Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Speech Communication xxx (2015) xxx–xxx

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Generalized *maximum a posteriori* spectral amplitude estimation for speech enhancement

Yu Tsao *, Ying-Hui Lai [1]

Research Center for Information Technology Innovation (CITI), Academia Sinica, No 128, Academia Road, Section 2, Nankang, Taipei 11529, Taiwan

Received 15 June 2015; received in revised form 20 September 2015; accepted 9 October 2015

Abstract

Spectral restoration methods for speech enhancement aim to remove noise components in noisy speech signals by using a gain function in the spectral domain. How to design the gain function is one of the most important parts for obtaining enhanced speech with good quality. In most studies, the gain function is designed by optimizing a criterion based on some assumptions of the noise and speech distributions, such as minimum mean square error (MMSE), maximum likelihood (ML), and maximum *a posteriori* (MAP) criteria. The MAP criterion shows advantage in obtaining a more reliable gain function by incorporating a suitable prior density. However, it has a problem as several studies showed: although MAP based estimator effectively reduces noise components when the signal-to-noise ratio (SNR) is low, it brings large speech distortion when the SNR is high. For solving this problem, we have proposed a generalized maximum *a posteriori* spectral amplitude (GMAPA) algorithm in designing a gain function for speech enhancement. The proposed GMAPA algorithm dynamically specifies the weight of prior density of speech spectra according to the SNR of the testing speech signals to calculate the optimal gain function. When the SNR is high, GMAPA adopts a small weight to prevent overcompensations that may result in speech distortions. On the other hand, when the SNR is low, GMAPA uses a large weight to avoid disturbance of the restoration caused by measurement noises. In our previous study, it has been proven that the weight of the prior density plays a crucial role to the GMAPA performance, and the weight is determined based on the SNR in an utterance-level. In this paper, we propose to compute the weight with the consideration of time–frequency correlations that result in a more accurate estimation of the gain function. Experiments were carried out to evaluate the proposed algorithm on both objective tests and subjective tests. The experimental results obtained from objective tests indicate that GMAPA is promising compared to several well-known algorithms at both high and low SNRs. The results of subjective listening tests indicate that GMAPA provides significantly higher sound quality than other speech enhancement algorithms.

Keywords: Speech enhancement; Spectral restoration; Generalized MAPA; GMAPA

1. Introduction

The corruption of speech signals by background noise may seriously degrade the speech quality and thus restrict the applicability of speech technology. The goal of speech enhancement is to reduce noise components and thus enhance the signal-to-noise ratio (SNR), intelligibility, and perceptual quality of degraded speech. A speech enhancement method usually serves as a preprocessor in many speech techniques. For example, speech coding systems often apply speech enhancement processes to increase the quality and intelligibility for speech communication (Lim and Oppenheim, 1979; Li et al., 2011a, 2011b). Hearing aids adopt speech enhancement methods to improve speech intelligibility for hearing-loss individuals in noisy environments (Venema, 2006; Levitt, 2001; Lai et al., 2013a).

* Corresponding author. Tel.: +886 2 2787 2390; fax: +886 2 2787 2315.
E-mail addresses: yu.tsao@citi.sinica.edu.tw (Y. Tsao), jackylai@citi.sinica.edu.tw (Y.-H. Lai).

[1] Tel.: +886 2 2787 2301; fax: +886 2 2787 2315.

http://dx.doi.org/10.1016/j.specom.2015.10.003
0167-6393/© 2015 Published by Elsevier B.V.

Various speech enhancement approaches have been proposed. A notable class is spectral restoration (Chen, 2008). Spectral restoration approaches aim to estimate a gain function for performing noise reduction in the frequency domain in order to obtain a clean speech spectrum from a noisy speech spectrum. Well-known spectral restoration approaches include spectral subtraction (SS) (Boll, 1979) and Wiener filter (Scalart and Filho, 1996) with their various extensions (Lu and Loizou, 2008; Li et al., 2008; Mittal and Phamdo, 2000). Additionally, some spectral restoration approaches are derived based on probabilistic models of speech and noise signals. Successful examples include minimum mean-square-error (MMSE) spectral estimator (Ephraim and Malah, 1984; Soon et al., 1999; Martin, 2005; Hansen et al., 2006; Malah et al., 1999; Cohen, 2002), maximum *a posteriori* spectral amplitude (MAPA) estimator (Plourde and Champagne, 2008; Lotter and Vary, 2005; Suhadi et al., 2011; Li et al., 2006; Xin et al., 2008), and maximum likelihood spectral amplitude (MLSA) estimator (McAulay and Malpass, 1980; Kjems and Jensen, 2012).

In probabilistic model based estimation, the maximum *a posteriori* (MAP) based criterion explicitly takes a certain prior distribution of signal in modeling, which results in a much more accurate estimation than that without taking the prior knowledge of signal distributions. The MAP criterion has been widely adopted to many speech processing tasks, such as acoustic model adaptation (Gauvain and Lee, 1994; Tsao et al., 2014a), feature compensation (Tsao et al., 2014b), voice conversion (Chen et al., 2003), and language model adaptation (Federico, 1996). When the available adaptation data is especially limited, ML-based approaches may encounter over-fitting issues and accordingly provide poor classification or regression performances. By incorporating the prior information, MAP-based approaches can effectively alleviate over-fitting to yield satisfactory performance. Meanwhile, specifying a suitable weight between likelihood and prior density according to the target task plays an important role to the achievable performance for the MAP criterion. For example, in automatic speech recognition, a suitable weight is favorable when combining scores from acoustic model and language model (Bahl et al., 1980; Ogawa et al., 1998). Moreover, it has confirmed effective to select an optimal regularization weight to control the residual and the stabilizing terms for a robust restoration of multichannel images (Zervakis, 1996). For the speech enhancement task, the ML criterion enhances speech signals based on the current data; however, due to variations caused by noise, the ML criterion may not be the optimal solution. On the other hand, the MAP criterion can compute more reliable gain function by incorporating a suitable prior density when the SNR is low. However, it is noted that the MAP criterion brings large speech distortion when the SNR is high. For solving this problem, we have proposed a generalized maximum *a posteriori* spectral amplitude (GMAPA) algorithm in designing a gain function for speech enhancement (Su et al., 2013; Lai et al., 2015). In the proposed GMAPA algorithm, the suitable weight of prior density is determined, based on the SNR of noisy speech, to calculate the gain function.

Several previous studies also intend to online update the gain function using the gamma modeling (Dat et al., 2006) and the multidimensional normal inverse Gaussian (MNIG) modeling (Hendriks and Martin, 2007). The distinct point of the GMAPA algorithm is that a regression scheme is directly applied to map the SNR to the weight of prior density. When the SNR is low, a large weight is obtained and used to avoid large residual noise in restoration. On the other hand, a small weight is obtained and used to prevent speech distortion when the SNR is high. In the previous study (Su et al., 2013), we confirmed that the weight of prior density is a crucial factor to the noise reduction capability of GMAPA, and the weight is computed based on the SNR of the whole utterance. As it is known, the statistical regularity of speech shows strong correlation or coherence between neighboring frequency bins and time frames (Lu et al., 2011, 2010; Fan et al., 2014). Based on these correlations, rich spectral–temporal structure can be explored in local frequency and time-varying spectral patches and used to improve the speech enhancement capability. In this study, we propose to further improve the GMAPA performance by incorporating such spectral–temporal information.

The present study expands upon our previous GMAPA study (Su et al., 2013) with three additional contributions. The first one is that we detail the derivations of the GMAPA algorithm and the procedures for estimating the mapping function that determines the weight of prior density. The second contribution is that we extend the original GMAPA algorithm with the consideration of spectral–temporal coherence, which results in more accurate estimation of gain function. To incorporate temporal and spectral characteristics, we proposed temporal grouping (TG) and spectral grouping (SG) techniques, respectively. When combining the TG and SG techniques, GMAPA effectively reduces the processing latency while provides better sound quality. The third contribution is that thorough experiments were carried out including objective and subjective tests to evaluate the capability of the GMAPA algorithm in more detail. For the objective evaluations we compared the speech distortion index (SDI) (Chen et al., 2006), perceptual estimation of speech quality (PESQ) (ITU-T, 2001; Rix et al., 2001; Hu and Loizou, 2008), and segmental SNR improvement (SSNRI) (Chen, 2008) using utterances from the Aurora-4 task (Hirsch and Pearce, 2000; Parihar et al., 2004). The subjective listening test was conducted with a single-blind design using a standardized Mandarin speech database (Lai et al., 2013b). The experimental results confirm the effectiveness of the proposed GMAPA algorithm relative to several well-known speech enhancement algorithms.

The remainder of this paper is organized as follows: Section 2 introduces the spectral restoration process and reviews several well-known probabilistic-model-based algorithms. Section 3 presents the GMAPA algorithm and TG and SG techniques. Section 4 reports our experimental setup and results. The conclusions are drawn in Section 5.

2. Spectral restoration techniques

This section reviews the overall spectral restoration process and three probabilistic-model-based algorithms, namely MMSE, MLSA, and MAPA estimators.

2.1. Spectral analysis

In the time domain, a noisy speech signal $y[n]$ is composed of a clean speech signal $s[n]$ corrupted by an additive noise signal $v[n]$:

$$y[n] = s[n] + v[n], \tag{1}$$

where n denotes the sampling time index. We assumed that the clean speech and noise signals are independent and additive, and that both of them are random processes. In the spectral domain, the noisy speech spectrum of the m-th frame, $Y[m,l]$, can be expressed as

$$Y[m,l] = S[m,l] + V[m,l], 0 \leqslant l \leqslant L-1, \tag{2}$$

where l is the frequency bin corresponding to frequency ω_l, where $\omega_l = 2\pi l/L$, $l = 0, 1, \ldots, L-1$; $S[m,l]$ and $V[m,l]$ are the speech and noise spectra, respectively.

Fig. 1 shows the overall speech enhancement system, which can be decomposed into noise tracking and gain estimation. The noise-tracking stage computes the noise power spectral density (PSD) from the noisy speech, $Y[m,l]$, to obtain *a priori* SNR and *a posteriori* SNR statistics (Cohen, 2002, 2003; Kum and Chang, 2009). Then gain estimation involves calculating a gain function, $G[m,l]$, based on the computed *a priori* SNR and *a posteriori* SNR statistics. Finally, the enhanced speech, $\widehat{S}[m,l]$, is obtained by filtering $Y[m,l]$ through $G[m,l]$. For the sake of brevity, in the following discussion $Y[m,l]$, $S[m,l]$, $V[m,l]$, and $G[m,l]$ are denoted as Y, S, V, and G, respectively.

