. When deploying STFT, the bandwidth is constant for each frequency bin, which is not the case for the human auditory system. Thus, a natural extension has been to use human auditory models in speech enhancement to improve the speech quality and intelligibility [8–11].

The human auditory spectrum model consists of a bank of bandpass filters, which follows a spectral bark scale or the so-called critical bands [11, 12]. In [11], a standard subtractive speech enhancement method is presented to eliminate the musical artifacts in very noisy situations. The masking properties of the auditory system are utilized to compute the subtraction parameter. In [13], a spectral

*Correspondence: l.alibreesm@postgrad.curtin.edu.au
[1] Department of Electrical Engineering, computing and Mathematical Sciences, Curtin University, Perth, Australia
Full list of author information is available at the end of the article

subtraction noise reduction method is proposed using a spatial weighting technique based on the inhibitory property of the auditory system, which results in improving the estimated speech while reducing the musical noise.

Speech enhancement algorithms calculate a gain function, which is in most cases a function of a posteriori signal to noise ratio (SNR) or a combination of a posteriori and a priori SNR [14]. One exemplary speech enhancement algorithm is the spectral subtraction (SS) method proposed by Boll [15]. This algorithm is the most commonly used mainly due to its straightforward implementation and low computational complexity. In this method, a clean speech estimate is obtained by subtracting an estimated noise power spectrum from the noisy speech power spectrum while keeping the phase of the degraded speech signal. The spectral subtraction method embeds erroneous estimation of noise statistics resulting in an annoying artifact in the estimated speech signal commonly known as musical noise, which can be masked using perceptual thresholds [11, 16].

In contrast, the minimum mean-square error log spectral amplitude (MMSE-LSA) estimator proposed by Ephraim [17] avoids the appearance of the musical noise artifact. This estimator uses a priori SNR estimation based on a decision -directed estimation, which involves a weighted sum of two terms, the a priori SNR estimate from the previous frame and the maximum likelihood (ML) SNR estimate from the current frame. This estimation technique reduces the variance of the a priori SNR estimates particularly during noise frames, and as a result, the musical noise artifact is eliminated [18]. However, the emphasis of the previous frame in the DD estimation has a consequence that it leads to a slow adaptation towards speech onsets and offsets. Moreover, as DD approach depends on the a priori SNR estimation in the previous frame, an extra one frame delay is obtained during speech transients and results in a degradation of the speech quality [7].

The a priori SNR estimation algorithm has been improved in many ways, e.g., Breithaupt et al. [19] proposed the temporal cepstrum smoothing (TCS) technique for speech enhancement. This technique improves the accuracy of the a priori SNR estimation by exploiting the a priori knowledge of speech and noise signal and selectively smoothing the maximum likelihood estimate in the cepstral domain. This allows the preservation of speech components while simultaneously achieving high noise attenuation. However, this method has limitations under low SNR conditions where the noise components cannot be separated from the speech components. Suhadi [20] suggested a data-driven technique employing two trained neural networks to estimate the a priori SNR, one for speech and one for noise. The use of neural networks requires a substantial training process for estimating the a priori SNR since the proposed method is not a robust estimator under different noise environments, which results in a degradation of the estimated speech quality under non-stationary noise conditions. Plapous [21] presented a two-step noise reduction technique (TSNR) to refine the estimation of the a priori SNR and increase the estimator adaptation speed. The main disadvantage when using this TSNR method is its sensitivity to the selection of the gain function. A different choice of the gain function gives very different estimation results [22, 23]. A modified decision-directed approach (MDD) proposed by Yong et al. [7] matches the current noisy speech spectrum with the current a priori SNR estimate rather than the delayed one. This reduces the one frame delay for speech onsets, but the tracking speed of the a priori SNR estimation is still slow compared to the true SNR change since the recursive smoothing factor is constant and close to one.

In this paper, we extend the research in [24], which includes an improved a priori SNR estimation based on modeling the speech absence probability with a sigmoid function. This sigmoid function was used to control the adaptation speed of the a priori SNR estimation. The sigmoid function operates as an adaptive weighting function that emphasizes either the DD term or the ML estimate in the a priori SNR estimate update. The rationale used when developing the weighting function was that for positive SNR values; the a priori and the a posteriori SNR estimates are almost the same. Accordingly, by adding flexibility to select either of the two terms for SNR values below or above a certain threshold, we provide a way to emphasize both estimates. By utilizing a threshold and the sigmoid shape, an improved adaptation of the a priori SNR estimate is obtained.

The choice of gain function plays an important role since it is included in the DD estimation resulting in different performance. Previously, only the Wiener gain function was considered. In this work, we propose an improved a priori SNR estimation [24] using different gain functions, namely, Wiener filter (WF) [25] and MMSE-LSA gain function [17]. A new evaluation technique referred to as the modified Hamming distance (HD) has also been proposed. In common objective measures, speech components are not emphasized since they have small amplitudes or small energy. The proposed modified Hamming distance is based on voice activity detection (VAD) decision information in each time-frequency bin. Since this information is binary, data scaling that depend on amplitude or energy is avoided; thus, we can compare to ideal VAD decisions. Also in this work, we utilize a critical band mapping for an STFT analysis-resynthesis system in the speech enhancement framework for human perceptual processing. Moreover, the utilized critical band processing helps to reduce computational complexity since it combines K FFT frequency bins into I critical bands instead ($I \ll K$).

The remainder of this paper is organized as follows. In Section 2, a single-channel speech enhancement framework with critical band processing is developed. Section 3 shows the decision-directed based a priori SNR estimators. Section 4 develops the proposed a priori SNR estimation approach together with an investigation on the effect of the key parameters of the sigmoid function. Section 5 demonstrates the evaluation methodology. Section 6 presents the experimental results and discussion while Section 7 concludes the paper.

2 Critical band speech enhancement

A natural way to process speech signals is to use a perceptual filter bank [26]. By employing the inhibitory property of the human auditory system and combining with the speech enhancement algorithms [11], the performance of the speech processing system can be improved. There are many perceptual frequency warping scales used for speech processing [27, 28]. In this work, we employed a bark scale filter bank with a non-uniform resolution and incorporated it in a speech enhancement framework with the proposed a priori SNR estimation method. We assume that the noise and speech are additive and uncorrelated; thus, the noisy speech signal is given by

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

where $s(n)$ and $v(n)$ denote the clean speech signal and noise, respectively. The block diagram for critical band speech processing is described in Fig. 1. In the sequel, we will outline the details of the processing.

In the first step, the noisy signal is transformed to the time-frequency domain by applying STFT with K frequency bins

$$Y(k,m) = S(k,m) + V(k,m) \tag{2}$$

where k is the frequency bin index and m is the time frame index. Then, in order to transform the output from the STFT $Y(k,m)$ into the critical band, an analytical function is used to express the transformation between frequency f (in Hz) and critical band z (in bark scale), which is defined by [29]

$$f = 600 \sinh\left(\frac{z}{6}\right). \tag{3}$$

The noisy spectrum is expressed in terms of the critical band numbers i and frame index m by combining the FFT frequency bins into I critical bands as follows:

$$Y_{\mathrm{CB}}(i,m) = \sum_{k=1}^{K/2+1} M(i,k)\,|Y(k,m)| \tag{4}$$

where $i = [1, 2, \cdots, I]$. The number of critical bands I is chosen with respect to the bark scale [29]. Here, $M(i,k)$ are the critical bandpass filter coefficients, which are defined as

$$M(i,k) = \begin{cases} 10^{(z(k) - z_{\mathrm{c}}(i) + 0.5)} & z(k) < z_{\mathrm{c}}(i) - 0.5 \\ 1 & z_{\mathrm{c}}(i) - 0.5 < z(k) < z_{\mathrm{c}}(i) + 0.5 \\ 10^{-2.5(z(k) - z_{\mathrm{c}}(i) - 0.5)} & z(k) > z_{\mathrm{c}}(i) + 0.5 \end{cases} \tag{5}$$

where $z_c(i)$ represents the center frequency of the ith critical band. A MATLAB implementation of the bark scale critical band processing is described in [30]. The main task of the speech enhancement scheme is to enhance the speech signal by applying a specific spectral gain function to the noisy spectrum. Let $\mathbf{G}_{\mathrm{CB}}(m)$ denotes the gain vector in the critical band for the mth frame

$$\mathbf{G}_{\mathrm{CB}}(m) = [\, G_{\mathrm{CB}}(1,m), G_{\mathrm{CB}}(2,m), ..., G_{\mathrm{CB}}(I,m)]^T .$$

There are many different gain functions proposed in the literature. Common gain function often can be expressed as a function of the a priori SNR $\xi(i,m)$, such as the WF method, which can be defined as [25]

$$G_{\mathrm{WF,CB}}(i,m) = \frac{\xi(i,m)}{1 + \xi(i,m)} \tag{6}$$

with $\xi(i,m)$ denoting the a priori signal-to-noise ratio SNR, which is defined as

$$\xi(i,m) = \frac{\lambda_s(i,m)}{\lambda_v(i,m)} \tag{7}$$

where $\lambda_v(i,m) = E\left[|V(i,m)|^2\right]$ and $\lambda_s(i,m) = E\left[|S(i,m)|^2\right]$ are the power spectral density of noise and clean speech, respectively.

MMSE-LSA [17] is another widely used speech estimator, which is obtained by minimizing the logarithm of the mean square error between original and enhanced speech spectra, and can be defined as a function of the priori SNR and the posteriori SNR, given by

$$G_{\mathrm{LSA,CB}}(i,m) = \frac{\xi(i,m)}{1 + \xi(i,m)} \exp\left\{ \frac{1}{2} \int_{\nu_k}^{\infty} \frac{e^{-t}}{t} dt \right\} \tag{8}$$

where the lower bound ν_k of the integral is given by

$$\nu_k = \frac{\xi(i,m)}{1 + \xi(i,m)} \gamma(i,m) \tag{9}$$

and $\gamma(i,m)$ denotes the a posteriori SNR defined as

$$\gamma(i,m) = \frac{|Y_{\mathrm{CB}}(i,m)|^2}{\lambda_v(i,m)}. \tag{10}$$

Once the gain vector $\mathbf{G}_{\mathrm{CB}}(m)$ in a critical band is calculated, it is interpolated back to the gain vector in the STFT domain $\mathbf{G}(m)$ through an interpolation matrix $\mathbf{A}$,

$$\mathbf{G}(m) = \mathbf{A}\mathbf{G}_{\mathrm{CB}}(m) \tag{11}$$

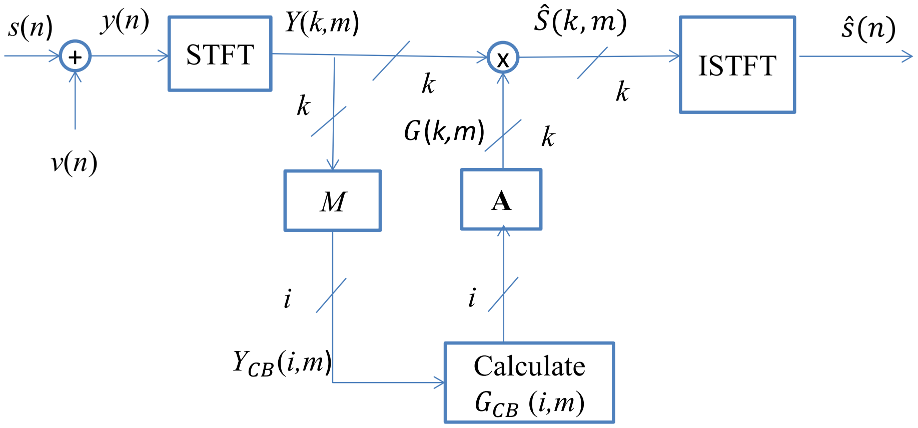

Fig. 1 Block diagram for the critical band processing

where the $\mathbf{A}$ matrix can be defined by least square approximation as $\mathbf{A} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T$ and $\mathbf{M}$ denotes the matrix with elements $M(i,k)$. From empirical findings, better results are obtained by simplifying the reconstruction matrix as

