A two-microphone subband noise reduction scheme with a new noise subtraction parameter for speech quality enhancement

T. T. Aung[1], H. Thumchirdchupong[2], N. Tangsangiumvisai[3], A. Nishihara[4]

[1,2,3]*Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University, Phyathai Road, Patumwan, Bangkok 10330, THAILAND*
[1]suuthiri@gmail.com, [2]hathaichanok_thum@hotmail.com, [3]Nisachon.T@chula.ac.th

[4]*The Center for Research and Development of Educational Technology, Tokyo Institute of Technology, Ookayama, Meguro-ku, Tokyo, 152-8552 JAPAN*
[4]aki@cradle.titech.ac.jp

Abstract

An improved subband noise reduction technique is proposed for two-microphone voice communication systems. The technique aims to enhance the speech quality by utilizing a subband structure with different noise reduction schemes for different frequency bands. In the low-frequency band where dominant cues of speech spectral components are usually located and the noise signals from the two channels are mainly correlated, the Spectral Subtraction (SS) method, together with a new variable noise subtraction parameter, is employed so that the noise attenuation performance and speech distortion are controllable. In the high-frequency band where less-dominant frequency information of speech spectrum is located, the Modified Cross-Spectral Subtraction (MCSS) technique is utilized to remove the high-frequency decorrelated noise spectral components. Extensive comparisons among various noise reduction techniques based on computer simulations demonstrate that the proposed two-microphone subband noise reduction scheme achieves excellent noise attenuation performance while preserving the speech quality.

Keywords: noise reduction, subband structure, spectral subtraction, two-microphone, subjective listening test.

1. Introduction

Noise Reduction (NR) techniques have become one of the most crucial functions for modern voice communication systems in order to alleviate the effect of disturbing background noise so that speech intelligibility is improved [1]-[5]. These NR techniques can be categorized into two groups depending on the number of microphones within the systems; i.e. single-microphone and multi-microphone techniques.

One of the single-microphone NR techniques that operate in the frequency-domain is called the Spectral Subtraction (SS) method [1]. The enhanced speech spectrum can be obtained by subtracting the estimated noise spectrum from the noisy speech spectrum. Various methods have also been proposed based on the conventional SS method due to its simplicity, low computational complexity, and high efficiency [3]-[5]. For the conventional SS method, speech distortion is, however, unavoidable. This is mainly due to the fixed choice of the noise subtraction parameter and the frequency characteristics of typical speech signals; i.e. the speech spectral components are located with larger power spectral density (PSD) in the low-frequency region than that in the high-frequency region [6]. As a result, when the noise subtraction parameter is chosen to be large, a large amount of noise reduction will be equally removed from the noisy spectral components at all frequencies. This may cause speech distortion due to the removal of speech spectral components especially the unvoiced speech spectral components that are located in the high-frequency region. On the other hand, by using smaller value of the noise subtraction parameter in order to preserve the speech spectral components, the additive background noise may not be sufficiently eliminated.

Several approaches have been proposed to vary the noise subtraction parameter of the SS-based NR methods, based upon the Signal-to-Noise Ratio (SNR), in order to control the amount of speech distortion while achieving sufficient level of noise reduction [7]-[10]. In [7], when the SNR is high, i.e. during speech presence, the noise subtraction parameter is small to minimize speech distortion. On the other hand, the noise subtraction parameter is large to greatly reduce the noise spectral components when the SNR is low, such as during speech pauses. This concept of variable noise subtraction parameter in [7] has been extended as presented in [8]-[10]. For multi-band SS method, the noise subtraction parameter is adapted by introducing a tweaking factor in order to customizing the noise removal in each frequency band [8]. Phase modification and magnitude compensation are employed for the multi-band SS method in [9]. To provide an additional degree of noise reduction control within each band, the perceptual weighting filter is used for multi-band SS method [10]. However, these methods in [8]-[10] still employ the conventional noise subtraction parameter in [7], which is a linear function of the SNR level. An enhanced variable noise subtraction parameter for the SS-based NR methods is therefore proposed in this paper to further improve the noise reduction performance of the conventional technique in [7] and to control the speech distortion level. In fact, the use of the proposed variable noise subtraction parameter, together with the multi-band structure of the SS methods such as those in [8]-[10], is entirely possible.

Another main reason for the speech distortion problem of the SS-based NR methods is the accuracy of the noise spectrum estimation. At each frequency, if the noise spectrum is overestimated, the speech spectral components will possibly be removed after the operation of noise subtraction of the SS-based NR methods. On the other hand, if the noise spectrum is underestimated, residual noise will occur. These residual spectral noise spikes at random frequencies may cause an annoying artifact known as ***musical noise***, which results in short sinusoids whose frequencies vary from frame to frame [11]. As a consequence, musical noise can be perceived during speech pauses and/or when it is *not* masked by the speech spectral components particularly at low SNR levels [12].

A better estimate of the noise spectrum can be obtained when using multiple microphone observations as in the *multi-microphone* NR techniques, due to the spatial filtering information which helps suppress the interfering signals. These multi-microphone NR techniques operate either in the time domain or in the frequency domain [13]-[27]. The Generalized Sidelobe Canceller (GSC), which is normally used to suppress coherent noise, is employed to estimate the noise spectrum in highly non-stationary multiple-noise-source environments in hearing-aid application [13], [14]. Based upon the Cross-Spectral Subtraction (CSS) method in [15], various approaches have been introduced to estimate the noise cross-PSD for two-microphone systems [16]-[20]. The two-sensor NR technique is proposed for noise reduction in hands-free car kits, utilizing the coherence function between the two noisy microphone signals as a filter in order to remove the decorrelated residual noise and the musical noise effect [16]. A noise cross-PSD estimation method based on a soft decision using minimum statistics is proposed in [17]. The CSS method in [15] is modified based on the a priori SNR and Wiener filter to control the amount of noise reduction [18], [19]. Another noise cross-PSD estimation approach using phase information is given in [20]. For noise reduction in reverberant rooms, a self-adaptive noise reduction system, based on a four-microphone array combined with an adaptive post-filtering scheme is proposed in [21], and its modified version in [22]. The two-channel version of the SS method is combined with a new estimator for noise power spectrum using two microphones [23]. It is also shown that non-stationary noise sources, as well as room reverberation, can be reduced [23]. Real-time implementation of a Singular Value Decomposition (SVD) – based optimal filtering technique is presented in [24] for noise reduction in a dual microphone behind-the-ear (BTE) hearing aid. Moreover, the noise reduction performance of an adaptive beam-former in a dual-microphone BTE hearing-aid application is evaluated in noisy environments [25]. Although coherence-based NR methods are designed for uncorrelated noise, a new coherence-based dual-microphone NR technique in [26] is shown to outperform the well-known GSC. The Power Level Difference (PLD) of the desired speech signals between the two microphones of mobile handsets is used for noise spectrum estimation [27]. As a result, to compromise between the noise attenuation performance and the computational cost, the two-microphone NR techniques should be chosen to alleviate the effect of additive background noise and musical noise, as compared to these multi-microphone ones or to the single-microphone ones.

In this paper, a two-microphone subband NR scheme for improving background noise attenuation performance, while preserving the speech quality, is introduced. Its main focus is for two-microphone voice communication systems, such as hearing aid applications where hearing-impaired people wear the devices in both ears, hands-free car kits and teleconferencing systems with two microphones and two loudspeakers to support stereo signals, etc. The proposed two-microphone subband NR scheme makes use of the conventional SS method and the modified version of the CSS method, which will be referred to as the modified CSS (MCSS) technique [16], for operating at two different frequency regions. This is due to distinct frequency characteristics of the noise and speech signals. In the low-frequency region, where the noise signals are mainly correlated and the speech dominant cues are located, the SS method, equipped with the proposed enhanced variable noise subtraction parameter, is applied separately to each channel to effectively suppress the noise spectral components with minimum speech distortion. As for the high-frequency region, the MCSS technique is employed in order to greatly remove the high-frequency decorrelated noise and the musical noise effect. Extensive comparisons among various NR techniques based on computer simulations demonstrate that the proposed two-microphone subband NR scheme obtains superior noise attenuation performance whereas the enhanced speech signal is minimally distorted.

This paper is organized as follows. In Section 2, the conventional SS method and the MCSS technique for noise reduction are summarized. In Section 3, the proposed two-microphone subband NR structure and the enhanced variable noise subtraction parameter for speech quality enhancement are described. Experimental results and discussions are given in Section 4, followed by the conclusion in Section 5.

2. Noise Reduction Methods

In this section, the conventional SS and the two-sensor modified Cross-Spectral Subtraction (MCSS) methods are outlined to provide a background for the proposed two-microphone subband NR technique described in Section 3.

2.1. The conventional SS method

The noisy microphone signal is modeled as a sum of the clean speech, $s(n)$, and the additive background noise, $b(n)$, as expressed by

$$x(n) \quad = \quad s(n)+b(n) \tag{1}$$

where n is the discrete-time index. By assuming that the speech signal, $s(n)$, and the additive background noise, $b(n)$, are uncorrelated to each other, the noisy microphone signal can be analyzed in the frequency-domain via the Short-time Fourier Transform (STFT) as given by

$$X(k,l) \quad = \quad S(k,l)+B(k,l) \tag{2}$$

where $l=1,2,\ldots$ is the analysis frame or block and the index $k=1,2,\ldots,N$ represents each frequency bin or frequency component of the analyzed spectrum while employing *N*-point STFT. From eq.(2), the enhanced speech spectrum $\hat{S}(k,l)$ can be obtained as follows:

$$|\hat{S}(k,l)|^p = \begin{cases} |X(k,l)|^p - \alpha \, |\hat{B}(k,l)|^p, & |\hat{S}(k,l)|^p \geq \beta \, |X(k,l)|^p \\ \beta \, |X(k,l)|^p , & \text{otherwise} \end{cases} \quad (3)$$

where p is the exponent and the noise subtraction parameter, α, controls the amount of noise spectral subtraction. When $p=1$, it is known as the Magnitude Spectral Subtraction (MSS) method while $p=2$, it will be referred to as the Power Spectral Subtraction (PSS) method. The spectral flooring parameter, β, is introduced to avoid negative estimated values of the magnitude or power spectrum after the subtraction process.

