Auditory-Based Spectral Amplitude Estimators for Speech Enhancement

Eric Plourde, *Student Member, IEEE*, and Benoît Champagne, *Senior Member, IEEE*

***Abstract*—We propose a new family of Bayesian estimators for speech enhancement where the cost function includes both a power law and a weighting factor. The parameters of the cost function, and therefore of the corresponding estimator gain, are chosen based on characteristics of the human auditory system, namely, the compressive nonlinearities of the cochlea, the perceived loudness and the ear's masking properties. It is found that choosing the parameters in this way results in a decrease of the estimator gain at high frequencies. This frequency dependence of the gain improves the noise reduction while limiting the speech distortion. Experimental results show that the new estimators achieve better enhancement performance than existing Bayesian estimators such as those based on the minimum mean-square error (MMSE) of the short-time spectral amplitude (STSA), the MMSE of the logarithm of the STSA (LSA) or the weighted euclidien (WE) error, both in terms of objective and subjective measures.**

***Index Terms*—Bayesian estimators, human auditory system, short-time spectral amplitude, speech enhancement.**

I. Introduction

In SPEECH enhancement, the general objective is to remove a certain amount of noise from a noisy speech signal while keeping the speech component as undistorted as possible. In Bayesian short-time spectral amplitude (STSA) estimation for speech enhancement, an estimate of the clean speech is derived by minimizing the expectation of a cost function that penalizes errors in the clean speech STSA estimate. Such estimators have been found in the past to perform better than most other methods including the spectral subtraction and subspace approaches [3].

A well-known Bayesian STSA estimator, the minimum mean square error (MMSE) of the STSA (i.e., MMSE STSA), is obtained when the chosen cost function is the squared error between the estimated and actual clean speech STSA [4]. Based on the assumption that the human auditory system performs a logarithmic compression of the STSA, and therefore that the logarithm of the STSA is more perceptually relevant than the STSA [5], the MMSE of the logarithm of the STSA (MMSE log-STSA or LSA) was proposed in [6]. In fact, one possible avenue for choosing an appropriate cost function is to consider the human hearing mechanism. In [7] and [8], masking thresholds were introduced in the Bayesian estimator's cost function to make it more perceptually significant while in [9], several perceptually relevant distortion metrics were considered as cost functions.

One of the cost functions which was found to yield the best results in [9] was based on the perceptually weighted error criterion used in speech coding. In that approach, the error spectrum is weighted by a filter which is the inverse of the original speech spectrum. This was adapted in [9] by proposing a generalization of the MMSE STSA cost function where the error between the estimated and actual clean speech STSA is weighted by the STSA of the clean speech raised to an exponent p; the resulting estimator is termed Weighted Euclidien (WE).

Another generalization of the MMSE STSA cost function was proposed by You *et al.* [10] in the β-Order STSA MMSE estimator; which we will denote as β-SA for convenience. The β-SA estimator applies a power law (i.e., an exponent β) to the estimated and actual clean speech STSA in the squared error of the cost function. While it was not interpreted as such in [10], this transformation can be seen as performing a nonlinear compression on the STSA. The dynamic range compression performed by the ear is a known characteristic of human hearing and, in fact, power laws have been used in the past to model this compression [11].

In this paper, to take advantage of both weighting *and* compression in the cost function, we propose a new family of Bayesian STSA estimators including both a weighting factor and a power law, which we will call the Weighted β-SA estimators (W β-SA). Moreover, we propose appropriate frequency-dependent values for the parameters entering in the W β-SA cost function, i.e., β and α (the latter is related to the WE estimator parameter p), based on characteristics of the human auditory system among which are the compressive nonlinearities of the cochlea, the perceived loudness, and the ear's masking properties.

It is found that choosing β and α according to the proposed approaches results in a decrease of the estimator's gain at high frequencies. This frequency dependence of the gain improves the noise reduction while limiting the speech distortion. Moreover, the new W β-SA estimator, with the proposed parameter values, shows improvements over the other Bayesian STSA estimators compared (i.e., MMSE STSA [4], LSA [6], and WE [9]) both in terms of objective and subjective measures.

The paper is organized as follows. Section II reviews existing Bayesian estimators while Section III derives the W β-SA family of estimators. In Sections IV-A and IV-B, the chosen

Manuscript received January 23, 2008; revised July 01, 2008. This work was supported in part by the Fonds Québécois de la Recherche sur la Nature et les Technologies and a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC). This paper was presented in part at EUSIPCO 2007 [1] and ICASSP 2008 [2]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yariv Ephraim.

The authors are with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC H3A 2A7, Canada (e-mail: eric.plourde@mail.mcgill.ca; benoit.champagne@mcgill.ca).

Digital Object Identifier 10.1109/TASL.2008.2004304

TABLE I
BAYESIAN STSA COST FUNCTIONS AND RESULTING ESTIMATOR GAINS G_k WITH EQUIVALENT W β-SA PARAMETER VALUES (β AND α).

	$C(\mathcal{X}_k, \hat{\mathcal{X}}_k)$	G_k	β	α
MMSE STSA [4]	$\left(\mathcal{X}_k - \hat{\mathcal{X}}_k\right)^2$	$\frac{\sqrt{\upsilon_k}}{\gamma_k}\Gamma(1.5)M\left(-0.5, 1; -\upsilon_k\right)$	1	0
LSA [6]	$\left(\log \mathcal{X}_k - \log \hat{\mathcal{X}}_k\right)^2$	$\frac{\upsilon_k}{\gamma_k}\exp\left\{\frac{1}{2}\int_{\upsilon_k}^{\infty}\frac{e^{-t}}{t}dt\right\}$	$\to 0$	0
β-SA [10]	$\left(\mathcal{X}_k^{\beta} - \hat{\mathcal{X}}_k^{\beta}\right)^2$	$\frac{\sqrt{\upsilon_k}}{\gamma_k}\left[\Gamma(\frac{\beta}{2}+1)M\left(-\frac{\beta}{2}, 1; -\upsilon_k\right)\right]^{1/\beta}, \quad \beta > -2$	β	0
WE [9]	$\mathcal{X}_k^p(\mathcal{X}_k - \hat{\mathcal{X}}_k)^2$	$\frac{\sqrt{\upsilon_k}}{\gamma_k}\frac{\Gamma\left(\frac{p+1}{2}+1\right)}{\Gamma\left(\frac{p}{2}+1\right)}\frac{M\left(-\frac{p+1}{2}, 1; -\upsilon_k\right)}{M\left(-\frac{p}{2}, 1; -\upsilon_k\right)}, \quad p > -2$	1	$-p/2$

values for β and α are discussed. Section V presents objective and subjective experimental results, while a conclusion follows in Section VI.

II. BAYESIAN STSA ESTIMATORS

Let the observed noisy speech of a particular frame i be

$$y_i[n] = x_i[n] + w_i[n] \quad 0 \le n < N \tag{1}$$

where $x_i[n]$ is the clean speech, $w_i[n]$ is the additive noise, and N is the length of the observation interval. Let $Y_{i,k}$, $X_{i,k}$, and $W_{i,k}$ denote the k^{th} complex spectral components of the noisy speech, clean speech, and noise, respectively, of the ith frame. To simplify the notation, we will usually omit the subscript i.