$$\mathbf{A} = \text{diag}\left(\frac{1}{\mathbf{1M}}\right)\mathbf{M}^T$$

where $\mathbf{1}$ is $1 \times I$ row vector. The estimated speech in the STFT domain is then reconstructed by applying the interpolated gain function $G(k,m)$ on the noisy signal in Eq. (2)

$$\hat{S}(k,m) = G(k,m)Y(k,m). \tag{12}$$

Finally, the speech estimate is obtained by taking the inverse STFT of the enhanced speech and using the overlap-add method

$$\hat{s}(n) = \text{ISTFT}\left(\hat{S}(k,m)\right). \tag{13}$$

3 Conventional a priori SNR estimation

In many speech enhancement algorithms, a priori SNR estimation is a dominant part of the gain function calculation as in Eqs. (6) and (8). Inaccuracies in the estimation of the a priori SNR can lead to audible speech distortion and musical noise. The state-of-the-art method to estimate the a priori SNR from noisy speech while avoiding musical noise is the DD approach [31]. In this method, the a priori SNR estimation is expressed as a weighting average of the amplitude estimate at the previous frame and the maximum likelihood estimate of the a priori SNR at the current frame. This method is defined by

$$\hat{\xi}_{\text{DD}}(i,m) = \beta\frac{|\hat{S}(i,m-1)|^2}{\hat{\lambda}_{\text{v}}(i,m-1)} + (1-\beta)P\left[\hat{\gamma}(i,m)-1\right] \tag{14}$$

where $\hat{S}(i,m-1)$ and $\hat{\lambda}_{\text{v}}(i,m-1)$ denote the amplitude estimate and the noise estimate at the previous frame, respectively. P is the half wave rectification to keep the a priori SNR value positive, and $0 < \beta < 1$ denotes a weighting factor that controls the trade-off between the a priori SNR from previous frame and the posteriori SNR at current frame, which can be defined as

$$\beta = \exp(-R/f_s t_s) \tag{15}$$

where R is the frame rate, t_s and f_s denote the time averaging constant and the sampling frequency, respectively. By setting the weighting factor close to 1, two different behaviors of the a priori SNR estimation can be observed as explained in [18]. In the noise frames, the a priori SNR estimate corresponds to a scaled version of the a posteriori SNR since the second term of the DD approach is equal to zero. Thus, a priori SNR estimation can be expressed by

$$\hat{\xi}_{\text{DD}}^{\downarrow}(i,m) \approx \beta G_{\text{CB}}^2(i,m-1)\hat{\gamma}(i,m-1).$$

This behavior reduces the variations in the a priori SNR estimate and thus reduces the amount of musical noise produced. In the frames with speech onsets, the a priori SNR follows the a posteriori SNR from the preceding frame as given by

$$\begin{aligned}\hat{\xi}_{\text{DD}}^{\uparrow\uparrow}(i,m) &= \beta\frac{G_{\text{CB}}^2(i,m-1)|Y_{\text{CB}}(i,m-1)|^2}{\hat{\lambda}_{\text{v}}(i,m)} \\ &\quad + (1-\beta)P\left[\hat{\gamma}(i,m)-1\right] \\ &\approx \beta G_{\text{CB}}^2(i,m-1)\hat{\gamma}(i,m-1) \\ &\quad + (1-\beta)P\left[\hat{\gamma}(i,m)-1\right]\end{aligned}$$

where the second term that indicates the ML estimate would only have little impact on the estimation process since β is close to 1. In this case, the tracking of change in the a priori SNR estimate is slow since the a priori SNR estimation mainly depends on the posteriori SNR estimation in the previous frame. This behavior can lead to speech transient distortion. In order to overcome this problem, the authors in [7] proposed a modified decision-directed (MDD) approach. In that method, the a priori SNR estimate at the current frame is matched with the a posteriori SNR in the current frame instead of the previous one. Thus, the one-frame delay is reduced, which results in less speech distortion compared to the conventional DD approach. The MDD a priori SNR estimate is given by

$$\hat{\xi}_{\mathrm{MDD}}(i,m) = \beta \frac{G^2_{\mathrm{CB}}(i,m-1)\,|Y_{\mathrm{CB}}(i,m)|^2}{\hat{\lambda}_{\mathrm{v}}(i,m)} + (1-\beta)P\big[\hat{\gamma}(i,m)-1\big]. \tag{16}$$

In addition, to maintain the advantage of the DD approach in eliminating the musical noise, the magnitude square of the noisy signal has been smoothed by using first-order recursive smoothing procedure as given by [7] to reduce the variance of the a priori SNR estimate. The first-order recursive averaging of the noisy signal is given by

$$\lambda_y(i,m) = \alpha_y\lambda_y(i,m-1) + (1-\alpha_y)\,|Y_{\mathrm{CB}}(i,m)|^2 \tag{17}$$

where α_y is a smoothing constant. The smoothed $|Y_{\mathrm{CB}}(i,m)|^2$ is replacing the instantaneous power estimate in the a posteriori SNR Eq. (10).

4 Proposed a priori SNR estimation

The drawback of the MDD approach is that the fix weighting factor β in Eq. (15) reduces the influence from the second term towards the a priori SNR update resulting in a scaled down a priori SNR estimate when compared to the true a priori SNR. In light of this, we can conclude that the fix weighting factor β gives low variability of the gain function during noise-only periods but does not provide a fast change of the gain function when a speech utterance comes. Thus, it is desirable to replace the fix weighting factor β with an adaptive weighting factor $\beta(i,m)$.

Recognizing that the speech absence probability is a key for the weighting according to Eq. (16), we model the speech absence probability based on a sigmoid function. As a remark, if the cumulative distribution function (CDF) is a sigmoid function, the probability density function (pdf) is similar to a Gaussian pdf but with larger tails, which is plausible for speech applications. The sigmoid consists of two parameters, σ to control transition speed and ρ to determine the threshold of active speech signal and noise [32]. The selection of these parameter values is based on the observation that the a priori SNR equals the posterior SNR for high SNRs. An adaptive weighting function $\hat{\beta}(i,m)$ is proposed based on the a posteriori SNR and is given by

$$\hat{\beta}(i,m) = \frac{\beta_0}{1+\exp[-\sigma(\tilde{\gamma}(i,m)-\rho)]} \tag{18}$$

where β_0 is a constant slightly larger than β. The modified a priori SNR estimation approach is then defined by

$$\hat{\xi}_{\mathrm{prop}}(i,m) = \hat{\beta}(i,m)\frac{G^2_{\mathrm{CB}}(i,m-1)\,|Y_{\mathrm{CB}}(i,m)|^2}{\hat{\lambda}_v(i,m)} + (1-\hat{\beta}(i,m))P\big[\tilde{\gamma}(i,m)-1\big] \tag{19}$$

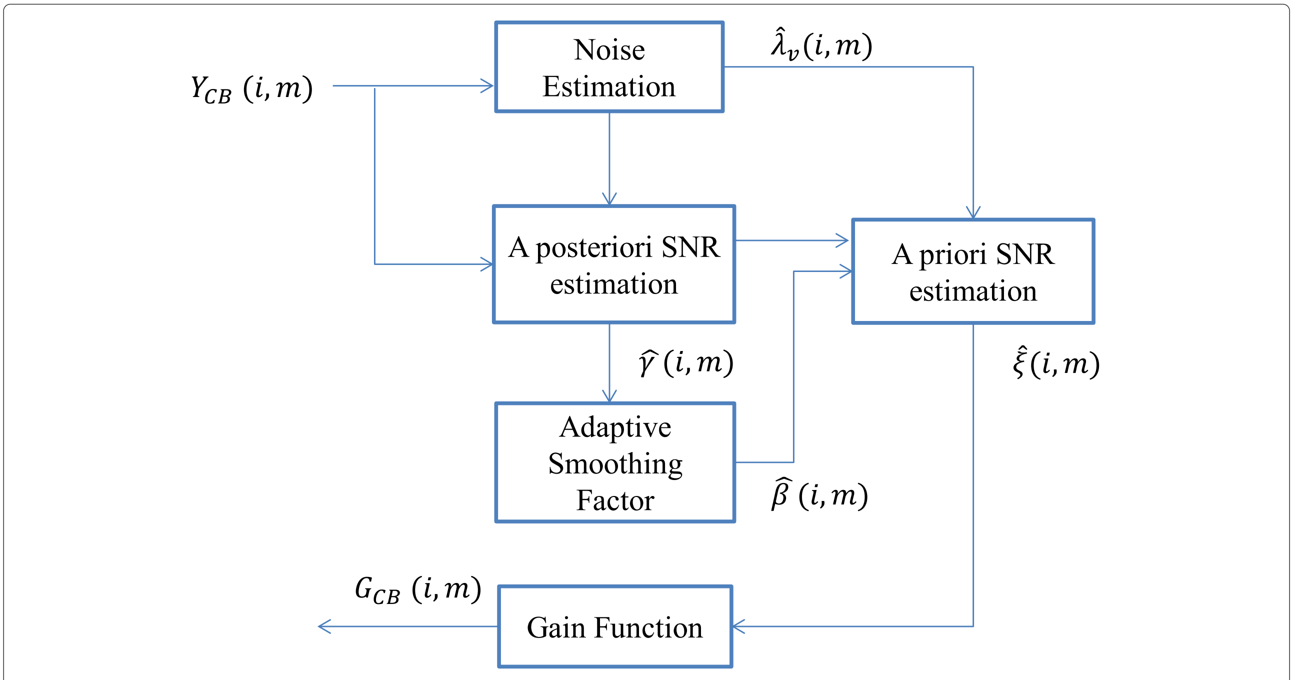

Fig. 2 Block diagram of the spectral gain function computation using the proposed a priori SNR estimation method

where $\tilde{\gamma}(i,m)$ is the a posteriori SNR estimate employing the smoothed estimate of the noisy speech from Eq. (17). Figure 2 describes the computation of the gain function by using the proposed method with an adaptive weighting function. In the following, we investigate the effect of two parameters σ and ρ on the proposed adaptive weighting function $\hat{\beta}(i,m)$.

