AN MMSE SPEECH ENHANCEMENT APPROACH INCORPORATING MASKING PROPERTIES

Chang Huai YOU[+]*, Soo Ngee KOH*, Susanto RAHARDJA*[+]

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

ABSTRACT

This paper describes a new speech enhancement approach which employs adaptive β-order minimum mean square error (MMSE) spectral estimation of short time spectral amplitude (STSA) of a speech signal. In the proposed approach, human perceptual auditory masking effect is incorporated into the speech enhancement algorithm. The relationship between the value of β and the noise masking threshold is considered. The algorithm is based on a criterion by which the audible noise may be masked rather than being attenuated, and thereby reducing the chance of distortion to speech. Performance assessment is given to show that our proposal can achieve a more significant noise reduction and a better spectral estimation of weak speech spectral components from a noisy signal as compared to many existing speech enhancement algorithms.

1. INTRODUCTION

There exist in the literature a great number of approaches for enhancement of noise-corrupted speech such as Ephraim-Malah (E-M) MMSE [1], log spectral amplitude (LSA) estimation [2], speech enhancement based on human auditory perceptual criteria [3] and configuration-based spectral estimation [4]. The objectives of speech enhancement are mainly to improve the perceptual quality, speech intelligibility and to reduce listener fatigue.

In this paper, single channel speech enhancement is studied. One of the main approaches of speech enhancement algorithms is to obtain the best possible estimates of the short time spectra of a speech signal from a given noisy speech. In [5], 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 gain function.

The advantage of the E-M noise suppression method is derived from 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 [5]; 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 [6], the characteristics and performance of a β-order MMSE method is studied. The method assumes that speech and noise spectral amplitudes are Gaussian distributions. The cost function $J = E\{(A_k^\beta - \hat{A}_k^\beta)^2\}$ is used as an estimation criterion, where $\hat{A}_k$ 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 proposed and its performance is analyzed.

In this paper, the human perceptual auditory masking effect is incorporated into the adaptive β-order MMSE method in order to achieve an optimal effectiveness for both high noise suppression and low perceptual speech distortion. There are many existing speech enhancement methods which exploit the properties of the human auditory system. The main aim of these methods is to find an optimal trade off between noise suppression, speech distortion and residual tonal noise level [3]. In addition, most of them use the *a posteriori* SNR to achieve noise suppression. In [3], a generalized spectral subtraction is proposed to adapt speech enhancement parameters based on the noise masking threshold. It is assumed that a human listener is unable to perceive additive noise as long as it remains below this threshold.

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 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 are looking for the estimate $\hat{A}_k$, which minimizes the following distortion measure

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

given the observed signal $\{x(t), 0 \le t \le T\}$, where β is the order of the spectral amplitude of the signal. Obviously, the estimator is given by

$$\hat{A}_k = \sqrt[\beta]{E\{A_k^\beta | x(t), 0 \le t \le T\}}. \tag{3}$$

With the Gaussian model assumption, the gain function is given by

$$G_\beta(\xi_k, \gamma_k) = \frac{\sqrt{\upsilon_k}}{\gamma_k}[\Gamma(\beta/2+1)M(-\beta/2; 1; -\upsilon_k)]^{1/\beta} \tag{4}$$

where $\Gamma(\cdot)$ is the gamma function and M($\alpha; \gamma; z$) is the confluent hypergeometric function, and υ_k is defined as follows

$$\upsilon_k = \frac{\xi_k}{1+\xi_k}\gamma_k \tag{5}$$

where ξ_k and γ_k represent the *a priori* SNR and *a posteriori* SNR respectively [1, 2], i.e.,

$$\xi_k = \frac{\eta_s(k)}{\eta_n(k)}, \quad \gamma_k = \frac{R_k^2}{\eta_n(k)}. \tag{6}$$

$\eta_n(k) = E\{|N_k|^2\}$, and $\eta_s(k) = E\{|S_k|^2\}$ are the variances of the kth spectral components of noise and speech, respectively.

The estimate of *a priori* SNR, ξ_k, can be given by the decision-directed approach proposed in [1] and is described as follows

$$\hat{\xi}_k(l) = (1-\alpha)max[\gamma_k(l) - 1, \quad 0] + \alpha\frac{|G_\beta(\hat{\xi}_k(l-1), \gamma_k(l-1))X_k(l-1)|^2}{\eta_n(k)}. \tag{7}$$

Normally, the parameter α is set to 0.98 [1, 5], and l denotes the frame number.

