A Fast-Converging Adaptive Frequency-Domain MVDR Beamformer for Speech Enhancement

Shengkui Zhao¹, Douglas L Jones²

¹Advanced Digital Sciences Center, Illinois at Singapore, Singapore
²Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Illinois, USA

shengkui.zhao@adsc.com.sg, jones@ifp.illinois.edu

Abstract

In this paper, we present a fast-converging adaptive frequency-domain minimum-variance-distortionless-response (MVDR) beamformer (FMV) for speech enhancement. The well-known FMV solution is optimum in the microphone array processing. However, the direct computation of the optimum FMV solution is often undesirable due to the inversion of the spatio-spectral correlation matrix which is often unstable and is expensive for large arrays. To avoid the matrix inversion, we develop a fast-converging conjugate gradient (CG) algorithm for iteratively computing the FMV solution. Compared to the existing steepest descent (SD) algorithm, the CG algorithm can dramatically improve the convergence speed for the case of multiple interfering signals in speech enhancement. Therefore, the computational load and processing time can be significantly reduced. The speech enhancement experiments using a four-channel acoustic-vector-sensor (AVS) microphone array are demonstrated for the target speech signal corrupted by two and five interfering speech signals and superior performance are achieved.

Index Terms: speech enhancement, microphone arrays, correlation, convergence, adaptive signal processing

1. Introduction

Microphone arrays have been widely studied for speech enhancement in many applications such as teleconferencing, hearing aids, voice communication and automatic speech recognition. In the literature, although the iterative-adaptive techniques of the linearly constrained adaptive beamformer by Frost [1] and the generalized sidelobe canceller (GSC) by Griffiths and Jim [2] have been widely studied for suppressing statistically stationary interfering sources that are uncorrelated with the target source, they tend to adapt slowly or inaccurately in the presence of nonstationary interfering sources [3]. Therefore, their performance are compromised for speech enhancement with multiple interfering speech sources. On the other hand, the nonadaptive techniques of the optimal Capon beamformer [4] and the frequency-domain minimum-variance-distortionless-response (MVDR) beamformer (FMV) [3] have been shown superior to the Frost’s algorithm and the GSC algorithm for suppressing the interfering speech sources. However, they require the inversion of the time-domain or spatio-spectral correlation matrix, which is often unstable and is expensive for large arrays.

Recently, an iterative adaptation procedure was presented in [5] and [6] for computing the optimum FMV solution. It was shown that the iterative algorithms can avoid the matrix inversion and stably converge to the FMV solution. However, the current algorithm bases on the steepest descent method [7] that tends to converge slowly in the speech enhancement of multiple interfering sources. As a result, the current steepest descent (SD) algorithm may ultimately produce unacceptable computational load and processing time. In this paper, we propose a fast-converging algorithm for the FMV solution based on the conjugate gradient method [8]. Theoretically, in the four-microphone case, the conjugate gradient (CG) algorithm is guaranteed to converge in 4 or less iterations for noiseless conditions regardless the number of interfering sources. Therefore, the total computational load and processing time can be significantly reduced. The proposed CG approach is evaluated in the speech enhancement experiments using a four-channel acoustic-vector-sensor (AVS) microphone array. Two and five interfering speech sources are tested. In addition, both the sample-averaging estimator and exponentially decaying estimator for the spatio-spectral correlation matrix are used for comparisons.

2. The fast-converging algorithm for FMV solution

2.1. The optimum FMV solution

We consider the M-microphone system as shown in Fig. 1. The microphones receive the target speech from the look direction contaminated by noise and interfering speech signals. By applying a short-time fast Fourier
Figure 1: M-Microphone array system and frequency-domain beamformer for speech enhancement in the look direction.

