Improved Adaptive β-Order MMSE Speech Enhancement

Chang Huai You*, Soo Ngee Koh†, Haizhou Li*, Susanto Rahardja*
* Institute for Infocomm Research, Singapore
E-mail: {echyou, hli, rsusanto}@i2r.a-star.edu.sg
† Nanyang Technological University, Singapore
E-mail: esnkoh@ntu.edu.sg

***Abstract*—This paper considers a single channel speech enhancement algorithm, which is based on our previous work on β-order minimum mean square error (MMSE) spectral estimation. We propose to make β a function of both local and frame signal-to-noise ratios (SNRs) in order to achieve more effective preservation of weak speech components. Moreover, by taking into account the speech-presence uncertainty in the adaptive β-order MMSE algorithm, we achieve a significant noise reduction and an improved spectral estimation of weak speech components. Experiments also show that the proposed estimator outperforms other well known speech enhancement algorithms.**

I. Introduction

For single microphone speech enhancement using spectral domain, many approaches, including Wiener filtering, spectral subtraction, maximum-likelihood noise attenuation, masking properties based over-subtraction method, the Ephraim-Malah (E-M) MMSE [1] and Log-Spectral Amplitude (LSA) [2] methods, as well as uncertainty of speech-presence [3], have been reported.

In [4], the elimination of musical noise phenomenon with the E-M suppression method is analyzed; it proves that the E-M noise suppressor is effective if a nonlinear smoothing procedure is used to obtain more consistent estimates of the *a priori* and *a posteriori* SNRs which are used to control the attenuation function. The advantage of the E-M noise suppression method is obtained from the non-linearity of the averaging procedure and the decision-directed *a priori* SNR estimation. 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 [4]; 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. In [5], an improved estimation is proposed over the decision-directed estimation by introducing a recursive method based on the assumption of statistical independence of not only spectral components but also time-domain components.

In this paper, based on the adaptive β-order MMSE method [6], we propose an improved method by considering the local factor to the process of estimation, which would preserve weak spectral components. Furthermore, speech-presence uncertainty is considered for the reduction of noise. Comparing with the masking-based β-order MMSE method [7] which preserves the weak spectral components by not suppressing the inaudible noise, we emphasize the local SNR information to preserve the weak detectable components, and use speech-presence probability to achieve a sufficient suppression of noise. This paper is aimed to reduce the additive noise on the frequency domain so that the time-domain speech signal can be enhanced for the human listening. Recently, MMSE has also been applied for the feature enhancement in cepstral domain [8]. However, the speech signal after feature enhancement processing can not return back to the time-domain speech signal for human listening. As a result, it only benefits the machine recognition.

II. An Improved Adaptive β-Order MMSE

A. 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 and $X_k = R_k e^{j\vartheta_k}$ denote the kth spectral component of the clean speech signal $s(t)$, noise $n(t)$ and the observed noisy speech $x(t)$, respectively. We have the β-order MMSE suppression gain function as follows [6]

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

where β is the order of the spectral amplitude of the signal while the MMSE criterion is used, $\Gamma(\cdot)$ is the gamma function and $M(\alpha;\gamma;z)$ is the confluent hypergeometric function. υ_k is defined by

$$\upsilon_k = \frac{\xi_k}{1+\xi_k}\gamma_k \tag{3}$$

where ξ_k and γ_k represent the *a priori* SNR and *a posteriori* SNR respectively. Let $\eta_n(k) = \mathbf{E}\{|N_k|^2\}$, and $\eta_s(k) = \mathbf{E}\{|S_k|^2\}$ denote the variances of the kth spectral components of noise and speech signal respectively, we have

$$\xi_k = \frac{\eta_s(k)}{\eta_n(k)}, \quad \gamma_k = \frac{R_k^2}{\eta_n(k)}. \tag{4}$$

The *a priori* SNR, ξ_k, is estimated by a causal recursive estimation proposed in [5] as follows

$$\hat{\xi}_k(l/l) = \frac{\hat{\xi}_k(l/l-1)}{1+\hat{\xi}_k(l/l-1)}(1+\frac{\hat{\xi}_k(l/l-1)}{1+\hat{\xi}_k(l/l-1)}\gamma_k), \quad (5)$$

$$\hat{\xi}_k(l/l-1) = \max\Big\{(1-\alpha)\hat{\xi}_k(l-1/l-1) + \alpha\frac{\hat{A}_k(l-1)^2}{\eta_n(k,l-1)}, \quad \xi_{min}\Big\}, \quad (6)$$

where l is the time frame index, and parameters ξ_{min} = -25 dB, and α = 0.9.