The noise spectrum estimate for each analysis frame can be obtained recursively during the non-speech-activity frame, as given by [28]

$$|\hat{B}(k,l)|^p = \gamma |\hat{B}(k,l-1)|^p + (1-\gamma)|X(k,l)|^p \quad (4)$$

where $0 < \gamma \leq 1$ is a forgetting factor. Hence, the use of a robust Voice Activity Detector (VAD) is necessary to determine whether the analysis frame represents the speech-activity or non-speech-activity (noise-only) frame. Thus, the noise spectrum estimator plays an important role in the SS method. The block diagram of the SS method can be illustrated in Fig. 1, where the enhanced speech signal, $\hat{s}(n)$, can be obtained by employing the inverse STFT (ISTFT). It is important to note that the SS method with a fixed value of the noise subtraction parameter, α, in eq.(3) suffers from the musical noise effect. To alleviate this problem, the variable noise subtraction parameter which is a function of the SNR was developed in [7]. This will be described in more detail in Section 3, along with the proposed noise subtraction parameter.

[Figure 1]

For noise reduction in two-microphone applications such as hearing aids where hearing-impaired people wear the devices in both ears, hands-free car kits, and teleconferencing systems with two microphones and two loudspeakers to support stereo signals, the SS method can be individually applied to each microphone signal.

2.2 The modified Cross-Spectral Subtraction (MCSS) technique

Based upon the CSS method in [15], its modified version in [16] was designed for noise reduction in hands-free car kits using two microphones. The coherence function between the two noisy microphone signals is used as a filter to remove the decorrelated residual noise and the musical noise effect. There are two main factors determining the setup of the two-microphone systems; one is the distance between the speaker and the two microphones, and the other is the microphones spacing. These factors have a great impact on the correlation between speech and noise signals received by the two microphones. It is assumed that the distance between the speaker and the two microphones is such that those microphones pick up a high portion of the direct desired signal [23]. Thus, the speech signals received by the two microphones are mutually correlated, i.e. the Magnitude-Squared-Coherence (MSC) function between the two microphone signals is close to one.

The two noise signals, $b_1(n)$ and $b_2(n)$, can be characterized as diffuse sound field [21]-[23]. The MSC of the two noise signals, $b_1(n)$ and $b_2(n)$, which is given by

$$\mathit{\Gamma}_{b_1 b_2}(k,\ell) = \frac{\left|\gamma_{b_1 b_2}(k,l)\right|^2}{\gamma_{b1}(k,l)\,\gamma_{b2}(k,l)}$$

exhibits strong correlation at the low frequency spectral region up to the "first minimum" frequency, $f_{\min}$, and shows decorrelation at the high frequency region beyond $f_{\min}$. The first minimum frequency of the MSC function, $\mathit{\Gamma}_{b_1 b_2}(k,\ell)$, can be obtained as $f_{\min} = c/2d$ where c

denotes the speed of sound and d is the microphone spacing [22]. A typical spacing between the two microphones results in $f_{\min}$ varying from 210 Hz at 80-cm spacing to 1700 Hz at 10-cm spacing [16], [22].

The two noisy signals picked up at the two microphones; $x_i(n)$, for $i = 1,2$, are modeled as shown in Fig. 2 to be the sum of the speech signal, $s_i(n)$, and additive background noise signals, $b_i(n)$, as given by

$$x_i(n) \quad = \quad s_i(n) + b_i(n). \tag{5}$$

Subsequently, the two noisy signals, $x_i(n)$, are analyzed in the frequency-domain via STFT, as given by

$$X_i(k,l) \quad = \quad S_i(k,l) + B_i(k,l). \tag{6}$$

[Figure 2]

A coherence filter, $G_{\mathrm{css}}(k,l)$, of the MCSS technique is employed to remove the decorrelated part of the noise signals beyond $f_{\min}$ [16]. It is based on the MSC function between the two microphone signals, $x_1(n)$ and $x_2(n)$, normalized by their auto-PSD product. The coherence filter, $G_{\mathrm{css}}(k,l)$, is formulated using the subtraction between the square root of the MSC function of $x_1(n)$ and $x_2(n)$, and the normalization of the norm of the noise cross-PSD, as given by

$$G_{\mathrm{css}}(k,\ell) \quad = \quad \frac{\left|\gamma_{x_1x_2}(k,l)\right| - \left|\gamma_{b_1b_2}(k,l)\right|}{\sqrt{\gamma_{x1}(k,l)\gamma_{x2}(k,l)}}. \tag{7}$$

To calculate the coherence filter in eq.(7), the noisy signal PSD of each microphone signal, $\gamma_{x_1}(k,l)$ and $\gamma_{x_2}(k,l)$, and the cross-PSD of the two microphone signals, $\gamma_{x_1x_2}(k,l)$, are estimated recursively from their corresponding spectra as follows:

$$\gamma_{x_i}(k,l) \quad = \quad \lambda(k,l)\gamma_{x_i}(k,l-1) + \left[1-\lambda(k,l)\right]X_i(k,l)X_i^*(k,l) \tag{8}$$

and

$$\gamma_{x_1x_2}(k,l) \quad = \quad \lambda(k,l)\gamma_{x_1x_2}(k,l-1) + \left[1-\lambda(k,l)\right]X_1(k,l)X_2^*(k,l) \tag{9}$$

for $i=1,2$, $\langle\cdot\rangle^*$ denotes the complex conjugate of the vector quantity, and $0<\lambda(k,l)\leq 1$ is the forgetting factor. It was suggested in [16] that the forgetting factor, $\lambda(k,l)$, for eq.(8) and eq.(9) depends on the SNR of the *first* microphone signal for each frequency bin, k, and analysis frame, l,as given by

$$\lambda(k,l) \quad = \quad 0.98-0.3\left(\frac{\mathrm{SNR}(k,l)}{1+\mathrm{SNR}(k,l)}\right). \tag{10}$$

Since the SNR is frequency-dependent, so is the forgetting factor, $\lambda(k,l)$. It can be seen from eq.(10) that, at low SNR, the forgetting factor, $\lambda(k,l)$, is large to allow smoothed estimation of the PSD, thus, musical noise can be controlled during non-speech-activity frames. On the other hand, when the SNR value is high, the forgetting factor, $\lambda(k,l)$, takes small values so that the PSD estimates follow fast speech variations during speech-activity frames.

By assuming that the SNR does not change rapidly from one frame to another, the ratio of the SNR in eq.(10) can then be expressed in terms of the function of the coherence filter in the previous frame [16] as

$$\frac{\mathrm{SNR}(k,l)}{1+\mathrm{SNR}(k,l)} \quad \simeq \quad G_{\mathrm{css}}(k,l-1). \tag{11}$$

Thus, the forgetting factor, $\lambda(k,l)$, can be written as a function of the coherence filter in the previous frame as

$$\lambda(k,l) \quad \simeq \quad 0.98-0.3G_{\mathrm{css}}(k,l-1). \tag{12}$$

However, with such choices of the forgetting factor, the authors in [16] found that the musical noise still occurred during speech-activity frames.

As for the estimation of the norm of the noise cross-PSD, $\left|\gamma_{b_1 b_2}(k,l)\right|$, in eq.(7), it is normally obtained during noise-only periods using VAD. It is suggested in [15], [16] to use the overestimated value, $\sqrt{\gamma_{b_1}(k,l)\gamma_{b_2}(k,l)}$, instead, as given in the following expressions.

$$\sqrt{\gamma_{b_1}(k,l)\gamma_{b_2}(k,l)} \quad = \quad d\left(\widetilde{\mathrm{SNR}}_{\mathrm{post}}(k,l)\right)\sqrt{\gamma_{b_1}(k,l-1)\gamma_{b_2}(k,l-1)} \tag{13}$$

where γ_{b_i} is the noise PSD in each channel, $i=1,2$, and the function $d\left(\widetilde{\mathrm{SNR}}_{\mathrm{post}}(k,l)\right)$ depends on real positive constants, $h,\ g,\ L$, and the posteriori modified SNR, $\widetilde{\mathrm{SNR}}_{\mathrm{post}}(k,l)$;

$$d\left(\widetilde{\mathrm{SNR}}_{\mathrm{post}}(k,l)\right)=L+\left(\frac{(1-L)}{1+1/\left(g\cdot\widetilde{\mathrm{SNR}}_{\mathrm{post}}(k,l)\right)}\right)\left(1+\frac{1}{1+g\cdot h\cdot\widetilde{\mathrm{SNR}}_{\mathrm{post}}(k,l)}\right). \tag{14}$$

The posteriori modified SNR, $\widetilde{\mathrm{SNR}}_{\mathrm{post}}(k,l)$, is given by

$$\mathrm{SNR}_{\mathrm{post}}(k,l) \quad = \quad \frac{\left|X_1(k,l)X_2(k,l)\right|}{\sqrt{\gamma_{b_1}(k,l-1)\gamma_{b_2}(k,l-1)}}. \tag{15}$$

Normally, $g = 1/(1-h)$ and $0 < L < 1$ were chosen in [16]. The advantage of this approach to *overestimate* the noise cross-PSD, $\gamma_{b_1 b_2}(k,l)$, is that it does *not* require any VAD in the system.

