Deng and Bao *EURASIP Journal on Audio, Speech, and Music Processing* (2015) 2015:28
DOI 10.1186/s13636-015-0073-6

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Speech enhancement based on Bayesian decision and spectral amplitude estimation

Feng Deng and Chang-Chun Bao*

Abstract

In this paper, a single-channel speech enhancement method based on Bayesian decision and spectral amplitude estimation is proposed, in which the speech detection module and spectral amplitude estimation module are included, and the two modules are strongly coupled. First, under the decisions of speech presence and speech absence, the optimal speech amplitude estimators are obtained by minimizing a combined Bayesian risk function, respectively. Second, using the obtained spectral amplitude estimators, the optimal speech detector is achieved by further minimizing the combined Bayesian risk function. Finally, according to the detection results of speech detector, the optimal decision rule is made and the optimal spectral amplitude estimator is chosen for enhancing noisy speech. Furthermore, by considering both detection and estimation errors, we propose a combined cost function which incorporates two general weighted distortion measures for the speech presence and speech absence of the spectral amplitudes, respectively. The cost parameters in the cost function are employed to balance the speech distortion and residual noise caused by missed detection and false alarm, respectively. In addition, we propose two adaptive calculation methods for the perceptual weighted order p and the spectral amplitude order β concerned in the proposed cost function, respectively. The objective and subjective test results indicate that the proposed method can achieve a more significant segmental signal-noise ratio (SNR) improvement, a lower log-spectral distortion, and a better speech quality than the reference methods.

Keywords: Speech enhancement; Bayesian decision; Spectral amplitude estimation; Combined Bayesian risk function; General weighted cost function

1 Introduction

Speech enhancement could improve the quality of noisy speech, which results in a broad range of applications, such as mobile speech communication, robust speech recognition, aids for the hearing impaired, and so on. Therefore, speech enhancement has widely attracted research, and a large number of speech enhancement algorithms, for example, spectral subtraction (SS) method [1], wavelet de-noising method [2], subspace method [3], speech enhancement based on human auditory perceptual model [4], the minimum mean square error (MMSE) estimator of Ephraim-Malah [5], log-spectral amplitude (LSA) estimator [6], and speech enhancement based on speech presence uncertainty [7], have been proposed. Some speech enhancement methods [1, 4–7] are often operated in the discrete Fourier transform (DFT) domain, that is, the enhanced speech is obtained by estimating DFT coefficients of clean speech from the noisy speech.

As we all know, speech signal is present only in some frames based on short-time analysis, and only some frequency bins contain significant energy in each frame. This means that the spectral amplitude of speech signal is generally sparse. However, the existing speech enhancement methods do not take the sparse characteristics into consideration and often only focus on estimating the spectral amplitude rather than detecting the speech presence or speech absence. Although the SS method [1] could detect the existence of speech by signal power in the frequency domain, it is so simple that SS method often randomly produces ‘music noise’ caused by falsely detecting noise peaks as speech. Under the assumption of speech presence uncertainty, Ephraim and Malah derived a short-time spectral amplitude (STSA) estimator [5] by applying speech presence uncertainty to the MMSE method, which

* Correspondence: baochch@bjut.edu.cn
Speech and Audio Signal Processing Lab, School of Electronic Information and Control Engineering, Beijing University of Technology, 100124 Beijing, China

can improve the enhancement performance of the MMSE method [5]. Furthermore, combining the speech presence uncertainty with the LSA estimator [6], the optimal modified log-spectral amplitude (OM-LSA) estimator [8] was proposed. These speech estimators based on the speech presence uncertainty can yield reasonable enhancement results for the stationary noise environments. However, under the non-stationary noise conditions, the performance of these estimators may be degraded since the time-varying noise energy results in a false calculation about speech presence probability. In addition, some speech enhancement methods employed voice activity detection (VAD) [9, 10] to detect the existence of speech, but with the decrease of the signal-noise ratio (SNR) and the increases of non-stationary characteristics of the noise, the performance of the VAD methods often become worse. Consequently, the performance of speech enhancement is decreased. Moreover, the VAD methods are usually carried out frame by frame, and therefore, they cannot detect the existence of speech in frequency bins. Considering the significance of speech detection and estimation for speech enhancement, a simultaneous detection and estimation approach (SDEA) for speech enhancement was presented [11], which includes the detection and estimation operations simultaneously. However, the quadratic spectral amplitude (QSA) error was used as its cost function, which limits the ability of noise reduction and affects the enhancement performance of the method.

In order to solve the aforementioned problems, we propose a single-channel speech enhancement method based on Bayesian decision and spectral amplitude estimation (BDSAE), in which the importance of the speech detection and estimation for speech enhancement are jointly considered. The speech detection module and spectral amplitude estimation module are included in this method, and the two modules are strongly coupled. First, the optimal speech amplitude estimators under each of the decisions (i.e., speech presence or speech absence) are obtained by minimizing a combined Bayesian risk function. Second, using the obtained spectral amplitude estimators, the optimal speech detector for the existence of speech signal in spectral amplitudes is achieved by further minimizing the combined Bayesian risk function. Finally, according to the results of speech detector, the decision rule is made, and thus the final optimal spectral amplitude estimator is selected for enhancing noisy speech. Furthermore, by taking into account both detection and estimation errors, we propose a combined cost function, in which the cost parameters are used to balance the speech distortion and residual noise caused by missed detection and false alarm, respectively. Moreover, the combined cost function consists of two general weighted distortion measures under the speech presence or speech absence of spectral amplitudes, in which the perceptual weighted order p [12–14] and the spectral amplitude order β [15, 16] are jointly used. In order to obtain more flexible and effective gain functions, the parameters p and β are adaptively estimated, that is, the parameter p is made to be a frequency-dependent value, and the value of β is calculated according to the posterior SNR. To summarize, the BDSAE method not only considers the sparse characteristics of spectral amplitudes of speech signal (i.e., speech detection) but also takes the full advantages of both the traditional perceptual weighted estimators [12, 14] and β-order spectral amplitude estimators [15, 16] (i.e., speech estimation), which can obtain more flexible and effective gain functions for speech enhancement. The experiment results indicate that the proposed BDSAE method can improve the quality of enhanced speech both in terms of subjective and objective measures.

The remainder of this paper is organized as follows. In Section 2, the proposed BDSAE speech enhancement method is described. In Section 3, we present the adaptive calculation methods for the perceptual weighted order p and the spectral amplitude order β, respectively. In Section 4, we describe the implementation of the proposed BDSAE method. The performance evaluation is presented in Section 5, and Section 6 gives the conclusions.

2 The proposed BDSAE speech enhancement method

In this section, we first present conventional spectral amplitude estimation scheme for speech enhancement. Then, the proposed speech enhancement scheme based on Bayesian decision and spectral amplitude estimation is described. Finally, we derive the optimal decision rule and spectral amplitude estimator by introducing general weighted cost functions.

2.1 Conventional spectral amplitude estimation scheme

Assuming that the clean speech signal $x(n)$ is contaminated by an uncorrelated additive noise $d(n)$, then the noisy speech signal $y(n)$ can be expressed as: $y(n) = x(n) + d(n)$. By taking a DFT of $y(n)$, we can obtain the following expression about $y(n)$ in frequency domain:

$$Y(\omega_k) = X(\omega_k) + D(\omega_k) \tag{1}$$

where n is the time domain index of the speech signal. $Y(\omega_k)$, $X(\omega_k)$, and $D(\omega_k)$ denote the kth DFT coefficients of noisy speech, clean speech, and noise signal, respectively. $\omega_k = 2\pi k/N$, k is the index of frequency bins, and N is the frame length.

Since the human auditory system is not sensitive to the phase spectrum, we can replace the phases of clean speech and noise signal by the one of the noisy speech, and then we can rewrite Eq. (1) in polar form as follows:

$$\begin{aligned} Y_k e^{j\theta_y(k)} &= X_k e^{j\theta_x(k)} + D_k e^{j\theta_d(k)} \\ &\approx X_k e^{j\theta_y(k)} + D_k e^{j\theta_y(k)} \\ &= (X_k + D_k) e^{j\theta_y(k)} \end{aligned} \tag{2}$$

where Y_k, X_k, and D_k denote the *k*th spectral magnitudes of the noisy speech, clean speech, and noise signal, respectively. $\theta_y(k)$, $\theta_x(k)$, and $\theta_d(k)$ are the phases corresponding to the frequency bin k of the noisy speech, clean speech, and noise signal, respectively.

From (2), we can obtain $Y_k = X_k + D_k$. That is to say, we can ignore the phases of clean speech and noise signal and mainly focus on estimating the spectral magnitude of clean speech from the noisy speech signal.

For the conventional Bayesian spectral amplitude estimation methods [4–8], the speech spectral amplitude estimation $\hat{X}_k$ is obtained by minimizing the expectation of a given cost function $C(X_k, \hat{X}_k)$, which can be defined as follows:

$$\hat{X}_k = \arg\min E\left\{C\left(X_k, \hat{X}_k\right)\right\} \tag{3}$$

where $E\{.\}$ denotes the statistical expectation. The $C\left(X_k, \hat{X}_k\right)$ is the cost function.

However, these methods do not take the sparse characteristics of spectral amplitudes of speech signal into consideration, and thus, they just focus on estimating the spectral amplitudes of speech signal rather than speech detection and spectral amplitude estimation simultaneously. That is, for the speech presence or speech absence of spectral amplitudes in each frequency bin, the cost function $C\left(X_k, \hat{X}_k\right)$ of the conventional methods is the same, which limits the performance of speech enhancement. Therefore, by taking into account both speech detection and estimation, we propose a new speech enhancement method based on Bayesian decision and spectral amplitude estimation.

2.2 Bayesian decision and spectral amplitude estimation scheme

In this section, we reformulate the speech enhancement as a Bayesian decision and estimation problem under two hypotheses with the framework of statistical decision theory [17, 18].

First, according to the sparsity of speech spectral magnitude, some frequency bins are speech dominant (i.e., speech presence) and some frequency bins are noise dominant (i.e., speech absence). In this way, for the *k*th spectral magnitude of the noisy speech Y_k, we let $Hk\ 0$ and $Hk\ 1$ denote, respectively, speech absence and speech presence hypotheses in the frequency bin k [11, 13]:

$$\begin{aligned} H_0^k &: \ Y_k = D_k \\ H_1^k &: \ Y_k = X_k + D_k \end{aligned} \tag{4}$$

Then, the Bayesian decision is employed to detect the two hypotheses, so we define two decision spaces $\eta k\ j$ $(j = 0,\ 1)$ for detecting the speech presence or speech absence in the frequency bin k. In this way, if the decision $\eta k\ 0$ is made, the speech hypothesis $Hk\ 0$ is accepted, which means speech is absent in the frequency bin k, and thus the corresponding enhanced speech $\hat{X}_k = \hat{X}_{k,0}$ is obtained. Similarly, if the decision $\eta k\ 1$ is made, the speech hypothesis $Hk\ 1$ is detected, which means speech is present in the frequency bin k, and then the corresponding speech estimation $\hat{X}_k = \hat{X}_{k,1}$ is achieved.