By decomposing the noisy and clean speech spectra (i.e., Y and S in Eq. (2)) into their amplitude and phase components, we have

$$Y = Y_k \exp(j\theta_Y), \tag{3}$$

$$S = S_k \exp(j\theta_S), \tag{4}$$

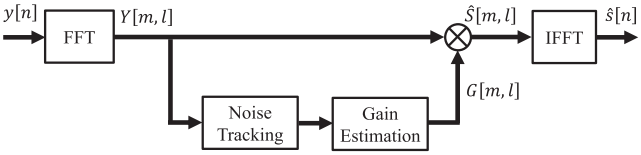

Fig. 1. Block diagram of a speech enhancement system.

where $Y_k = |Y|$ and $S_k = |S|$ are the amplitudes, and $\theta_Y = \angle Y$ and $\theta_S = \angle S$ are the phases of Y and S, respectively. In the following discussion, subscript k denotes the amplitude part of a signal. To reconstruct S from Y, we first estimated the phase of the clean speech spectrum as

$$\exp(j\widehat{\theta}_S) = \arg \min_{\exp(j\widehat{\theta}_S^*)} E\left[\left|\exp(j\theta_S) - \exp(j\widehat{\theta}_S^*)\right|^2\right]. \tag{5}$$

The prior density of θ_S is modeled by a uniform distribution

$$p(\theta_S) = \frac{1}{2\pi}, \tag{6}$$

over $(-\pi, \pi)$ (Chen, 2008). From Eqs. (5) and (6), we then have

$$\exp(j\widehat{\theta}_S) = \exp(j\theta_Y). \tag{7}$$

Full details of the phase estimation process can be found elsewhere (Chen, 2008). Accordingly, the clean speech spectrum is estimated as

$$\widehat{S} = \widehat{S}_k \exp(j\theta_Y) = G \cdot Y, \tag{8}$$

The goal of spectral restoration is to compute the gain function G.

2.2. Related work

This section introduces three well-known methods of gain estimation: MMSE, MLSA, and MAPA. The methods used to calculate the noise power and gain function for these algorithms are derived based on two assumptions: (i) both the speech and noise signals are random processes; (ii) the speech and noise signals are independent, and the noise signal is additive. Two statistics for SNR, namely *a priori* SNR ξ_k and *a posteriori* SNR γ_k, are defined as $\xi_k = \sigma_s^2/\sigma_v^2$ and $\gamma_k = Y_k^2/\sigma_v^2$, where $\sigma_s^2 = E[|S|^2]$ and $\sigma_v^2 = E[|V|^2]$. ξ_k and γ_k are denoted as ξ and γ, respectively, in the discussion below for the sake of simplicity.

2.2.1. General formulations

By assuming that both the clean speech and noise spectra are modeled by Gaussian distributions, the conditional probability density function (PDF), $p(Y|S_k, \theta_S)$, can be derived as

$$p(Y|S_k, \theta_S) = \frac{1}{\pi\sigma_v^2} \exp\left(-\frac{|V|^2}{\sigma_v^2}\right). \tag{9}$$

The amplitude and phase components of complex Gaussian random variables with zero mean are statistically independent (Chen, 2008). Thus, $p(S_k, \theta_S)$ becomes

$$p(S_k, \theta_S) = p(S_k) \cdot p(\theta_S), \tag{10}$$

where $p(S_k)$ is modeled by the Rayleigh distribution

$$p(S_k) = \frac{2S_k}{\sigma_s^2} \exp\left(-\frac{S_k^2}{\sigma_s^2}\right), \tag{11}$$

where σ_s^2 denotes the hyper-parameter in the density.

2.2.2. *MMSE algorithm*

The spectral amplitude of the MMSE estimator is given by the conditional expectation

$$\widehat{S}_k = E[S_k|Y] = \int_0^\infty S_k p(S_k|Y) dS_k$$
$$= \frac{\int_0^\infty \int_0^\pi S_k p(Y|S_k,\theta_S) p(S_k,\theta_S) d\theta_S dS_k}{\int_0^\infty \int_0^\pi p(Y|S_k,\theta_S) p(S_k,\theta_S) d\theta_S dS_k}. \quad (12)$$

By substituting Eqs. (9) and (10) into Eq. (12) combined with some derivations, the MMSE-based gain function, G_{MMSE}, can be expressed as

$$G_{MMSE} = \Gamma\left(\frac{3}{2}\right) \frac{\sqrt{\delta}}{\gamma} \exp\left(\frac{-\delta}{2}\right) \left[(1+\delta) I_0\left(\frac{\delta}{2}\right) + \delta I_1\left(\frac{\delta}{2}\right)\right], \quad (13)$$

where $\delta = [\xi/(1+\xi)]\gamma$; $\Gamma(\cdot)$, $I_0(\cdot)$, and $I_1(\cdot)$ denote the gamma function, zero-order modified Bessel function, and first-order modified Bessel function, respectively. The enhanced speech spectrum for the MMSE estimator can then be estimated by $\widehat{S} = G_{MMSE} \cdot Y$.

2.2.3. *MLSA algorithm*

The MLSA estimator computes the spectral amplitude, $\widehat{S}_k$, as

$$\widehat{S}_k = \arg\max_{S_k} J_{MLSA}(S_k), \quad (14)$$

where $J_{MLSA}(S_k)$ is the MLSA cost function and is defined as

$$J_{MLSA}(S_k) = \ln\{p(Y|S_k)\}. \quad (15)$$

By differentiating the log-likelihood function of Eq. (15) with respect to S_k, and equating the result to zero, we can obtain the gain function of the MLSA estimator:

$$G_{MLSA} = \frac{1 + \sqrt{1 - (1/\gamma)}}{2}. \quad (16)$$

With the estimated G_{MLSA}, the enhanced speech spectrum for the MLSA estimator can be computed by $\widehat{S} = G_{MLSA} \cdot Y$.

2.2.4. *MAPA algorithm*

The MAPA estimator computes the spectral amplitude, $\widehat{S}_k$, as

$$\widehat{S}_k = \arg\max_{S_k} J_{MAPA}(S_k). \quad (17)$$

$J_{MAPA}(S_k)$ is the MAPA cost function and can be expressed as

$$J_{MAPA}(S_k) = \ln\{p(Y|S_k)p(S_k)\}. \quad (18)$$

By differentiating the MAPA cost function of Eq. (18) with respect to S_k, and equating the result to zero, we can obtain the MAPA-based gain function, G_{MAPA}, as

$$G_{MAPA} = \frac{\xi + \sqrt{\xi^2 + (1+\xi)\xi/\gamma}}{2(1+\xi)}. \quad (19)$$

The enhanced speech spectrum for the MAPA estimator can be computed as $\widehat{S} = G_{MAPA} \cdot Y$.

3. The GMAPA algorithm

In this section, we first introduce the proposed GMAPA algorithm and the mapping function to determine the weight of prior density. Then we present the TG and SG techniques, which are used to prepare suitable temporal and spectral information to facilitate GMAPA better noise reduction capability.

3.1. *GMAPA algorithm*

The main difference between the MLSA and MAPA estimators is in the prior density, $p(S_k)$. For the GMAPA estimator, the spectral amplitude, $\widehat{S}_k$, is calculated in the same way:

$$\widehat{S}_k = \arg\max_{S_k} J_{GMAPA}(S_k), \quad (20)$$

where $J_{GMAPA}(S_k)$ is the GMAPA cost function, which is defined as

$$J_{GMAPA}(S_k) = \ln\{p(Y|S_k)[p(S_k)]^\alpha\}, \quad (21)$$

where α is the weight of prior density. Please note that $J_{GMAPA}(S_k)$ becomes $J_{MAPA}(S_k)$ in Eq. (18) when setting $\alpha = 1$ in Eq. (21), while it becomes $J_{MLSA}(S_k)$ in Eq. (15) when setting $\alpha = 0$ in Eq. (21).

By differentiating the GMAPA cost function of Eq. (21) with respect to S_k, and equating the result to zero, we obtain the GMAPA-based gain function as

$$G_{GMAPA} = \frac{\xi + \sqrt{\xi^2 + (2\alpha - 1)(\alpha + \xi)\xi/\gamma}}{2(\alpha + \xi)}. \quad (22)$$

The enhanced speech spectrum for the GMAPA estimator thus can be obtained by $\widehat{S} = G_{GMAPA} \cdot Y$. The derivation of the GMAPA gain function in Eq. (22) is provided in more detail in Appendix A.

Notably in Eq. (21), α controls the weight between likelihood $p(Y|S_k)$ and prior density $p(S_k)$. The likelihood part yields the restoration solution based on the current speech data, whereas the prior density indicates a regularization on the solution. For high SNR conditions, the noise-tracking stage in Fig. 1 can estimate noise statistics accurately, and thus spectral restoration gain function computed based on the likelihood without any regularization can achieve satisfactory performance. However in low SNR conditions, the huge variations caused by noise may cause noise-tracking estimator to perform poorly, and thus an unsatisfactory gain function estimation might be incurred if only considering the likelihood part in Eq. (21). This issue can be handled by incorporating

Please cite this article in press as: Tsao, Y., Lai, Y.-H., Generalized *maximum a posteriori* spectral amplitude estimation for speech enhancement, Speech Comm. (2015), http://dx.doi.org/10.1016/j.specom.2015.10.003

suitable prior information. The statistics used to prepare the prior density can be estimated using the training data in the offline phase (Gauvain and Lee, 1994). In this study, the prior density is estimated by the previous speech statistics and updated sequentially. Therefore, the prior density in Eq. (21) actually acts as a smoothness (or stabilizing) constraint. A large weight results in a strong smoothness constraint while a small weight induces a weak smoothness constraint on the enhancement processing. In view of the foregoing considerations, a larger α should be used in low SNR conditions to avoid disturbance caused by measurement noises, and a smaller α should be used in high SNR conditions to prevent overcompensations that may result in speech distortions.

3.2. Determining the weight of prior density

If the SNR of the testing condition is known (e.g., by using an additional mechanism to predict SNR, or when the environment is familiar), α in Eq. (21) can be predefined directly. However, the optimal α for the gain function needs to be calculated when applying GMAPA in an unknown environment. Since α is dependent on the SNR of the testing condition, the training data can be used to find a mapping function that determines the correlation between the optimal α and the SNR of the test utterance. In this study we used a sigmoid function (Alippi and Storti-Gajani, 1991; Zhang et al., 1996) to optimally determine the weight α for G_{GMAPA} in Eq. (21) according to

$$\alpha = \frac{\alpha_{max}}{1 + \exp[-b(\bar{\gamma} - c)]}, \tag{23}$$

where α_{max} is the maximum value (upper bound) for α, b and c are coefficients of the sigmoid function, and $\bar{\gamma}$ is the mean of the *a posteriori* SNR for a given utterance, where $\bar{\gamma} = 1/T\sum_{t=1}^{T}\gamma^{(t)}$, where T is the total number of frames for this utterance, and $\gamma^{(t)}$ is the *a posteriori* SNR for the t-th frame. We use the training data to determine $\{\alpha_{max}, b, c\}$. Fig. 2 shows the designed function for determining the value of weight α from $\bar{\gamma}$, and indicates that α is larger when the SNR is lower. For ease of presentation, we denote the mapping function of Eq. (23) as STW (SNR-to-weight) function. In Section 4.2, we will confirm that the STW function is a proper choice to model α and introduce the criterion that we used to determine $\{\alpha_{max}, b, c\}$.

