ADAPTIVE β-ORDER MMSE ESTIMATION FOR SPEECH ENHANCEMENT

ChangHuai You[+], *SooNgee Koh**, *Susanto Rahardja*[+]

[+] Institute for Infocomm Research, Singapore 119613
* School of EEE, Nanyang Technological University, Singapore 639798

ABSTRACT

This paper introduces an adaptive β-order minimum mean square error (MMSE) spectral estimator for the short time spectral amplitude (STSA) of speech. The characteristic of β-order MMSE attenuation function is introduced and analyzed. The performance of the proposed adaptive β-order MMSE has been thoroughly examined by a large number of computer simulations. The new proposed scheme has been found to outperform the conventional a priori SNR Wiener filtering, the Ephraim & Malah STSA-MMSE and Log Spectral Amplitude (LSA) schemes. It can achieve a more significant noise reduction and a better spectral estimation for weak speech components from a noisy speech signal as compared to the conventional schemes.

1. INTRODUCTION

The main objective of speech enhancement is to reduce the corrupting noise component of a noisy speech signal while preserving the original clean speech quality as much as possible. Though research on speech enhancement has been going on for a long time, issues such as distortions to the original speech signal and residual noise sometimes in the form of musical tones created by the enhancement algorithms remain unsolved. A large number of research papers on different approaches and methods have managed to address these problems with varying degrees of success.

In most practical situations, speech signals are degraded by additive noise in a car environment, air traffic control communication, cocktail party environment, etc. Speech enhancement is needed in these situations. It is also needed for speech coding and speech recognition systems to improve their coding and recognition performance.

For single microphone speech enhancement, many algorithms, including conventional spectral subtraction [1], speech estimation based on uncertainty of speech presence [2], model based speech enhancement, the Ephraim & Malah (EM) MMSE [3] and LSA [4], have been reported.

One of the main approaches of speech reduction algorithms is to obtain the best possible estimate of the short time spectra of a speech signal from a given noisy speech. In [5], the proposed approach is to estimate the short time spectral magnitudes, $|S_k|$, of a speech signal by minimizing $[|X_k|^\beta - E\{|N_k|^\beta\}]^{1/\beta}$ for some constant values of β. k is the frequency bin index, S_k, X_k and N_k are Fourier transforms of a windowed segment of speech, noisy speech and noise, respectively.

The advantage of the EM noise suppression method is built upon the non-linearity of the averaging procedure. When the speech level is well above the noise level, the a priori SNR estimation equation involves a mere one-frame delay, and the estimate is no longer a smoothed SNR estimate, which is important in the case of non-stationary signal [6]; when the speech signal level is close to or below the noise level, the a priori SNR estimation equation has a smoothing property and the musical tone phenomenon is greatly reduced. Therefore, the total effect of noise suppression is improved as compared to other conventional methods.

In order to investigate the characteristics and performance of the β-order MMSE method based on the assumption that speech and noise spectral amplitudes are Gaussian distributed, the use of the cost function $J = E\{(A_k^\beta - \hat{A}_k^\beta)^2\}$ as an estimation criterion is considered in this paper. $\hat{A}_k$ is the estimate of spectral amplitude of the speech signal whose spectral component is $S_k = A_k e^{j\alpha_k}$. To obtain a more accurate estimate and achieve sufficient suppression of noise as well as minimal musical tones in the residual signal, an adaptive β-order MMSE method is discussed and its performance analyzed.

2. β-ORDER MMSE SHORT-TIME SPECTRAL SUPPRESSION

An observed noisy speech signal $x(t)$ is assumed to be a clean speech signal $s(t)$ degraded by uncorrelated additive noise $n(t)$, i.e.,

$$x(t) = s(t) + n(t), \quad 0 \le t \le T. \tag{1}$$

Let $S_k = A_k e^{j\alpha_k}$, N_k, $X_k = R_k e^{j\vartheta_k}$ denote the kth spectral component of the clean speech signal $s(t)$, noise $n(t)$ and the observed noisy speech $x(t)$, respectively. On the basis of the Gaussian statistical model [3], we are looking for the estimate $\hat{A}_k$, which minimizes the following dis-

tortion measure:

$$J = E\{(A_k^\beta - \hat{A}_k^\beta)^2\}. \quad (2)$$

Under the assumed Gaussian statistical model, we obviously have [3]

$$\hat{A}_k = \sqrt[\beta]{E\{A_k^\beta | X_k\}}. \quad (3)$$

Based on the above assumption, the evaluation of $E(A_k^\beta|X_k)$ is given by

$$E\{A_k^\beta|X_k\} = \frac{\int_0^\infty \int_0^{2\pi} a_k^\beta p(X_k|a_k,\alpha_k)p(a_k,\alpha_k)\, d\alpha_k\, da_k}{\int_0^\infty \int_0^{2\pi} p(X_k|a_k,\alpha_k)p(a_k,\alpha_k)\, d\alpha_k\, da_k}. \quad (4)$$

With the Gaussian model assumption, $p(X_k|a_k,\alpha_k)$ and $p(a_k,\alpha_k)$ are derived in [3] and repeated below:

$$p(X_k|a_k,\alpha_k) = \frac{1}{\sqrt{\pi\eta_n(k)}} exp\{-\frac{|X_k - a_k e^{j\alpha_k}|^2}{\eta_n(k)}\}, \quad (5)$$

$$p(a_k,\alpha_k) = \frac{a_k}{\sqrt{\pi\eta_s(k)}} exp\{-\frac{a_k^2}{\eta_s(k)}\}, \quad (6)$$

where $\eta_n(k) = E[|N_k|^2]$, and $\eta_s(k) = E[|S_k|^2]$ are the variances of the kth spectral components of noise and the speech signal, respectively. On substituting Eqs. (5) and (6) into Eq. (4), and using the integral representation of the modified Bessel function of zero order $I_0(.)$ [[7], Eqs. 8.406.3, 8.411.1, 6.631.1, 9.212.1] [4], we obtain

$$\begin{aligned} &E\{A_k^\beta|X_k\} \\ &= \frac{\int_0^\infty a_k^{\beta+1} exp(-a_k^2/\eta(k)) I_0(2a_k\sqrt{\upsilon_k/\eta(k)})\, da_k}{\int_0^\infty a_k exp(-a_k^2/\eta(k)) I_0(2a_k\sqrt{\upsilon_k/\eta(k)})\, da_k} \\ &= \eta(k)^{\beta/2}\Gamma(\beta/2+1)M(-\beta/2;1;-\upsilon_k). \end{aligned} \quad (7)$$

$\Gamma(\cdot)$ is the gamma function and M($\alpha;\gamma;z$) is the confluent hypergeometric function [[7], Eq. 9.210.1], and $\eta(k)$ and υ_k are defined as follows:

$$\eta(k) = (\frac{1}{\eta_s(k)} + \frac{1}{\eta_n(k)})^{-1}, \quad \upsilon_k = \frac{\xi_k}{1+\xi_k}\gamma_k, \quad (8)$$

where ξ_k and γ_k represent the a priori SNR and a posteriori SNR respectively [3, 4],

$$\xi_k = \frac{\eta_s(k)}{\eta_n(k)}, \quad \gamma_k = \frac{|X_k|^2}{\eta_n(k)}. \quad (9)$$

The estimate of the amplitude of the speech signal is then obtained by

$$\begin{aligned} \hat{A}_k &= [E\{A_k^\beta|X_k\}]^{1/\beta} \\ &= \eta(k)^{1/2}[\Gamma(\beta/2+1)M(-\beta/2;1;-\upsilon_k)]^{1/\beta}, \end{aligned} \quad (10)$$

and the estimate of speech signal is given as follows:

$$\hat{S}_k = G_\beta(\xi_k,\gamma_k)X_k, \quad (11)$$

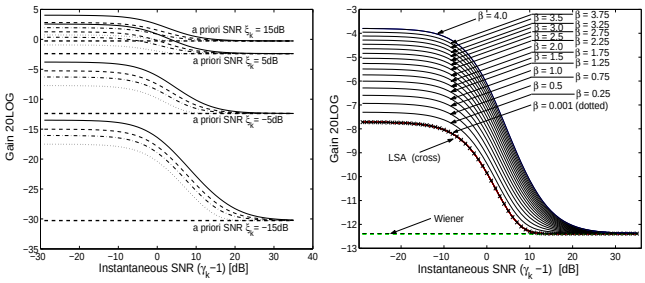
here $G_\beta(\xi_k,\gamma_k)$ is the gain function which is given by

$$G_\beta(\xi_k,\gamma_k) = \frac{\sqrt{\upsilon_k}}{\gamma_k}[\Gamma(\beta/2+1)M(-\beta/2;1;-\upsilon_k)]^{1/\beta}. \quad (12)$$