The time-domain beamformer output \( y(k,n) \) can be obtained by taking the short-time inverse Fourier transform of \( Y(k,n) \) and \( Y_h(k,n) \), respectively, and \( o(k) \) is the speech power spectral density and \( e(k) \) is the interference-plus-noise signal. The FMV beamformer output is given by

\[
\mathbf{w}_{F MV} = \frac{\mathbf{R}_{xx}^{-1} \mathbf{e}}{\mathbf{e}^H \mathbf{R}_{xx} \mathbf{e}}
\]

(5)

For nonstationary speech signals, this optimal weight vector (5) must be computed for each frame estimate of the correlation matrix at each frequency bin, with a computational complexity of \( \mathcal{O}(KM^3) \) for matrix inversions where \( K \) is the FFT length. This could introduce very high computational load for battery-powered devices of limited power or performance, or for high-dimensional array processing applications. In addition, the computation of the matrix inverse may also introduce stability problems. The work in [6] shows that even if the correlation matrix of \( \mathbf{R}_{xx} \) is replaced by the interference-plus-noise correlation matrix \( \mathbf{R}_{vv} \), the use of a pre-computed \( \mathbf{R}_{xx}^{-1} \) often produces a poor performance in the actual noise field. Therefore, it is highly desirable to reduce the computational load while avoiding the stability problems of the matrix inversion. The work of [5] and [6] proposed to use an iterative steepest descent (SD) algorithm. However, the SD algorithm has slow convergence problem for multiple interfering sources.

2.2. Fast-converging algorithm for FMV solution

To overcome the above-mentioned issues, in this section we design a fast-converging algorithm for iteratively computing the FMV solution using the conjugate gradient (CG) method that has been widely used in adaptive filtering [8]. To solve the constrained minimization problem (4), we consider the following quadratic cost function by the method of Lagrange multipliers:

\[
\mathcal{F}(\mathbf{w}) = \frac{1}{2} \mathbf{w}^H \mathbf{R}_{xx} \mathbf{w} + \operatorname{Re}\{\lambda(\mathbf{w}^H \mathbf{e} - 1)\}
\]

(6)

where \( \lambda \in \mathbb{C} \) is Lagrange multiplier, and \( \operatorname{Re}\{\cdot\} \) denotes the real value. To find the optimal solution of (6), we use the CG method to iteratively update the weight vector as:

\[
\mathbf{w}_{l+1} = \mathbf{w}_l + \alpha_l \mathbf{p}_l
\]

(7)

\[
\mathbf{p}_{l+1} = \mathbf{g}_{l+1} + \beta_l \mathbf{p}_l
\]

(8)

where \( l \) is the iteration index, \( \alpha_l, \beta_l \) are the step size factors, \( \mathbf{p}_l \) and \( \mathbf{g}_{l+1} \) are the weight update vector and residual vector, respectively. The weight vector is initialized as the conventional matched filter \( \mathbf{w}_1 = \mathbf{e}/\|\mathbf{e}\|^2 \) the same as in the SD algorithm [5]. \( \mathbf{p}_1 \) is initialized as \( \mathbf{p}_1 = \mathbf{g}_1 \) and the residual vector \( \mathbf{g}_l \) is derived from the derivation of (6) as

\[
\mathbf{g}_l = -\nabla_{\mathbf{w}_l} \mathcal{F}(\mathbf{w}_l) = -\mathbf{R}_{xx} \mathbf{w}_l - \lambda \mathbf{e}
\]

(9)

where * denotes complex conjugate. To eliminate \( \lambda \) from (9), we substitute (7) into the constraint equation \( \mathbf{w}_{l+1}^H \mathbf{e} = 1 \) and obtain that \( \mathbf{p}_l^H \mathbf{e} = 0 \). By noting
that the weight update vector $p_l$ spans the subspace of $\{g_1, \ldots, g_l\}$ and $g_i^H R_j = 0, i \neq j$, we obtain that $g_i^H e = 0$. By eliminating $\lambda$ from (9), the residual vector is finally expressed as

$$g_l = -Rw_l$$  \hspace{1cm} (10)

where $R \triangleq \left( I - \frac{ee^H}{\|e\|^2} \right) R_{xx}$.

