Multichannel Noise Reduction using low order RTF estimate

Subhojit Chakladar, Nam Soo Kim, Yu Gwang Jin, Tae Gyoon Kang

School of Electrical Engineering and INMC
Seoul National University, Seoul 151-742, Korea
{subhojit, nkim}@snu.ac.kr, {ygjin, tgkang}@hi.snu.ac.kr

Abstract

The relative transfer function generalized sidelobe canceler (RTF-GSC) is a popular method for implementing multichannel speech enhancement. However, an accurate estimation of channel transfer function ratios pose a challenge, especially in noisy environment. In this work, we demonstrate that even a very low order RTF estimate can give superior performance in terms of noise reduction without incurring excessive speech distortion. We show that noise reduction is dependent on the correlation between the input noise and the noise reference generated by the Blocking Matrix (BM), and that a low order RTF estimate preserves this correlation better than a high order one. The performance for both high order and low order RTF estimates are compared using output SNR, Noise Reduction and a perceptual measure for speech quality.

Index Terms: Noise reduction, GSC, Correlation, Relative transfer function.

1. Introduction

The generalized sidelobe canceler (GSC)[1] is one of the most popular multi-channel noise reduction algorithms. It implements linearly constrained minimum variance (LCMV) beamforming[2] by decoupling the constraint and the minimization. The GSC consists of three building blocks. The first component is a fixed beamformer (FBF) which satisfies the desired constraint. The second component is a blocking matrix (BM) which blocks the desired signal to produce noise reference signals. The final stage consists of an unconstrained LMS type algorithm which minimizes the output noise power. The relative transfer function GSC (RTF-GSC) algorithm exploits nonstationarity of the desired speech signal to estimate the channel transfer function ratios and is effective in removal of stationary noise[3]. Though this method gives good noise reduction with minimum signal distortion, its performance is heavily dependent on the accuracy of the RTF estimation which can be quite challenging, especially in noisy environments. Previous works [4], [5] have stressed on the importance of FBF and BM stages of the GSC.

The actual RTF, which are very long length sequences, are approximated by FIR filters in the GSC implementation. Ideally, when the RTF is approximated as a FIR filter with a large number of taps, excellent noise reduction and little speech distortion is obtained, both of which are very desirable. However, in this paper, we study the performance of the final component of the GSC algorithm - the Adaptive Noise Canceler (ANC). This stage implements a NLMS algorithm using the FBF output as the input signal and the BM output as the reference signal. The aim of this stage is to minimize the noise power. In adaptive filtering, greater the correlation between the noise component of the input signal and the noise reference signal, greater is the noise reduction.

In this work, we first show that the output noise power is dependent on the correlation between the noise components in the FBF output and the BM output. It will be shown that greater the correlation between these two components, the lower is the output noise power. Next, we explore how this can be achieved. Following the RTF-GSC method, we show that correlation is better preserved when a low order RTF estimate is used in the FBF and the BM. This then has the dual advantage of making the estimation of RTF easier and giving superior noise reduction performance at the same time. We also evaluate a perceptual measure for speech quality and show that the distortion incurred for low order RTF estimate is almost equal to or in some cases less than that for a high order RTF estimate.

The paper is organized as follows. In section 2, we introduce the signal model and formulate the problem. This is followed by mathematical analysis of our proposed algorithm in section 3, where we show the dependence of output noise power on correlation between FBF and BM noise output. We also prove that in the case of diffuse white noise, this happens when using a low order RTF estimate. In section 4, we present the simulation results. The final section 5 is devoted to discussions and conclusions.

2. Problem Formulation

The sensor array consists of M microphones. The received signal consists of two components - the desired speech signal and the undesired noise signal. The aim here is to reduce the noise component as far as possible. The 1st channel is assumed to be the reference channel. We use the following notation:

- $Z_m(n)$ is the *m*th sensor signal;
- *S(n)* is the desired speech source;
- $N_m(n)$ is the noise component at the *m*th sensor;
- $a_m(n)$ is the channel transfer function (TF) between the desired speech source and the *m*th sensor.
- $U_m(n)$ is the noise reference from the *m*th channel;
- *D(n)* or $Y_{FBF}(n)$ is the fixed beamformer (FBF) output;
- $h_m(n)$ is the relative transfer function (RTF) from the *m*th channel and the reference channel;
- $\widetilde{h}_m(n)$ is the RTF from the reference channel to the *m*th channel.