By applying the coherence filter to one of the noisy spectra, $X_i(k,\ell)$, for $i = 1,2$, the enhanced speech spectra is therefore obtained as

$$\hat{S}_i(k,\ell) \quad = \quad G_{\mathrm{CSS}}(k,\ell) X_i(k,\ell). \tag{16}$$

The enhanced speech signal, $\hat{s}_i(n)$, can therefore be obtained by employing the inverse STFT process. Fig. 2 shows the block diagram of the MCSS technique when the coherence filter is applied to channel 1 of the system. Similarly, if the coherence filter is applied to the noisy spectrum from the second microphone, $\hat{S}_2(k,\ell)$, the enhanced speech signal, $\hat{s}_2(n)$, will be obtained.

It is important to note that the MCSS technique, however, has a limitation in that the spacing between the two microphones of the system must be sufficiently far apart so that the noise signals from two microphones are decorrelated.

3. The proposed two-microphone subband NR scheme

3.1 Two-microphone subband structure

The proposed two-microphone subband NR scheme is developed by exploiting the distinct frequency characteristics of the noise and speech signals. The noise characteristics from the two microphones are considered as diffuse where their MSC function exhibits high coherence in the low-frequency band up to the first-minimum frequency, $f_{\min}$, and small coherence in the high-frequency band beyond $f_{\min}$. Note that, $f_{\min}$ is in turn dependent upon the microphone spacing [22], as described in Section 2.2. Thus, for a microphone spacing of about 15 cm, the first-minimum frequency, $f_{\min}$, of 1 *k*Hz is obtained. For the speech characteristics, the MSC function of the speech signals received by the two microphones exhibits high coherence close to unity, and their dominant cues are usually located in the low-frequency region below 1 *k*Hz [6].

Following the described characteristics, the proposed two-microphone subband NR scheme makes use of the SS and the MCSS methods for operating at two different frequency regions. The SS method is employed to effectively suppress the noise spectral components in the low frequency band below $f_{\min}$, where the noise signals are mainly correlated and the speech dominant cues are mainly located, i.e. the MSC functions of the speech signals and the noise signals are close to one. In addition, an enhanced variable noise subtraction parameter, $\alpha(k,l)$, as introduced in Section 3.2, is suggested to be equipped with the SS method in order to obtain superior noise attenuation performance with minimal distortion on the quality of the enhanced speech signal. On the other hand, in the high frequency band above $f_{\min}$, where the noise characteristics exhibit small coherence and the speech signals have less spectral components, the proposed two-microphone subband NR scheme exploits the MCSS technique which is much more efficient for removal of the decorrelated noise signals and the musical noise effect.

[Figure 3]

The block diagram of the proposed two-microphone subband NR technique is shown in Fig. 3. The noisy signal from each microphone, $x_i(n)$, for $i = 1, 2$, is divided into two frequency bands, namely the low-frequency (*L*) band and the high-frequency (*H*) band. The low-frequency noisy speech signal is obtained by passing the noisy signal from each microphone, $x_i(n)$, through a linear-phase lowpass Finite Impulse Response (FIR) filter with the order of 127, and the cutoff frequency of 1 *k*Hz (shown as 'LPF' in Fig. 3). Similarly, by passing the noisy signal from each microphone, $x_i(n)$, through a linear-phase highpass FIR filter with the same order and the cutoff frequency (shown as 'HPF' in Fig. 3), the high-frequency noisy speech signal is obtained. As a result, near perfect reconstruction of the sum of the low-frequency and high-frequency signals can be obtained since the sum of the lowpass and highpass filter responses is almost flat. Consequently, this yields practically minimum impact on the speech quality of the reconstructed signal.

For noise reduction in the low-frequency band, the low-frequency noisy signals from each microphone, $x_{L,i}(n)$, for $i = 1, 2$, are sent to the SS method that employs the proposed variable noise subtraction parameter, $\alpha(k,l)$, as shown in Fig. 3. As for the high-frequency band, the noisy signals from each microphone, $x_{H,i}(n)$, are sent to the MCSS technique to remove especially the decorrelated noise components. Consequently, the frequency components of the enhanced speech signals in both low-frequency and high-frequency bands for each i^{th} channel, $\hat{s}_{L,i}(n)$, and $\hat{s}_{H,i}(n)$, respectively, are combined to obtain the full-band enhanced speech signals, $\hat{s}_i(n)$, for each channel, $i = 1, 2$.

Since both the SS and MCSS techniques are implemented in the proposed two-microphone subband NR scheme, the computational complexity will certainly be increased as compared to those of single-microphone NR methods. Nevertheless, with the state-of-the-art technology for integrated circuits (IC) implementation, the additional computational complexity should no longer be an important practical issue.

3.2 A new noise subtraction parameter for speech quality enhancement

In this section, the use of variable noise subtraction parameters for the SS method is described to improve the speech quality as compared to the conventional case with a fixed subtraction parameter. In addition, a new noise subtraction parameter is proposed for enhanced speech quality.

In order to suppress the effect of the musical noise, the SS method employs a variable noise subtraction parameter which is dependent on the SNR levels [7]. This is based on the fact that, at each particular frequency bin, if the speech spectral component is much more dominant the spectral component of the noise, a high SNR level, the speech will partially *mask* the noise. Thus, a "small" value of noise subtraction parameter should be chosen for minimum speech distortion. On the other hand, when the SNR level is low, the value of the noise subtraction parameter should be "high" so that a large amount of noise subtraction is obtained. It is therefore desirable for the SS method to employ variable noise subtraction parameter, $\alpha(k,l)$, that is a function of SNR for each frequency bin, k, and analysis frame, l, instead of a fixed value α in eq.(3) of the conventional SS method.

In [7], the so-called noise over-subtraction parameter, $\alpha_{\mathrm{OS}}(k,l)$, is suggested to be a linear function of the posteriori SNR, as given by

$$\alpha_{\mathrm{OS}}(k,l) \quad = \quad \alpha_0(k,l) - \left(\frac{3}{20}\right)\mathrm{SNR}_{\mathrm{post}}(k,l) \tag{17}$$

where the posteriori SNR is obtained at frequency bin, k, and analysis frame l by

$$\mathrm{SNR}_{\mathrm{post}}(k,l) \quad = \quad 10 \times \log_{10}\left\{\frac{|X(k,l)|^2}{\left|\hat{B}(k,l)\right|^2}\right\}. \tag{18}$$

In eq.(17), $\alpha_0(k,l)$ is the value of $\alpha_{\mathrm{OS}}(k,l)$ when $\mathrm{SNR}_{\mathrm{post}}(k,l)$ is equal to $0\ dB$. Based on extensive experiments, it was suggested in [7] that $\alpha(k,l)=1$ for $\mathrm{SNR}(k,l) \geq 20\,dB$ and $\alpha(k,l)$ is not allowed to increase for $\mathrm{SNR}(k,l) \leq -5\,dB$. Fig. 4 illustrates the value of $\alpha(k,l)$ as a function of $SNR_{\mathrm{post}}(k,l)$ when $\alpha_0(k,l)=4$.

[Figure 4]

This concept of variable noise subtraction parameter in [7], $\alpha_{\mathrm{OS}}(k,l)$, has been extended as presented in [8]-[10]. In [8], the variable noise subtraction parameter, $\alpha_{\mathrm{OS}}(k,l)$, was employed in the multi-band SS method for the case of the colored-noise-corrupted speech. To enable control over each of the multi-band frequencies, a tweaking factor, $\delta_{\mathrm{MB},q}$, was also multiplied with $\alpha_{\mathrm{MB},q}$ for customizing the noise removal in each q^{th} band. At low-frequency band, $\delta_{\mathrm{MB},q}$, is suggested to be low so as to minimize the speech distortion, whereas higher values of $\delta_{\mathrm{MB},q}$ can be applied for higher frequencies. It is noted that the product of the noise subtraction parameter and the tweaking factor, $\alpha_{\mathrm{MB},q}\delta_{\mathrm{MB},q}$, employed in [8] for the low-frequency band is exactly the same as $\alpha_{\mathrm{OS}}(k,l)$ in [7].

To further enhance the speech quality, a new variable noise subtraction parameter $\alpha(k,l)$ is introduced in this paper. The underlying idea is to further reduce the values of $\alpha(k,l)$, particularly at high SNR levels as compared to $\alpha_{\mathrm{OS}}(k,l)$ in eq.(17), while still maintaining the high values of $\alpha(k,l)$ at low SNR levels. In this way, the amount of speech distortion can be better controlled while still achieving sufficient level of noise reduction with a consequent benefit to enhancing the speech quality. The above requirement suggests that the SNR-dependent relationship be different from the linear function in eq.(17). Thus, in this work, $\alpha(k,l)$ is proposed to be *inversely* proportional to the posteriori SNR as described by

$$\alpha(k,l) = \begin{cases} \alpha_{\min}, & \alpha(k,l) < \alpha_{\min} \\ \alpha_{\mathrm{int}}(k) + \dfrac{\sigma}{\varepsilon + \mathrm{SNR}_{\mathrm{post}}(k,l)}, & \alpha_{\min} \le \alpha(k,l) < \alpha_{\max}, \\ \alpha_{\max}, & \text{otherwise} \end{cases} \qquad (19)$$

where $\alpha_{\text{int}}(k)$ is the initial value of the noise subtraction parameter for the k^{th} frequency bin of all frames, $l = 1, 2, \ldots$ is the analysis frame, and σ is a positive integer to enable control of the noise removal at each frequency bin. The parameter $\alpha_{\min}$ is chosen as a lower limit of $\alpha(k,l)$ to ensure sufficient amount of noise reduction at high SNRs. It is suggested in this paper that $\alpha_{\min}$ is greater than zero, but smaller than one, i.e. $0 < \alpha_{\min} < 1$. The parameter $\alpha_{\max}$ is the upper limit of $\alpha(k,l)$ and it is suggested that over-subtraction is allowed, i.e. $\alpha_{\max} > 1$. The parameter ε in eq.(19) is to prevent $\alpha(k,l)$ from being excessively large when the term $\text{SNR}_{\text{post}}(k,l)$ is close or equal to zero, and is determined by the following equation;

$$\varepsilon = \frac{1}{\alpha_{\max} - \alpha_{\min}} + \text{SNR}' \tag{20}$$

where $\alpha(k,l)$ is not allowed to increase beyond $\alpha_{\max}$ for $\text{SNR}_{\text{post}}(k,l) \leq \text{SNR}'$. In this work, it is suggested that $\alpha_{\min}$, $\alpha_{\max}$, and σ are chosen to be the same for all frequency bins, $k = 1, \ldots, N$. The posteriori SNR, $\text{SNR}_{\text{post}}(k,l)$, is defined as given in eq.(18). The enhanced speech spectrum in the i^{th} channel for $i = 1, 2$ when employing the SS method with $p = 2$, jointly with the proposed noise subtraction parameter, $\alpha(k,l)$, becomes

$$\left|\hat{S}_i(k,l)\right|^2 = \begin{cases} \left|X_i(k,l)\right|^2 - \alpha(k,l)\left|\hat{B}_i(k,l)\right|^2, & \left|\hat{S}_i(k,l)\right|^2 \geq \beta\left|X_i(k,l)\right|^2 \\ \beta\left|X_i(k,l)\right|^2, & \text{otherwise} \end{cases}. \tag{21}$$

