Further Analysis of the β-Order MMSE STSA Estimator for Speech Enhancement

Eric Plourde and Benoît Champagne

Department of Electrical and Computer Engineering

McGill University

Montreal, Quebec, Canada, H3A 2A7

Email: eric.plourde@mail.mcgill.ca, benoit.champagne@mcgill.ca

Abstract— In Bayesian approaches for speech enhancement, the clean speech is estimated by minimizing the expectation of a desired cost function. In the β-order MMSE STSA (βSA) Bayesian estimator, the cost function is the squared difference between the estimated and actual clean speech short-time spectral amplitude (STSA), both to the power $\beta > 0$. In this paper we propose an extension of the analysis of the βSA estimator for values of $\beta < 0$. We find that when $\beta < 0$, a normalization occurs in the βSA estimator which produces more noise reduction as β is reduced at the expense of additional speech distortion. Furthermore, the βSA estimator with $\beta = -1$ slightly outperforms the well known MMSE STSA and MMSE log-STSA (LSA) estimators in terms of the PESQ, for the two noises studied, while the overall MOS appreciation for $\beta = -1$ is found to be better than both MMSE STSA and LSA for white noise.

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. Many approaches have been proposed to achieve this goal, such as the spectral subtraction, Bayesian or subspace approaches [1]. In Bayesian approaches, an estimate of the clean speech is derived by minimizing the expectation of a defined cost function. Well-known Bayesian estimates are minimum mean square error (MMSE) estimates of the short-time spectral amplitude (STSA) derived by Ephraim and Malah ([2], [3]). In the so-called MMSE STSA estimator [2], the chosen cost function involves the difference between the estimated and actual clean speech STSA, while in MMSE log-STSA (LSA) [3], the difference is taken between the logarithm of the estimated and actual clean speech STSA. Recently, these estimators were generalized under the β-order MMSE STSA (βSA) estimator [4] which applies a positive exponent, β, to both the actual and estimated clean speech STSA.

In this paper, we propose an extension of the βSA estimator analysis performed in [4] to negative values of β. After exposing general aspects of MMSE speech enhancement, we show that for a decreasing β, both the speech distortion and noise reduction increases. Furthermore, compared to the MMSE STSA and LSA estimators, the βSA estimator with $\beta = -1$ yields better results in terms of PESQ for the two noises studied while the overall MOS appreciation is found to be better than both MMSE STSA and LSA for white noise.

II. THE β-ORDER MMSE STSA ESTIMATOR

Let the observed noisy speech be

$$y(t) = x(t) + n(t) \qquad 0 \le t \le T \tag{1}$$

where $x(t)$ is the clean speech, $n(t)$ is the additive noise and $[0, T]$ is the observation interval. Let Y_k, X_k and N_k denote the kth complex spectral components of the noisy speech, clean speech and noise respectively.

In Bayesian STSA estimation for speech enhancement, the goal is to obtain the estimator $\hat{\mathcal{X}}_k$ of $\mathcal{X}_k \triangleq |X_k|$ which minimizes $E\{C(\mathcal{X}_k, \hat{\mathcal{X}}_k)\}$ where $C(\mathcal{X}_k, \hat{\mathcal{X}}_k)$ is a chosen cost function and E denotes statistical expectation. In MMSE-STSA, $C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = (\mathcal{X}_k - \hat{\mathcal{X}}_k)^2$ while in LSA, $C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = (\log(\mathcal{X}_k) - \log(\hat{\mathcal{X}}_k))^2$.

The MMSE-STSA estimator was generalized under the βSA estimator in [4][1] by modifying the cost function as $C(\mathcal{X}_k, \hat{\mathcal{X}}_k, \beta) = (\mathcal{X}_k^{\beta} - \hat{\mathcal{X}}_k^{\beta})^2$ where the exponent β is a positive real parameter. The βSA estimator is expressible as:

$$\hat{\mathcal{X}}_k = G_k |Y_k| \tag{2}$$

$$G_k = \frac{\sqrt{\upsilon_k}}{\gamma_k} \left[\Gamma\left(\frac{\beta}{2} + 1\right) M\left(-\frac{\beta}{2}; 1; -\upsilon_k\right) \right]^{1/\beta} \tag{3}$$

with:

$$\upsilon_k = \frac{\xi_k}{1 + \xi_k}\gamma_k, \quad \xi_k = \frac{E\{\mathcal{X}_k^2\}}{E\{|N_k|^2\}}, \quad \gamma_k = \frac{|Y_k|^2}{E\{|N_k|^2\}}$$

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

When $\beta = 1$, the βSA estimator is identical to the MMSE STSA estimator. Furthermore, You *et al.* suggested in [4] that when $\beta \to 0$, the βSA estimator is equivalent to the LSA estimator. In fact, when comparing Equation (27) in [5] and Equation (19) in [3], one concludes that the two estimators are actually identical. Therefore, the MMSE STSA and LSA estimators are both subsets of the more general βSA estimator.

