AN ACOUSTIC MIMO FRAMEWORK FOR ANALYZING MICROPHONE-ARRAY BEAMFORMING

Jingdong Chen[1]*, Jacob Benesty*[2]*, and Yiteng (Arden) Huang*[1]

[1]: Bell Labs, Lucent Technologies
600 Mountain Avenue
Murray Hill, NJ, 07974, USA
e-mail:{jingdong,arden}@research.bell-labs.com

[2]: Université du Québec, INRS-EMT
800 de la Gauchetière Ouest, Suite 6900
Montréal, Québec, H5A 1K6, Canada
e-mail: benesty@emt.inrs.ca

ABSTRACT

Although a significant amount of research attention has been devoted to microphone-array beamforming, the performance of all the developed algorithms in practical acoustic environments is still far from meeting our expectation. So further research efforts on this topic are indispensable. In this paper, we treat a microphone array as a multiple-input multiple-output (MIMO) system and develop a general framework for analyzing performance of beamforming algorithms based on the acoustic MIMO channel impulse responses. Under this framework, we study the bounds for the length of beamforming filter, which in turn shows the performance bounds of beamforming in terms of speech dereverberation and interference suppression. We also discuss the intrinsic relationships among different classical beamforming techniques and explain, from the channel condition point of view, what the prerequisites have to be fulfilled in order for those techniques to work.

Index Terms— Microphone Arrays, Beamforming, LCMV, MINT, MVDR.

1. PROBLEM FORMULATION

The problem considered in this paper can be described as an $M \times N$ system, where we have M sources in the sound field and we use N microphones to observe signals from their field of view. The output of the nth microphone is given by

$$x_n(k) = \sum_{m=1}^{M} \mathbf{h}_{nm}^T \mathbf{s}_m(k) + b_n(k), \qquad n = 1, 2, \cdots, N, \tag{1}$$

where

$$\mathbf{h}_{nm} = \left[\begin{array}{cccc} h_{nm,0} & h_{nm,1} & \cdots & h_{nm,L_h-1} \end{array} \right]^T$$

is the acoustic channel impulse response from Source m to Microphone n, L_h is the length of the longest channel impulse response,

$$\mathbf{s}_m(k) = \left[\begin{array}{cccc} s_m(k) & s_m(k-1) & \cdots & s_m(k-L_h+1) \end{array} \right]^T,$$

$b_n(k)$ is the noise observed at the nth microphone, and ^{T} denotes the transpose of a vector or a matrix.

Given the above signal model, the array processing is to estimate some of the M source signals from the microphone observations $x_n(k)$ $(n = 1, 2, \cdots, N)$. Suppose that there are P $(P > 0)$ desired signals that we want to estimate. Without loss of generality, we assume that the first P signals, i.e., $s_p(k)$, $p = 1, 2, \cdots, P$, are the desired sources while the other Q source signals $s_{P+q}(k),\ q = 1, 2, \cdots, Q$, are the interferers, where $P + Q = M$. Then the objective of the array processing becomes to extract the signals $s_p(k),\ p = 1, 2, \cdots, P$, from the given observation signals $x_n(k)$, $n = 1, 2, \cdots, N$. For ease of analysis, let us neglect the noise terms $b_n(k)$ in (1). In this case, the estimation of the source signals would involve two processing operations: dereverberation and interference suppression.

Now suppose that we can achieve an estimate of $s_p(k)$ by applying N filters to the N microphone outputs, i.e.,

$$y_p(k) = \sum_{n=1}^{N} \mathbf{g}_{pn}^T \mathbf{x}_n(k), \qquad p = 1, 2, \cdots, P, \tag{2}$$

where

$$\begin{aligned} \mathbf{g}_{pn} &= \left[\begin{array}{cccc} g_{pn,0} & g_{pn,1} & \cdots & g_{pn,L_g-1} \end{array} \right]^T, \\ \mathbf{x}_n(k) &= \left[\begin{array}{cccc} x_n(k) & x_n(k-1) & \cdots & x_n(k-L_g+1) \end{array} \right]^T, \end{aligned}$$

L_g is the length of the $\mathbf{g}$ filters. A legitimate question then arises: is it possible to find $\mathbf{g}_{pn}$ in such a way that $y_p(k) = s_p(k-\tau_p)$ (where τ_p is a delay constant)? In other words, is it possible to perfectly recover $s_p(k)$ (up to a constant delay)? We will answer this question in the following sections. But before continuing, we make another assumption. We assume that the number of microphones used is greater than, or at least equal to the number of sound sources, i.e., $N \geq M$.