In Bayesian STSA estimation for speech enhancement, the goal is to obtain the estimator $\hat{\mathcal{X}}_k^o$ of $\mathcal{X}_k \triangleq |X_k|$, i.e., the STSA of X_k, which minimizes the expectation of a given cost function $C(\mathcal{X}_k, \hat{\mathcal{X}}_k)$

$$\hat{\mathcal{X}}_k^o = \underset{\hat{\mathcal{X}}_k}{\operatorname{argmin}} E\{C(\mathcal{X}_k, \hat{\mathcal{X}}_k)\} \tag{2}$$

where E denotes statistical expectation. This estimator is then combined with the phase of the noisy speech $\angle Y_k$ to yield the estimator of the complex spectrum of the clean speech

$$\hat{X}_k = \hat{\mathcal{X}}_k^o e^{j\angle Y_k}. \tag{3}$$

The time-domain estimate $\hat{x}[n]$ is obtained by performing an inverse Fourier transform of $\hat{X}_k$ for each frame which are then combined using the overlap–add method.

In the MMSE STSA estimator [4]

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = (\mathcal{X}_k - \hat{\mathcal{X}}_k)^2 \tag{4}$$

while in the MMSE log-STSA (LSA) estimator [6],

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = (\log(\mathcal{X}_k) - \log(\hat{\mathcal{X}}_k))^2. \tag{5}$$

In the derivation of these two estimators, the complex spectrums (i.e., the Fourier expansion coefficients) of the clean speech and noise were considered to be independent, identically distributed (i.i.d.) Gaussian random variables with zero mean and variances $\sigma_x^2 = E\left\{\mathcal{X}_k^2\right\}$ and $\sigma_w^2 = E\left\{|W_k|^2\right\}$, respectively.

Recently, the MMSE STSA estimator was generalized [10] by modifying the cost function (4) as

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = (\mathcal{X}_k^{\beta} - \hat{\mathcal{X}}_k^{\beta})^2 \tag{6}$$

where the exponent β is a real parameter whose purpose is to control the associated estimator gain function and, consequently, the tradeoff between speech distortion and noise reduction. Only the case $\beta > 0$ was considered in [10] while the case $-2 < \beta < 0$ was considered in [12]. We will refer to this estimator as the β-SA estimator. Interestingly, it was observed through gain curves in [10] that when $\beta \to 0$, the β-SA estimator tends to the LSA estimator (we provide a formal proof in the Appendix).

In [9], the following weighted form of the MMSE STSA cost function was proposed:

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = \mathcal{X}_k^p(\mathcal{X}_k - \hat{\mathcal{X}}_k)^2 \tag{7}$$

where p is a real parameter with $p > -2$. This estimator is termed the WE estimator and takes advantage of the masking properties of the ear. In fact, for $p < 0$, it forces a better clean speech estimation in regions where the STSA is smaller, and therefore less likely to mask noise remaining in the clean speech estimate. Similar to β in the β-SA estimator, p was also found to control the tradeoff between speech distortion and noise reduction when the corresponding estimator's gain is smaller than 1. In particular, a value of p closer to -2 was found to produce more noise reduction but also introduced greater speech distortions.

Using the statistical model in [4], gains G_k can be obtained from the previous cost functions such that

$$\hat{\mathcal{X}}_k^o = G_k|Y_k| \tag{8}$$

where $\hat{\mathcal{X}}_k^o$ is the corresponding optimal STSA estimator (2). Table I presents several Bayesian cost functions along with their associated estimator's gain (the values of α and β in Table I will be discussed in the next section). The gain parameters are

$$\upsilon_k = \frac{\xi_k}{1+\xi_k}\gamma_k, \ \xi_k = \frac{E\left\{\mathcal{X}_k^2\right\}}{E\left\{|W_k|^2\right\}}, \ \gamma_k = \frac{|Y_k|^2}{E\left\{|W_k|^2\right\}} \tag{9}$$

where $\Gamma(x)$ is the gamma function and $M(a, b; z)$ is the confluent hypergeometric function [13]. Moreover, $\gamma_k - 1$ can be interpreted as the instantaneous signal-to-noise ratio (SNR) while ξ_k acts as a long-term estimator of the SNR.

III. WEIGHTED β-SA ESTIMATOR

In this paper, we seek to combine the β-SA and WE cost functions into a single cost function to take advantage of the

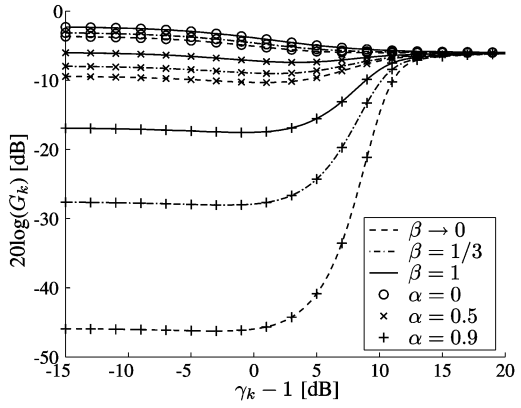

Fig. 1. W β-SA estimator gain ($20\log(G_k)$) versus instantaneous SNR ($\gamma_k - 1$) for several β and α values ($\xi_k = 0$ dB).

interpretations that can be given to the parameters β and α as will be discussed in the next section. The proposed cost function is therefore

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = \left(\frac{\mathcal{X}_k^{\beta} - \hat{\mathcal{X}}_k^{\beta}}{\mathcal{X}_k^{\alpha}} \right)^2 \tag{10}$$

where we used $\alpha = -p/2$ for convenience and β and α are real parameters whose ranges are discussed below.

By using (10) in (2), we obtain the corresponding Bayesian estimator

$$\hat{\mathcal{X}}_k^o = \left(\frac{E\left\{ \mathcal{X}_k^{\beta-2\alpha} | Y_k \right\}}{E\left\{ \mathcal{X}_k^{-2\alpha} | Y_k \right\}} \right)^{1/\beta}. \tag{11}$$

Using the Gaussian statistical model in [4] and [6] (i.e., clean speech and noise spectrums are i.i.d. Gaussian random variables with zero mean), we know (see [6] and in [9, App. A]) that

$$E\left\{ \mathcal{X}_k^m | Y_k \right\} = \lambda_k^{m/2} \Gamma\left(\frac{m}{2} + 1 \right) M\left(-\frac{m}{2}, 1; -\upsilon_k \right) \tag{12}$$

where $m > -2$ and

$$\frac{1}{\lambda_k} = \frac{1}{E\left\{ |W_k|^2 \right\}} + \frac{1}{E\left\{ {\mathcal{X}_k}^2 \right\}}.$$

Using (12) in (11) with the appropriate values of the parameter m (i.e., $m = \beta - 2\alpha$ for the numerator and $m = -2\alpha$ for the denominator), we can show that

$$\hat{\mathcal{X}}_k^o = G_k |Y_k|$$

where

$$G_k = \frac{\sqrt{\upsilon_k}}{\gamma_k} \left(\frac{\Gamma\left(\frac{\beta-2\alpha}{2} + 1 \right) M\left(-\frac{\beta-2\alpha}{2}, 1; -\upsilon_k \right)}{\Gamma(-\alpha + 1) M(\alpha, 1; -\upsilon_k)} \right)^{1/\beta} \tag{13}$$

and $\beta > 2(\alpha - 1)$, $\alpha < 1$. We will denote this new estimator as the Weighted β-SA estimator (W β-SA).