B. Proposed Adaptive β Value

It is noted that the suppression gain function (Eq. (2)) is obtained based on the assumption that speech and noise spectral components are statistically independent of each other. In other words, the estimate of a speech spectral amplitude of a certain frequency bin is only determined by the *a priori* and *a posteriori* SNRs of that frequency bin, and is independent of the SNRs of the other frequency bins. However, it is well-known that a speech signal is not a chaotic signal and the statistical independence assumption is not sufficiently satisfied in practice. Based on the above analysis, we can make β a function of frame-SNR $\Xi(l)$ and local-SNR $\Lambda(l,k)$, i.e., $\beta(l,k) = \mathbf{F}\big(\Xi(l),\Lambda(l,k)\big)$. The frame-SNR and local-SNR of the current frame l are defined respectively by

$$\Xi(l) = 10\log_{10}\frac{\sum_{k=0}^{N/2}\psi(l,k)^2}{\sum_{k=0}^{N/2}\eta_n(l,k)} \quad (7)$$

$$\Lambda(l,k) = 10\log_{10}\frac{\sum_{i=-b}^{b}\sum_{j=-a}^{d} w(i,j)\psi(l-j,k-i)^2}{\sum_{i=-b}^{b}\sum_{j=-a}^{d} w(i,j)\eta_n(l-j,k-i)} \quad (8)$$

where w is a time-frequency window with size $(2b+1)\times(a+d+1)$, d and a denote the numbers of time-domain past and future samples used for smoothing for frequency bin k at time l, and $\psi(l,k)$ is defined as follows

$$\psi(l,k) = \max[R_k(l) - \sqrt{\eta_n(l,k)}, \quad \varepsilon]. \quad (9)$$

In our study, b = 1, a = 1, d = 2 and the matrix w is given by

$$w = \begin{bmatrix} 0.35 & 0.5 & 0.7 & 0.5 \\ 0.50 & 0.8 & 1.0 & 0.8 \\ 0.35 & 0.5 & 0.7 & 0.5 \end{bmatrix}. \quad (10)$$

We may express the β value as a function of the two variables in polynomial form, i.e.,

$$\beta(l,k) = \mathbf{F}\big(\Xi(l),\Lambda(l,k)\big) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} c_{ij}\Xi(l)^i\Lambda(l,k)^j$$
$$= \tau_0 + \tau_1\Xi(l) + \tau_2\Lambda(l,k) + \tau_3\Xi(l)\Lambda(l,k) + \Phi(O_h). \quad (11)$$

By ignoring the high-order polynomial terms $\Phi(O_h)$, and making β a monotonically non-decreasing function of $\Xi(l)$ and $\Lambda(l,k)$, we can express $\beta(l,k)$ approximately in the following form

$$\beta(l,k) \cong \tau_0 + \tau_1\Xi(l) + \tau_2\Lambda(l,k) + \tau_3\max[\Xi(l)+\tau_4,0]\max[\Lambda(l,k)+\tau_5,0] \quad (12)$$

where c_{ij} and τ_i $(i,j = 0,1,2,...,\infty)$ denote the polynomial coefficients.

From Eq. (2), we can see that the gain increases as the value of β increases, and the smaller the value of the instantaneous SNR $(\gamma_k - 1)$ is, the bigger the increment of gain (in dB) will be as β increases. This particular property of gain is useful to bring about the retrieval of speech for the weak speech spectral components. For a low 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 may disappear. When the value of β is high, the strong speech spectral components remain at almost the same enhancement level as in the case of low β value because the gain always converges to the Wiener gain value when the instantaneous SNR is big enough. However, the weak spectral components (i.e. low instantaneous SNR, $(\gamma_k - 1)$) which exist in the same frame as strong spectral components of speech signal may be appropriately enhanced because the gain has a big value for the low instantaneous SNR spectral components with high β value.