2. LEAST-SQUARES AND MINT APPROACHES

The microphone signals can be rewritten in the following form,

$$\mathbf{x}_n(k) = \sum_{m=1}^{M} \mathbf{H}_{nm} \mathbf{s}_{L,m}(k),\ n = 1, 2, \cdots, N, \tag{3}$$

where

$$\mathbf{H}_{nm} = \left[\begin{array}{cccccc} h_{nm,0} & \cdots & h_{nm,L_h-1} & 0 & \cdots & 0 \\ 0 & h_{nm,0} & \cdots & h_{nm,L_h-1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & h_{nm,0} & \cdots & h_{nm,L_h-1} \end{array} \right]$$

is a Sylvester matrix of size $L_g \times L$, with $L = L_g + L_h - 1$, and $\mathbf{s}_{L,m}(k) = [s_m(k)\ s_m(k-1)\ \cdots\ s_m(k-L+1)]^T$, $m = 1, 2, \cdots, M$. Plugging (3) into (2), we find that

$$y_p(k) = \sum_{m=1}^{M} \left[\sum_{n=1}^{N} \mathbf{g}_{pn}^T \mathbf{H}_{nm} \right] \mathbf{s}_{L,m}(k),\ p = 1, 2, \cdots, P. \tag{4}$$

From the above expression, we see that in order to perfectly recover $s_p(k)$, the following condition has to be satisfied:

$$\mathbf{H}^T\mathbf{g}_p = \mathbf{u}'_p, \tag{5}$$

where

$$\begin{aligned}
\mathbf{H} &= \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} & \cdots & \mathbf{H}_{1M} \\ \mathbf{H}_{21} & \mathbf{H}_{22} & \cdots & \mathbf{H}_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{H}_{N1} & \mathbf{H}_{N2} & \cdots & \mathbf{H}_{NM} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{H}_{:1} & \mathbf{H}_{:2} & \cdots & \mathbf{H}_{:M} \end{bmatrix}, \\
\mathbf{g}_p &= \begin{bmatrix} \mathbf{g}_{p1}^T & \mathbf{g}_{p2}^T & \cdots & \mathbf{g}_{pN}^T \end{bmatrix}^T, \\
\mathbf{u}'_p &= \begin{bmatrix} \underbrace{\mathbf{0}_{L\times 1}^T \cdots \mathbf{0}_{L\times 1}^T}_{(p-1)L} & \mathbf{u}_p^T & \underbrace{\mathbf{0}_{L\times 1}^T \cdots \mathbf{0}_{L\times 1}^T}_{(M-p)L} \end{bmatrix}^T,
\end{aligned}$$

and

$$\mathbf{u}_p = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix}^T$$

is a vector of length L, whose τ_pth component is equal to 1. The channel matrix $\mathbf{H}$ is of size $NL_g \times ML$. Depending on the values of N and M, we have two cases, namely, $N = M$ and $N > M$.

Case 1: $N = M$.

In this case, $ML = NL = NL_g + NL_h - N$. Since $L_h > 1$, we have $ML > NL_g$. This means that the number of rows of $\mathbf{H}^T$ is always larger than its number of columns. Now let's assume that the matrix $\mathbf{H}^T$ has full column rank. In this situation, the best estimator that we can derive from (5) is the least-squares solution, i.e.,

$$\mathbf{g}_p^{\mathrm{LS}} = \left[\mathbf{H}\mathbf{H}^T\right]^{-1}\mathbf{H}\mathbf{u}'_p. \tag{6}$$

However, this solution may not be good enough in practice for the following reasons: 1. we do not know how to determine L_g; 2. the whole impulse response matrix $\mathbf{H}$ must be known to find the optimal filter in the LS sense, and thus there is very little flexibility with this method.

Case 2: $N > M$.