Then the *m*th sensor signal can be denoted by

$$Z_m(n) = a_m(n) * S(n) + N_m(n) \quad (1)$$

where * denotes convolution. In frequency domain this becomes

$$Z_m(\omega) = a_m(\omega)S(\omega) + N_m(\omega) \quad (2)$$

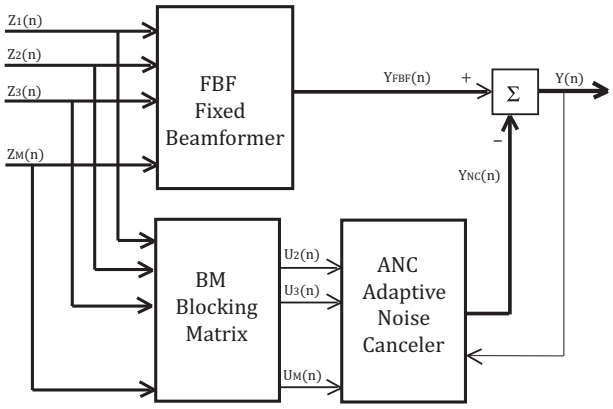

Figure 1: *Various constituent blocks of GSC.*

The RTFs are given by

$$a_m(n) = h_m(n) * a_1(n) \tag{3}$$

$$a_1(n) = \widetilde{h}_m(n) * a_m(n) \tag{4}$$

The FBF output is obtained by aligning the input signals from the sensors with the reference channel and summing them up

$$D(n) = \sum_{i=0}^{M} \widetilde{h}_i(n) * Z_i(n) \tag{5}$$

In frequency domain this becomes

$$D(\omega) = \mathbf{H}^{\dagger}(\omega)\mathbf{Z} \tag{6}$$

where **H** is the FBF vector, †represents conjugate transpose, **Z** is the input signal vector. If $\widetilde{\mathbf{S}}$ is the reverberated signal component and **N** is the noise component, then **Z** can be written as

$$\mathbf{Z}(\omega) = \widetilde{\mathbf{S}}(\omega) + \mathbf{N}(\omega) \tag{7}$$

The vectors have the following structure

$$\mathbf{Z}^T(\omega) = [Z_1(\omega) \;\; Z_2(\omega) \;\; \ldots \;\; Z_M(\omega)]$$

$$\mathbf{N}^T(\omega) = [N_1(\omega) \;\; N_2(\omega) \;\; \ldots \;\; N_M(\omega)]$$

$$\mathbf{H}^T(\omega) = [\widetilde{H}_1(\omega) \;\; \widetilde{H}_2(\omega) \;\; \ldots \;\; \widetilde{H}_M(\omega)]$$

where $\widetilde{H}_m(\omega)$ is the Fourier Transform of $\widetilde{h}_m$(n) from (4) The BM output consists of (M-1) channels and each channel is obtained by subtracting aligned version of the reference channel from the signal in that channel.

$$U_m(n) = Z_m(n) - h_m(n) * Z_1(n) \tag{8}$$

where m can vary from 2 to M. In the frequency domain, if **B** represents the BM, then the above equation looks like

$$\mathbf{U}(\omega) = \mathbf{B}^{\dagger}(\omega)\mathbf{Z}(\omega) \tag{9}$$

The final output is obtained from the multichannel ANC, which implements a NLMS algorithm. The FBF output is the input to it and the BM output acts as the reference signal.

3. Multichannel Noise Reduction

In this section, we will derive conditions for minimization of output noise power. The first part derives the output noise power in terms of the signals generated in the intermediate stages of the algorithm - namely the FBF output and the BM output. Since this deals with noise reduction, we focus on the silent intervals where only noise signal is present.