While only the case $\beta > 0$ was considered in [4], the resulting expression for the gain in (3) is in fact valid for

---

[1]An equivalent estimator for the power spectra of the clean speech, $\hat{\mathcal{X}}_k^2$, was also derived in [5] and termed *Generalized MMSE*.

$\beta > -2$. Therefore, the study for the case $-2 < \beta < 0$ is missing in [4] and will be the subject of the remainder of this paper.

III. THE CASE $\beta < 0$

A. A normalization interpretation

In the case $-2 < \beta < 0$, $\beta = -|\beta|$ and the βSA cost function becomes:

$$
\begin{align}
C(\mathcal{X}_k, \hat{\mathcal{X}}_k, \beta) &= \left(\frac{1}{\mathcal{X}_k^{|\beta|}} - \frac{1}{\hat{\mathcal{X}}_k^{|\beta|}} \right)^2 \tag{4} \\
&= \left(\frac{\hat{\mathcal{X}}_k^{|\beta|} - \mathcal{X}_k^{|\beta|}}{\mathcal{X}_k^{|\beta|} \hat{\mathcal{X}}_k^{|\beta|}} \right)^2 \tag{5} \\
&= \frac{C(\mathcal{X}_k, \hat{\mathcal{X}}_k, |\beta|)}{\left(\mathcal{X}_k \hat{\mathcal{X}}_k \right)^{2|\beta|}} \tag{6}
\end{align}
$$

Therefore, we see that using a negative β amounts to normalizing the cost function for a positive β, i.e. $C(\mathcal{X}_k, \hat{\mathcal{X}}_k, |\beta|)$, by $\left(\mathcal{X}_k \hat{\mathcal{X}}_k \right)^{2|\beta|}$. The denominator in (6) can be thought of as an approximation of the power spectrum to the exponent $2|\beta|$. Therefore, taking $\beta < 0$ has the effect of normalizing the cost function by the estimated power spectrum to the exponent $2|\beta|$. This normalization penalizes the estimation error more heavily when the power spectrum is small, which corresponds to spectral valleys, than when it is large, i.e. spectral peaks. Therefore, it takes advantage of the masking properties of the human ear: in fact, more noise is likely to be audible in the speech spectral valleys (i.e. for low speech spectral amplitude values) than in the speech spectral peaks where the noise is more likely to be masked by the speech. The βSA estimator for $\beta < 0$ will therefore favor a more accurate estimation of the speech in the spectral valleys.

It is worth noting that the βSA cost function for the case $\beta < 0$ is closely related to one proposed by Loizou in [6]:

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

where q is a real, positive or negative, parameter[2]. Loizou's motivation to derive that estimator was in fact to exploit the masking properties of the ear. In fact, the arguments we just exposed were also proposed by Loizou in [6] for $q > 0$. Both estimator will therefore have a similar behavior and favor an accurate estimation of the speech in the spectral valleys. However, in the case of (6), a $|\beta|$ exponent appears in the numerator and, also, the normalization is performed using both the estimated and actual speech spectral amplitudes.

B. Analysis of the estimator

1) Gain vs. instantaneous SNR: Figure 1 shows the gain, $20 \log(G_k)$, versus the instantaneous SNR, $\gamma_k - 1$, when ξ_k is fixed, for several estimators: MMSE STSA (or βSA with $\beta = 1$), LSA (or βSA with $\beta \to 0$) and βSA with

[2]In [6], the cost function is defined using $p = -q$.

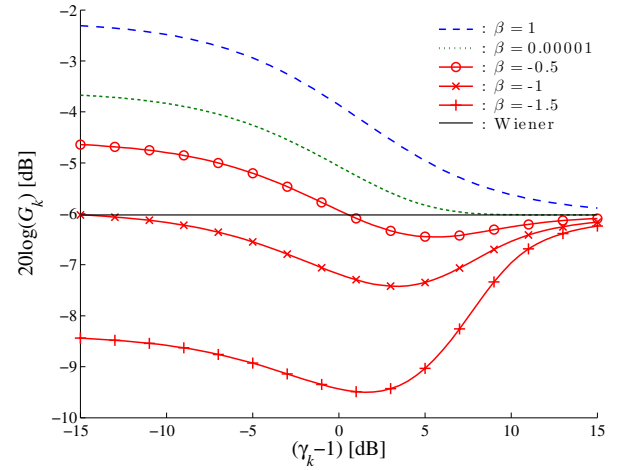$\beta = -0.5, -1, -1.5$. As can be seen, the gain decreases as β decreases.

