ADAPTIVE MMSE SPEECH SPECTRAL AMPLITUDE ESTIMATOR UNDER SIGNAL PRESENCE UNCERTAINTY

Behdad Dashtbozorg, and Hamid Reza Abutalebi

Speech Processing Research Lab, Electrical and Computer Eng. Dept., Yazd University, Yazd, Iran
Phone: +(98) 351 8122396, Fax: +(98) 351 8210699, email: habutalebi@yazduni.ac.ir
web: http://pweb.yazduni.ac.ir/pages/faculties/engineering/elec/SPRL

ABSTRACT

In this paper, we present an adaptive speech spectral amplitude estimator that minimizes the Mean Square Error (MSE) under signal presence uncertainty. The spectral gain function, that is an adaptive variable-order Minimum MSE (MMSE) estimator, is obtained as a weighted geometric mean of hypothetical gains associated with speech presence and absence. The proposed estimator uses MMSE estimator considering various orders (powers) for the spectrum. The order of estimator and speech presence probability are estimated for each time frame and each frequency component individually. Our evaluations confirm superiority of the proposed method in noise suppression from corrupted speech.

1. INTRODUCTION

The speech collected from a noisy environment is predominantly degraded by additive background noise. Speech enhancement methods improve the performance of the speech-based systems by enhancing the speech components in noisy speech. The main objective of speech enhancement methods is to improve one or more perceptual aspects of speech, such as speech quality and intelligibility.

Here, we focus on the class of speech enhancement systems that capitalizes on major efficacy of Short-Time Spectral Amplitude (STSA) of speech signal on its perception. In these systems, the STSA of (clean) speech signal is estimated and combined with the short-time phase of the degraded speech, to construct the enhanced signal.

This research follows the works done by Ephraim and Malah (E-M MMSE) [1] for derivation of an MMSE STSA. Ephraim and Malah also proposed a modified version of the estimator for the Log-Spectral Amplitude (LSA) [2]; the modified technique minimizes MSE of the log-spectra based on Gaussian model for the spectral components of speech signal.

The LSA estimator proved very efficient in reducing musical noise. Its modification under signal presence uncertainty is obtained by multiplication of spectral gain and conditional speech presence probability. This probability is estimated for each frequency bin and each frame [3]. Cohen proposed an Optimally Modified LSA (OM-LSA) estimator [4] that considers optimal spectral gain function as a weighted geometric mean of hypothetical gains associated with the speech presence uncertainty. The exponential weight of each hypothetical gain was its corresponding probability, conditional on the observed signal.

On the other side, in [5], You *et al.* proposed β-order MMSE speech enhancement approach for estimating variable (β)-order STSA of the speech signal. You *et al.* examined the effectiveness of various ranges of β in MMSE estimating of STSA. In their work, β value was adapted using the frame Signal-to-Noise Ratio (SNR). Evaluation results demonstrate the superiority of this method in noise reduction and a better spectral estimation of weak speech spectral components compared to that of [1].

In this paper, we propose a novel method, called Adaptive MMSE (AMMSE) that is a hybrid version of OM-LSA and β-order MMSE methods. By using the ideas of these two methods, we present an adaptive speech spectral amplitude estimator that minimizes the MMSE of speech signal spectral amplitude under signal presence uncertainty. In the proposed estimator, we simultaneously search for the optimal values of 1) probability of speech presence, and 2) the order of MMSE estimation for each frame and each frequency component. This leads us to an enhancement system with significant noise reduction in both high and low input SNRs. The system has also less residual noise compared to state-of-the-art methods.

This paper has been organized as follows. β-order MMSE estimator is explained in section 2. In section 3, we discuss the issue of speech presence uncertainty and present the proposed AMMSE estimator. The procedure for determining proper value of β is expressed in section 4. In section 5, we explain simulations and performance evaluations. Finally, section 6 consists of some concluding remarks.

2. BETA-ORDER MMSE ESTIMATOR

Let $x(t)$ and $d(t)$ denote the speech and uncorrelated additive noise signal, respectively. The observed signal $y(t)$ is given by

$$y(t) = x(t) + d(t) . \quad (1)$$

After dividing into overlapping frames, we apply Short Time Fourier Transform (STFT) on each frame. In this regard, $A(k,l) = |X(k,l)|$, $|D(k,l)|$ and $R(k,l) = |Y(k,l)|$ denote spectral amplitude of speech, noise and observation signals, respectively; k is the frequency bin index and l is the time frame index.

