INTERSPEECH 2009 BRIGHTON

Joint Noise Reduction and Dereverberation of Speech Using Hybrid TF-GSC and Adaptive MMSE Estimator

Behdad Dashtbozorg[1], *Hamid Reza Abutalebi*[1]

[1] Speech Processing Research Lab, Electrical and Computer Engineering Department, Yazd University, Yazd, Iran

dashtbozorg@stu.yazduni.ac.ir, habutalebi@yazduni.ac.ir

10.21437/Interspeech.2009-420

Abstract

This paper proposes a new multichannel hybrid method for dereverberation of speech signals in noisy environments. This method extends the use of a hybrid noise reduction method for dereverberation which is based on the combination of Generalized Sidelobe Canceller (GSC) and a single-channel noise reduction stage. In this research, we employ Transfer Function GSC (TF-GSC) that is more suitable for dereverberation. The single-channel stage is an Adaptive Minimum Mean-Square Error (AMMSE) spectral amplitude estimator. We also modify the AMMSE estimator for dereverberation application. Experimental results demonstrate superiority of the proposed method in dereverberation of speech signal in noisy environments.

Index Terms: Dereverberation, Spectral estimator, TF-GSC, AMMSE

1. Introduction

The main objective of speech enhancement is to reduce the corrupting noise and reverberation from received speech signal while preserving the original speech quality as much as possible. Some of de-noising methods can also be used in dereverberation [1]. These systems mainly estimate the amplitude of short-time spectrum of clean signal; then, the phase of received signal will be added to the estimated amplitude in order to obtain the enhanced signal. By modifying these algorithms, we can also reduce the (late) reverberation in noisy environments.

In [1], Habets *et al.* presented a method for dereverberation which was based on Optimally Modified – Log Spectral Amplitude (OM-LSA) estimator. On the other side, in [2], Abutalebi *et al.* proposed a hybrid de-noising method, called GSC-AMMSE, which is the combination of Generalized Sidelobe Canceller (GSC) with a single-channel noise reduction stage. The employed single-channel noise reduction stage is an Adaptive MMSE (AMMSE) which is a hybrid version of OM-LSA and β-order MMSE methods [3], [4]. AMMSE estimator is an adaptive speech spectral amplitude estimator that minimizes the MMSE of speech signal spectral amplitude under signal presence uncertainty. AMMSE estimator simultaneously searches for the optimal values of 1) probability of speech presence, and 2) the order of MMSE estimation for each frame.

In this paper, we have proposed a hybrid method, called TF-GSC-AMMSE which is an adaptive beamformer followed by a post processor. The adaptive beamformer is the Transfer-Function GSC (TF-GSC), which was proposed by Gannot *et al.* in [5]. Considering joint noise reduction and dereverberation application, we modify AMMSE method and employ it as the post processor. Also, we propose a new method for estimating the order of AMMSE estimator. This leads us to an enhancement system with significant noise and reverberation reduction in both high and low input SNRs. The system has also very low residual noise compared to the state-of-the-art methods.

Objective and subjective evaluation of TF-GSC-AMMSE method was performed under various conditions. We used Segmental Signal to Interference Ratio (SegSIR), Log-Likelihood Ratio (LLR) distance and Perceptual Evaluation of Speech Quality (PESQ) [6], [7] in the evaluations.

In section 2, we review the main concepts regarding TF-GSCs and discuss the combining of TF-GSC and spectral estimator. We propose a new spectral estimator for dereverberation in section 3. In section 4, the performance of the proposed method is evaluated and compared with the conventional TF-GSC. Finally, section 5 consists of some concluding remarks.

2. Hybrid TF-GSC and spectral estimator

In [2], Abutalebi *et al.* proposed two hybrid methods that use OM-LSA or AMMSE estimators as a post-processor for GSC. In these methods, GSC beamformer is firstly applied on the microphone array signals. Then, a single channel spectral amplitude estimator improvers the output of GSC.