3.1. Output Noise Power

The ANC employs NLMS algorithm. It is expected that the filter coefficients would asymptotically converge to the Wiener solution[6]. Let D = $\mathbf{H}^{\dagger}\mathbf{N}$ denote the noise component of the FBF output and **U** = $\mathbf{B}^{\dagger}\mathbf{N}$ denote the noise component in the BM output. When D is the input signal and **U** is the noise reference signal, the Wiener filter coefficients are given by

$$\mathbf{G} = \Phi_{UU}^{-1}\Phi_{UD} \tag{10}$$

where

$$\Phi_{UD} = E(\mathbf{U}D^{\dagger}) = \mathbf{B}^{\dagger}E(\mathbf{N}\mathbf{N}^{\dagger})\mathbf{H} = \mathbf{B}^{\dagger}\phi_{NN}\mathbf{H} \tag{11}$$

$$\Phi_{UU} = E(\mathbf{U}\mathbf{U}^{\dagger}) = \mathbf{B}^{\dagger}E(\mathbf{N}\mathbf{N}^{\dagger})\mathbf{B} = \mathbf{B}\phi_{NN}\mathbf{B} \tag{12}$$

E denoting the expectation operator and ϕ_{NN} is the input noise power. The noise at the output of the ANC can then be written as

$$N_0 = D - \mathbf{G}^{\dagger}\mathbf{U} \tag{13}$$

Expanding the constituent terms we get

$$\begin{aligned} N_0 &= \mathbf{H}^{\dagger}\mathbf{N} - (\mathbf{B}^{\dagger}\phi_{NN}\mathbf{H})^{\dagger}(\mathbf{B}^{\dagger}\phi_{NN}\mathbf{B})^{-1}(\mathbf{B}^{\dagger}\mathbf{N}) \\ &= \mathbf{H}^{\dagger}\mathbf{N} - \mathbf{H}^{\dagger}\phi_{NN}\mathbf{B}(\mathbf{B}^{\dagger}\phi_{NN}\mathbf{B})^{-1}(\mathbf{B}^{\dagger}\mathbf{N}) \\ &= \mathbf{H}^{\dagger}(I - \phi_{NN}\mathbf{B}(\mathbf{B}^{\dagger}\phi_{NN}\mathbf{B})^{-1}\mathbf{B}^{\dagger})\mathbf{N} = \mathbf{H}^{\dagger}\mathbf{F}\mathbf{N} \end{aligned} \tag{14}$$

where **I** denotes an M×M identity matrix and **F** is the term inside the brackets. Output noise power is given by

$$\Phi_{NN} = E(N_0 N_0^{\dagger}) = \mathbf{H}^{\dagger}\mathbf{F}\phi_{NN}\mathbf{F}^{\dagger}\mathbf{H} \tag{15}$$

Now observe that $\mathbf{F}\phi_{NN}\mathbf{F}^{\dagger} = \mathbf{F}\phi_{NN}$. This means that the power of the residual output noise can be expressed as

$$\begin{aligned} \Phi_{NN} &= \mathbf{H}^{\dagger}\mathbf{F}\phi_{NN}\mathbf{H} \\ &= \mathbf{H}^{\dagger}\phi_{NN}\mathbf{H} - (\mathbf{B}^{\dagger}\phi_{NN}\mathbf{H})^{\dagger}(\mathbf{B}\phi_{NN}\mathbf{B})^{-1}(\mathbf{B}^{\dagger}\phi_{NN}\mathbf{H}) \end{aligned} \tag{16}$$

Substituting (11) and (12) in (16) we get

$$\Phi_{NN} = \Phi_{DD} - \Phi_{UD}^{\dagger}\Phi_{UU}^{-1}\Phi_{UD} \tag{17}$$

which can also be expressed as

$$\Phi_{NN} = \Phi_{DD}(1 - \rho_{UD}^2) \tag{18}$$

where Φ_{DD} is the noise power at the FBF and ρ_{UD} is related to the correlation coefficient between the noise at the FBF output and BM output. Since the absolute value of the correlation coefficient has an upper limit of 1, the output noise power will be lower as the correlation coefficient increases. Thus we have established the dependence of the output noise power on the correlation coefficient.