Fig. 1(a) shows the gain curves of the β-order MMSE estimator in comparison with the Wiener estimator as a function of γ_k and ξ_k. Fig. 1(b) shows the gain curves with different β values for ξ_k equals to -5dB. The a priori SNR, ξ_k, can be best estimated by the decision-directed approach proposed in [3] and is described as follows:

$$\begin{aligned} \hat{\xi}_k(l) &= (1-\alpha)max(\gamma_k(l)-1,0) \\ &+ \alpha\frac{|G_\beta(\hat{\xi}_k(l-1),\gamma_k(l-1))X_k(l-1)|^2}{\eta_n(k)}. \end{aligned} \quad (13)$$

Normally, the parameter α is set to 0.98 [3, 6].



(a) (b)

Fig. 1. (a) Gain versus instantaneous SNR ($\gamma_k - 1$) in comparison with the Wiener gain (bold-dashed) for $\beta = 0.001$ (dotted), $\beta = 1.0$ (dotdash), $\beta = 2.0$ (dashed) and $\beta = 4.0$ (solid), and a priori SNR $\xi_k = -15, -5, 5, 15dB$; (b) Gain versus instantaneous SNR ($\gamma_k - 1$), for different β values and a priori SNR $\xi_k = -5dB$.

3. DISCUSSION ON CONSTANT β VALUE AND A NEW PROPOSAL OF ADAPTIVE β VALUE

From Fig. 1(a), it is noted that whatever β value is, the gain always converges to the Wiener gain value if the instantaneous SNR ($\gamma_k - 1$) is big enough for a certain a priori SNR ξ_k value. In Fig. 1(b), we can see that when β is very close to 0 ($\beta = 0.001$), the gain curve is very close to the LSA (EM) [4] gain curves. When $\beta = 1$, it is exactly the same as the STSA-MMSE (EM) [3] gain curve.

The elimination of musical tones of the STSA-MMSE (EM) [3] scheme is described in [6]. It is mainly due to the effectiveness of the so called 'decision-directed approach'

for estimating the a priori SNR, ξ_k. Obviously, this musical tone elimination feature can also be applied to the β-order MMSE described by Eq. (12) for any value of β.

Fig. 2(a) shows the gain curves as a function of β for different ξ_k values when the instantaneous SNR ($\gamma_k - 1$) is equal to 0 dB. Fig. 2(b) shows the gain curves as a function of β value, for different a posteriori SNR, γ_k, for the case of a priori SNR, ξ_k, equals to 0 dB.

Fig. 2(a) shows that gain increases as the value of β increases. Fig. 2(b) shows that the smaller the value of the instantaneous SNR ($\gamma_k - 1$) is, the bigger the increment of gain (in dB) is as β increases, for the case of 0 dB a priori SNR. This particular characteristic of gain is very important to speech recovery for the weak speech spectral components. Usually, for conventional speech enhancement methods, when a speech spectral component with high SNR value is suitably enhanced by attenuating the spectral component of the noisy speech signal with the appropriate gain value, the weak speech spectral components will be overly attenuated if the attenuation is applied to all the spectral components of the same frame. In β-order MMSE, we can exploit the relation between β and gain by adjusting β to a proper value in order not to over-attenuate the weak speech components in a frame. In other words, for a small value of β used on a frame of noisy speech samples, the strong speech spectral components can be appropriately enhanced but the weak speech spectral components will be lost; when the value of β is big, the strong speech spectral components remain almost unchanged when attenuating the noise part because the gain always converges to the Wiener gain value when the a posteriori SNR is big enough. However the weak spectral components may be appropriately enhanced because the gain has a bigger increment as β value is bigger for a low a posteriori SNR.

