OPTIMAL MULTI-MICROPHONE SPEECH ENHANCEMENT IN CARS

Lae-Hoon Kim and Mark Hasegawa-Johnson

University of Illinois at Urbana-Champaign, United States

ABSTRACT

Hands-free speech telephony and speech recognition in cars suffer from additive noise and reverberation. We propose an iterative blind channel estimation algorithm based on an analysis-by-synthesis loop closed around a multipath Generalized Sidelobe Canceller (GSC). By combining a post-filter with the proposed scheme, optimal speech enhancement in practical situations can be achieved. The algorithm is tested using simulated data and using real speech recordings from the AVICAR database.

***Index Terms*—** dereverberation, speech enhancement, robust beamforming, blind channel identification

1. INTRODUCTION

In recent years, many systems have used multi-microphone arrays for the task of speech enhancement [1, 2, 3] and robust speech recognition [4, 5]. However, few approaches have presented a theoretical basis for the multi-microphone speech signal processing under the assumed statistical models of source speech signal, room impulse responses (RIRs) and noise. One of the few published systems that considers a theoretical basis for speech enhancement is that of Balan and Rosca [1], which showed that multi-microphone MMSE spectral amplitude estimation can be factored into a sufficient statistic followed by a single-microphone postfilter. If we can assume that we know the RIRs, optimal estimation of the speech signal can be done using a simple two-step method: first the sufficient statistic is computed, then the classical MMSE estimator. The two-step method actually guarantees optimality in the sense of the one channel estimator. However, it is actually not easy to satisfy the assumption of known RIRs. Inspired by the sufficient statistic factorization approach [1], we address a realistic implementation of the sufficient statistic.

In an acoustic echo cancellation scenario, if we know the source signal, we can adaptively estimate the channel response [6]. Because more correctly beamformed output is much nearer to the source signal, we might be able to use the beamformed output as an input to estimate the channel response from the output signal. Good channel estimation makes the beamformer based on multipath GSC more accurate, and this again guarantees better channel estimation, where the multipath GSC is different with the conventional GSC in the sense that it reflects the multipath effect to the constraint part realized as the fixed beamformer (FBF). Until we can get a satisfactory channel estimation result, in other words, a satisfactory beamforming result (satisfactory deconvolution), we keep iterating this adaptive procedure with some reasonable channel constraint. The iterative procedure can be used for the multi-channel identification as well as the optimal beamforming. Even though we may not get perfect channel identification, still this is a useful scheme in multipath GSC, because we might use the converged multichannel information as a coefficient vector for the FBF, rather than using a naive delay and sum beamformer as in the conventional GSC and by leveraging the converged channel we actually mitigate the inherent signal cancellation problem due to the reverberation.

To visualize the situation more or less in a simple and tractable way, we first show the convergence of a simplified version of the proposed scheme. The preliminary simulation test has been conducted to show how this concept is working. The result of the simple preliminary simulation shows that this method seems to achieve sufficient blind deconvolution at the output of FBF after enough iterations. We expand the proposed algorithm into the realistic environment in a car.

2. PROPOSED METHOD

2.1. Multipath GSC

Multi-path GSC can be formulated as an optimization problem as in (1), which is a generalized version of generalized sidelobe canceller (GSC) [7] under multi-path acoustic environment.

$$\arg\min_{\underline{w}} E\left\{\underline{w}^T \underline{yy}^T \underline{w}\right\} \text{ subject to } C^T \underline{w} = \underline{f}, \qquad (1)$$

where $\hat{s}(n) = \underline{w}^T \underline{y}$ is the time-domain estimated source signal and $\underline{y}$ is is the noisy signal vector, superscript T is transpose, the array filter coefficient $\underline{w} = [\underline{w}_1^T\ \underline{w}_2^T \cdots \underline{w}_N^T]^T$ and $\underline{w}_i^T = [w_{(i-1)L+1}\ w_{(i-1)L+2} \cdots w_{(i-1)L+L}]$ where $i = 1, 2, \cdots, N$, $\underline{y} = [y_{1,[1:L]}\ y_{2,[1:L]} \cdots y_{N,[1:L]}]^T$ where $y_{i,[1:L]} = [y_i(n-(i-1)n_0)\ y_i(n-(i-1)n_0-1) \cdots y_i(n-(i-1)n_0-(L-1))]$, $i = 1, 2, ...N$ and n is the current time index, steered to a look direction of $\theta = \arcsin(n_0 F_s d/c)$ for microphone spacing d and sampling rate F_s, and $C^T\underline{w} = \underline{f}$ is a linear constraint. To derive multi-path GSC, we simply need to manipulate the constraint part in (1). The constraint part has the following convolutional form:

$$\begin{bmatrix} C_{h_1} \; C_{h_2} \; \cdots \; C_{h_N} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_{NL} \end{bmatrix} = \underline{f} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad (2)$$

where $l_h + L - 1$ by L matrix C_{h_i} is constructed from the room response $h_i(n)$:

$$\begin{aligned} C_{h_i} &= \begin{bmatrix} h_i(0) & 0 & \cdots & 0 \\ h_i(1) & h_i(0) & \ddots & \vdots \\ \vdots & h_i(1) & \ddots & 0 \\ h_i(l_h-1) & \vdots & \ddots & h_i(0) \\ 0 & h_i(l_h-1) & \vdots & h_i(1) \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & h_i(l_h-1) \end{bmatrix}, \\ i &= 1, 2, \cdots, N \qquad (3) \end{aligned}$$

(3) is a typical linear convolution matrix which has Toeplitz structure. The constraint in (2) is therefore equivalent to the requirement that $\underline{W}$ inverts $\underline{H}$, i.e., (2) is actually a channel deconvolution in the look-direction [8]. Note that the length of the constraint vector $\underline{f}$ is determined as $l_h + L - 1$ in (3). We can simply find that we can get the optimal estimator in Bayesian sense via multi-path GSC followed by Bayesian estimation. The beauty of this approach is in the FBF part and the subsequent blocking matrix, which can be constructed as the null space of the multi-channel convolution matrix. The FBF can be simply regarded as a multi-channel deconvolution problem. Therefore, we might be able to apply any kind of multi-channel deconvolution scheme [8, 9, 10] for this fixed part. The blocking matrix can also be constructed by using an echo cancellation scheme [6], because ideally the output of the fixed beamformer is the deconvolved and beamformed speech signal. Although we might be able to apply any kind of multi-channel deconvolution scheme for this fixed part, in the subsequent sections we propose a blind multi-channel RIR identification algorithm, which in fact exploits the structure of multipath GSC.

2.2. Iterative blind estimation of RIR based on Multipath GSC

2.2.1. Problem formulation

The channel response estimation follows the optimization process below.

$$\begin{aligned} \hat{h}_i(t) &= \arg\min_{\hat{h}_i(t)} \|s(t) * (h_1(t) * w_1(t) + \\ &\cdots \quad + h_N(t) * w_N(t)) * \hat{h}_i(t) - s(t) * h_i(t)\|^2 \end{aligned} \qquad (4)$$

where $\underline{\hat{h}}_i$ is the estimated channel.

$$\underline{\hat{h}}_i = \arg\min_{\underline{\hat{h}}_i} \|\hat{C}\underline{\hat{h}}_i - \underline{h}_i\|^2 = (\hat{C}^T\hat{C})^{-1}\hat{C}^T\underline{h}_i \qquad (5)$$

where $\hat{C}$ is the convolution matrix obtained with the beamformed output of impulse responses as input for the beamformer. Ideally if $\hat{C} = I$, in other words if the FBF with RIR as input produces perfectly deconvolved output, then we can obtain the real channel response. As in (5), the estimated channel responses are obtained in the optimal sense of least squares hereafter with the constraint of forcing to have zero values as estimated RIR except the estimated time stamps of each dominant reflection in RIRs. This is similar with the concept of acoustic echo cancellation except the constraint part.

2.2.2. Algorithm

The proposed algorithm is introduced below step by step. In here, we just care about the deconvolution, because the noise suppression after the deconvolution is quite straightforward. Based on the assumption that we know the time stamps for the reflections, we can successfully estimate the magnitude of the reflections using the following algorithm, where the number of microphones is N. The way of estimating the time position for the reflections are discussed later in the subsequent section.