Finally, using the speech presence or not hypotheses $Hk\ i(i = 0,\ 1)$ and decision spaces $\eta k\ j(j = 0,\ 1)$, we can reformulate the speech enhancement as the Bayesian decision and spectral amplitude estimation problem, which is presented as follows:

Let the cost function $C_j\left(X_k, \hat{X}_k\right)$ denote the cost for making a decision $\eta k\ j$(and choosing the speech estimator $\hat{X}_{k,j}$), and we can consider the detection decision η ($\eta k\ 0$ or $\eta k\ 1$) as the function of $Y(\omega_k)$, i.e., $\eta = \psi(Y(\omega_k))$. Therefore, for making a decision $\eta = \psi(Y(\omega_k))$, the combined cost function $\tilde{C}\left(X_k, \hat{X}_k \middle| \psi(Y(\omega_k))\right)$ can be presented as follows:

$$\tilde{C}\left(X_k, \hat{X}_k \middle| \psi(Y(\omega_k))\right) = \sum\nolimits_{j=0}^{1} p\left(\eta_j^k \middle| Y(\omega_k)\right) C_j\left(X_k, \hat{X}_k\right) \tag{5}$$

where $p(\eta k\ j|Y(\omega_k))$ is a conditional decision probability. For notation simplification, we omit the frequency bin indices later.

Applying the combined cost function of (5) into (3), the combined Bayesian risk function R can be defined by the following:

$$\begin{aligned} R &= E\left[\tilde{C}\left(X, \hat{X} \middle| \psi(Y(\omega))\right)\right] \\ &= \int_{\Omega_y}\int_{\Omega_x} \tilde{C}\left(X, \hat{X} \middle| \psi(Y(\omega)\right) p(Y(\omega), X) dX dY(\omega) \\ &= \int_{\Omega_y}\int_{\Omega_x} \tilde{C}\left(X, \hat{X} \middle| \psi(Y(\omega)\right) p(Y(\omega)|X) p(X) dX dY(\omega) \end{aligned} \tag{6}$$

where Ω_x and Ω_y denote the spaces of clean speech and noisy speech, respectively. $p(X)$ is the priori probability of spectral magnitude which can be defined as follows [11, 13]:

$$p(X) = qp(X|H_1) + (1-q)p(X|H_0) \tag{7}$$

where $q = p(H_1)$ denotes the priori speech presence probability, and $p(X|H_0) = \delta(X)$ is the Dirac delta function [11].

Since the cost functions $C_j\left(X, \hat{X}\right)$ are different for speech hypothesis H_0 and hypothesis H_1, we let $C_{ij}\left(X, \hat{X}\right) = C_j\left(X, \hat{X}|H_i\right)$ denote the cost that is conditioned on the true H_i and the decision η_j. Namely, the cost function relies on both the true speech X under H_i and the estimated speech $\hat{X}$ under decision η_j. Thus, the cost function couples the two modules of speech detection and spectral amplitude estimation. By substituting (7) into (6), we can get

$$\begin{aligned} R = \int_{\Omega_y} dY(\omega) \int_{\Omega_x} dX p(Y(\omega)|X) \{ \\ \times p\left(\eta_0|Y(\omega)\right)\left[qp(X|H_1)C_{10}\left(X,\hat{X}\right) + (1-q)p(X|H_0)C_{00}\left(X,\hat{X}\right)\right] \\ + p(\eta_1|Y(\omega))\left[qp(X|H_1)C_{11}\left(X,\hat{X}\right) + (1-q)p(X|H_0)C_{01}\left(X,\hat{X}\right)\right]\} \end{aligned} \tag{8}$$

Given the hypothesis-decision pair $\{H_i, \eta_j\}$, we define the risk $r_{ij}(Y(\omega))$ as follows [11]:

$$r_{ij}(Y(\omega)) = \int_{\Omega_x} C_{ij}\left(X, \hat{X}\right) p(X|H_i) p(Y(\omega)|X) dX \tag{9}$$

According to (9), the combined Bayesian risk function R in (8) can be rewritten as:

$$\begin{aligned} R = \int_{\Omega_y} dY(\omega) \{ p\left(\eta_0|Y(\omega)\right)[qr_{10}(Y(\omega)) + (1-q)r_{00}(Y(\omega))] \\ + p(\eta_1|Y(\omega))[qr_{11}(Y(\omega)) + (1-q)r_{01}(Y(\omega))]\} \end{aligned} \tag{10}$$

In (10), since the decision probability $p(\eta_j|Y(\omega)) \in \{0, 1\}$ is binary, for minimizing the combined Bayesian risk function R, we first estimate the optimal spectral amplitude $\hat{X}_j$ under each of the decisions η_j. Second, using the obtained $\hat{X}_j$, the optimal speech presence decision η_j can be derived by further minimizing the combined Bayesian risk function R. Namely, according to the two-stage minimization process of (10), the optimal speech decision rule can be given by:

$$\begin{aligned} \eta_j \\ = \begin{cases} \eta_1, \text{ if } q[r_{10}(Y(\omega)) - r_{11}(Y(\omega))] \geq (1-q)[r_{01}(Y(\omega)) - r_{00}(Y(\omega))] \\ \qquad \eta 0, \text{ otherwise} \end{cases} \end{aligned} \tag{11}$$

Under the speech presence decision η_j, the spectral amplitude estimation $\hat{X}_j$ can be obtained from (10) by:

$$\hat{X}_j = \mathit{arg\,min}\{qr_{1j}(Y(\omega)) + (1-q)r_{0j}(Y(\omega))\}, j = 0, 1 \tag{12}$$

Figure 1 shows the comparison of two schemes of speech presence decision and spectral amplitude estimation. Figure 1a is a conventional independent detection and estimation system that consists of an estimator and a detector. The estimator and detector are not coupled which independently choose to accept or reject the estimator output, such as the well-known SS method [1]. The SS method estimates the speech spectrum by subtracting the estimated noise spectrum from the noisy speech spectrum [1] and thresholding the result according to some desired residual noise level. In fact, the thresholding process is a detector in the frequency bins: the speech spectral coefficients are assumed to be present in noisy speech spectral coefficients if their energies are above the threshold; otherwise, the speech spectral coefficients are considered to be absent in noisy speech spectral coefficients. That is to say, the speech estimator and detector are independent.

Figure 1b is the proposed speech detection and estimation scheme, where the estimator is obtained by (12) and the interrelated decision rule of (11) is used to choose the appropriate estimator, $\hat{X}_0$ or $\hat{X}_1$, for minimizing the combined Bayesian risk R. Since the risk $r_{ij}(Y(\omega))$ existing both in (11) and (12) is a function of the speech estimation

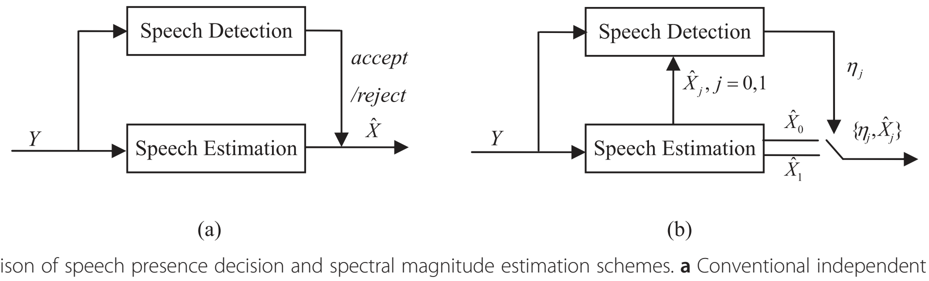

Fig. 1 The comparison of speech presence decision and spectral magnitude estimation schemes. **a** Conventional independent detection and estimation system. **b** Strongly coupled detection and estimation system

$\hat{X}_j$, the decision rule of (11) requires information of the speech estimator under each of its own decisions, which we can see the arrows between speech detection block and speech estimation block in Fig. 1b. That is, the speech detection and speech estimation are strongly coupled in the proposed scheme. In this way, if the decision η_0 is made, the speech hypothesis H_0 is accepted, which means speech is absent in noisy speech spectral coefficients, and thus the corresponding enhanced speech $\hat{X} = \hat{X}_0$ is obtained. Similarly, if the decision η_1 is made, the speech hypothesis H_1 is detected, which means speech is present in noisy speech spectral coefficients, therefore, the corresponding speech estimation $\hat{X} = \hat{X}_1$ is achieved.

2.3 The derivation of BDSAE based on the general weighted cost functions

In this section, based on a general weighted cost functions, we first derive the optimal speech existence decision rule of (11) and spectral amplitude estimators of (12) for the BDSAE system by minimizing the combined Bayesian risk function *R*. Then the gain's change process of the BDSAE system is analyzed. Next, we discuss the influences of cost parameters in weighted cost functions for the BDSAE system. Finally, the influences of p and β parameters for the BDSAE system are demonstrated.

From (11) and (12), we can see that both the optimal speech detector and spectral amplitude estimator contain the risk $r_{ij}(Y(\omega))$ which depends on the cost function $C_{ij}(X,\ \hat{X})$. The cost function plays a significant role in the Bayesian spectral amplitude estimator. For different cost function, we can derive various kinds of spectral amplitude estimators and obtain different speech enhancement performance. In this paper, not only the speech estimation error need to be considered but also the speech detection error should be taken into account. Therefore, we present the cost function associated with the hypothesis-decision pair $\{H_i, \eta_j\}$ [11, 13]:

$$C_{ij}\left(X,\hat{X}\right) = c_{ij}d_{ij}\left(X,\hat{X}\right) \tag{13}$$

where i and j are the indices of speech hypothesis and decision space, respectively; $d_{ij}\left(X,\hat{X}\right)$ is the distortion measure which is defined in (14); and c_{ij} is the cost parameter which is used to balance the costs associated with the hypothesis-decision pair $\{H_i, \eta_j\}$. The cost parameters c_{00} and c_{11} indicate the decision is correct, namely, there is no cost need to balance, so their values are equal to 1 here; c_{01} is used to balance the cost of false alarm (i.e., the speech absence is detected as speech presence), which can avoid too much noise residual in the enhanced speech; and c_{10} is used to balance the cost of miss detection (i.e., the speech presence is detected as speech absence), which can control speech distortion in the enhanced speech.

For speech hypothesis H_i ($i = 0, 1$), the general weighted distortion measure $d_{ij}\left(X,\hat{X}\right)$ is defined as follows:

$$d_{ij}\left(X,\hat{X}\right) = \begin{cases} X^p\left(X^\beta - \hat{X}_j^\beta\right)^2, & \text{if } H_i = H_1 \\ \left(\left(G_f Y\right)^\beta - \hat{X}_j^\beta\right)^2, & \text{if } H_i = H_0 \end{cases} \tag{14}$$

where i and j are the indices of speech hypothesis and decision space; G_f denotes gain floor factor, p is the perceptual weighted order, and β is the spectral amplitude order.

From (14) we can see that, for speech hypothesis H_1, the perceptual weighted order p [12–14] and the spectral amplitude order β [15, 16] are jointly incorporated into the distortion measure. For speech hypothesis H_0, the gain floor factor G_f is employed to the distortion measure which allows some comfort background noise level in the enhanced speech.