In order to utilize effectively the perceptual properties of the human auditory system, we consider the flooring effect by using the masking threshold. Therefore, we re-define the gain function, $G_{\beta_p}(\xi_k, \gamma_k)$, as follows

$$G_{\beta_p}(\xi_k, \gamma_k) = \begin{cases} \frac{\sqrt{\upsilon_k}}{\gamma_k}[\Gamma(\beta/2+1)M(-\beta/2; 1; -\upsilon_k)]^{1/\beta}, \\ \qquad \text{if} \quad \gamma_k > \rho_1(k) + \rho_2(k) \\ \{\rho_2(k)\frac{1}{\gamma_k}\}^{\frac{1}{2}}, \qquad \text{otherwise} \end{cases} \tag{8}$$

where $\rho_1(k)$ and $\rho_2(k)$ are determined by the masking threshold [3] and elaborated further in the following page. The estimate of speech signal therefore is given as follows

$$\hat{S}_k = G_{\beta_p}(\xi_k, \gamma_k)X_k. \tag{9}$$

3. PROPOSED ADAPTIVE β VALUE BASED ON MASKING PROPERTIES

It is noted that the gain always converges to the Wiener gain value if the value of instantaneous SNR ($\gamma_k - 1$) is large enough for a certain *a priori* SNR ξ_k value regardless of the β value. When β approaches 0 ($\beta = 0.001$), the gain curve is very close to the E-M LSA [2] gain curve. While $\beta = 1$, it is exactly the same as the E-M STSA-MMSE [1] gain curve.

The elimination of musical tones using the E-M STSA-MMSE [1] estimator is described in [5], and obviously, it can be applied to the β-order MMSE described by Eq. (8) for any value of β.

Fig. 1 (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, and it indicates that gain increases as the value of β increases. Fig. 1 (b) shows the gain curves as a function of β value, for different instantaneous SNR, (γ_k-1), for the case of *a priori* SNR, ξ_k, equals to 0 dB. It illustrates that 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. According to psychoacoustics theory, masking plays a very important role in human hearing. Some audio components cannot be heard because they are masked by other audio components. This implies that human listeners cannot notice the difference between the original speech and the speech distorted by a processing step if the distortions caused by the processed speech are masked by some parts of the original speech retained in the processed speech. The maximum allowable noise spectrum (or distortion spectrum) where the distortion is not discernible by a human listener is called the masking threshold. We can use auditory masking effects in noise suppression to overcome the classic tradeoff between noise reduction and speech distortion, wherein the audible noise is masked rather than being suppressed further. This reduces chances of further distortion of speech.

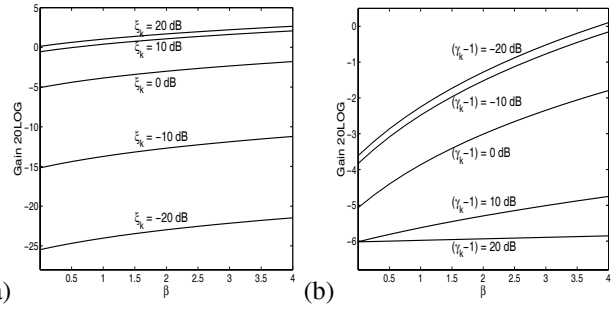

Fig. 1. (a). Gain versus β value for *a priori* SNR ξ_k = -20, -10, 0, 10, 20 dB and instantaneous SNR (γ_k-1) = 0 dB. (b). Gain versus β value for instantaneous SNR (γ_k-1) = -20, -10, 0, 10, 20 dB and *a priori* SNR ξ_k = 0 dB.

Based on the analysis of the characteristic of β-Order STSA MMSE estimator, we arrive at the appropriate β value

for a particular frame l, i.e., we can make β a function of frame SNR $\Xi(l)$, which is defined as follows

$$\Xi(l) = 10\log_{10}\max\Big\{\frac{\sum_{k=0}^{N/2}\{max[R_k(l)-\sqrt{\eta_n(l,k)},0]\}^2}{\sum_{k=0}^{N/2}\eta_n(l,k)},\varepsilon\Big\} \tag{10}$$

where ε is a very small positive number, in practice $\varepsilon = 2.22\times10^{-16}$. From the masking threshold , we can determine the power and spectral shape of noise that might be inaudibly inserted into the speech signal. Consequently, the frequency-dependent masking threshold can be regarded as the desired re-shaping of the noise spectrum. We propose that the value of β depends not only on the frame SNR, but also changes according to a particular masking threshold in the corresponding Bark band. This approach has shown to achieve more improvements in both objective evaluation and subjective assessment as compared to the case which does not use masking properties. We may express the β value as a function of the two variables in polynomial form