3.2. Monotonicity of Correlation Coefficient

Here we try to find the condition under which the correlation coefficient is high. Since the estimated RTF is the constituting element of both the FBF and the BM, the structure of the RTF can affect the correlation coefficient. We show that the correlation coefficient is a monotonically decreasing function of the order of the estimated RTF for diffuse noise. Let $U_N^K(n)$ and $D_N^K(n)$ denote the BM noise output and FBF noise output using a Kth order estimate of the RTF. We carry out the derivation for a 2 channel case for simplicity. Let the 1st channel be the reference channel and the 2nd one be any arbitrary channel m. Then

$$U_N^K(n) = N_m(n) - \Sigma_{i=0}^K h(i) N_1(n-i) \tag{19}$$

$$D_N^K(n) = N_1(n) + \Sigma_{i=0}^K \widetilde{h}(i) N_m(n+i) \tag{20}$$

where h and $\widetilde{h}$ are given by (3) and (4).

$$\begin{aligned} E(U_N^K D_N^K) = E([N_m(n) - \Sigma_{i=0}^K h(i) N_1(n-i)] \\ [N_1(n) - \Sigma_{i=0}^K \widetilde{h}(i) N_m(n+i)]) \end{aligned} \tag{21}$$

which when expanded gives

$$\begin{aligned} E(U_N^K D_N^K) = R_{NmN1}(0) \\ + \sum_{i=0}^K \widetilde{h}(i) R_{NmNm}(i) - \sum_{j=0}^K h(j) R_{N1N1}(j) \\ - \sum_{i=0}^K h(i) \sum_{j=0}^K \widetilde{h}(j) R_{NmN1}(i+j) \end{aligned} \tag{22}$$

If the noise is diffuse and white, then the cross correlation will be 0 and the autocorrelation will have only 1 non zero term at 0 lag. So this reduces to

$$E(U_N^K D_N^K) = \widetilde{h}(0) R_{NmNm}(0) - h(0) R_{N1N1}(0) \tag{23}$$

which is independent of the order of the estimated RTF. Now

$$\begin{aligned} E((U_N^K)^2) = R_{NmNm}(0) - 2 \sum_{j=0}^K h(j) R_{NmN1}(j) \\ + \sum_{i=0}^K h(i) \sum_{j=0}^K h(j) R_{N1N1}(i-j) \end{aligned} \tag{24}$$

The noise is white and diffuse, so the last term is non-zero only when i=j. Thus

$$E((U_N^K)^2) = R_{NmNm}(0) + \sum_{i=0}^K h^2(i) R_{N1N1}(0) \tag{25}$$

Similarly,

$$E((D_N^K)^2) = R_{N1N1}(0) + \sum_{i=0}^K \widetilde{h}^2(i) R_{N1N1}(0) \tag{26}$$

These 2 quantities are monotonically increasing with the increase in order of the estimated RTF which is K. Let

$$\rho_{UkDk} = E(U_N^K D_N^K) / \sqrt{E((U_N^K)^2) E((D_N^K)^2)} \tag{27}$$

be the correlation coefficient when using a Kth order estimate of the RTF. Since (23) appear in the numerator and (25),(26) appear in the denominator of (27), the correlation coefficient is a monotonically decreasing function of K. For p > q, then we have

$$\rho_{UpDp} < \rho_{UqDq} \tag{28}$$

Combining (18) and (28) we get

$$\Phi_{NN}^q < \Phi_{NN}^p \tag{29}$$

where Φ_{NN}^q stands for output noise power obtained by using an RTF estimate of order q. This means that a lower order RTF estimate gives lower output noise power.