With more microphones than sources, is it possible to find a better solution than the LS one? Let $M = N - K$, $K > 0$. In fact, requiring $\mathbf{H}^T$ to have a number of rows that is equal to or larger than its number of columns, we find this time an upper bound for L_g,

$$L_g \leq (N/K - 1)(L_h - 1). \tag{7}$$

If we take

$$L_g = (N/K - 1)(L_h - 1), \tag{8}$$

and if L_g is an integer, $\mathbf{H}^T$ is now a square matrix. Therefore,

$$\mathbf{g}_p^{\mathrm{MINT}} = \left[\mathbf{H}^T\right]^{-1}\mathbf{u}'_p. \tag{9}$$

This expression is exactly the MINT method [2], which can perfectly recover the signal of interest $s_p(k)$ if $\mathbf{H}$ is known or can be accurately estimated. Of course, we supposed that $\mathbf{H}^T$ is of full rank, which is equivalent to saying that the polynomials formed from $h_{1m}, h_{2m}, \cdots, h_{Nm}$, $m = 1, 2, \cdots, M$, share no common zeroes.

It is very interesting to see that, if we have more microphones than sources, we have more flexibility in estimation of the signals of interest and have a better idea for the choice of L_g.

3. FROST ALGORITHM

Following (3), if concatenating the N observation vectors together, we get:

$$\mathbf{x}(k) = \begin{bmatrix} \mathbf{x}_1^T(k) & \mathbf{x}_2^T(k) & \cdots & \mathbf{x}_N^T(k) \end{bmatrix}^T = \mathbf{H}\,\mathbf{s}_{ML}(k),$$

where $\mathbf{s}_{ML}(k) = \begin{bmatrix} \mathbf{s}_{L,1}^T(k) & \mathbf{s}_{L,2}^T(k) & \cdots & \mathbf{s}_{L,M}^T(k) \end{bmatrix}^T$. The covariance matrix corresponding to $\mathbf{x}(k)$ is:

$$\mathbf{R}_{xx} = E\left\{\mathbf{x}(k)\mathbf{x}^T(k)\right\} = \mathbf{H}\mathbf{R}_{ss}\mathbf{H}^T, \tag{10}$$

with $\mathbf{R}_{ss} = E\left\{\mathbf{s}_{ML}(k)\mathbf{s}_{ML}^T(k)\right\}$. We assumed that $\mathbf{R}_{xx}$ is invertible, which is equivalent to stating that the $\mathbf{R}_{ss}$ matrix is of full rank and $\mathbf{H}^T$ matrix has full column rank. We are now ready to study two interesting cases.

Case 1: Partial Knowledge of the Impulse Response Matrix.

In this case, we wish to extract source $s_p(k)$ with only the knowledge of $\mathbf{H}_{:p}$, i.e., the impulse responses from that source to the N microphones. With this information, the LCMV filter is obtained by solving the following problem [1]:

$$\min_{\mathbf{g}_p} \mathbf{g}_p^T\mathbf{R}_{xx}\mathbf{g}_p \quad \text{subject to} \quad \mathbf{H}_{:p}^T\mathbf{g}_p = \mathbf{u}_p. \tag{11}$$

Hence,

$$\mathbf{g}_p^{\mathrm{LCMV1}} = \mathbf{R}_{xx}^{-1}\mathbf{H}_{:p}\left[\mathbf{H}_{:p}^T\mathbf{R}_{xx}^{-1}\mathbf{H}_{:p}\right]^{-1}\mathbf{u}_p. \tag{12}$$

We refer to this approach as the LCMV1, where a necessary condition for $\left[\mathbf{H}_{:p}^T\mathbf{R}_{xx}^{-1}\mathbf{H}_{:p}\right]$ to be nonsingular is to have $NL_g \geq L$, which implies that

$$L_g \geq (L_h - 1)/(N - 1). \tag{13}$$

An important thing to observe is that the minimum length required for the filters $\mathbf{g}_{pn}^{\mathrm{LCMV1}}$, $n = 1, 2, \cdots, N$, decreases as the number of microphones increases. As a consequence, the Frost filter has the potential to significantly reduce the effect of the interferers with a large number of microphones.