$$\beta(l,k) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} c_{ij}\Xi(l)^i A_{p_f}(l,k)^j = \tau_0+\tau_1\Xi(l)+\tau_2 A_{p_f}(l,k)+\tau_3\Xi(l)A_{p_f}(l,k)+\Phi(O_h) \tag{11}$$

where A_{p_f} represents the perceptual factor of the human auditory system in the frequency domain. c_{ij} and τ_i $(i,j = 0,1,2,...,\infty)$ denote the polynomial coefficients, and $\Phi(O_h)$ denotes the high-order polynomial terms.

By ignoring the high-order polynomial items, and assuring β is the monotonic non-decreasing function of $\Xi(l)$, we express $\beta(l,k)$ approximately in the following form

$$\beta(l,k) = \tau_0+\tau_1\Xi(l)+\tau_2 A_{p_f}(l,k) + \tau_3 max[(\Xi(l)+\tau 4),0]A_{p_f}(l,k) \tag{12}$$

where τ_i $(i = 0,1,2,3,4)$ denotes the polynomial coefficients. If the masking threshold is high, residual noise will naturally be masked and become inaudible. In this case, the β value should be high in order to reduce the chance of speech distortion. However, if the masking threshold is low, residual noise will be annoying to a human listener and it is thus necessary to process it. This is done by reducing β value appropriately.

Normally, we adjust our speech enhancement parameters based on the principle that the residual noise stays below the masking threshold of the auditory system. According to the above analysis, we adapt the high β value to high SNR condition, and also high β value to high masking threshold. We observe that too high β value will cause apparent distortion of speech signal. Here, we give the empirical polynomial coefficients for the β function as follows

$$\beta(l,k) = min\{max\{0.942+0.121\Xi(l)+0.981A_{p_f}(l,k) + 0.187max[(\Xi(l)+6.7),0]A_{p_f}(l,k), \quad 0.001\}, \quad 4\}. \tag{13}$$

As speech is quasi-stationary in nature, it is not surprising that there is a significant correlation in the perceptual properties of two adjacent frames. In addition, the effect of pre-masking and post-masking occurs predominantly in the same critical band. Thus, we propose the following adjustment to the relevant factor of perceptual properties in the Bark domain, A_p, as follows

$$A_p(l,\lambda) = \delta A_p(l-1,\lambda)+(1-\delta)\hat{T}_p(l,\lambda) \tag{14}$$

where δ is a forgetting factor equals to or less than 0.05, which is obtained empirically, and the value of the normalized masking threshold, $\hat{T}_p$, is determined by

$$\hat{T}_p(l,\lambda) = \frac{T(l,\lambda)-T_{min}(l)}{T_{max}(l)-T_{min}(l)} \tag{15}$$

where $T_{max}(l)$ and $T_{min}(l)$ are the maximal and minimal values of noise masking threshold $T(l,\lambda)$ at current frame l.

The perceptual factors, A_{p_f}, and the normalized masking threshold, $\hat{T}_p$, in the frequency domain and A_p in the Bark domain, have the following relationship

$$A_{p_f}(l,k) = A_p(l,\lambda) \tag{16}$$

$$\hat{T}_{p_f}(l,k) = \hat{T}_p(l,\lambda). \tag{17}$$

Just like most masking-based speech enhancement methods, noise masking threshold $T(l,\lambda)$ has to be obtained from a speech signal. In this paper we determine $T(l,\lambda)$ from an estimated speech signal which is obtained using the normal *a posteriori* spectral subtraction method.

The parameters $\rho_1(k)$ and $\rho_2(k)$ are obtained as follows

$$\rho_1(k) = \hat{T}_{p_f}(l,k)(\rho_{1max}-\rho_{1min})+\rho_{1min} \tag{18}$$

$$\rho_2(k) = \hat{T}_{p_f}(l,k)(\rho_{2max}-\rho_{2min})+\rho_{2min} \tag{19}$$

where ρ_{1min}, ρ_{2min} and ρ_{1max}, ρ_{2max} denote the minimal and maximal values of the suppression parameters respectively. For Eq. (8) case, we select ρ_{1min}=1, ρ_{1max}=6.28, ρ_{2min}=0 and ρ_{2max}=0.015.