It is expected that β will increase as $\Xi(l)$ increases and vice versa. From Eq. (12), we can see that the parameters $\tau_i, i = 0,1,...,5$ are very important to the speech enhancement system. Therefore, the process of obtaining the parameters is a key issue in our adapted β-order design. Actually, τ_0 is a floor used in determining the total level of β value, and τ_1 and τ_2 is used to determine the level of influence the frame and local SNRs have on the suppression gain value respectively. τ_3 is used to constrain the overall effect of frame and local SNRs on the β value. τ_4 and τ_5 are to set a lower bound contributions to the gain function from frame and local SNRs. Through a large number of computer simulation using real speech data, the τ_i $(i = 0,1,2,...)$ parameters are appropriately adjusted and the concrete β function is given as follows

$$\beta(l,k) = \begin{cases} 0.01\max[\Xi(l)+15,0]\max[\Lambda(l,k)+1,0] + 0.53\Xi(l) \\ \quad + 0.2\Lambda(l,k) + 2.95, & \text{if} \quad \Xi(l) > \text{-5 (dB)} \\ 0.01\max[\Xi(l)+15,0]\max[\Lambda(l,k)+1,0] + 0.53\Xi(l) \\ \quad + 0.3\max[\Lambda(l,k)+5,0] + 2.95, & \text{otherwise.} \end{cases} \quad (13)$$

To emphasize the dynamic range of the β value, we express the final form of β as follows

$$\hat{\beta}(l,k) = \min\big\{\max\{\beta(l,k), \quad \tau_6\}, \quad \tau_7\big\} \quad (14)$$

where τ_6 = 0 and τ_7 = 4 are used in our simulation experiments.

C. Incorporating Speech-Presence Uncertainty

1) Gain: In this paper, we further consider the effect of the speech-presence uncertainty to the adaptive β-order MMSE. Let H_0 and H_1 represent speech-absence and speech-presence respectively in a binary hypothesis model. The probability density functions (pdfs) of $X_k(l)$ with Gaussian distribution during speech-absence and speech-presence are respectively given by

$$p(X_k(l)|H_0(k,l)) = \frac{1}{\pi\eta_n(k,l)} exp\big[-\frac{|X_k(l)|^2}{\eta_n(k,l)}\big] \tag{15}$$

and

$$\begin{aligned} &p(X_k(l)|H_1(k,l)) \\ &= \frac{1}{\pi(\eta_n(k,l)+\eta_s(k,l))} exp\big[-\frac{|X_k(l)|^2}{\eta_n(k,l)+\eta_s(k,l)}\big]. \end{aligned} \tag{16}$$

Based on the Bayes' theorem, we obtain the following equation

$$\frac{p(H_1(k,l)|X_k(l))p(X_k(l))}{p(H_0(k,l)|X_k(l))p(X_k(l))} = \frac{p(X_k(l)|H_1(k,l))p(H_1(k,l))}{p(X_k(l)|H_0(k,l))p(H_0(k,l))}. \tag{17}$$

Since $p(H_1(k,l)|X_k(l)) + p(H_0(k,l)|X_k(l)) = 1$, we have

$$\begin{aligned} p(H_1(k,l)|X_k(l)) &= \frac{1}{1+\frac{p(X_k(l)|H_0(k,l))p(H_0(k,l))}{p(X_k(l)|H_1(k,l))p(H_1(k,l))}} \\ &= \frac{1-q_k(l)}{1-q_k(l)+q_k(l)(1+\xi_k)exp(-\upsilon_k)} \end{aligned} \tag{18}$$

where $q_k(l) = p(H_0(k,l))$ and $p(H_1(k,l)) = 1 - q_k(l)$. The estimate of speech amplitude is then given by

$$\begin{aligned} \hat{A}_k^\beta =& \mathbf{E}\{A_k^\beta|X_k(l), H_1(k,l)\}p(H_1(k,l)|X_k) \\ &+ \mathbf{E}\{A_k^\beta|X_k(l), H_0(k,l)\}p(H_0(k,l)|X_k). \end{aligned} \tag{19}$$

Generally, as illustrated in [1], the second term is 0 and so the estimate of speech amplitude can be obtained by

$$\hat{A}_k = G_\beta(\xi_k,\gamma_k)p(H_1(k,l)|X_k)^{1/\beta}R_k. \tag{20}$$

The gain function with the factor of uncertainty of speech-presence is obtained by

$$\begin{aligned} G_M(\xi_k,\gamma_k,\beta) &= G_\beta(\xi_k,\gamma_k)p(H_1(k,l)|X_k)^{1/\beta} \\ &= \frac{\sqrt{\upsilon_k}}{\gamma_k}\big[\Gamma(\beta/2+1)M(-\beta/2;1;-\upsilon_k) \\ &\quad \times \frac{1-q_k(l)}{1-q_k(l)+q_k(l)(1+\xi_k)exp(-\upsilon_k)}\big]^{1/\beta}. \end{aligned} \tag{21}$$