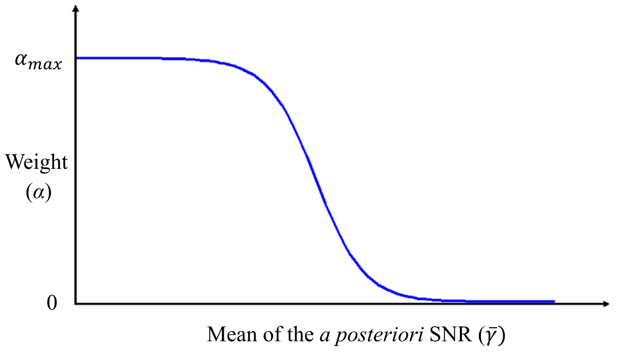

Fig. 2. STW function of α and mean of the *a posteriori* SNR.

3.3. Incorporating the temporal and spectral information

The original GMAPA algorithm computes a gain function for each frame and frequency bin for spectral restoration, while the same weight of prior density, α, is used throughout the utterance. In this study, we extend the original GMAPA by incorporating temporal and spectral information to attain better enhancement capability. To effectively utilize the temporal and spectral information, we derived TG and SG techniques. Fig. 3 shows these two techniques applied on one speech utterance. In the following, we will detail these two techniques and the integration of them with GMAPA.

3.3.1. GMAPA with TG

Speech patterns, such as phones, syllables, or words, usually occupy a set of consecutive frames. The objective of TG is to exploit such information by using a temporal window to embrace the neighboring spectra and to capture the invariant structure of speech signals. As shown in Fig. 3, the TG technique segments the speech utterance by applying temporal windows, each containing M_c frames. Hereafter, GMAPA with TG is termed GMAPA + TG. In original GMAPA, we need to compute the utterance-level *a posteriori* SNR, $\bar{\gamma}$ to determine α in Eq. (23) for computing the gain function for spectral restoration. For GMAPA + TG, each temporal window determines a distinct weight of the prior density, α_c, $c = 1, \ldots, N_{TC}$, where N_{TC} is the total number of temporal windows. In this way, the local temporal information could be incorporated in the gain function estimation, enabling GMAPA to achieve better noise reduction performance. Moreover, the utterance-based SNR calculation used in original GMAPA will slow down the online enhancement process. Since GMAPA + TG updates α_c in a window-wise manner, the processing latency can be effectively reduced. When performing speech enhancement, GMAPA + TG uses a sliding-window with M_c frames. For the first M_c frames ($t \leqq M_c$) we set $\alpha = 1$ in Eq. (22) and keep computing $\gamma^{(t)}$. When $t \geqq M_c + 1$, we then compute the average *a posteriori* SNR for the t-th frame as $\bar{\gamma}^{(t)} = \beta \cdot (1/M_c\sum_{k=t-M_c+1}^{k=t}\gamma^{(k)}) + (1 - \beta) \cdot \bar{\bar{\gamma}}$, where $\bar{\bar{\gamma}}$ is the average *a posteriori* SNR of the previous sliding-windows, and β is a coefficient. Then GMAPA + TG determines the optimal α_c based on Eq. (23) with the computed $\bar{\gamma}^{(t)}$. In this way, both the computation of $\bar{\gamma}^{(t)}$ and the determination of α_c are performed in the real-time phase, thus reducing the processing latency.

3.3.2. GMAPA with SG

The objective of SG is to decompose the full-band spectra into several sub-bands for a better characterization of local spectral structure. In Fig. 3, we prepare N_{SC} sub-bands from the original speech: $N_{SC} = Ceil(L/L_c)$, where

Please cite this article in press as: Tsao, Y., Lai, Y.-H., Generalized *maximum a posteriori* spectral amplitude estimation for speech enhancement, Speech Comm. (2015), http://dx.doi.org/10.1016/j.specom.2015.10.003

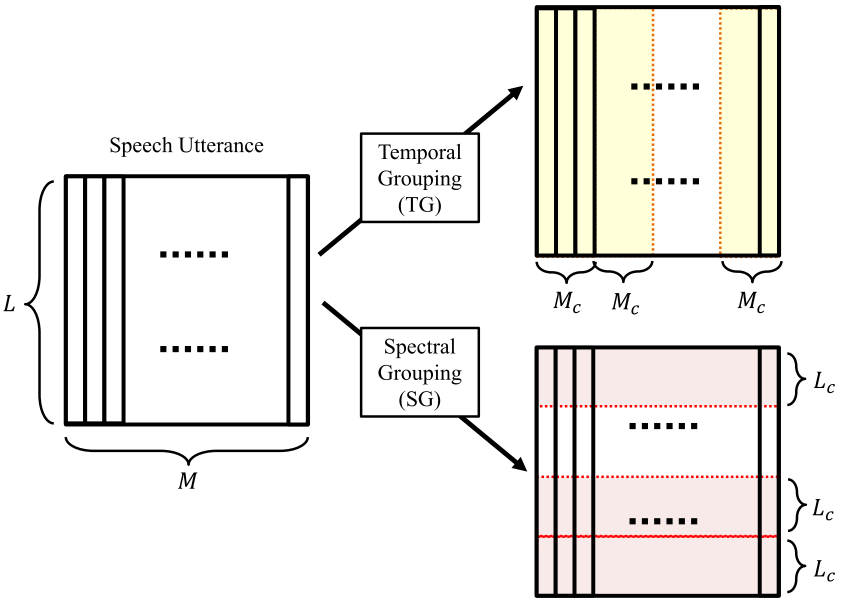

Fig. 3. The TG and SG techniques. M and L, respectively, are the numbers of frames and frequency bins in this utterance; M_c and L_c, respectively, are the numbers of frames and frequency bins in each temporal group and spectral group.

L_c and $Ceil(\cdot)$ denote the number of frequency bins in one sub-band and the ceiling function, respectively. For simplicity, GMAPA with SG is termed GMAPA + SG hereafter. In the offline phase, GMAPA + SG prepares N_{SC} STW functions (accordingly N_{SC} sets of $\{\alpha_{max}, b, c\}$). In the online phase, GMAPA + SG determines N_{SC} weights: α_c, $c = 1, \ldots, N_{SC}$ based on the prepared N_{SC} STW functions; then these N_{SC} weights are used to compute the GMAPA gain functions based on Eq. (22) for spectral restoration. When setting $L_c = L$ (namely, $N_{SC} = 1$), all of the frequency bins share the same weight (the same as that used in original GMAPA). On the other hand, when setting $L_c = 1$ (namely, $N_{SC} = L$), each frequency bin has a specific weight.

4. Experiments

This section first uses an example to show how to determine the optimal parameter set $\{\alpha_{max}, b, c\}$ in Eq. (23) and Fig. 2. The spectrograms for speech processed by different speech enhancement algorithms are then presented, followed by evaluation results from the objective and subjective experiments.

4.1. Speech enhancement system configuration

This section introduces the speech enhancement system configuration that was used for the performance evaluations in this study. In addition to the proposed GMAPA algorithm, we conducted experiments using three related spectral restoration algorithms for comparison: MMSE, MLSA, and MAPA algorithms. The corresponding gain functions are presented in Table 1.

Table 1
Gain functions of MMSE, MLSA, MAPA, and GMAPA algorithms.

MMSE	$\Gamma\left(\frac{3}{2}\right)\frac{\sqrt{\delta}}{\gamma}\exp\left(-\frac{\delta}{2}\right)\left[(1+\delta)I_0\left(\frac{\delta}{2}\right)+\delta I_1\left(\frac{\delta}{2}\right)\right]$
MLSA	$\frac{1+\sqrt{(Y_k^2-\sigma_v^2)/Y_k^2}}{2}$
MAPA	$\frac{\xi+\sqrt{\xi^2+(1+\xi)\xi/\gamma}}{2(1+\xi)}$
GMAPA	$\frac{\xi+\sqrt{\xi^2+(2\alpha-1)(\alpha+\xi)\xi/\gamma}}{2(\alpha+\xi)}$

$\delta = [\xi/(1+\xi)]\gamma$.

We also incorporated the speech-presence uncertainty to improve the speech enhancement performance (Scalart and Filho, 1996; Malah et al., 1999):

$$G_{cor} = \frac{\Lambda}{1+\Lambda} G, \tag{24}$$

where G_{cor} is the corrected gain function that was finally used to perform speech enhancement, and Λ is a simplified notation for $\Lambda[m,k]$, where:

$$\Lambda = \frac{(1-q)}{q}\frac{p(Y|H_1)}{p(Y|H_0)} = \frac{(1-q)}{q}\frac{\exp(\delta)}{1+\xi}, \tag{25}$$

where q is a simplified notation for $q[m,k]$, representing the speech absent probability (SAP) (Lotter and Vary, 2005; Choi and Kang, 2005), and $\delta = [\xi/(1+\xi)]\gamma$.

4.2. STW function estimation

In this section we introduce the procedure used to find the optimal parameter set $\{\alpha_{max}, b, c\}$ in Eq. (23). The same procedure was used for all of the evaluations in the experiments.

We first prepared a set of training data using 36 utterances made by 3 males and 3 females. These utterances were excerpted from the Aurora-4 (Hirsch and Pearce, 2000; Parihar et al., 2004) clean training set. Aurora-4 includes speech data at two sampling rates, 8 kHz and 16 kHz, and the 8 kHz data was selected in this study. The 36 utterances were used to artificially generate noisy speech signals containing white Gaussian noise (WGN), exhibition noise, and train station noise at six SNRs: 0, 5, 10, 15, and 20 SNRs; these noise types represent color noise, stationary noise, and non-stationary noises, respectively (Hirsch and Pearce, 2000). We prepared 36 utterances for each SNR. Next, we calculated the mean *a posteriori* SNR for each SNR using $\bar{\gamma}_u = 1/T_u \sum_{t=1}^{T_u} \gamma^{(t)}$, where T_u is the total number of frames of the utterances for the u-th SNR ($u = 0$, 5, 10, 15, and 20). Then, for each SNR, we computed the average *a posteriori* SNR over the 36 utterances. Table 2 lists the estimated means of *a posteriori* SNR $\bar{\gamma}$ for different input SNRs, where the minimum statistics (MS) method (Martin, 1994, 2001) was used as the noise estimator in this example. In Table 2, for input signals with SNRs of 0 and 20 dB, the estimated average $\bar{\gamma}_{0\ \mathrm{dB}}$ and $\bar{\gamma}_{20\ \mathrm{dB}}$ values were 17.73 and 22.76, respectively.