1. Initialize the magnitude of the time location of reflections with $epsilon$ and 0 otherwise.

2. Perform multipath GSC to get output $\hat{s}$ and update $\hat{h}_1(r)$ with solution of (4).

3. Set the updated magnitude $\hat{h}_1$ of the time location with 0 if those are not the position for the designated reflections.

4. Iterate 2 and 3, until there is no more significant change in the magnitude of the reflection.

If you follow the first iteration, you will get the first update of $\hat{h}_1(r) \approx h_1(r) - \frac{1}{N}(h_1(r) - \epsilon + h_2(r) + \cdots + h_N(r))$ and if this number is bigger than ϵ it will be updated until there is no change of $\hat{h}_1(r)$ and this is going to be:

$$\hat{h}_1(r) = h_1(r) - \frac{1}{N-1}(h_2(r) + \cdots + h_N(r)) \qquad (6)$$

In the early part of the RIR, reflections are infrequent, therefore typically $h_2(r) = \cdots = h_N(r) = 0$ or at least $h_2(r), \cdots, h_N(r) << h_1(r)$ and therefore $\hat{h}_1(r) \cong h_1(r)$ in (6). Even if there exists noise, because we take a mean of iteration measurements we can regard it as zero since we can easily assume that the noise process is zero mean. However, in reality, because of low-pass filtering for sampling and other low-pass filtering effects acting on the reflections, the response will not contain perfect impulses, and this imperfection will produce some errors. Therefore, reflections with similar direction of arrival (DOA) will not be estimated correctly using this scheme. This intuitively makes sense; the benefits of using beamforming are reduced when the direction of interference is in the DOA of the source.
Figure 1 shows the converged result of a 2 channel measurement with a seven reflection RIR, including one negative component and one merged component in the RIR.

$$x1 = [1\ 0\ 0\ 0\ 0.5\ 0\ 0\ 0.4\ 0\ 0.05\ 0.3\ 0\ 0\ -0.1\ 0.09\ 0\ 0\ 0.04]^T$$
$$x2 = [1\ 0\ 0\ 0\ 0\ 0\ 0.5\ 0\ 0.45\ 0\ 0\ 0.3\ -0.1\ 0\ 0\ 0.09\ 0.04\ 0]^T$$

The first three reflection time stamps are assumed to be known and others are set as zero. We can confirm that by having correct time stamp for some early reflection, not all, we can estimate the channel responses up to the given reflection points and at the same time the deconvolution can be performed up to the reflection points. This results are very promising because we can track and deconvolve dominant early reflections, which is usually sparse enough and deterministically treatable within reasonably small amount of time frame where we can assume that the early responses are time invariant.

2.2.3. *Algorithm with reflection time stamp estimation*

In this section, we propose a heuristic dominant reflection time stamp estimation together with the proposed algorithm. Algorithm is as follows.

1. Initially we choose DSB as a first FBF and perform normalized least mean square algorithm to estimate the RIR FIR coefficients using the output of DSB.

2. Select the time stamps in which the estimated RIR magnitudes are above a predefined threshold, which determines the significance level of the reflection.

3. Perform the proposed algorithm.

4. Iterate 2 and 3 enough

Figure 2 shows the converged result, where the simulated output of two channel have been obtained by convolving the channel response with a white Gaussian noise source and the threshold has been set as “0.08”. Note that most of the significant reflection points above the threshold can be estimated almost correctly.

3. EXPERIMENT WITH REAL CAR DATA

In this section, we test the proposed algorithm using the real multi-channel sources measured in cars. The whole procedure for testing can be summarized as following.

1. Interchannel delay is estimated using GCC-PHAT method [11] and adjust the delay to formulate DSB.

2. Perform the proposed algorithm.

Figure 3 shows the 2-channel identification results using one of single digit utterance in AVICAR database [12], and no distinctive reflection other than direct path has been estimated. Possible explanation about this result can be the fact that the space inside of car is too small for having distinctive reflections which could be sparsely separable using the proposed algorithm. However, because this fact also means that there dose not exist significantly correlated reflections in the original signals with the beamformed output using the direct path information (DSB), we can avoid the signal canceling problem when we use GSC structure with DSB as FBF. Optimal signal enhancement result and isolated digit recognition result with conventional GSC have been reported in [5].