To retain a similar property as a constant weighting factor β for speech-only and noise-only frames, we impose constraints on $\hat{\beta}(i,m)$ as:

$$\hat{\beta}(i,m) = \begin{cases} \beta, & \text{for noise-only frames or when } \tilde{\gamma}(i,m) = 1 \\ 1-\beta, & \text{for speech-only frames or when } \tilde{\gamma}(i,m) = \gamma_u,\ \gamma_u >> 1. \end{cases} \tag{20}$$

which lead to

$$\begin{cases} \frac{\beta_0}{1+\exp(-\sigma(1-\rho))} &= \beta \\ \frac{\beta_0}{1+\exp(-\sigma(\gamma_u-\rho))} &= 1-\beta \end{cases} \tag{21}$$

or

$$\begin{cases} \sigma\,(1-\rho) &= -\ln\left(\frac{\beta_0}{\beta}-1\right) \\ \sigma\,(\gamma_u-\rho) &= -\ln\left(\frac{\beta_0}{1-\beta}-1\right). \end{cases} \tag{22}$$

We now calculate the parameters σ and ρ directly for different levels of γ_u. From Eq. (22), we have

$$\frac{1-\rho}{\gamma_u-\rho} = \frac{\ln\left(\frac{\beta_0}{\beta}-1\right)}{\ln\left(\frac{\beta_0}{1-\beta}-1\right)}. \tag{23}$$

As such, the parameter ρ can be obtained from γ_u as

$$\rho = \frac{1-\gamma_u\frac{\ln\left(\frac{\beta_0}{\beta}-1\right)}{\ln\left(\frac{\beta_0}{1-\beta}-1\right)}}{1-\frac{\ln\left(\frac{\beta_0}{\beta}-1\right)}{\ln\left(\frac{\beta_0}{1-\beta}-1\right)}}. \tag{24}$$

The parameter σ can be calculated as

$$\sigma = \frac{-\ln\left(\frac{\beta_0}{\beta}-1\right)}{1-\rho}. \tag{25}$$

Figure 3 shows the pdf of a posteriori SNR for different noise types for β =0.98 and β_0=0.983, mapped with different adaptive smoothing factors calculated at several posteriori SNR values, γ_u: (i) at γ_u=5 dB SNR with $\sigma = -4.469$, $\rho = 2.295$; (ii) at $\gamma_u = 7$ dB SNR with $\sigma = -2.408$, $\rho = 3.402$; (iii) at $\gamma_u = 9$ dB SNR with $\sigma = -1.391$, $\rho = 5.159$; and (iv) at $\gamma_u = 15$ dB SNR with $\sigma = -0.315$, $\rho = 19.344$. Adaptive smoothing factors with different parameters (slopes and means) can control the trade-off between the musical noise and the ability to preserve speech components. In pink noise case, the SNR estimate in noise-only case is distributed approximately between 0 and 1. According to Eq. (20), the adaptive smoothing factor is approximately β during this period to reduce the SNR variance. This can be noted from the figure (first plot on the left), where the adaptive smoothing factor is almost 0.983, which explains the ability of the proposed method to maintain the advantage of the conventional decision-directed method in reducing musical noise at low SNRs. Moreover, in the factory noise case where the SNR estimate is distributed between 0 and 2 during noise-only periods, the proposed smoothing factors designed at $\gamma_u = 9$ dB and $\gamma_u = 15$ dB reached the imposed constraint (0.983) during the noise variance, whereas adaptive factors designed at $\gamma_u = 5$ dB and $\gamma_u = 7$ dB are lower than 0.983 during noise periods, which leads to an increase in musical noise.

For the babble noise case, the figure on the left shows the pdf of a posteriori SNR estimate during a noise-only period. It can be observed that the pdf has a large spread because of the non-stationary character of the babble noise, which means that an adaptive smoothing factor designed at higher a posteriori SNR γ_u is required to reduce the SNR variance during noise-only frames and reducing the effect of musical noise. From the figure, it can be clearly noted that adaptive smoothing factor designed at $\gamma_u = 15$ dB is the best among the designed factors since it attained a higher value over the a posteriori SNR distribution during the noise-only frames.

In addition, it can be noted that the weighting factor is inversely proportional to the a posteriori SNR γ. Thus, during the noise frames, γ takes small values. Consequently, the resulting weighting factor $\hat{\beta}(i,m)$ is close to 1, which means that the proposed method will have identical behavior as the DD and the MDD methods. This explains the ability of the proposed method to maintain the advantage of the DD method in reducing musical noise in the low SNRs. Since the second term is zero, the a priori SNR estimate in noise frames will be given by

$$\hat{\xi}^{\downarrow}_{\text{prop}}(i,m) = \hat{\beta}(i,m)G^2_{\text{CB}}(i,m-1)\hat{\gamma}(i,m). \tag{26}$$

During speech activity frames, the resulting weighting factor takes values close to 0. In that scenario, the first term of Eq. (19) is almost negligible, and the a priori SNR estimate in speech activity frames will correspond to a smoothed version of the maximum likelihood estimate as given by

$$\hat{\xi}^{\uparrow\uparrow}_{\text{prop}}(i,m) = (1-\hat{\beta}(i,m))P[\,\tilde{\gamma}(i,m)-1\,]\,. \tag{27}$$

During a speech transition, the weighting factor decreases with each increment of the instantaneous SNR. As a consequence, the a priori SNR estimation corresponds to a combination of the first and second terms in Eq. (19) as given by

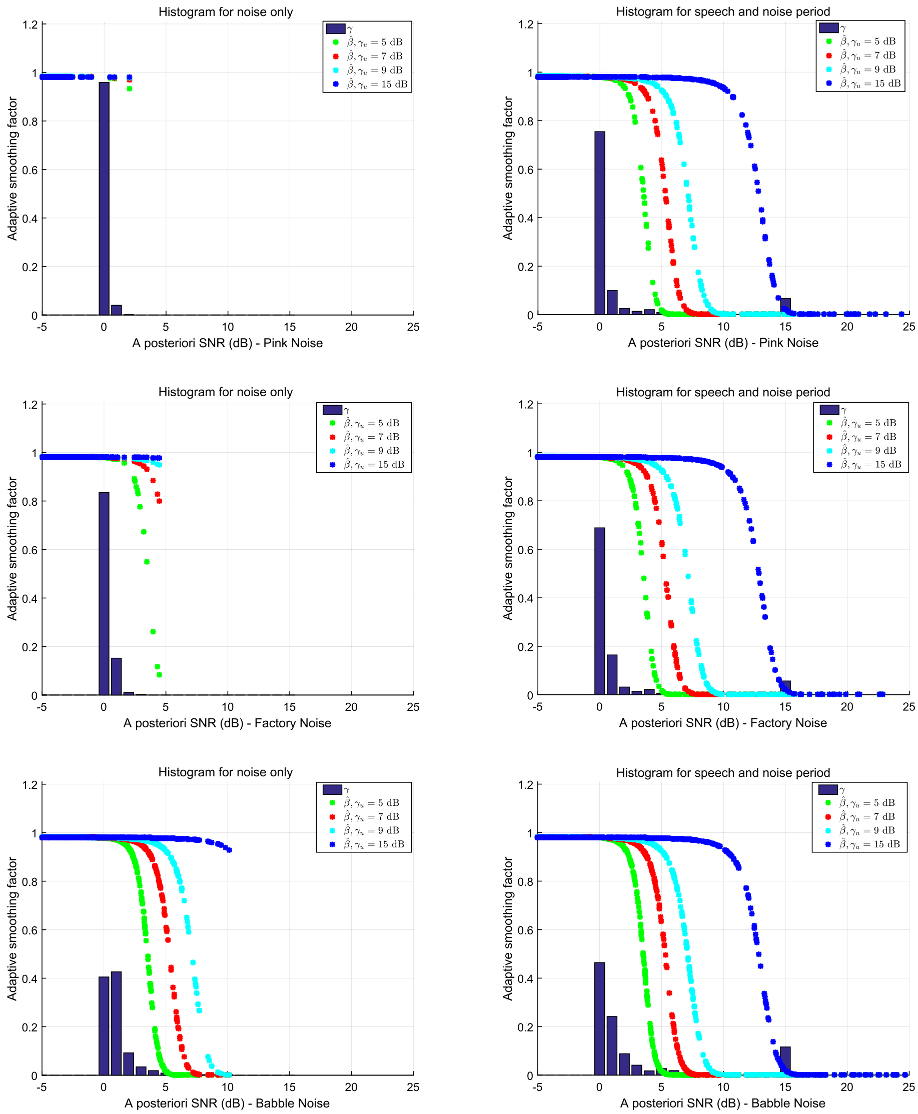

Fig. 3 Histograms of a posteriori SNR estimate for different background noises (1st row) for pink noise, (2nd row) for factory noise, and (last row) for babble noise at the 9th critical band mapped with adaptive smoothing factor calculated with different sets of parameters (adaptive smoothing factor calculated at (i) γ_u=5 dB, (ii) γ_u=7 dB, (iii) γ_u=9 dB, and (iv) γ_u=15 dB). Left figures for noise-only periods and right figures for speech-and-noise periods

$$\hat{\xi}^{\uparrow}_{\text{prop}}(i,m) = \hat{\beta}(i,m)G^2_{\text{CB}}(i,m-1)\hat{\gamma}(i,m) + (1-\hat{\beta}(i,m))P\big[\tilde{\gamma}(i,m)-1\big]. \qquad (28)$$

From (19), it can be noticed that the second term will have a varying impact on the a priori SNR updating process depending on the instantaneous SNR estimate. It is here the proposed method makes a difference in tracking any abrupt SNR changes. The apparent result is that more speech components are preserved as well as a reduction in the speech transient distortion.

5 Evaluation methodology

Speech quality evaluation can be classified into two categories: objective measurement and subjective measurement [3]. The first category is based on a mathematical comparison between the original and the enhanced speech signals. Many objective measurements have been proposed in the literature, such as the perceptual evaluation of speech quality measure (PESQ) [33, 34], segmental SNR measure SNR_{seg} [35, 36], and kurtosis ratio measure (KurtR) [37]. In addition, we propose a new evaluation method based on the Hamming distance as a speech preservation measure. The Hamming distance is a measure that takes into account speech presence or not for each time-frequency point. By measuring the difference between a clean speech binary mask and a processed speech binary mask, the measure takes into account the presence of speech in each time-frequency bin without amplitude weighting.

The perceptual evaluation of speech quality measure (PESQ) is the speech quality assessment recommended by ITU-T P.862 for its ability to predict the speech quality with a high correlation versus subjective listening tests [38]. PESQ implementation consists of first, estimating the bark spectrum of the input and the degraded signals by using a perceptual model in order to compute the loudness spectra and then compare between them to predict the perceived quality of the degraded signal. This objective means of quality assessment is expressed in terms of the mean opinion scores (MOS), measured from 1 to 5, where higher scores indicate higher quality. Here, we are using the implementation provided by Loizou [3].

Time domain-based segmental SNR is one of the widely used objective measures to evaluate the performance of speech enhancement algorithms, which is formed by averaging the frame level of SNR estimate [36] as given by

$$\mathrm{SNR_{seg}} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\|\mathbf{s}(m)\|^2}{\left\|\mathbf{s}(m)-\hat{\mathbf{s}}(m)\right\|^2} \tag{29}$$

where M denotes the number of frames, while $\hat{\mathbf{s}}(m)$ and $\mathbf{s}(m)$ are the estimated and original speech vectors, respectively, in time domain. The segmental SNR values are limited in the range of $[-10, 35]$ dB in order to exclude frames with no speech.

In addition, to further investigate the performance of the a priori SNR estimation methods, we utilized the segmental speech preservation $\mathrm{SNR_{seg,sp}}$ and segmental noise reduction $\mathrm{SNR_{seg,noise}}$ as in [39]. These two measures give indications whether the improvement in $\mathrm{SNR_{seg}}$ is due to more noise reduction or more speech preservation and they can be defined as follows:

$$\mathrm{SNR_{seg,sp}} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\|\mathbf{s}(m)\|^2}{\left\|\mathbf{s}(m)-\tilde{\mathbf{s}}(m)\right\|^2} \tag{30}$$

$$\mathrm{SNR_{seg,noise}} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\|\mathbf{v}(m)\|^2}{\|\tilde{\mathbf{v}}(m)\|^2} \tag{31}$$

where $\tilde{\mathbf{s}}(m)$ and $\tilde{\mathbf{v}}(m)$ denote the mth frame of the filtered clean speech and noise signals with the same gain function used to enhance the noisy signal.