In this study we used the speech distortion index (SDI) (Chen et al., 2006) as the reference to obtain the optimal α for the GMAPA algorithm. The SDI corresponds to the ratio of the energies of the residual speech and clean speech signals:

$$\mathrm{SDI} = \frac{E[(s[n] - \hat{s}[n])^2]}{E[s^2[n]]}, \tag{26}$$

where $s[n]$ and $\hat{s}[n]$ are the clean and enhanced speech signals, respectively. A lower SDI indicates a smaller difference between the enhanced and clean speech signals. To estimate the STW function of Eq. (23), the same utterances used to obtain the *a posteriori* SNRs listed in Table 2 were first adopted to compute SDIs, with various α values at specific SNRs. The average SDI values for the GMAPA algorithm using different α for 0–20 dB SNRs over WGN, exhibition noise, and train station noise are listed in Fig. 4. For each specific SNR, the result of the lowest SDI is marked with "↓". From the figure, the optimal α values for input SNRs of 0 dB and 20 dB SNRs were 2.0 and 0.5, respectively. From Table 2 and Fig. 4, we confirm that the STW function in Eq. (23) is a proper choice.

Table 2
Mean *a posteriori* SNRs for 0–20 dB SNR conditions.

SNR	0 dB	5 dB	10 dB	15 dB	20 dB
$\bar{\gamma}$	17.73	18.84	19.99	21.45	22.76

The values of $\bar{\gamma}$ and optimal α collected in Table 2 and Fig. 4, respectively, are then used to calculate $\{\alpha_{max}, b, c\}$ in Eq. (23) using the following $\{\bar{\gamma}_i,\ \alpha_i\}$ pairs, $i = 1, 2, \ldots,$ I, where I denotes the number of training data sets, and $I = 5$ in the example of Table 2 and Fig. 4. The pairs are $\{17.73, 2.0\}$, $\{18.84, 2.0\}$, $\{19.99, 1.5\}$, $\{21.45, 1.0\}$, and $\{22.76, 0.5\}$. Then the optimal set $\{\hat{\alpha}_{max}, \hat{b}, \hat{c}\}$ is estimated by

$$\{\hat{\alpha}_{max}, \hat{b}, \hat{c}\} = \arg\min_{\alpha_{max},b,c} J_{CF}\ (\alpha_{max}, b, c|\bar{\gamma}_i, \alpha_i,\ i = 1, 2 \ldots, I) \tag{27}$$

where

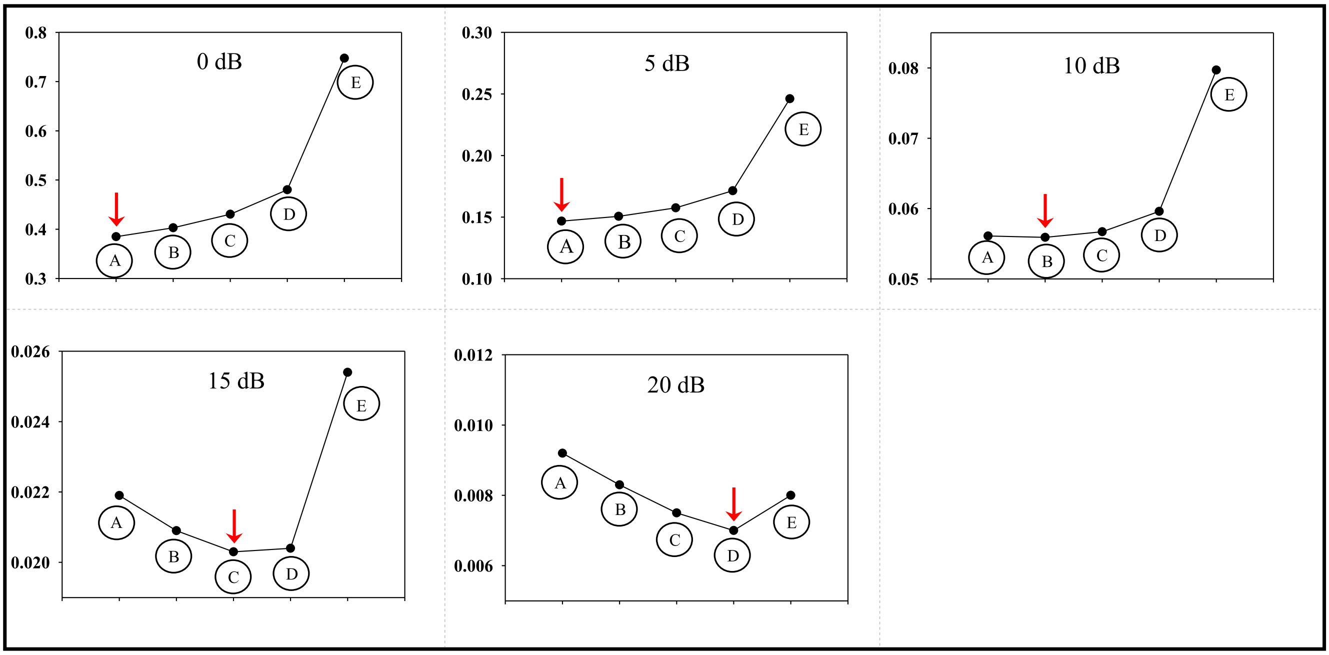

Fig. 4. Average SDI values (over WGN, exhibition noise, and train station noise at 0–20 dB SNRs) for the GMAPA algorithm using different α: For Ⓐ, Ⓑ, Ⓒ, Ⓓ, and Ⓔ, α = 2.0, 1.5, 1.0, 0.5, 0.0, respectively. For each SNR, the result of the lowest SDI is marked with "↓".

$$J_{CF}(\alpha_{max}, b, c|\bar{\gamma}_i, \alpha_i, i = 1, 2 \dots I) = \sum_{i=1}^{I} \left\| \frac{\alpha_{max}}{1 + \exp[-b(\bar{\gamma}_i - c)]} - \alpha_i \right\|^2. \tag{28}$$

The MMSE algorithm is adopted to compute $\{\alpha_{max}, b, c\}$ based on (27) and (28) (Cavallini, 1993). For the results listed in Table 2 and Fig. 4, we obtained $\{\alpha_{max}, b, c\} = \{2.195, -0.7787, 21.18\}$. The parameter set is predetermined using the available training data in the offline phase. When performing GMAPA speech enhancement in the online phase, we first calculate $\bar{\gamma}$ for the testing utterance and determine the corresponding α based on Eq. (23). The gain function in Eq. (22) is then computed, and finally speech enhancement is performed.

4.3. Spectrogram analyses

In this section we use visual presentations to compare different speech enhancement approaches. A spectrogram is often used to display how the frequencies present in a speech signal varying over time (Flanagan, 1972; Haykin, 1991). To investigate the speech enhancement capability of GMAPA, we prepared testing data contaminated by babble and car noises, which represent non-stationary and stationary noises, respectively (Hirsch and Pearce, 2000). Fig. 5(a) and (b) respectively present spectrograms of the clean speech and the noisy speech contaminated by babble noise at an SNR of 5 dB. The content of the speech was "*To Mr. Hawke that is as it should be*". The spectra of the enhanced speech obtained using the MLSA, MAPA, and GMAPA estimators are shown in Fig. 5(c), (d), and (e), respectively. Meanwhile, Fig. 6 shows the same five spectrogram plots as those in Fig. 5 when using the car noise.

Figs. 5 and 6 indicate that both the MLSA and MAPA algorithms can reduce the noise components, with MAPA exhibiting superior performance [compare panels (c) and (d) in both figures]. Moreover, the GMAPA algorithm showed an even greater noise reduction capability than the MAPA algorithm [compare panels (d) and (e) in both figures]. In the following, we will present the objective and subjective evaluation results to show the effectiveness of the proposed GMAPA approach.

4.4. Objective evaluations

In this section, we introduce the experimental setup and the results obtained in the objective evaluations.

4.4.1. Experimental setup

The objective evaluations were carried out using speech signals excerpted from the Aurora-4 task. We selected 36 utterances made by 6 speakers (3 males and 3 females) from the clean training set. The 6 speakers were different from those who pronounced the training data, and thus these 36 utterances were not used in the training set for optimizing $\{\alpha_{max}, b, c\}$ as introduced in Section 4.2. We intentionally selected utterances of various lengths, with around 8.3 s in average. With the 36 clean utterances, we generated two sets of noisy speech signals (babble and

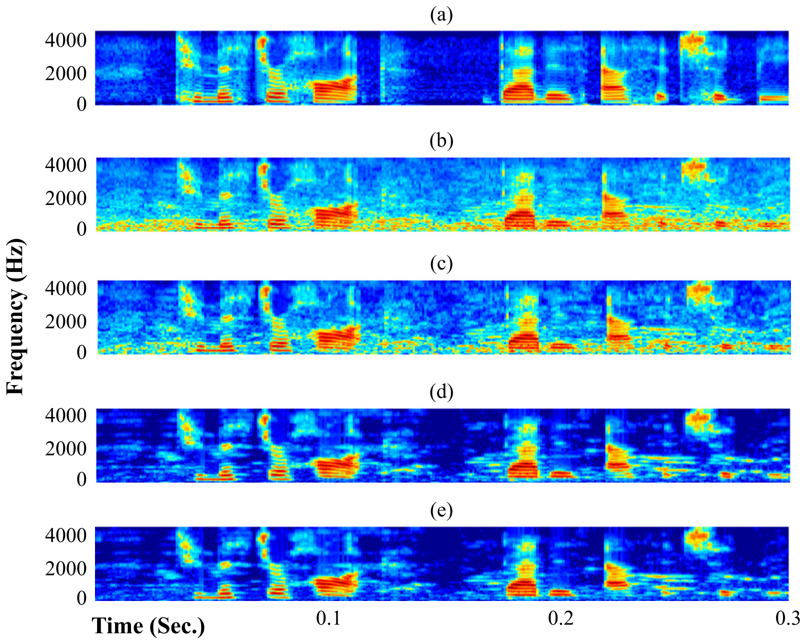

Fig. 5. Spectrograms for babble noise: (a) clean speech; (b) 5 dB SNR noisy speech; (c), (d), and (e) enhanced speech obtained using the MLSA, MAPA, and GMAPA algorithms, respectively.