We are looking for the estimate of $A(k,l)$, named $\hat{A}(k,l)$, that minimizes the mean-square error between β-order (clean) speech spectral amplitude and the β-order estimated speech spectral amplitude, i.e.

$$E\{(A(k,l)^{\beta} - \hat{A}(k,l)^{\beta})^2\}, \quad (2)$$

where $E\{.\}$ denotes the expectation operator. The β-order MMSE estimator is given by [5]:

$$\hat{A}(k,l) = \sqrt[\beta]{E\{A(k,l)^{\beta} \mid Y(k,l)\}}. \quad (3)$$

Assuming complex Gaussian pdf for each individual spectral component of speech and noise, by minimizing the cost function in (2) with respect to $\hat{A}(k,l)$, the estimate of the spectral amplitude of speech signal is then obtained by [5]:

$$\hat{A}(k,l) = \eta(k,l)^{1/2}[\Gamma(\frac{\beta}{2}+1)M(-\frac{\beta}{2};1;-\nu(k,l))]^{1/\beta}, \quad (4)$$

where $\Gamma(.)$ is the gamma function, $M(\alpha;\gamma;z)$ is the confluent hyper-geometric function, and $\eta(k,l)$ and $\nu(k,l)$ are defined as follows:

$$\eta(k,l) = [\frac{1}{\lambda_x(k,l)} + \frac{1}{\lambda_d(k,l)}]^{-1}, \quad (5)$$

$$\nu(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)}\gamma(k,l). \quad (6)$$

$\lambda_d(k,l)$ and $\lambda_x(k,l)$ are the variances of noise and speech, and $\xi(k,l)$ and $\gamma(k,l)$ represent the *a priori* SNR and *posteriori* SNR, respectively [1]:

$$\xi(k,l) = \frac{\lambda_x(k,l)}{\lambda_d(k,l)}, \quad \gamma(k,l) = \frac{R^2(k,l)}{\lambda_d(k,l)}. \quad (7)$$

The variance of noise, $\lambda_d(k,l)$, is estimated using a Voice Activity Detector (VAD). Finally, the estimate of speech spectral amplitude component, $\hat{A}(k,l)$, is given as follows:

$$\hat{A}(k,l) = G_{\beta}(k,l)R(k,l), \quad (8)$$

where $G_{\beta}(k,l)$ is the gain function of β-order MMSE estimator given by

$$G_{\beta}(k,l) = \frac{\sqrt{\nu(k,l)}}{\gamma(k,l)}\left[\Gamma(\frac{\beta}{2}+1)M\left(-\frac{\beta}{2};1;-\nu(k,l)\right)\right]^{1/\beta}. \quad (9)$$

You *et al.* adapted the value of β (used in (9)) according to the frame SNR. Based on the analysis of the characteristics of β-order MMSE, it is desirable for β to increase as frame SNR $(\Xi(l) = \sum_{k=0}^{N/2}\lambda_x(k,l) \Big/ \sum_{k=0}^{N/2}\lambda_d(k,l))$ increases, and to decrease when $\Xi(l)$ decreases. You *et al*. proposed following semi-linear relationship between β and $\Xi(l)$:

$$\beta(l) = \max\{\min[\mu_1\Xi(l)+\mu_2, \mu_3], \mu_4\}, \quad (10)$$

where μ_1, μ_2, μ_3 and μ_4 denote linear coefficients. This approach has been found to be experimentally effective in achieving good simulation results for estimator [5].

3. SIGNAL PRESENCE UNCERTAINTY

In this section, we introduce a new method for estimating the speech signal spectral amplitude under signal presence uncertainty. The proposed technique is similar to one described by Cohen in [4], but instead of using LSA estimator, it uses β-order MMSE estimator.