If we take the minimum required length for L_g, i.e., $L_g = (L_h - 1)/(N-1)$ and assume that L_g is an integer, $\mathbf{H}_{:p}$ turns to be a square matrix and (12) becomes:

$$\mathbf{g}_p^{\mathrm{LCMV1}} = \left[\mathbf{H}_{:p}^T\right]^{-1}\mathbf{u}_p = \begin{bmatrix} \mathbf{H}_{1p}^T & \mathbf{H}_{2p}^T & \cdots & \mathbf{H}_{Np}^T \end{bmatrix}^{-1}\mathbf{u}_p, \tag{14}$$

which is the MINT method [2]. So the MINT method is a particular case of the Frost algorithm. Although never shown before, this result should not come as a surprise since the motivation behind the two approaches is similar.

We assumed in (14) that $\mathbf{H}_{:p}$ has full rank, which requires that the N polynomials formed from $h_{1p}, h_{2p}, \cdots, h_{Np}$ share no common zeros. From (10), we can deduce that a necessary condition for $\mathbf{R}_{xx}$ to be invertible is to have $NL_g \leq ML$. When $M = N$, i.e., the number of sources is equal to the number of microphones, this condition is always true, which means that there is no upper bound for L_g. When $N > M$, assume that $M = N - K$, $K > 0$, this condition becomes

$$L_g \leq (N/K - 1)(L_h - 1). \tag{15}$$

Combining (15) and (13), we see how L_g is bounded, i.e.,

$$(L_h - 1)/(N - 1) \leq L_g \leq (N/K - 1)(L_h - 1). \tag{16}$$

Case 2: Full Knowledge of the Impulse Response Matrix and $N > M$.

Here, we wish to extract source $s_p(k)$ with the full knowledge of the impulse response matrix $\mathbf{H}$, with $M = N - K$, $K > 0$. Taking all this information into account in our optimization problem,

$$\min_{\mathbf{g}_p} \mathbf{g}_p^T \mathbf{R}_{xx} \mathbf{g}_p \quad \text{subject to} \quad \mathbf{H}^T \mathbf{g}_p = \mathbf{u}'_p, \tag{17}$$

we find the solution,

$$\mathbf{g}_p^{\mathrm{LCMV2}} = \mathbf{R}_{xx}^{-1} \mathbf{H} \left[\mathbf{H}^T \mathbf{R}_{xx}^{-1} \mathbf{H} \right]^{-1} \mathbf{u}'_p. \tag{18}$$

We refer to this approach as the LCMV2, where we assumed that both $\mathbf{R}_{xx}$ and $\left[\mathbf{H}^T \mathbf{R}_{xx}^{-1} \mathbf{H} \right]$ are nonsingular and their inverse matrices exist. From the previous analysis, we know that in order for $\mathbf{R}_{xx}$ to be invertible the condition in (15) has to be true. Also, a necessary condition for $\left[\mathbf{H}^T \mathbf{R}_{xx}^{-1} \mathbf{H} \right]$ to be nonsingular is to have $NL_g \geq ML$, which implies that

$$L_g \geq (N/K - 1)(L_h - 1). \tag{19}$$

Therefore, the only condition for (18) to exist is that:

$$L_g = (N/K - 1)(L_h - 1), \tag{20}$$

and this value needs to be an integer. In this case, $\mathbf{H}$ is a square matrix and (18) becomes:

$$\mathbf{g}_p^{\mathrm{LCMV2}} = \left[\mathbf{H}^T \right]^{-1} \mathbf{u}'_p, \tag{21}$$

which is also the MINT solution [2].

4. GENERALIZED SIDELOBE CANCELLER

The generalized sidelobe canceller (GSC) transforms the LCMV algorithm from a constrained problem into an unconstrained form [3]. Consider the linearly constrained optimization problem given in (11). If we assume that $L_g > (L_h - 1)/(N - 1)$ so that the nullspace of $\mathbf{H}_{:p}^T$ not to be equal to zero (this indicates that the GSC structure makes sense only for the LCMV1 filter), the GSC method can be formulated as [4]:

$$\mathbf{g}_p = \mathbf{f}_p - \mathbf{B}_p \mathbf{w}_p, \tag{22}$$

where

$$\mathbf{f}_p = \mathbf{H}_{:p} \left[\mathbf{H}_{:p}^T \mathbf{H}_{:p} \right]^{-1} \mathbf{u}_p \tag{23}$$

is the minimum-norm solution of $\mathbf{H}_{:p}^T \mathbf{f}_p = \mathbf{u}_p$ and $\mathbf{B}_p$ is the blocking matrix that spans the nullspace of $\mathbf{H}_{:p}^T$, i.e. $\mathbf{H}_{:p}^T \mathbf{B}_p = \mathbf{0}$. The size of $\mathbf{B}_p$ is $NL_g \times (NL_g - L)$, where $NL_g - L$ is the dimension of the nullspace of $\mathbf{H}_{:p}^T$. Therefore, $\mathbf{w}_p$ is a vector of length $NL_g - L = (N-1)L_g - L_h + 1$, which is obtained from the following unconstrained optimization problem:

$$\min_{\mathbf{w}_p} (\mathbf{f}_p - \mathbf{B}_p \mathbf{w}_p)^T \mathbf{R}_{xx} (\mathbf{f}_p - \mathbf{B}_p \mathbf{w}_p), \tag{24}$$

and the solution is:

$$\mathbf{w}_p^{\mathrm{GSC}} = \left[\mathbf{B}_p^T \mathbf{R}_{xx} \mathbf{B}_p \right]^{-1} \mathbf{B}_p^T \mathbf{R}_{xx} \mathbf{f}_p. \tag{25}$$

It has been shown that,

$$\begin{aligned} \mathbf{g}_p^{\mathrm{LCMV1}} &= \mathbf{R}_{xx}^{-1} \mathbf{H}_{:p} \left[\mathbf{H}_{:p}^T \mathbf{R}_{xx}^{-1} \mathbf{H}_{:p} \right]^{-1} \mathbf{u} \\ &= \left\{ \mathbf{I} - \mathbf{B}_p \left[\mathbf{B}_p^T \mathbf{R}_{xx} \mathbf{B}_p \right]^{-1} \mathbf{B}_p^T \mathbf{R}_{xx} \right\} \mathbf{f}_p = \mathbf{g}_p^{\mathrm{GSC}} \end{aligned} \tag{26}$$

so the LCMV and GSC algorithms are equivalent.

Expressions (22) and (26) have a very nice physical interpretation [compared to (12)]. The LCMV filter $\mathbf{g}_p^{\mathrm{LCMV}}$ is the sum of two orthogonal vectors $\mathbf{f}_p$ and $-\mathbf{B}_p \mathbf{w}_p^{\mathrm{GSC}}$, which serve for different purposes. The objective of the first vector, $\mathbf{f}_p$, is to perform dereverberation on the signal $s_p(k)$, while the objective of the second vector, $-\mathbf{B}_p \mathbf{w}_p^{\mathrm{GSC}}$, is to reduce the effect of the interference. Increasing the length L_g of the filters $\mathbf{g}_p^{\mathrm{LCMV}}$ from its minimum value $(L_h - 1)/(N - 1)$ will not change anything on the dereverberation part. However, increasing L_g will augment the dimension of the nullspace of $\mathbf{H}_{:p}^T$, and hence the length of $\mathbf{w}_p^{\mathrm{GSC}}$. As a result, better interference suppression is expected. It is obvious, from a theoretical point of view, that perfect dereverberation is possible (if $\mathbf{H}_{:p}$ is known or can be accurately estimated) but perfect interference suppression is not. In practice, if all the impulse responses h_{np} $(n = 1, \cdots, N)$ can be estimated, we can expect good dereverberation but interference suppression may be limited for the simple reason that it will be very hard to make L_g much larger than L_h (the length of the impulse responses h_{np}).

To find the bounds for the length of $\mathbf{w}_p^{\mathrm{GSC}}$, we consider two situations. The first one is when $N = M$. In this case, we know from the previous section that there is no upper bound for L_g. This implies that $\mathbf{w}_p^{\mathrm{GSC}}$ can be taken as large as we wish. As a result, we can expect better interference suppression as L_g is increased. By increasing the number of microphones (with $N = M$), the minimum length required for L_g will decrease compared to L_h, which is a very good thing because in practice acoustic impulse responses can be very long.