If a big value of β is used for a speech absence frame, the musical tones will be enlarged. If too small a β value is used in speech presence frame, the weak speech spectral components will be lost although the musical tones will also be very much attenuated in the same frame. If we increase β value in a speech presence frame, the spectral components with weak SNR will be raised and therefore they could be better estimated, although the noise (musical tones) will be enlarged correspondingly to a certain degree. Fortunately, according to the acoustic masking principle, the weak musical tones could be masked by the strong speech signal at the same time. The effects of big β value acting on the strong speech frame and small β value acting on the low SNR frame are the salient attributes of the proposed β-order MMSE method. Based on the above discussion, we arrive at the appropriate β value for a particular frame l, i.e., we can make β a function $\mathbf{F}$ of frame SNR $\Xi(l)$. We express the equation as follows:

$$\beta(l) = \mathbf{F}(\Xi(l)). \tag{14}$$

The frame SNR of the current frame l can be defined as:

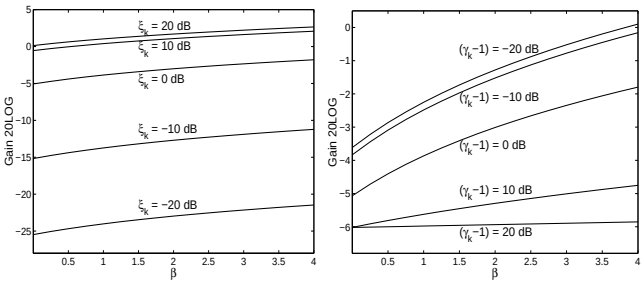
$$\Xi(l) = 10\log_{10}\frac{\sum_{k=0}^{N/2}\left|R_k(l) - \sqrt{\eta_n(l,k)}\right|^2}{\sum_{k=0}^{N/2}\eta_n(l,k)}. \tag{15}$$

Based on the above observation, it is expected that β will increase as $\Xi(l)$ increases, and it will decrease otherwise. Here a linear relationship between β and $\Xi(l)$ is applied as follows:

$$\beta(l) = \alpha_1\Xi(l) + \alpha_2, \tag{16}$$

where α_1 and α_2 denote linear coefficients. From Eq. (12), we have a constraint condition: i.e., $\beta > 0$. In practical application, we can define the dynamic range of β value as

$$\beta(l) = max(min(\alpha_1\Xi(l) + \alpha_2, \alpha_3), \alpha_4). \tag{17}$$



(a) (b)

Fig. 2. (a) Gain versus β value for a priori SNR ξ_k = -20, -10, 0, 10, 20 dB and instantaneous SNR (γ_k-1) = 0 dB; (b) Gain versus β value for instantaneous SNR (γ_k-1) = -20, -10, 0, 10, 20 dB and a priori SNR ξ_k = 0 dB.

4. PERFORMANCE EVALUATION

We use noise data from the NOISEX-92 database in our performance evaluation. The frame size for 16 kHz sampling rate case is 512 samples, which are Hamming windowed with 75% overlap between adjacent frames. The evaluation parameters are as follows: α_1=0.25, α_2=1.75, α_3=4.00, α_4=0.001, α=0.98.

A total of 10 different utterances from the TIMIT database are used in our evaluation. Half of the utterances are male and the other half female. Fig. 3 shows the average segmental SNR improvement arising from the use of Eq. (12) with different values of β as well as in the case of our adaptive β-order method based on Eqs. (12) through (17), where the input segmental SNR is adjusted to -10dB, -5dB, 0dB, 5dB, and 10dB respectively. From the figure, we can see that

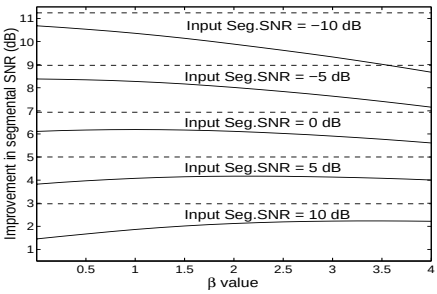


Fig. 3. Performance of the speech estimator based on Eq. (12) for different β values (solid), and the proposed adaptive β-order method (dashed) at 16 kHz sampling rate for the F16 cockpit noise case.

the proposed β-order method always outperforms the other methods in terms of average segmental SNR improvement.