Given two hypotheses, $H_0(k,l)$ and $H_1(k,l)$, respectively indicating speech absence and presence in the k-th frequency bin of l-th frame, we have

$$\begin{aligned} H_0(k,l) &: Y(k,l) = D(k,l), \\ H_1(k,l) &: Y(k,l) = X(k,l) + D(k,l). \end{aligned} \quad (11)$$

We assume that the STFT coefficients, for both speech and noise, are complex Gaussian variables. Accordingly, the conditional pdfs of observed signal are given by

$$\begin{aligned} p(Y(k,l) \mid H_0)) &= \frac{1}{\pi\lambda_d(k,l)}\exp\{-\frac{|Y(k,l)|^2}{\lambda_d(k,l)}\}, \\ p(Y(k,l) \mid H_1)) &= \frac{1}{\pi(\lambda_x(k,l)+\lambda_d(k,l))} \\ &\quad \times\exp\{-\frac{|Y(k,l)|^2}{\lambda_x(k,l)+\lambda_d(k,l)}\}. \end{aligned} \quad (12)$$

Appling Bayes rule for the conditional speech presence probability, we have

$$\begin{aligned} P(H_1(k,l) \mid Y(k,l)) &= \{1+\frac{q(k,l)}{1-q(k,l)}(1+\xi(k,l)) \\ &\quad \times\exp(-\nu(k,l))\}^{-1} \triangleq p(k,l), \end{aligned} \quad (13)$$

where $q(k,l) \triangleq P(H_0(k,l))$ is the *a priori* probability for speech absence.

Based on the binary hypothesis model and equation (2),

$$\begin{aligned} &E\{A(k,l)^{\beta} \mid Y(k,l)\} \\ &= E\{A(k,l)^{\beta} \mid Y(k,l), H_1(k,l)\}p(k,l) \\ &+ E\{A(k,l)^{\beta} \mid Y(k,l), H_0(k,l)\}(1-p(k,l)). \end{aligned} \quad (14)$$

Using (3), we have

$$\begin{aligned} \hat{A}(k,l) = \{&E\{A(k,l)^{\beta} \mid Y(k,l), H_1(k,l)\}p(k,l) \\ &+ E\{A(k,l)^{\beta} \mid Y(k,l), H_0(k,l)\}(1-p(k,l))\}^{1/\beta}. \end{aligned} \quad (15)$$

During speech absence, the gain is constrained to be larger than a threshold $G_{\min}$, that is determined by subjective criteria for the noise naturalness. Let

$$E\{A(k,l)^{\beta} \mid Y(k,l), H_0(k,l)\} = (G_{H_0}|Y(k,l)|)^{\beta}. \quad (16)$$

When speech is present, the conditional estimation of spectral component is defined by

$$E\{A(k,l)^{\beta} \mid Y(k,l), H_1(k,l)\} = (G_{\beta}(k,l)|Y(k,l)|)^{\beta}, \quad (17)$$

where $G_{\beta}(k,l)$ is the gain function of β-order MMSE estimator, that was obtained in (9). Substituting (16) and (17) into (15), the spectral gain is determined via

$$G(k,l)=\{G_{\beta}(k,l)^{\beta}\,p(k,l)+{G_{H_0}}^{\beta}(1-p(k,l))\}^{1/\beta}$$
$$=\left[\left(\frac{\sqrt{\nu(k,l)}}{\gamma(k,l)}\left[\Gamma(\frac{\beta}{2}+1)M\left(-\frac{\beta}{2};1;-\nu(k,l)\right)\right]^{1/\beta}\right)^{\beta}\right. \quad (18)$$
$$\left.\times\,p(k,l)+{G_{H_0}}^{\beta}\left(1-p(k,l)\right)\right]^{1/\beta}.$$

Equation (18) presents the gain function for our proposed estimator (namely, AMMSE). Compared to the basic MMSE estimators (such as EM-MMSE [1] and LSA [2]), AMMSE has two additional parameters; the first parameter, β, is the order of MMSE estimator computed in a manner explained in the next section. The second parameter, $p(k,l)$, is the estimation of conditional speech presence probability that is obtained by local and global spectral averaging in frequency domain [4]. These parameters make the speech signal amplitude more accurate; resulting excellent noise suppression, while retaining weak speech components and avoiding the musical residual noise.

