Signal Processing 88 (2008) 2764–2776

Contents lists available at ScienceDirect

Signal Processing

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

Adaptive β-order generalized spectral subtraction for speech enhancement

Junfeng Li [a,*,1], Shuichi Sakamoto [a], Satoshi Hongo [b], Masato Akagi [c], Yôiti Suzuki [a]

[a] Research Institute of Electrical Communication, Tohoku University, 2-1-1 Katahira, Sendai, Japan
[b] Department of Design and Computer Application, Miyagi National College of Technology, 48 Nodayama, Natori, Japan
[c] School of Information Science, JAIST, 1-1 Asahidai, Nomi, Japan

ARTICLE INFO

Article history:
Received 10 November 2007
Received in revised form
9 April 2008
Accepted 3 June 2008
Available online 12 June 2008

Keywords:
Generalized spectral subtraction
Adaptive spectral order
Sigmoid function
Data-driven optimization
Speech enhancement

ABSTRACT

The performance degradation of speech communication systems in noisy environments inspired increasing research on speech enhancement and noise reduction. As a well-known single-channel noise reduction technique, spectral subtraction (SS) has widely been used for speech enhancement. However, the spectral order β set in SS is always fixed to some constants, resulting in performance limitation to a certain degree. In this paper, we first analyze the performance of the β-order generalized spectral subtraction (GSS) in terms of the gain function to highlight its dependence on the value of spectral order β. A data-driven optimization scheme is then introduced to quantitatively determine the change of β with the change of the input signal-to-noise ratio (SNR). Based on the analysis results and considering the non-uniform effect of real-world noise on speech signal, we propose an adaptive β-order GSS in which the spectral order β is adaptively updated according to the local SNR in each critical band frame by frame as in a sigmoid function. The performance of the proposed adaptive β-order GSS is finally evaluated objectively by segmental SNR (SEGSNR) and log-spectral distance (LSD), and subjectively by spectrograms and mean opinion score (MOS), using comprehensive experiments in various noise conditions. Experimental results show that the proposed algorithm yields an average SEGSNR increase of 2.99 dB and an average LSD reduction of 2.71 dB, which are much larger improvement than that obtained with the competing SS algorithms. The superiority of the proposed algorithm is also demonstrated by the highest MOS ratings obtained from the listening tests.

1. Introduction

Acoustic background noises present in daily-life environments significantly degrade the performance of many speech applications, such as speech communication systems and sound-based human–machine interaction. To solve these problems and further improve the performance of these applications in adverse environments, it is therefore essential to apply effective speech enhancement algorithms as a front-end processor [1].

A variety of speech enhancement algorithms have been reported in the literature [1]. Among them, single-channel techniques play a crucial role, since the number of microphones is limited on account of some practical requirements. As a well-known single-channel speech enhancement method, *spectral subtraction* (SS) was first proposed more than 20 years ago and has been widely used due to its simplicity in implementation and its effectiveness in reducing additive noise [2]. The basic concept underlying SS is to subtract the

* Corresponding author. Tel.: +81 761 51 1228; fax: +81 761 51 1149.
E-mail address: junfeng@jaist.ac.jp (J. Li).
[1] Is currently with the Graduate School of Information Science, Japan Advanced Institute of Science and Technology, 1-1, Asahidai, Nomi, Ishikawa 923-1292, Japan.

0165-1684/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2008.06.005

estimated spectrum of the noise signal, which is originally calculated in non-speech periods, from that of the noisy signal [2].

SS has recently been modified and improved to overcome its shortcomings (e.g., low noise reduction performance and "musical" noises) in different ways [2–8]. Boll [2] applied several secondary procedures to the processed signals after SS to further attenuate residual noise. To enhance noise reduction ability and mitigate the "musical" noise, Berouti et al. [3] introduced two additional parameters, an oversubtraction factor that controls the amount of noise to be subtracted and a spectral flooring factor that mitigates the "musical" noise. Furthermore, Schless et al. suggested to set both the oversubtraction factor and the spectral flooring factor based on the current *signal-to-noise ratio* (SNR), achieving higher noise reduction performance [4]. Evans et al. [9] gave a possible explanation for the higher noise reduction performance and the lower "musical" noise achieved with that method [4] in the context of automatic speech recognition. Kamath et al. proposed to empirically set different oversubtraction factors in different subbands, resulting in improved speech quality and largely reduced "musical" noise in colored noise conditions [5]. Moreover, the performance of SS was also enhanced by continuously updating noise estimates even in speech-presence periods, by using a minimum statistic noise estimation technique [10,11] or a quantile-based noise estimation technique [12], instead of noise estimation in non-speech periods only. More recently, Sim et al. derived a short-time spectral amplitude (STSA) estimator of the speech signal based on a parametric formulation of the *generalized spectral subtraction* (GSS) by minimizing the *mean-square error* (MSE) between the speech spectrum and its estimate [6].

Following the work presented in [6], the original SS and its modifications can be summarily represented as $|\hat{S}|^\beta = a|X|^\beta - b|\hat{N}|^\beta$, where $\hat{S}$, X and $\hat{N}$ are the *short-time Fourier transforms* (STFTs) of the clean speech estimate, the noisy signal and of the noise estimate; a and b are the parameters, $|\cdot|$ denotes the module operator, and β is the spectral order on which we focus in this research. In the traditional SS algorithms mentioned above, the spectral order β is always fixed to some constant. For example, $\beta = 1.0$ corresponds to the amplitude SS [2,7] and $\beta = 2.0$ corresponds to the power SS [3,5,10–12]. Constant values of β involve a low computational cost and result in a certain degree of noise reduction. Interestingly, the results of two researches [6,8] indicate that the SS method using β with a relatively small value (e.g., $\beta = 1.0$ or 0.5) reduces a large degree of noise components, leading to a "cleaner" processed signal in conditions of low SNR. For high SNRs, a large value of β (e.g., $\beta = 2.0$) is preferred to preserve the speech components. Therefore, the appropriate value of β is dependent on the noise conditions that are considered (i.e., the SNR). You et al. suggested to determine a constant value for the spectral order β in each frame according to the frame SNR as in a linear function under the STSA estimator scenario [13,14]. Obviously, this algorithm assumes a uniform effect of noise on speech signals in the frequency domain and a linear dependence of the β value on the local input SNR conditions, which are not satisfied in practical environments.

Overall, the fixed β values set in the traditional algorithms imply that the noise signal affects the desired signal to a constant SNR degree, which is not reasonable in relation to real-world noise signals and limits the performance improvement of the SS algorithms. In practical environments, however, the desired signal is always contaminated by noise in a time-varying frequency-dependent way, resulting in varying SNRs in the time–frequency domain. Therefore, it is believed that adaptively assigning different appropriate values to the spectral order β according to the current noise conditions (e.g., SNR) is useful in improving the performance of SS algorithms.

To overcome the drawbacks of the traditional SS methods, in this paper, we investigate the performance dependence of the β-order GSS on the value of spectral order β, describe this dependence quantitatively through a data-driven optimization procedure, and further propose an adaptive β-order GSS for speech enhancement. The characteristics of the β-order GSS are analyzed in terms of the gain function, with a special focus on the impact of the β value on the noise reduction performance of the β-order GSS. Results of the analysis indicate that the value of β should be increased as the input SNR increases to preserve speech components, and should be decreased as the input SNR decreases to enhance noise reduction performance. The change tendency of β with the local input SNR is quantitatively derived through a data-driven optimization procedure, which shows that the optimized β varies with the local SNR in a way that can be approximated by a sigmoid function. Moreover, considering the non-uniform effect of real-world noise on speech signal in the time–frequency domain, we propose to determine the value of the spectral order β adaptively according to the local input SNR in each subband frame by frame as in the sigmoid function. Experimental results in various noise conditions demonstrate that the proposed method outperforms the traditional SS algorithms in terms of both objective and subjective evaluation measures.

The remainder of this paper is organized as follows. Section 2 formulates the problem to solve and describes the β-order GSS. Section 3 discusses the characteristics of the β-order GSS to highlight the impact of the spectral order β on the noise reduction performance of the β-order GSS. In Section 4, the non-uniform effect of real-world noise on speech signal in the time–frequency domain is discussed, a data-driven optimization procedure is then introduced to derive the quantitative change tendency of the value of β with the local input SNR, and an adaptive β-order GSS is finally proposed in which the spectral order β is adaptively adjusted according to the local input SNRs as in the sigmoid function. Section 5 details the implementation of the proposed adaptive β-order GSS. Experimental results are provided in Section 6 followed by general discussions on the traditional and proposed SS algorithms in Section 7. Conclusions are finally given in Section 8.