1. *Speech estimator*: Assuming both $X(\omega)$ and $D(\omega)$ are zero-mean, complex Gaussian variables with variances $\lambda_x = E\{X^2\}$ and $\lambda_d = E\{D^2\}$, respectively. By substituting (9), (13), and (14) into (12), we have

$$\begin{aligned}\hat{X}_j = \ & \arg\min\Big\{qc_{1j}\int_{\Omega_x} X^p\left(X^\beta - \hat{X}_j^\beta\right)^2 p(X|H_1)p(Y(\omega)|X)dX \\ & + (1-q)c_{0j}\int_{\Omega_x}\left(\left(G_f Y\right)^\beta - \hat{X}_j^\beta\right)^2 p(X|H_0)p(Y(\omega)|X)dX\Big\}\end{aligned} \tag{15}$$

According to Bayesian criterion, by taking the derivative of (15) with respect to $\hat{X}_j$ and setting it to zero, we can get

$$\begin{aligned} & qc_{1j}\int_{\Omega_x} -2\beta\hat{X}_j^{\beta-1}X^p\left(X^\beta - \hat{X}^\beta\right)p(X|H_1)p(Y(\omega)|X)dX \\ & - (1-q)c_{0j}2\beta\hat{X}_j^{\beta-1}\left(\left(G_f Y\right)^\beta - \hat{X}_j^\beta\right)p(Y(\omega)|H_0) = 0\end{aligned} \tag{16}$$

By solving (16), we have

$$\begin{aligned} & \hat{X}_j^\beta\left[qc_{1j}\int_{\Omega_x} X^p p(X|H_1)p(Y(\omega)|X)dX + (1-q)c_{0j}p(Y(\omega)|H_0)\right] \\ & = \left[qc_{1j}\int_{\Omega_x} X^{p+\beta}p(X|H_1)p(Y(\omega)|X)dX + (1-q)c_{0j}\left(G_f Y\right)^\beta p(Y(\omega)|H_0)\right]\end{aligned} \tag{17}$$

Dividing (1-q)$p(Y(\omega)|H_0)$ on both sides of (17), we can obtain

$$\begin{aligned}&\hat{X}_j^{\beta}\left[c_{1j}\Lambda(Y(\omega))\int_{\Omega_x}X^p p(X|Y(\omega))dX+c_{0j}\right]\\&\quad=c_{1j}\Lambda(Y(\omega))\int_{\Omega_x}X^{p+\beta}p(X|Y(\omega))dX+c_{0j}\left(G_fY\right)^{\beta}\end{aligned}\tag{18}$$

where $\Lambda(Y(\omega))=\frac{q}{1-q}\frac{p(Y(\omega)|H_1)}{p(Y(\omega)|H_0)}$ is the generalized likelihood ratio.

By solving (18) for $\hat{X}_j^{\beta}$, we have

$$\hat{X}_j^{\beta}=\frac{c_{1j}\Lambda(Y(\omega))\int_{\Omega_x}X^{p+\beta}p(X|Y(\omega))dX+c_{0j}\left(G_fY\right)^{\beta}}{c_{1j}\Lambda(Y(\omega))\int_{\Omega_x}X^p p(X|Y(\omega))dX+c_{0j}}\tag{19}$$

According to [16], we have

$$\int_0^{\infty}X^{\mu}p(X|Y(\omega))dX=\phi^{\mu/2}\Gamma\left(\frac{\mu}{2}+1\right)\Phi\left(-\frac{\mu}{2},1;-\nu\right)\tag{20}$$

where μ denotes the spectral amplitude order. $\Gamma(\cdot)$ is the gamma function, and $\Phi(\cdot)$ denotes the confluent hypergeometric function. For $\phi^{\mu/2}$ of (20), we can simplify it as follows:

$$\begin{aligned}\phi^{\mu/2}&=\left(\frac{\lambda_x\lambda_d}{\lambda_x+\lambda_d}\right)^{\mu/2}\\&=\left(\frac{\lambda_x}{1+\xi}\right)^{\mu/2}=\left(\frac{\xi\lambda_d}{1+\xi}\right)^{\mu/2}\\&=\left(\frac{\xi Y^2}{(1+\xi)\gamma}\frac{\gamma}{\gamma}\right)^{\mu/2}=\left(\frac{\sqrt{\nu}}{\gamma}Y\right)^{\mu}\end{aligned}\tag{21}$$

where $\lambda_x=E\{X^2\}$ and $\lambda_d=E\{D^2\}$ are the speech and noise variances, respectively. ξ is a priori SNR, γ is a posteriori SNR, and ν is the function of ξ and γ. Here, ξ, γ, and ν are defined as follows [12]:

$$\xi=\frac{\lambda_x}{\lambda_d},\ \gamma=\frac{Y^2}{\lambda_d},\nu=\frac{\xi}{1+\xi}\gamma\tag{22}$$

By substituting (20), (21), and (22) into (19), we can derive the optimal spectral amplitude estimation $\hat{X}_j$ under the speech decision η_j ($j=0$, 1):

$$\begin{aligned}\hat{X}_j&=\left(\frac{c_{1j}\Lambda(Y(\omega))\left[\left(\frac{\sqrt{\nu}}{\gamma}\right)^{\beta}\Gamma\left(\frac{p+\beta}{2}+1\right)\Phi\left(-\frac{p+\beta}{2},1;-\nu\right)\right]+c_{0j}G_f^{\beta}}{c_{1j}\Lambda(Y(\omega))\left[\Gamma\left(\frac{p}{2}+1\right)\Phi\left(-\frac{p}{2},1;-\nu\right)\right]+c_{0j}}\right)^{1/\beta}Y\\&=\mathrm{G}_j(\xi,\gamma,p,\beta)\cdot Y\end{aligned}\tag{23}$$

where $G_j(\xi,\ \gamma,\ p,\ \beta)$ is the gain function of BDSAE method under the speech decision η_j.

2. *Speech detector*: From (11), we can find that, in order to obtain an optimal speech presence decision rule, the risk $r_{ij}(Y(\omega))$ requires to be calculated, so for speech hypothesis H_1, we have

$$\begin{aligned}&r_{1j}(Y(\omega))\\&=\frac{c_{1j}\exp\left(-\frac{\gamma}{1+\xi}\right)}{\pi\lambda_d(1+\xi)}\\&\times\left[\begin{array}{l}\phi^{(p/2+\beta)}\Gamma\left(\frac{p}{2}+\beta+1\right)\Phi\left(-\left(\frac{p}{2}+\beta\right),1;-\nu\right)\\+\left(G_jY\right)^{2\beta}\phi^{(p/2)}\Gamma\left(\frac{p}{2}+1\right)\Phi\left(-\frac{p}{2},1;-\nu\right)\\-2\left(G_jY\right)^{\beta}\phi^{(p/2+\beta/2)}\Gamma\left(\frac{p+\beta}{2}+1\right)\Phi\left(-\frac{p+\beta}{2},1;-\nu\right)\end{array}\right]\end{aligned}\tag{24}$$

where ξ is a priori SNR, γ is a posteriori SNR, and ν is the function of ξ and γ, which have been defined in (22). $\phi=\lambda_x\lambda_d/(\lambda_x+\lambda_d)$, the variances of speech and noise λ_x and λ_d can be expressed as $\lambda_x=E\{X^2\}$, $\lambda_d=E\{D^2\}$, respectively. The detailed procedure for deriving risk $r_{1j}(Y(\omega))$ is given in Appendix 1.

For speech hypothesis H_0, we can obtain

$$\begin{aligned}r_{0j}(Y(\omega))&=c_{0j}\left(\left(G_fY\right)^{\beta}-\left(G_jY\right)^{\beta}\right)^2\frac{1}{\pi\lambda_d}\exp\left(-\frac{Y^2}{\lambda_d}\right)\\&=\frac{c_{0j}}{\pi\lambda_d}\left(G_f^{\beta}-G_j^{\beta}\right)^2Y^{2\beta}\exp(-\gamma)\end{aligned}\tag{25}$$

where γ is a posteriori SNR and λ_d is the variance of the noise. The derivation details of risk $r_{0j}(Y(\omega))$ is given in Appendix 2. Therefore, by substituting $r_{1j}(Y(\omega))$and $r_{0j}(Y(\omega))$ into (11), we can obtain the optimal speech presence decision rule.

To conclude the above results, BDSAE from noisy speech requires the following:

(a) Calculating the gain function under each of the speech decisions η_j by (23);
(b) Finding the optimal decision η_j by (11) according to (24) and (25), then the optimal gain function associated with the optimal decision η_j is achieved. Therefore, the corresponding speech estimation is obtained by applying the gain to the noisy speech.

3. *Gains analysis*: Figure 2 demonstrates the gain's change process of BDSAE system versus the value of ($\gamma-1$) that referred to as the instantaneous SNR, where the parameters $c_{01}=1.5$, $c_{10}=5$, $q=0.8$, and $G_f=-15$ dB, respectively. Here, we just call the gain

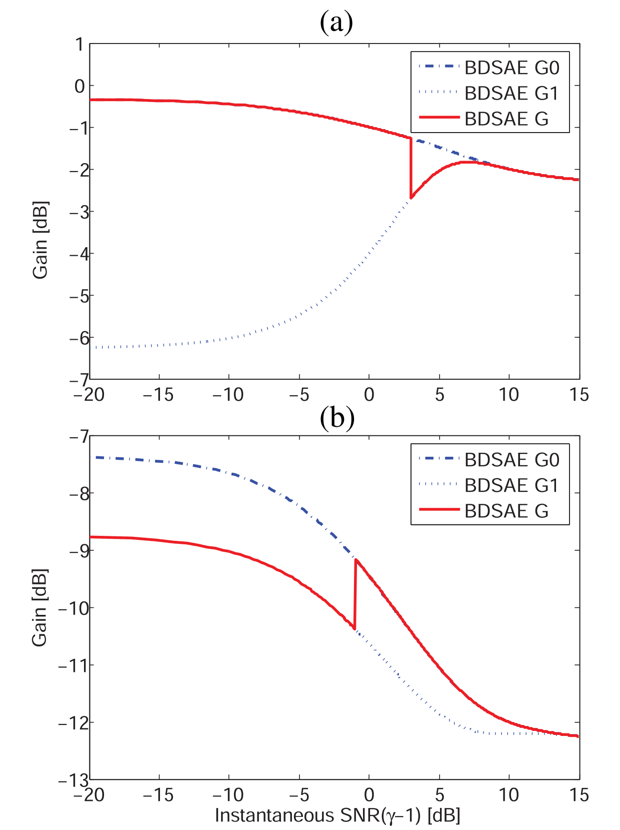

Fig. 2 Gains of Bayesian decision and estimation system versus instantaneous SNR ($\gamma-1$). **a** System gain change from G_0 to G_1, for the case of $p = 0.5$, $\beta = 0.5$, and a priori SNR $\xi = 5$ dB. **b** System gain change from G_1 to G_0, for the case of $p = -0.5$, $\beta = 1.5$, and a priori SNR $\xi = -5$ dB

function $G_j(\xi, \gamma, p, \beta)$ as G_j for convenience. The G_0 (blue dashed line) and G_1 (blue dotted line) are the gains under the decision η_0 and η_1, respectively. The G (red solid line) denotes the gain of BDSAE system.

As shown in Fig. 2a, for a priori SNR of $\xi = 5$ dB, as long as the instantaneous SNR is higher than about 3 dB, the speech decision changes from η_0 to η_1, thus the optimal system gain G changes from G_0 to G_1. Similarly, as shown in Fig. 2b, for a priori SNR of $\xi = -5$ dB, as long as the instantaneous SNR is higher than about −1 dB, the speech decision changes from η_1 to η_0, thus the optimal system gain G changes from G_1 to G_0. Note that if there is an ideal speech detector, a more significantly non-continuous gain would be obtained. However, in the proposed BDSAE scheme, although the speech detector is not ideal, it is optimized to minimize the combined Bayesian risk function R, that is, the non-continuous system gain G of the proposed BDSAE is optimal, which could obtain good enhancement performance shown in Section 5.