Fig. 1. Estimator gain ($20 \log(G_k)$) vs instantaneous SNR ($\gamma_k - 1$) ($\xi_k = 0$dB)

For $\beta > 0$, the gain is a monotonically decreasing function of $\gamma_k - 1$. However, when $\beta < 0$, the βSA gain is not a monotonically decreasing function anymore. Furthermore, in [4], it was noted that the gain tended to the Wiener filter gain (which is $\xi_k/(1 + \xi_k)$) as the instantaneous SNR increased. For $\beta < 0$, we see that it is still the case, however, the gain can now become less than the Wiener filter's.

2) Noise reduction vs. speech distortion: As observed in Fig. 1, the gain decreases as β decreases, therefore, more noise reduction and also more speech distortion should be expected for lower β values. In order to study the speech distortion and noise reduction properties of the estimator, we adapted distortion metrics presented in [7] for the time domain to act respectively as a speech distortion metric ($\Upsilon(G_k)$) and noise reduction metric ($\Psi(G_k)$) in the frequency domain:

$$
\Upsilon(G_k) = E\left\{ [\mathcal{X}_k - G_k \mathcal{X}_k]^2 \right\} \tag{8}
$$

$$
\Psi(G_k) = \frac{1}{E\left\{ [G_k |N_k|]^2 \right\}} \tag{9}
$$

In Eq. 8, $\Upsilon(G_k)$ measures the clean speech distortion energy and, therefore, its value increases for increasing speech distortions. In Eq. 9, $\Psi(G_k)$ reflects the inverse of the noise energy remaining in the enhanced speech and increases for increasing noise reduction.

Figure 2 plots $\Upsilon(G_k)$ and $\Psi(G_k)$ vs. the frequency for different gains G_k as given by the associated estimators (average of 30 sentences, white noise, SNR = 0 dB).

As expected, we can see that the speech distortion increases as β is decreased while the noise reduction also increases. Therefore, the βSA estimator introduces more speech distortion energy but also performs better at noise reduction as β is decreased. The use of negative β enables extension of the trade-off between speech distortion and noise reduction. Since
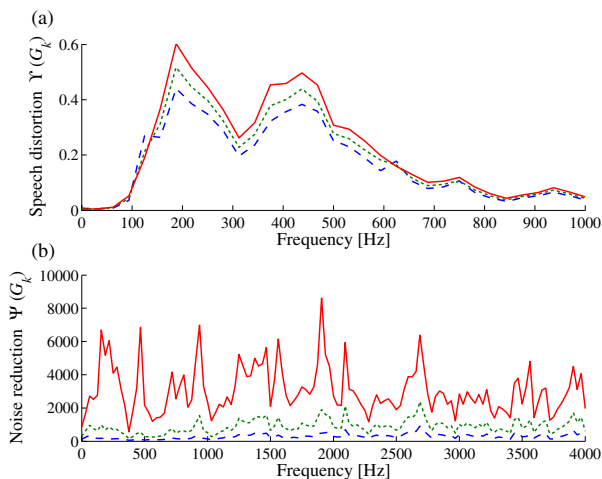
1595

Fig. 2. (a) Speech distortion metric vs. frequency (0 - 1000 Hz) (b) Noise reduction metric vs. frequency (0 - 4000 Hz) (average of 30 sentences, white noise, SNR = 0 dB) (– – – MMSE STSA; - - - LSA; ——— βSA, $\beta = -1$).

the energy of the speech signal is mainly at low frequencies, the speech distortion energy is also mainly located at low frequencies.

It was found, through informal listening tests, that the value of $\beta = -1$ was a good compromise between noise reduction and speech distortion, on the other hand, serious speech distortions were introduced when β became smaller than -1.5.