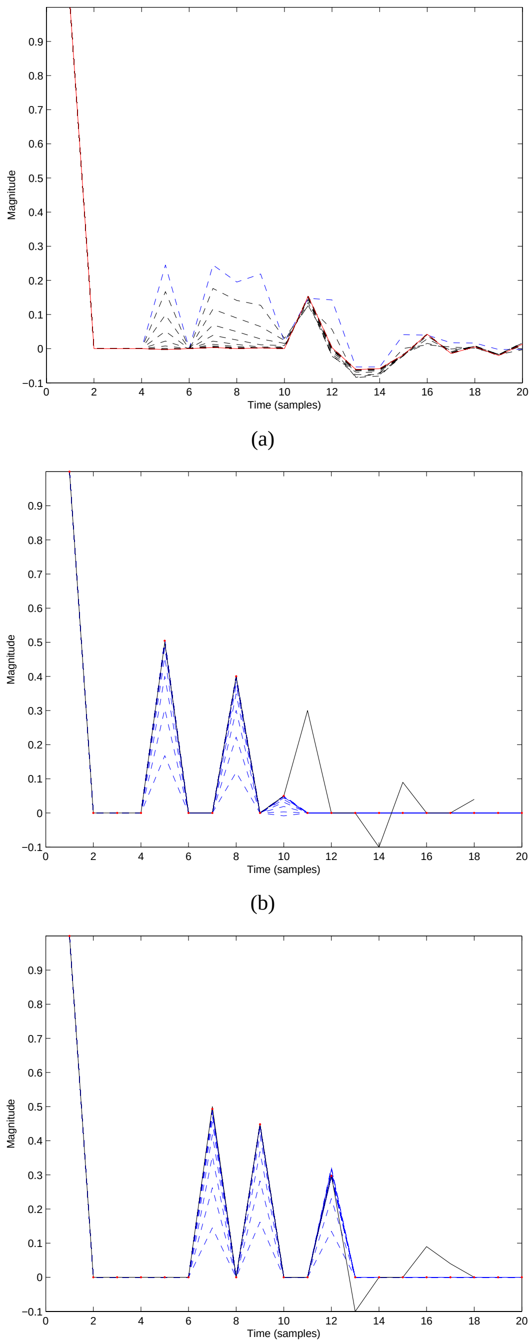
4. CONCLUSION

In this paper, we propose the multipath GSC based blind channel identification method, which can be plugged in as a realistic replacement of the sufficient statistic for optimal speech enhancement. The simulation with artificially generated sparse channels show that the proposed algorithm can converge into the original channel responses above a predefined significance threshold. Channel estimation experiment with real data measured in a car results in the fact that there exists no distinctive significant reflection which can contribute to the signal cancellation when we use GSC structure.

5. REFERENCES

[1] R. Balan and J. Rosca, “Microphone array speech enhancement by bayesian estimation of spectral amplitude and phase,” in *Proc. Sensor Array and Multichannel Signal Process. Workshop*, 2002.

[2] S. Gannot and I. Cohen, “Speech enhancement based on the general transfer function gsc and postfiltering,” *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 561–571, November 2004.

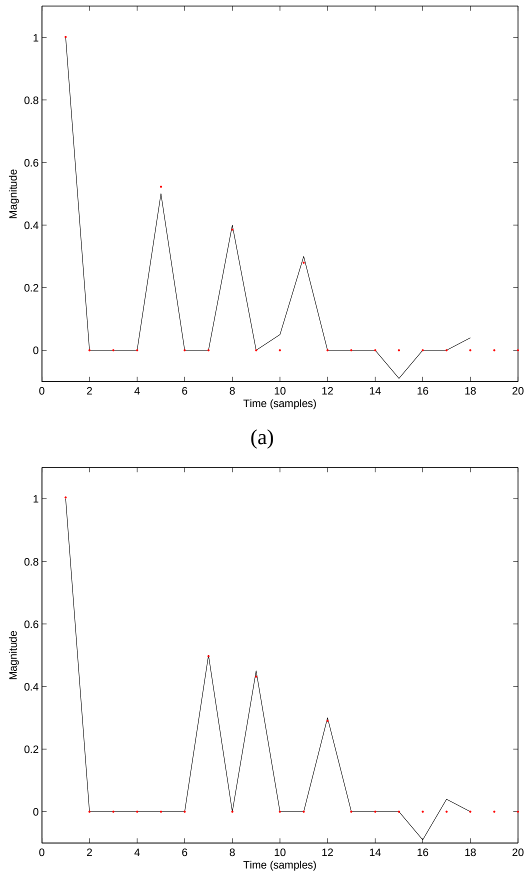


(a)