The Kurtosis ratio measure is a mathematical measure used to calculate the musical noise. Such measure defines by the estimated speech signal and the noisy speech signal during noise frames only [37]. In order to detect the speech silence and presence, a VAD decision was employed [40], given two hypotheses $\mathcal{H}_0(k,m)$ and $\mathcal{H}_1(k,m)$ indicate the speech absence and presence, respectively. VAD decision is given by

$$D(k,m) = \begin{cases} 1 \text{ if } & \mathcal{H}_1(k,m) \\ 0 \text{ if } & \mathcal{H}_0(k,m) \end{cases} \tag{32}$$

and $V(k,m) = 1 - D(k,m)$ denotes the activity detection of the noise periods. In order to avoid the miss-detection of speech components, the reference VAD were generated with 50 dB global SNR. Kurtosis ratio can be defined by

$$\mathrm{KurtR} = E\left\{\frac{\kappa_{\hat{s}}(k)}{\kappa_y(k)}\right\} \tag{33}$$

where $\kappa_{\hat{s}}(k)$ and $\kappa_y(k)$ indicate the kurtosis of the enhanced signal and the noisy signal at the kth frequency bin, respectively. They are defined as follows:

$$\kappa_{\hat{s}}(k) = \frac{\sum_{m=1}^{M}\left|\hat{S}_s(k,m)V(k,m)\right|^4}{\left\{\sum_{m=1}^{M}\left|\hat{S}_s(k,m)V(k,m)\right|^2\right\}^2} - 2 \tag{34}$$

and

$$\kappa_y(k) = \frac{\sum_{m=1}^{M}|Y(k,m)V(k,m)|^4}{\left\{\sum_{m=1}^{M}|Y(k,m)V(k,m)|^2\right\}^2} - 2. \tag{35}$$

We proposed an evaluation method to measure the capability of the speech enhancement technique for preserving more weak speech components, referred to as the modified Hamming distance. It is determined by the difference of the time-frequency points detected using VAD decision [41] applied on the clean speech signal and the estimated speech signal. The detection of the VAD decisions for the noisy speech signal and estimated speech signal was performed only based on full-band VAD decisions for clean speech frames. The rationale for developing this new measure is that the result is amplitude invariant, which is important when measuring speech components. Those speech components otherwise would be overshadowed by strong amplitude components. The modified Hamming distance measure is calculated as

$$\mathrm{HD} = \frac{2}{KM}\sum_{m=1}^{M}\sum_{k=1}^{K/2}\left(\hat{D}(k,m)\oplus D(k,m)\right). \tag{36}$$

where $\oplus$ performs a logical exclusive OR operation that returns output containing elements set to either logical 1 (true) or logical 0 (false). Here, $D(k,m)$ denotes the voice activity detection of the clean signal and $\hat{D}(k,m)$ denotes the VAD of the estimated speech signal conditioned on clean speech detected, which is computed initially by testing each sub-band independently for speech activity using the decision device and then analyzed by further logic to reduce false alarms. A lower HD score indicates more speech components are preserved.

The second category of evaluations is based on subjective listening tests, which are considered more accurate and reliable [42]. For the subjective listening test, 10 subjects (5 males and 5 females) were recruited to compare and rate the estimated speech signals, the noisy signals, and the clean speech signals under different SNR conditions. Three different utterances were concatenated to be used for this test. They were corrupted with different sources of noise at 10 dB SNR. In this paper, we used pink noise source which is a shaped and filtered version of the white noise, and babble noise source that represents a group of people speaking in a canteen. The listening test was performed in a quiet office room using a DT-880 Beyerdynamic headphones. A laptop was connected through the USB interface to the headphones via a Topping VX-1 amplifier to provide good quality audio and consistent sound level. The sound clips were embedded in a PowerPoint document, which was also used for recording the results. The listeners were required to listen to the sentences enhanced by the different methods (DD, MDD, and the proposed method) and rated them on a scale from 1 to 5 by steps of 1. This rating takes into account three criteria: speech quality, background noise, and the musical noise levels [3] and [7]. The ranking instruction can be found in Table 1, which describes the scale of the criteria used in the listening test. This methodology helps to reduce the listeners uncertainty in rating which speech enhancement method is better in terms of the aforementioned criteria and referce to the clean speech signals and the noisy signals.

Table 1 Scale description of the listening test criteria [43]

Rating	Description
Speech	
5	Not degraded
4	Little degraded
3	Somewhat degraded
2	Fairly degraded
1	Very degraded
Background noise	
5	Not noticeable
4	Somewhat noticeable
3	Noticeable but not intrusive
2	Fairly conspicuous, somewhat intrusive
1	Very conspicuous, very intrusive
Musical noise	
5	Not noticeable
4	Somewhat noticeable
3	Noticeable but not intrusive
2	Fairly conspicuous, somewhat intrusive
1	Very conspicuous, very intrusive

6 Experimental results and discussion

6.1 Experimental setup

In this section, extensive experiments were conducted to evaluate the performance of the proposed approach in different scenarios. First, the performance of the proposed method was compared to the performances of the DD approach [31], the MDD approach [7], and the TSNR approach [21]. Second, we demonstrated the robustness of the proposed a priori SNR estimator by employing different gain functions. The speech sequences and noise were extracted from the NOISEUS and NOISEX databases, respectively [3]. In this work, 30 speech sentences were used (15 male speakers and 15 female speakers). Four different background noise types employed, which include pink noise, F16 Cockpit noise, factory noise, and babble noise. The noisy signal was obtained by combining the speech sequences with background noise at input SNRs of 0, 5, and 10 dB. All the sequences had been re-sampled to $f_s = 8000$ Hz. The values of the a priori SNR estimations had a floor ($\xi_0 = -25$ dB). In order to limit the noise reduction, we employed a noise residual floor $\epsilon = -20$ dB.

For the TSNR approach, by using the DD approach to estimate the a priori SNR in the first step, we computed the gain function by using WF as in Eq. (6). In the second step, we used different gain functions to enhance the noisy speech signal. A minimum mean-square error (MMSE) noise power estimator based on the speech presence probability [44] was employed to estimate the noise PSD for all the a priori SNR estimators, starting with a noise only period of 1 s. The value of the smoothing constant in Eq. (17) was chosen as $\alpha_y = 0.3$. Three different STFT analysis window cases had been considered for the evaluation of the proposed a priori SNR estimation method as shown in Table 2.

6.2 Case 1

Consider STFT analysis window with length $K = 256$ (32 ms) and a frame rate of $R = 128$ (50% overlap) with square-root Hanning window [7].

Table 2 Smoothing parameters of the a priori SNR estimation for different STFT analysis specifications

Case	Window length K	Window overlap	β	β_0
Case 1	256	50%	0.960	0.963
Case 2	256	75%	0.980	0.983
Case 3	512	50%	0.922	0.925

Based on these values, the frequency bins of the noisy spectrum were then grouped into $I = 17$ critical bands as shown in Eq. (4). The fixed weighting constants for DD and MDD approaches were chosen as $\beta = 0.96$ as shown in Table 2. The time averaging constant t_s is calculated as 0.391 s, by using Eq. (15) and STFT analysis parameters. As discussed in Section 4, the level γ_u in Eq. (20) for the adaptive smoothing factor is chosen according to the noise characteristics. As such, for pink noise, white noise, and factory noise, an adaptive smoothing factor is obtained with $\gamma_u = 9$ dB, resulting in $\sigma = -1.28$ and $\rho = 5.496$. For highly varying background noise such as babble noise, the adaptive smoothing factor is obtained with $\gamma_u = 15$ dB resulting in $\sigma = -0.290$ and $\rho = 20.831$ to keep the weighting factor close to 1 during noise frames, which helps to increase the robustness of the a priori SNR estimation against the SNR fluctuations.

6.2.1 Evaluation of a priori SNR estimation

Figure 4 demonstrates the behavior of the DD, the MDD, the TSNR, and the proposed a priori SNR estimators at 10 dB input SNR and under pink, factory, and babble background noise conditions, respectively. Speech enhancement is performed by using a WF [25] as shown in the subfigures on the left side and MMSE- LSA [17] as shown in the subfigures on the right side. It is clearly visible that during noise-only periods where the a posteriori SNR is sufficiently low, the DD, the MDD, and the TSNR methods represent a smoothed version of the a posteriori SNR. The proposed method has identical behavior as DD and MDD since $\hat{\beta}$ is very close to 1, which is aligned with Eq. (20). This explains the ability of the proposed method to eliminate the musical noise. During the speech onset, the proposed a priori SNR estimation with different gain functions responds more quickly to abrupt changes in the a posteriori SNR when compared to the other a priori SNR estimators. In terms of tracking ability, in pink and factory noise cases, it can be observed that the DD, the MDD, and the TSNR a priori SNR estimations follow the a posteriori SNR with a delay in the speech onset frames, which results in a speech distortion, whereas the proposed a priori SNR estimation reduces the delay and preserves speech components. For babble noise case, the MDD, the TSNR, and the proposed methods achieve slightly higher adapting speed than the DD approach during speech transitions.

Furthermore, we calculate mean square error (MSE) between the smoothed version of the true a priori SNR and the aforementioned estimation methods as depicted in Table 3. The results show the average MSE over the total number of frames and frequency bins. It can be clearly seen that the proposed estimator has the lowest estimation error compared to the DD, the MDD, and the TSNR estimators. Hence, the proposed a priori SNR estimator results in less estimation errors.

6.2.2 Objective results

The performance of the proposed a priori SNR estimation method is evaluated and compared to the performance of different a priori SNR estimators such as the DD, the MDD, and the TSNR methods for different noise types and under various SNR conditions. The clean speech is corrupted by pink, F16 cockpit, factory, and babble noise at 0, 5, and 10 dB input SNRs.

Tables 4, 5, 6, and 7 show the mean objective results for the stationary background noise case (pink), and non-stationary background noise cases (F16 cockpit, factory, and babble), respectively, with the DD, the MDD, the TSNR, and the proposed a priori SNR estimation methods combined with WF or MMSE-LSA gain functions.

The improvement in terms of speech quality of the proposed method is affirmed by the perceptual evaluation of speech quality PESQ measures. The proposed a priori SNR estimator outperforms the aforemention estimators in terms of speech quality, indicated by higher PESQ measures as depicted in Tables 4, 5, and 6. In the babble noise case, PESQ measures reveal that the proposed estimator achieves significantly better results than the TSNR method and approximately the same speech quality improvement as the MDD approach when combined with MMSE-LSA gain function. However, in WF gain function case, all a priori SNR estimators achieve approximately the same results.

The Kurtosis ratio results show the ability of the proposed method to maintain the advantage of DD and MDD methods in reducing the musical noise. Under different types of noise and SNR conditions, the proposed a priori SNR estimation method delivers lower Kurtosis ratio scores than the conventional DD approach. Although the TSNR method combined with the WF gain function achieves lower Kurtosis ratio compared to other estimators, it prones to generate more musical noise when combined with the MMSE-LSA gain function. This indicates that the performance of the TSNR method is not steady with different gain functions.

Beside PESQ and Kurtosis ratio measures, we also evaluate the speech preservation performance of the different a priori SNR estimation methods. For WF gain function case, the HD measure results indicate that the proposed method has slightly lower HD scores than the compared methods. For MMSE-LSA gain function, although results show that the proposed method achieves lower HD scores than the MMD and TSNR methods, it has slightly higher scores than the DD approach especially for babble noise case.