Our second situation is when we have more microphones than sources. Assume that $M = N - K$, $K > 0$. Using (16) and the fact that $L_{\mathbf{w}_p} = (N-1)L_g - L_h + 1$, we can easily deduce the bounds for the length of $\mathbf{w}_p^{\mathrm{GSC}}$:

$$\begin{aligned} 0 < L_{\mathbf{w}_p} &\leq \frac{N}{K}(N - K - 1)(L_h - 1) \\ &\leq \frac{N}{K}(N - K - 1)(N - 1)L_g. \end{aligned} \tag{27}$$

This means that there is a limit to interference suppression. Consider the scenario where we have one desired source only ($P = 1$) and Q interferers. We have $M = Q + 1 = N - K$ and (27) is now:

$$0 < L_{\mathbf{w}_p} \leq \frac{NQ}{N - Q - 1}(L_h - 1) \leq \frac{N(N-1)Q}{N - Q - 1} L_g. \tag{28}$$

We see from (28) that if L_h and Q remain the same, when we increase the number of microphones, it will allow us to use a larger value for $L_{\mathbf{w}_p}$ to augment the speech-dereverberation and interference-suppression performance.

5. MVDR APPROACH

The minimum variance distortionless response (MVDR) method, due to Capon [5], is a particular case of the LCMV1. In the original formulation of MVDR, the observation signals were assumed free of reverberation so it applies only one constraint to the direct path of

the desired source. In the presence of reverberation, the constraint for MVDR should be modified as follows,

$$\mathbf{h}_{:p}^T(\kappa_p)\mathbf{g}_p = 1, \tag{29}$$

where $\mathbf{h}_{:p}(\kappa_p)$ is the κ_pth column of the matrix $\mathbf{H}_{:p}$. The aim of this constraint is to align the desired source signal, $s_p(k)$, at the output of the beamformer. Hence, in the MVDR approach, we have the following optimization problem:

$$\min_{\mathbf{g}_p} \mathbf{g}_p^T \mathbf{R}_{xx} \mathbf{g}_p \quad \text{subject to} \quad \mathbf{h}_{:p}^T(\kappa_p)\mathbf{g}_p = 1, \tag{30}$$

whose solution is:

$$\mathbf{g}_p^{\mathrm{MVDR}} = \frac{\mathbf{R}_{xx}^{-1}\mathbf{h}_{:p}(\kappa_p)}{\mathbf{h}_{:p}^T(\kappa_p)\mathbf{R}_{xx}^{-1}\mathbf{h}_{:p}(\kappa_p)}. \tag{31}$$

The minimum required length for the filters $\mathbf{g}_{pn}^{\mathrm{MVDR}}$ is $L_g = \kappa_p$. In this case, the performance of the MVDR beamformer is similar to that of the classical delay-and-sum beamformer. This method does not require the full knowledge of the impulse responses but only the relative delays among microphones. However, it may have the problem of signal self cancellation.

6. EXPERIMENTS

This section compares different algorithms and studies the effect of filter length on beamforming performance. We set up a microphone array system in the varechoic chamber at Bell Labs [which is a room measures 6600 mm long, 5850 mm wide, and 2750 mm high ($x \times y \times z$)]. The array consists of 4 microphones, placed, respectively, at (2437, 5600, 1400), (2537, 5600, 1400), (2637, 5600, 1400), and (2737, 5600, 1400). There are three sources (loudspeakers) in the sound field: one target [$s_1(k)$, is located at (3337, 1438, 1600)], and two interferers [$s_2(k)$ and $s_3(k)$ are placed at (1337, 2938, 1600) and (5337, 2938, 1600) respectively]. The reverberation time T_{60} is controlled to be approximately 0.35 seconds. To make the experiments repeatable, the impulse response from each source to each microphone was measured. These measured impulse responses are then treated as the true ones. In our experiment, the sampling rate is 8 kHz. The impulse responses are truncated to 128 points ($L_h = 128$). $s_1(k)$ is a prerecorded speech signal from a male speaker, and both $s_2(k)$ and $s_3(k)$ are speech signals from a same female speaker. The microphone outputs are obtained by convolving the sources with the impulse responses. The input SIR is -8.25 dB.