The W β-SA estimator gain G_k depends on the parameters of the cost function (i.e., β and α) as well as on γ_k and ξ_k. Fig. 1 presents gain curves as a function of the instantaneous SNR $\gamma_k - 1$ for a fixed $\xi_k = 0$ dB and several β and α values. As can be observed, the estimator's gain decreases when α increases and increases when β increases. It is worth noting that, below the ideal value of $\mathcal{X}_k/|Y_k|$, a decrease in the gain will result in more noise reduction but will invariably introduce more speech distortion. Also, since the proposed estimator generalizes both the β-SA and WE estimators, the gains of the later can be obtained by setting $\alpha = 0$ for β-SA and $\beta = 1$, $\alpha = -p/2$ for WE (see Table I).

It was shown in [9] that the WE estimator tends to a Wiener estimator as the instantaneous SNR tends to infinity. In fact, the more general W β-SA estimator also tends to a Wiener filter: we know from [14, (13.1.5)] that as $\gamma_k \to \infty$, the confluent hypergeometric function $M(-m/2, 1; -\upsilon_k)$, where υ_k is functionally related to γ_k through (9), can be written as

$$M\left(-\frac{m}{2}, 1; -\upsilon_k \right) = \frac{\upsilon_k^{m/2}}{\Gamma\left(\frac{m}{2} + 1 \right)}. \tag{14}$$

Using (14) in (13) with the appropriate values of the parameter m, we have

$$G_k = \frac{\xi_k}{1 + \xi_k} \tag{15}$$

which is a Wiener filter gain. Interestingly, since a Wiener filter results from the MMSE estimator of the complex spectral components [15], the use of the W β-SA cost function is therefore equivalent to $C(X_k, \hat{X}_k) = (X_k - \hat{X}_k)^2$ for high instantaneous SNR.

IV. Choosing Appropriate β and α Values

The parameter values of speech enhancement algorithms have been chosen in the past based on frame SNR such that a higher gain is obtained for higher SNR and vice versa [10], [16]. This had the effect of removing less noise at higher SNR to prevent speech distortion and more noise at low SNR.

Rather than considering the frame's SNR, we choose to consider the human auditory system to select appropriate values for β and α; in which case β and α will be fixed for all frames. In the first part of this section, we will present two different choices for β according to, first, the perceived loudness of sound and, second, the compressive nonlinearities of the cochlea. In the second part of this section, we will choose values of α considering the masking properties of the human auditory system. These different values will be compared through experimental results in Section V to assess their relevance in speech enhancement.

A. Choosing Appropriate β Values

In the LSA estimator [6], the logarithm of the spectral amplitude was considered. This was based on the fact that the MMSE of the logarithm of the spectral amplitude was thought to be more perceptually relevant than the spectral amplitude itself. It is known that loudness is more perceptually relevant than the sound's intensity. Therefore, a cost function which would consider the difference in terms of the perceived loudness would seem preferable to cost functions which consider the difference

in terms of the sound intensity. Power laws have been used in the past to model the nonlinear relation between the intensity of sound and its perceived loudness [17], [18]. An exponent of 1/3 (i.e., cubic root) has been used in [18] to approximate the nonlinear transformation between intensity and perceived loudness. An appropriate value for β would therefore be $\beta = 1/3$. This value will be further assessed experimentally in Section V.

An important factor that plays a role in the perception of loudness is the dynamic range compression performed by the ear [19]. This dynamic range compression is thought to be due to many factors among which are the cochlea's compressive nonlinearities. Compression rates of 0.2 dB/dB were measured at the base of the mammalian cochlea (i.e., for high frequencies) for intensities between 40 and 90 dB sound pressure level (SPL) [20] (conversational speech is at 60 dB SPL). These compression rates can be easily incorporated in the proposed Bayesian cost function (10). In fact, β can be directly interpreted as the compression rate, in dB/dB, of the input spectral amplitudes and thus set to corresponding physiologically meaningful values. Therefore, instead of motivating the value of β strictly in terms of loudness perception, we can also look at the physiology of the cochlea, which can explain to some extent the loudness perception of the human auditory system, and propose other relevant values for β.

The cochlea's compressive nonlinearities are well documented and accepted at high frequencies; however, there is no consensus on the degree of nonlinearity at lower frequencies (i.e., at the apex of the cochlea) [19], [20]. There would seem to be less compression at lower frequencies than at higher frequencies. In fact, some research even fails to show any compression (i.e., compression rate of 1 dB/dB) at low frequencies or even show an expansion (i.e., compression rate > 1 dB/dB) [20]. Here, we will assume no compressive nonlinearity at the low-frequency limit. Since the compression rates will be different at low and high frequencies, the values of β will therefore be frequency dependent, i.e., β_k.

To propose adequate values for the β_k's, we need to define the cochlea's rate of compression for every frequency k. Since for low frequency we consider the absence of compressive nonlinearity, we will choose β_k at the low-frequency limit as $\beta_{\text{low}} = 1$. As indicated previously, the compressive nonlinearity at high frequencies is thought to have a compression rate of approximately 0.2 dB/dB. For high frequencies, it therefore seems plausible to set the high-frequency limit of the β_k value as $\beta_{\text{high}} = 0.2$.

Physiological experiments on the cochlear rate of compression at intermediate frequencies (i.e., between the apex and the base of the cochlea) are extremely scarce [19]. Therefore, we propose to interpolate β_k for intermediate frequencies based on the following approach. We consider the fact that each frequency corresponds to a position on the basilar membrane following the so-called tonotopic mapping [20]. One such tonotopic mapping, proposed in [21], is given by

$$d_k = \frac{1}{\eta} \log_{10}\left(\frac{f_k}{A} + l\right) \tag{16}$$

where d_k is the position on the basilar membrane in millimeters, $\eta = 0.06\ \text{mm}^{-1}$, $A = 165.4$ Hz, $l = 1$ are parameters set as per [21], and f_k is the frequency in Hz corresponding to spectral component k, i.e., $f_k = kF_s/N$, where F_s is the sampling frequency set to 16 kHz in this study.

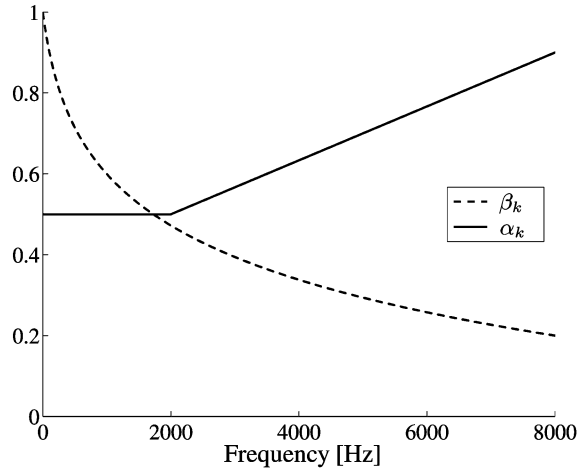

Fig. 2. β_k and α_k versus frequency [Hz].