Consequently, the estimate of a speech spectral component is expressed as

$$\hat{S}_k = G_M(\xi_k,\gamma_k,\beta)X_k. \tag{22}$$

2) Estimation of Speech Absence Probability: In [1], $q_k(l)$ could be fixed to a value of 0.2. If a more accurate estimate of the time-varying value of $q_k(l)$ is used, the estimate of the speech signal will be closer to its original value. Hence, we adopt the following method to estimate $q_k(l)$

$$p_0(k,l) = (1-\alpha_p)p_0(k,l-1) + \alpha_p\Omega(k,l) \tag{23}$$

$$\Omega(k,l) = \begin{cases} 0, & \text{if } \rho_{cur}(k,l) > \hbar_p \\ 1, & \text{if } \rho_{cur}(k,l) \le \hbar_p \end{cases} \tag{24}$$

$$q_k(l) = 0.15 + 0.15p_0(k,l) \tag{25}$$

where the current-SNR, $\rho_{cur}(k,l)$, is defined as

$$\rho_{cur}(k,l) = 10\log_{10}\frac{\{\max[R_k(l) - \sqrt{\eta_n(k,l)}, \quad \varepsilon]\}^2}{\eta_n(k,l)}. \tag{26}$$

Usually, $\alpha_p = 0.85$ and $\hbar_p = 2 \sim 6$ (dB) are suggested; ε is a very small positive number, e.g., $\varepsilon = 2.22 \times 10^{-16}$.

III. Performance Evaluation

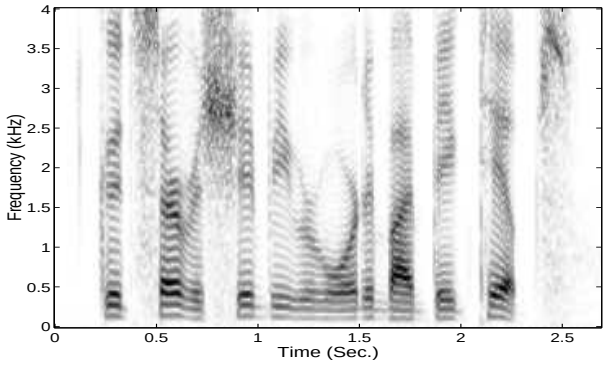
The performance evaluation is based on spectrogram analysis, segmental SNR measure, and listening tests. Five different noise types, taken from the NOISEX-92 database, are used. They are white Gaussian noise, Babble noise, interior Volvo car noise, Leopard noise and F16 cockpit noise. 30 utterances from the TIMIT database and 20 utterances from a Mandarin speech database are used in the simulation. Half of the utterances are male and the other half are female. The effectiveness of the proposed enhancement algorithm is evaluated for the 8 kHz sampling rate with the frame size of 256 samples, which are Hamming windowed with 75% overlap.

Fig. 1 shows the recovered spectral components, which are corrupted by F16 noise, obtained by the proposed speech enhancement methods in comparison with other estimators.

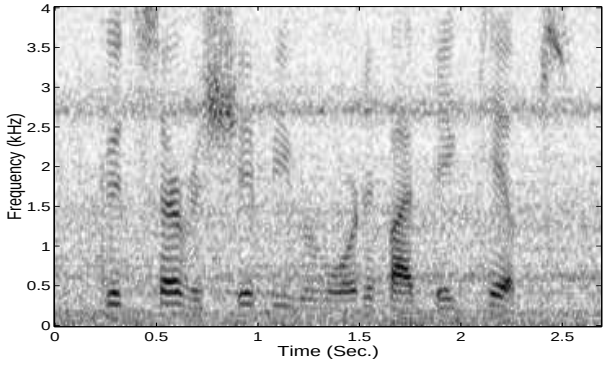
Fig. 2 shows the segmental SNR improvement performances of the different speech estimation algorithms which include MMSE [1], LSA [2], OM-LSA [3], conventional adaptive β-order method [6], and the proposed method. From these figures, we can see that the proposed method always outperforms the other methods in terms of average segmental SNR improvement for the case of white noise. Its performance is even more impressive for car interior noise. Listening tests also confirm the advantages of our adaptive β-order method based on uncertainty of speech-presence.