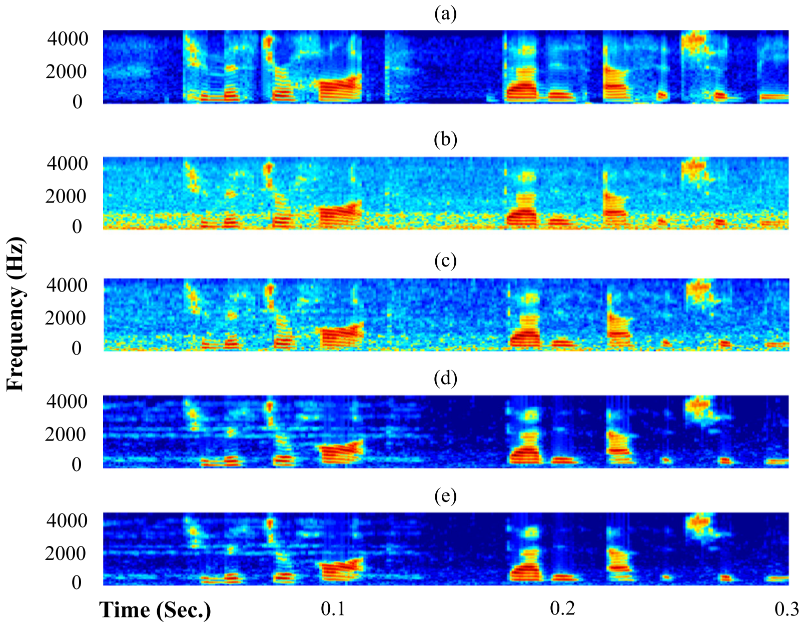

Fig. 6. Spectrograms for car noise: (a) clean speech; (b) 5 dB SNR noisy speech; (c), (d), and (e) enhanced speech obtained using the MLSA, MAPA, and GMAPA algorithms, respectively. (More samples: https://goo.gl/w86vpm.)

car) by contaminating the clean utterances at five SNRs (i.e., 0, 5, 10, 15, and 20 dB SNRs). Finally, we prepared 10 different conditions, each containing 36 utterances. In this study we used the PESQ, SSNRI, and SDI for objective evaluations.

The PESQ was proposed by the International Telecommunication Union (ITU) as a measure of the difference between enhanced and clean speech signals. PESQ scores range from −0.5 to 4.5, with a higher score indicating that the enhanced speech signal is closer to the clean speech signal (Loizou, 2013).

The SSNRI represents the difference in the segmental SNR between the enhanced speech and noisy speech:

$$\mathrm{SSNRI} = \mathrm{SSNR}_{enhanced} - \mathrm{SSNR}_{noisy}, \tag{29}$$

where $\mathrm{SSNR}_{enhanced}$ is the SSNR of enhanced speech:

$$\mathrm{SSNR}_{enhanced} = \frac{1}{M}\sum_{m=1}^{M} 10 \times \log_{10}\left[\frac{\sum_{n=mN}^{mN+N-1} s^2[n]}{\sum_{n=mN}^{mN+N-1}(s[n]-\hat{s}[n])^2}\right], \tag{30}$$

and SSNR_{noisy} is the SSNR of noisy speech:

$$SSNR_{noisy} = \frac{1}{M}\sum_{m=1}^{M} 10 \times \log_{10}\left[\frac{\sum_{n=mN}^{mN+N-1} s^2[n]}{\sum_{n=mN}^{mN+N-1}(s[n]-y[n])^2}\right], \tag{31}$$

where M is the number of frames, and N is the frame size. A higher SSNRI score indicates that the speech signal has been enhanced more effectively.

4.4.2. Results of GMAPA and related approaches

Figs. 7–9 show the PESQ, SSNRI, and SDI scores for the MLSA, MAPA, MMSE, and GMAPA algorithms. In each figure, the left and right panels show the average results of babble and car noises, respectively, at a specific SNR. The MS method was used in this set of experiments. In Figs. 7–9, we also show the average results (denoted as "Average") of 0–20 dB SNRs for comparison. The result of original noisy speech is denoted as "Original noisy". The results for GMAPA were obtained by using $\{\alpha_{max} = 2.195, b = -0.7787, c = 21.18\}$.

The results of babble and car noises in Figs. 7–9 are quite consistent. From the results, we first observe that the proposed GMAPA algorithm performs the best in terms of the PESQ, SSNRI, and SDI objective evaluations across 0–20 dB SNR conditions. The better results achieved by GMAPA over MLSA and MAPA confirm that a suitable weight for prior density plays an important role in the gain function estimation for speech enhancement. Moreover, GMAPA provides clearer improvements over related algorithms in lower SNR conditions (0 dB, 5 dB, and 10 dB SNRs); the improvements become marginal in higher SNR conditions (15 dB and 20 dB SNRs). The result suggests that GMAPA with dynamically determining suitable weights gains more benefits under more adverse (low SNR) conditions.

4.4.3. Results of GMAPA with TG and SG techniques

In this section, we present the results of GMAPA with the TG and SG techniques. We intend to investigate that if TG and SG can enable more accurate estimations of weights of prior density and accordingly better speech

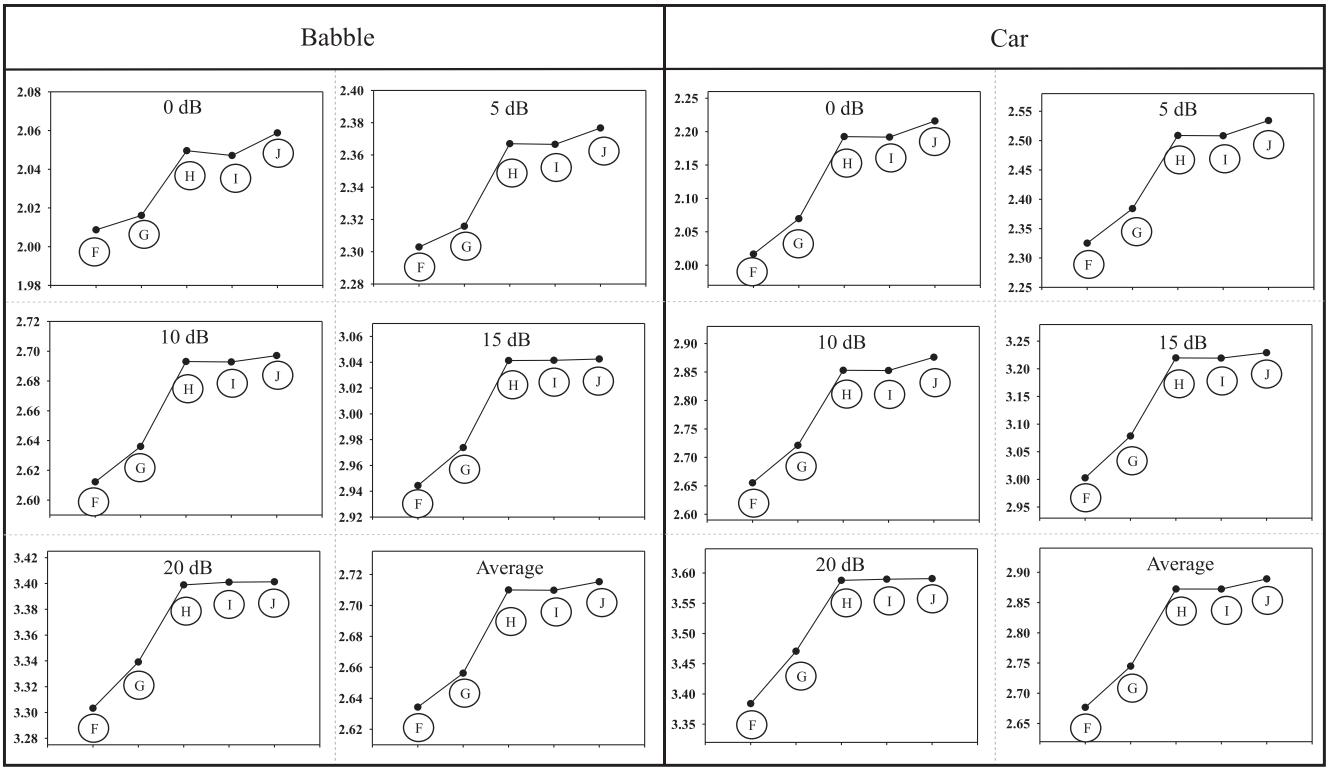

Fig. 7. PESQ values for Ⓕ Original noisy, Ⓖ MLSA, Ⓗ MAPA, Ⓘ MMSE, and Ⓙ GMAPA, at 0–20 dB SNRs and Average of babble noise (in the left side) and car noise (in the right side).

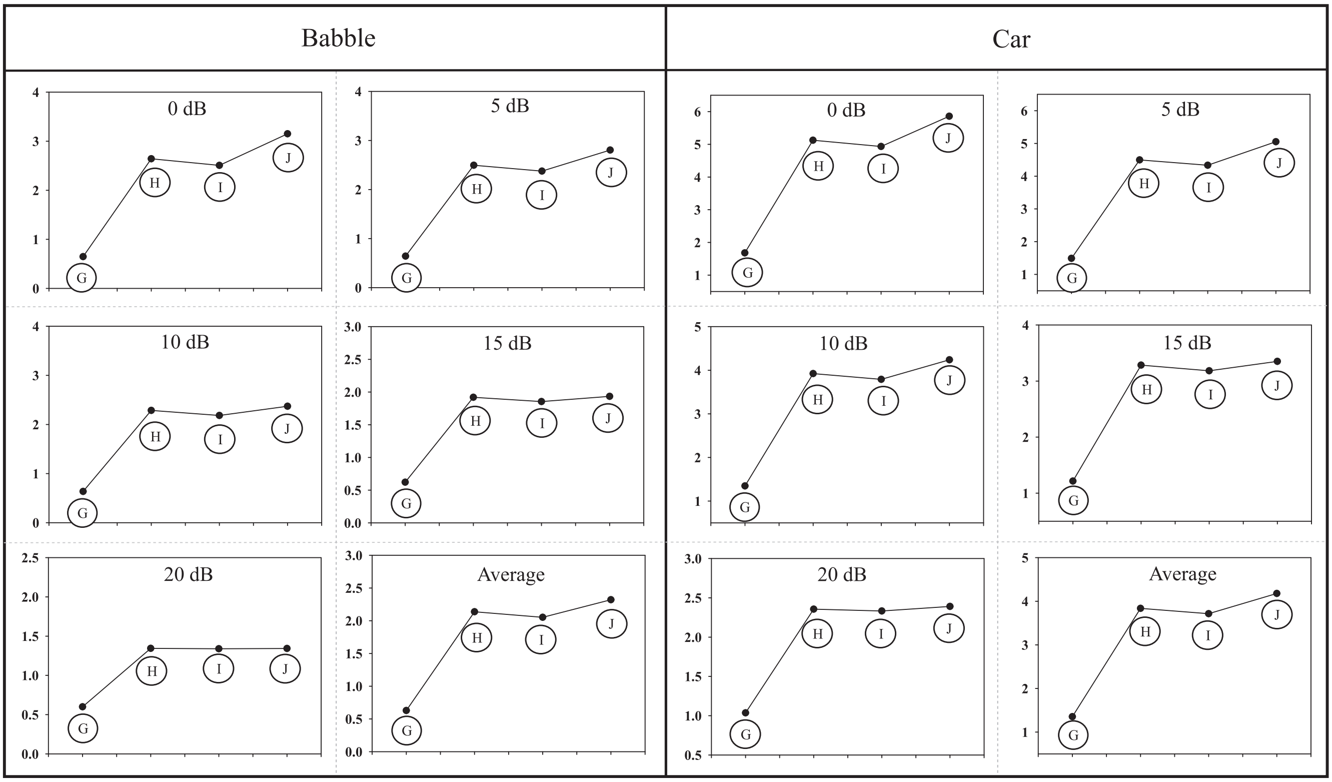

Fig. 8. SSNRI values for Ⓖ MLSA, Ⓗ MAPA, Ⓘ MMSE, and Ⓙ GMAPA, at 0–20 dB SNRs and Average of babble noise (in the left side) and car noise (in the right side).