We will therefore consider the compression rate to vary linearly with respect to the position d_k on the basilar membrane, corresponding to frequency f_k as given by the tonotopic mapping. In fact, the compressive nonlinearity is thought to be caused by the active process of the outer hair cells, and it is known that the hair cells follow a tonotopic organization where they are optimally sensitive to a particular frequency according to their position on the basilar membrane [22]. The complete set of β_k values are thus derived by linearly interpolating between β_{low} and β_{high} according to d_k

$$\beta_k = d_k \frac{(\beta_{\text{high}} - \beta_{\text{low}})}{\frac{1}{\eta}\log_{10}\left(\frac{F_s}{2A} + l\right)} + \beta_{\text{low}}. \tag{17}$$

Fig. 2 represents the different values of β_k as a function of the frequency.

In the first part of this section, we proposed the use of an exponent value $\beta = 1/3$ as a simple model for approximating the nonlinear transformation between intensity and perceived loudness. It is interesting to note that more elaborate loudness models lead to a similar pattern in compression as the one described in the second part of this section. In fact, in [23], an exponent of 0.2 is used at high frequencies to preform compression while it is increased for lower frequencies.

B. Choosing Appropriate α Values

The WE estimator [9] takes advantage of the masking properties of the ear. In fact, one of the motivations for deriving the WE estimator was to favor a more accurate estimation of smaller STSA since they are less likely to mask noise remaining in the clean speech estimate. This was done by choosing a fixed value of p that increased the weight of smaller STSA in the cost function (e.g., $p = -1$).

Since most of the speech energy is located at lower frequencies [24], higher frequencies should contain mainly small STSA. Therefore, it would be relevant to further increase the weights of the smaller STSA in the cost function for higher frequencies. This can be done by increasing α for higher frequencies (or equivalently decreasing p since $\alpha = -p/2$). We therefore propose, instead of using a fix value of α as in [9],

to modify α as a function of frequency, i.e., α_k, increasing its value for higher frequencies.

To do so, we need to choose appropriately the values of α_k for each frequency. In [9], the value of $p = -1$ (corresponding to $\alpha = 0.5$) has been suggested as a good compromise between the desired noise reduction performed by the estimator and the speech distortion introduced. This value can also be regarded as being a good compromise between increasing the weight of smaller STSA while keeping an appropriate estimation error for larger STSA. Since the main part of the speech energy, which will contain most of the larger STSA, is approximately located below 2000 Hz [24] (which also includes most of the first two formants [25]), we will choose the value of $\alpha = 0.5$ up to 2000 Hz. For higher frequencies, we want to further increase the weights of smaller STSA. Since, on average, the total speech energy decreases as frequency increases, we therefore propose to linearly increase the value of α as a function of the frequency. The W β-SA estimator restricts α to $\alpha < 1$, based on experimentations, we choose $\alpha = 0.9$ as the high frequency limit. Choosing higher values (e.g., $\alpha = 0.99$) did not introduce significant noise reduction while unnecessarily distorting the speech. Therefore α_k will be given by

$$\alpha_k = \begin{cases} \alpha_{\text{low}}, & f_k \leq 2\,\text{kHz} \\ \frac{(f_k - 2000)(\alpha_{\text{high}} - \alpha_{\text{low}})}{\frac{F_s}{2} - 2000} + \alpha_{\text{low}}, & \text{else} \end{cases} \tag{18}$$

where $\alpha_{\text{low}} = 0.5$ and $\alpha_{\text{high}} = 0.9$ (see Fig. 2).

As can be seen when observing the gain curves in Fig. 1 for the chosen values of β_k and α_k in Fig. 2, both approaches suggest a decrease in the gain for high frequencies. Further justifications for such processing will be given in Section V-B.

V. Experimental Results

In this section, we will study the W β-SA estimator with the proposed β and α values and compare it to the MMSE STSA, LSA, and WE Bayesian estimators using both objective and subjective measures. The value of p in the WE estimator will be set to $p = -1$ as proposed in [9]. There is no constant value of β proposed in [10] against which to compare the proposed algorithms. However, it is important to note that the case $\alpha = 0$ corresponds to the β-SA estimator.

Three types of noises from the Noisex database [26] are used in the experiments: a so-called white noise and two colored noises, that is a pink noise and an aircraft cockpit noise (buccaneer-1). Other noise types were considered during the experimentation and lead to the same conclusions as the ones drawn next. The normalized average spectrum magnitudes of the different noises used here are shown in Fig. 3. Noisy speech signals were created according to ITU-T standard P.56 [27]. The number of noisy sentences used, respectively, in the objective and subjective evaluations will be specified in the corresponding subsections below. All speech signals were sampled at 16 kHz and a raised-cosine window [28] was used (512 samples, 32 ms) in the STSA computation. All frames were zero-padded from 512 samples to 1024 samples to limit temporal aliasing [29], and the corresponding signals were phase shifted by 256 samples to avoid discontinuities between blocks as similarly proposed in [30]. A 75% overlap was used in the overlap-add synthesis method as in [4]. All algorithms used the *decision-directed* approach for the estimation of ξ_k [4], and a voice activity detector proposed in [31] was used to evaluate the noise spectral amplitude variance. None of the estimators considered the uncertainty of speech signal presence in the noisy observations.

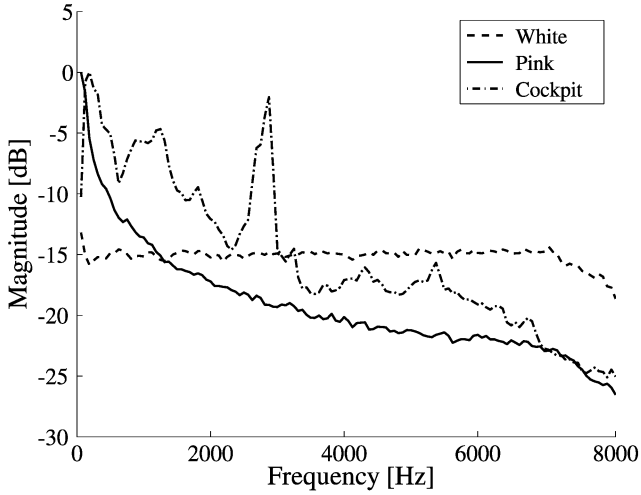

Fig. 3. Normalized average noise spectrum magnitudes [dB] versus frequency [Hz].

A. Objective Results

Many objective measures are available to assess speech enhancement algorithms. Some of them are more correlated with the speech distortion introduced by the enhancement while others are more correlated with the noise reduction. A study of the correlation between the mean opinion score (MOS) and some objective measures was presented in [32]. On the one hand, it was shown that the segmental SNR $(\text{SNR}_{\text{seg}})$ measure is best correlated with background noise reduction and poorly correlated with the speech distortion. On the other hand, the log-likelihood-ratio measure (LLR) was found to be best correlated with the speech distortion and poorly correlated with the noise reduction. We will use both SNR_{seg} and LLR to get a more complete evaluation of the performance of the algorithms. Moreover, we will also use the Perceptual Evaluation of Speech Quality (PESQ) measure which, while it was not originally intended to assess speech enhancement algorithms [33], has been found to have a good correlation overall with MOS [32] and was recently used in several speech enhancement studies [10], [34], [35]. The PESQ attempts to predict MOS scores and yields a result from 1 to 4.5, the higher score being the best result.

