On Microphone-Array Beamforming From a MIMO Acoustic Signal Processing Perspective

Jacob Benesty, *Senior Member, IEEE*, Jingdong Chen, *Member, IEEE*, Yiteng (Arden) Huang, *Member, IEEE*, and Jacek Dmochowski

***Abstract*—Although many microphone-array beamforming algorithms have been developed over the past few decades, most such algorithms so far can only offer limited performance in practical acoustic environments. The reason behind this has not been fully understood and further research on this matter is indispensable. In this paper, we treat a microphone array as a multiple-input multiple-output (MIMO) system and study its signal-enhancement performance. Our major contribution is fourfold. First, we develop a general framework for analyzing performance of beamforming algorithms based on the acoustic MIMO channel impulse responses. Second, we study the 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. Third, we address the connection between beamforming and the multiple-input/output inverse theorem (MINT). Finally, we discuss the intrinsic relationships among different classical beamforming techniques and explain, from the channel condition perspective, what the prerequisites are for those techniques to work.**

***Index Terms*—Beamforming, Frost, linearly constrained minimum variance (LCMV), microphone arrays, multiple-input/output inverse theorem (MINT), minimum variance distortionless response (MVDR).**

I. Introduction

Microphone arrays, which consist of sets of microphone sensors that are spatially arranged in specific patterns, have been studied for more than three decades. Such systems have already played, and will continue to play an important role in applications like audio-bridging and teleconferencing where distant or hands-free audio acquisition is required [1], [2].

One of the most important functionalities of microphone arrays is to extract the speech of interest from its observations corrupted by noise, reverberation, and competing sources. The typical method for this is to form a beam and point it to a desired direction. As a result, signals from this so-called look direction is reinforced, while signals from all the other directions are attenuated. Suppose that we have an array consisting of N microphones and denote the microphone outputs as $x_n(k)$ $(n = 1, 2, \ldots, N)$. Beamforming is achieved based on manipulating the signals $x_n(k)$. Many algorithms have been developed over the three decades. The simplest one is the delay-and-sum beamformer as shown in Fig. 1, which was originally investigated in the underwater acoustics and radar antenna areas [3]. The basic idea is to delay (or advance) each microphone output by a proper amount of time so that the signal components from the desired source are synchronized across all sensors. These delayed (or advanced) signals are then weighted and summed together. Since they add up together coherently, the desired signal components are reinforced. In contrast, the other sources and noise are suppressed or even eliminated as they are added together destructively. The weighting coefficients g_n can be either fixed or adaptively determined. The latter leads to an adaptive beamformer [4]–[9]. The advantage of adaptive beamforming over nonadaptive beamforming can roughly be interpreted as follows. From the configuration shown in Fig. 1, at a single frequency, a total of $N-1$ nulls can be formed in the directivity pattern. If we adjust g_n adaptively by taking into account the signal and noise characteristics, the $N-1$ nulls can be properly designed and placed so that noise and interference can be better rejected.

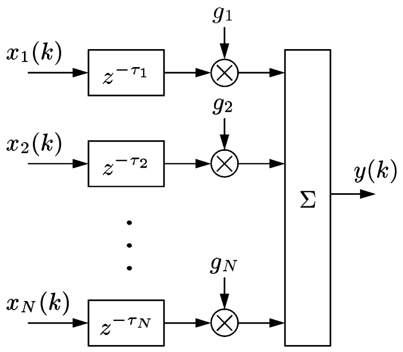


Fig. 1. Structure of a delay-and-sum beamformer.

The above technique was developed for narrowband signals. Although it serves as the basis for any array beamforming, this technique is not very useful for acoustic applications since speech is a typical broadband signal. Consequently, the directivity pattern of the delay-and-sum beamformer would not be the same across a broad frequency band. If we use such a beamformer, then noise and interference signals coming from a direction different from the beamformer's look direction will not be uniformly attenuated over its entire spectrum. This "spectral tilt" results in a disturbing artifact in the array output [10].

Manuscript received March 31, 2006; revised August 24, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hiroshi Sawada.

J. Benesty and J. Dmochowski are with the Université du Québec, INRS-EMT, Montréal, QC H5A 1K6 Canada.

J. Chen and Y. Huang are with Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA (e-mail: jingdong@research.bell-labs.com).

Digital Object Identifier 10.1109/TASL.2006.885251

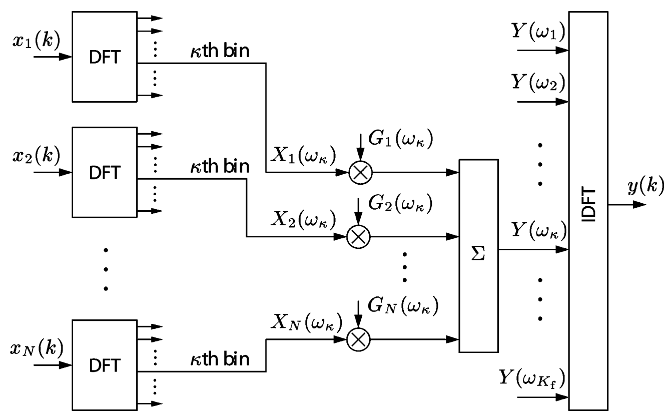


Fig. 2. Structure of a frequency-domain broadband beamformer by narrowband decomposition.

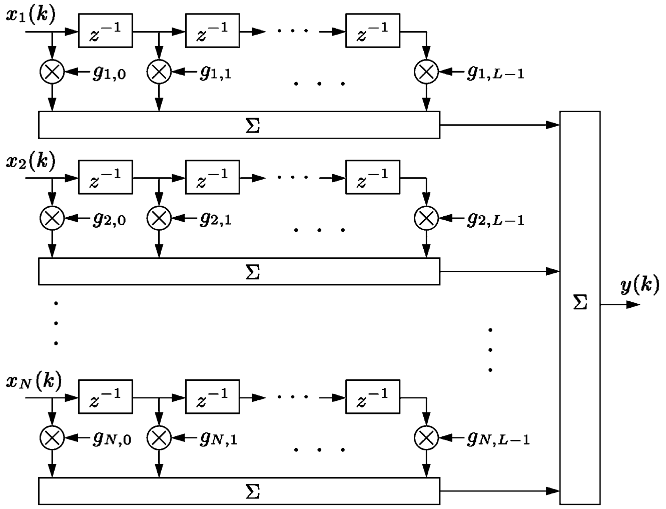


Fig. 3. Structure of a time-domain filter-and-sum beamformer.