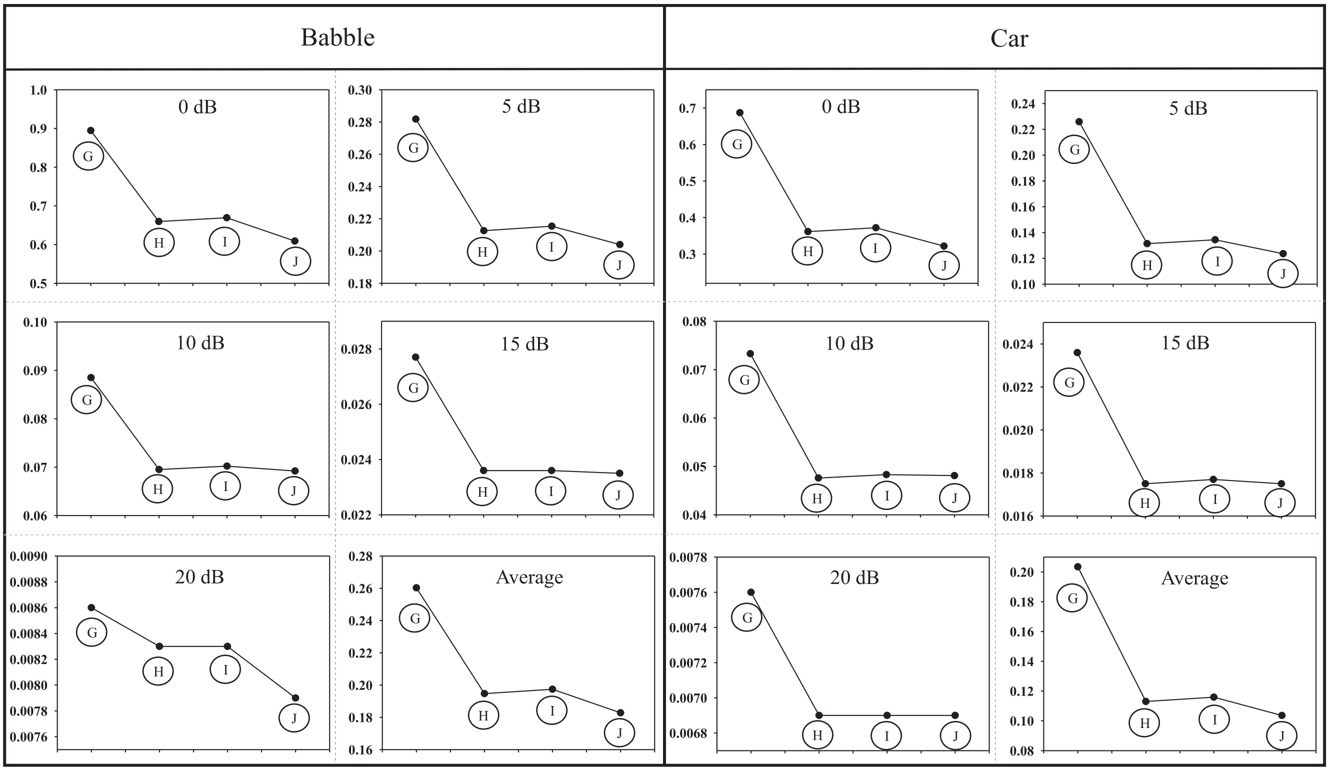

Fig. 9. SDI values for Ⓖ MLSA, Ⓗ MAPA, Ⓘ MMSE, and Ⓙ GMAPA, at 0–20 dB SNRs and Average of babble noise (in the left side) and car noise (in the right side).

enhancement performance for GMAPA. Please note that the statistics of ξ and γ in Eq. (22) are the same for GMAPA, GMAPA + TG and GMAPA + SG, while GMAPA + TG dynamically determines the weights (α in Eq. (22)) in a sliding-window manner, and GMAPA + SG estimates multiple weights, each for a particular sub-band. Fig. 10(a) shows the average PESQ scores (over babble and car noises at 0–20 dB SNRs) of GMAPA + TG using $M_c = 10$, 20, 30, 40, and M, where M is the total number of frames in the utterance. As presented in Section 3.3.1, when $M_c = M$, the whole utterance is used to determine the weight of prior density. From the results in Fig. 10(a), we note that $M_c = 10$, 20, 30, 40 all outperform that of $M_c = M$. Moreover, when $M_c = 30$, GMAPA + TG reaches the highest PESQ score; the performance maintains unchanged even though the size of sliding-window is increased (i.e., $M_c = 40$). The results confirm the effectiveness of using local temporal information for a more accurate estimation of the weight of prior density for GMAPA. Moreover, because TG computes the weight in a sliding-window manner, the processing latency of GMAPA + TG is smaller as compared to the utterance-based estimation used in original GMAPA.

In Fig. 10(b), we present the average PESQ results of GMAPA with SG (denoted as GMAPA + SG). As shown in Section 3.3.2, when $L_c = L$, the entire frequency bins share the same weight of prior density; when $L_c = 1$, each frequency bin has a specific weight of prior density. Fig. 10(b) also presents the results of GMAPA using $L_c = L/5$, $L/20$, and $L/80$. The results in Fig. 10(b) show that applying SG to combine the entire frequency bins into several groups can enable GMAPA to achieve better performance. It is also noted that when $L_c = L/5$, GMAPA + SG can achieve the best performance. The result suggests that it may not be necessary to specify a weight for each frequency bin; instead, the same weights shared by a reasonable group of frequency bins can already give satisfactory performance.

Finally we intend to compare GMAPA with TG and SG and related speech enhancement algorithms. The best setup ($M_c = 30$ for TG and $L_c = L/5$ for SG from Fig. 10) was used for GMAPA with TG and SG (denoted as GMAPA + TGSG). Fig. 11(a), (b), and (c), respectively, show the average PESQ, SSNRI, and SDI scores (over babble and car noises at 0–20 dB SNRs) of original noisy speech, MLSA, MAPA, MMSE, and GMAPA + TGSG. Notably, the TG and SG techniques were only involved to estimate α for computing the gain function in Eq. (22), and the same $\widehat{S} = G \cdot Y$ process in Eq. (8) was used for GMAPA + TGSG and related algorithms. By comparing the results in Fig. 11(a), (b), and (c), we noted that GMAPA + TGSG can provide better PESQ, SSNRI, and SDI scores than MLSA, MAPA, and MMSE.

4.5. Subjective listening tests

The purpose of the subjective listening tests was to evaluate the sound quality produced by the GMAPA algorithm

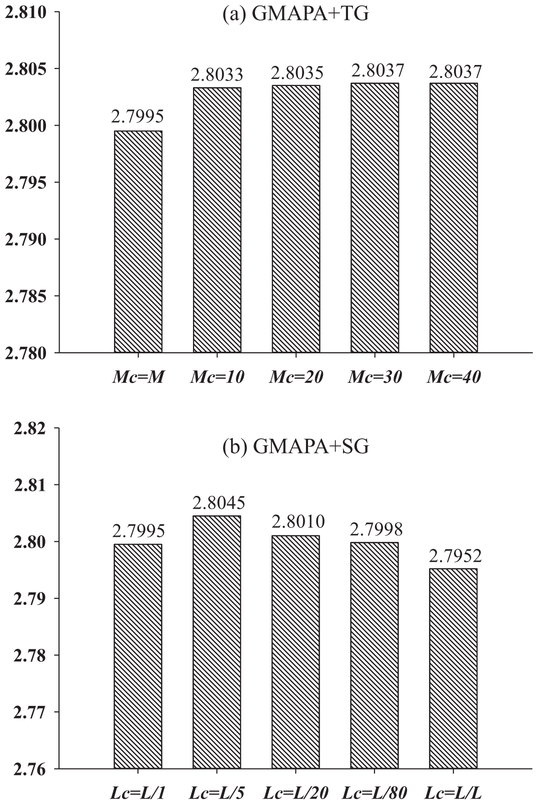

Fig. 10. Average PESQ scores (over babble and car noises at 0–20 dB SNRs) of (a) GMAPA + TG; (b) GMAPA + SG.

and compare it with those for the MLSA, MAPA, and MMSE algorithms in human subjects. We also intended to compare the subjective and objective assessment results, as described in detail below.

4.5.1. Experimental setup

The subjective listening test had a single-blind design in that subjects did not know which speech enhancement method was used when they were presented with the test sounds. The experimental platform was implemented using LabVIEW software, and the TDH-50P earphone (Valente et al., 1997) provided the original noisy and processed speech sounds to the human subjects. The TDH-50P earphone is commonly used in audiometer for hearing test because of its flat and wide frequency responses (100–8 kHz), high sensitivity [106 ± 2 dB sound pressure level (SPL) output with 1 milliwatt], linearity output (linear for power input from 0 to 400 mW), and low distortion (less than 1 percent). Before conducting the listening test, we first calibrated the output sound level based on ANSI (American National Standards Institute) S3.7 standard (ANSI, 1995 (R2008)) to ensure that the output levels were as intended. Four SNRs (−10, 0, 10, and 20 dB) were used to form the test set. The source signals of the test were concatenated segments of Mandarin spoken sentences combined with pink noise segments lasting for 15 s. The sentences were extracted from Taiwan news broadcasts spoken by a female newscaster (Lai et al., 2013c). The speech and noise signals were adjusted simultaneously by the same absolute amounts when producing the different SNRs; for example, when the input SNR was increased by 10 dB, the speech was increased by 5 dB and the noise was decreased by 5 dB, and the combined sounds were then adjusted to 65 dB SPL, which is a moderate listening level. In addition, all measurement procedures were performed in a quiet environment in which the background noise level was below 45 dB SPL. Finally, the original noisy signal and the noisy signals produced by each speech enhancement algorithm were played to the subject at 65 dB SPL. The purpose of this test was to investigate the subjective