2. β-Order GSS

2.1. Problem formulation

Let $s(t)$ and $n(t)$ denote the speech signal and the uncorrelated additive noise signal, respectively, where t is the discrete-time index. The observed noisy signal $x(t)$ is the sum of the original clean speech signal $s(t)$ and the disturbing noise $n(t)$, given by

$$x(t) = s(t) + n(t). \tag{1}$$

Applying the STFT, the observed signal in the time–frequency domain is represented as

$$X(k,\ell) = S(k,\ell) + N(k,\ell), \tag{2}$$

where k and ℓ are the frequency bin index and the time frame index, respectively; $X(k,\ell)$, $S(k,\ell)$ and $N(k,\ell)$ are STFTs of the corresponding signals.

The β-order GSS method is defined as [6]

$$|\hat{S}_\beta(k,\ell)|^\beta = a_\beta(k,\ell)|X(k,\ell)|^\beta - b_\beta(k,\ell)E[|N(k,\ell)|^\beta], \tag{3}$$

where β is the spectral order, $\hat{S}_\beta(k,\ell)$ denotes the estimate of the speech spectrum using the β-order GSS method, $a_\beta(k,\ell)$ and $b_\beta(k,\ell)$ are the parameters used in the β-order GSS method, and $E[\cdot]$ is the expectation operator. Note that the speech spectrum estimate $\hat{S}_\beta(k,\ell)$ and the parameters $a_\beta(k,\ell)$ and $b_\beta(k,\ell)$ are dependent not only on time and frequency, but also on the spectral order β. The formulation in Eq. (3) describes the original SS method and a number of its modifications. If $\beta = 1.0$, the expression in Eq. (3) represents the amplitude SS [2,7]. If $\beta = 2.0$, the expression represents the power SS [3,5,10–12].

2.2. β-Order GSS

The parameters $a_\beta(k,\ell)$ and $b_\beta(k,\ell)$ in the β-order GSS are determined and optimized by minimizing the MSE $e_\beta(k,\ell)$ between the β-order speech spectrum amplitude $|S(k,\ell)|^\beta$ and its estimate $|\hat{S}_\beta(k,\ell)|^\beta$, that is,

$$(a_\beta(k,\ell), b_\beta(k,\ell)) = \arg\min_{a,\ b} E\{[e_\beta(k,\ell)]^2\}, \tag{4}$$

where

$$e_\beta(k,\ell) = |S(k,\ell)|^\beta - |\hat{S}_\beta(k,\ell)|^\beta. \tag{5}$$

To simplify the derivation procedure, it is further assumed that each individual spectral component of speech and noise signals is a statistically independent complex Gaussian random variable, as considered in [15,16]. Under this complex Gaussian assumption and substituting Eqs. (2) and (3) into Eq. (5), the optimal parameters are then determined by differentiating Eq. (5) with respect to the individual parameters followed by setting the results equal to zero. As a result, the parameters $a_\beta(k,\ell)$ and $b_\beta(k,\ell)$ are derived as [6]

$$a_\beta(k,\ell) = \frac{[\xi_\beta(k,\ell)]^\beta}{1+[\xi_\beta(k,\ell)]^\beta}, \tag{6}$$

$$b_\beta(k,\ell) = \frac{[\xi_\beta(k,\ell)]^\beta}{1+[\xi_\beta(k,\ell)]^\beta}(1-[\xi_\beta(k,\ell)]^{-\beta/2}), \tag{7}$$

where $\xi_\beta(k,\ell)$ is referred to as the *a priori* SNR [15], defined as

$$\xi_\beta(k,\ell) = \frac{E[|\hat{S}_\beta(k,\ell)|^2]}{E[|\hat{N}(k,\ell)|^2]}, \tag{8}$$

where $\hat{N}(k,\ell)$ is the estimate of the noise spectrum calculated in speech pauses in the current implementation. The estimate of the *a priori* SNR, $\xi_\beta(k,\ell)$, is updated in a decision-directed scheme [15] which significantly decreases the residual "musical" noise as detailed in [17]. Substituting the optimal parameters in Eqs. (6) and (7) into Eq. (3), the gain function of the β-order GSS method is finally represented as [6]

$$\begin{aligned}\hat{G}_\beta(k,\ell) &= \frac{|\hat{S}_\beta(k,\ell)|}{|X(k,\ell)|} \\ &= \left\{\frac{[\xi_\beta(k,\ell)]^\beta}{1+[\xi_\beta(k,\ell)]^\beta}\right\}^{1/\beta}\left\{1-(1-[\xi_\beta(k,\ell)]^{-\beta/2})\right. \\ &\quad \left.\times\Gamma\left(\frac{\beta}{2}+1\right)\left(\frac{1}{\gamma(k,\ell)}\right)^{\beta/2}\right\}^{1/\beta},\end{aligned} \tag{9}$$

where $\Gamma(\cdot)$ denotes the Gamma function, and $\gamma(k,\ell)$ is the *a posteriori* SNR [15] defined as

$$\gamma(k,\ell) = \frac{E[|X(k,\ell)|^2]}{E[|\hat{N}(k,\ell)|^2]}. \tag{10}$$

Note that in addition to the *a priori* SNR $\xi_\beta(k,\ell)$ and the *a posteriori* SNR $\gamma(k,\ell)$ on which the traditional SS algorithms are always dependent, the gain function of the β-order GSS given in Eq. (9) is further a function of the spectral order β. Moreover, in practical implementation, in order to avoid severe speech distortion, the gain function is usually restricted by a threshold with low value (e.g., $G_{\min} = 0.01$).

3. Analysis of β-order GSS

In this section, we analyze the performance of the β-order GSS in terms of the gain function that is defined in Eq. (9) to highlight its dependence on the value of the spectral order β. For clarity, the frequency bin index k and the frame index ℓ are omitted in this section.

In order to make the following discussion more easily understandable, it should first be noted a the low gain function normally leads to a high noise reduction and concomitantly a high speech distortion; on the other hand, a high gain function generally yields a low speech distortion and a low noise reduction. Concerning the analysis of the β-order GSS, Fig. 1(a) plots the gain of the β-order GSS versus the *a priori* SNR ξ for some typical values of β when $\gamma = 10$ dB; Fig. 1(b) plots the gain of the β-order GSS versus the *a posteriori* SNR γ when $\xi = 10$ dB. Fig. 1 demonstrates that for a fixed β, the gain of the β-order GSS increases as the *a priori* SNR and the *a posteriori* SNR increase, resulting in a decreasing noise reduction. This observation is reasonable because a high degree of noise reduction is only needed in low input SNR conditions. More importantly, for different typical values of β, the different gains of the β-order GSS result in

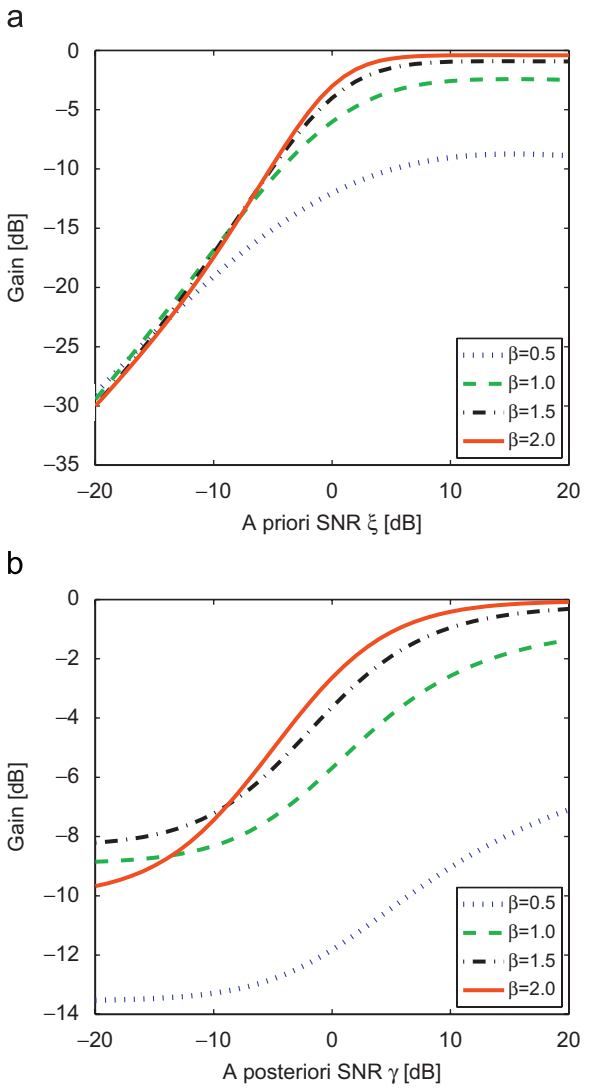

Fig. 1. (a) β-Order GSS gain versus the *a priori* SNR ξ for different values of β when $\gamma = 10$ dB; (b) β-order GSS gain versus the *a posteriori* SNR γ for different values of β when $\xi = 10$ dB.