GSC is a beamforming structure that is used for implementing a variety of linearly constrained adaptive array processors, including Frost's algorithm [8]. GSC does the adaptive beamforming via two processing paths; in the first path, a signal-independent fixed beamformer enhances desired signal components. The second path consists of the blocking matrix and the adaptive portion, which provides a set of filters that adaptively minimize the noise power in the output [9].

The standard GSC structure assumes the received signals as simple delayed versions of source signal. In real room situations where this assumption is not valid, the desired signal leaks into the adaptive path of the GSC structure. This results in the distortion or cancellation of desired signal. As a remedy, Gannot *et al.* [5] proposed an improved version of GSC, called TF-GSC, which considers arbitrary transfer functions between source and microphones. A sub-optional solution for these arbitrary transfer functions was also proposed using transfer function ratios that were estimated online. The blocking matrix was constructed using same

transfer function ratios, thereby significantly reducing the leakage of the desired signal.

Although TF-GSC can be used in moderate reverberant environments, it should be noted it does not reduce the amount of reverberation. As a remedy, the use of some post-filters on the output of GSC (or TF-GSC) has been already proposed.

In this research, we propose a hybrid method that uses modified AMMSE estimator as the post-processor for TF-GSC. Ws also modify the AMMSE estimator and show that it has noticeable capability in dereverberation.

3. Single-channel dereverberation stage

AMMSE estimator [2] is a hybrid version of OM-LSA and β-order MMSE methods for noise reduction. In this section, we modify AMMSE estimator and present an adaptive speech spectral amplitude estimator for dereverberation in noisy environments. Note the proposed estimator will be applied on the output of the TF-GSC, and it is a single-channel method.

3.1. Problem formulation

Reverberation is the process of multi-path propagation of an acoustic sound from its source to microphone. Let $x(n)$ denote the clean speech signal, the observed signal $y(n)$ is given by

$$y(n) = \underbrace{x(n) + r(n)}_{z(n)} + d(n) \tag{1}$$

where $r(n)$ is a non-stationary interference (or signal reflections) and $d(n)$ is a stationary interference (or background noise). Also, $z(n)$ is the sum of direct signal and its reflections. The reverberant signal can be expressed as

$$z(n) = \sum_{j=0}^{L_h-1} h_j(n)x(n-j) \tag{2}$$

where $h_j(n)$ denotes the j^{th} coefficient of the impulse response of the acoustic plant between source and microphone, and L_h denotes the length of the impulse response.

The aim of dereverberation is to form $\hat{x}(n)$, an estimate of $x(n)$, from $y(n)$. This is a blind problem since neither the signal $x(n)$ nor the acoustic impulse responses $h_j(n)$ are available. The acoustic impulse response is split into two segments, $h_e(n)$ and $h_l(n)$, so that

$$h_j(n) = \begin{cases} h_{e,j}(n), & 0 \le j < N_l \\ h_{l,j}(n), & N_l \le j < L_h - 1 \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The parameter N_l can be chosen depending on the application or subjective preference. Equation (1) is re-written using equation (3):

$$y(n) = \underbrace{\sum_{j=0}^{N_l-1} h_j(n)x(n-j)}_{z_e(n)} + \underbrace{\sum_{j=N_l}^{L_h-1} h_j(n)x(n-j)}_{z_l(n)} + d(n) \tag{4}$$

where signal $z_e(n)$ consists direct signal and early reflections and $z_l(n)$ consists late reflections (reverberations). The purpose of dereverberation is to reduce the effect of late reverberation. . Due to the non-stationarity of the source and the statistical properties of the acoustic impulse response, we can assume that the early and late reflections are statistically independent. Therefore, we can suppress the late reverberant signal by treating it as an additive noise term [1].

The observed signal $y(n)$ is transformed into the time-frequency domain by applying the short-time Fourier transform (STFT). Firstly, the signal $Y(k,l)$ is used to estimate the noise spectral variance $\lambda_d(k,l) = E\left\{\left|D(k,l)\right|^2\right\}$, where $D(k,l)$ is defined as the STFT of the noise signal $d(n)$. Secondly, considering $Z_l(k,l)$ as the STFT of the signal component $z_l(n)$, the late reverberant spectral variance, $\lambda_{z_l}(k,l) = E\left\{\left|Z_l(k,l)\right|^2\right\}$ is estimated.