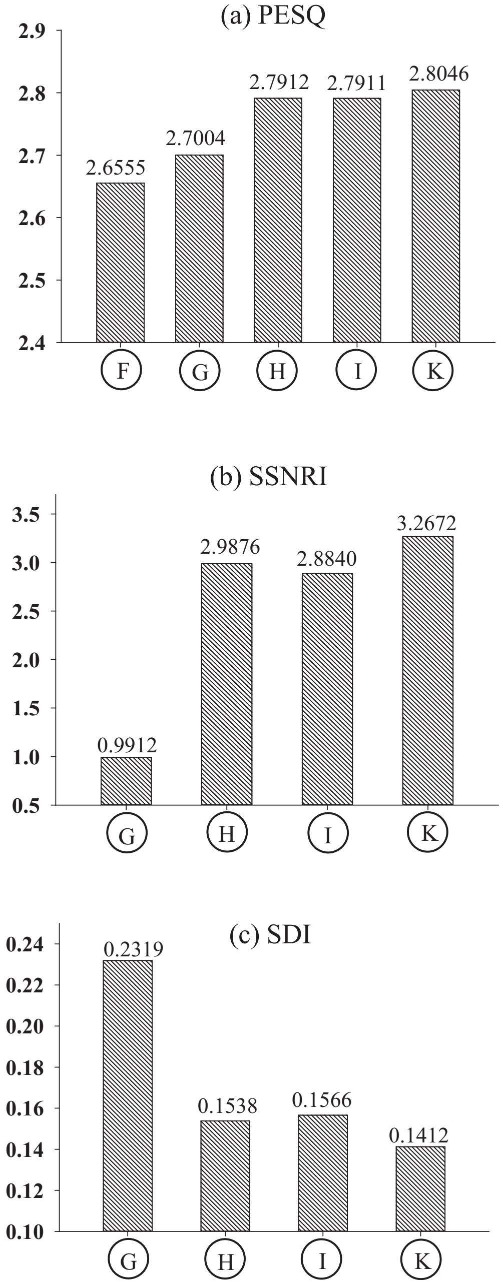

Fig. 11. Average objective evaluation scores (over babble and car noises at 0–20 dB SNRs) of Ⓕ Original noisy, Ⓖ MLSA, Ⓗ MAPA, Ⓘ MMSE, and Ⓚ GMAPA + TGSG.

sound quality performance in each speech enhancement algorithm. A standard sound quality questionnaire in clinical trials was used to rate scores (Lai et al., 2013b). The questionnaire included the following five statements: (Q1) I think this method provides high sound quality, (Q2) I think this method provides a natural sound, (Q3) I can hear very clearly when I use this method, (Q4) I feel comfortable when I use this method, and (Q5) I can't hear noise when I use this method; the subject was asked to score these statements for each of the speech signals using a mean opinion score (MOS) comparison. The MOS rating is the most widely used measure for subjective quality tests, in which the subjects rate the test speech scale from 5 to 1 for "strongly agree," "agree," "neither agree nor disagree," "disagree," and "strongly disagree," respectively (Quackenbush et al., 1988). Nine subjects with normal hearing (i.e., pure-tone hearing thresholds of better than 20 dB HL) whose native language is Mandarin participated in this study. The subjects were aged from 23 to 37 years, with a mean of 27 years.

4.5.2. Subjective listening test results

Fig. 12 shows the average MOS rating results from the sound quality questionnaire for the five different processing methods for nine subjects, where a higher score indicates that the enhancement method was preferred. The MCRA noise estimator was used in this set of experiments. The test results for GMAPA + TGSG in Fig. 11 are presented and denoted as GMAPA in Fig. 12. The sound quality scores for the original noisy signals and the outputs of the MLSA, MAPA, MMSE, and GMAPA algorithms were 4.11, 4.76, 4.87, 4.89, 4.93, respectively, for an input SNR of 20 dB; 3.66, 3.58, 4.16, 4.64, 4.76 for an input SNR of 10 dB; 2.78, 2.82, 3.56, 3.56, 3.71 for an input SNR of 0 dB; 1.80, 1.60, 1.96, 2.04, 2.60 for an input SNR of −10 dB. The paired *t*-test (Hayter, 2006) revealed significant differences for the original noisy signal versus the GMAPA algorithm ($p < 0.001$), MLSA versus GMAPA algorithms ($p < 0.001$), MAPA versus GMAPA algorithms ($p = 0.002$), and MMSE versus GMAPA algorithms ($p = 0.049$). The results confirm that the MOS values for the subjective listening tests were higher for sounds enhanced by the proposed GMAPA algorithm than for those enhanced using the other tested methods. Notably the results of subjective listening tests in Fig. 12 were actually consistent to the objective results in Figs. 7–9.

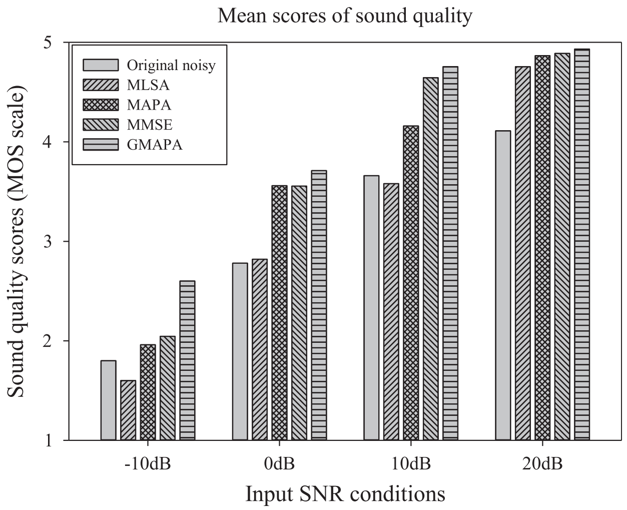

Fig. 12. Average MOS rating results over five questions for different types of speech enhancement method and different input SNRs.

5. Conclusion

This paper presents the GMAPA algorithm for implementing spectral restoration for speech enhancement. The GMAPA algorithm uses a weight, α, to determine the prior information for calculating the gain function. An STW mapping function was designed to determine the optimal α value according to the estimated SNR of the noisy speech signals. The objective evaluation results from PESQ, SSNRI, and SDI confirmed that the GMAPA algorithm outperformed the MMSE, MLSA, and MAPA algorithms at both lower and higher SNR levels. Moreover, the GMAPA performance can be further improved by considering the temporal and spectral information into the gain function estimation. In the meanwhile, the MOS scores obtained in a set of subjective listening tests showed that the GMAPA algorithm provided significantly higher satisfaction than the MMSE, MLSA, and MAPA algorithms at four SNR levels. These results suggest that the GMAPA algorithm can provide higher speech quality for human listening than other spectral restoration algorithms under various noisy conditions.

Recently, environment-aware applications have caused people's high attention. This study extends the traditional spectral restoration approaches to be capable of performing environment-aware speech enhancement based on the sigmoid-based STW function, which characterizes the correlation of *a posteriori* SNR statistics of the testing data and the weight of prior density for the gain estimation. In the future, we plan to further enhance the capability of environment-aware speech enhancement for GMAPA by using a better STW function with two directions. First, we will explore to use a dynamic model, such as hidden Markov model or particle filter, to more accurately characterize the time-variant characteristic of speech and nonstationary noise signals. Second, we plan to design the STW function by considering other acoustic properties, such as noise types and reverberation conditions, as well as other objective goals (rather than SDI used in this study). Moreover, this study uses fixed numbers of frames and frequency bins for TG and SG, respectively. GMAPA with variable sizes of temporal group and spectral group is also a worthwhile future work.

Acknowledgement

This work was supported by the Ministry of Science and Technology of Taiwan under Project MOST 104-2221-E-001-026-MY2.

Appendix A.

Here we derive the GMAPA gain function listed in Eq. (22). Assume that the noisy speech spectrum, Y, and the amplitude of the clean speech spectrum, S_k, are uncorrelated. Then the conditional PDF, $p(Y|S_k)$, can be written as

$$p(Y|S_k) = \frac{1}{\pi\sigma_v^2}\exp\left(-\frac{Y_k^2 + S_k^2}{\sigma_v^2}\right) I_0\left(\frac{2S_kY_k}{\sigma_v^2}\right), \tag{32}$$

where $I_0(\cdot)$ is the zero-order modified Bessel function. When the *a priori* SNR is higher (i.e., $\frac{2S_kY_k}{\sigma_v^2} \gg 1$), Eq. (32) can be approximated as

$$p(Y|S_k) \approx \frac{1}{2\pi\sigma_v\sqrt{\pi S_kY_k}}\exp\left(-\frac{Y_k^2 - 2S_kY_k + S_k^2}{\sigma_v^2}\right). \tag{33}$$

The PDF of the clean speech signal is

$$p(S_k) = \frac{2S_k}{\sigma_s^2}\exp\left(-\frac{S_k^2}{\sigma_s^2}\right). \tag{34}$$

Substituting Eq. (33) and Eq. (34) into Eq. (21), the cost function can be written as

$$\begin{aligned} J_{GMAPA}(S_k) = ln\left(\frac{1}{2\pi\sigma_v\sqrt{\pi S_kY_k}}\right) - \left(\frac{Y_k^2 - 2S_kY_k + S_k^2}{\sigma_v^2}\right) \\ + \alpha ln\left(\frac{2S_k}{\sigma_s^2}\right) - \frac{\alpha S_k^2}{\sigma_s^2}. \end{aligned} \tag{35}$$

Let

$$\frac{\partial J_{GMAPA}(S_k)}{\partial S_k} = 0; \tag{36}$$

then

$$\frac{-1}{2S_k} - \frac{-2Y_k + 2S_k}{\sigma_v^2} + \frac{\alpha}{S_k} - \frac{2\alpha S_k}{\sigma_s^2} = 0. \tag{37}$$

Multiplying Eq. (37) by $2S_k$ yields

$$-1 - \frac{-4Y_kS_k + 4S_k^2}{\sigma_v^2} + 2\alpha - \frac{4\alpha S_k^2}{\sigma_s^2} = 0. \tag{38}$$

After some derivations, Eq. (38) can be written as

$$S_k = \frac{Y_k\sigma_s^2 + \sqrt{Y_k^2\sigma_s^4 + (2\alpha - 1)(\alpha\sigma_v^2 + \sigma_s^2)\sigma_v^2\sigma_s^2}}{2(\alpha\sigma_v^2 + \sigma_s^2)}. \tag{39}$$

This is equivalent to

$$S_k = \frac{\xi + \sqrt{\xi^2 + [(2\alpha - 1)(\alpha + \xi)\sigma_s^2/Y_k^2]}}{2(\alpha + \xi)} Y_k, \tag{40}$$

and hence

$$S_k = \frac{\xi + \sqrt{\xi^2 + [(2\alpha - 1)(\alpha + \xi)\xi/\gamma]}}{2(\alpha + \xi)} Y_k. \tag{41}$$

Finally, we obtained the gain function G_{GMAPA} in Eq. (22).