The SNR_{seg} measure can be expressed as [36]

$$\text{SNR}_{\text{seg}} = \frac{1}{L}\sum_{i=0}^{L-1} 10\log_{10}\frac{\|\mathbf{x}_i[n]\|^2}{\|\mathbf{x}_i[n] - \hat{\mathbf{x}}_i[n]\|^2} \tag{19}$$

where $\mathbf{x}_i[n]$ and $\hat{\mathbf{x}}_i[n]$ are the N-dimensional vector comprising of the clean and enhanced speech at frame i, respectively, and L is the number of frames in the speech signal. As proposed in [36], each frame SNR was thresholded by a -20-dB lower bound and a 35-dB higher bound.

TABLE II
SNR_{seg} FOR SEVERAL β AND α VALUES (WHITE NOISE, 0 dB)

	$\beta = 1$	$\beta = \beta_k$	$\beta = 1/3$	$\beta \to 0$
$\alpha = 0$	-0.79 (MMSE STSA)	0.21	0.36	0.79 (LSA)
$\alpha = 0.5$	1.77 (WE $p = -1$)	2.15	2.13	2.32
$\alpha = \alpha_k$	2.19	2.50	2.55	2.77

TABLE III
SNR_{seg} FOR SEVERAL β AND α VALUES (PINK NOISE, 0 dB)

	$\beta = 1$	$\beta = \beta_k$	$\beta = 1/3$	$\beta \to 0$
$\alpha = 0$	-1.34 (MMSE STSA)	-0.65	-0.51	-0.22 (LSA)
$\alpha = 0.5$	0.36 (WE $p = -1$)	0.48	0.57	0.69
$\alpha = \alpha_k$	0.42	0.52	0.63	0.77

TABLE IV
SNR_{seg} FOR SEVERAL β AND α VALUES (COCKPIT NOISE, 0 dB)

	$\beta = 1$	$\beta = \beta_k$	$\beta = 1/3$	$\beta \to 0$
$\alpha = 0$	-1.72 (MMSE STSA)	-0.93	-1.00	-0.74 (LSA)
$\alpha = 0.5$	-0.22 (WE $p = -1$)	-0.05	-0.03	0.08
$\alpha = \alpha_k$	-0.16	-0.01	0.03	0.16

TABLE V
LLR FOR SEVERAL β AND α VALUES (WHITE NOISE, 0 dB)

	$\beta = 1$	$\beta = \beta_k$	$\beta = 1/3$	$\beta \to 0$
$\alpha = 0$	1.99 (MMSE STSA)	1.78	1.88	1.84 (LSA)
$\alpha = 0.5$	1.77 (WE $p = -1$)	1.64	1.75	1.73
$\alpha = \alpha_k$	1.38	1.40	1.36	1.59

TABLE VI
LLR FOR SEVERAL β AND α VALUES (PINK NOISE, 0 dB)

	$\beta = 1$	$\beta = \beta_k$	$\beta = 1/3$	$\beta \to 0$
$\alpha = 0$	1.41 (MMSE STSA)	1.19	1.35	1.33 (LSA)
$\alpha = 0.5$	1.29 (WE $p = -1$)	1.22	1.29	1.30
$\alpha = \alpha_k$	1.23	1.53	1.40	1.71

TABLE VII
LLR FOR SEVERAL β AND α VALUES (COCKPIT NOISE, 0 dB)

	$\beta = 1$	$\beta = \beta_k$	$\beta = 1/3$	$\beta \to 0$
$\alpha = 0$	1.45 (MMSE STSA)	1.25	1.40	1.38 (LSA)
$\alpha = 0.5$	1.37 (WE $p = -1$)	1.30	1.37	1.38
$\alpha = \alpha_k$	1.33	1.58	1.48	1.74

The LLR measure can be expressed as [37]

$$\text{LLR} = \log\left(\frac{\hat{\mathbf{a}}\mathbf{R}\hat{\mathbf{a}}^T}{\mathbf{a}\mathbf{R}\mathbf{a}^T}\right) \tag{20}$$

where $\mathbf{a}$ is the linear predictive coding (LPC) coefficient row vector of the original clean speech signal, $\hat{\mathbf{a}}$ is the LPC coefficient row vector of the enhanced speech signal, $\mathbf{R}$ is the autocorrelation matrix of the original clean speech signal, and T indicates the transpose operator. To remove unrealistically high distortion levels, the mean of all frames is evaluated by ignoring the frames with LLR greater than 5σ [38] (this corresponded typically to less than 1% of the frames). A lower LLR score indicates a better performance.

Tables II–IV present the SNR_{seg} results for white, pink, and cockpit noises, respectively, at an SNR of 0 dB. All SNR_{seg}, LLR, and PESQ results are averages obtained from 60 Harvard sentences (six males, six females, five sentences each) [39]. The columns and lines of the tables are structured in somewhat decreasing β and increasing α order where $0.2 \leq \beta_k \leq 1$ and $0.5 \leq \alpha_k \leq 0.9$.

As reported in [9] for the WE estimator and in [10] for the β-SA estimator, we can observe that the SNR_{seg} generally increases for a decreasing β and an increasing α. This result is easily explained since for a decreasing β and an increasing α, the gain function of the estimator G_k decreases (see Fig. 1) which produces more noise reduction and, as we mentioned previously, the SNR_{seg} is better correlated with noise reduction. The best result is therefore obtained for the smallest β (i.e., $\beta \to 0$) and biggest α (i.e., $\alpha = \alpha_k$).

We present LLR results in Tables V–VII for white, pink, and cockpit noises, respectively, at an SNR of 0 dB. For the white noise case, the best results were obtained for $\alpha = \alpha_k$. For the colored noises, the best results were obtained for $\beta = \beta_k$, $\alpha = 0.5$. Setting $\alpha = \alpha_k$ reduces greatly the noise at high frequency since it decreases the gain, but it simultaneously introduces some speech distortions, especially when combined with smaller β values. Those high-frequency speech distortions were less perceptible in white noise which has a high-frequency content. However, for the colored noises used here, which have a small high-frequency content, the speech distortions became more perceptible.