4. PERFORMANCE EVALUATION

Various types of noise, taken from the NOISEX-92 database, are used in our performance evaluation. A total of 30 phonetically balanced speech utterances from the TIMIT database are used in the evaluation. The effectiveness of the proposed enhancement algorithm is evaluated at the sampling rate of 8kHz with a frame size of 256 samples (32 ms). The samples

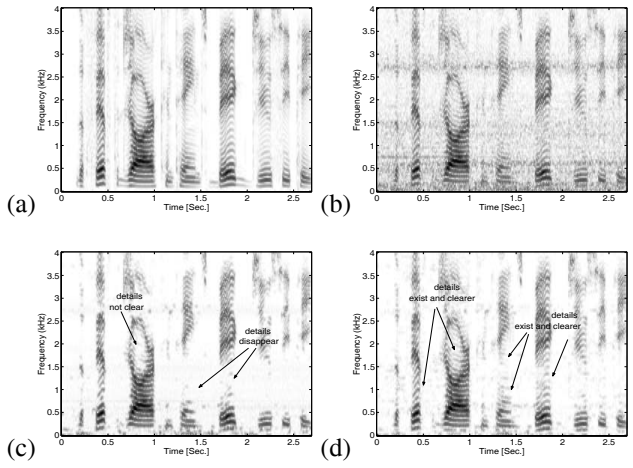

Fig. 2. Speech spectrograms (a) Clean Speech (b) Noisy speech; (f16 noise) (c) Over-subtraction masking method; (d) Our proposed masking-based adaptive β-order method.

Methods	SS	LSA	MMSE	OM	BM
MOS	2.01	3.25	2.98	2.81	3.52

Table 1. MOS values for the following methods. 1. SS: Spectral subtraction; 2. LSA: E-M LSA; 3. MMSE: E-M MMSE; 4. OM: Over-subtraction based on masking properties; 5. BM: Our proposed masking-based adaptive β-order MMSE.

are Hamming windowed with 75% overlap between adjacent frames.

Fig. 2 shows the spectrogram comparison, and Fig. 3 shows the improvement in segmental SNR performance and Itakura-Saito distortion measure for the different speech estimation algorithms. From these figures, we can see that the proposed method always outperforms the other methods. Table 1 gives the MOS listening test results which confirm that our proposed masking-based enhancement method leads to the best performance for human listeners as compared to other enhancement methods.

5. CONCLUSION

In this paper, a method that combines the adaptive β-order STSA-MMSE speech enhancement estimator and perceptual properties of the human auditory system is proposed. Based on the perceptual properties of the human auditory system, a method to determine the value of beta is introduced. The method is verified by computer simulations. From the simulations results, it is shown that the proposed algorithm outperforms many existing speech enhancement methods. In addition, it possesses a good tradeoff in minimizing both speech distortion and residual noise simultaneously. This characteristic is especially useful for the case of weak spectral components of speech signal that is corrupted by noise.

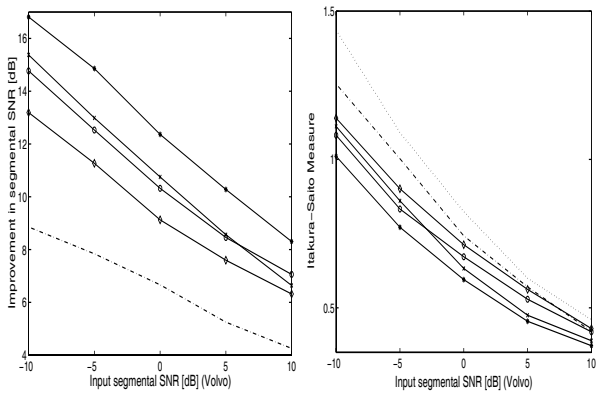

Fig. 3. (a) Segmental SNR improvement and (b) Itakura-Saito distortion measure for noisy input (.) with Volvo car noise. The tested methods are as following (-.) *a posteriori* spectral subtraction, (x) Over-subtraction masking, (o) LSA, ($\diamond$) E-M MMSE, (*) Our proposed masking-based adaptive β-order method.

6. 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] N. Virag, "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System," *IEEE Trans. Speech and Audio Processing*, Vol. 7, No.2, pp. 126-137, Mar. 1999.

[4] I.Y. Soon and S.N. Koh, "Low Distortion Speech Enhancement," *IEE Proceedings, Vision, Image and Signal Processing*, Vol. 147, No. 3, pp. 247-253, Jun. 2000.

[5] 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.

[6] C.H. You, S.N. Koh, and S. Rahardja, "Adaptive β-Order MMSE Estimation for Speech Enhancement", *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-03, Vol. 1, pp. 852-855, 2003.