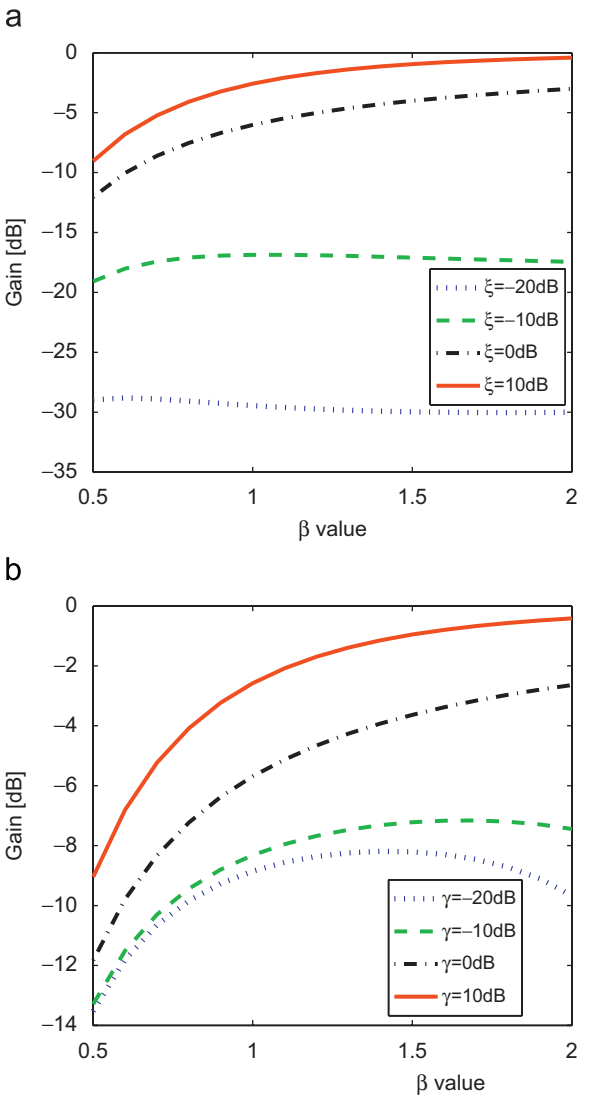

Fig. 2. (a) β-Order GSS gain versus the value of β for different the *a priori* SNRs when $\gamma = 10$ dB; (b) β-order GSS gain versus the value of β for different the *a posteriori* SNRs when $\xi = 10$ dB.

different degrees of noise reduction performance. The dependence of the performance of the β-order GSS on the value of β is further highlighted in Fig. 2, which plots the gain function versus the values of β for different the *a priori* SNRs and for different the *a posteriori* SNRs. Fig. 2 indicates that the gain function of the β-order GSS increases (i.e., the degree of noise reduction decreases) as the value of β increases, especially at relatively high input SNRs. While, in the speech-stop following pause intervals where the *a priori* SNR is relatively high due to the decision-direction estimation technique and the *a posteriori* SNR is extremely low due to the absence of speech signal, the gain function might be of the relatively low gain especially for larger value of spectral order β. Moreover, a low spectral order β generates a low gain function, corresponding to a high noise reduction and a large speech distortion. These analytical results indicate that the noisereduction performance of the β-order GSS is greatly dependent on the value of the spectral order β, and that the β value that offers the best performance in different SNR conditions should also be different (i.e., the β value is dependent on the current SNR). Therefore, it is feasible for the β-order GSS to achieve higher noise reduction with lower speech distortion by adjusting the spectral order β to an appropriate value.

In the traditional SS methods [2–8], however, the spectral order β is always fixed to certain constants (e.g., $\beta = 1.0$ and 2.0) for all frames and all frequency bins. The β value that is computed to give an acceptable noise reduction performance at a low SNR, corresponding to a low gain function, normally introduces severe speech distortion at a high SNR. On the other hand, the β value that is calculated to yield an acceptable performance at a high SNR, corresponding to a high gain function, generally

provides low noise reduction performance at a low SNR. More specifically, if a extremely high β value is used in a speech-absence frame/band, residual noises will be observed due to the use of a large gain function. If an extremely low β value is used in a speech-presence frame/band, part of the speech component will be removed, although the noise components are also attenuated through an extremely low gain function. As a result, the fixed values of β, which are independent of the current noise conditions, result in the performance limitation of the traditional SS methods, especially in time-varying real-world noise conditions.

As discussed above, the value of the spectral order β has a great influence on the performance of the β-order GSS. Considering the fact that the SNRs in practical conditions are time-varying and frequency-dependent (see detailed discussion in Section 4.1), the value of β should be adaptively updated according to the current SNR, finally improving the performance of the β-order GSS. In our proposed adaptive β-order GSS, therefore, we can exploit the relationship between the performance of the β-order GSS and the β value by adjusting β to an appropriate value to attenuate noise components as much as possible while preserving speech components. More specifically, a small value of β is used in a speech-absence frame/band (i.e., a low SNR) to enhance noise reduction through a low gain, and large value of β is utilized in a speech-presence frame/band (i.e., a high SNR) to preserve speech components through a high gain. In this case, though noise components are also less attenuated, they are always masked by the strong concurrent speech components. Therefore, the influence of the β value on the gain of the β-order GSS provides flexibility in suppressing noise components and preserving speech components, by adjusting the spectral order β to an appropriate value according to local SNRs in the time–frequency domain.

4. Adaptive β-order GSS

As discussed in Section 2, the gain function of the β-order GSS in Eq. (9) was derived under the assumption that each spectral component of speech and noise signals follows a statistically independent Gaussian distribution [6]. Though the independent Gaussian distribution is statistically reasonable and widely used due to its simplicity [15], the strong correlation of the spectral components between adjacent frequency bins has recently been taken into account [16,18]. Considering these strong correlations, the appropriate value of β should be determined depending not only on the knowledge of the current frequency bin under consideration but also on that of the neighboring bins. This is believed to greatly improve the robustness and accuracy of the estimated β value.

4.1. Non-uniform effect of noise on speech in the time–frequency domain

In the traditional SS methods, the constant values of the spectral order β imply that the desired speech signal is contaminated by additive noise at a constant SNR across all frequency bins in a time-invariant way. This is not the case, however, with real-world noise (e.g., car noise and babble noise). In real-world environments, noise signals that are mostly time-varying and colored affect target speech signal non-uniformly over the entire signal and the whole spectrum. In the time domain, speech signals are a highly non-stationary signal whose characteristics vary greatly over time, e.g., the speech-presence period is completely different from the speech-absence period. Moreover, characteristics of noise signals also change over time due to the time-varying properties of the noise sources. The time-varying characteristics of speech and noise signals result in the change of the local SNR over time, further indicating that the value of β should vary with time. In the frequency domain, the energy of speech is not uniformly distributed over all frequency bins, e.g., the frequency components corresponding to the formants are generally characterized by a high energy of speech. Moreover, the colorness of real-world noise generally affects a speech signal to a different degree in the different frequency bins. Thus, the frequency-dependent properties of speech and noise signals result in changes of the local SNR with frequency, further indicating that the value of β should vary with frequency. As a result, the value of β should be adaptively updated according to the current local SNR in the time–frequency domain.

These characteristics are highlighted by a typical example shown in Fig. 3, in which the speech signal (“Sekando arubamu o happyou shitekara tuâ ni deru.”) is corrupted by real-world car noise at a global SNR of 10 dB. Fig. 3 demonstrates that the local SNR greatly varies with time due to the time-varying characteristics of the speech and noise signals, and also changes significantly for different subbands because of the colorness of the noise signal and of the non-uniform spectral energy distribution of the speech signal. As a result, the speech signal

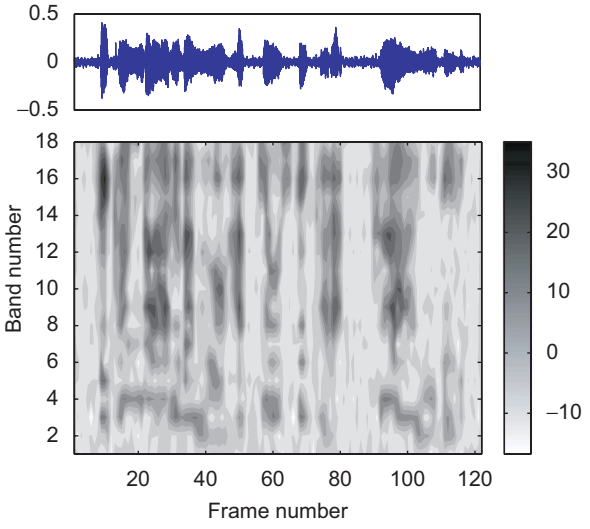

Fig. 3. Local SNRs (dB) in different critical bands and different frames for the utterance (“Sekando arubamu o happyou shitekara tuâ ni deru.”) corrupted by real-world car noise for a global SNR of 10 dB.

corrupted by real-world noise is normally characterized by different local SNRs in different partitions in the time–frequency domain. Accordingly, the appropriate value of the spectral order β must be adaptively determined according to the local SNR in the time–frequency domain.