We next compare the PESQ (Perceptual Evaluation of Speech Quality—ITU-T Recommendation P.862) [33] results of the proposed estimator, with the MMSE STSA, LSA, and WE $(p = -1)$ estimators at noisy speech SNRs between -5 and 5 dB. Fig. 4 shows the PESQ improvements over the noisy speech signal PESQ values for the given estimators and SNR values. White noise [Fig. 4(a)], pink noise [Fig. 4(b)], and aircraft cockpit noise [Fig. 4(c)] are presented. The noisy speech PESQ values were 0.94 at -5 dB and 1.52 at 5 dB

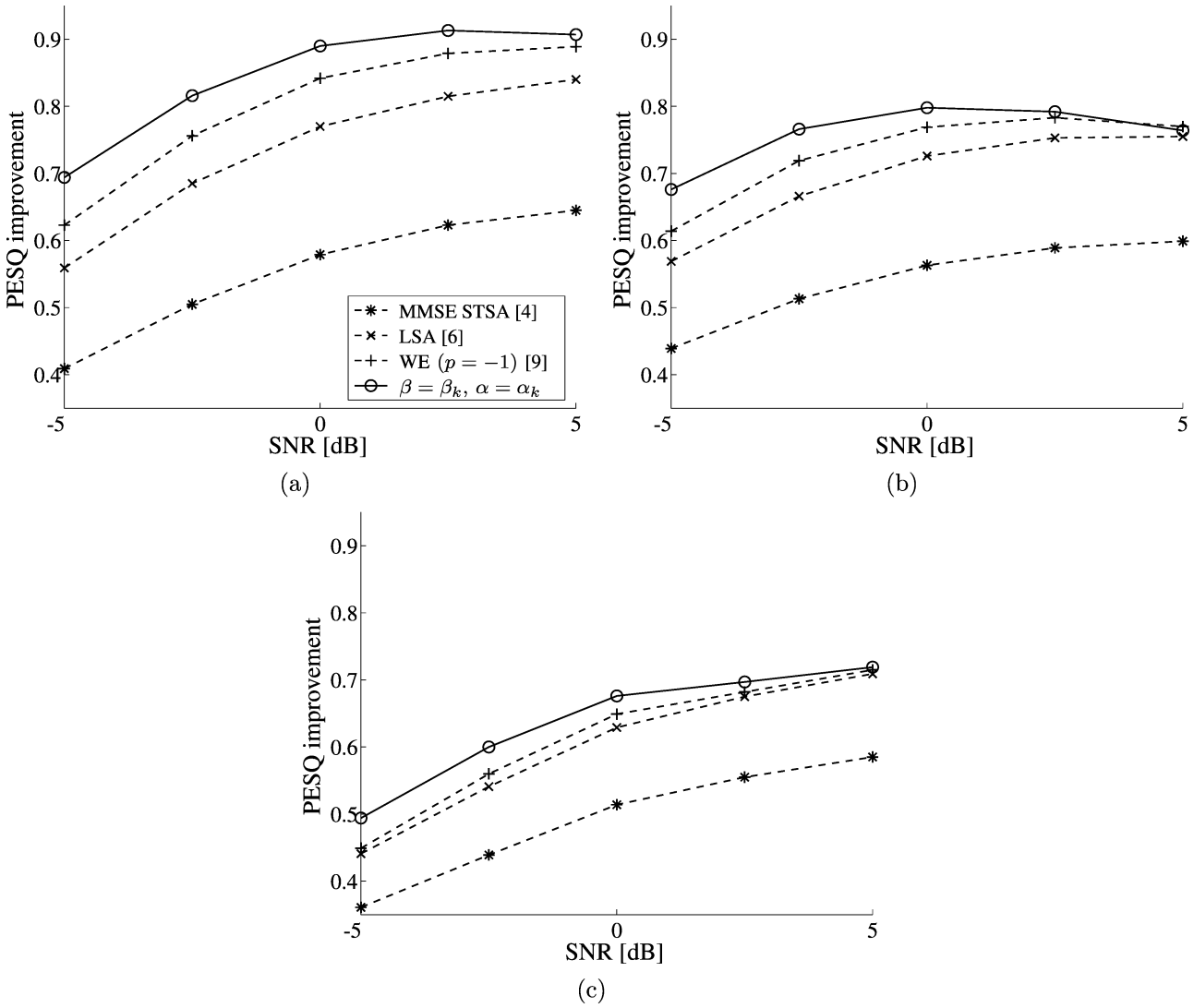

Fig. 4. PESQ improvement over noisy signal versus SNR. (a) White noise. (b) Pink noise. (c) Aircraft cockpit noise.

(evaluated as averages of all three noise types). For clarity purposes, only the $\beta = \beta_k$, $\alpha = \alpha_k$ case is plotted.

The case $\beta = \beta_k$, $\alpha = \alpha_k$ was found to be better than the MMSE STSA, LSA, and WE $(p = -1)$ estimators. While the results are not presented here, the cases $\beta = \beta_k$, $\alpha = 0.5$ and $\beta = 1$, $\alpha = \alpha_k$ where found to be better overall than WE $(p = -1)$ and LSA but worse than $\beta = \beta_k$, $\alpha = \alpha_k$ whereas the cases $\beta \to 0$, $\alpha = \alpha_k$, and $\beta = 1/3$, $\alpha = \alpha_k$ performed better than LSA and WE $(p = -1)$ at an SNR of -5 dB but worse at higher SNRs. In fact, while the case $\beta \to 0$, $\alpha = \alpha_k$ had the highest SNR_{seg} score, it introduces significant speech distortion (as identified by the LLR results) and shows a poor PESQ value, in particular at higher SNRs. The best compromise therefore seems to be with $\beta = \beta_k$, $\alpha = \alpha_k$.

While the results for male and female spoken utterances are grouped together in the previous tables and figures, an analysis was performed where the results were separated according to the speaker's gender. Results in terms of LLR where similar for both male and female while SNR_{seg} results from the sentences spoken by males were approximately 1-dB inferior to the ones spoken by females; however, the conclusions did not change when comparing the different estimators in each gender group. PESQ values were found to be slightly inferior for females when compared to males for all estimators. Again, the same ordering of the different estimators was obtained in each group. The only exception was for the cockpit noise and male utterances where the LSA estimator was found to be better than WE for all SNRs and also better than W β-SA ($\beta = \beta_k$, $\alpha = \alpha_k$) for an SNR of 5 dB.

B. Subjective Results

As a subjective measure, we used a test setup similar to the MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) (ITU-R Recommendation BS.1534-1) [40] method as implemented in [41]. In MUSHRA, the subjects are provided with the test utterances plus one reference and one hidden anchor and are asked to rate the different signals on a scale of 0 to 100, 100 being the best score. As the hidden anchor, we used a signal having an SNR of 5 dB less than the noisy signal to be enhanced. The listeners were allowed to listen to each sentence several times and always had access to the clean signal reference.

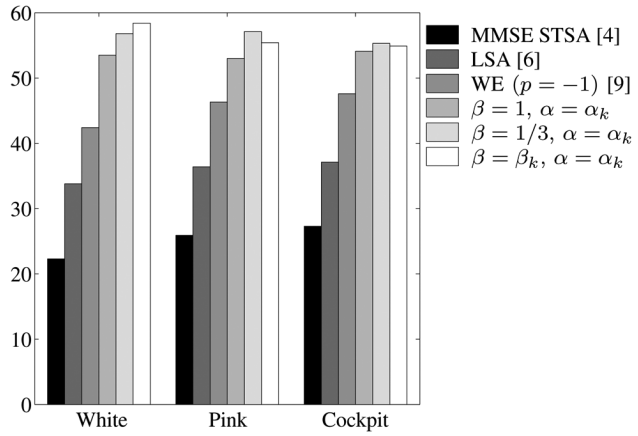

Fig. 5. Comparative subjective results for white, pink, and cockpit noises (0 dB).