As the segmental SNR does not reveal the spectral characteristics of the residual noise, another comparison was made using speech spectrograms. The perceptual quality of the enhanced speech is also assessed by means of subjective listening tests. From the speech spectrograms shown in Fig. 4, it is clear that the proposed adaptive β-order MMSE estimator could restore more spectral details of the original speech as compared to the LSA scheme. Our informal subjective listening tests also confirmed the better performance of the proposed scheme.

5. CONCLUSION

The focus of our speech enhancement study is to develop an optimal noise reduction algorithm that would maximize noise reduction while minimizing speech distortion. In this paper, the characteristics of β-order STSA MMSE is introduced, and we propose an adaptive β-order STSA-MMSE speech enhancement method based on the characteristics of β-order MMSE to achieve very effective speech enhancement. We compare the new adaptive β-order method with other speech enhancement estimators through computer simulations to show the effectiveness of the proposed adaptive β-order MMSE method. It has been verified through a large amount of computer simulations that the proposed adaptive β-order method outperforms many conventional methods and has potential for minimizing both speech distortion and residual noise, especially for the case of weak spectral components of speech signal corrupted by noise.

6. REFERENCES

[1] J.S. Lim and A.V. Oppenheim, "Enhancement And Band-Width Compression Of Noisy Speech," *Proceedings of The IEEE*, Vol. 67, No. 12, pp. 1586-1604, Dec. 1979.

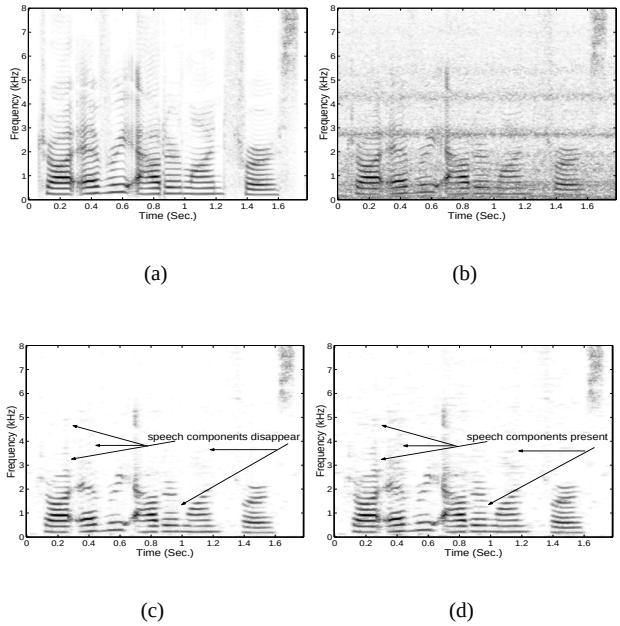


Fig. 4. Speech spectrograms: (a) Clean speech (16kHz sampling rate) (b) Noisy speech (F16 noise) with SegSNR = 0 dB (c) LSA estimated speech signal (SegSNR = 7.06dB) (d) Proposed adaptive β-order estimated speech signal (SegSNR=7.98dB).

[2] I.Y. Soon, S.N. Koh, and C.K. Yeo, "Improved Noise Suppression Filter Using Self-Adaptive Estimator Of Probability Of Speech Absence," *Signal Processing*, Vol.75, No.2, pp 151-159, June 1999.

[3] Y. Ephraim and D. Malah, "Speech Enhancement Using A Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-32, No. 6, pp. 1109-1121, Dec. 1984.

[4] Y. Ephraim and D. Malah, "Speech Enhancement Using A Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-33, No. 2, pp. 443-445, Apr.1985.

[5] J.S. Lim, "Evaluation Of A Correlation Subtraction Method Enhancing Speech Degraded By Additive White Noise," *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-26, No. 5, pp. 471-472, Oct. 1978.

[6] O. Cappé, "Elimination Of The Musical Noise Phenomenon With The Ephraim And Malah Noise Suppressor," *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 2, pp. 345-349, 1994.

[7] I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series, and Products*. New York: Academic, 1980.