4.2. Adaptive scheme for an appropriate value of spectral order β

4.2.1. Data-driven optimization of spectral order β

The discussions on the performance dependence of the β-order GSS on the value of β in Section 3 indicate that it is desirable to increase the value of β as the SNR increases to preserve the speech components, and to decrease the value of β as the SNR decreases to enhance the noise reduction performance. After knowing these qualitative results, one further problem that has to be solved is to determine how the appropriate (optimized) value of β quantitatively changes as the local input SNR changes. Since it is difficult to theoretically solve this problem, we turn to model this change tendency through a data-driven optimization approach.

The basic idea of the data-driven optimization that we exploited is to find the appropriate value of the spectral order β that is able to minimize the distance between the spectral amplitude of the clean signal and that of its estimate. In our data-driven optimization procedure, 10 speech sentences were randomly selected from the NTT database [19], and two noise signals (“car” and “babble”) were taken from the NOISEX-92 database [20]. The speech and noise signals were first downsampled to 8 kHz and then mixed with an global SNR ranging from −40 to 40 dB. We assume that the noise spectrum is known a prior in this optimization procedure. For a given value of β, the gain function of the β-order GSS is calculated using Eq. (9), and then used to enhance the target speech signal. Furthermore, considering the mechanism of human perception, we propose to optimize the spectral order β in each critical subband by minimizing the distance between the spectral amplitude $|S(k,\ell)|$ of the clean signal and that of its estimate $|\hat{S}_\beta(k,\ell)|$ in the corresponding subband, that is,

$$\beta_m^{\text{opt}} = \arg\min_{0.1\leqslant\beta\leqslant 3.0}\left(\sum_{k=\omega_m}^{\omega_{m+1}} ||S(k,\ell)| - |\hat{S}_\beta(k,\ell)||\right), \tag{11}$$

where the range of β is empirically confined to $[0.1, 3.0]$, and ω_m denotes the boundary frequency of the m-th critical band. Though only 10 speech sentences are used in the optimization procedure, we should note that the optimization is performed in the following scenarios: in each frames (each speech sentence is divided into 220–350 overlapping frames by windowing before Fourier transform), in each critical subband (e.g., 18 subbands) and at the different global SNR conditions (that is, −40 to 40 dB with the step of 10 dB). It is therefore believed that the optimization of spectral order β is sufficient in the statistical sense, and the parameters obtained from the this optimization procedure might be able to be applied in other different conditions.

Fig. 4 shows the scatter plot of the optimized β value against the local input SNR (defined in Eq. (12) below) and the mean curve, as well as the fitted sigmoid function in the “car” and “babble” noise conditions. The results shown in Fig. 4 indicate that: (1) the optimal β value should increase (decrease) as the local input SNR increases (decreases), which proves the analysis results discussed in Section 3; (2) there exists a strong correlation between the optimal β value and the local input SNR; (3) most importantly, the change tendency of the appropriate value of β with the change of the local input SNR can be approximated by a sigmoid function defined in Eq. (13) below, which quantitatively describes the dependency of the noise reduction performance of the β-order GSS on the spectral order β and motivates us to model this change tendency with a sigmoid function in our proposed algorithm.

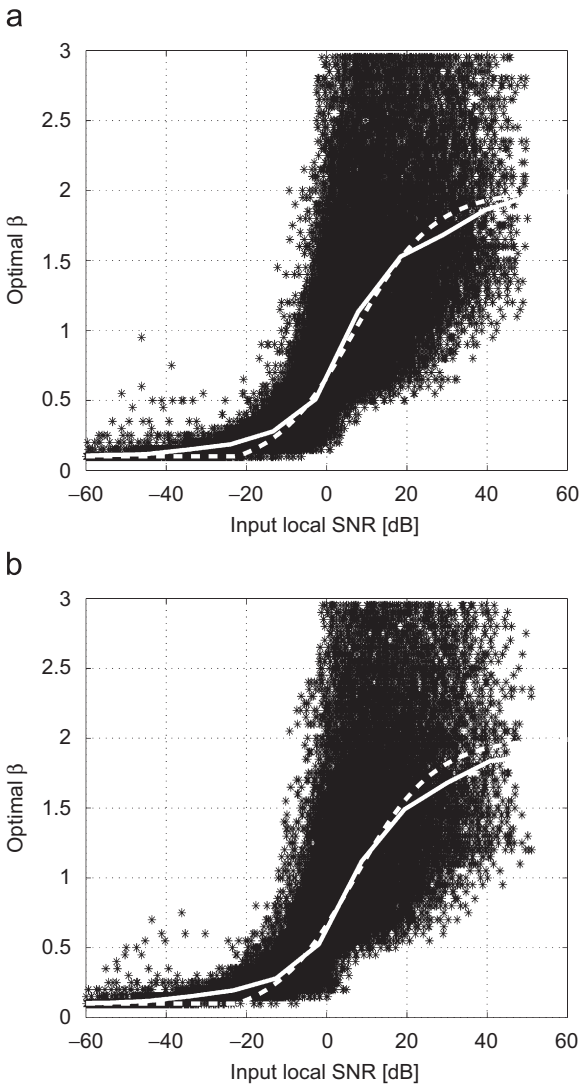

Fig. 4. Scatter plot of the optimized β value with respect to the input local SNR, the mean of the scattered data (solid line) and the fitted sigmoid function (dashed line) with the parameters $A = 0.1$, $B = 2.0$ and $D = 7$, (a) in the car noise condition and (b) in the babble noise condition.

4.2.2. Adaptive determination of spectral order β

Considering the strong correlation of spectral components between adjacent frequency bins [16,18], we propose to update the value of β according to the local SNR in each subband instead of the instantaneous SNR in each individual frequency bin. This estimation scheme greatly improves the robustness and accuracy of the β estimate due to the use of the information of the neighboring frequencies. Considering the mechanism of human perception, the whole spectrum is first divided into subbands according to the critical-band scale [21]. Then, the local SNR $\rho(m,\ell)$ in the m-th critical band and the ℓ-th frame is calculated as

$$\rho(m,\ell)=10\log_{10}\left(\frac{\sum_{k=\omega_m}^{\omega_{m+1}}||X(k,\ell)|-|\hat{N}(k,\ell)||^2}{\sum_{k=\omega_m}^{\omega_{m+1}}|\hat{N}(k,\ell)|^2}\right), \tag{12}$$

where ω_m denotes the boundary frequency of the m-th critical band.

As shown in Section 4.2.1, the average change of the optimized β value with the change of the local input SNR is well approximated by a sigmoid function. Therefore, we propose to determine the appropriate estimate of the spectral order $\tilde{\beta}(m,\ell)$ according to the local SNR $\rho(m,\ell)$ in each critical band frame by frame by use of the sigmoid function, given by

$$\tilde{\beta}(m,\ell)=\frac{B}{1+\mathrm{e}^{-A[\rho(m,\ell)-D]}}, \tag{13}$$

where the parameter A controls the changing speed of the value of $\tilde{\beta}(m,\ell)$ with respect to the local SNR $\rho(m,\ell)$, B determines the range of the value of $\tilde{\beta}$, and D denotes the shift along the SNR axis.

In addition to updating the spectral order $\tilde{\beta}(m,\ell)$ according to the local SNR in the time–frequency domain, we limit the value of $\tilde{\beta}(k,\ell)$ to a minimum value $\beta_{\min}$, since an extremely low β value introduces severe speech distortion through an extremely low gain function. As a result, the appropriate value of the spectral order $\hat{\beta}(m,\ell)$ is finally determined as

$$\hat{\beta}(m,\ell)=\max[\tilde{\beta}(m,\ell),\beta_{\min}]. \tag{14}$$

5. Implementation of adaptive β-order GSS

The proposed adaptive β-order GSS consists of three main blocks: spectral analysis/synthesis, adaptive spectral order β estimation and noise suppression, illustrated in Fig. 5.