The spectral floor parameter is chosen to be $\beta = 0.01$ [1], [28]. Fig. 5 illustrates the plot of the proposed $\alpha(k,l)$ versus $\text{SNR}_{\text{post}}(k,l)$ when $\alpha_{\min} = 1$, $\alpha_{\max} = 2$, $\sigma = 1$, and

$\text{SNR}' = 0\,dB$. This yields $\varepsilon = 1$ as calculated by eq.(20). As evident from the plot, $\alpha(k,l)$ is kept close to $\alpha_{\min} = 1$ at high SNR levels, and there is a sharp increase in $\alpha(k,l)$ as the SNR approaches the defined value at $\text{SNR}' = 0\,dB$ where it is kept at $\alpha_{\max} = 2$ for $\text{SNR}(k,l) \le 0\,dB$.

[Figure 5]

Note that, the use of the proposed variable noise subtraction parameter together with the multi-band structure of the SS method is entirely possible.

4. Experimental results and discussions

Throughout all the experiments, the following setups were applied. A clean speech signal of a female speaker, which was obtained from a standard speech database [29], at the sampling frequency of 8 *k*Hz, was degraded by a number of additive background noise signals at various input SNR levels ranging from 5 dB to 20 dB.

The two-microphone babble noise signals were recorded in a moving car with a microphone spacing of 15 cm, yielding the first minimum frequency, $f_{\min}$, of about 1 *k*Hz [22]. Given also that the dominant cues of the speech signals were mainly located in the low-frequency region below 1 *k*Hz [6], the crossover frequency between the low-frequency and the high-frequency bands of the two-microphone subband NR scheme was thus set at 1 *k*Hz.

In addition, the single-microphone pink noise, factory noise, and F16 cockpit noise were obtained from the NOISEX-92 database [30]. To generate their two-microphone versions, these noise signals were time-shifted to effect a microphone spacing of 15 cm, and subsequently filtered by two impulse responses that represent the signal path between these noise source and two microphones. Moreover, White Gaussian Noise (WGN), was also included in order to allow observation of the noise attenuation performance of the investigated NR techniques. The speech and background noise signals were assumed to be uncorrelated to each other. The analysis frame for STFT was 25 ms with the 12.5 ms frame shift and the Hamming window was employed. The energy-based VAD scheme in [31] was employed in all experiments.

4.1 Objective results

Three experiments were carried out. The first experiment was to find out the appropriate choice of the upper limit for the proposed variable noise subtraction parameter, $\alpha_{\max}$, of the SS method. In the second experiment, the proposed variable noise subtraction parameter and the noise over-subtraction technique in [7] for the Spectral Subtraction (SS) method (when $p = 2$) were compared with the conventional SS method. The setup for these two experiments was a single-microphone system. The parameters $\alpha_{\min}$, $\alpha_{\max}$, σ, of the proposed variable noise subtraction parameter, $\alpha(k,l)$, were chosen according to Fig. 5. The choice of the fixed α values of the conventional SS method was chosen to obtain the best noise attenuation performance at each input SNR level. In the final experiment, the two-microphone subband NR scheme with the proposed variable noise subtraction parameter, $\alpha(k,l)$, was compared to all other investigated NR techniques; the MCSS technique, the SS method with the proposed noise subtraction parameter, $\alpha(k,l)$, and the SS method with the noise over-subtraction technique in [7], $\alpha_{\mathrm{OS}}(k,l)$, in a two-microphone system.

The performances of the investigated NR techniques are given via a number of objective measures. The output SNR indicates the noise attenuation performances of those NR techniques, i.e. a higher output SNR level than the input SNR one suggests improved noise attenuation performance. On the other hand, a lower output SNR level than the input SNR one signifies that the enhanced speech signal has residual noise. The output SNR is defined as

$$\mathrm{SNR}_{\mathrm{o/p}}(\mathrm{dB}) \quad = \quad 10 \times \log_{10} \left\{ \frac{\sum_{n=0}^{M-1} s^2(n)}{\sum_{n=0}^{M-1} \left(s(n) - \hat{s}(n) \right)^2} \right\} \tag{22}$$

where $s(n)$ is the clean speech signal, $\hat{s}(n)$ is the enhanced speech signal, and M is the total number of speech samples.

The Log Spectral Distance (LSD) is used as a speech distortion measure. It is described as the difference between the log power spectrum of clean speech signal and that of the enhanced signal, given by

$$\mathrm{LSD}\,(dB) \quad = \quad \frac{10}{J} \sum_{l=0}^{J-1} \left\{ \log_{10}(\left|S(k,l)\right|^2) - \log_{10}(\left|\hat{S}(k,l)\right|^2) \right\} \tag{23}$$

where J is the number of speech-activity frames. A low LSD level suggests a low level of speech distortion.

The Perceptual Evaluation of Speech Quality (PESQ) score, based on ITU-T Recommendation P.862, is an objective measure for predicting the speech quality of the narrow-band (3.1 *k*Hz) speech signals; noisy and enhanced speech signals, that will be obtained in a subjective listening test [32]. The PESQ scores are normally in the range of -0.5 to 4.5.

Note that, these values, which were obtained from the enhanced signal from the first microphone (channel 1), were used for the calculation of the output SNR, LSD, and the PESQ scores of the two-microphone system.

Experiment 1: The upper limit for the proposed variable noise subtraction parameter

In order to find out a suitable choice of the upper limit for the variable noise subtraction parameter, $\alpha_{\max}$, it is shown in Fig. 6 (a) that as $\alpha_{\max}$ is increased beyond two, there is no further improvement on the noise attenuation performance obtained, via the output SNR. Similarly, Fig. 6 (b) demonstrates that the LSD value is no longer decreased as $\alpha_{\max}$ is increased beyond two. Therefore, the upper limit of the variable noise subtraction parameter, $\alpha(k,l)$, in eq.(19) will be selected at $\alpha_{\max}=2$ for the rest of the experiments.

[Figure 6]

Experiment 2: Fixed noise subtraction parameter vs. Variable noise subtraction parameters for the SS method

The proposed variable noise subtraction parameter, $\alpha(k,l)$, and the noise over-subtraction technique in [7], $\alpha_{\mathrm{OS}}(k,l)$, for the SS method (when $p=2$) were compared with the conventional case with a fixed noise subtraction parameter α. The single-microphone system was considered and the additive background noise signals were WGN and babble noise with the input SNR level of 10 dB.

For the WGN case, the NA performance of the SS method with the proposed variable noise subtraction parameter, $\alpha(k,l)$, outperforms the noise over-subtraction technique in [7], $\alpha_{\mathrm{OS}}(k,l)$, and the conventional SS method with a fixed noise subtraction parameter, α, as shown in Table 1. It is also demonstrated that the proposed variable noise subtraction parameter introduces lower level of speech distortion than the noise over-subtraction technique in [7] and the fixed noise subtraction parameter, as shown in Table 2. Similar results are obtained for the case of babble noise. It is found that, when the input SNR is 20 dB, the conventional SS method yields the best NA performance. However, the conventional SS method does not introduce the lowest level of speech distortion.

[Table 1]

[Table 2]

Experiment 3: Comparison among the proposed two-microphone NR scheme and the other investigated techniques

In this experiment, the proposed two-microphone subband NR scheme with the proposed variable noise subtraction parameter, $\alpha(k,l)$, was mainly investigated and compared with the MCSS technique, and the SS method using the proposed variable noise subtraction parameter, $\alpha(k,l)$, and the noise over-subtraction technique in [7], $\alpha_{\mathrm{OS}}(k,l)$. Note also that, the NA performance and speech distortion measures of the enhanced speech spectra could also be observed from their spectrogram plots, as illustrated in Fig. 7 and Fig. 8 for the cases of WGN and babble noise, respectively. The spectrogram plots of the noisy speech signals were as given in Fig. 7 (a) for the WGN case and Fig. 8 (a) for the babble noise case, respectively. From these plots, it is clearly shown that the spectral noise components of babble noise are only located at low frequencies, i.e. below 1 kHz, while those of WGN are equally distributed across the whole frequency region.

It is demonstrated in Table 3 that the NA performance of the two-microphone MCSS technique is better than the single-microphone SS method. The proposed two-microphone subband NR scheme yields the best NA performance among the other investigated NR techniques for the cases of WGN, pink noise, and F16 cockpit noise. Furthermore, the proposed two-microphone subband NR scheme introduces the least amount of speech distortion, as shown via the LSD measure in Table 4. The spectrogram plots in Fig. 7 (d) for the WGN case and Fig. 8 (d) for the case of babble noise also confirm similar indication for the ability to preserve speech spectral components of the proposed two-microphone subband NR scheme, as compared to the other investigated techniques.