3.2. Modifying MMSE estimator for dereverberation

In this section, we consider an adaptive method that estimates the early reflections spectral components $Z_e(k,l)$. This estimation is like AMMSE method in [2].

The gain function of β-order MMSE estimator $G_\beta(k,l)$ for reverberant speech signal can be written as

$$G_\beta(k,l) = \frac{\sqrt{\nu(k,l)}}{\gamma(k,l)}\left[\Gamma(\frac{\beta}{2}+1)M\left(-\frac{\beta}{2};1;-\nu(k,l)\right)\right]^{1/\beta} \tag{5}$$

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

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

where $\xi(k,l)$ represent the *a priori* SIR defined as:

$$\begin{aligned} &\frac{1}{\xi(k,l)} = \frac{1}{\xi_{z_l}(k,l)} + \frac{1}{\xi_d(k,l)} \\ &\xi_d(k,l) = \frac{\lambda_{z_e}(k,l)}{\lambda_d(k,l)}, \quad \xi_{z_l}(k,l) = \frac{\lambda_{z_e}(k,l)}{\lambda_{z_l}(k,l)} \end{aligned} \tag{7}$$

$\gamma(k,l)$ denotes the *a posteriori* SIR

$$\gamma(k,l) = \frac{\left|Y(k,l)\right|^2}{\lambda_d(k,l) + \lambda_{z_l}(k,l)} \tag{8}$$

where $\lambda_d(k,l)$, $\lambda_{z_l}(k,l)$ and $\lambda_{z_e}(k,l)$ are noise spectral variance, late reverberation spectral variance and early reverberation spectral variance (direct signal and early reflections), respectively. The early speech spectrum $Z_e(k,l)$ is constructed by applying a time and frequency dependent gain function $G_\beta(k,l)$ to $Y(k,l)$,

$$\hat{Z}_e(k,l) = G_\beta(k,l)Y(k,l) \tag{9}$$

3.3. Adaptive estimator

In this section, we introduce a new dereverberation method that estimates the early speech spectral amplitude under signal presence uncertainty. This is an extension of one proposed for noise reduction in [2]. The estimator is similar to one described by Cohen in [3], 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

$$\begin{aligned} H_0(k,l) &: Y(k,l) = Z_l(k,l) + D(k,l) \\ H_1(k,l) &: Y(k,l) = Z_e(k,l) + Z_l(k,l) + D(k,l) \end{aligned} \tag{10}$$

We assume that the STFT coefficients, for both speech and noise, are complex Gaussian variables.
Based on the binary hypothesis model,

$$\begin{aligned} &E\{A(k,l)^\beta \mid Y(k,l)\} \\ &= E\{A(k,l)^\beta \mid Y(k,l), H_1(k,l)\}\, p(k,l) \\ &+ E\{A(k,l)^\beta \mid Y(k,l), H_0(k,l)\}(1 - p(k,l)) \end{aligned} \tag{11}$$

where $A(k,l) = |Z_e(k,l)|$ is the STFT of the speech signal and $p(k,l)$ is speech presence probability. we have

$$\begin{aligned} \hat{A}(k,l) = \{&E\{A(k,l)^\beta \mid Y(k,l), H_1(k,l)\}\, p(k,l) \\ &+ E\{A(k,l)^\beta \mid Y(k,l), H_0(k,l)\}(1 - p(k,l))\}^{1/\beta} \end{aligned} \tag{12}$$

By considering speech absence, we have

$$E\{A(k,l)^\beta \mid Y(k,l), H_0(k,l)\} = (G_{H_0} |Y(k,l)|)^\beta \tag{13}$$

During speech absence, the gain is constrained to be larger than a threshold $G_{\min}$, which is determined by subjective criteria for the noise naturalness. The lower-bound constraint does not result in the desired result because reverberation can still be clearly audible. Hence to suppress the non-stationary interference, we obtain G_{H_0} as following

$$G_{H_0}(k,l) = G_{\min} \frac{\lambda_d(k,l)}{\lambda_d(k,l) + \lambda_{z_l}(k,l)} \tag{14}$$