In the proposed method, the observed noisy signal is first windowed by a half-overlapped hann window of 512 samples, and then spectrally analyzed using the *fast Fourier transform* (FFT). With the 8 kHz sampling rate used in our implementation, this gives a window length of 64 ms.

To adaptively determine the appropriate value of the spectral order β, the frequency components are grouped into 18 subbands with the boundary frequencies calculated in the critical-band scale that has been shown to be advantageous for human perception. The noise spectrum estimation is performed in the speech pauses with the help of a *voice activity detector* (VAD) that is manually done in the current implementation. Based on the local SNR $\rho(m,\ell)$ in Eq. (12), we compute the appropriate spectral order estimate $\hat{\beta}(k,\ell)$ using Eqs. (13) and (14).

To perform noise suppression, the *a priori* SNR and the *a posteriori* SNR are first calculated with Eqs. (8) and (10), respectively. With the newly obtained spectral order estimate $\hat{\beta}(m,\ell)$, the gain function can be computed using Eq. (9) and used to suppress noise signals.

At the last step, the enhanced speech spectrum is generated by combining the enhanced speech amplitude spectrum with the phase of the noisy inputs. The enhanced speech signal is finally synthesized by computing the inverse STFT, and by overlapping and adding two consecutive frames according to the overlap-and-add method.

6. Experiments and results

To validate the usefulness of the proposed adaptive β-order generalized spectral subtraction (PRO-SS), its performance was investigated in various noise conditions and further compared to that of the traditional SS algorithms, including the power SS (POW-SS) by setting $\beta=2.0$ in [6] , the amplitude SS (AMP-SS) by setting $\beta=1.0$ in [6] and the SS algorithm (SR-SS) by setting $\beta=0.5$ in [6]. Note that the traditional SS algorithms were implemented by setting the spectral order β to the fixed values (2.0, 1.0 and 0.5) in the gain function of the GSS defined in Eq. (9). The reasons for the implementation and comparison are: (i) The PRO-SS method is a derivation of the GSS algorithm in [6]. The comparison would highlight the added-value of the PRO-SS algorithm that we propose here. (ii) Since, in our PRO-SS algorithm, the "decision-directed" scheme [15] is used to estimate the *a priori* SNR, our algorithm should be compared to those that also use the "decision-directed" SNR estimation. This comparison will avoid the necessity to measure the contribution of the "decision-directed" SNR estimation, instead of the contribution of the proposed adaptive scheme for the spectral order. The performance was evaluated in terms of both objective and subjective speech quality measures.

6.1. Experimental configuration

We assess the performance of the PRO-SS objectively and subjectively with the following experiments: we randomly selected 40 clean continuous speech sentences produced by two females and two males from the NTT speech database with a sampling rate of 44.1 kHz at 16 bits [19]. Three types of noise sources, "car", "babble (speech-like)" and "train", were chosen from the NOISEX-92 database [20]. The clean speech and noise signals were first downsampled to 8 kHz. The noise signals were then scaled to obtain a prescribed input SNR, before they were added to the clean speech signal. We generated noisy speech signals artificially by adding various noise signals to the clean signals at different SNRs ranging from 0 to 15 dB with a 5-dB step size. Note that the car noise was a stationary signal, whereas the babble and

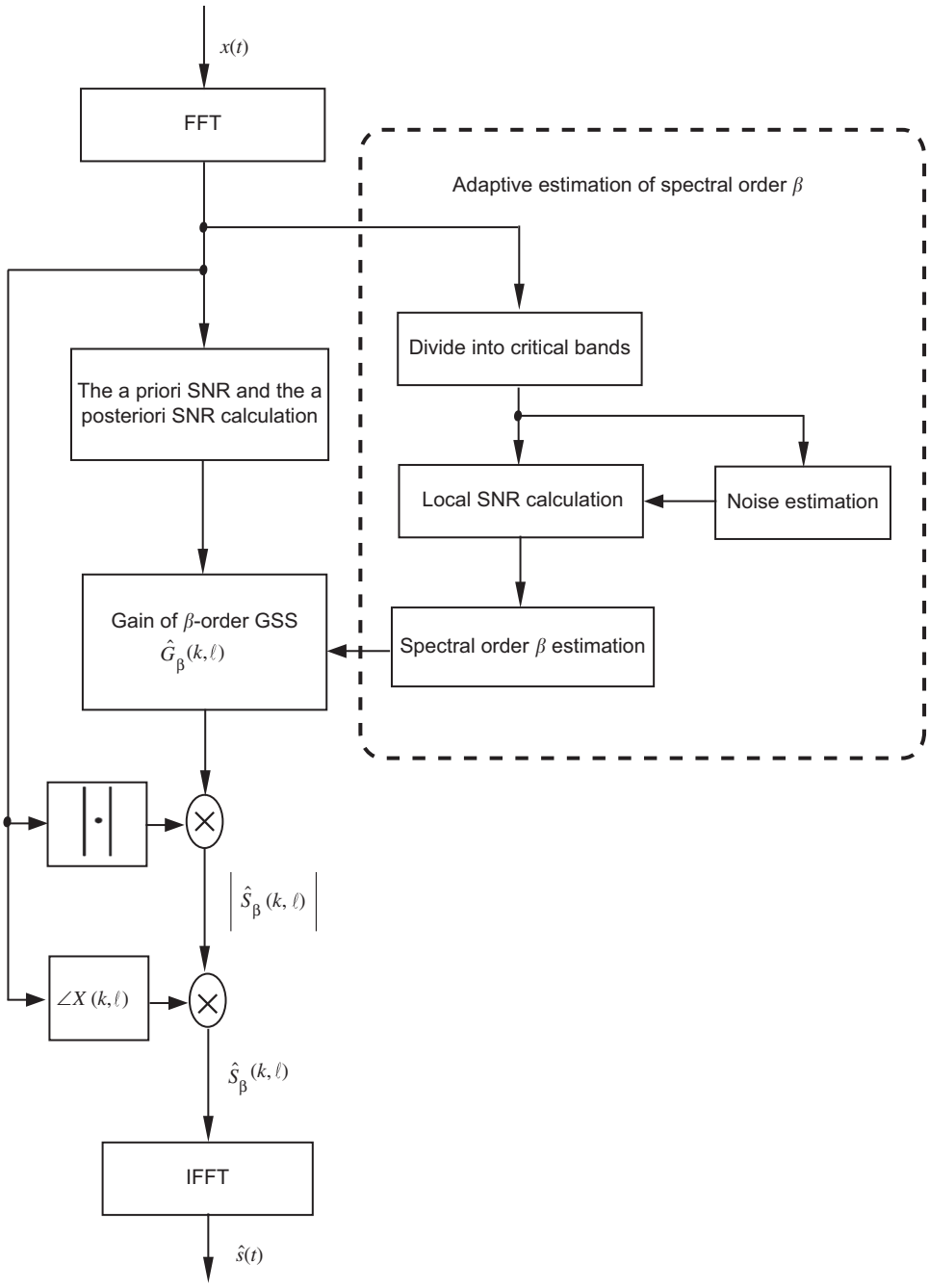

Fig. 5. Block diagram of the proposed adaptive β-order generalized spectral subtraction.

train noises were non-stationary signals with different degrees of non-stationarity. The parameters used in the PRO-SS method which were introduced in Section 4.2.2 and used in Fig. 4 were set as follows: $A = 0.1$, $B = 2.0$, $D = 7$ and $\beta_{\min} = 0.1$; the spectral floor $G_{\min} = 0.01$ corresponding to a maximal noise attenuation of roughly 40 dB; and a commonly used value of 0.98 for the smoothing factor in the decision-directed approach of the *a priori* SNR, as in [15–17].

6.2. Objective evaluation

6.2.1. Evaluation measures

To evaluate the SS methods for speech enhancement, two objective speech quality measures were used: *segmental SNR* (SEGSNR) and *log-spectral distance* (LSD).

The first, SEGSNR, is a widely used objective evaluation measure for speech enhancement algorithms and has been proved to be closely correlated to subjective speech

quality [22]. SEGSNR is defined as the ratio of the power of clean speech to that of noise signal embedded in a noisy signal or an enhanced signal by the studied algorithms, given by [23]

$$\text{SEGSNR} = \frac{10}{L}\sum_{\ell=0}^{L-1}\log_{10}\left(\frac{\sum_{k=0}^{K-1}[s(\ell K+k)]^2}{\sum_{k=0}^{K-1}[\hat{s}(\ell K+k)-s(\ell K+k)]^2}\right), \quad (15)$$

where $s(\cdot)$ is the reference speech signal, $\hat{s}(\cdot)$ is the noisy signal or the enhanced signal processed by the tested algorithms, and L and K represent the number of frames in the signal and the size of STFT, respectively. The SNR in each frame is limited to the perceptually meaningful range between 35 and -10 dB. This prevents the SEGSNR measure from being biased in the frames that do not contribute significantly to overall speech quality [24]. Note that a higher SEGSNR means a higher speech quality of the enhanced signal.