4. *Influence of cost parameters*: In addition, from (23) we can see that, in the proposed BDSAE method, the non-continuous system gain G depends on the cost parameters c_{ij} as well as parameters p and β. If the cost parameter c_{01} associated with false alarm is much less than the generalized likelihood ratio $\Lambda(Y(\omega))$, that is, the speech is definitely present in the spectral amplitude, the BDSAE gain function $G_1(\xi, \gamma, p, \beta)$ under the decision η_1 can be approximated as follows:

$$G_{\mathrm{appr}}(\xi,\gamma) \approx \left(\frac{\left[\left(\frac{\sqrt{\nu}}{\gamma}\right)^{\beta} \Gamma\left(\frac{p+\beta}{2}+1\right) \Phi\left(-\frac{p+\beta}{2}, 1; -\nu\right) \right]}{\left[\Gamma\left(\frac{p}{2}+1\right) \Phi\left(-\frac{p}{2}, 1; -\nu\right) \right]} \right)^{1/\beta} \tag{26}$$

In this way, the gain function $G_1(\xi, \gamma, p, \beta)$ is equal to $G_{\mathrm{appr}}(\xi, \gamma)$, which means a good enhancement effect can be obtained under correct decision η_1. However, if the cost parameter c_{01} is much larger than the generalized likelihood ratio $\Lambda(Y(\omega))$, the speech is absent in the spectral amplitude. In this case, the BDSAE gain function $G_1(\xi, \gamma, p, \beta)$ under the decision η_1 (i.e., false alarm) is equal to G_f approximately (i.e., $G_1(\xi, \gamma, p, \beta) \approx G_f$), and thus the cost of false alarm is compensated and the residual noise in the enhanced speech signals can be reduced effectively. On the other hand, if the cost parameter c_{10} associated with missed detection is much smaller than the inverse of generalized likelihood ratio $\Lambda(Y(\omega))$, the BDSAE gain function $G_0(\xi, \gamma, p, \beta)$ under the decision η_0 is equal to G_f approximately (i.e., $G_0(\xi, \gamma, p, \beta) \approx G_f$). Therefore, it can remove noise greatly when speech is definitely absent. On the contrary, if the cost parameter c_{10} is much greater than the inverse of $\Lambda(Y(\omega))$, the BDSAE gain function $G_0(\xi, \gamma, p, \beta)$ under the decision η_0 (i.e., miss decision) is equivalent to the gain function $G_{\mathrm{appr}}(\xi, \gamma)$ of (26) (i.e., $G_0(\xi, \gamma, p, \beta) \approx G_{appr}(\xi, \gamma)$), so the cost of miss decision can be compensated and the speech distortion can be reduced as well. Here, in order to obtain a better trade-off between speech distortion and noise reduction, the empirical values of cost parameters c_{01} and c_{10} are chosen the same as 1.5.

5. *Influence of p and β parameters*: Furthermore, the p and β parameters are also more important to system gain G of the BDSAE method. Figure 3 shows their influences on gain function $G_j(\xi, \gamma, p, \beta)$ for different p and β values, where the parameters $c_{01} = 1.5$, $c_{10} = 5$, $q = 0.8$, and $G_f = -15$ dB, respectively. Here, the value of $(\gamma - 1)$ is referred to as the instantaneous SNR.

As shown in Fig. 3a, given a fixed parameter $\beta = 0.5$ and the a priori SNR $\xi = -5$ dB, the gain G_0 and G_1 of BDSAE estimator always increase with the increasing of parameter

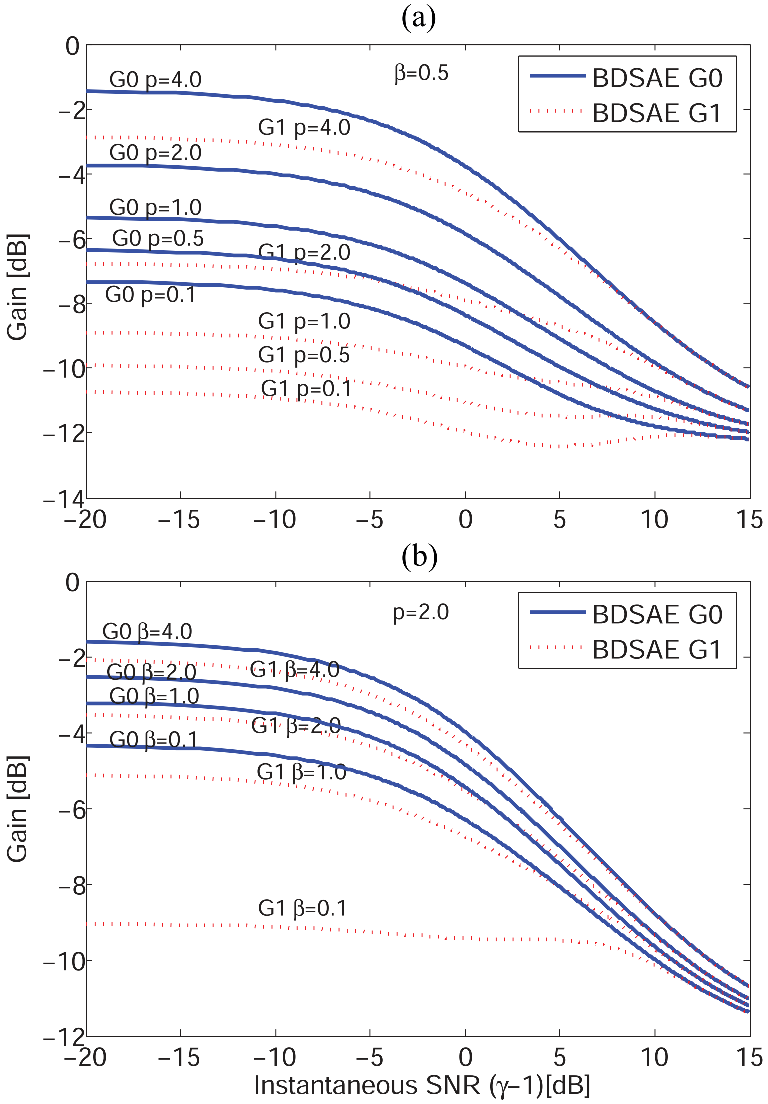

Fig. 3 Gains of BDSAE estimator versus instantaneous SNR ($\gamma - 1$) for different p and β values. **a** Gains versus instantaneous SNR ($\gamma - 1$) for different p values, for the case of $\beta = 0.5$ and a priori SNR $\xi = -5$ dB. **b** Gains versus instantaneous SNR ($\gamma - 1$) for different β values, for the case of $p = 2$ and a priori SNR $\xi = -5$ dB

p when instantaneous SNR ($\gamma - 1$) varies from −20 to 15 dB. That is, for different p values, we can obtain different system gain G values, and the corresponding noise reductions can be achieved.

From Fig. 3b, we can see that the gain G_0 and G_1 of BDSAE estimator also always increase with the increasing of parameter β for a fixed $p = 2$ and the a priori SNR $\xi = -5$ dB when instantaneous SNR ($\gamma - 1$) varies from −20 to 15 dB. Namely, for the different β values, the system gain G values are different, and the noise reduction obtained is also different. In this way, we can obtain the appropriate system gain G values by adaptively choosing the right p and β values, which can yield effective noise reduction and good speech enhancement performance. The adaptive calculation methods of p and β parameters will be presented in Section 3.

3 Adaptive calculation of *p and β* parameters

From the aforementioned analysis, we can see that the perceptually weighted order p and the spectral amplitude order β play an important role in speech enhancement, which can result in a better enhancement performance by choosing appropriate values for p and β. Therefore, in this section, we will present an adaptive calculation method for p and β, respectively.

3.1 Adaptive calculation of parameter *p*

For the calculation of parameter p, in [12], the method did not consider the variability of p, and just a fixed p

value was chosen for the trade-off between noise reduction and speech distortion. Since no flexible gain was introduced, the enhancement performance of the estimator was limited. In [14], the variability of parameter p was considered, and an adaptive calculation method of p was presented, in which the parameter p was considered as a polynomial of the sub-band SNR and auditory perceptual parameter. In this way, for a larger STSA and a smaller STSA, the speech estimation errors can be penalized differently. However, the method shown in [14] needs to calculate the masking thresholds, and thus the pre-enhancement process is required, which increases the computational complexity greatly.

Since most of the speech energy is located at the lower frequencies (i.e., larger STSA) and at the higher frequencies, the speech energy is weakened (i.e., smaller STSA) [19], for the lower frequencies, the value of parameter p should be high and vice versa for the higher frequencies. That is, the estimation error at the higher frequencies is penalized more heavily than that at the lower frequencies. In this way, the residual noise can be suppressed effectively at the higher frequencies, and the speech distortion at the lower frequencies can be reduced at the same time. Therefore, on the basis of such idea, we propose a new adaptive calculation method for parameter p.

First, the appropriate lower bound and higher bound of parameter p for high frequency and low frequency require to be chosen, respectively. As discussed in [12], the p value with more negative produced more noise reduction but the greater speech distortion was introduced as well. Moreover, the $p = -1$ was suggested as a good trade-off between the noise reduction and speech distortion in [12]. Therefore, we choose $p_{\min} = -1$ as the lower bound of parameter p for high frequency. According to [14], in order to reduce the speech distortion at lower frequencies, $p_{\max}$ is set up to 4.0 as the upper bound of parameter p for low frequency in this paper.

Second, since the speech energy usually decreases as frequency increases, for the calculation of p value at the intermediate frequencies, the linear decreasing of p is proposed as a function of the frequency, i.e.,

$$p(k) = p_{\max} - \frac{k(p_{\max} - p_{\min})}{N} \tag{27}$$

where k is the index of frequency bins, N is the frame length, and $p(k)$ denotes the p value of the kth frequency bin.

According to (27), we can obtain the decreased gain from lower frequency to higher frequency, and a larger noise reduction can be achieved at high frequencies, and thus the speech distortion at the higher frequencies is inevitable because the larger STSA sometimes exists at the higher frequencies. In order to reduce the speech distortion at the higher frequencies, we employ the sub-band SNR to modify p. First, the 21 critical sub-bands [20] are divided for each frame of noisy observation. Then the variable $\tilde{p}$ is assumed to be a linear function of the critical sub-band SNR $\Xi(b, k)$, where b is the index of the critical bands. Finally, the range of $\tilde{p}$ is limited as $[\tilde{p}_{\min}, \tilde{p}_{\max}]$ to obtain a trade-off between the noise reduction and speech distortion [16]. In this way, the value of $\tilde{p}$ can be calculated by the following:

$$\tilde{p}(k) = \max\{\min[\mu \cdot \Xi(b,k) + v, \tilde{p}_{\max}], \tilde{p}_{\min}\} \tag{28}$$

where b denotes the index of the critical bands and k is the index of frequency bins. $\Xi(b, k)$ denotes the kth sub-band SNR that belongs to the bth band. The constants μ and v are set to 0.45 and 1.5, respectively, and the minimum and maximum values of $\tilde{p}$ are set to 0.4 and 4.0, respectively, i.e., $\tilde{p}_{\min} = 0.4$ and $\tilde{p}_{\max} = 4.0$.