One way to overcome this problem is to use harmonically nested subarrays [11], [12]. Every subarray is designed for operating at a single frequency. However, such a solution requires a large array with a great number of microphones, and the array geometry is unusual. Another way to circumvent this problem is to perform narrowband decomposition and design narrowband beamformers independently at each frequency, as shown in Fig. 2. With a simple inspection of Fig. 2, we see that this broadband beamformer is equivalent to applying a finite-duration impulse response (FIR) filter to each microphone output and then summing the filtered signals together as illustrated in Fig. 3. This filter-and-sum algorithm was originally proposed by Frost [13] in the early 1970s and has been intensively studied since then [1], [14]–[32]. In addition to producing a broadband directivity pattern, the effectiveness of this beamformer can also be explained in the following way. As mentioned earlier, the delay-and-sum structure can only produce $N-1$ nulls at a single frequency. By applying FIR filters of length L to N channels, we can now produce $N-1$ nulls at $L-1$ different frequencies. Therefore, this technique offers more flexibility in rejecting noise and interference than the delay-and-sum beamformer. Similar to the delay-and-sum case, the filters' coefficients in Fig. 3 can also be determined either in a nonadaptive way or adaptively [14]–[32]. With adaptive algorithms, the nulls can be properly designed and placed at directions and frequencies for better interference and noise attenuation. However, adaptive beamformers may not be as robust as their nonadaptive counterparts and often lead to signal self cancellation, which is undesirable.

Although so many efforts have been devoted to microphone array processing, the performance of most microphone array beamformers in practical acoustic environments still cannot meet expectation. Apparently, the potential of microphone arrays has not been fully realized as expected. The reasons behind this are very sophisticated and have not been fully understood thus far. Therefore, further research in this area is indispensable.

This paper deals with the signal enhancement problem using microphone arrays. We treat a microphone array as a multiple-input multiple-output (MIMO) system. A general framework based on the MIMO channel impulse responses is then developed for analyzing beamforming performance. Under this framework, we study the bounds for the length of beamforming filter, which in turn show the performance bounds of beamforming in terms of speech dereverberation and interference suppression. We address the connection between beamforming and the multiple-input/output inverse theorem (MINT) [33], which was originally developed to achieve the exact inverse filtering of the room acoustics. We also discuss the intrinsic relationships among different classical beamforming techniques and explain, from the channel condition point of view, what the necessary conditions are for those techniques to work.

The rest of this paper is organized as follows. In Section II, we briefly describe the signal model used in this paper. We then formulate a MIMO framework and discuss the signal estimation problem for an N-element microphone array with M sources $(M \leq N)$ in Section III. We will see that the proposed MIMO framework has more degrees of freedom than the existing approaches. As a result, good performances from a theoretical point of view are possible. In Section IV, we present some experimental results. Finally, some conclusions will be provided in Section V.

II. Problem Description

The problem considered in this paper is illustrated in Fig. 4, where we have M sources in the sound field and we use N microphones to collect signals. We assume that the number of microphones used is greater than, or at least equal to the number of sound sources, i.e., $N \geq M$. The output of the nth microphone is given by

$$x_n(k) = \sum_{m=1}^{M} h_{nm} * s_m(k) + b_n(k), \quad n = 1, 2, \ldots, N \quad (1)$$

where $*$ denotes convolution, $s_m(k)$ is the mth source signal, h_{nm} is the acoustic channel impulse response from the source

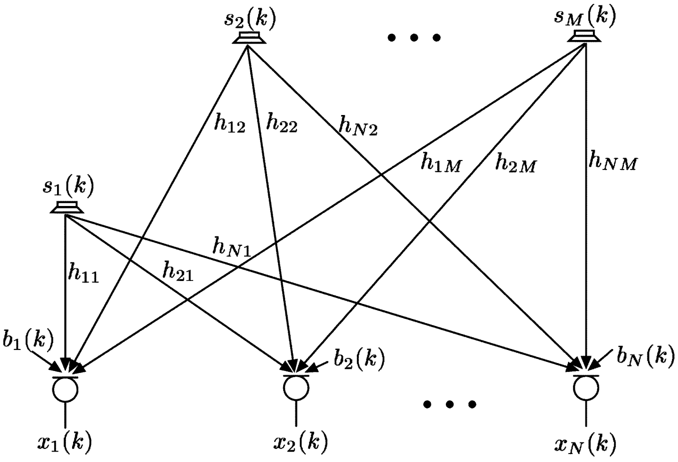


Fig. 4. Illustration of a microphone array system.

m to microphone n, and $b_n(k)$ is the noise observed at the nth microphone. In vector/matrix form, this signal model can be rewritten as

$$\mathbf{x}_{\mathrm{a}}(k) = \mathbf{H}\mathbf{s}(k) + \mathbf{b}(k) \tag{2}$$

where

$$\begin{aligned}
\mathbf{x}_{\mathrm{a}}(k) &= [x_1(k) \quad x_2(k) \quad \ldots \quad x_N(k)]^T \\
\mathbf{H} &= [\mathbf{H}_1 \quad \mathbf{H}_2 \quad \ldots \quad \mathbf{H}_M] \\
\mathbf{H}_m &= \begin{bmatrix} \mathbf{h}_{1m}^T \\ \mathbf{h}_{2m}^T \\ \vdots \\ \mathbf{h}_{Nm}^T \end{bmatrix} \\
\mathbf{h}_{nm} &= [h_{nm,0} \quad h_{nm,1} \quad \ldots \quad h_{nm,L_h-1}]^T \\
n &= 1,2,\ldots,N, \quad m = 1,2,\ldots,M \\
\mathbf{b}(k) &= [b_1(k) \quad b_2(k) \quad \ldots \quad b_N(k)]^T \\
\mathbf{s}(k) &= [\mathbf{s}_1^T(k) \quad \mathbf{s}_2^T(k) \quad \ldots \quad \mathbf{s}_M^T(k)]^T \\
\mathbf{s}_m(k) &= [s_m(k) \quad s_m(k-1) \quad \ldots \quad s_m(k-L_h+1)]^T
\end{aligned}$$

where L_h is the length of the longest channel impulse response, and ^{T} denotes the transpose of a vector or a matrix. Given this signal model, the array processing is to estimate some of the M source signals.