The second measure is LSD, which is often used to assess the distortion of the desired speech signal [25]. LSD is defined as the difference between the log spectrum of the clean speech and that of the noisy signal or the enhanced signal by the studied algorithms, given by [25]

$$\text{LSD} = \frac{10}{L}\sum_{\ell=0}^{L-1}\left(\frac{1}{K}\sum_{k=0}^{K}[10\log_{10}\mathscr{A}S(k,\ell) - 10\log_{10}\mathscr{A}\hat{S}(k,\ell)]^2\right)^{1/2}, \quad (16)$$

where $\mathscr{A}S(k,\ell)\triangleq\max\{|S(k,\ell)|^2,\delta\}$ is the clipped spectral power, such that the log-spectrum dynamic range is confined to about 50 dB (that is, $\delta = 10^{-50/10}\max_{k,\ell}|S(k,\ell)|^2\}$). Note that a lower LSD level indicates less speech distortion.

6.2.2. Evaluation results

The experimental results of SEGSNR and LSD averaged across all sentences in the three noise conditions are summarized in Tables 1 and 2, respectively. Table 1 presents, except for the SR-SS, the three other tested algorithms (i.e., POW-SS, AMP-SS and PRO-SS) all improve the SEGSNR to different degrees in all noise conditions at all SNR levels, especially at the low SNRs. Using the SR-SS method, slight SEGSNR improvements are observed in the low SNR conditions and disappear in the middle and high SNR conditions, resulting in the worst speech enhancement performance among the tested algorithms. This is because the SR-SS reduces the noise signal at the cost of severe target signal distortion, which is demonstrated with waveforms and spectrograms in Section 6.3.1. With regard to the POW-SS and AMP-SS methods, the PRO-SS consistently yields the highest SEGSNR improvements in all conditions. In car noise conditions, for instance, compared with the noisy inputs, the POW-SS and the AMP-SS, the average SEGSNR improvements achieved by our PRO-SS method amount to 3.83, 1.45 and 0.48 dB, respectively. On average, in comparison of noisy inputs, the PRO-SS yields a SEGSNR improvement of about 2.99 dB averaged across all tested conditions, which is much higher than those of traditional algorithms (i.e., 1.77 dB for the POW-SS, 2.62 dB for the AMP-SS, and -0.56 dB for the SR-SS). Therefore, the PRO-SS generates

Table 1
Segmental SNR (dB) of the noisy signal, the traditional power SS output when $\beta = 2.0$ (POW-SS), the traditional amplitude SS output when $\beta = 1.0$ (AMP-SS), the traditional SS output when $\beta = 0.5$ (SR-SS) and of the proposed adaptive β-order GSS (PRO-SS) output, in the car, babble and train noise conditions

Algorithm	Car				Babble				Train			
	0	5	10	15	0	5	10	15	0	5	10	15
Noisy	−3.23	0.01	3.50	7.22	−3.01	0.22	3.71	7.43	−3.06	0.18	3.67	7.40
POW-SS	0.02	2.97	5.77	8.26	−0.77	2.11	4.98	7.55	−0.37	2.31	4.97	7.46
AMP-SS	0.95	3.83	6.77	9.37	0.22	2.94	5.76	8.38	0.63	3.06	5.55	7.98
SR-SS	−0.65	0.73	2.29	3.89	−0.88	0.60	2.19	3.80	−0.78	0.56	2.03	3.56
PRO-SS	**1.33**	**4.49**	**7.39**	**9.62**	**0.72**	**3.36**	**6.03**	**8.48**	**1.05**	**3.42**	**5.85**	**8.19**

Table 2
Log-spectral distance (dB) of the noisy signal, the traditional power SS output when $\beta = 2.0$ (POW-SS), the traditional amplitude SS output when $\beta = 1.0$ (AMP-SS), the traditional SS when $\beta = 0.5$ (SR-SS) and of the proposed adaptive β-order GSS (PRO-SS) output, in the car, babble and train noise conditions

Algorithm	Car				Babble				Train			
	0	5	10	15	0	5	10	15	0	5	10	15
Noisy	8.34	6.00	4.22	2.88	9.03	6.40	4.36	2.86	13.00	9.40	6.37	4.01
POW-SS	5.43	4.02	2.95	2.18	6.44	4.59	3.28	2.35	8.02	5.90	4.22	2.94
AMP-SS	4.73	3.60	2.72	2.11	5.43	4.07	3.03	2.29	6.42	4.96	3.79	2.90
SR-SS	5.75	5.02	4.33	3.77	5.88	5.11	4.42	3.86	6.48	5.68	5.01	4.45
PRO-SS	**4.67**	**3.47**	**2.60**	**2.08**	**5.12**	**3.91**	**2.96**	**2.21**	**5.98**	**4.85**	**3.67**	**2.87**

the output signal with the highest speech quality. This achievement can be attributed to the utilization of the time-varying and frequency-dependent spectral order β in the proposed method. Moreover, from a careful observation of Table 1, we note that the SEGSNR improvements achieved with the PRO-SS method in the car noise conditions are higher than those in the babble and train noise conditions. This is because the noise signal in the car environment is much more stationary than the noise signals in the babble and train noise conditions, and the noise estimation that is performed in non-speech periods in the current implementation is more suitable for stationary noise.

Concerning the results of LSD given in Table 2, we can observe that, except for the SR-SS, the three other tested algorithms lead to a decrease of LSDs in all noise conditions at all SNRs, especially at low SNRs. The SR-SS gives higher LSD results, even compared with the noisy input signals, especially in high SNR conditions, owing to severe speech distortion. With respect to the traditional POW-SS and AMP-SS methods, the proposed PRO-SS consistently yields the lowest LSDs for all conditions and all SNRs. On average, compared to the noisy inputs, the PRO-SS produces a reduction of LSD of about 2.71 dB averaged across all tested conditions, which is much higher than that obtained with the traditional algorithms (i.e., 2.05 dB for the POW-SS, 2.57 dB for the AMP-SS and 1.43 dB for the SR-SS). That is, the enhanced speech signal processed with the adaptive PRO-SS method contains the lowest speech distortion in comparison to those obtained with other traditional SS methods. This achievement can specifically be attributed to the use of high gains in speech-presence periods due to high β values.

6.3. Subjective evaluation

6.3.1. Waveforms and spectrograms

The first subjective evaluation of the studied SS algorithms was performed using waveforms and spectrograms. Typical examples of waveforms and spectrograms, corresponding to the sentence "Sekando arubamu o happyou shitekara tuâ ni deru.", corrupted by car noise at 10 dB, are plotted in Fig. 6. Fig. 6(b) shows that the target speech signal is highly corrupted by the car noise in all frequency bins and all time frames, especially in the low frequencies. Fig. 6(c) shows that the output of the traditional POW-SS (i.e., $\beta = 2.0$) is still characterized by high-level noise. This is because a high gain function is exploited in the POW-SS, due to the use of a relatively high value of β. The noise components are further reduced by using the lower gain function of the traditional AMP-SS (i.e., $\beta = 1.0$) when the β value is low, as shown in Fig. 6(d). The SR-SS greatly reduces the noise components as well as the speech components, leading to severe speech distortion, as shown in Fig. 6(e). In contrast, Fig. 6(f) demonstrates that our adaptive PRO-SS is able to suppress the car noise components with very low speech distortion in the time–frequency domain through the use of the time-varying frequency-dependent spectral order β.

6.3.2. Listening tests

The performance of the adaptive PRO-SS was also evaluated using the listening tests. To reduce the length of the subjective evaluations, only a subset of 12 sample sentences, uttered by two female speakers and two male speakers, were drawn and contaminated by car noise, babble noise and train noise at two different SNR levels, 5 and 10 dB.

The resulting 72 noisy speech sentences were then processed with the four algorithms: POW-SS, AMP-SS, SR-SS and PRO-SS.

The listening tests involved eight volunteers of between 22 and 30 years of age. The test for each listener lasted approximately 2 h, consisting of four 20-min sessions, separated by a short 10-min break. The tested speech materials were randomly presented to each listener through a headphone, and the listeners were free to adjust the sound volume to a comfortable level. The listeners were instructed to rate the quality of the enhanced output signals based on their preference in terms of *mean opinion score* (MOS): 1 = bad, 2 = poor, 3 = fair, 4 = good and 5 = excellent.