C. PESQ and MOS test results

We present comparative results for three estimators: MMSE STSA (or βSA with $\beta = 1$), LSA (or βSA with $\beta \to 0$) and βSA with $\beta = -1$. Thirty sentences from the TIMIT database, each sampled at 8 kHz, were used where 3 men and 3 women each spoke 5 sentences. Two types of noise were used from the NOISEX database (white and buccaneer-1 noises, the latter having mainly low frequency noise and a high frequency component) [8]. The observation frames were of 32ms and a 50% overlap was used between all frames in the overlap-add method used for the reconstruction of the enhanced speech. All algorithms used the *decision-directed* approach for the estimation of ξ_k [2] and a voice activity detector proposed in [9] was used to evaluate the noise spectral amplitude variance.

Table I presents the PESQ results on a scale from 1 to 4.5, with 4.5 being the best score. As can be seen, the βSA with $\beta = -1$ slightly outperforms MMSE STSA and LSA, in terms of PESQ, for almost all cases, being equivalent to LSA for buccaneer-1 noise at 10 dB SNR.

While PESQ is widely used to assess speech enhancement algorithm performances, it cannot handle all artifacts caused by such algorithms [10] and further assessment needs to be made using subjective tests.

In order to compare the results obtained with PESQ, we performed informal MOS subjective listening tests on 6 subjects using a subset of 4 sentences from the initial 30, each spoken by a different individual (2 men, 2 women). Therefore,

		Noisy speech	MMSE STSA	LSA	βSA ($\beta = -1$)
white	0 dB	1.29	1.39	1.44	1.47
	5 dB	1.37	1.60	1.70	1.72
	10 dB	1.58	1.83	1.95	1.96
buccaneer-1	0 dB	1.29	1.46	1.53	1.57
	5 dB	1.44	1.67	1.78	1.81
	10 dB	1.67	1.91	2.03	2.03

the average for each final MOS score is made over 24 scores. As suggested by ITU-T P.835 [11], MOS tests included an assessment of the speech distortion (5 = Not distorted, 1 = Very distorted), background noise (5 = Not noticeable, 1 = Very intrusive) and overall speech quality (5 = Excellent, 1 = Bad). Tests were performed in an isolated acoustic room using *beyerdynamic DT880* headphones.

		Noisy speech	MMSE STSA	LSA	βSA ($\beta = -1$)
white	Speech	3.9	2.4	2.9	2.8
	Background	1.2	2.2	2.5	2.9
	Overall	1.7	2.1	2.5	2.7
buccaneer-1	Speech	3.7	2.8	3.1	2.8
	Background	1.2	2.4	2.8	2.9
	Overall	1.8	2.4	2.8	2.5

When comparing LSA and βSA with $\beta = -1$ MOS test results (Table II), we see that the latter demonstrated more speech distortion but also more noise reduction than the former for both noises, as expected from section III-B. However, the overall perception was not the same for both noises. In fact, βSA with $\beta = -1$ was thought to be better than LSA for white noise but the inverse was found for buccaneer-1 noise. Also, based on Figure 2, MMSE STSA should have yielded less speech distortion than the other two estimators, however, this is not what we have observed. This could be due to the fact that, when a frame overlap of 50% is used, a perceivable echo is present in the MMSE STSA enhanced signal which is quite less perceivalble in LSA and βSA and may not have been well taken into account by Equation 8.

IV. CONCLUSION

In this paper, we proposed an extension of the βSA estimator analysis for $\beta < 0$. We showed a normalization effect as well as an increasing noise reduction for a decreasing β accompanied by increasing speech distortion. Furthermore, when setting $\beta = -1$ in the βSA estimator, we showed that it achieved better results in terms of PESQ than MMSE STSA and LSA. Finally, overall MOS appreciation for $\beta = -1$

is found to be better for white noise but inferior than LSA for the *buccaneer-1* noise. One interesting avenue will be to investigate the use of different values of $\beta > -2$ across the frequency axis.

REFERENCES

[1] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, Springer, 2005.

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

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

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

[5] J. H. L. Hansen, V. Radhakrishnan, and K. Hoberg Arehart, "Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system," *IEEE Trans. Audio Speech Language Processing*, vol. 14, no. 6, pp. 2049–2063, 2006.

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

[7] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio Speech Language Processing*, vol. 14, no. 4, pp. 1218–1234, July 2006.

[8] Rice University, "Signal processing information base: Noise data," [Online] Available http://spib.rice.edu/spib/select_noise.html, Accessed December 20, 2006.

[9] J. Sohn and N. S. Kim, "A statistical model-based voice activity detection," *IEEE Signal Processing Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[10] "P.862.3: Application guide for objective quality measurement based on recommendations P.862, P.862.1 and P.862.2 (prepublication)," Tech. Rep., ITU-T, 2005.

[11] "P.835: Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," Tech. Rep., ITU-T, 2003.