III. Signal Estimation Based on a MIMO Framework

In this section, we discuss how to estimate the desired source signals from the microphone observations. For ease of analysis, let us neglect the noise terms $\mathbf{b}(k)$ in (2). In this situation, the output of the nth microphone at time k is written as

$$x_n(k) = \sum_{m=1}^{M} \mathbf{h}_{nm}^T \mathbf{s}_m(k), \qquad n = 1,2,\ldots,N. \tag{3}$$

Suppose that among the M sources 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., $s_p(k)$, $p = 1,2,\ldots,P$, are the desired sources while the other Q source signals $s_{P+q}(k)$, $q = 1,2,\ldots,Q$, are the interferers, where $P + Q = M$. So the objective of the array processing is to extract the signals $s_p(k)$, $p = 1,2,\ldots,P$ from the given observation signals $x_n(k)$, $n = 1,2,\ldots,N$. This would involve two processing operations: dereverberation and interference suppression. Suppose that we can achieve an estimate of $s_p(k)$ by applying N filters to the N microphone outputs, i.e.,

$$y_p(k) = \sum_{n=1}^{N} \mathbf{g}_{pn}^T \mathbf{x}_n(k), \quad p = 1,2,\ldots,P \tag{4}$$

where

$$\begin{aligned}
\mathbf{g}_{pn} &= [g_{pn,0} \quad g_{pn,1} \quad \cdots \quad g_{pn,L_g-1}]^T \\
p &= 1,2,\ldots,P, \quad n = 1,2,\ldots,N
\end{aligned}$$

are PN filters of length L_g and

$$\begin{aligned}
\mathbf{x}_n(k) &= [x_n(k) \quad x_n(k-1) \quad \ldots \quad x_n(k-L_g+1)]^T \\
n &= 1,2,\ldots,N.
\end{aligned}$$

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 Sections III-A–D.

A. Least-Squares and MINT Approaches

The microphone signals can be rewritten in the following form:

$$\mathbf{x}_n(k) = \sum_{m=1}^{M} \mathbf{H}_{nm} \mathbf{s}_{L,m}(k), \quad n = 1,2,\ldots,N \tag{5}$$

where (see the equation at the bottom of the page) is a Sylvester matrix of size $L_g \times L$, with $L = L_g + L_h - 1$, and

$$\begin{aligned}
\mathbf{s}_{L,m}(k) &= [s_m(k) \quad s_m(k-1) \quad \ldots \quad s_m(k-L+1)]^T \\
m &= 1,2,\ldots,M.
\end{aligned}$$

$$\mathbf{H}_{nm} = \begin{bmatrix}
h_{nm,0} & \ldots & h_{nm,L_h-1} & 0 & 0 & \ldots & 0 \\
0 & h_{nm,0} & \ldots & h_{nm,L_h-1} & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \ldots & 0 & h_{nm,0} & \ldots & h_{nm,L_h-1}
\end{bmatrix}$$

Plugging (5) into (4), we find that

$$y_p(k) = \sum_{m=1}^{M} \left[\sum_{n=1}^{N} \mathbf{g}_{pn}^T \mathbf{H}_{nm} \right] \mathbf{s}_{L,m}(k), \quad p = 1, 2, \ldots, P. \tag{6}$$

From the previous expression, we see that in order to perfectly recover $s_p(k)$, the following M conditions have to be satisfied:

$$\sum_{n=1}^{N} \mathbf{H}_{np}^T \mathbf{g}_{pn} = \mathbf{u}_p, \tag{7}$$

$$\sum_{n=1}^{N} \mathbf{H}_{nm}^T \mathbf{g}_{pn} = \mathbf{0}_{L \times 1}, \quad m = 1, 2, \ldots, M, \ m \neq p \tag{8}$$

where

$$\mathbf{u}_p = [0 \ \ \ldots \ \ 0 \ \ 1 \ \ 0 \ \ \ldots \ \ 0]^T$$

is a vector of length L, whose τ_pth component is equal to 1. In matrix/vector form, the M previous conditions are

$$\mathbf{H}^T \mathbf{g}_p = \mathbf{u}_p' \tag{9}$$

where we have the following equations shown at the bottom of the page. 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 $N > M$.

1) 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 (9) is the least-squares (LS) solution, i.e.,

$$\mathbf{g}_p^{\mathrm{LS}} = \left[\mathbf{H}\mathbf{H}^T\right]^{-1} \mathbf{H}\mathbf{u}_p'. \tag{10}$$

However, this solution may not be good enough in practice for several reasons. First, we do not know how to determine L_g. Second, 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. In addition, it does not seem easy to quantify the amount of dereverberation and interference suppression separately.

2) 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

$$L_g \leq \left(\frac{N}{K} - 1\right)(L_h - 1). \tag{11}$$

If we take

$$L_g = \left(\frac{N}{K} - 1\right)(L_h - 1) \tag{12}$$

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

$$\mathbf{g}_p^{\mathrm{MINT}} = \left[\mathbf{H}^T\right]^{-1} \mathbf{u}_p'. \tag{13}$$

This expression is exactly the MINT method [33], [34], 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 $h_{1m}, h_{2m}, \ldots, h_{Nm}$, $m = 1, 2, \ldots, M$, 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.

B. Frost Algorithm

Following (5), if we concatenate the N observation vectors together, we get

$$\begin{aligned} \mathbf{x}(k) &= \left[\mathbf{x}_1^T(k) \ \ \mathbf{x}_2^T(k) \ \ \ldots \ \ \mathbf{x}_N^T(k)\right]^T \\ &= \mathbf{H}\mathbf{s}_{ML}(k) \end{aligned}$$

where

$$\mathbf{s}_{ML}(k) = \left[\mathbf{s}_{L,1}^T(k) \ \ \mathbf{s}_{L,2}^T(k) \ \ \ldots \ \ \mathbf{s}_{L,M}^T(k)\right]^T.$$

The covariance matrix corresponding to $\mathbf{x}(k)$ is

$$\mathbf{R}_{xx} = E\left\{\mathbf{x}(k)\mathbf{x}^T(k)\right\} = \mathbf{H}\mathbf{R}_{ss}\mathbf{H}^T \tag{14}$$

with $\mathbf{R}_{ss} = E\left\{\mathbf{s}_{ML}(k)\mathbf{s}_{ML}^T(k)\right\}$. We assume that $\mathbf{R}_{xx}$ is invertible, which is equivalent to stating that the $\mathbf{R}_{ss}$ matrix is of

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

full rank and $\mathbf{H}^T$ matrix has full column rank. We are now ready to study two interesting cases.