Moreover, the results show that the proposed method achieves better noise reduction as it outperforms the conventional DD, MDD, and TSNR methods in terms of

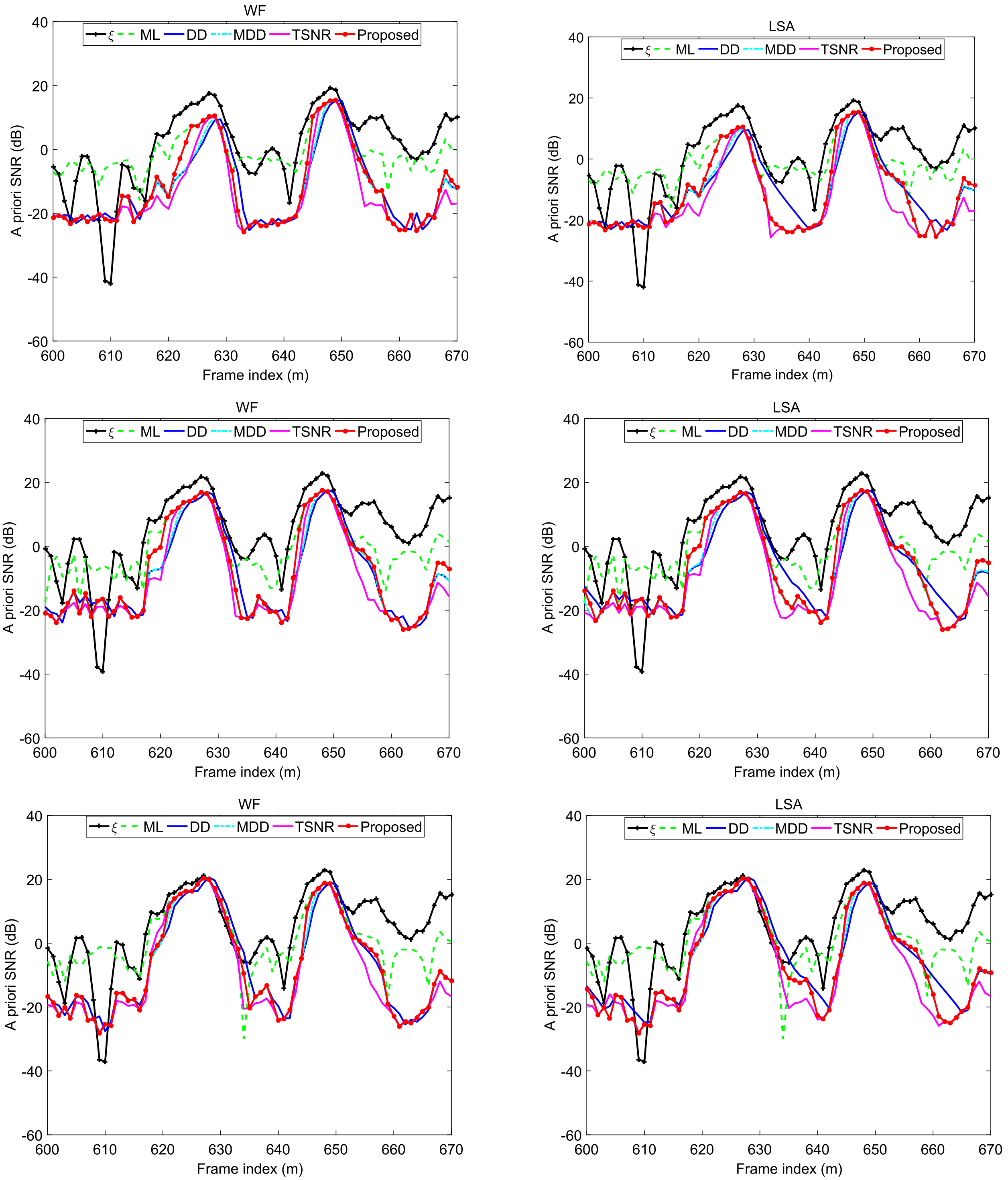

Fig. 4 Comparison of the a priori SNR estimation over a short-time period between the true a priori SNR ξ (black solid line with a marker), ML a priori SNR estimate (green dashed line), $\hat{\xi}_{\mathrm{DD}}$ (blue solid line), $\hat{\xi}_{\mathrm{MDD}}$ (cyan dot solid line), $\hat{\xi}_{\mathrm{TSNR}}$ (magenta solid line) and $\hat{\xi}_{\mathrm{prop}}$ (red solid line with a marker), at 9th critical band and 10 dB SNR under different background noise: 1st row for pink noise, 2nd row for factory noise, and last row for babble noise

Table 3 Mean square error comparison of different a priori SNR estimation methods for different noise types

Noise type	DD	MDD	TSNR	Proposed
Pink	32.18	30.48	32.09	**30.16**
Factory	48.53	47.36	47.50	**47.20**
Babble	41.19	40.46	40.93	**40.33**

Bold values denote the best performance

segmental SNR for all noise types and SNR conditions. In addition to the quality measure of the speech components, we further evaluate the ability of the a priori SNR estimation methods in maintaining soft speech components and noise reduction by utilizing the segmental speech SNR and the segmental noise SNR. Figures 5 and 6 show the $\mathrm{SNR}_{\mathrm{seg,sp}}$ and $\mathrm{SNR}_{\mathrm{seg,noise}}$ scores for babble and factory noise and under varying input SNRs. It can be clearly seen that the proposed a priori SNR estimation method in general has better speech preservation capability as indicated by the higher scores of $\mathrm{SNR}_{\mathrm{seg,sp}}$ compared to the DD approach especially at high SNR (>10 dB), due to its improving tracking speed to abrupt changes in the speech onset. In terms of noise reduction performance, the results reveal that all the a priori SNR estimators achieve approximately same scores in babble noise case.

6.2.3 *Spectrograms*

Figures 7 and 8 highlight the ability of the proposed a priori SNR estimator in preserving more speech components than the decision-directed (DD) and modified decision-directed (MDD) a priori SNR estimators for different noise types. The clean speech signal is corrupted by either pink noise or factory noise with a 10 dB SNR condition. It can be observed that the proposed a priori SNR estimator preserves more speech components than DD and MDD a priori SNR estimators.

6.3 Case 2

Consider the case of STFT analysis window with length $K = 256$ (32 ms) and a frame rate $R = 64$ (75% overlap) together with Hamming window.

Based on [31], the fixed weighting constants for DD and MDD approaches used as $\beta = 0.98$ as shown in Table 2, which corresponds to a time averaging constant t_s =0.396 s. Accordingly, the adaptive smoothing factor for factory noise is obtained with $\gamma_u = 9$ dB, resulting in $\sigma = -1.391$ and $\rho = 5.159$.

Table 8 shows the mean objective results for factory noise at different input SNRs. According to the speech quality results, the proposed method has better performance in terms of higher PESQ scores for evaluated gain functions. In addition, it can be clearly observed that the proposed method results in better noise reduction compared to the other a priori SNR estimators, indicated by higher segmental SNR scores. In terms of speech preservation, results of HD measure reveal that the proposed method has slightly better scores than the other a priori SNR estimation methods especially when combined with WF gain function.

On the other hand, results demonstrate that the proposed method always maintain the advantage of the DD approach in reducing the musical noise generation. For low SNR (< 5 dB) with WF gain function, the proposed method has slightly higher KurtR scores than the DD approach due to its sensitivity towards noise variance. In the MMSE-LSA gain function case, the TSNR approach has the highest KurtR scores for varying input SNRs when compared to the DD, the MDD, and the proposed a priori SNR estimators.

6.4 Case 3

Consider STFT analysis window with length $K = 512$ (64 ms) and a frame rate of $R = 256$ (50% overlap) with square-root Hanning window [7].

In this case, the fixed weighting constant for DD and MDD approaches is chosen as $\beta = 0.922$ as shown in Table 2. Accordingly, the adaptive smoothing factor for pink noise is obtained with $\gamma_u = 9$ dB, resulting in $\sigma = -1.168$ and $\rho = 5.902$. For babble noise with $\gamma_u = 15$ dB , the parameters of the smoothing factor are as follows: $\sigma = -0.264$ and $\rho = 22.620$.

Table 4 Mean objective results for pink noise

Gain	SNR	PESQ				$\mathrm{SNR}_{\mathrm{seg}}$				HD				KurtR			
		DD	MDD	TSNR	Prop	DD	MDD	TSNR	Prop	DD	MDD	TSNR	Prop	DD	MDD	TSNR	Prop
WF	0	1.887	1.874	1.879	**1.965**	0.253	0.257	0.057	**0.662**	0.405	0.417	0.407	**0.403**	1.030	1.014	**1.013**	1.026
	5	2.355	2.344	2.327	**2.399**	2.671	2.710	2.948	**3.379**	0.389	0.394	0.399	**0.381**	1.173	1.067	**1.057**	1.102
	10	2.735	2.725	2.696	**2.782**	5.616	5.683	6.037	**6.388**	0.364	0.372	0.375	**0.326**	1.554	1.231	**1.208**	1.339
LSA	0	1.978	1.969	1.768	**2.043**	0.184	0.364	−0.399	**0.387**	0.399	0.397	0.395	**0.390**	1.245	**1.052**	3.724	1.077
	5	2.405	2.437	2.266	**2.497**	2.608	2.899	2.532	**3.396**	**0.360**	0.381	0.369	0.361	1.795	**1.141**	3.799	1.200
	10	2.762	2.805	2.670	**2.869**	5.637	5.966	5.837	**6.720**	**0.320**	0.354	0.350	0.343	2.708	**1.367**	3.941	1.512

Bold values denote the best performance

Table 5 Mean objective results for F16 Cockpit noise

Gain	SNR	PESQ				SNR_{seg}				HD				KurtR			
		DD	MDD	TSNR	Prop	DD	MDD	TSNR	Prop	DD	MDD	TSNR	Prop	DD	MDD	TSNR	Prop
WF	0	1.893	1.883	1.888	**1.966**	−0.233	−0.181	−0.037	**0.363**	0.404	0.406	0.407	**0.400**	1.031	**1.010**	1.011	1.022
	5	2.338	2.330	2.318	**2.384**	2.509	2.581	2.746	**3.132**	0.389	0.390	0.392	**0.382**	1.200	1.066	**1.056**	1.102
	10	2.714	2.705	2.687	**2.755**	5.392	5.488	5.687	**6.112**	0.356	0.366	0.376	**0.341**	1.626	1.226	**1.209**	1.355
LSA	0	1.969	1.970	1.813	**2.038**	0.033	0.211	−0.381	**0.564**	**0.382**	0.397	0.387	0.384	1.303	**1.056**	3.559	1.083
	5	2.376	2.414	2.268	**2.496**	2.361	2.660	2.348	**3.132**	**0.351**	0.376	0.365	0.355	1.958	**1.154**	3.845	1.218
	10	2.732	2.798	2.654	**2.831**	5.223	5.576	5.462	**6.134**	**0.309**	0.346	0.342	0.338	2.839	**1.375**	3.996	1.545

Bold values denote the best performance

Table 6 Mean objective results for factory noise

Gain	SNR	PESQ				SNR_{seg}				HD				KurtR			
		DD	MDD	TSNR	Prop	DD	MDD	TSNR	Prop	DD	MDD	TSNR	Prop	DD	MDD	TSNR	Prop
WF	0	2.447	2.441	2.423	**2.489**	1.564	1.589	1.792	**2.178**	0.368	0.377	0.379	**0.356**	1.666	1.251	**1.211**	1.402
	5	2.795	2.791	2.772	**2.830**	4.444	4.499	4.843	**5.206**	0.331	0.344	0.345	**0.328**	2.429	1.654	**1.582**	1.970
	10	3.170	3.204	3.180	**3.220**	7.586	7.660	7.965	**8.368**	0.299	0.303	0.316	**0.283**	3.475	2.385	**2.203**	2.979
LSA	0	2.484	2.529	2.388	**2.565**	1.388	1.670	1.565	**2.117**	**0.320**	0.356	0.355	0.344	2.712	**1.499**	2.854	1.704
	5	2.803	2.880	2.754	**2.902**	4.312	4.645	4.665	**5.127**	**0.273**	0.319	0.323	0.299	3.416	**1.942**	3.337	2.299
	10	3.091	3.257	3.114	**3.260**	7.564	7.874	7.833	**8.374**	**0.227**	0.278	0.285	0.259	3.816	**2.624**	3.914	3.159

Bold values denote the best performance

Table 7 Mean objective results for babble noise

Gain	SNR	PESQ				SNR_{seg}				HD				KurtR			
		DD	MDD	TSNR	Prop	DD	MDD	TSNR	Prop	DD	MDD	TSNR	Prop	DD	MDD	TSNR	Prop
WF	0	1.934	1.931	1.922	**1.936**	−0.325	−0.305	−0.650	**−0.266**	**0.398**	0.401	0.402	0.400	1.257	1.181	**1.171**	1.192
	5	2.280	2.285	2.273	**2.285**	1.628	1.679	1.795	**1.887**	**0.374**	0.381	0.382	0.379	1.509	1.273	**1.232**	1.293
	10	2.598	2.605	2.593	**2.670**	4.350	4.459	4.664	**4.687**	**0.338**	0.350	0.351	0.347	2.056	1.576	**1.479**	1.612
LSA	0	1.966	1.979	1.842	**1.981**	−0.426	−0.273	−0.963	**−0.243**	**0.361**	0.383	0.382	0.380	1.620	**1.408**	2.205	1.423
	5	2.290	2.332	2.199	**2.335**	1.528	1.788	1.490	**1.870**	**0.326**	0.358	0.360	0.357	2.041	**1.539**	2.252	1.556
	10	2.611	2.673	2.561	**2.674**	4.288	4.636	4.473	**4.796**	**0.281**	0.323	0.322	0.321	2.504	**1.819**	2.575	1.849