To quantitatively assess the performance of interference suppression and speech dereverberation, we evaluate two criteria, namely signal-to-interference ratio (SIR) and speech spectral distortion. The input and output SIRs are given by (see [6] for more details):

$$\mathrm{SIR}^{\mathrm{in}} \triangleq \frac{1}{N}\sum_{n=1}^{N} \frac{E\left\{[h_{n1} * s_1(k)]^2\right\}}{\sum_{m=2}^{M} E\left\{[h_{nm} * s_m(k)]^2\right\}} \tag{32}$$

$$\mathrm{SIR}^{\mathrm{out}} \triangleq \frac{E\left\{[\phi_1 * s_1(k)]^2\right\}}{\sum_{m=2}^{M} E\left\{[\phi_m * s_m(k)]^2\right\}} \tag{33}$$

where $\phi_m = \sum_{n=1}^{N} g_{1n} * h_{nm}$ denotes the impulse response of the equivalent channel between the mth source and the beamforming output. For spectral distortion, we evaluate the Itakura-Saito (IS) distance between $s_1(k)$ and $s_1(k) * \phi_m$, which should measure the amount of reverberation present in the estimated speech signal after beamforming.

Table 1. Performance of different algorithms when the MIMO impulse responses are known *a priori* ("*" indicates the maximum value that the L_g can take for the condition and "×" means that the L_g cannot take this value for the method in the given condition).

	Direct		LCMV1		LCMV2		GSC		MVDR	
L_g	SIR (dB)	IS	SIR (dB)	IS	SIR (dB)	IS	SIR (dB)	IS	SIR (dB)	IS
765*	117.2	0.0	7.9	0.0	117.2	0.0	1.5	0.0	4.4	7.7
700	24.8	0.0	1.3	0.0	×	×	1.3	0.0	4.4	7.6
600	11.2	0.2	0.1	0.0	×	×	0.1	0.0	4.5	7.4
300	4.0	0.2	-6.7	0.0	×	×	-6.7	0.0	3.0	9.1

During beamforming, we assumed that the MIMO impulse response were know *a priori*. Table 1 summarizes the experimental results. Many observations can be made from this table. 1. As the length of the impulse responses, i.e., L_h, increases, the maximum achievable (with the maximum L_g) gain in SIR decreases. This occurs to all the algorithms. 2. In the ideal condition where impulse responses are known and L_g is set to its maximum value, both the direct and LCMV2 (or MINT) techniques can achieve almost perfect interference suppression and speech dereverberation. Similar to the direct and LCMV2 methods, the LCMV1 and GSC can also perform perfect speech dereverberation, but their interference suppression performance is limited. This is mainly because LCMV1 and GSC did not use the channel information from the interferers to the microphones. 3. In each reverberant condition (a fixed L_h), if we reduce the length of the g_1 filter, the amount of interference suppression decreases significantly. 4. The MVDR method is relatively robust to the length of the g_1 filter, but it suffers dramatic signal distortion.

7. CONCLUSIONS

This paper developed a general acoustic MIMO framework for microphone array beamforming. Under this general framework, we analyzed the lower and upper bounds for the length of the beamforming filter, which in turn shows the performance bounds of beamforming in terms of speech dereverberation and interference suppression. We addressed the connection between beamforming and the multiple-input/output inverse theorem (MINT), which was originally developed to achieve the exact inverse filtering of the room acoustics. We also discussed the intrinsic relationships among the most classical beamforming techniques and explained, from the channel condition point of view, what the necessary conditions have to be fulfilled in order for the different beamforming techniques to work.

8. REFERENCES

[1] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, pp. 926–935, Aug. 1972.

[2] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-36, pp. 145–152, Feb. 1988.

[3] 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, Jan. 1982.

[4] C. W. Jim, "A comparison of two LMS constrained optimal array structures," *Proc. IEEE*, vol. 65, pp. 1730–1731, Dec. 1977.

[5] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, pp. 1408–1418, Aug. 1969.

[6] 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 Trans. Speech Audio Process.*, vol. 13, pp. 882–895, Sept. 2005.