1) 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 linearly constrained minimum variance (LCMV) filter is obtained by solving the following problem:

$$\min_{\mathbf{g}_p} \mathbf{g}_p^T \mathbf{R}_{xx} \mathbf{g}_p \quad \text{subject to} \quad \mathbf{H}_{:p}^T \mathbf{g}_p = \mathbf{u}_p. \tag{15}$$

Hence

$$\mathbf{g}_p^{\mathrm{LCMV1}} = \mathbf{R}_{xx}^{-1} \mathbf{H}_{:p} \left[\mathbf{H}_{:p}^T \mathbf{R}_{xx}^{-1} \mathbf{H}_{:p}\right]^{-1} \mathbf{u}_p. \tag{16}$$

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

$$L_g \geq \frac{L_h - 1}{N - 1}. \tag{17}$$

An important thing to observe is that the minimum length required for the filters $\mathbf{g}_{pn}^{\mathrm{LCMV1}}$, $n = 1, 2, \ldots, N$, 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 (16) becomes

$$\begin{aligned} \mathbf{g}_p^{\mathrm{LCMV1}} &= \left[\mathbf{H}_{:p}^T\right]^{-1} \mathbf{u}_p \\ &= \left[\mathbf{H}_{1p}^T \quad \mathbf{H}_{2p}^T \quad \ldots \quad \mathbf{H}_{Np}^T\right]^{-1} \mathbf{u}_p \end{aligned} \tag{18}$$

which is the MINT method [33], [34]. We assumed in (18) that $\mathbf{H}_{:p}$ has full rank, which is equivalent to saying that the N polynomials formed from $h_{1p}, h_{2p}, \ldots, h_{Np}$ share no common zeros. Mathematically, this condition is expressed as follows:

$$\begin{aligned} &\gcd\left[H_{1p}(z), H_{2p}(z), \ldots, H_{Np}(z)\right] = 1 \\ &\quad \Leftrightarrow \exists G_{p1}(z), G_{p2}(z), \ldots, G_{pN}(z) \\ &\quad : \sum_{n=1}^{N} H_{np}(z) G_{pn}(z) = 1 \end{aligned} \tag{19}$$

where $\gcd[\cdot]$ denotes the greatest common divisor of the polynomials involved and, $H_{np}(z)$ and $G_{pn}(z)$ are the z-transforms of h_{np} and g_{pn}, respectively.

From (14), 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

$$L_g \leq \left(\frac{N}{K} - 1\right)(L_h - 1). \tag{20}$$

Combining (20) and (17), we see how L_g is bounded, i.e.,

$$\frac{L_h - 1}{N - 1} \leq L_g \leq \left(\frac{N}{K} - 1\right)(L_h - 1). \tag{21}$$

2) 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

$$\min_{\mathbf{g}_p} \mathbf{g}_p^T \mathbf{R}_{xx} \mathbf{g}_p \quad \text{subject to} \quad \mathbf{H}^T \mathbf{g}_p = \mathbf{u}_p' \tag{22}$$

we find the solution

$$\mathbf{g}_p^{\mathrm{LCMV2}} = \mathbf{R}_{xx}^{-1} \mathbf{H} \left[\mathbf{H}^T \mathbf{R}_{xx}^{-1} \mathbf{H}\right]^{-1} \mathbf{u}_p'. \tag{23}$$

We refer to this approach as the LCMV2, where we assume 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 (20) 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

$$L_g \geq \left(\frac{N}{K} - 1\right)(L_h - 1). \tag{24}$$

Therefore, the only condition for (23) to exist is that

$$L_g = \left(\frac{N}{K} - 1\right)(L_h - 1) \tag{25}$$

and this value needs to be an integer. In this case, $\mathbf{H}$ is a square matrix and (23) becomes

$$\mathbf{g}_p^{\mathrm{LCMV2}} = \left[\mathbf{H}^T\right]^{-1} \mathbf{u}_p' \tag{26}$$

which is also the MINT solution [33], [34]. Indeed, it was shown in [35] how to convert an $M \times N$ MIMO system (with $M < N$) into M interference-free SIMO systems. The MINT method is then applied in each one of these SIMO systems to remove the channel effect. So this two-step approach is equivalent to the LCMV2.

It's quite remarkable that 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.

C. Generalized Sidelobe Canceller Structure

The generalized sidelobe canceller (GSC) transforms the LCMV algorithm from a constrained problem into an unconstrained form. Therefore, the GSC and LCMV beamformers are essentially the same while the former has some implementation advantages [14], [15], [23], [24]. Given the channel impulse responses, the GSC method can be formulated by dividing the filter vector $\mathbf{g}$ into two components operating on orthogonal subspaces, as illustrated in Fig. 5. Consider the linearly constrained optimization problem given in (15). 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

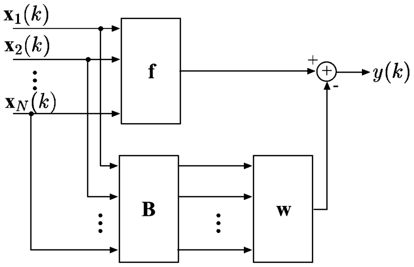


Fig. 5. The structure of a generalized sidelobe canceller.

only for the LCMV1 filter), the GSC method can be formulated as [14], [23], [24]

$$\mathbf{g}_p = \mathbf{f}_p - \mathbf{B}_p\mathbf{w}_p \tag{27}$$

where

$$\mathbf{f}_p = \mathbf{H}_{:p}\left[\mathbf{H}_{:p}^T\mathbf{H}_{:p}\right]^{-1}\mathbf{u}_p \tag{28}$$

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:

$$\min_{\mathbf{w}_p} \left(\mathbf{f}_p - \mathbf{B}_p\mathbf{w}_p\right)^T \mathbf{R}_{xx}\left(\mathbf{f}_p - \mathbf{B}_p\mathbf{w}_p\right) \tag{29}$$

and the solution is

$$\mathbf{w}_p^{\mathrm{GSC}} = \left[\mathbf{B}_p^T\mathbf{R}_{xx}\mathbf{B}_p\right]^{-1}\mathbf{B}_p^T\mathbf{R}_{xx}\mathbf{f}_p. \tag{30}$$