A total of eight listeners (seven males, one female ages in the mid 20s to low 30s with a background in either speech processing or telecommunications) participated in the test of which half where judged to be experienced listeners. A subset of two sentences (one male speaker, one female speaker) were chosen randomly from the sentences used previously for the objective evaluation. The two sentences were corrupted by the same three noise types as before and enhanced using several estimators, the same sentences were used for all subjects. Tests were performed in an isolated acoustic room using Beyerdynamic DT880 headphones. The average duration of a test was approximately 30 min per subject.

Fig. 5 presents the comparative subjective results for the MMSE STSA, LSA, and WE estimators along with those of the W β-SA estimator with proposed values $\beta = 1$, $\alpha = \alpha_k$; $\beta = 1/3$, $\alpha = \alpha_k$, and $\beta = \beta_k$, $\alpha = \alpha_k$. As can be observed, the sentences enhanced using the W β-SA estimator were rated higher than those enhanced by the other estimators for all noise types. Two-tailed paired t-tests [42] revealed the advantage of the W β-SA estimator with the proposed values ($\beta = 1$, $\alpha = \alpha_k$; $\beta = 1/3$, $\alpha = \alpha_k$; $\beta = \beta_k$, $\alpha = \alpha_k$) over the WE ($p = -1$) to be statistically significant for all three noise types within a 95% confidence interval.

We observe that listeners in the previous experiment preferred an enhanced speech having more high-frequency noise reduction than one having less high-frequency speech distortion. It was observed in [43] that listeners seem to be more sensitive to speech distortion than noise reduction when participating in a subjective evaluation of enhanced speech. This conclusion was based on experiments with sampled speech at 8 kHz, whereas we used a 16-kHz sampling rate. Therefore, the conclusions of [43] only applies to the lower frequency portion of the spectrum considered in our work. Based on our experimental work with 16 kHz and the result in [43], it would seem that the high-frequency speech distortion is less important in subjective evaluations than the low-frequency speech distortion.

Additional subjective tests (not shown here), using a smaller subset of the previous subjects, were also performed for an SNR of 5 dB. The W β-SA algorithms still received higher scores than all the other algorithms. However, while a substantial advantage of the W β-SA estimators was still found over LSA, the difference between the W β-SA estimators and the WE ($p = -1$) estimator was found to be narrower than for the 0-dB case. Moreover, an analysis where the results were grouped according to the speaker's gender was also performed for the subjective results. No differences were observed in the comparative results except that the three W β-SA estimators (i.e., $\beta = 1$, $\alpha = \alpha_k$; $\beta = 1/3$, $\alpha = \alpha_k$, and $\beta = \beta_k$, $\alpha = \alpha_k$) were interchanged for the cockpit noise and male spoken utterances.

C. Discussion

The human ear is more sensitive between 3 and 4 kHz, as can be observed from an equal loudness curve [19], and will therefore perceive weaker sounds in that frequency band. Therefore, it would seem advantageous to improve the estimation of those weaker sounds in the frequency band between 3 and 4 kHz. Additional experiments where conducted where we locally increased the value of α for those frequencies, therefore giving more importance to weaker sounds. We compared this approach with the approach using the proposed α_k values. A slight improvement was observed in terms of PESQ for the white noise as well as in terms of LLR for the colored noise cases; all SNR_{seg} values as well as the other PESQ and LLR values showed no significant differences. Moreover, informal listening experiments revealed marginal differences between the two approaches.

We chose the W β-SA estimator parameters based on characteristics of the human auditory system. It turn out that, both the approaches using β_k and α_k produce a decrease in the gain G_k at high frequencies compared to lower frequencies (as can be observed from the gains in Fig. 1 with the values of β_k and α_k as in Fig. 2). This decrease in G_k generates more noise reduction at high frequencies but has the simultaneous effect of producing more speech distortions. The speech distortions are however minimized at low frequencies, where the main speech energy is located, by keeping β high and α low, therefore producing a higher gain. The proposed β_k and α_k values will therefore be more advantageous when the noise has high-frequency content, such as white noise, in which case more noise will be removed while speech distortions will be less perceptible. This explains why the proposed algorithms obtained the best performance in white noise.

Moreover, the distortions of the high-frequency contents of speech, such as fricatives, will be less perceptible in heavy noise (i.e., low SNRs) but they could become more perceptible in regions or sentences where the noise is weak. This could explain why the estimators are more advantageous at smaller SNRs, as observed. It is important to note, however, that the gain is mostly decreased for low instantaneous SNRs. In fact, for high instantaneous SNRs, all estimators tend toward the Wiener gain therefore reducing the speech distortions. For low instantaneous SNRs, the heavy noise will mask the speech signal; since these cannot be restored, the estimator will apply a small gain which will remove much of the noise.

VI. Summary and Conclusion

We proposed a new family of estimators for speech enhancement, the W β-SA, where the cost function included both a

power law and a weighting factor. The corresponding estimator's gain parameters (i.e., β and α) were chosen according to characteristics of the human auditory system. It is found that doing so suggests a decrease in the gain at high frequencies which limits the speech distortions at low frequencies while increasing the noise reduction at high frequencies. Improvements over existing Bayesian estimators such as the MMSE STSA, LSA, and WE estimators were reported, particularly for noise having high-frequency content and at low SNRs, both in terms of objective (SNR_{seg}, LLR and PESQ) and subjective measures. In particular, choosing $\beta = \beta_k$ and $\alpha = \alpha_k$ was found to yield good overall results.

APPENDIX

In this appendix, we show that the W β-SA estimator with $\alpha = 0$ and $\beta \to 0$ (or β-SA estimator with $\beta \to 0$) is equivalent to the LSA estimator.

Proof: Setting $\alpha = 0$ in (13) we get the β-SA estimator gain, which is expressible as

$$G_{\beta,k} = \frac{\sqrt{\upsilon_k}}{\gamma_k}\left[\Gamma\left(\frac{\beta}{2}+1\right) M\left(-\frac{\beta}{2},1;-\upsilon_k\right)\right]^{1/\beta}.$$

Using 8.342.1 from [13], which states

$$\ln\Gamma(z+1) = -\gamma z + \sum_{k=2}^{\infty}(-1)^k \frac{z^k}{k}\zeta(k) \quad |z|<1$$

where γ is Euler's constant and $\zeta(k)$ is given in [13], we have

$$G_{\beta,k} = \frac{\sqrt{\upsilon_k}}{\gamma_k}\exp\left\{-\frac{\gamma}{2} + \frac{1}{\beta}\sum_{k=2}^{\infty}(-1)^k \frac{\left(\frac{\beta}{2}\right)^k}{k}\zeta(k) + \frac{1}{\beta}\ln M\left(-\frac{\beta}{2},1;-\upsilon_k\right)\right\}.$$

Therefore

$$\lim_{\beta\to 0} G_{\beta,k} = \frac{\sqrt{\upsilon_k}}{\gamma_k} e^{-\gamma/2} \lim_{\beta\to 0}\exp\left\{\frac{\ln M\left(-\frac{\beta}{2},1;-\upsilon_k\right)}{\beta}\right\} = \frac{\sqrt{\upsilon_k}}{\gamma_k} e^{-\gamma/2}\exp\left\{\lim_{\beta\to 0}\frac{\frac{\partial}{\partial\beta}M\left(-\frac{\beta}{2},1;-\upsilon_k\right)}{M\left(-\frac{\beta}{2},1;-\upsilon_k\right)}\right\}$$

where L'Hopital's rule has been used.