The parameter $\alpha_l$ is selected such that the next-step residual vector $g_{l+1}$ is orthogonal to the current weight update vector $p_l$, that is

$$p_l^H g_{l+1} = 0$$  \hspace{1cm} (11)

Substituting (7) into (10) for $g_{l+1}$ and from (11), we have

$$p_l^H g_l - \alpha_l p_l^H R p_l = 0$$  \hspace{1cm} (12)

Simply manipulating (12), we obtain the step size $\alpha_l$ as

$$\alpha_l = \frac{p_l^H g_l}{p_l^H R p_l + \gamma_1}$$  \hspace{1cm} (13)

where $\gamma_1$ is a small constant to avoid division by zero.

The step size $\beta_l$ is selected such that the next-step weight update direction vector $p_{l+1}$ is linearly independent and $R$-conjugate to the previous direction vector:

$$p_{l+1}^H R p_l = 0$$  \hspace{1cm} (14)

Substituting (8) into (14), the step size $\beta_l$ is given by

$$\beta_l = \frac{g_{l+1}^H R p_l}{p_l^H R p_l + \gamma_2},$$

where $\gamma_2$ has the same function as $\gamma_1$.

Simplifying $\beta_l$ by the equations $g_{l+1} = g_l - \alpha_l R p_l$ and $g_{l+1} = 0$, we finally have

$$\beta_l = \frac{g_{l+1}^H g_{l+1}}{\alpha_l p_l^H R p_l + \gamma_2}$$  \hspace{1cm} (15)

The complete proposed CG algorithm for the FMV solution is listed in Table 1 and the SD algorithm [5] is also shown for comparison. The asymptotic convergence of the CG algorithm was proved in [9]. Note that the alternative expressions for $\alpha_l$ and $\beta_l$ in the literature can also be applied. For each iteration the CG algorithm has a computational complexity of $0.5M^2 + 6.5M + 1$ multiplications and 2 divisions and the SD algorithm $M^2 + 4M$ multiplications and 1 divisions.

### 2.3. The estimators for correlation matrix

When implementing the FMV beamformer, the spatio-spectral correlation matrix $R_{xx}$ is usually unknown and it needs to be estimated for each date update. The conventional estimation approaches are shown as follows:

1) The sample-averaging (SA) estimator

$$R_{sa}(k,n) = \frac{1}{C} \sum_{c=0}^{C-1} x(k,n-c)x^H(k,n-c)$$  \hspace{1cm} (16)

### Table 1: The SD and the proposed CG algorithm for iterative computation of the FMV solution.

<table>
<thead>
<tr>
<th>SD Algorithm</th>
<th>CG Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Initialization: $w_1 = \frac{e}{|e|^2}$</td>
<td>1. Initialization: $w_1 = \frac{e}{|e|^2}$</td>
</tr>
<tr>
<td>$R = \left( I - \frac{ee^H}{|e|^2} \right) R_{xx}$</td>
<td>$R = \left( I - \frac{ee^H}{|e|^2} \right) R_{xx}$</td>
</tr>
<tr>
<td>$p_1 = g_1 = -Rw_1$</td>
<td>$p_1 = g_1 = -Rw_1$</td>
</tr>
<tr>
<td>2. Iterative computation:</td>
<td>2. Iterative computation:</td>
</tr>
<tr>
<td>For $l = 1, 2, \ldots$</td>
<td>For $l = 1, 2, \ldots$</td>
</tr>
<tr>
<td>$g_l = Rw_l$</td>
<td>$\alpha_l = \frac{g_l^H g_l}{p_l^H R p_l + \gamma_1}$</td>
</tr>
<tr>
<td>$\mu_l = g_l^H R_{xx} w_l$</td>
<td>$w_{l+1} = w_l + \alpha_l p_l$</td>
</tr>
<tr>
<td>$w_{l+1} = w_l - \mu_l g_l$</td>
<td>$g_{l+1} = g_l - \alpha_l R p_l$</td>
</tr>
<tr>
<td>$\beta_l = \frac{g_{l+1}^H g_{l+1}}{\alpha_l^2 p_l^H R p_l + \gamma_2}$</td>
<td>$\beta_l = \frac{g_{l+1}^H g_{l+1}}{\alpha_l^2 p_l^H R p_l + \gamma_2}$</td>
</tr>
<tr>
<td>$p_{l+1} = g_{l+1} + \beta_l p_l$</td>
<td>$p_{l+1} = g_{l+1} + \beta_l p_l$</td>
</tr>
<tr>
<td>3. Terminate if $g_l \approx 0$</td>
<td>3. Terminate if $g_l \approx 0$</td>
</tr>
</tbody>
</table>