Equation (29) is equivalent to finding a weighting vector $\mathbf{w}_p$ that minimizes $E\left[e_p^2(k)\right]$, where

$$e_p(k) = \mathbf{x}^T(k)\mathbf{f}_p - \mathbf{x}^T(k)\mathbf{B}_p\mathbf{w}_p \tag{31}$$

is the error signal between the outputs of the two filters $\mathbf{f}_p$ and $\mathbf{B}_p\mathbf{w}_p$. In [25], it is shown that

$$\begin{aligned}\mathbf{g}_p^{\mathrm{LCMV1}} &= \mathbf{R}_{xx}^{-1}\mathbf{H}_{:p}\left[\mathbf{H}_{:p}^T\mathbf{R}_{xx}^{-1}\mathbf{H}_{:p}\right]^{-1}\mathbf{u}\\ &= \left\{\mathbf{I} - \mathbf{B}_p\left[\mathbf{B}_p^T\mathbf{R}_{xx}\mathbf{B}_p\right]^{-1}\mathbf{B}_p^T\mathbf{R}_{xx}\right\}\mathbf{f}_p\\ &= \mathbf{g}_p^{\mathrm{GSC}}\end{aligned} \tag{32}$$

so the LCMV and GSC algorithms are equivalent.

Expressions (27) and (32) have a very nice physical interpretation [compared to (16)]. 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} $(n = 1,\ldots,N)$ 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}). In other words, as reverberation of the room increases, interference suppression decreases. This result was shown experimentally in [26] and [27]. One possible way for improvement is to process the observation signals in two steps: the LCMV filter for dereverberation (first step) followed by a Wiener filter for noise reduction (second step); see, for example, the methods used in [28]–[30]. This method may be very effective from a noise reduction point of view but it will distort the desired signal $s_p(k)$.

To find the bounds for the length of $\mathbf{w}_p^{\mathrm{GSC}}$, we consider two situations. The first one is when the number of microphones is equal to the number of sources[1] $(N = M)$. In this case, we know from the previous subsection 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 (21) 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 $\mathbf{w}_p^{\mathrm{GSC}}$

$$\begin{aligned}0 < L_{\mathbf{w}_p} &\le \frac{N}{K}(N-K-1)\\ (L_h-1) &\le \frac{N}{K}(N-K-1)(N-1)L_g.\end{aligned} \tag{33}$$

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 (33) is now

$$0 < L_{\mathbf{w}_p} \le \frac{NQ}{N-Q-1}(L_h-1) \le \frac{N(N-1)Q}{N-Q-1}L_g. \tag{34}$$

We see from (34) that the upper bound of $L_{\mathbf{w}_p}$ depends on three factors: the reverberation condition (L_h), the number of interference sources (Q), and the number of microphones (N). When Q and N are fixed, if the length of the room impulse response L_h increases, this indicates that the environment is more reverberant and the dereverberation problem will become more difficult, so we have to increase $L_{\mathbf{w}_p}$ to compensate for the additional reflections. In case that L_h and N remain the same, but the number of interference sources Q increases, this implies that we have more interferers to cope with so we have to use a larger

[1] There is no distinction here between the interference and desired sources. By extracting the signal of interest $s_p(k)$ from the rest, the algorithm will see the other desired sources as interferences. We assume that all sources are active at the same time; if it is not the case, we will be in a situation where we have more microphones than sources.

$L_{\mathbf{w}_p}$. Now, suppose that L_h and Q remain the same, if we increase the number of microphones, this will allow us to use a larger value for $L_{\mathbf{w}_p}$. We should, however, make a distinction between this case and the former two situations. When we have more microphones, we achieve more realizations of the source signals. So we can increase $L_{\mathbf{w}_p}$ to augment the speech-dereverberation and interference-suppression performance. But in the former two situations, we would expect some degree of performance degradation since the problem becomes more difficult to solve as L_h and Q increase.

D. Minimum Variance Distortionless Response 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:

$$\mathbf{h}_{:p}^T(\kappa_p)\mathbf{g}_p = 1 \tag{35}$$

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:

$$\min_{\mathbf{g}_p} \mathbf{g}_p^T \mathbf{R}_{xx} \mathbf{g}_p \quad \text{subject to} \quad \mathbf{h}_{:p}^T(\kappa_p)\mathbf{g}_p = 1 \tag{36}$$

whose solution is

$$\mathbf{g}_p^{\mathrm{MVDR}} = \frac{\mathbf{R}_{xx}^{-1}\mathbf{h}_{:p}(\kappa_p)}{\mathbf{h}_{:p}^T(\kappa_p)\mathbf{R}_{xx}^{-1}\mathbf{h}_{:p}(\kappa_p)}. \tag{37}$$

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. As L_g is increased compared to κ_p, the signal of interest will be still aligned at the output of the beamformer, while other signals will tend to be attenuated.

This method may be the most useful in practice, since it does not require the full knowledge of the impulse responses but only the relative delays among microphones. However, an adaptive implementation of the MVDR may cancel the desired signal [36].

IV. Experiments

In this section, we will study the effect of filter length on beamforming performance and compare different algorithms via simulations in realistic acoustic environments.

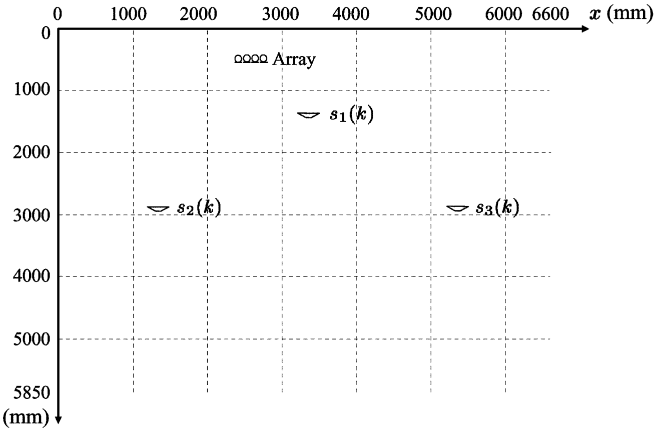


Fig. 6. Layout of the experimental setup in the varechoic chamber (coordinate values measured in millimeters). The three sources are placed, respectively, at (3337, 1438, 1600), (1337, 2938, 1600), and (5337, 2938, 1600). The four microphones in the linear array are located, respectively, at (2437, 5600, 1400), (2537, 5600, 1400), (2637, 5600, 1400), and (2737, 5600, 1400).