As for the cases of babble noise and factory noise, the proposed two-microphone subband NR technique exhibits comparable performance, in terms of the output SNR, to the

MCSS technique. In addition, the LSD measure of the proposed two-microphone subband NR scheme is close to that of the SS method which employed the proposed variable noise subtraction parameter, particularly at high input SNR levels. By considering at the spectrogram plots of the noisy speech signals, as given in Fig. 7 (a) for the WGN case and Fig. 8 (a) for the babble noise case, respectively, it is evident that the MCSS technique considerably removes the speech spectral components of the enhanced signal, particularly in the high-frequency region. As a result, the MCSS technique introduces a significant amount of speech distortion. As shown via the LSD measures in Table 4, the LSD measures of the MCSS technique is higher than those of the SS method, especially for the case of babble noise and factory noise.

[Table 3]

[Table 4]

[Figure 7]

[Figure 8]

The PESQ scores of the noisy and the enhanced speech signals of the investigated NR techniques are given in Table 5, for various noise types with the input SNR level of 10 dB. It is observed that the proposed two-microphone subband NR technique with the proposed scheme for variable noise subtraction parameter yields the best PESQ scores among the other investigated NR techniques. The SS method with the variable noise subtraction parameter in [7] achieves the second place ranking, followed by the proposed scheme for variable noise subtraction parameter. The PESQ scores of the MCSS technique are higher than those of the conventional SS method.

[Table 5]

4.2 Subjective results

A subjective listening test was undertaken with 15 listeners, based on the ITU-T Recommendation P.800 [33]. The 5-point-scale Mean Opinion Score (MOS) score is a measure of the overall quality of the processed speech signals, ranging from 1 (bad) to 5 (excellent). In this subsection, the MOS scores of the enhanced speech signals of all of the investigated NR techniques are compared with their noisy versions at the input SNR level of 10 dB. The MOS score of the clean speech signal is found to be 4.85.

[Table 6]

It is shown in Table 6 that the enhanced speech signal of the proposed two-microphone subband NR scheme yields the highest MOS scores among the investigated NR techniques for almost all of the additive noise types. The MOS scores of the SS method with the variable noise subtraction parameter in [7] and the SS method with the proposed scheme for variable noise subtraction parameter offer the second best performance in terms of speech quality. The MOS score of the MCSS technique is always less than that of the proposed two-microphone subband NR technique, but is higher than that of the conventional SS method, except for the case of babble noise and factory noise. In general, it is also observed that these listeners find the factory noise, pink noise, and F16 cockpit noise less annoying than the WGN and the babble noise. In addition, it is found that the PESQ scores obtained in Table 5 are in agreement with the actual subjective MOS scores obtained by the subjective listening test.

5. Conclusions

A two-microphone subband NR scheme with a new variable noise subtraction parameter has been proposed in this paper to offer advantages of the SS and the MCSS techniques. In the low-frequency region, the SS method is employed, while, in the high-frequency region, the MCSS technique is operated. As demonstrated via experimental results of both the objective and subjective performances, the proposed two-microphone subband NR scheme is able to control the amount of noise reduction, while introducing less amount of speech distortion as compared to the other investigated NR techniques.

Acknowledgements

This research work was financially supported by AUN/SEED-Net project, JICA.

References

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Transactions on Acoustics, Speech and Signal Processing, 1979, ASSP-27, (2), pp. 113-120.

[2] K. Chung, F.G. Zeng and S. Waltzman, "Utilizing hearing aid directional microphones and noise reduction algorithms to improve speech understanding and listening preferences for cochlear implant users," International Congress Series, 2004, 1273, pp.89-92.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator", IEEE Transactions on Acoustics, Speech, and Signal Processing, 1984, 32, (6), pp. 1109-1121.

[4] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement", IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14, (6), pp. 2098-2108.

[5] R. Thoonsaengngam and N. Tangsangiumvisai, "The a priori SDR estimation technique with reduced speech distortion for Acoustic Echo and Noise Suppression", IEICE Transactions on Communications, 2009, E92-B, (10), pp. 3022-3033.

[6] F.L. Wightman and D.J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization", Journal on Acoustical Society of America, 1992, 91, (3), pp. 1648-1661.

[7] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise”, in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Washington, USA., April 1979, pp. 208-211.

[8] S.D. Kamath and P.C. Loizou, “A multi-band spectral subtraction method for enhancing speech corrupted by colored noise”, in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, U.S.A., May 2002, pp. IV-4164.

[9] C. Li and W.J. Liu “A novel multi-band spectral subtraction method based on phase modification and magnitude compensation”, in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Prague, Czech Republic, May 2011, pp. 4760-4763.

[10] S. Hayashi, H. Inukai, and M. Suguimoto, “A Subtractive-Type Speech Enhancement Using the Perceptual Frequency-Weighting Function”, IEICE Transactions on Fundamentals, 2009, E92-A, (1), pp. 226-234.

[11] T. Esch and P. Vary, “Efficient musical noise suppression for speech enhancement systems”, in Proc. International Conference on Acoustic, Speech, and Signal Processing (ICASSP), Taipei, Taiwan, April 2009, pp. 4409-4412.

[12] J. Chen, J. Benesty, Y. Huang, and S. Doclo, “New insights into the noise reduction Wiener filter”, IEEE Transactions on Audio, Speech, Language Processing, 2006, 14, (4), pp. 1218-1234.

[13] J. Li, M. Akagi and Y. Suzuki, “A two-microphone noise reduction method in highly non-stationary multiple-noise-source environments”, IEICE Transactions on Fundamentals, 2008, E91-A, (6), pp. 1337-1346.

[14] J. Li, M. Akagi and Y. Suzuki, “Extension of the Two-Microphone Noise Reduction Method for Binaural Hearing Aids,” in Proc. International Conference on Audio, Language, and Image Processing (ICALIP), Shanghai, China, July 2008, pp. 97-101.

[15] A. A. Azirani, R.L. Bouquin-Jeannes, and G. Faucon, “Enhancement of speech degraded by coherent and inconherent noise using cross-spectral subtraction”, IEEE Transactions on Speech, and Audio Processing, 1997, 5, (5), pp. 484-487.

[16] A. Guerin, R.L. Bouquin-Jeannes, and G. Faucon, “A two-sensor noise reduction system: Applications for hands-free car kit”, EURASIP Journal on Applied Signal Processing, 2003, 2003, (11), pp. 1125-1134.

[17] X. Zhang and Y. Jia, “A soft-decision based noise power cross power spectral density estimation for two-microphone speech enhancement systems”, in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, USA., April 2005, pp. 813-816.

[18] M. Rahmani, A. Akbari, B. Ayad, M. Mazoochi, and M.S. Moin, “A modified coherent based method for dual microphone speech enhancement”, in Proc. International Conference on Signal Processing and Communications (ICSPC), Dubai, United Arab Emirates, November 2007, pp. 225-228.

[19] B. Zamani, M. Rahmani, and A. Akbari, “Residual Noise Control for Coherence Based Dual Microphone Speech Enhancement”, in Proc. International Conference on Computer and Electrical Engineering (ICCEE), Phuket, Thailand, March 2008, pp. 601-605.

[20] M. Rahmani, A. Akbari, B. Ayad, and B. Lithgow, “Noise cross PSD estimation using phase information in diffuse noise field,” Signal Processing, 2009, 89, (5), pp. 703–709.

[21] R. Zelinski, “A microphone array with adaptive post-filtering for noise reduction in reverberant rooms”, in Proc. International Conference on Acoustic, Speech, and Signal Processing (ICASSP), New York, USA., April 1988, pp. 2578-2581.

[22] I. A. McCowan and H. Bourlard, “Microphone array post-filter for diffuse noise field”, in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, USA., May 2002, pp. 905-908.

[23] M. Dorbecker and S. Ernst, “Combination of Two-channel Spectral Subtraction and Adaptive Wiener Post-Filtering for Noise Reduction and Dereverberation”, in Proc. European Signal Processing Conference (EUSIPCO), Trieste, Italy, September 1996, pp. 995-998.

[24] J.B. Maj, M. Moonen, and J. Wouters, “SVD-Based Optimal Filtering Technique for Noise Reduction in Hearing Aids Using Two Microphones”, EURASIP Journal on Applied Signal Processing, 2002, 2002, (4), pp. 432-443.

[25] J.B. Maj, J. Wouters, M. Moonen, “Noise Reduction Results of an Adaptive Filtering Technique for Dual-Microphone Behind-the-Ear Hearing Aids”, Ear & Hearing : Journal of the American Auditory Society, 2004, 25, (3), pp. 215-229.

[26] N. Yousefian, K. Kokkinakis, and P.C. Loizou, “A coherence-based algorithm for noise reduction in dual-microphone applications”, in Proc. European Signal Processing Conference (EUSIPCO), Aalborg, Denmark, August 2010, pp. 1904-1908.

[27] M. Jeub, C. Herglotz, C. Nelke, C. Beaugeant, and P. Vary, “Noise reduction for dual-microphone mobile phones exploiting power level differences”, in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, Japan, March 2012, pp. 1693-1696.

[28] S.V. Vaseghi, “Advanced Digital Signal Processing and Noise Reduction”, (John Wiley and Sons Ltd., 2nd edn. 2000), Chapter 11.

[29] “TIMIT (Texas Instruments (TI) and Massachusetts Institute of Technology (MIT)) speech database”, which is sponsored by DARPA (The Defense Advanced Research Projects Agency of the United States Department of Defense), http://www.ldc.upenn.edu/, accessed June 2013.