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

$$E\{A(k,l)^\beta \mid Y(k,l), H_1(k,l)\} = (G_{H_1}(k,l) |Y(k,l)|)^\beta \tag{15}$$

where $G_{H_1}(k,l)$ is the gain function of β-order MMSE estimator, that was obtained in (5). Substituting (13) and (10) into (12), the spectral gain is determined via

$$\begin{aligned} G(k,l) &= \{G_{H_1}(k,l)^\beta\, p(k,l) + {G_{H_0}}^\beta (1 - p(k,l))\}^{1/\beta} \\ &= \left[\left(\frac{\sqrt{v(k,l)}}{\gamma(k,l)} \left[\Gamma(\frac{\beta}{2} + 1) M\left(-\frac{\beta}{2}; 1; -v(k,l) \right) \right]^{1/\beta} \right)^\beta \right. \\ &\left. \times p(k,l) + \left(G_{\min} \frac{\lambda_d(k,l)}{\lambda_d(k,l) + \lambda_{z_l}(k,l)} \right)^\beta (1 - p(k,l)) \right]^{1/\beta} \end{aligned} \tag{16}$$

Equation (16) presents the gain function for our proposed estimator (namely, AMMSE). It 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 [3]. These parameters make the speech signal amplitude more accurate; resulting excellent noise and reverberation suppression, while retaining weak speech components and avoiding the musical residual noise.

3.4. Proper value for the order of AMMSE (β)

In the proposed formula by You *et al.* [4], the value of β is adapted semi-linearly according to the frame SNR. 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.

In this research, we propose the adaptation of β according to the probability of speech presence, $p(k,l)$. Simulation results show that there is a direct relation between the value of β and the value of $p(k,l)$.

Assuming $\beta > 0$, we write the cost function of estimator as:

$$C\left(A(k,l), \hat{A}(k,l), \beta\right) = \left(A(k,l)^\beta - \hat{A}(k,l)^\beta\right)^2. \tag{17}$$

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

$$C\left(A(k,l), \hat{A}(k,l), \beta\right) = \frac{C\left(A(k,l), \hat{A}(k,l), |\beta|\right)}{\left(A(k,l)\hat{A}(k,l)\right)^{2|\beta|}}. \tag{18}$$

The denominator in (18) 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 (17) (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. Correspondingly, the proposed estimator performs more accurate in the spectral valleys.

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

$$\beta(k,l) = \alpha \times p(k,l) \qquad , -1 \le \alpha < 0 \tag{19}$$

where α is the linear coefficient.

There is two important points here: 1) unlike the method by You *et al.* [6] 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.

4. Performance evaluation

In order to examine the performance of TF-GSC-AMMSE method, we have considered a setup with the main speech source at 90° and the noise source at 15°. The sampling rate in the entire system is 16 kHz. In this evaluation, we have used white Gaussian noise with various SNRs (-10dB, -5dB, 0dB, 5dB, 10dB, 15dB, 20dB) in two different values for reverberation time (RT_{60} = 200ms, 500ms). Also, the value of α (in equation (19)) has been empirically set to (-0.8).

To evaluate the performance of the proposed method, we have used three objective measures: SegSIR, LLR distance, and PESQ [6], [7].

In simulation, an eight-channel linear microphone array has been used. The spacing between microphones is 5 cm. We have compared the performance of hybrid method TF-GSC-AMMSE with the input signal and the output of the TF-GSC (alone). The results have been drawn in Figures 1, 2 and 3 for SegSIR, LLR distance, and PESQ, respectively. As shown, the proposed method has superior performance in terms of all three quality measurements in various input SNRs and in two different reverberant situations (RT_{60}= 200ms and RT_{60}= 500ms).

5. Conclusions

In this paper, we firstly modified a signal amplitude estimator (called AMMSE) for single channel speech dereverberation. The proposed estimator adaptively works based on a minimum mean-square error under speech presence uncertainty. This method has noticeable noise reduction and dereverberation in single microphone applications. Then, we used AMMSE algorithms as post-processors in the TF-GSC structure. This resulted in more dereverberation. It was shown that these combinations, give rise to improve dereverberation performance of the TF-GSC in noisy environments.