4. PROPER VALUE FOR THE ORDER OF AMMSE ESTIMATOR

In the proposed formula by You [5], the value of β is adapted semi-linearly according to the frame SNR (see Eq. (10)). It results in an equivalent value of β for all the spectral components of a frame. Here, we propose a method for estimating the value of β for each frame and each spectral component, individually, which makes the estimation more accurate.

As mentioned before, You *et al*. [5] proposed and validated a direct relation between β and the frame SNR. Also, it is obvious that there is a direct relation between SNR and speech presence probability ($p(k,l)$). Consequently, there is a direct relation between β and $p(k,l)$. Simplifying the issue, we consider a linear relation between β and $p(k,l)$, and propose the adaptation of β according to the value of $p(k,l)$. By applying a linear relation between these two parameters, musical noise will be decreased and speech intelligibility will be increased considerably.

Considering equation (2), and assuming $\beta>0$, we re-write the cost function of estimator as:

$$C\left(A(k,l),\hat{A}(k,l),\beta\right)=\left(A(k,l)^{\beta}-\hat{A}(k,l)^{\beta}\right)^{2}. \quad (19)$$

Now, let $-1<\beta<0$, so $\beta=-|\beta|$ and the cost function can be re-written as:

$$C\left(A(k,l),\hat{A}(k,l),\beta\right)=\left(\frac{1}{A(k,l)^{|\beta|}}-\frac{1}{\hat{A}(k,l)^{|\beta|}}\right)^{2}$$
$$=\left(\frac{A(k,l)^{|\beta|}-\hat{A}(k,l)^{|\beta|}}{A(k,l)^{|\beta|}\,\hat{A}(k,l)^{|\beta|}}\right)^{2} \quad (20)$$
$$=\frac{C\left(A(k,l),\hat{A}(k,l),|\beta|\right)}{\left(A(k,l)\hat{A}(k,l)\right)^{2|\beta|}}.$$

The denominator in (20) is an approximation of power spectrum to the exponent of $2|\beta|$. Therefore, taking a negative value for β has the effect of normalizing the cost function (19) (for positive $|\beta|$) by the estimated power spectrum to the exponent of $2|\beta|$. This normalization increases the contribution of spectral valleys in the cost function (estimation error) compared to that of spectral peaks. Actually, this employs masking properties of human hearing system that more noise is likely to be audible in speech spectral valleys than in speech spectral peaks. Consequently, the proposed estimator performs more accurate in the spectral valleys.

Considering above explanations, we simplify the relationship between the value of β and $p(k,l)$ as following linear function:

$$\beta(k,l)=\alpha\times p(k,l), \quad (21)$$

where $-1\le\alpha<0$ is the linear coefficient.

There is two important points here: 1) unlike the method by You *et al.* [5] that estimates β value for each frame, our proposed method determines the value of β for each frame and each frequency component and its value is obtained by a linear relationship with the probability of speech presence; and 2) we consider negative values for β, that make our estimation more accurate in spectral valleys.

5. PERFORMANCE EVALUATION

For simulation, we have used eight (clean) speech signal samples (at the sampling rate of 16 kHz from TIMIT database [7]) and made these signals noisy with white Gaussian noise. A wide range of input SNRs (-10dB, -5dB, 0dB, 5dB, 10dB, 15dB, 20dB) has been considered in the experiments. Also, the value of α (in equation (21)) has been empirically set to (-0.8).

To evaluate the performance of the proposed method, we have used three objective measures: SegSNR, LLR distance, and PESQ [7]-[9]. We have compared the output of AMMSE algorithm with those for OM-LSA and β-order MMSE methods. The results have been drawn in figures 1, 2, and 3 for SegSNR, LLR distance, and PESQ, respectively. As shown, the proposed method has superior performance in terms of all three quality measurements in various input SNRs.

We have also repeated the evaluations for the speech corrupted by low-pass noises (such as pink and F16 noises from

Noisex database [10]). Similar comparative results demonstrate the superiority of the proposed estimator.

6. CONCLUSION

In this research, we proposed an adaptive minimum mean-square error (AMMSE) spectral amplitude estimator under speech signal presence uncertainty. The estimator is used for the enhancement of noisy speech.