The MOS results presented in Table 3 show that among the tested algorithms, the PRO-SS consistently produced the highest MOS ratings in three noise conditions and at two SNR levels. With respect to the POW-SS, the PRO-SS offers on average approximately a one-point improvement, and a 0.5-point improvement compared with the AMP-SS averaged in all tested conditions. The highest MOS result indicates that our adaptive PRO-SS produces the enhanced signal of the highest speech quality, being preferred by the listeners. Moreover, from a careful observation of Table 3, we see that, compared with other traditional SS algorithms, the MOS rates achieved with the PRO-SS are reduced in the babble and train noise conditions. This performance reduction is due to the VAD-based noise estimation approach and the highly non-stationary characteristics of the babble and train noises. These results are consistent with the objective results discussed in the previous section.

The subjective MOS results were also statistically analyzed by performing multiple paired comparisons (Tukey's HSD) between the MOS ratings obtained with speech signals enhanced by use of the three tested SS algorithms. The results are listed in Table 4. In the table, asterisks indicate significant differences between the MOS scores of two enhanced signals processed by the corresponding two algorithms. Table entries denoted as "n.s." stand for a non-significant difference between the MOS results of two enhanced signals.

Table 4 presents that the PRO-SS provides statistically significant improvement of speech quality compared with the POW-SS in all tested noise conditions. Compared to the AMP-SS, significant improvements were observed in a few conditions, the car and train noise conditions at 10 dB. However, when compared to the SR-SS, no statistically

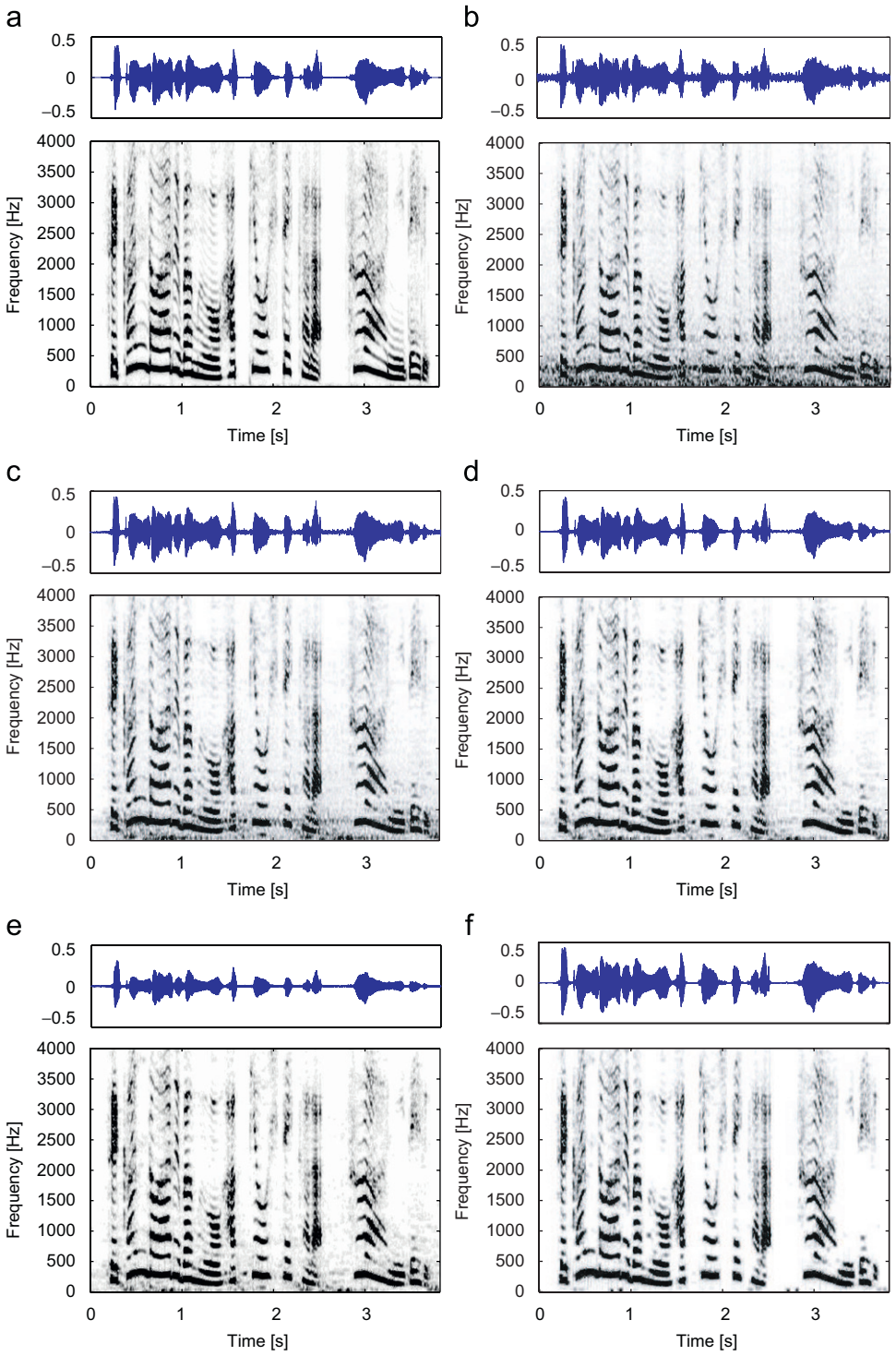

Fig. 6. Speech waveforms and spectrograms. (a) Original clean speech signal: "Sekando arubamu o happyou shitekara tuâ ni deru.". (b) Noisy signal in the car noise condition at the SNR of 10 dB. (c) Enhanced signal processed by the traditional power SS when $\beta = 2.0$ (POW-SS). (d) Enhanced signal processed by the traditional amplitude SS when $\beta = 1.0$ (AMP-SS). (e) Enhanced signal processed by the traditional SS when $\beta = 0.5$ (SR-SS). (f) Enhanced signal processed by the proposed adaptive β-order GSS (PRO-SS).

significant improvement was obtained with our proposed algorithm in all tested conditions. This is possibly because that the SR-SS produces the less musical noise compared to the other standard approaches, and the involved speech distortion was not exhibited through freely controlling the volume.

Table 3
Mean opinion score of the traditional power SS output when $\beta = 2.0$ (POW-SS), the traditional amplitude SS output when $\beta = 1.0$ (AMP-SS), the traditional SS with $\beta = 0.5$ (SR-SS), and of the proposed adaptive β-order GSS (PRO-SS) output, in the car, babble and train noise conditions

Algorithm	Car		Babble		Train	
	5	10	5	10	5	10
POW-SS	2.10	2.44	2.28	2.28	2.37	2.49
AMP-SS	2.71	3.09	2.64	2.89	2.70	2.95
SR-SS	2.65	2.98	2.70	3.00	2.67	3.02
PRO-SS	**3.12**	**3.43**	**2.98**	**3.20**	**3.15**	**3.52**

Table 4
Results obtained from the statistical analysis of MOS ratings for the traditional power SS when $\beta = 2.0$ (POW-SS), the traditional amplitude SS when $\beta = 1.0$ (AMP-SS), the traditional SS with $\beta = 0.5$ (SR-SS) and the proposed adaptive β-order GSS (PRO-SS)

Algorithm pairs	Car		Babble		Train	
	5-dB	10-dB	5-dB	10-dB	5-dB	10-dB
(PRO-SS, POW-SS)	*	*	*	*	*	*
(PRO-SS, AMP-SS)	n.s.	*	n.s.	n.s.	n.s.	*
(PRO-SS, SR-SS)	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.

Table entries denoted with asterisks "*" indicate significant differences ($p = 0.05$) between the MOS ratings of two signals enhanced by the two corresponding algorithms; those denoted with "n.s." indicate non-significant differences between the MOS ratings.

As a result, the proposed adaptive PRO-SS gives a statistically significant performance improvement of speech enhancement in almost all the tested environments compared with the traditional SS algorithms.

7. Discussion

In this section, we present a general discussion on the traditional SS algorithms and the proposed adaptive β-order GSS.

In the traditional SS algorithms, the spectral order β is usually fixed to some constants (e.g., 1.0 and 2.0). The traditional SS algorithms with the constant values of the spectral order β demonstrate a certain degree of noise reduction. One of merits of these algorithms is the low computational cost, i.e., its simplicity in implementation. Although an approach in which the value of β is updated according to the frame SNR in a linear function has recently been reported by You et al. [14], this approach neglects the non-uniform effect of noise on speech signals in the time–frequency domain and the non-linear dependence of the value of β on the local SNR in real conditions.