4. Experimental Results

The simulation was conducted on a MATLAB©platform. The room was modeled as an enclosure of dimension 5m × 4m × 2.8m with the speech source located at1.58m × 2.08m × 1.5m. The sensor array consisted of 5 microphones placed near the middle of the room. The room impulse response (RIR) was generated using the image method[7] with reverberation time of 0.2s and convolved with the clean speech signals. The clean speech signals were obtained from both male and female speakers and were 5 to 10s long. The diffuse white noise was computer generated from the NOISEX-92 database [8]and added to the reverberated speech signals to form the corrupted signals.
Two cases were investigated. In the first case, the estimated RTF was modeled as a FIR filter with 300 coefficients. This represented the high order RTF estimate and corresponds to RTF_{high}. In the second case the estimated RTF was modeled as an FIR filter with just one non zero coefficient. This resembles the case when the signals in various channels are related by a simple delay and a scaling factor. This represented the low order RTF estimate and corresponds to RTF_{low}. The RTFs were assumed to be known but both of them can be estimated using the procedure mentioned in[9].The NLMS filters in the ANC were modeled as FIR filters with 256 coefficients for each channel.
The adaptation of the filter coefficients in ANC were done only during silent intervals to reduce speech distortion. The simulation was conducted for seven different input SNR- -5dB, -3dB, -1dB, 0dB, 1dB, 3dB and 5dB. The overall performance was measured in terms of average output SNR which is defined for the output signal y(n) as

$$SNR_{avg} \triangleq \frac{(\sum_{n \in Ts} y^2(n)/Ts - \sum_{n \in Tn} y^2(n)/Tn)}{(\sum_{n \in Tn} y^2(n)/Tn)}$$

where *Ts* is the speech interval (when the desired speech signal is active) and *Tn* denotes silent interval (when the desired speech signal is absent and only noise is present). We also measured the noise reduction which is defined for the silent interval as

$$NR \triangleq \frac{\sum_{n \in Tn} z_1^2(n)}{\sum_{n \in Tn} y^2(n)}$$

where *y(n)* is the output noise and $z_1(n)$ is the input noise in the reference channel (since we are dealing with silent interval only, the input signal in reference channel is equivalent to the input noise). Finally, the speech distortion was evaluated using a perceptual measure for speech quality given in [10]. This measure is a mean opinion score (MOS), of rating the output signal on a five-point scale. This scale is described in Table 1.

In Fig. 2 we see that RTF_{low} estimate clearly gives superior performance in terms of output SNR, noise reduction and similar (slightly better till about 3dB input SNR) performance in the case of speech quality. We also see that RTF_{low} is particularly

Rating	Description
5	Very natural, no degradation
4	Fairly natural, little degradation
3	Somewhat natural, somewhat degraded
2	Fairly unnatural, fairly degraded
1	Very unnatural, very degraded

Table 1: *Scale of signal distortion.*

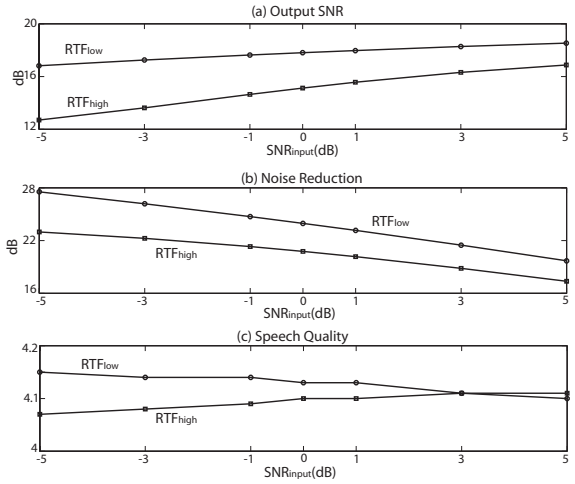

Figure 2: *Output performance for two different types of RTF.(a) Output SNR measured in dB, (b) Noise Reduction measured in dB, (c) Speech quality measured on a perceptual scale*

effective in very low input SNR cases (<0dB) where the gap in performance with RTF$_{high}$ is higher.