According to (27) and (28), the final parameter p is obtained by weighting p and $\tilde{p}$:

$$\hat{p}(k) = \varepsilon \cdot \tilde{p}(k) + (1 - \varepsilon) \cdot p(k) \tag{29}$$

where the weighting factor ε is related to the sub-band SNR $\Xi(b, k)$, which is defined by the following:

$$\varepsilon = \frac{1}{1 + \exp(-(\Xi(b,k) - \Xi_0))} \tag{30}$$

where Ξ_0 is a constant. Here, $\Xi_0 = 3.22$ and $\Xi(b, k)$ is defined as follows:

$$\Xi(b,k) = \frac{\sum_{k=B_{\text{low}}(b)}^{B_{\text{up}}(b)} \left| Y(b,k) - \sqrt{\lambda_d(b,k)} \right|^2}{\sum_{k=B_{\text{low}}(b)}^{B_{\text{up}}(b)} \lambda_d(b,k)} \tag{31}$$

in which b denotes the index of the critical bands, k is the index of frequency bins. $B_{\text{up}}(b)$ and $B_{\text{low}}(b)$ denote the upper and lower frequency bound of the bth critical band, respectively. $Y(b, k)$ denotes the kth spectral amplitude of noisy speech that belongs to the bth band, and $\lambda_d(b, k)$ is the kth noise variances that belongs to the bth band.

3.2 Adaptive calculation of parameter β

For the calculation of β, in [15] and [16], the calculation methods of parameter β are based on overall SNR of each frame, and a linear relationship between β and frame SNR was applied. The β only monotonically increases or decreases with the frame SNR increases or decreases. That is, the value of β is fixed and does not vary with the frequency bins in each frame, so it cannot obtain flexible gain, the enhancement performance is limited. For this problem, a solution was proposed in [14], in which the parameter β was interpreted as the compression rate of the spectral amplitude and calculated based on the critical band. That is, the β value is different for different critical band, which can result in a more flexible gain. However, there is no consensus on the degree of compressive nonlinearity at the lower and intermediate frequencies, which might influence the accuracy of β value. Therefore, in this paper, we propose a new calculation method for the parameter β that varies with the frequency bins.

As we know, the higher the a posterior SNR $\gamma(k)$ of (12), the larger the speech presence probability, so β should be larger for reducing speech distortion and vice versa. Therefore, according to $\gamma(k)$, we can employ a monotonically increasing sigmoid function [21] to calculate the value of β.

First, since the strong correlation exists between the adjacent frequency bins, the average posterior SNR $\tilde{\gamma}(k)$ is obtained by applying a normalized window to $\gamma(k)$,

$$\tilde{\gamma}(k) = \sum_{i=-L_h}^{L_h} h(i)\gamma(k-i) \tag{32}$$

where h is a normalized hamming window with length $2L_h + 1$ and $L_h = 5$.

Second, the β value is often limited to the range of [0.001, 4.0] for the trade-off between noise reduction and speech distortion [14–16]. Therefore, the sigmoid function is employed to map the $\tilde{\gamma}(k)$ into (0, $\beta_{\max}$) for parameter β, then we have

$$\beta(k) = \frac{\beta_{\max}}{1 + \exp\left(-\alpha\left(\tilde{\gamma}(k) - \gamma_0\right)\right)} \tag{33}$$

where α is used to control the steepness of the sigmoid function, γ_0 is the position of the inflection point, and $\beta_{\max}$ is a constant. They are set to 0.42, 3.5, and 4.0, respectively. And $\beta(k)$ denotes the β value of the kth frequency bin.

Finally, we limit the minimum value of β to 0.001, that is, the final β is obtained by:

$$\hat{\beta}(k) = \max\{\beta(k), \beta_{\min}\} \tag{34}$$

where k is the index of frequency bins, $\beta_{\min} = 0.001$. Figure 4 gives the variation of β values versus a posterior SNR $\gamma(k)$.

As shown in Fig. 4, the β value of the proposed method is a monotonically increasing function with respect to $\gamma(k)$. Namely, the larger noise reduction can be yielded as $\gamma(k)$ decreases, and the lower speech distortion can be achieved when $\gamma(k)$ increases.

4 Implementation of the proposed method

In this section, we present the implementation of the proposed BDSAE method. The block diagram of the implementation is given in Fig. 5. Firstly, the noisy speech is windowed and transformed into frequency domain by DFT. Secondly, the minima controlled recursive averaging (MCRA) method [22] is employed to estimate the noise power spectrum. Thirdly, using the spectral amplitude of the noisy speech and the estimated noise power spectrum, the critical sub-band SNRs are obtained. Fourthly, the parameters p and β are adaptively calculated according to the critical sub-band SNRs and a posteriori SNRs. Finally, combining a posteriori SNR and a priori SNR obtained by a decision-directed (DD) method [5], the optimal spectral amplitude estimator and decision rule of the BDSAE method are derived by further minimizing the combined Bayesian risk function R, which are used to enhance DFT coefficients of the noisy speech. Then the inverse Fourier transform and the overlap-adding algorithm are performed to obtain the enhanced speech signal in the time domain.

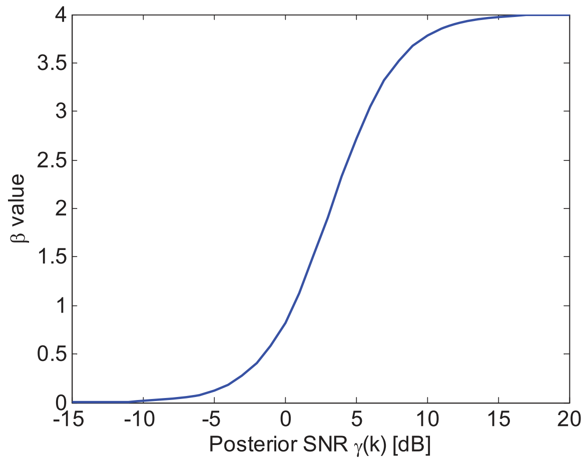

Fig. 4 The variation of β values versus the a posterior SNR $\gamma(k)$

5 Performance evaluation

In this section, we discuss the performance evaluation of the proposed BDSAE method. First, the experimental setup of the proposed method is described. Then, we compare the objective and subjective experimental results between the proposed method and the reference methods.

5.1 Experimental setup

In order to evaluate the performance of the proposed BDSAE method for speech enhancement, white Gaussian noise, street noise, and interior Volvo car noise from ITU-T noise database and babble noise, factory noise, and Fl6 cockpit noise from NOISEX-92 [23] database were used in the test experiments. Twenty-four speech sentences were taken from the Chinese sub-database of NTT speech database, where 12 sentences produced by two female speakers (i.e., six sentences for each female speaker)

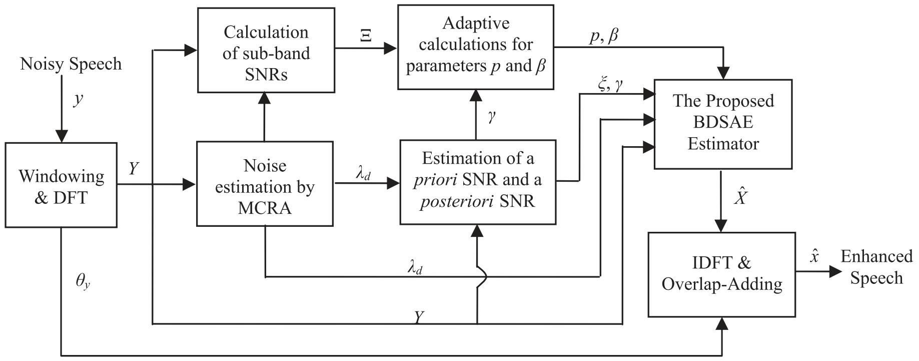

Fig. 5 The block diagram of the proposed method

and another 12 sentences produced by two male speakers (i.e., six sentences for each male speaker). All these speech signals and noises are re-sampled at 16 kHz, and all the signals were 8 s in duration. Frame size N is 512 samples, and the samples are sine windowed with 50 % overlap between adjacent frames. The noisy speech signals were produced according to ITU-T P.56 standard [24], and the input SNRs of noisy speech are 0, 5, 10, and 15 dB, respectively.

In the experiments, the MMSE STSA estimator [5], the SDEA estimator [11], the weighted Euclidean distortion measure (WEDM) estimator [12], and the β-STSA estimator [16] are chosen as the reference methods for comparing with the proposed BDSAE method. The DD method [5] is applied to all these reference methods and the BDSAE method. The MCRA algorithm [22] is used for these methods to estimate noise power spectrum from noisy speech signals. All the reference methods we used were implemented according to the referenced papers, and the corresponding parameters of the methods were not tuned as well.

For the performance evaluation of the speech enhancement methods, the segmental SNR (SNR$_{\text{seg}}$) measure [25], the log-spectral distortion (LSD) measure [11], and the perceptual evaluation of speech quality (PESQ) [26] were used as objective quality evaluation methods. Furthermore, the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) listening test [27] was employed to evaluate the subjective quality.

5.2 Objective quality tests

In this subsection, we describe various objective quality tests. Before we provide rigorous quantitative results, we briefly discuss spectrograms of the signals processed by the proposed system and the reference systems.

1. *Spectrograms*: Figure 6 shows the spectrograms of the input noisy speech (mixed with white Gaussian noise for 0 dB) and the enhanced speech signals obtained by the various enhancement methods. From Fig. 6, we can see that the proposed method outperforms the reference methods.
2. *Segmental SNR*: The SNR$_{\text{seg}}$ [25] measure can be employed to evaluate the objective quality of enhanced speech signals of different speech enhancement methods. The SNR$_{\text{seg}}$ is measured by calculating the SNR for each frame of speech and averaging these SNRs over all test speech sequences, which can be defined by the following:

$$SNR_{\text{seg}} = \frac{1}{L}\sum_{l=0}^{L-1} 10 \cdot \log_{10}\left(\frac{\sum_{n=Nl}^{Nl+N-1} x^2(n)}{\sum_{n=Nl}^{Nl+N-1} [x(n) - \hat{x}(n)]^2}\right) \tag{35}$$

where n denotes the index of signal samples, N is the frame length. l is frame index, and L is the total number of frames. $x(n)$ denotes the clean speech signal, and $\hat{x}(n)$ denotes the enhanced speech signal.

For different input SNRs (i.e., 0, 5, 10, and 15 dB), Fig. 7 gives the comparison of SNR$_{\text{seg}}$ improvement for different enhancement methods under White Gaussian noise. Figure 8 gives the comparison of SNR$_{\text{seg}}$ improvement for different enhancement methods under factory noise.

From Figs. 7 and 8, we can see that, in the case of white Gaussian noise and Factory noise, the SNR$_{\text{seg}}$ improvement of the WEDM method is much better than the other reference methods, but a little worse than the proposed BDSAE method. The SNR$_{\text{seg}}$ improvements of the BDSAE method are nearly 5.0 and 3.0 dB larger than the WEDM method in the white Gaussian noise and factory noise conditions, respectively.