[30] “NOISEX-92 : database of recording of various noises”, http://sipl.technion.ac.il/Info/Downloads_DataBases_NoiseX92_e.shtml, accessed June 2013.

[31] D. Enqing, L. Guizhong, Z. Yatong, and C. Yu, “Voice Activity Detection Based on Short-Time Energy and Noise Spectrum Adaptation” in Proc. International Conference on Signal Processing (ICSP), Beijing, China, October 2002, pp. 464-467.

[32] “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs”, ITU-T Recommendation, P.862, 2001.

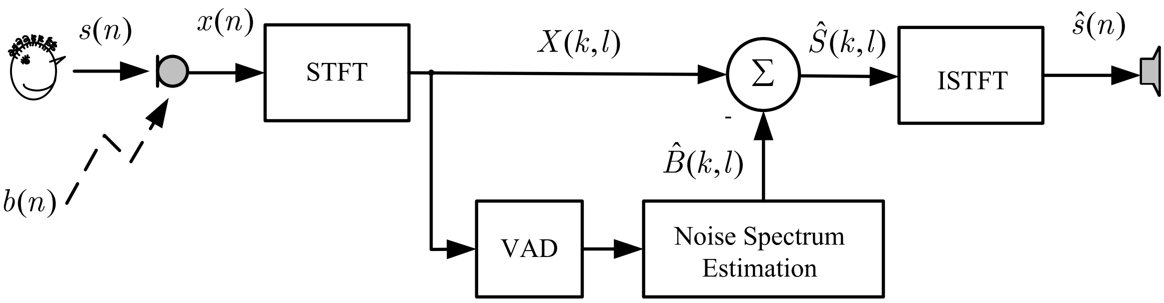
[33] “Methods for subjective determination of transmission quality”, ITU-T Recommendation, P.800, 1996.

List of Figures

List of Tables

page

Figure 1



s(n)
x(n)
STFT
X(k,l)
Σ
Ŝ(k,l)
ISTFT
ŝ(n)
b(n)
-
B̂(k,l)
VAD
Noise Spectrum
Estimation

Figure 2

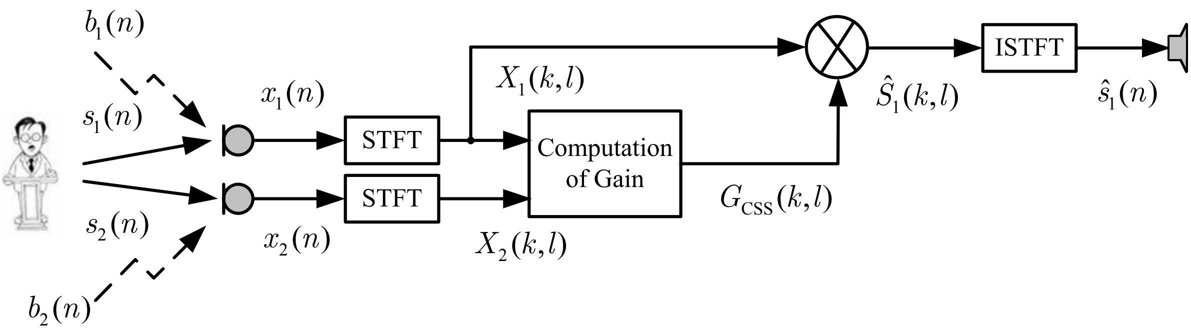
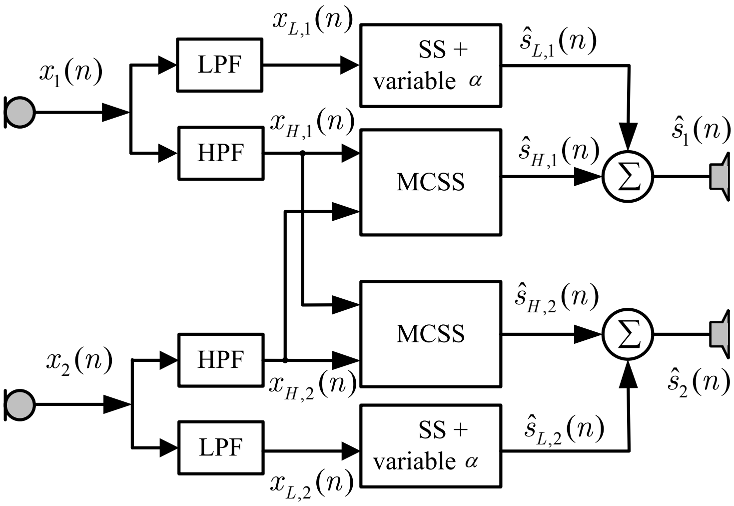

$b_1(n)$
$s_1(n)$
$s_2(n)$
$b_2(n)$
$x_1(n)$
$x_2(n)$
STFT
STFT
$X_1(k,l)$
$X_2(k,l)$
Computation of Gain
$G_{\mathrm{CSS}}(k,l)$
$\hat{S}_1(k,l)$
ISTFT
$\hat{s}_1(n)$

Figure 3



$x_1(n)$
LPF
HPF
$x_{L,1}(n)$
$x_{H,1}(n)$
SS + variable α
MCSS
$\hat{s}_{L,1}(n)$
$\hat{s}_{H,1}(n)$
$\hat{s}_1(n)$
Σ
$x_2(n)$
HPF
LPF
$x_{H,2}(n)$
$x_{L,2}(n)$
MCSS
SS + variable α
$\hat{s}_{H,2}(n)$
$\hat{s}_{L,2}(n)$
$\hat{s}_2(n)$
Σ