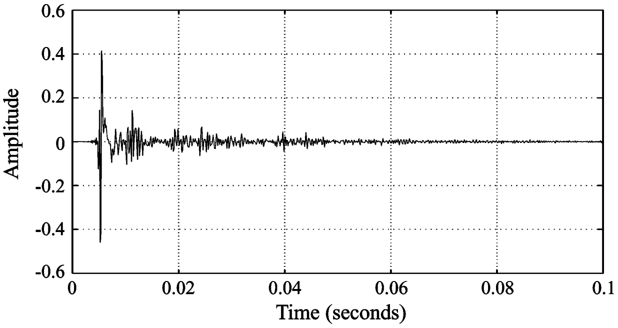


Fig. 7. One measured impulse response from $s_1(k)$ to microphone 1.

A. Experimental Setup

The experiments were conducted with the acoustic impulse responses measured in the varechoic chamber at Bell Laboratories [37]. The chamber is a rectangular room, which measures 6.7 m long by 6.1 m wide by 2.9 m high $(x \times y \times z)$ and is equipped with 368 electronically controlled panels. Each panel consists of two perforated sheets whose holes, if aligned, expose sound absorbing material (fiberglass) behind, but if shifted to misalign, form a highly reflective surface. Every panel can be controlled individually so that the holes on a particular panel are either fully open (absorbing state) or fully closed (reflective state). As a result, a total of 2^{368} different room characteristics can be generated by varying the binary states of the 368 panels in different combination [38]. For a detailed description about the varechoic chamber and how the reverberation time is controlled, see [37] and [38].

The layout of the experimental setup is illustrated in Fig. 6, where a linear array which consists of four omni-directional microphones were employed with their positions being, respectively, at (2437, 5600, 1400), (2537, 5600, 1400), (2637, 5600, 1400), and (2737, 5600, 1400). We have three sources in the sound field: one target $s_1(k)$ (a signal from a male speaker) is located at (3337, 1438, 1600), and two interferers (two signals from a female speaker), $s_2(k)$ and $s_3(k)$, are placed at (1337, 2938, 1600) and (5337, 2938, 1600), respectively. The objective of this study is to investigate how the desired signal $s_1(k)$ can be dereverberated and the two interference sources, $s_2(k)$ and $s_3(k)$, can be suppressed or canceled when four microphones

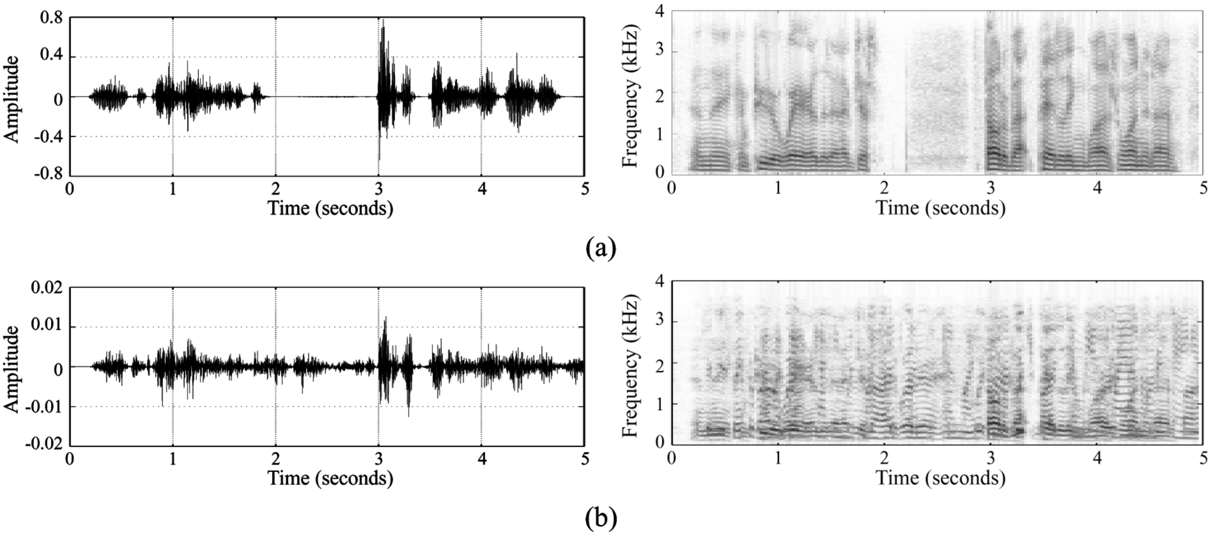


Fig. 8. Time sequence and the corresponding spectrogram of (a) the desired source signal $s_1(k)$ (from a male speaker) and (b) the output of microphone 1, i.e., $x_1(k)$.

are used. For ease of analysis, we neglect the ambient noise effect. The reverberation is controlled such that T_{60} is approximately equal to 0.35 s. To make the experiments repeatable, the impulse response from each source to each microphone was measured (the impulse response was first measured at 48 kHz and then downsampled to 8 kHz). As an example, Fig. 7 plots an impulse response measured from $s_1(k)$ to the first microphone. These measured impulse responses will be treated as the true ones in our experiments.

B. Experimental Results

To visualize the performance of different beamforming algorithms, we first conduct a simple experiment where all the impulse responses are truncated to only 64 points (the zeros shared by all the impulse responses at the beginning are also removed). All three source signals are prerecorded speech sampled at 8 kHz, where $s_1(k)$ is from a male speaker and both $s_2(k)$ and $s_3(k)$ are from the same female speaker. The waveform and spectrogram of the first 5 s of $s_1(k)$ are shown in Fig. 8(a). The microphone outputs are obtained by convolving the three source signals with the corresponding impulse responses. Fig. 8(b) plots the first 5 s of the signal observed at the first microphone.