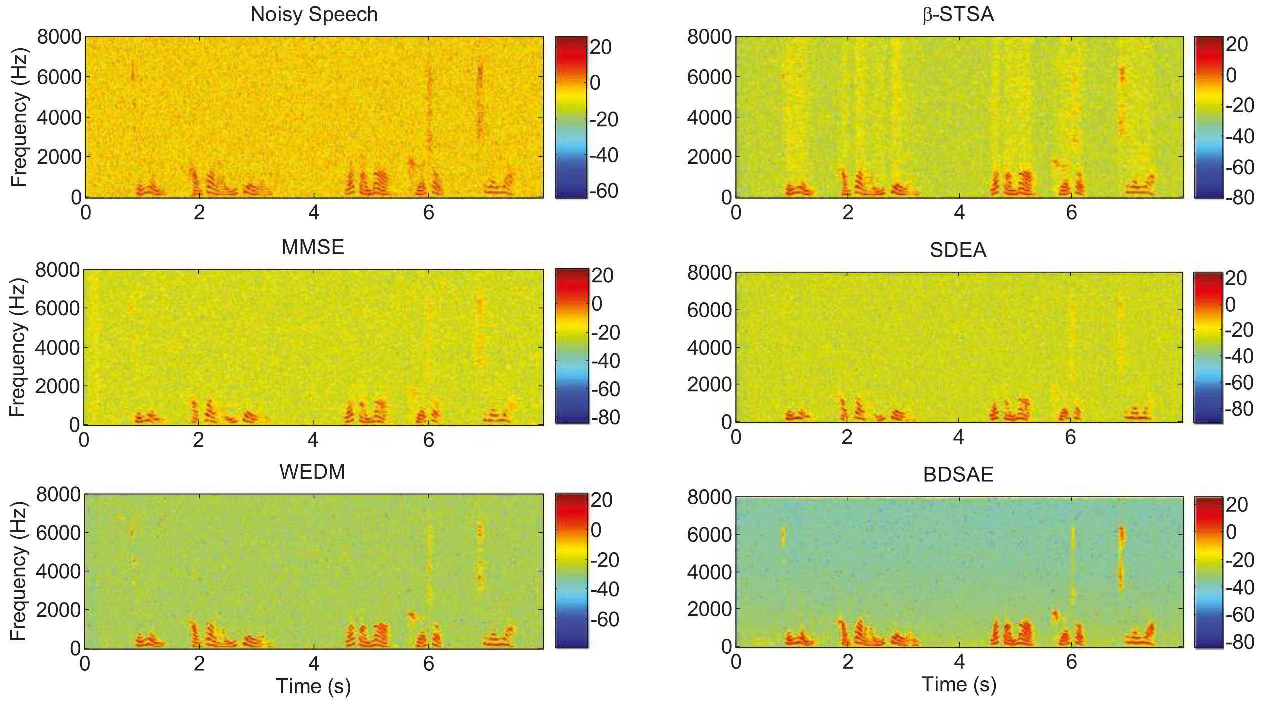

Fig. 6 The speech spectrogram comparison for the various enhancement methods

For each input SNR, the average $\mathrm{SNR}_{\mathrm{seg}}$ improvement of various enhancement methods for six types of noise are presented in Table 1.

From Table 1, we can find that BDSAE method produce much higher average $\mathrm{SNR}_{\mathrm{seg}}$ improvement than the reference methods. Furthermore, for each input SNR, in comparison with the WEDM method whose performance is better than the other reference methods, the average $\mathrm{SNR}_{\mathrm{seg}}$ improvement of the BDSAE method is increased about 3.3 dB for all test noise signals. Therefore, according to the experimental results of Figs. 7 and 8 and Table 1, it is obvious that the BDSAE method performs better than the reference methods.

3. *LSD*: The LSD measure [11] is also used to evaluate the objective quality of the enhanced speech, which measures the similarity between the clean speech spectrum and the estimated speech spectrum. The definition of LSD is given as:

$$d_{\mathrm{LSD}} = \frac{1}{L}\sum_{l=0}^{L-1}\sqrt{\frac{1}{N}\sum_{k=0}^{N-1}\left[10\log_{10}\frac{|\hat{X}(l,k)|^2}{|X(l,k)|^2}\right]^2} \tag{36}$$

where l is the frame index, k is the index of frequency bins, L is the total frames, and N is the frame length.

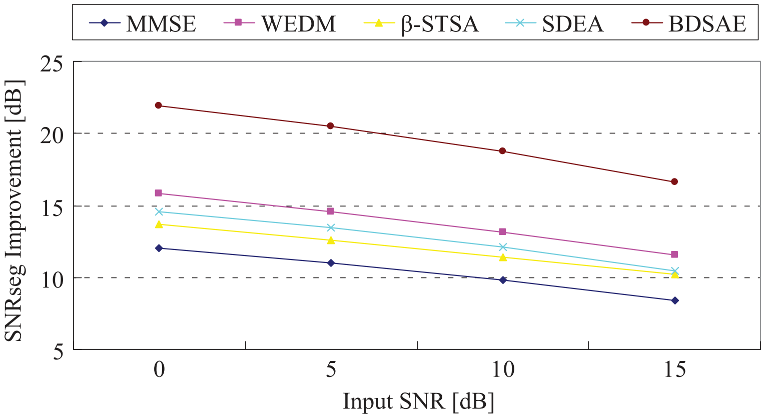

Fig. 7 Comparison of $\mathrm{SNR}_{\mathrm{seg}}$ improvement under White Gaussian noise

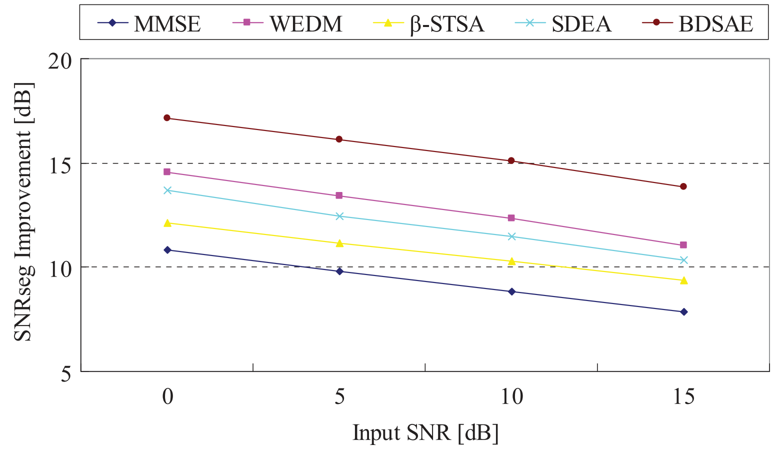

Fig. 8 Comparison of SNRseg improvement under factory noise

$X(l, k)$ denotes the DFT coefficient of the clean speech signal and $\hat{X}(l,k)$ denotes the DFT coefficient of the enhanced speech signal.

According to the idea of [11], the log-spectrum dynamic range of speech signal is confined to about 50 dB for the LSD experiments. The LSD test results are shown in Figs. 9 and 10 for the case of white Gaussian noise and factory noise, respectively. The average LSD test results are given in Table 2 for different input SNRs.

From Figs. 9 and 10, we can see that, in the white Gaussian noise and Factory noise conditions, all speech enhancement methods can obviously reduce the LSD comparing with the noisy speech, where the BDSAE method can obtain much lower LSD than the other reference methods for four SNR conditions. By comparing with the WEDM method whose LSD is lower than the other reference methods, the average LSDs of the BDSAE method are decreased about 3.0 and 1.5 dB for different input SNRs in the white Gaussian noise and factory noise conditions, respectively.

From Table 2, we can see that, by comparing with the LSD of noisy speech, all speech enhancement methods can reduce the LSD to some extent. The average LSD of the BDSAE method is lower than the reference methods in various input SNRs. That is, the proposed method outperforms the reference methods.

Table 1 Test results of SNRseg improvement

Enhancement methods	SNRseg improvement			
	0 dB	5 dB	10 dB	15 dB
MMSE	11.22	10.23	9.25	8.13
WEDM	14.15	13.22	12.16	10.85
β-STSA	11.97	11.13	10.31	9.44
SDEA	13.99	12.86	11.78	10.54
BDSAE	17.45	16.53	15.64	14.48

4. *PESQ*: The PESQ [26] is widely used to assess the objective quality of speech signals, and a higher PESQ score corresponds to a better speech quality. For the case of white Gaussian noise and factory noise, the PESQ test results are compared in Figs. 11 and 12, respectively. The total average PESQ scores of six types of noise are given in Table 3 for four kinds of input SNRs.

From Figs. 11 and 12, we can see that, for the case of white Gaussian noise and Factory noise, in comparison with the reference methods, the BDSAE method yields higher average PESQ scores for various input SNR conditions.

From Table 3, we can find that the average PESQ scores of the enhanced speech signals are all higher than noisy speech signals, which illustrates that the quality of enhanced speech signals produced by all kinds of enhancement methods are improved obviously. In addition, by comparing with the reference methods, the BDSAE method produces higher average PESQ scores for various input SNR conditions. Therefore, it is further confirmed that the proposed BDSAE method is superior to the reference algorithms.

5.3 Subjective quality tests

The quality of enhanced speech is generally assessed by subjective perception, such as speech intelligibility, naturalness, and articulation. The MUSHRA listening test [27] is a commonly used method for the subjective evaluation of audio quality. It requires fewer participants to obtain a statistically significant result [27] reference. Therefore, we employed the MUSHRA listening test to evaluate the subjective quality of enhanced speech. In the MUSHRA test, the subjects are provided with the signals under test as well as one reference and a hidden anchor. The subjects are asked to grade the different signals on a quality scale between 0 and 100,

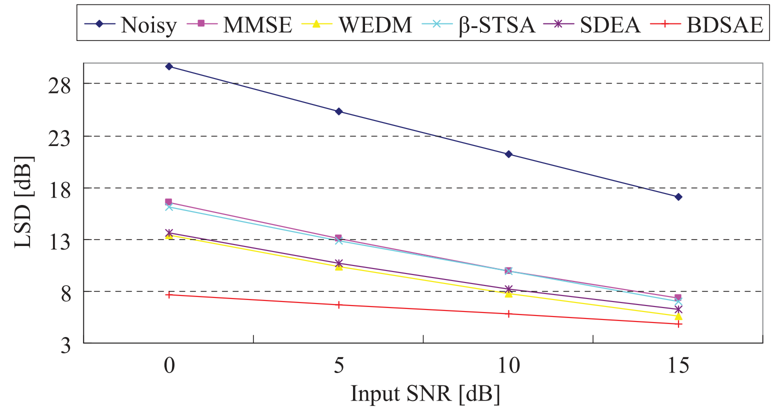

Fig. 9 Comparison of LSD under white Gaussian noise

100 being the best score. As the hidden anchor, we used a speech signal having an SNR of 5 dB less than the noisy speech to be enhanced [20]. The listeners were allowed to listen to each test speech several times and always had access to the clean speech reference.

Six male and four female listeners whose ages are from 20 to 30 years old participated in the MUSHRA tests. Two speech sentences (i.e., one male speaker, one female speaker) were randomly chosen from the aforementioned twenty-four speech sentences, and the corresponding noisy speech sentences contaminated by the aforementioned six types of noise under the different input SNRs (i.e., 0, 5, 10, and 15 dB) were chosen from noisy speech data set which is discussed in Section 5.1. All these noisy speeches were enhanced by the speech enhancement methods and were used for the MUSHRA test. After all the listeners had graded the test signals, a statistical analysis of the results was conducted for the different speech enhancement methods for different input SNRs. Figure 13 shows the MUSHRA listening test results, with the average MUSHRA scores together with the 95 % confidence intervals.

From Fig. 13, we can find that, for four input SNR conditions, the WEDM method yields higher average MUSHRA scores than the other reference methods but lower than the BDSAE method. That is, the proposed BDSAE method performs better than the state-of-the-art reference methods for the subjective quality.

5.4 Discussion

From the aforementioned experimental results, we can see that the proposed BDSAE approach performs better than the reference methods. Herein, we discuss its advantages to the reference methods.