Deriving $M\left(-\beta/2,1;-\upsilon_k\right)$ term by term as in [6] and since $\lim_{\beta\to 0} M\left(-\beta/2,1;-\upsilon_k\right) = 1$ we have

$$\lim_{\beta\to 0} G_{\beta,k} = \frac{\sqrt{\upsilon_k}}{\gamma_k}\exp\left\{-\frac{\gamma}{2} - \frac{1}{2}\sum_{r=1}^{\infty}\frac{(-\upsilon_k)}{r!}\frac{1}{r}\right\} = \frac{\sqrt{\upsilon_k}}{\gamma_k}\exp\left\{\frac{1}{2}\left(\ln(\upsilon_k) + \int_{\upsilon_k}^{\infty}\frac{e^{-t}}{t}dt\right)\right\} = \frac{\upsilon_k}{\gamma_k}\exp\left\{\frac{1}{2}\int_{\upsilon_k}^{\infty}\frac{e^{-t}}{t}dt\right\}$$

where 8.214.1 from [13] was used in the second line and the last line is the LSA gain from [6]

ACKNOWLEDGMENT

The authors would like to thank the participants in the listening tests for their precious help as well as the anonymous reviewers for their helpful comments.

REFERENCES

[1] E. Plourde and B. Champagne, "Integrating the cochlea's compressive nonlinearity in the bayesian approach for speech enhancement," in *Proc. 15th Eur. Signal Process. Conf.*, Poznań, Poland, 2007, pp. 70–74.

[2] E. Plourde and B. Champagne, "Perceptually based speech enhancement using the weighted β-SA estimator," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, 2008, pp. 4193–4196.

[3] Y. Hu and P. C. Loizou, "Subjective comparison of speech enhancement algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Toulouse, France, 2006, pp. 153–156.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[5] Y. Ephraim and I. Cohen, *The Electrical Engineering Handbook*, 3rd ed. Boca Raton, FL: CRC, 2005.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[7] P. J. Wolfe and S. J. Godsill, "Towards a perceptually optimal spectral amplitude estimator for audio signal enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Istanbul, Turkey, 2000, pp. 821–824.

[8] P. J. Wolfe and S. J. Godsill, "A perceptually balanced loss function for short-time spectral amplitude estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Hong Kong, 2003, pp. 425–428.

[9] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 857–869, Sep. 2005.

[10] C. H. You, S. N. Koh, and S. Rahardja, " β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 475–486, Jul. 2005.

[11] R. Meddis and L. P. O'Mard, "A computational algorithm for computing nonlinear auditory frequency selectivity," *J. Acoust. Soc. Amer.*, vol. 109, no. 6, pp. 2852–2861, Jun. 2001.

[12] E. Plourde and B. Champagne, "Further analysis of the β-order MMSE STSA estimator for speech enhancement," in *Proc. 20th IEEE Canadian Conf. Elect. Comput. Eng.*, Vancouver, BC, Canada, 2007, pp. 1594–1597.

[13] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 6th ed. New York : Academic, 2000.

[14] *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, M. Abramowitz and I. A. Stegun, Eds. Washington, DC: U.S. Gov. Print. Off., 1964.

[15] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 845–856, Sep. 2005.

[16] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Washington, DC, 1979, pp. 208–211.

[17] T. L. Petersen and S. F. Boll, "Acoustic noise suppression in the context of a perceptual model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Atlanta, GA, 1981, pp. 1086–1088.

[18] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[19] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. New York: Academic, 2004.

[20] L. Robles and M. A. Ruggero, "Mechanics of the mammalian cochlea," *Physiological Rev.*, vol. 81, no. 3, pp. 1305–1352, Jul. 2001.

[21] D. D. Greenwood, "A cochlear frequency-position function for several species—29 years later," *J. Acoust. Soc. Amer.*, vol. 87, no. 6, pp. 2592–2605, Jun. 1990.

[22] E. R. Kandel, J. H. Schwartz, and T. M. Jessell, *Principles of Neural Science*, 4th ed. New York: McGraw-Hill, 2000.

[23] B. C. J. Moore, B. R. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *J. Audio Eng. Soc.*, vol. 45, no. 4, pp. 224–240, Apr. 1997.
[24] C. Formby and R. B. Monsen, "Long-term average speech spectra for normal and hearing-impaired adolescents," *J. Acoust. Soc. Amer.*, vol. 71, no. 1, pp. 196–202, 1982.
[25] D. O'Shaughnessy, *Speech Communications: Human and Machine*, 2nd ed. Piscataway, NJ: IEEE Press, 2000.
[26] "Signal processing information base: Noise data," Rice University, Houston, TX [Online]. Available: http://spib.rice.edu/spib/select_noise.html.
[27] "Recommendation P.56: Objective measurement of active speech level," ITU-T, 1993.
[28] P. Kabal, Windows for transform processing McGill Univ., Montreal, QC, Canada, Tech. Rep., 2005 [Online]. Available: http://www-mmsp.ece.mcgill.ca/Documents/Reports/2005/KabalR2005a.pdf.
[29] J. B. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-25, no. 3, pp. 235–238, Jun. 1977.
[30] H. Gustafsson, S. E. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 799–807, Nov. 2001.
[31] J. Sohn and N. S. Kim, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.
[32] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in *Proc. 9th Int. Conf. Spoken Lang. Process.—Interspeech*, Pittsburgh, PA, 2006, pp. 1447–1450.
[33] "Recommendation P.862: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU-T, 2001.
[34] N. Ma, M. Bouchard, and R. A. Goubran, "Speech enhancement using a masking threshold constrained kalman filter and its heuristic implementations," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 1, pp. 19–32, Jan. 2006.
[35] V. Grancharov, J. Samuelsson, and B. Kleijn, "On causal algorithms for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 3, pp. 764–773, May 2006.
[36] P. E. Papamichalis, *Practical Approaches to Speech Coding*. New York: Prentice-Hall, 1987.
[37] S. R. Quackenbush, T. P. Barnwell, III, and M. A. Clements, *Objective Measures of Speech Quality*. New York: Prentice-Hall, 1988.
[38] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. 5th Int. Conf. Spoken Lang. Process.*, Sydney, Australia, 1998, pp. 2819–2822.
[39] "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AU-17, no. 3, pp. 225–246, Sep. 1969.
[40] "Recommendation BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems," ITU-R, 2001.
[41] E. Vincent, "MUSHRAM: A MATLAB interface for MUSHRA listening tests," [Online]. Available: http://www.elec.qmul.ac.uk/people/emmanuelv/mushram/
[42] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*, 3rd ed. New York: Wiley, 2003.
[43] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

Eric Plourde, photograph and biography not available at the time of publication.

Benoît Champagne, photograph and biography not available at the time of publication.