Bold values denote the best performance

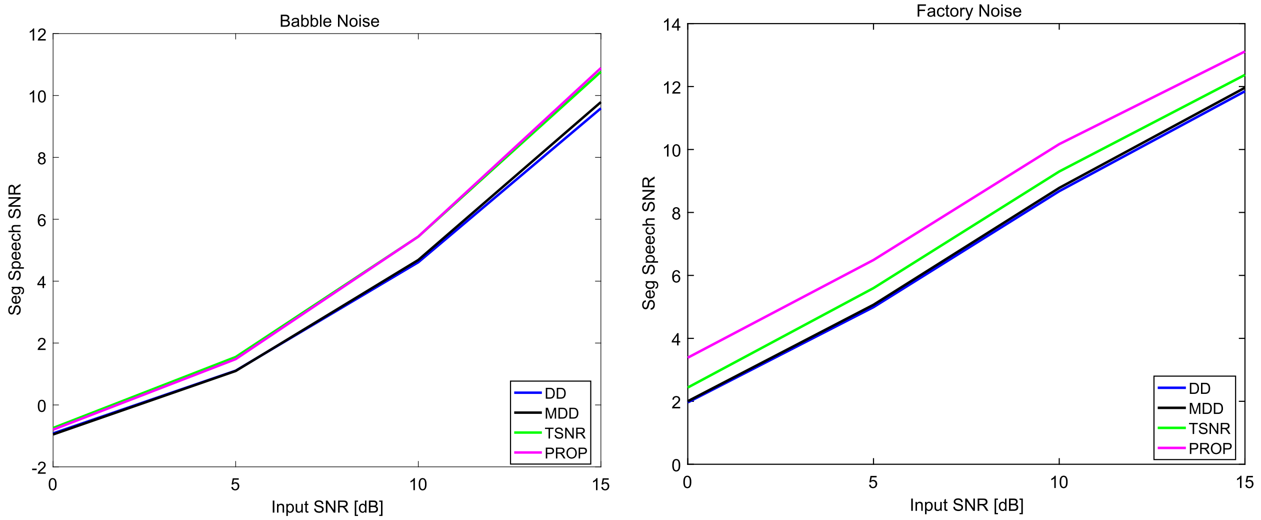

Fig. 5 Speech distortion measure for noisy speech corrupted with different noise types and under different SNR levels enhanced by Wiener filter speech estimation technique

6.4.1 *Objective results*

The performance of the proposed a priori SNR estimation method is evaluated and compared to the performance of the DD and the MDD methods for different noise types and under various SNR conditions. The clean speech is corrupted by pink and babble noise at 0, 5, and 10 dB input SNRs.

Tables 9 and 10 show the mean objective results for the stationary background noise case (pink) and non-stationary background noise case (babble), respectively, with the DD, the MDD, and the proposed a priori SNR estimation methods combined with WF or MMSE-LSA gain functions.

From the PESQ measures, it can be clearly noticed that the proposed a priori SNR estimator results in better speech quality than the conventional DD and the MDD approaches, indicated by higher PESQ measures. However, in babble noise case, PESQ measures reveal that the proposed estimator achieves approximately the same speech quality improvement as MDD approach and better than the conventional DD approach.

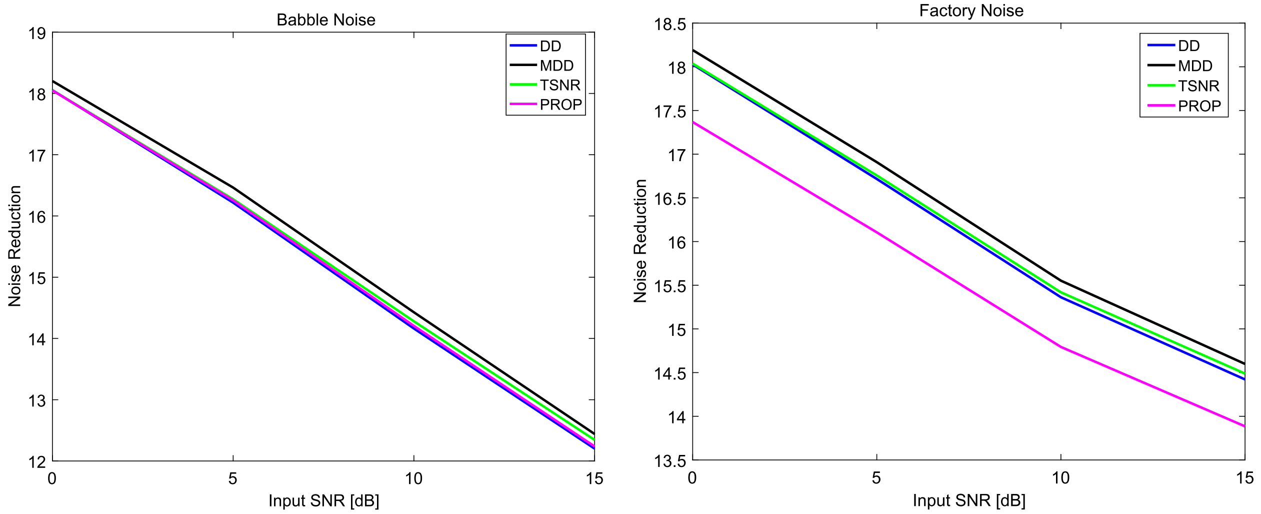

Fig. 6 Noise reduction measure for noisy speech corrupted with different noise types and under different SNR levels enhanced by Wiener filter speech estimation technique

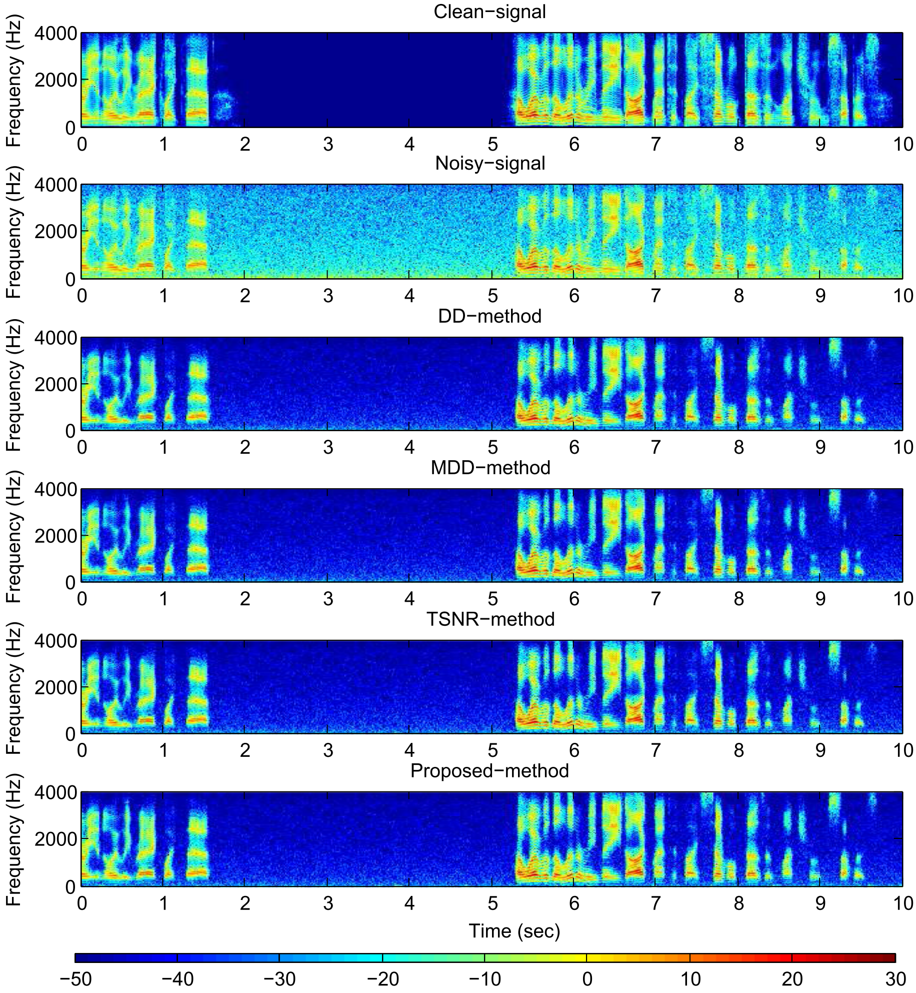

Fig. 7 Speech spectrograms for noisy speech corrupted with pink noise at 10 dB enhanced by Wiener filter speech estimation technique

In addition, the proposed method achieves better noise reduction as it outperforms the conventional DD and MDD approaches in terms of segmental SNR for all noise types and SNR conditions.

In terms of speech preservation performance, HD measure results indicate that the proposed method has slightly lower scores than the conventional DD and MDD approaches. In babble noise case, although it achieves better results than MDD approach, it has slightly higher HD measures than DD approach.

Moreover, Kurtosis ratio results show the ability of the proposed method to maintain the advantage of DD and MDD methods in reducing the musical noise under different types of noise and SNR conditions.

6.4.2 Evaluation of listening tests

Tables 11 and 12 demonstrate the average results of the listening test in terms of speech quality, background noise, musical noise, and the overall performance of each estimation method by determining the mean of the rating scores. Ten normal hearing participants in the age of (20–35) took part in this test. They were asked to rate speech signals estimated by three different a priori SNR estimators in terms of speech, background noise, and musical noise as explained in the previous section. The speech and background results show that the proposed method outperformed the DD and the MDD methods, which aligned with the objective results of PESQ and segmental SNR. Moreover, for the musical noise ratings, the proposed method combined with different gain functions and different background noise scored approximately the same as the MDD method, which is slightly better than the DD method. This means that the proposed method maintains the advantage of the DD approach in generating less musical noise which can be observed from the objective measurement Kurtosis ratio.