IV. Conclusion

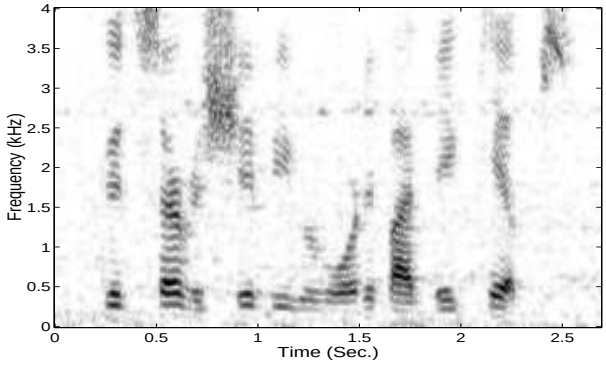
In this paper, a new adaptive β-order MMSE algorithm is proposed. The new algorithm demonstrates a considerable improvement over the conventional adaptive β-order MMSE algorithm. A significant improvement is also achieved when the algorithm incorporates speech presence uncertainty. The paper shows, through computer simulations, that the proposed method outperforms many conventional methods and has the potential of minimizing both speech distortion and residual



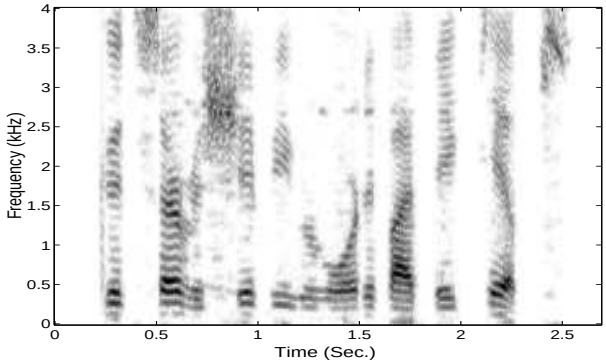
(a)



(b)



(c)



(d)

Fig. 1. Speech spectrograms: (a) Clean speech (8kHz sampling rate) (b) Noisy speech (F16 noise) with segSNR = -7.91 dB (c) LSA estimated speech signal (segSNR = 1.60 dB) (d) Proposed adaptive β-order with speech-presence uncertainty estimated speech signal (segSNR = 5.81 dB).

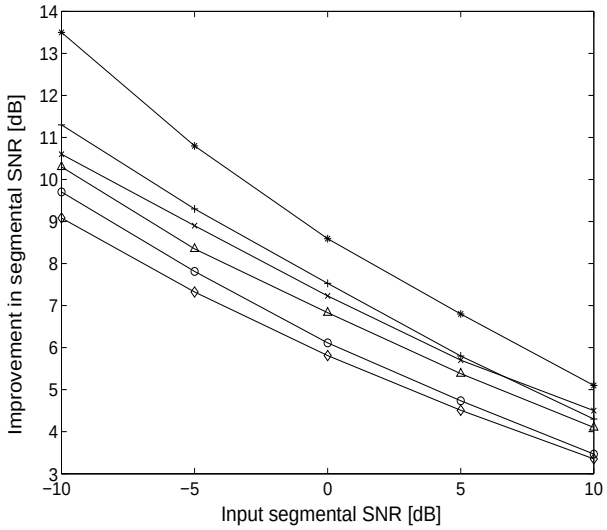


Fig. 2. Performance of the different speech estimation algorithms at 8kHz sampling rate; MMSE ($\diamondsuit$), LSA (o), OM-LSA (+), conventional adaptive β-order ($\triangle$), proposed adaptive β-order (x) and proposed speech-presence uncertainty based adaptive β-order (*). It shows the average segmental SNR improvement for White Gaussian noise.

noise. In particular, the enhancement effect is more significant for the case of weak spectral components of a speech signal corrupted by noise.

REFERENCES

[1] 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.

[2] 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.

[3] I. Cohen and B. Berdugo, "Speech Enhancement for Non-Stationary Noise Environments," *Signal Processing*, Vol. 81, No. 11, pp. 2403-2418, Nov. 2001.

[4] 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.

[5] I. Cohen, "On the Decision-Directed Estimation Approach of Ephraim and Malah," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-04, Vol. 1, pp. 293-296, May 2004.

[6] C.H. You, S.N. Koh and S. Rahardja, "β-Order MMSE Spectral Amplitude Estimation for Speech Enhancement," *IEEE Trans. on Speech and Audio Processing*, Vol.13, pp.475-486, Jul. 2005.

[7] C.H. You, S.N. Koh, and S. Rahardja, "Masking-Based β-Order MMSE Speech Enhancement", *Speech Communication*, Vol. 48, Issue 1, pp. 57-70, Jan. 2006.

[8] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, "Robust speech recognition using cepstral minimum-mean-square-error noise suppressor," *in IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 5, July 2008.