where $C$ is number of data samples.

2) The exponentially decaying (ED) estimator

$$R_{ed}(k,n) = \lambda R_{ed}(k,n-1) + x(k,n)x^H(k,n)$$  \hspace{1cm} (17)

where $\lambda$ is the smoothing factor. Note that the matrix regularization (diagonal loading) should be used in order to get a robust beamformer [3],

$$\hat{R}_{xx}(k,n) = R_{sa}(k,n) + \gamma I$$  \hspace{1cm} (18)

where $\hat{R}_{xx}(k,n)$ is the estimates of $R_{xx}(k,n)$. $I$ is the identity matrix and $\gamma$ a small constant.

### 3. Experimental results

The proposed CG approach is evaluated for speech enhancement using a microphone array pictured in Fig. 2. In the array, three bidirectional microphones X, Y, Z and one omnidirectional microphone are mounted to approximate an AVS sensor. The microphone response were obtained through calibrations as detailed in [3].

To obtain the simulated azimuth of the microphone signals, a series of high-context speech sentences from BBC radio were convolved with the corresponding microphone impulse response. The microphone impulse responses were truncated to 128 samples. For each source, there are 20 sentences including 10 male voices and 10 female voices used in our tests. Each sentence is 10 seconds long and quantized at 16 bits with a sampling rate of 44.1 kHz. The noisy microphone signals were the digital superposition of the target speech source, the interfering speech sources and the additive Gaussian noise.
time-domain input signals were transformed into the frequency domain via a 512-point FFT, using a Hann window with a length of 512 samples and 75% overlap. The parameters were set as $C = 10$, $\lambda = 0.99$, $\gamma = 0.01$. All the algorithms were implemented in Matlab. Two tests of different number of interfering sources were demonstrated. In test #1, the test azimuth of the target speech source was 0 degree right facing the microphone X and the two interfering sources were 60 and 90 degrees to the right. In test #2, the target source was the same angle and the five interfering sources were 40, 60 and 90 degrees to the left and 60 and 90 degrees to the right.

Fig. 3 shows the convergence to the FMV solution of the SD algorithm and CG algorithm in the two tests with input SNR of 0 dB and the estimator $R_{sa}$. It is observed that the CG algorithm has a fast convergence compared to the SD algorithm. In case of more interfering sources in test #2, the SD algorithm converges very slowly. Our tests show that the CG algorithm converges in 3 iterations for both estimators and different input SNRs. Fig. 4 shows the output signal-to-noise ratio (SNR) enhancement results for different input SNRs using the proposed CG algorithm after 3-iteration runs. In test #1, the ED estimator is clearly superior to the SA estimator and produces SNR improvements of 10 to 15 dB more. In test #2, the SA estimator performs better in low input SNRs, while the ED estimator is better for high input SNRs.

4. Conclusions

We have presented a CG approach for iteratively computing the optimum solution of the FMV beamformer. The proposed CG approach significantly improves algorithm convergence for speech enhancement of multiple interfering sources while avoiding stability problem in direct computation of matrix inversion. The evaluations of the proposed CG approach for speech enhancement using a VAS microphone array show big improvements in the output SNRs for the less-sources-than-sensors case and moderate improvements for the more-sources-than-sensors case.

5. Acknowledgements

This study is supported by the research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapore’s Agency for Science, Technology and Research (A*STAR).

6. References