In this method, we use an MMSE estimator, whose order is adapted according to the probability of speech presence in each frame and each frequency component. The spectral gain function is obtained by modifying the gain function of the β-order estimator, based on binary hypothesis model. The modification includes a lower bound for the gain that is determined by subjective criteria for the noise naturalness, and exponential weights, which are given by the conditional speech presence probability. We also demonstrated that using negative value for the order of estimator makes the estimation more accurate in speech spectral valleys and consequently, results in greater speech quality improvement.

The proposed method has been evaluated and compared to the conventional estimators, in various noise types and levels in terms of SegSNR, LLR, and PESQ measures. Results show that the proposed estimator achieves better performance under all tested environment conditions. In this method, excellent noise suppression is obtained, while retaining weak speech components and avoiding the musical residual noise phenomena.

ACKNOWLEDGMENT

The authors would like to thank Iran Tele-communication research Center (ITRC) that funded this research.

REFERENCES

[1] 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

[2] 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.

[3] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech presence uncertainty to improve speech enhancement in non-stationary noise environments", in *Proc. of ICASSP*, pp. 789–792, March 1999.

[4] I. Cohen, "On speech enhancement under signal presence uncertainty," in *Proc. of ICASSP*, May 2001.

[5] C. H. You, S. N. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech, Audio Process*., Jul. 2005.

[6] J. S. Garofolo, *et al.*, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, Philadelphia, 1993.

[7] J. H. L. Hansen, and B. and Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. of ICSLP*, Dec. 1998.

[8] F. J. Fraga, C. A. Ynoguti, and A. G. Chiovato, "Further investigation on the relationship between objective measures of speech quality and speech recognition rate in noisy environments," in *Proc. of ICSLP*, 2006.

[9] ITU-T P.862, "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *ITU-T Recommendation*, P.862, 2000.

[10] http://www.noisex.com

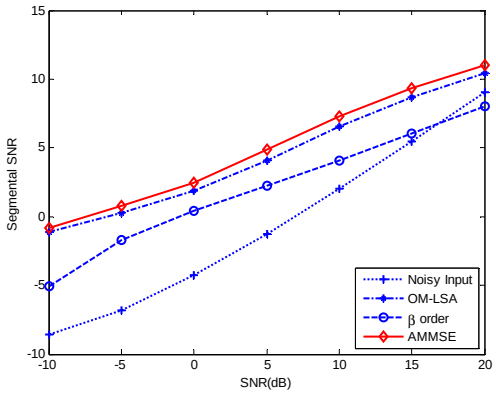

Figure 1 – The SegSNR comparison of proposed method with OM-LSA and β-order MMSE methods for white Gaussian noisy speech signal.

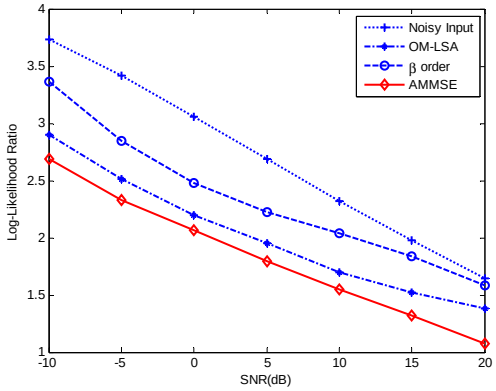

Figure 2 – The LLR comparison of proposed method with OM-LSA and β-order MMSE methods for white Gaussian noisy speech signal.

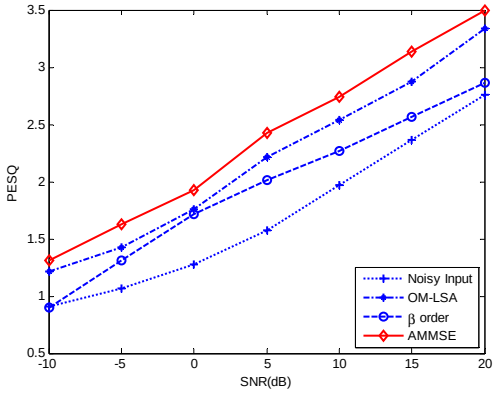

Figure 3 – The PESQ comparison of proposed method with OM-LSA and β-order MMSE methods for white Gaussian noisy speech signal.