As we know, the spectral amplitudes of speech signal are generally sparse since only some frequency bins contain significant energy in each speech frame. However, the reference methods do not take the sparse characteristics into consideration and often only focus on estimating the speech spectral amplitude rather than detecting their existence in the frequency bins. In this way, for the speech presence or speech absence in the frequency bins, they only use the same gain function to estimate

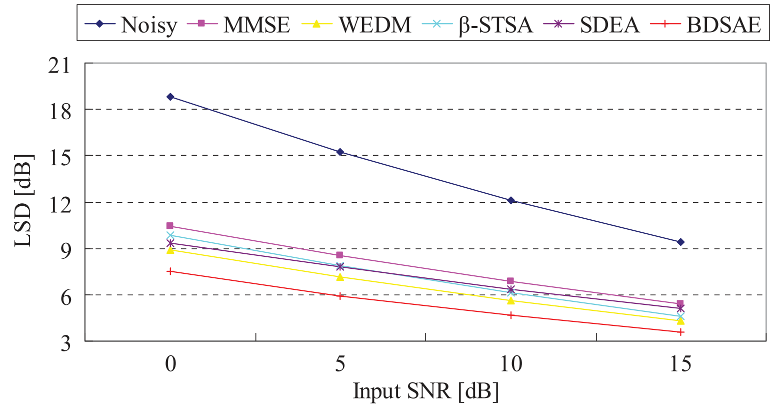

Fig. 10 Comparison of LSD under factory noise

Table 2 Test results of LSD

Enhancement methods	LSD			
	0 dB	5 dB	10 dB	15 dB
Noisy speech	19.62	16.14	12.95	10.11
MMSE	11.07	8.89	7.04	5.50
WEDM	9.45	7.51	5.84	4.46
β-STSA	10.73	8.52	6.56	4.89
SDEA	9.68	7.90	6.37	5.09
BDSAE	7.49	6.08	4.85	3.86

clean speech from noisy speech, which limits their enhancement performance.

For the proposed BDSAE approach, the sparse characteristics of spectral amplitudes of speech signal are considered. That is, under speech presence or speech absence in frequency bins, the cost functions are different which result in different gain functions. Then the speech detector is derived to choose the optimal gain function to estimate clean speech. In this way, for speech presence or speech absence in frequency bins, the gain functions are different and optimal, respectively, which can yield better speech enhancement performance. Moreover, the speech distortion and residual noise resulted from the detector error (i.e., missed detection and false alarm) can be compensated by cost parameters c_{ij}, which is discussed in Section 2.3 (i.e., (4) *Influence of cost parameters*).

In addition, the p and β parameters are induced to cost functions of the BDSAE approach, and the values of p and β are adaptive calculation as the frequency bins. Therefore, we can obtain more flexible and effective gain functions under speech presence and speech absence in frequency bin, which can yield effective noise reduction and good speech enhancement performance.

As can be seen from (23), the proposed BDSAE approach requires the calculation of two gain functions, G_0 and G_1, and the decision rule, in which the mainly computational complexity is focus on calculating the gamma function $\Gamma(\cdot)$ and the confluent hyper-geometric function $\Phi(\cdot)$. However, for the four reference methods (i.e., MMSE, WEDM, β-STSA, and SDEA) listed in Section 5.1, they also require to calculate the two functions of $\Gamma(\cdot)$ and $\Phi(\cdot)$. Therefore, the computational complexity of the proposed BDSAE approach is at the same level compared to the four reference methods. In addition, the proposed BDSAE approach is implemented frame by frame, and thus, there is no any delay existed.

To implement the proposed BDSAE method for real-time realization, the computational complexity involved in (23) could be further simplified. Here, we apply the idea of looking up a table [14, 16] for simplifying the gain function $G_j(\xi, \gamma, p, \beta)$ of (23). For the numerator and the denominator of (23), the algebraic product of the gamma function $\Gamma(.)$ and the confluent hyper-geometric function $\Phi(.)$ can be considered as the function of variables ϕ and ν, namely, $\Psi(\phi, \nu) = \Gamma(\phi + 1)\Phi(-\phi, 1; \nu)$. The variable ϕ is the function of parameters p and β in the BDSAE estimator, i.e., $\phi_1 = (p + \beta)/2$, $\phi_2 = p/2$. In this way, the gain function $G_j(\xi, \gamma, p, \beta)$ of Eq. (23) can be simplified as follows:

$$G_j(\zeta, \gamma, p, \beta) = \left(\frac{c_{1j}\Lambda(Y(\omega))\left(\sqrt{v}/_{\gamma}\right)^{\beta}\Psi(\phi_1, -v) + c_{0j}G_f^{\beta}}{c_{1j}\Lambda(Y(\omega))\Psi(\phi_2, -v) + c_{0j}} \right)^{1/_{\beta}}$$

Therefore, according to [14] and [16], the $\Psi(\phi, \nu)$ is designed for looking up a table which relies on variables ϕ and ν. The computational complexity of the proposed method is reduced greatly by the above simplification.

6 Conclusions

We present a single-channel speech enhancement method based on BDSAE. The optimal speech decision rule and

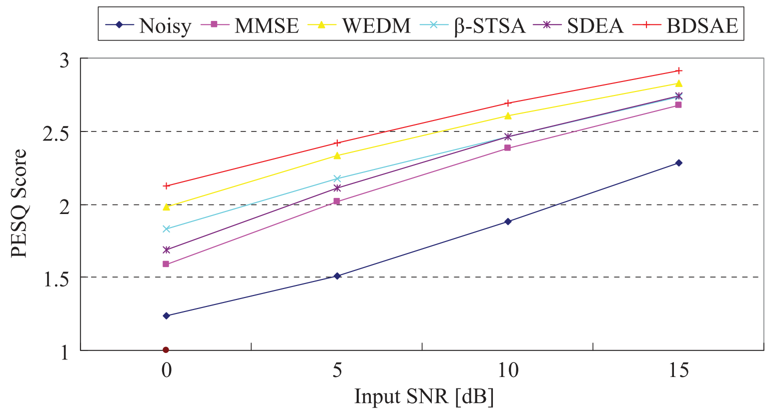

Fig. 11 Comparison of PESQ under White Gaussian noise

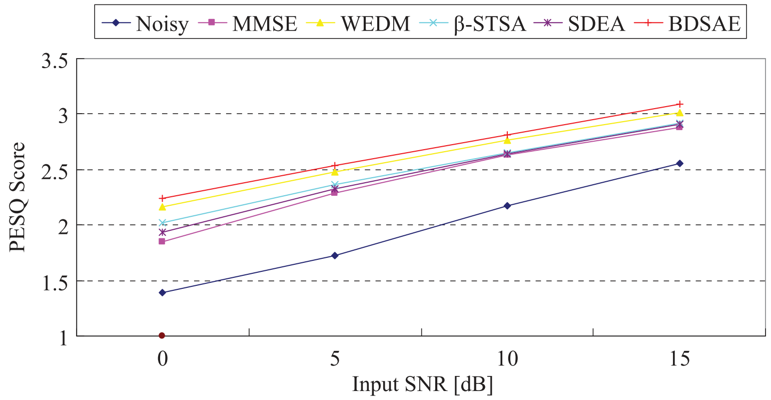

Fig. 12 Comparison of PESQ under factory noise

spectral amplitude estimator are derived by jointly minimizing the combined Bayesian risk function which considers both detection and estimation errors. Under presence and absence of spectral amplitude, the general weighted cost function is proposed, in which the perceptually weighted order p and the spectral amplitude order β are jointly used. In order to obtain flexible gain values for the BDSAE method, the adaptive estimation methods for the p and β parameters are presented, respectively. Furthermore, the cost parameters in the cost function are employed to balance the speech distortion and residual noise caused by missed detection and false alarm, respectively. Therefore, the BDSAE method not only considers the sparse characteristics of the spectral amplitudes of speech signal but also takes the full advantages of both the traditional perceptual weighted estimators and β-order spectral amplitude estimators, which can obtain more flexible and effective gain functions. Finally, we took the objective and subjective quality tests for the enhanced speech based on SNR_{seg}, LSD, PESQ, and MUSHRA listening tests, respectively. The test results indicate that the proposed BDSAE method can achieve a more significant performance improvement than the reference methods.

Table 3 Test results of PESQ

Enhancement methods	PESQ scores			
	0 dB	5 dB	10 dB	15 dB
Noisy speech	1.615	1.945	2.316	2.686
MMSE	2.013	2.403	2.716	2.973
WEDM	2.239	2.558	2.828	3.089
β-STSA	2.160	2.482	2.774	3.042
SDEA	2.105	2.451	2.752	3.025
BDSAE	2.340	2.651	2.926	3.189

7 Appendix 1—the derivation procedure of $r_{1j}(Y(\omega_k))$ of Eq. (24)

In this appendix, we derive the $r_{1j}(Y(\omega_k))$ of Eq. (24) and ignore the frequency bin k for notation simplification. Under speech hypothesis H_1, by substituting speech presence cost function $d_{1j}(X, X)$ into $r_{1j}(Y(\omega))$ of (9), we can obtain

$$r_{1j}(Y(\omega)) = c_{1j}\int\limits_{\Omega_x} X^p\left(X^\beta - \left(G_jY\right)^\beta\right)^2 p(X|H_1)p(Y(\omega)|X)dX \tag{37}$$

where we just call $G_j(\xi, \gamma, p, \beta)$ as G_j for convenience and $\hat{X} = G_jY$.

According to [28], we can get the multiplication of the two probability density functions as follows:

$$p(X|H_1)p(Y(\omega)|X) = \int\limits_0^{2\pi} p(x,\theta|H_1)p(Y(\omega)|x,\theta)d\theta \tag{38}$$

where x is the implementation of amplitude variable X and θ is the implementation of phase variable of $X(\omega)$. In this way, the (37) can be rewritten as follows:

$$r_{1j}(Y(\omega)) = c_{1j}\int\limits_0^\infty\int\limits_0^{2\pi} x^p\left(x^\beta - \left(G_jY\right)^\beta\right)^2 p(x,\theta|H_1)p(Y(\omega)|x,\theta)d\theta dx \tag{39}$$

where the probability density functions of (39) can be defined as follows [28]:

$$p(x,\theta|H_1) = \frac{x}{\pi\lambda_x}\exp\left(-\frac{x^2}{\lambda_x}\right) \tag{40}$$

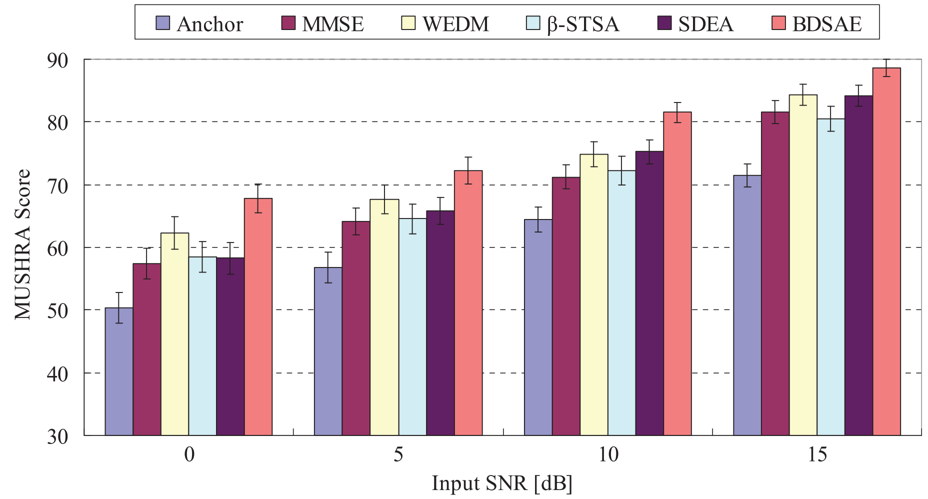

Fig. 13 Comparison of the MUSHRA score

$$p(Y(\omega)|x,\theta) = \frac{1}{\pi\lambda_d}\exp\left(-\frac{|Y(\omega)-X(\omega)|^2}{\lambda_d}\right) \tag{41}$$

By applying the two probability density functions of (40) and (41), we can obtain

$$\begin{aligned} &p(x,\theta|H_1)p(Y(\omega)|x,\theta) \\ &= \frac{x}{\pi^2\lambda_x\lambda_d}\exp\left(-\left(\gamma+\frac{x^2}{\lambda_x}-\frac{2Yx\cos\left(\theta-\theta^{'}\right)}{\lambda_d}\right)\right) \end{aligned} \tag{42}$$

According to [11, 29], we have

$$\int_0^{2\pi}\exp\left(\frac{2Yx\cos\left(\theta-\theta^{'}\right)}{\lambda_d}\right)d\theta = 2\pi J_0\left(i\frac{2Y}{\lambda_d}x\right) \tag{43}$$

where $J_0(.)$ denotes the zero-order Bessel function.