(b)

(c)

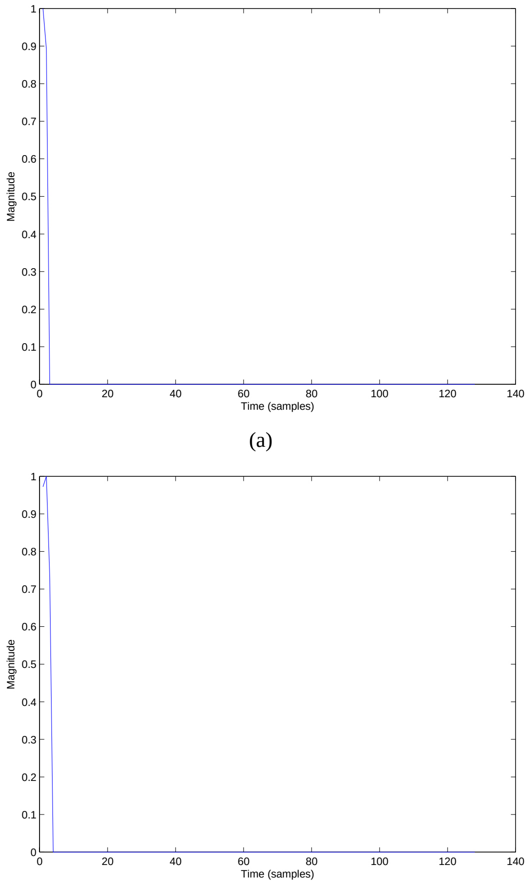
Fig. 1. (a)FBF output: Blue dotted line is DSB output, black dotted lines are updated FBF output, red line is final FBF output after 20 iteration. Updated FBF output produces more impulse-like output by eliminating the effect of the designated reflections, in other words, more deconvolved output. (b) Estimated channel h_1 (c) Estimated channel h_2: Red dots show the converged channel response after 20 iteration and the blue dotted lines are updated responses. The black line is for original RIR. The designated channel responses are almost perfectly identified.



(a)

(b)

Fig. 2. (a) Estimated channel h_1 (b) Estimated channel h_2: Red dots show the converged channel response after 20 iteration and the black line is for original RIR. The designated channel responses above the predefined threshold are almost correctly identified.



(a)

(b)

Fig. 3. (a) Estimated channel h_1 (b) Estimated channel h_2

[3] L.-H. Kim and M. Hasegawa-Johnson, “Optimal speech estimator considering room response as well as additive noise: Different approaches in low and high frequency range,” in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 2008.

[4] M. L. Seltzer, *Microphone Array Processing for Robust Speech Recognition*, PhD dissertation, Carnegie Mellon University, 2003.

[5] L.-H. Kim, M. Hasegawa-Johnson, and K.-M Sung, “Generalized optimal multi-microphone speech enhancement using sequential minimum variance distortionless response(mvdr) beamforming and postfiltering,” in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 2006.

[6] S. L. Gay and J. Benesty, *Acoustic Signal Processing for Telecommunication*, Kluwer Academic Publishers, Norwell, MA, 2000.

[7] L. J. Griffiths and C. W. Jim, “An alternative approach to linearly constrained adaptive beamforming,” *IEEE Trans. Antennas Propagat.*, vol. AP-30, pp. 27–34, January 1982.

[8] M. Miyoshi and U. Kaneda, “Inverse filtering of room acoustics,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, pp. 145–152, February 1988.

[9] Y. Huang, J. Benesty, and J. Chen, “A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 882–895, 2005.

[10] M. Delcroix, T. Hikichi, and M. Miyoshi, “Precise dereverberation using multichannel linear prediction,” *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 430–440, December 2007.

[11] G. H. Knapp and G. C. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-24, pp. 320–327, 1976.

[12] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, “Avicar: An audiovisual speech corpus in a car environment,” in *Proc. Int. Conf. Spoken Language Processing*, 2004.