6. Acknowledgements

The authors would like to thank Iran Telecommunication Research Center (ITRC) that funded this research.

7. References

[1] E.A.P. Habets, I. Cohen, S. Gannot, and P.C.W. Sommen, "Joint Dereverberation and Residual Echo Suppression of Speech Signals in a Noisy Environment," IEEE Trans. of Audio, Speech, and Language Process., June 2006.

[2] H. R. Abutalebi, B. Dashtbozorg and T. Zare, "Speech enhancement using hybrid generalized sidelobe canceller and spectral estimator," *in Proc. of Int. Symposium on Telecommunications,* pp. 564-569, *2008.*

[3] I. Cohen, "On speech enhancement under signal presence uncertainty," IEEE Trans. Acoust., Speech, Signal Process. ICASSP-01, May 2001.

[4] C. H. You, S. N. Koh, S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process*., July 2005.

[5] S. Gannot, D. Burshtein and E. Weintein, "Signal enhancement using beamforming and nonstionary with applications to speech," IEEE Trans. Signal processing, vol. 49, no. 8, pp. 1614-1621, 2001.

[6] J. H. L. Hansen, B. Pellom, "An Effective Quality Evaluation Protocol for Speech Enhancement Algorithms," *in Proc. of ICSLP*, Dec. 1998.

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

[8] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," Proceedings of the IEEE, vol. 60, pp. 926-935, Aug. 1972.

[9] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," IEEE Trans. on Signal Processing, vol. 47, no. 10, pp. 2677–2684, 1999.

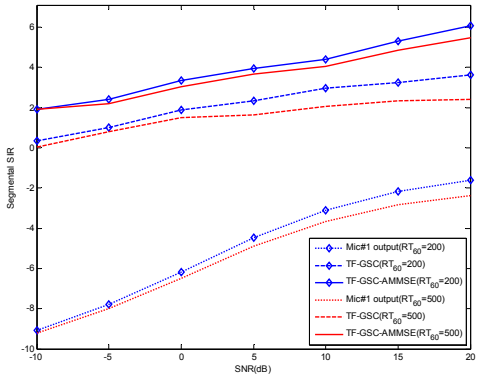

Figure 1: *The SegSIR evaluation results of TF-GSC-AMMSE method in RT_{60}=200ms and RT_{60}=500ms.*

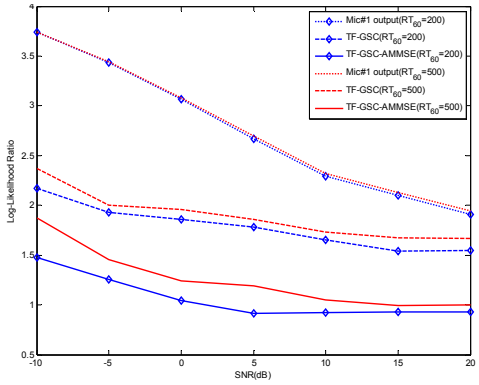

Figure *2: The LLR evaluation results of TF-GSC-AMMSE method in RT_{60}=200ms and RT_{60}=500ms.*

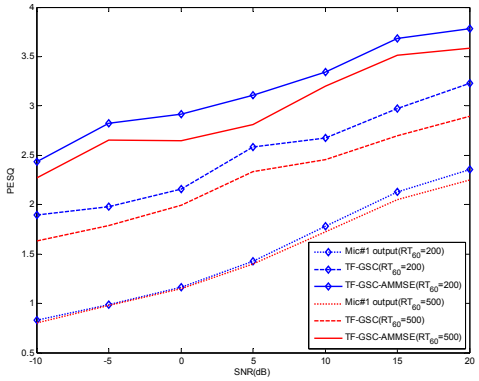

Figure 3: *The PESQ evaluation results of TF-GSC-AMMSE method in RT_{60}=200ms and RT_{60}=500ms.*