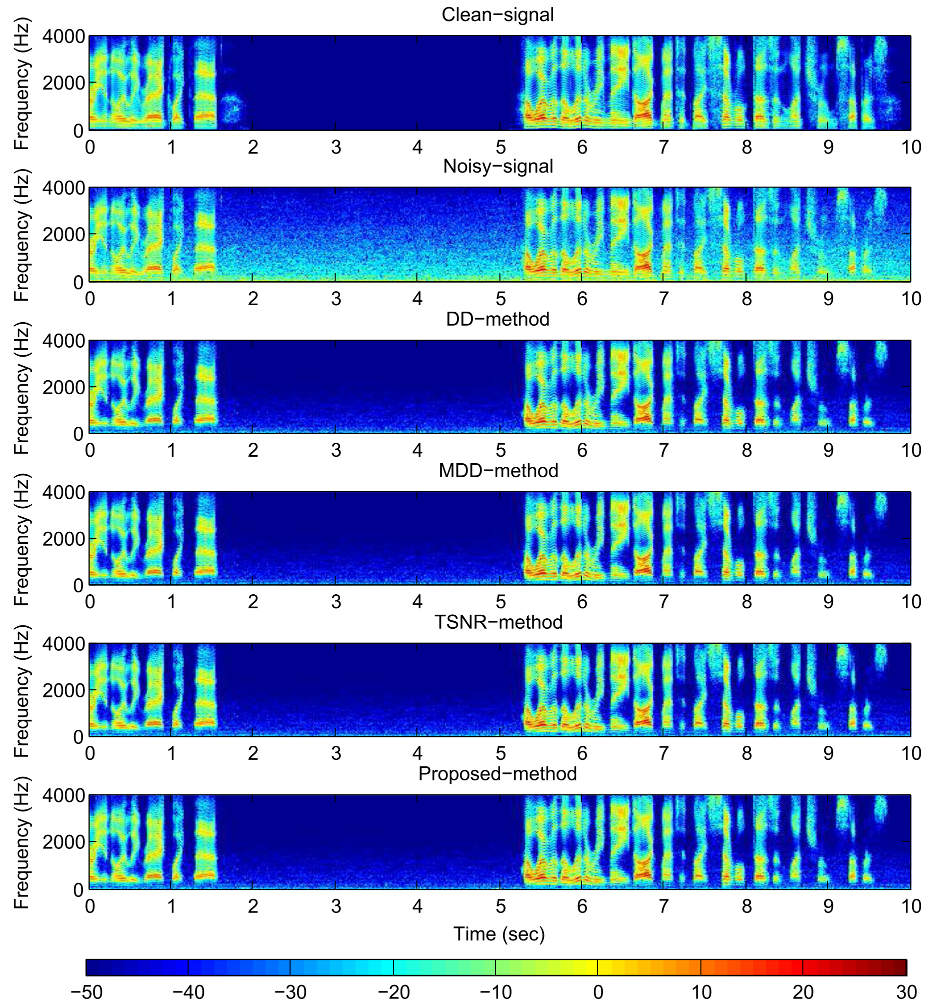

Fig. 8 Speech spectrograms for noisy speech corrupted with factory noise at 10 dB enhanced by WF speech estimation technique

Furthermore, the overall results of the 10 participants have been evaluated using a statistical analysis to assess the differences between the ratings obtained for each a priori SNR estimation method in terms of overall quality. For this purpose, we used analysis of variance (ANOVA) to indicate a significant difference between scores if the level of significance is smaller than 0.05. A significant difference between scores has been noted when the MMSE-LSA was combined with all the different a priori SNR estimation methods under different noise conditions as shown in Table 13. Moreover, a significant difference noted when the WF was employed with all the different a priori SNR estimation methods in pink noise case. The 90% confidence interval (CI) for the overall scores of the

Table 8 Mean objective results with 75% overlap for factory noise

Gain	SNR	PESQ				SNR$_{seg}$				HD				KurtR			
		DD	MDD	TSNR	Prop	DD	MDD	TSNR	Prop	DD	MDD	TSNR	Prop	DD	MDD	TSNR	Prop
WF	0	2.356	2.354	2.341	**2.398**	1.201	1.260	1.358	**1.652**	0.357	0.375	0.355	**0.348**	1.075	1.064	**1.0621**	1.081
	5	2.693	2.687	2.670	**2.738**	4.035	4.096	4.182	**4.504**	0.339	0.340	0.339	**0.330**	1.178	1.138	**1.137**	1.158
	10	3.099	3.091	3.066	**3.118**	7.089	7.159	7.250	**7.655**	0.314	0.317	0.326	**0.300**	1.483	1.340	**1.327**	1.435
LSA	0	2.392	2.406	2.238	**2.452**	1.169	1.374	1.067	**1.674**	0.345	0.352	0.362	**0.344**	1.153	**1.079**	1.383	1.112
	5	2.722	2.749	2.641	**2.799**	4.029	4.237	3.970	**4.556**	0.323	0.332	0.315	**0.320**	1.401	**1.165**	1.677	1.231
	10	3.067	3.144	3.023	**3.169**	7.142	7.335	7.149	**7.747**	**0.287**	0.305	0.297	0.296	1.917	**1.402**	1950	1.507

Bold values denote the best performance

Table 9 Mean objective results for pink noise

Gain	SNR	PESQ			SNR_{seg}			HD			KurtR		
		DD	MDD	Prop	DD	MDD	Prop	DD	MDD	Prop	DD	MDD	Prop
WF	0	1.960	1.944	**2.004**	−0.202	−0.138	**0.326**	0.441	0.451	**0.401**	1.064	**1.009**	1.028
	5	2.401	2.412	**2.462**	2.530	2.662	**3.148**	0.399	0.395	**0.358**	1.253	**1.070**	1.103
	10	2.748	2.776	**2.799**	5.461	5.606	**6.085**	0.348	0.363	**0.325**	1.660	**1.235**	1.436
LSA	0	2.018	2.054	**2.099**	−0.291	0.149	**0.505**	0.371	0.396	**0.328**	1.416	**1.073**	1.096
	5	2.388	2.497	**2.513**	2.334	2.992	**3.344**	0.326	0.368	**0.318**	1.857	**1.181**	1.242
	10	2.715	2.861	**2.870**	5.378	5.948	**6.346**	**0.271**	0.329	0.316	2.331	**1.406**	1.528

Bold values denote the best performance

Table 10 Mean objective results for babble noise

Gain	SNR	PESQ			SNR_{seg}			HD			KurtR		
		DD	MDD	Prop	DD	MDD	Prop	DD	MDD	Prop	DD	MDD	Prop
WF	0	1.971	1.978	**1.998**	−0.775	−0.693	**−0.604**	**0.392**	0.401	0.400	1.264	**1.148**	1.152
	5	2.289	2.316	**2.328**	1.537	1.665	**1.723**	**0.354**	0.396	0.369	1.531	**1.279**	1.287
	10	2.603	2.632	**2.694**	4.229	4.411	**4.523**	**0.303**	0.323	0.321	1.902	**1.534**	1.554
LSA	0	1.954	2.009	**2.011**	−1.007	−0.604	**−0.570**	**0.319**	0.362	0.360	1.481	**1.344**	1.347
	5	2.253	2.344	**2.399**	1.251	1.801	**1.864**	**0.271**	0.328	0.309	1.715	**1.450**	1.456
	10	2.572	2.669	**2.696**	4.032	4.658	**4.743**	**0.217**	0.282	0.208	1.936	**1.634**	1.648

Bold values denote the best performance

Table 11 Listening test results for pink noise at 10 dB input SNR

Gain	Categories	Pink noise		
		DD	MDD	Prop
WF	Speech	3.4	3.8	3.8
	Background noise	3.1	3.7	3.8
	Musical noise	4.5	4.8	4.9
	Over all	3.7	4.1	4.2
LSA	Speech	3.2	4.2	4.3
	Background noise	3.1	3.8	4.0
	Musical noise	4.1	4.6	4.5
	Over all	3.5	4.2	4.3

Table 12 Listening test results for babble noise at 10 dB input SNR

Gain	Categories	Babble noise		
		DD	MDD	Prop
WF	Speech	3.6	3.9	4.1
	Background noise	3.2	3.5	3.6
	Musical noise	4.0	4.2	4.4
	Over all	3.6	3.9	4.0
LSA	Speech	3.2	4.0	4.2
	Background noise	2.8	3.5	3.8
	Musical noise	3.9	4.4	4.5
	Over all	3.3	4.0	4.2

Table 13 Statistical analysis for subjective listening test under different noise conditions

Noise	Gain function	*p*-Value	DD		MDD		Proposed	
			Lower	Upper	Lower	Upper	Lower	Upper
Pink	WF	0.041	3.51	3.81	4.03	4.17	4.12	4.26
	LSA	0.001	3.24	3.68	4.14	4.27	4.27	4.33
Babble	WF	0.060	3.49	3.63	3.77	3.92	3.91	4.11
	LSA	0.001	3.11	3.45	3.86	4.06	4.08	4.25

subjective listening test provides further statistic analysis for performance comparison among different speech enhancement techniques and under varying noise types as depicted in Table 13. It is noted that the 90% CI of the proposed a priori SNR estimation method combined with MMSE-LSA gain function does not overlap with the other compared methods under varying noise conditions. Moreover, the 90% CI of the proposed method with different gain functions does not overlap with DD approach. This means that the proposed method statistically outperforms DD approach in different noise cases. Hence, the proposed method has better performance than the other methods in terms of overall quality.

6.5 Evaluation of the effect of the bark scale frequency warping on the noise characteristics

An experiment is conducted to prove the efficiency of the proposed bark scale-based frequency warping method in eliminating the musical noise. For this experiment, we have utilized the normal (Gaussian) distribution [45] and the Weibull distribution [46] is used to compare the noise distribution before and after the frequency warping. Figure 9 shows comparisons between the distribution of different types of noise before frequency warping at frequency 546.87 Hz and after the frequency warping at the 6th critical band. In the pink and factory noise cases, it can be clearly seen that the noise histograms fit well to a

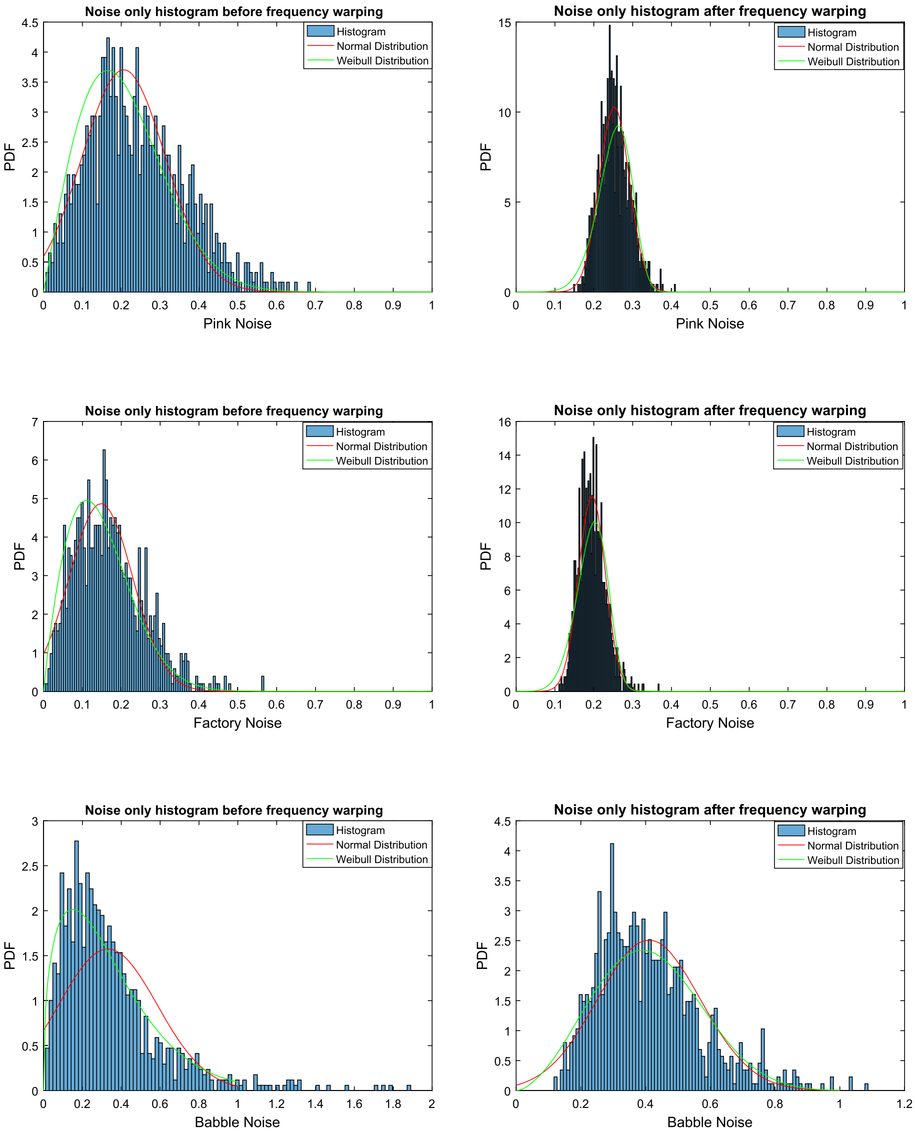

Fig. 9 Evaluation of bark scale-based frequency warping at the 6th critical band under different background noise: 1st row for pink noise, 2nd row for factory noise, and last row for babble noise

Table 14 Noise variance comparison before and after frequency warping and for different noise types

Noise type	Variance before frequency warping	Variance after frequency warping
Pink	0.1077	0.0387
Factory	0.0820	0.0343
Babble	0.2535	0.1590

Gaussian distribution, which is the common assumption in most noise estimation methods. This can help to reduce the musical noise by reducing the bias and provide a more precise estimate. Whereas in babble noise case, although a Gaussian distribution does not really fit the noise distribution after the frequency warping, it becomes more concentrated with a shorter tail compared to the noise distribution before the frequency warping.