In the proposed adaptive β-order GSS algorithm, the spectral order β is adaptively determined according to the local SNR in the time–frequency domain, which is motivated by the following facts: the performance of the GSS algorithm is dependent on the value of the spectral order β, as analyzed in Section 3, and the background noise affects the target signal non-uniformly in the time–frequency domain, as discussed in Section 4.1. Compared to the traditional SS algorithms with fixed β values, the proposed adaptive β-order GSS algorithm demonstrates the following beneficial characteristics: (1) the spectral order β is time-varying and frequency-dependent due to the change of acoustic environments; (2) the β value is updated according to the local SNR, which indicates the degree of corruption of target signal by the background noise. Compared with You's algorithm, the proposed adaptive β-order GSS has the following advantages: (1) the non-uniform effect of noise on speech signals in the time–frequency domain is fully taken into account; (2) the average change of the spectral order β with the local SNR is described by a sigmoid function that is derived through a data-driven optimization procedure, instead of using a linear function. The effectiveness and superiorities of the propose adaptive β-order GSS algorithm in reducing noise signal and preserving the target speech signal have been confirmed by comprehensive experiments.

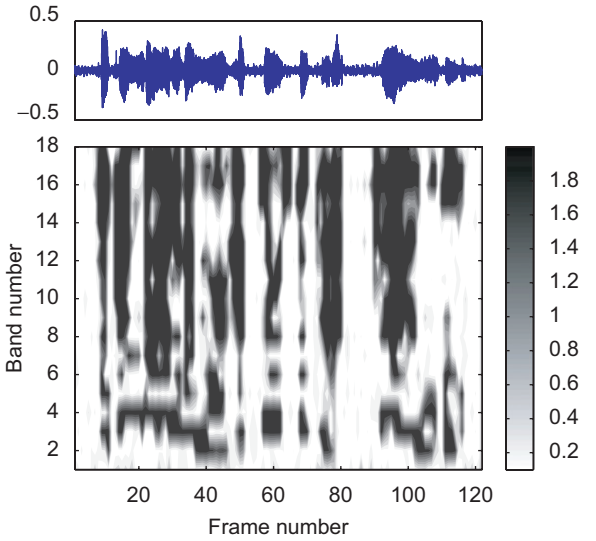

Fig. 7. Adaptation of the spectral order β in different critical bands and different frames for the same utterance as that used in Fig. 3.

To demonstrate the change of the spectral order β with the local input SNR, as an example, the adaptation of the spectral order β for the same utterance as that used in Fig. 3 is exemplified in Fig. 7. From Figs. 3 and 7, it can be seen that with the proposed adaptive β-order GSS, the spectral order β is adaptively adjusted to a small value at low SNR and to a large value at high SNR, according to the local SNRs in the time–frequency domain. The proposed adaptive β-order GSS is superior in suppressing noise components and preserving speech components under real-world noise conditions, because it takes the frequency-dependent and time-varying characteristics of the spectral order β into account.

8. Conclusion

In this paper, we first qualitatively analyzed the performance of the β-order GSS and highlighted its dependence on the value of the spectral order β, which

provides the theoretical principle of this research. The quantitative relationship between the optimal β value and the local SNR was learned through a data-driven optimization procedure. We then proposed an adaptive β-order GSS for speech enhancement in which the value of β is adaptively updated according to the local SNRs in the time–frequency domain as in the sigmoid function. Comprehensive experimental results in various noise conditions show that the proposed adaptive β-order GSS outperforms the traditional SS methods in terms of both objective and subjective speech quality measures.

In the current implementation, the spectrum of the noise signal is estimated in the speech-absence periods with the help of a voice activity detector. Future work on this algorithm will include the integration of advanced noise estimation approaches, such as the minimum statistic tracking noise estimation approach [10] and/or the improved minima controlled recursive averaging approach [26], as well as a further improvement by considering the human auditory-motivated mechanisms. Given the good preliminary performance of the adaptive β-order GSS, future work in this direction is expected to lead to additional promising performance improvements in speech enhancement.

Acknowledgments

This research is supported by the Sendai Intelligent Knowledge Cluster and Grant-in-Aid for Young Scientists (B) (No. 19700156) from the Ministry of Education, Science, Sports and Culture of Japan. The authors would like to thank all the listeners who took part in the listening tests and the anonymous reviewers for their helpful comments in improving the quality of this paper.

References

[1] J. Benesty, S. Makino, J. Chen (Eds.), Speech Enhancement, Springer, Berlin, 2005.
[2] S.F. Boll, Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoust. Speech Signal Process. ASSP-27 (2) (1979) 113–120.
[3] M. Berouti, R. Schwartz, J. Makhoul, Enhancement of speech corrupted by acoustic noise, in: Proceedings of the ICASSP, 1979, pp. 208–211.
[4] V. Schless, F. Class, SNR-dependent flooring and noise overestimation for joint application of spectral subtraction and model combination, in: Proceedings of the International Conference on Spoken Language Processing, 1998, pp. 721–725.
[5] Sunil D. Kamath, Philipos C. Loizou, A multi-band spectral subtraction method for enhancing speech corrupted by colored noise, in: Proceedings of the ICASSP2002, 2002, pp. 4164–4167.
[6] B.L. Sim, Y.C. Tong, J.S. Chang, C.T. Tan, A parametric formulation of the generalized spectral subtraction, IEEE Trans. Speech Audio Process. 6 (4) (1998) 328–337.
[7] P. Sovka, Extended spectral subtraction, in: Proceedings of the European Conference on Signal Processing and Communication, 1996.
[8] J.S. Lim, Evaluation of a correlation subtraction method enhancing speech degraded by additive white noise, IEEE Trans. Acoust. Speech Audio Process. ASSP-26 (5) (1978) 471–472.
[9] N.W.D. Evans, J.S. Mason, An assessment of local nonlinear spectral subtraction for remote speech recognition, in: Proceedings of the 1st Meeting on Speech Technology, Seville, 2000.
[10] R. Martin, Spectral subtraction based on minimum statistics, in: Proceedings of the EUSIPCO94, 1994, pp. 1182–1185.
[11] G. Doblinger, Computationally efficient speech enhancement by spectral minima tracking in subbands, in: Proceedings of the Eurospeech95, 1995, pp. 1513–1516.
[12] V. Stahl, A. Fischer, R. Bippus, Quantile based noise estimation for spectral subtraction and Wiener filtering, in: Proceedings of the ICASSP, 2000, pp. 1875–1878.
[13] C.H. You, S.N. Koh, S. Rahardja, β-Order MMSE spectral amplitude estimation for speech enhancement, IEEE Trans. Speech Audio Process. 13 (4) (2005) 475–486.
[14] C.H. You, S.N. Koh, S. Rahardja, Masking-based β-order speech enhancement, Speech Commun. 48 (1) (2006) 57–70.
[15] Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, IEEE Trans. Acoust. Speech Signal Process. 32 (6) (1984) 1109–1121.
[16] I. Cohen, B. Berdugo, Speech enhancement for non-stationary noise environments, Signal Process. (2001) 2403–2418.
[17] O. Cappe, Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor, IEEE Trans. Acoust. Speech Signal Process. 2 (2) (1994) 345–349.
[18] J. Li, M. Akagi, A noise reduction system based on hybrid noise estimation technique and post-filtering in arbitrary noise environments, Speech Commun. 48 (2006) 111–126.
[19] ⟨http://www.ntt-at.com/products_e/speech2002/⟩.
[20] A. Varga, H.J.M. Steeneken, Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems, Speech Commun. 12 (1993) 247–251.
[21] E. Zwicker, E. Terhardt, Analytical expressions for critical band rate and critical bandwidth as a function of frequency, J. Acoust. Soc. Am. 68 (1980) 1523–1525.
[22] J. Faneuff, D.R. Brown III, Noise reduction and increased VAD accuracy using spectral subtraction, in: Proceedings of the International Signal Processing Conference, 2003, p. 213.
[23] J.H.L. Hansen, B. Pellom, An effective quality evaluation protocal for speech enhancement algorithms, in: Proceedings of the International Conference on Spoken Language Processing, vol. 7, 1998, pp. 2819–2822.
[24] Y. Hu, P. Loizou, Subjective comparison of speech enhancement algorithms, in: ICASSP, 2006, pp. 153–156.
[25] I. Cohen, Multi-channel post-filtering in non-stationary noise environments, IEEE Trans. Signal Process. 52 (5) (2004) 1149–1160.
[26] I. Cohen, Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging, IEEE Trans. Speech Audio Process. 11 (4) (2003) 466–475.