References

Alippi, C., Storti-Gajani, G., 1991. Simple approximation of sigmoidal functions: realistic design of digital neural networks capable of learning. In: Proceedings of ISCAS'91, vol. 12, pp. 1505–1508.

ANSI, 1995 (R2008). ANSI S3.7-American National Standard Method for Coupler Calibration of Earphones.

Bahl, L.R., Bakis, R., Jelinek, F., Mercer, R.L., 1980. Language-model/acoustic channel balance mechanism. IBM Tech. Disclosure Bull. 23 (7B), 3464–3465.

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. 27, 113–120.

Cavallini, F., 1993. Fitting a logistic curve to data. College Math. J. 24, 247–253.

Chen, J., 2008. Fundamentals of Noise Reduction in Springer Handbook of Speech Processing. Springer (Chapter 43).

Chen, J., Benesty, J., Huang, Y., Doclo, S., 2006. New insights into the noise reduction Wiener filter. IEEE Trans. Audio Speech Lang. Process. 14, 1218–1234.

Chen, Y., Chu, M., Chang, E., Liu, J., Liu, R., 2003. Voice conversion with smoothed GMM and MAP adaptation. In: Proceedings of Interspeech'03, vol. 12, pp. 2413–2416.

Choi, M.S., Kang, H.G., 2005. An improved estimation of a priori speech absence probability for speech enhancement: in perspective of speech perception. In: Proceedings of ICASSP'05, vol. 12, pp. 1117–1120.

Cohen, I., 2002a. Noise estimation by minima controlled recursive averaging for robust speech enhancement. IEEE Signal Process. Lett. 9, 12–15.

Cohen, I., 2002b. Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. IEEE Signal Process. Lett. 9 (4), 113–116.

Cohen, I., 2003. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process. 11, 466–475.

Dat, T.H., Takeda, K., Itakura, F., 2006. Gamma modeling of speech power and its on-line estimation for statistical speech enhancement. IEICE Trans. Inform. Syst. 89 (3), 1040–1049.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 32, 1109–1121.

Fan, H.T., Hung, J.W., Lu, X., Wang, S.S., Tsao, Y., 2014. Speech enhancement using segmental nonnegative matrix factorization. In: Proceedings of ICASSP'14, vol. 12, pp. 4483–448.

Federico, M., 1996. Bayesian estimation methods for N-gram language model adaptation. In: Proceedings of ICSLP'96, vol. 12, pp. 240–243.

Flanagan, J.L., 1972. Speech Analysis, Synthesis and Perception. Springer-Verlag.

Gauvain, J.L., Lee, C.-H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech Audio Process. 2 (2), 291–298.

Hansen, J.H.L., Radhakrishnan, V., Arehart, K.H., 2006. Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system. IEEE Trans. Audio Speech Lang. Process. 14, 2049–2063.

Haykin, S., 1991. Advances in Spectrum Analysis and Array Processing. Prentice-Hall.

Hayter, A.J., 2006. Probability and Density for Engineers and Scientists, third ed. Duxbury Press.

Hendriks, R.C., Martin, R., 2007. MAP estimators for speech enhancement under normal and Rayleigh inverse Gaussian distributions. IEEE Trans. Audio Speech Lang. Process. 15 (3), 918–927.

Hirsch, H.G., Pearce, D., 2000. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: ISCA ITRW ASR2000.

Hu, Y., Loizou, P.C., 2008. Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang. Process. 16, 229–238.

ITU-T Recommendation P.862, 2001. Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-To-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs.

Kjems, U., Jensen, J., 2012. Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement. In: Proceedings of EUSIPCO'12, vol. 12, pp. 295–299.

Kum, J.M., Chang, J.H., 2009. Speech enhancement based on minima controlled recursive averaging incorporating second-order conditional MAP criterion. IEEE Signal Process. Lett. 16, 624–627.

Lai, Y.-H., Tsao, Y., Chen, F., 2013a. A study of adaptive WDRC in hearing aids under noisy conditions. Int. J. Speech Lang. Pathol. Audiol. 1 (2), 43–51.

Lai, Y.-H., Liu, T.-C., Li, P.-C., Shih, W.-T., Young, S.-T., 2013b. Development and preliminary verification of a Mandarin-based hearing-aid fitting strategy. PLOS ONE 8.

Lai, Y.-H., Li, P.-C., Tsai, K.-S., Chu, W.-C., Young, S.-T., 2013c. Measuring the long-term SNRs of static and adaptive compression amplification techniques for speech in noise. J. Am. Acad. Audiol. 24 (8), 671–683.

Lai, Y.-H., Wang, S.-S., Li, P.-C., Tsao, Y., 2015. A discriminative post-filter for speech enhancement in hearing aids. In: Proceedings of ICASSP'15, vol. 12, pp. 240–243.

Levitt, H., 2001. Noise reduction in hearing aids: an overview. J. Rehab. Res. Dev. 38 (1), 111–121.

Li, J., Sakamoto, S., Hongo, S., Akagi, M., Suzuki, Y., 2008. Adaptive β-order generalized spectral subtraction for speech enhancement. Signal Process. 88 (11), 2764–2776.

Li, J., Yang, L., Hu, Y., Akagi, M., Loizou, P.C., Zhang, J., Yan, Y., 2011a. Comparative intelligibility investigation of single-channel noise reduction algorithms for Chinese, Japanese and English. J. Acoust. Soc. Am. 129 (5), 3291–3301.

Li, J., Sakamoto, S., Hongo, S., Akagi, M., Suzuki, Y., 2011b. Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication. Speech Commun. 53 (5), 677–689.

Li, W., Itou, K., Takeda, K., Itakura, F., 2006. Single-channel multiple regression for in-car speech enhancement. IEICE Trans. Inform. Syst. 89-D (3), 1032–1039.

Lim, J.S., Oppenheim, A.V., 1979. Enhancement and bandwidth compression of noisy speech. Proc. IEEE 67, 1586–1604.

Loizou, P.C., 2013. Speech Enhancement: Theory and Practice. CRC Press.

Lotter, T., Vary, P., 2005. Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model. EURASIP J. Appl. Signal Process., 1110–1126

Lu, X., Matsuda, S., Unoki, M., Nakamura, S., 2010. Temporal contrast normalization and edge-preserved smoothing of temporal modulation structures of speech for robust speech recognition. Speech Commun. 52 (1), 1–11.

Lu, X., Unoki, M., Nakamura, S., 2011. Subband temporal modulation envelopes and their normalization for automatic speech recognition in reverberant environments. Comput. Speech Lang. 25 (3), 571–584.

Lu, Y., Loizou, P.C., 2008. A geometric approach to spectral subtraction. Speech Commun. 50, 453–466.

Malah, D., Cox, R.V., Accardi, A.J., 1999. Tracking speech-presence uncertainty to improve speech enhancement non-stationary noise environments. In: Proceedings of ICASSP'99, vol. 12, pp. 789–792.

Martin, R., 1994. Spectral subtraction based on minimum statistics. In: Proceedings of EUSIPCO'94, vol. 12, pp. 1182–1185.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. 9, 504–512.

Martin, R., 2005. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. IEEE Trans. Speech Audio Process. 13, 845–856.

McAulay, R.J., Malpass, M.L., 1980. Speech enhancement using a soft-decision noise suppression filter. IEEE Trans. Acoust. Speech Signal Process. 28, 137–145.

Mittal, U., Phamdo, N., 2000. Signal/noise KLT based approach for enhancing speech degraded by colored noise. IEEE Trans. Speech Audio Process. 8, 159–167.

Ogawa, A., Kazuya, T., Itakura, F., 1998. Balancing acoustic and linguistic probabilities. In: Proceedings of ICASSP'98, vol. 12, pp. 181–184.

Parihar, N., Picone, J., Pearce, D., Hirsch, H.G., 2004. Performance analysis of the Aurora large vocabulary baseline system. In: Proceedings of EUSIPCO'04, vol. 12, pp. 553–556.

Plourde, E., Champagne, B., 2008. Auditory-based spectral amplitude estimators for speech enhancement. IEEE Trans. Audio, Speech, Lang. Process. 16 (8), 1614–1622.

Quackenbush, S.R., Barnwell, T.P., Clements, M.A., 1988. Objective Measures of Speech Quality. Prentice Hall, Engle-wood Cliffs, NJ.

Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P., 2001. Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs. In: Proceedings of ICASSP'01, pp. 749–752.

Scalart, P., Filho, J.V., 1996. Speech enhancement based on a priori signal to noise estimation. In: Proceedings of ICASSP'96, pp. 629–632.

Soon, I.Y., Koh, S.N., Yeo, C.K., 1999. Improved noise suppression filter using self-adaptive estimator of probability of speech absence. Signal Process. 75 (2), 151–159.

Su, Y.- C., Tsao, Y., Wu, J.- E., Jean, F.- R., 2013. Speech enhancement using generalized maximum a posteriori spectral amplitude estimator. In: Proceedings of ICASSP'13, pp. 7467–7471.

Suhadi, S., Last, C., Fingscheidt, T., 2011. A data-driven approach to a priori SNR estimation. IEEE Trans. Audio Speech Lang. Process. 19, 186–195.

Tsao, Y., Matsuda, S., Hori, C., Kashioka, H., Lee, C.H., 2014a. A MAP-based online estimation approach to ensemble speaker and speaking environment modeling. IEEE Trans. Audio Speech Lang. Process. 22, 403–416.

Tsao, Y., Lu, X., Dixon, P., Hu, T.-Y., Matsuda, S., Hori, C., 2014b. Incorporating local information of the acoustic environments to MAP-based feature compensation and acoustic model adaptation. Comput. Speech Lang. 28, 709–726.

Valente, M., Potts, L.G., Valente, L.M., 1997. Differences and intersubject variability of loudness discomfort levels measured in sound pressure level and hearing level for TDH-50P and ER-3A earphones. J. Am. Acad. Audiol. 8, 59–67.

Venema, T., 2006. Compression for Clinicians. Thomson Delmar Learning, Chapter 7.

Xin, Z., Jancovic, P., Ju, L., Kokuer, M., 2008. Speech signal enhancement based on MAP algorithm in the ICA space. IEEE Trans. Signal Process. 56, 1812–1820.

Zervakis, M.E., 1996. Generalized maximum a posteriori processing of multichannel images and applications. Circ. Syst. Signal Process. 15, 233–260.

Zhang, M., Vassiliadis, S., Delgado-Frias, J.G., 1996. Sigmoid generators for neural computing using piecewise approximations. IEEE Trans. Comput. 45, 1045–1049.

Please cite this article in press as: Tsao, Y., Lai, Y.-H., Generalized *maximum a posteriori* spectral amplitude estimation for speech enhancement, Speech Comm. (2015), http://dx.doi.org/10.1016/j.specom.2015.10.003