5. Conclusion and Discussion

We analyzed dependence of the output noise power on the signals generated in the intermediate stages of the algorithm. We found that the overall noise reduction performance was dependent on the correlation between the noise component in the FBF output and the noise component in the BM output. We showed that the output noise power is a function of the correlation coefficient. Greater the correlation, lower is the output noise power. As the correlation coefficient approaches 1, the output noise power approaches 0. The RTF is the central component of the FBF and BM blocks. So the structure of the RTF would affect the correlation. We then demonstrated that when RTF is modeled as a FIR filter with fewer number of coefficients, the correlation is better preserved and hence the noise reduction is also more. However, the greater noise reduction was not obtained at the cost of speech distortion, as the perceptual speech quality scores show.

Normally, when using a low order RTF estimate, there is greater desired signal leakage in the output of the BM, which is the primary cause of speech distortion in the output. In this case, however, since the adaption of the NLMS filter coefficients were done only during the silent intervals, this problem could be avoided. It should be noted that when a voice activity detector (VAD) is employed and if the speech interval is the null hypothesis, then the significance level should be set very low. In other words, wrongly classifying a silent interval will have less effect on speech distortion than the opposite case. One of the assumptions made here is that the noise is stationary. This allows us to stop NLMS filter coefficient adaptation during the speech intervals and use the previously computed coefficients to carry out the filtering in that duration. Apart from that, the RTF$_{low}$ in this case is a linear phase filter and itself introduces no additional distortion.

The RTF$_{low}$ thus has two advantages over RTF$_{high}$. First is the improved performance in terms of noise reduction without any added speech distortion. Second is the fact that having lesser number of coefficients, RTF$_{low}$ is computationally easier to estimate than RTF$_{high}$. RTF$_{low}$ can also be modeled as a FIR filter of arbitrary low order or even a sparse filter and both of them shows better performance than RTF$_{high}$. The particular structure of RTF$_{low}$ chosen in this work was done to demonstrate the extreme case. In future, this analysis will be extended to directional noise cases.

6. Acknowledgements

This work was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2009-0083044) and by National Research Foundation of Korea(NRF) grant funded by the Korean Government(MEST) (NRF-2009-0085162).

7. References

[1] Griffiths, L.; Jim, C., "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas and Propagation*, vol.30, no.1, pp. 27- 34, Jan 1982.

[2] Frost, O.L., III, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE* , vol.60, no.8, pp. 926-935, Aug. 1972.

[3] Gannot, S.; Burshtein, D.; Weinstein, E., "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. on Signal Processing*, vol.49, no.8, pp.1614-1626, Aug 2001.

[4] Jianfeng Chen; Shue, L.; Senjin Liu, "A fixed blocking matrix for robust microphone array beamforming," *Seventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings.* , vol.2, no., pp. 407- 410 vol.2, 1-4 July 2003.

[5] Herbordt, W.; Kellermann, W., "Analysis of blocking matrices for generalized sidelobe cancellers for non-stationary broadband signals," *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02).*, vol.4, no., pp. IV-4187 vol.4, 2002.

[6] Widrow, B.*et al.*, "Adaptive noise cancelling: Principles and applications," *Proceedings of the IEEE* , vol.63, no.12, pp. 1692-1716, Dec. 1975.

[7] Allen, J.B. and Berkley, D. A., "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Amer.*, Vol. 65, no. 4, pp. 943–950, 1979

[8] Varga, A. and Steeneken, H. J. M.,"Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, Vol. 12, No. 3, pp. 247-251, Jul. 1993.

[9] Shalvi, O. and Weinstein, E., "System identification using nonstationary signals", *IEEE Trans. on Signal Processing*,Vol. 44, pp. 2055–2063, 1996.

[10] Hu, Y. and Loizou, P., "Evaluation of objective measures for speech enhancement," *Proceedings of INTERSPEECH-2006,* September 2006.