Figure 4

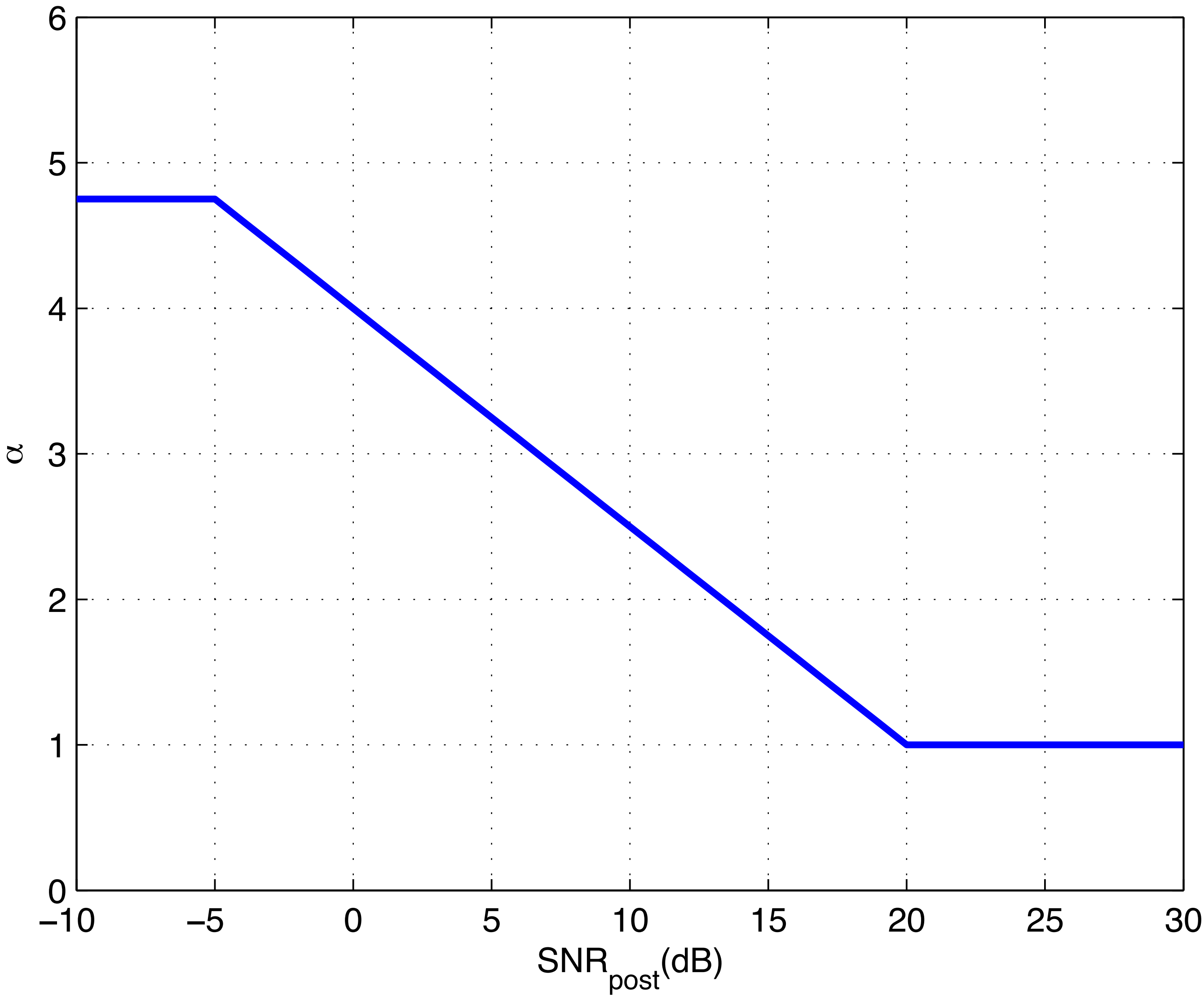


Figure 5

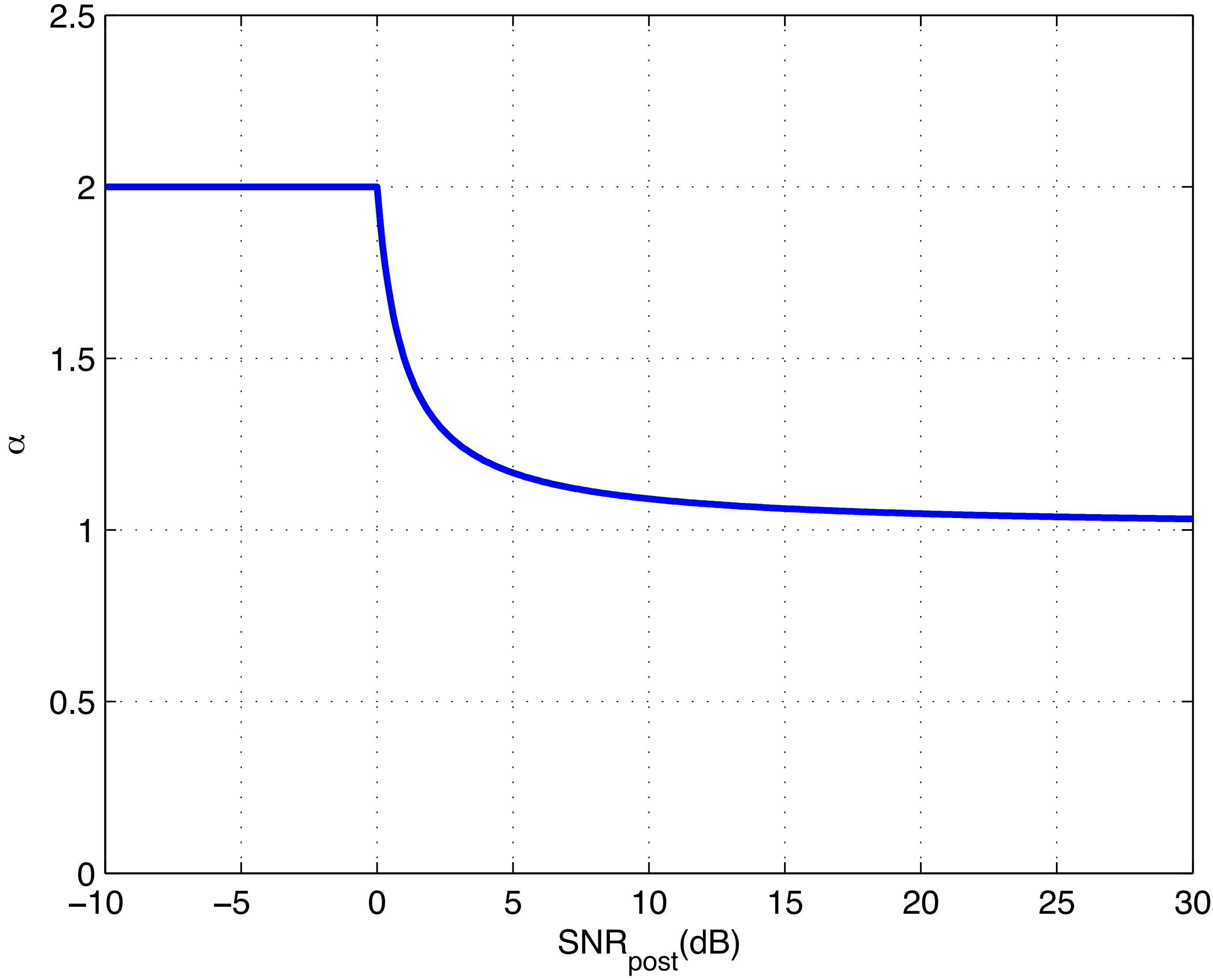
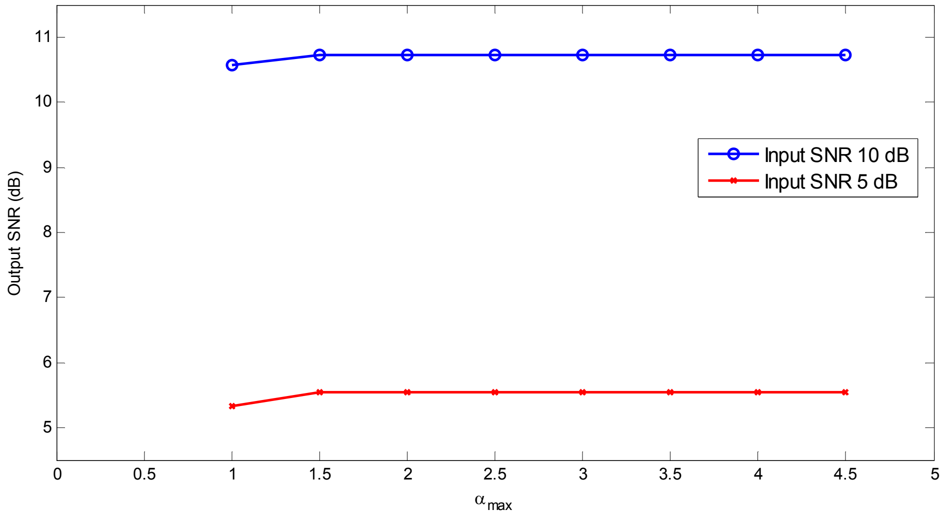
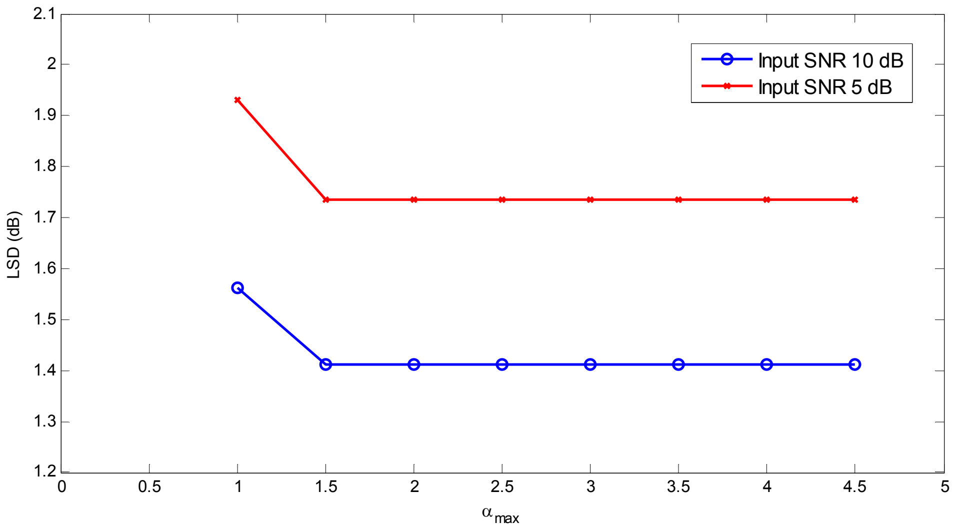


Figure 6



Output SNR (dB)
11
10
9
8
7
6
5
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
α_{max}
Input SNR 10 dB
Input SNR 5 dB

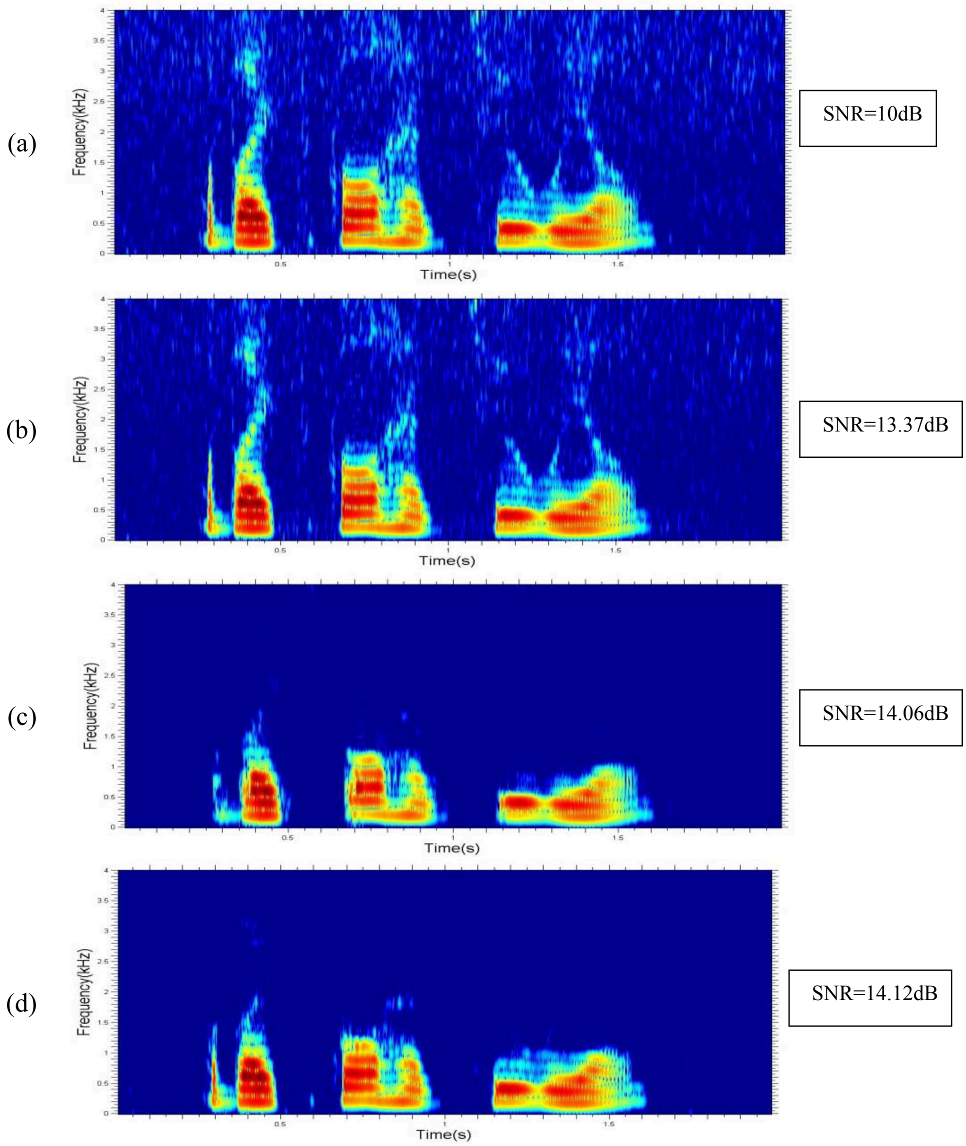
(a)