To extract $s_1(k)$, we need to estimate the $\mathbf{g}_1$ filter. This would require knowledge about the impulse responses from the three sources to the four microphones. In this experiment, we assume that the impulse responses are known *a priori*, so the results in this case demonstrate the upper limit of each algorithm for a given condition. Another parameter that has to be determined is the length of the $\mathbf{g}_1$ filter, i.e., L_g. Throughout the text, we have analyzed the bounds of L_g for different algorithms. In this experiment, L_g is chosen as its maximum value that can be taken according to (12), (20), (25), and (33) and is set to be the same for all the algorithms. Note that with this optimum choice of L_g, the LS and LCMV2 methods will produce the same results because under this condition the pseudoinverse in the LS method is equal to the exact inverse in the LCMV2 approach. In addition, we already see from Section III that LCMV2 and MINT are the same. The outputs of different beamformers are plotted in Fig. 9.

Comparing Figs. 9 and 8 reveals that both the LS and LCMV2 (MINT) approaches have achieved almost perfect interference suppression and speech dereverberation. However, the outputs of LCMV1 and GSC still consist of a small amount of interference signals. Apparently, LCMV1 and GSC are less effective than the LS and LCMV2 (MINT) techniques in terms of interference suppression. This is understandable since LCMV1 and GSC employ only the channel information from the desired source to the microphones while both the LS and LCMV2 (MINT) techniques use not only the impulse responses from the desired source but also those from all the interferers. In addition, we see that MVDR is inferior to all the other studied techniques in performance. This result is not surprising since MVDR poses less constraints than the other techniques.

To quantitatively assess the performance of interference suppression and speech dereverberation, we now evaluate two criteria, namely signal-to-interference ratio (SIR) and speech spectral distortion. For the notion of SIR, see [35], [39]. In this study, though we have M sources, our interest is in extracting only the target signal, i.e., the first source $s_1(k)$, so the average input SIR at microphone n is defined as

$$\mathrm{SIR}_n^{\mathrm{in}} \triangleq \frac{E\left\{[h_{n1} * s_1(k)]^2\right\}}{\sum_{m=2}^{M} E\left\{[h_{nm} * s_m(k)]^2\right\}}, \quad n = 1, 2, \ldots, N. \tag{38}$$

The overall average input SIR is then given by

$$\mathrm{SIR}^{\mathrm{in}} \triangleq \frac{1}{N} \sum_{n=1}^{N} \mathrm{SIR}_n^{\mathrm{in}}. \tag{39}$$

The output SIR is defined using the same principle but the expression will be slightly more complicated. For a concise presentation, we denote the impulse response of the equivalent channel between the mth source and the beamforming output

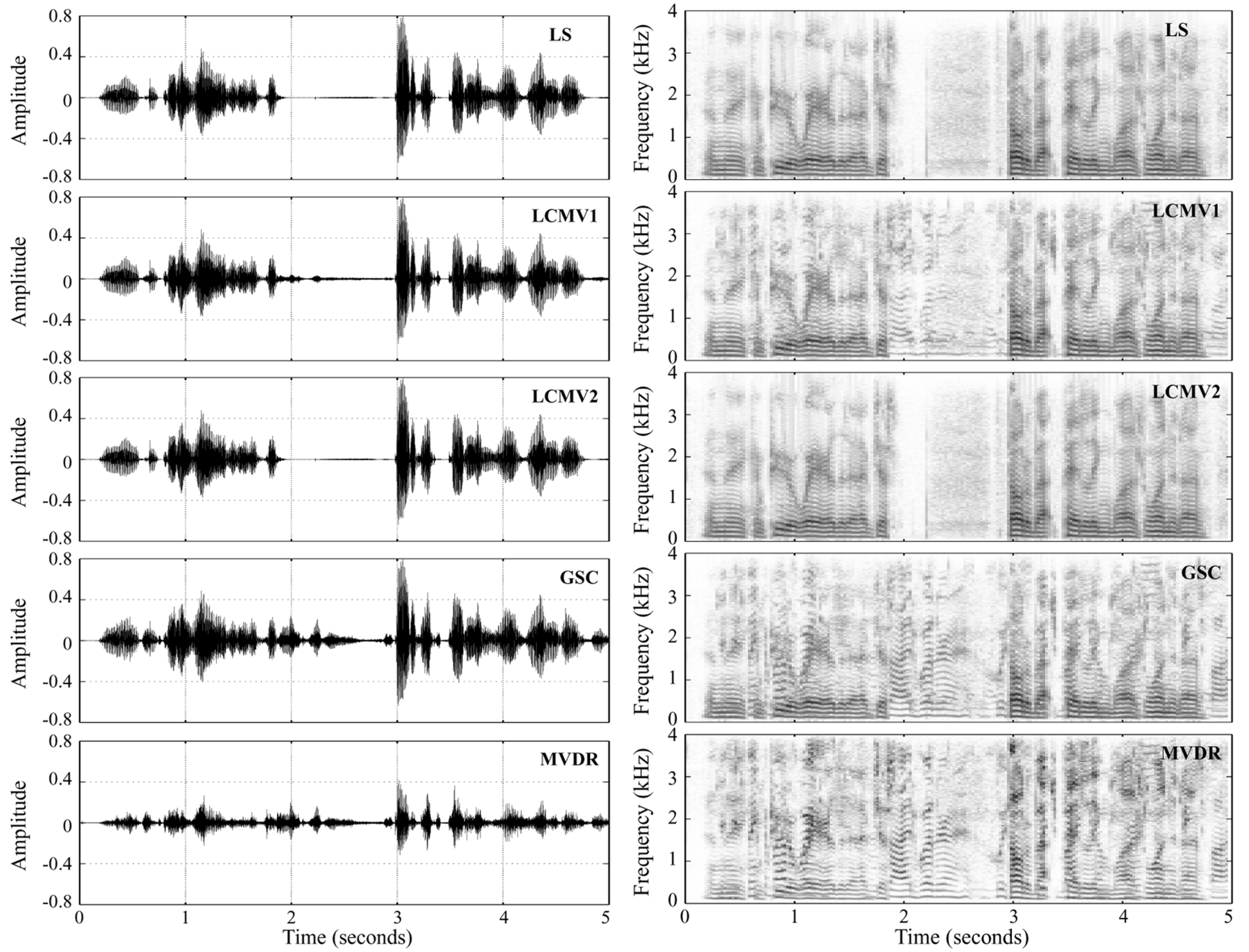


Fig. 9. Time sequence and the corresponding spectrogram of different beamforming algorithms, where $L_h = 64$ and $L_g = 189$ for all the algorithms. Note that under this condition, the LS, LCMV2, and MINT methods are theoretically the same.

as ϕ_m, which can be expressed as

$$\phi_m = \sum_{n=1}^{N} g_{1n} * h_{nm} \tag{40}$$

where g_{1n} is the filter between microphone n and the beamforming output, and h_{nm} is the impulse response between source m and microphone n. The output SIR can then be written as

$$\mathrm{SIR}^{\mathrm{out}} \triangleq \frac{E\left\{[\phi_1 * s_1(k)]^2\right\}}{\sum_{m=2}^{M} E\left\{[\phi_m * s_m(k)]^2\right\}}. \tag{41}$$

If we express both $\mathrm{SIR}^{\mathrm{out}}$ and $\mathrm{SIR}^{\mathrm{in}}$ in decibels, the difference between the two reflects the performance of interference suppression.