In order to highlight the ability of the bark scale based frequency warping in reducing the effect of the musical noise, a variance comparison of the noise PDF before and after the frequency warping is presented for different noise types as shown in Table 14. It can be clearly observed that the bark scale based frequency warping has the ability to significantly reduce the noise variance. This helps to reduce the musical noise effect and make it unnoticeable.

6.6 Evaluate the benefit of the critical band processing

In order to demonstrate the benefit of using critical band processing as a preprocessor to the speech enhancement framework, we present a comparison of objective measurement before and after applying the critical band processing with the proposed a priori SNR estimation method combined with different gain functions. Table 15 presents the performance comparison in terms of KurtR and PESQ under different input SNRs. Objective results reveal that CB processing significantly outperforms STFT in reducing the musical noise with the lowest scores of KurtR under different input SNRs. This is due to its ability in reducing the noise variance. In terms of speech quality, PESQ scores indicate improved speech quality for the CB processing in low SNR, while in high SNR, it approximately achieved the same reults as STFT.

Table 15 Objective results comparison as a function of input SNR for babble noise

Gain function	Input SNR (dB)	KurtR		PESQ	
		CB	STFT	CB	STFT
WF	0	1.192	3.708	1.936	1.884
	5	1.293	3.811	2.285	2.297
	10	1.612	3.874	2.670	2.602
LSA	0	1.423	3.231	1.981	1.908
	5	1.556	3.326	2.335	2.308
	10	1.849	3.399	2.674	2.694

7 Conclusions and future work

In this paper, an adaptive a priori SNR estimator has been extended and evaluated for different speech enhancement gain functions. As a basis for the adaptation, the a priori SNR estimation employs a model of speech absence probability based on a sigmoid function. The sigmoid function can be tuned to provide a trade-off between the speech onset sensitivity and the annoying noise artifacts also known as musical noise. Moreover, we have developed an objective measurement to evaluate the capability of the speech enhancement technique in preserving soft speech components known as modified hamming distance. In combination with different gain functions including the WF and the MMSE-LSA, the objective results show that the proposed method outperforms the conventional DD and MDD approaches with higher scores in PESQ and SNR_{seg}. Furthermore, the proposed bark scale-based frequency warping helps to reduce the effect of the musical noise and make it unnoticeable because of the significant reduction in the noise variance, which helps in the noise estimation needed for the SNR estimation. The obtained objective evaluation results are supported by the averaged results from the subjective listening tests, as the proposed method was preferred by the listeners. Future work will include different choices of filter banks and also a possible low delay implementation.

Acknowledgements
Not applicable.

Funding
Not applicable.

Availability of data and materials
All the speech utterances used in the evaluation and the subjective tests can be found at the website: http://ecs.utdallas.edu/loizou/speech/noizeus/index.html. Noise data can be found at the website:http://spib.linse.ufsc.br/noise.html.

Authors' contributions
LN carried out the main conception and design, evaluation, and interpretation of the results. PY carried out the critical band concept and its implementation to the speech enhancement framework. SN and HHD gave academic guidance to this research work. All authors discussed the final results, read, and approved the final manuscript.

Ethics approval and consent to participate
This study obtained Human Research Ethics Office approval (HRE2017-0030) and informed consent was obtained from each participant before participation.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Department of Electrical Engineering, computing and Mathematical Sciences, Curtin University, Perth, Australia. [2]Nuheara Ltd., Perth, Australia.

Received: 27 July 2018 Accepted: 30 April 2019
Published online: 07 June 2019

References

1. Doclo S., Kellermann W., Makino S., Nordholm S. E. (2015) Multichannel signal enhancement algorithms for assisted listening devices: exploiting spatial diversity using multiple microphones. IEEE Sig Process Mag 32(2):18–30
2. Benesty J., Makino S., Chen J. (2005) Speech Enhancement. Springer, New York
3. Loizou P. C. (2013) Speech Enhancement: Theory and Practice. CRC press, Boca Raton
4. McAulay R., Malpass M. (1980) Speech enhancement using a soft-decision noise suppression filter. IEEE Trans Acoust. Speech. Sig. Process. 28(2):137–145
5. Cohen I., Berdugo B. (2001) Speech enhancement for non-stationary noise environments. Sig. Process. 81(11):2403–2418
6. Gustafsson H., Nordholm S., Claesson I. (2001) Spectral subtraction using reduced delay convolution and adaptive averaging. IEEE Trans. Speech. Audio Process. 9(8):799–807
7. Yong P. C., Nordholm S., Dam H. H. (2013) Optimization and evaluation of sigmoid function with a priori SNR estimate for real-time speech enhancement. Speech Commun. 55(2):358–376
8. Irino T., Patterson R. D. (2006) A dynamic compressive gammachirp auditory filterbank. IEEE Trans Acoust. Speech. Sig. Process. 14(6):2222–2232
9. Kortlang S., Ewert S. D., Gerkmann T. (2014) Single channel noise reduction based on an auditory filterbank. In: 14th International Workshop on Acoustic Signal Enhancement (IWAENC), Juan-les-Pins. pp 283–287
10. Tsoukalas D. E., Mourjopoulos J. N., Kokkinakis G. (1997) Speech enhancement based on audible noise suppression. IEEE Trans Acoust. Speech. Sig. Process. 5(6):497–514
11. Virag N. (1999) Single channel speech enhancement based on masking properties of the human auditory system. IEEE Trans Acoust. Speech. Sig. Process. 7(2):126–37
12. Laine U. K., Karjalainen M., Altosaar T. (1994) Warped linear prediction (WLP) in speech and audio processing. In: Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 3, Adelaide. pp III-349
13. Haque S., Togneri R. (2010) A psychoacoustic spectral subtraction method for noise suppression in automatic speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas. pp 1618–1621
14. Breithaupt C., Martin R. (2011) Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient conditions. IEEE Trans Acoust. Speech. Sig. Process. 19(2):277–289
15. Boll S., Suppression of acoustic noise in speech using spectral subtraction (1979). IEEE Trans Acoust. Speech. Sig. Process. 27(2):113–120
16. Höglund N., Nordholm S. (2009) Improved a priori SNR estimation with application in Log-MMSE speech estimation. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz. pp 189–192
17. Ephraim Y., Malah D. (1985) Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans Acoust. Speech. Sig. Process. 33(2):443–445
18. Cappe O. (1994) Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Trans. Speech. Audio. Process. 2(2):345–349
19. Breithaupt C., Gerkmann T., Martin R. (2008) A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing. In: IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas. pp 4897–4900
20. Suhadi S., Last C., Fingscheidt T. (2011) A data-driven approach to a priori SNR estimation. IEEE Trans. Audio. Speech. Lang. Process. 19(1):186–195
21. Plapous C., Marro C., Scalart P. (2006) Improved signal-to-noise ratio estimation for speech enhancement. IEEE Trans. Audio. Speech. Lang. Process. 14(6):2098–2108
22. Alam M. d., Jahangir, Chowdhury M. d., Fasiul Alam M. d. (2009) Comparative study of a priori signal to noise ratio (SNR) estimation approaches for speech enhancement. IU-J. Electr Electron. Eng. 9(1):809–817
23. Ou S., Zhao X., Gao Y. (2007) Speech enhancement employing modified a priori SNR estimation. In: IEEE Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD 2007) (Vol. 3), Qingdao. pp 827–831
24. Nahma L., Yong P. C., Dam H. H., Nordholm S. (2017) Convex combination framework for a priori SNR estimation in speech enhancement. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp 4975–4979
25. Lim J. S., Oppenheim A. V. (1979) Enhancement and bandwidth compression of noisy speech. Proc. of the IEEE 67(12):1586–1604
26. Deller J. R., Hansen J. H. L., Proakis J. G. (2000) Discete-Time Processing of Speech Signals. Wiley, NY
27. Zwicker E. (1961) Subdivision of the audible frequency range into critical bands (Frequenzgruppen). J. Acoust. Soc. Am. 33(2):248–248
28. Jepsen M.L., Ewert S.D., Dau T. (2008) A computational model of human auditory signal processing and perception. J. Acoust. Soc. Am. 124(1):422–38
29. Sekey A., Hanson B. A. (1984) Improved 1 bark bandwidth auditory filter. J. Acoust. Soc. Am. 75(6):1902–1904
30. Hermansky H. (1990) Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. 87(4):1738–1752
31. Ephraim Y., Malah D. (1984) Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans Acoust. Speech. Sig. Process. 32(6):1109–1121
32. Yong P. C., Nordholm S., Dam H. H., Low S. Y. (2011) On the optimization of sigmoid function for speech enhancement. In: Proc. 19th European Signal Processing Conference (EUSIPCO), Barcelona, Spain. pp 211–215
33. Voran S. (1999) Objective estimation of perceived speech quality. I. Development of the measuring normalizing block technique. IEEE Trans. Speech. Audio. Process. 7(4):371–382
34. Rix A. W., Beerends J. G., Hollier M. P., Hekstra A. P. (2001) Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), vol. 2. pp 749–752
35. Lotter T., Vary P. (2005) Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model. EURASIP J. Appl. Sig. Process. 7:1110–1126
36. Hansen J. H., Pellom B. L. (1998) An effective quality evaluation protocol for speech enhancement algorithms. pp 2819–2822
37. Uemura Y., Takahashi Y., Saruwatari H., Shikano K., Kondo K. (2008) Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics. In: Proc. International Workshop on Acoustic Echo and Noise Control (IWANEC), Seattle, USA
38. Beerends J. G., Hekstra A. P., Rix A. W., Hollier M. P. (2002) Perceptual evaluation of speech quality (PESQ). The New ITU standard for end-to-end speech quality assessment, part ii: psychoacoustic model. J. Audio Eng. Soc. 50(10):765–778
39. Hendriks R. C., Martin R. (2007) MAP estimators for speech enhancement under normal and Rayleigh inverse Gaussian distributions. IEEE Trans. Audio. Speech. Lang. Process. 15(3):918–27
40. Davis A., Nordholm S., Low S. Y., Togneri R. (2006) A multi-decision sub-band voice activity detector. In: Proc. 14th European Conference on Signal Processing (EUSIPCO), Florence, Italy
41. Sohn J., Sung W. (1998) A voice activity detector employing soft decision based noise spectrum adaptation. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), vol.1. pp 365–368
42. Quackenbush S. R., Barnwell T. P., Clements M. A. (1988) Objective measures of speech quality. Prentice-Hall, Englewood Cliffs, NJ
43. Hu Y., Loizou P. C. (2008) Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio. Speech. Lang. Process. 16(1):229–238
44. Gerkmann T., Hendriks R. C. (2011) Noise power estimation based on the probability of speech presence. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz. pp 145–148
45. Shakil M., Kibria B. G. (2009) Exact distributions of the linear combination of gamma and Rayleigh random variables. Austrian J. Stat. 38(1):33–44
46. Hitczenko P. (1998) A note on a distribution of weighted sums of iid Rayleigh random variables. Sankhya: Indian J. Stat., Series A 60:171–175