By substituting (42) and (43) into (39), we can get

$$\begin{aligned} r_{1j}(Y(\omega)) &= \frac{2c_{1j}\exp(-\gamma)}{\pi\lambda_x\lambda_d}\int_0^\infty x^{p+1}\left(x^\beta-(G_jY)^\beta\right)^2\exp\left(-\frac{x^2}{\lambda_x}\right)J_0\left(i\frac{2Y}{\lambda_d}x\right)dx \\ &= \frac{2c_{1j}\exp(-\gamma)}{\pi\lambda_x\lambda_d}\int_0^\infty\left[x^{2\beta+p+1}+x^{p+1}(G_jY)^{2\beta}-2x^{\beta+p+1}(G_jY)^\beta\right] \\ &\qquad \exp\left(-\frac{x^2}{\lambda_x}\right)J_0\left(i\frac{2Y}{\lambda_d}x\right)dx \end{aligned} \tag{44}$$

where λ_x and λ_d are the speech and noise variances and γ is a *posteriori* SNR.

Simplifying (44), we obtain

$$\begin{aligned} r_{1j}(Y(\omega)) &= \frac{2c_{1j}\exp(-\gamma)}{\pi\lambda_x\lambda_d} \\ &\cdot\left[\begin{aligned} &\int_0^\infty x^{2\beta+p+1}\exp\left(-\frac{x^2}{\lambda_x}\right)J_0\left(i\frac{2Y}{\lambda_d}x\right)dx \\ &+(G_jY)^{2\beta}\int_0^\infty x^{p+1}\exp\left(-\frac{x^2}{\lambda_x}\right)J_0\left(i\frac{2Y}{\lambda_d}x\right)dx \\ &-2(G_jY)^{\beta}\int_0^\infty x^{\beta+p+1}\exp\left(-\frac{x^2}{\lambda_x}\right)J_0\left(i\frac{2Y}{\lambda_d}x\right)dx \end{aligned}\right] \end{aligned} \tag{45}$$

Following ([29], eq. 6.631.1), we have

$$\begin{aligned} &\int_0^\infty x^\mu\exp(-ax^2)J_\nu(bx)dx \\ &= \frac{b^\nu\Gamma(0.5\nu+0.5\mu+0.5)}{2^{\nu+1}a^{0.5(\nu+\mu+1)}\Gamma(\nu+1)}\Phi\left(\frac{\mu+\nu+1}{2},\nu+1;-\frac{b^2}{4a}\right) \end{aligned} \tag{46}$$

By substituting (46) into (45), we can obtain (47) which is the same with (24).

$$\begin{aligned} &r_{1j}(Y(\omega)) \\ &= \frac{c_{1j}\exp\left(-\frac{\gamma}{1+\xi}\right)}{\pi\lambda_d(1+\xi)} \\ &\cdot\left[\begin{aligned} &\phi^{(p/2+\beta)}\Gamma\left(\frac{p}{2}+\beta+1\right)\Phi\left(-\left(\frac{p}{2}+\beta\right),1;-\nu\right) \\ &+(G_jY)^{2\beta}\phi^{(p/2)}\Gamma\left(\frac{p}{2}+1\right)\Phi\left(-\frac{p}{2},1;-\nu\right) \\ &-2(G_jY)^{\beta}\phi^{(p/2+\beta/2)}\Gamma\left(\frac{p+\beta}{2}+1\right)\Phi\left(-\frac{p+\beta}{2},1;-\nu\right) \end{aligned}\right] \end{aligned} \tag{47}$$

8 Appendix 2—the derivation procedure of $r_{0j}(Y(\omega))$ of Eq. (25)

In this appendix, we derive the $r_{0j}(Y(\omega))$ of Eq. (25) and call $G_j(\xi,\gamma,p,\beta)$ as G_j for convenience. Under hypothesis

H_0, by substituting speech absence cost function $d_{0j}(X, X)$ into $r_{0j}(Y(\omega))$ of (9), we can obtain

$$r_{0j}(Y(\omega)) = c_{0j}\int_0^\infty \left((G_fY)^\beta - (G_jY)^\beta\right)^2 p(X|H_0)p(Y(\omega)|X)dX \tag{48}$$

Following (7), we have $p(X|H_0) = \delta(X)$. Then the Dirac delta function is substituted into (48), we can obtain

$$r_{0j}(Y(\omega)) = c_{0j}\left((G_fY)^\beta - (G_jY)^\beta\right)^2 p(Y(\omega)|H_0) \tag{49}$$

According to [28], the $p(Y(\omega)|H_0)$ of (49) can be defined as:

$$p(Y(\omega)|H_0) = \frac{1}{\pi\lambda_d}\exp\left(-\frac{Y^2}{\lambda_d}\right) \tag{50}$$

where λ_d denotes the variance of noise signal.

By substituting (50) into (49), we can obtain (51) which is (25).

$$\begin{aligned} r_{0j}(Y(\omega)) &= c_{0j}\left((G_fY)^\beta - (G_jY)^\beta\right)^2 \frac{1}{\pi\lambda_d}\exp\left(-\frac{Y^2}{\lambda_d}\right) \\ &= \frac{c_{0j}}{\pi\lambda_d}\left(G_f^\beta - G_j^\beta\right)^2 Y^{2\beta}\exp(-\gamma) \end{aligned} \tag{51}$$

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61471014).

Received: 25 May 2015 Accepted: 28 September 2015
Published online: 07 October 2015

References

1. SF Boll, Suppression of acoustic noise in speech using spectral subtraction [J]. IEEE Trans. Acoust., Speech Signal Process **27**(2), 113–120 (1979)
2. DL Donoho, De-noising by soft-thresholding [J]. IEEE Trans. Inf. Theory **41**, 613–627 (1995)
3. Y Ephraim, HL Van Trees, A signal subspace approach for speech enhancement [J]. IEEE Trans. Speech Audio Process. **3**(4), 251–266 (1995)
4. N Virag, Single channel speech enhancement based on masking properties of the human auditory system [J]. IEEE Trans. Speech Audio Process. **7**(2), 126–137 (1999)
5. Y Ephraim, D Malah, Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator [J]. IEEE Trans. Acoust. Speech Signal Process **32**, 1109–1121 (1984)
6. Y Ephraim, D Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator [J]. IEEE Trans. Acoust. Speech Signal Process **33**, 443–445 (1985)
7. Malah, D., Cox, R. V., Accardi, A. J., Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments [A]. In Proc. 24th IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP'99 [C], Phoenix, AZ, 1999: 789–792.
8. I Cohen, B Berdugo, Speech enhancement for non-stationary environments [J]. Signal Process. **81**, 2403–2418 (2001)
9. M Fujimoto, S Watanabe, T Nakatani, Frame-wise model re-estimation method based on Gaussian pruning with weight normalization for noise robust voice activity detection [J]. Speech Comm. **54**(2), 229–244 (2012)
10. CH Hsieh, TY Feng, PC Huan, Energy-based VAD with grey magnitude spectral subtraction [J]. Speech Comm. **51**(9), 810–819 (2009)
11. A Abramson, I Cohen, Simultaneous detection and estimation approach for speech enhancement [J]. IEEE Trans. Speech Audio Process. **15**, 2348–2359 (2007)
12. P Loizou, Speech enhancement based on perceptually motivated Bayesian estimators of the speech magnitude spectrum [J]. IEEE Trans. Speech Audio Process. **13**(5), 857–869 (2005)
13. Deng, F., Bao, C. C., Bao, F., A speech enhancement method by coupling speech detection and spectral amplitude estimation [A]. 14th Annual Conference of the International Speech Communication Association, Interspeech [C], Lyon, France, 2013: 3234–3238.
14. F Deng, F Bao, CC Bao, Speech enhancement using generalized weighted β-order spectral amplitude estimator [J]. Speech Comm. **59**, 55–68 (2014)
15. CH You, SN Koh, S Rahardja, Masking-based β-order MMSE speech enhancement [J]. Speech Comm. **48**(1), 57–70 (2006)
16. CH You, SN Koh, S Rahardja, β-order MMSE spectral amplitude estimation for speech enhancement [J]. IEEE Trans. Speech Audio Process. **13**(4), 475–486 (2005)
17. A Fredriksen, D Middleton, D Vandelinde, Simultaneous signal detection and estimation under multiple hypotheses [J]. IEEE Trans. Inf. Theory **18**(5), 607–614 (1968)
18. AG Jaffer, SC Gupta, Coupled detection-estimation of Gaussian processes in Gaussian noise [J]. IEEE Trans. Inf. Theory **18**(1), 106–110 (1972)
19. E Plourde, B Champagne, Auditory-based spectral amplitude estimators for speech enhancement [J]. IEEE Trans. Speech Audio Process. **16**(8), 1614–1623 (2008)
20. JD Johnston, Transform coding of audio signal using perceptual noise criteria [J]. IEEE J. Select. Areas Comm. **6**, 314–323 (1988)
21. Fu Z. H., Wang J. F., Speech presence probability estimation based on integrated time-frequency minimum tracking for speech enhancement in adverse environments [A]. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP [C], Dallas, Texas, USA, 2010: 4258–4261.
22. I Cohen, B Berdugo, Noise estimation by minima controlled recursive averaging for robust speech enhancement [J]. IEEE Signal Processing Letters **9**(1), 12–15 (2002)
23. A Varga, H Steeneken, Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems [J]. Speech Comm. **12**(3), 247–251 (1993)
24. ITU-T, Recommendation P.56: Objective measurement of active speech level, 1993.
25. SR Quackenbush, TP Barnwell, MA Clements, *Objective measures of speech quality [J]* (Englewood Cliffs, NJ, Prentice Hall, 1988)
26. ITU-T, Recommendation P.862: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, 2001.
27. E. Vincent, 2005. MUSHRAM: a MATLAB interface for MUSHRA listening tests. [Online] http://www.elec.qmul.ac.uk/people/emmanuelv/mushram.
28. Loizou PC, Speech enhancement: theory and practice [M]. Boca Raton: CRC Press; 2007. 213–225.
29. IS Gradshteyn, IM Ryzhik, *Table of integrals, series, and products [M]*, 6th edn. (Academic, New York, 2000)