LSD (dB)
2.1
2
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
α_{max}
Input SNR 10 dB
Input SNR 5 dB

(b)

Figure 7



(a)
Frequency(kHz)
Time(s)
SNR=10dB
(b)
Frequency(kHz)
Time(s)
SNR=13.37dB
(c)
Frequency(kHz)
Time(s)
SNR=14.06dB
(d)
Frequency(kHz)
Time(s)
SNR=14.12dB

Figure 8

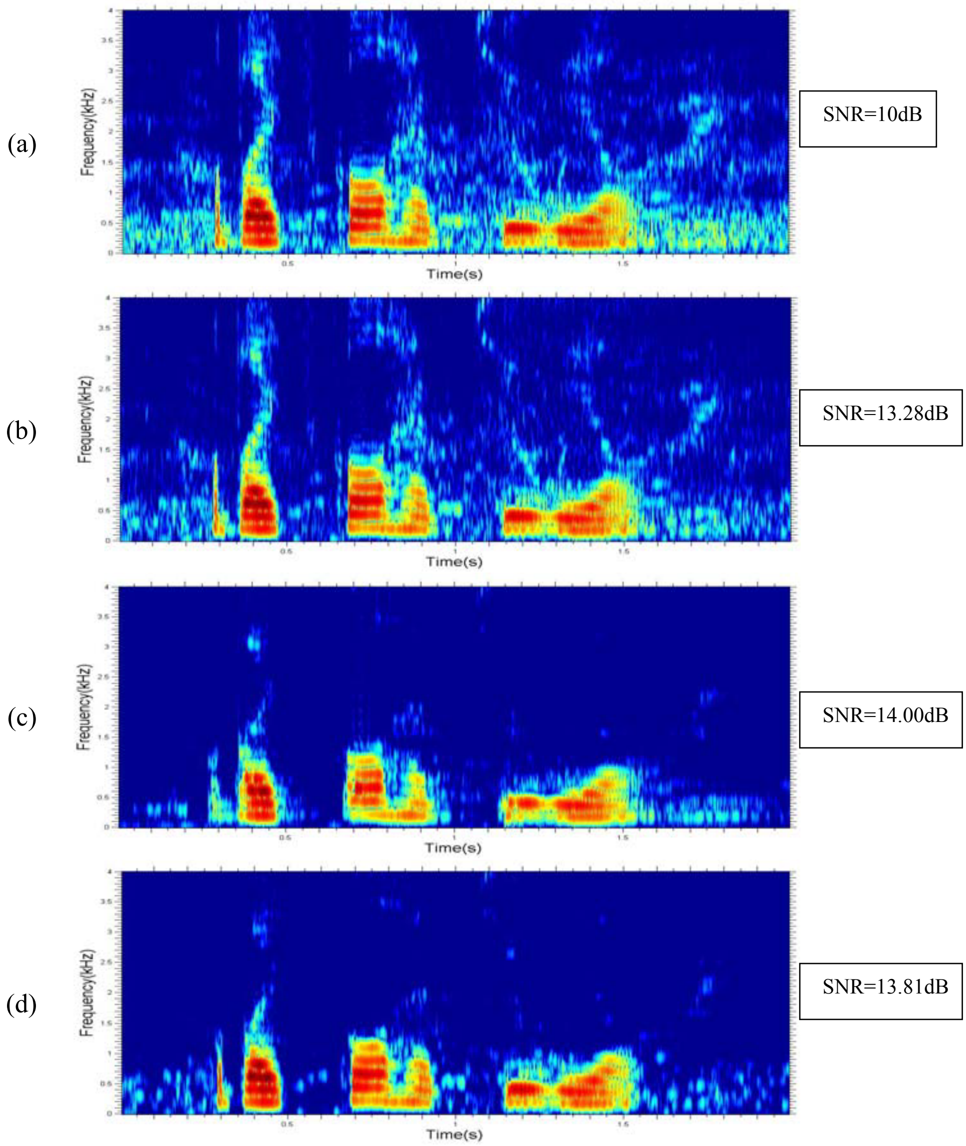

(a)
SNR=10dB
(b)
SNR=13.28dB
(c)
SNR=14.00dB
(d)
SNR=13.81dB
Frequency(kHz)
Time(s)

Table 1

Types of Noise	Input SNR (dB)	Output SNR (dB)		
		α	$\alpha_{\mathrm{OS}}(k,l)$	proposed $\alpha(k,l)$
WGN	5	6.21	8.55	**8.60**
	10	13.37	13.57	**13.62**
	15	18.62	18.83	**18.92**
	20	**23.87**	22.91	23.51
Babble	5	5.59	7.84	**7.92**
	10	13.28	13.46	**13.50**
	15	17.01	18.29	**18.98**
	20	22.30	23.51	**23.59**

Table 2

Types of Noise	Input SNR (dB)	LSD (dB)		
		α	$\alpha_{\mathrm{OS}}(k,l)$	proposed $\alpha(k,l)$
WGN	5	1.90	1.89	**1.85**
	10	1.45	1.41	**1.40**
	15	1.41	1.38	**1.35**
	20	0.99	0.95	**0.97**
Babble	5	2.10	2.09	**1.76**
	10	1.78	1.60	**1.39**
	15	1.66	1.54	**1.27**
	20	1.37	1.21	**1.02**

Table 3

Types of Noise	Input SNR (dB)	Output SNR (dB)			
		SS + $\alpha_{OS}(k,l)$	SS + proposed $\alpha(k,l)$	MCSS	Subband + proposed $\alpha(k,l)$
WGN	5	8.55	8.60	9.01	**9.12**
	10	13.57	13.62	14.06	**14.12**
	15	18.83	18.92	19.85	**19.92**
	20	22.91	23.51	24.00	**24.13**
Babble	5	7.84	7.92	**8.15**	8.01
	10	13.46	13.50	**14.00**	13.81
	15	18.29	18.98	19.01	**19.15**
	20	23.51	23.59	**24.68**	24.59
Factory	5	7.88	8.01	**8.75**	8.42
	10	13.50	13.66	**14.52**	14.30
	15	18.92	19.00	**19.86**	19.63
	20	24.21	24.38	**25.20**	25.01
Pink	5	8.90	9.00	9.32	**9.46**
	10	14.03	14.15	15.30	**15.41**
	15	19.21	19.30	19.98	**20.00**
	20	23.54	23.69	**24.02**	23.99
F16 cockpit	5	7.64	7.72	8.45	**8.53**
	10	13.00	13.21	14.60	**14.73**
	15	16.95	17.40	17.78	**17.80**
	20	22.32	22.52	23.07	**23.14**

Table 4

Types of Noise	Input SNR (dB)	LSD (dB)			
		SS + $\alpha_{\text{OS}}(k,l)$	SS + proposed $\alpha(k,l)$	MCSS	Subband + proposed $\alpha(k,l)$
WGN	5	1.89	1.85	1.77	**1.54**
	10	1.41	1.40	1.39	**1.33**
	15	1.38	1.35	1.31	**1.27**
	20	0.95	0.97	0.52	**0.34**
Babble	5	2.09	**1.76**	1.81	1.77
	10	1.60	**1.39**	1.55	1.45
	15	1.54	**1.27**	1.32	1.29
	20	1.21	**1.02**	1.26	1.17
Factory	5	2.57	**2.24**	2.49	2.32
	10	2.50	**2.25**	2.51	2.44
	15	1.98	**1.62**	1.80	1.77
	20	1.53	**1.24**	1.40	1.33
Pink	5	1.71	1.60	1.55	**1.37**
	10	1.51	1.49	1.38	**1.10**
	15	1.29	1.26	1.14	**1.00**
	20	0.14	0.13	0.09	**0.07**
F16 cockpit	5	2.51	2.43	2.30	**2.18**
	10	2.00	1.89	1.31	**1.19**
	15	1.30	1.28	1.15	**1.10**
	20	1.01	0.99	0.89	**0.53**

Table 5

Signals	PESQ scores				
	WGN	Babble noise	Factory noise	Pink noise	F16 cockpit noise
Noisy speech signal	2.06	2.60	2.50	2.31	2.43
SS method (fixed α)	2.25	2.47	2.43	2.40	2.44
SS method + variable $\alpha_{\mathrm{OS}}(k,l)$ in [7]	**3.10**	3.28	3.15	3.16	3.22
SS method + proposed $\alpha(k,l)$	2.94	2.95	3.06	3.04	3.01
MCSS technique	2.59	2.93	2.78	2.95	3.00
The proposed two-microphone subband NR + proposed $\alpha(k,l)$	2.95	**3.31**	**3.39**	**3.35**	**3.38**

Table 6

Signals	MOS scores				
	WGN	Babble noise	Factory noise	Pink noise	F16 cockpit noise
Noisy speech signal	2.10	2.47	3.15	2.90	3.30
SS method (fixed α)	2.30	2.75	1.75	2.40	2.60
SS method + variable $\alpha_{\mathrm{OS}}(k,l)$ in [7]	3.10	2.90	3.20	3.05	3.75
SS method + proposed $\alpha(k,l)$	2.30	2.30	**3.85**	3.80	3.05
MCSS technique	2.75	2.60	1.60	2.90	3.40
The proposed two-microphone subband NR + proposed $\alpha(k,l)$	**3.90**	**3.05**	3.25	**3.85**	**3.80**