To evaluate speech dereverberation, we investigate the Itakura–Saito (IS) distortion measure, which performs a comparison of spectral envelopes [autoregressive (AR) parameters] between the clean and the processed speech. For a detailed description of the IS distance, we refer to [40], [41]. Studies have shown that the IS measure is highly correlated (0.59) with subjective quality judgements [42]. A recent report reveals that the difference in mean opinion score (MOS) between two processed speech signals would be less than 1.6 if their IS measure is less than 0.5 for various codecs [43]. Many other reported experiments confirmed that two spectra would be perceptually nearly identical if their IS distance is less than 0.1. All this evidence indicates that the IS distance is a reasonably good objective measure of speech quality. In our experiment, the IS measure is calculated between $s_1(k)$ and $s_1(k) * \phi_m$; therefore, it evaluates the amount of reverberation present in the estimated speech signal after beamforming. The smaller the IS distance, the more effective will be the beamforming algorithm in dereverberation.

Table I summarizes the experimental results, where the source signals are the same as used in the previous experiment. Many observations can be made from this table. First of all, 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. For example, when $L_h = 64$, the LCMV1 algorithm improves SIR from $-$ 9.23 dB

TABLE I
PERFORMANCE OF INTERFERENCE SUPPRESSION AND SPEECH DEREVERBERATION USING DIFFERENT BEAMFORMING ALGORITHMS WHERE THE MIMO IMPULSE RESPONSES ARE KNOWN *A PRIORI*

			LS		LCMV1		LCMV2/MINT		GSC		MVDR	
$\mathrm{SIR}^{\mathrm{in}}$ (dB)	L_h	L_g	$\mathrm{SIR}^{\mathrm{out}}$ (dB)	IS	$\mathrm{SIR}^{\mathrm{out}}$ (dB)	IS	$\mathrm{SIR}^{\mathrm{out}}$ (dB)	IS	$\mathrm{SIR}^{\mathrm{out}}$ (dB)	IS	$\mathrm{SIR}^{\mathrm{out}}$ (dB)	IS
-9.23	64	189*	187.63	0.00	17.99	0.00	187.63	0.00	14.45	0.00	4.76	6.28
		150	9.28	0.02	9.10	0.00	×	×	9.10	0.00	4.28	6.65
		100	7.20	0.08	-0.46	0.00	×	×	-0.45	0.00	3.4	7.86
		50	4.52	0.13	-8.00	0.00	×	×	-8.00	0.00	2.66	8.07
-8.04	128	381*	171.26	0.00	9.58	0.00	171.26	0.00	4.09	0.00	4.20	6.86
		360	24.65	0.01	3.93	0.00	×	×	3.93	0.00	4.20	6.86
		320	14.27	0.01	2.81	0.00	×	×	2.81	0.00	4.22	6.75
		200	3.82	0.13	-3.85	0.00	×	×	-3.85	0.00	3.32	7.22
-8.25	256	765*	117.24	0.00	7.90	0.00	117.24	0.00	1.51	0.00	4.37	7.68
		700	24.77	0.03	1.26	0.00	×	×	1.26	0.00	4.40	7.56
		600	11.24	0.23	0.12	0.00	×	×	0.12	0.00	4.46	7.38
		300	4.0	0.15	-6.74	0.00	×	×	-6.74	0.00	3.00	9.07

NOTES: *: the maximum value that the L_g can take for the condition;
×: the L_g cannot take this value for the method in the given condition.

TABLE II
PERFORMANCE OF INTERFERENCE SUPPRESSION AND SPEECH DEREVERBERATION USING DIFFERENT BEAMFORMING ALGORITHMS WHERE THE CHANNEL IMPULSE RESPONSES ARE ESTIMATED USING A BLIND TECHNIQUE

				LS		LCMV1		LCMV2/MINT		GSC		MVDR	
$\mathrm{SIR}^{\mathrm{in}}$ (dB)	L_h	$L_{\hat{h}}$	L_g	$\mathrm{SIR}^{\mathrm{out}}$ (dB)	IS	$\mathrm{SIR}^{\mathrm{out}}$ (dB)	IS	$\mathrm{SIR}^{\mathrm{out}}$ (dB)	IS	$\mathrm{SIR}^{\mathrm{out}}$ (dB)	IS	$\mathrm{SIR}^{\mathrm{out}}$ (dB)	IS
-9.23	64	64	189	140.1	0.00	14.45	0.00	140.1	0.00	14.45	0.00	4.76	6.28
		50	147	-5.93	6.10	8.31	0.57	×	×	8.96	0.46	4.34	7.04
-8.04	128	128	381	133.04	0.0	4.09	0.00	133.04	0.00	4.09	0.00	4.20	6.86
		100	297	-4.73	5.85	4.87	0.85	×	×	3.63	0.90	4.11	7.06

NOTES: L_h: the length of true impulse responses;
$L_{\hat{h}}$: the length of the channel impulse responses used during blind channel identification.

(input SIR) to 17.99 dB (output SIR). The gain is approximately 27 dB. When L_h is increased to 256, the maximum SIR gain with the same technique is only 16 dB. This result should not come as surprise. As L_h increases, each microphone receives more reflections (with longer delays) from both the desired and interference sources. As a result, the received speech becomes more distorted and the estimation problem becomes more difficult. Second, in the ideal condition where impulse responses are known and L_g is set to its maximum value, both the LS and LCMV2 (MINT) techniques can achieve almost perfect interference suppression and speech dereverberation. The SIR gains are more than 100 dB and the IS distances are approximately zero. Similar to the LS and LCMV2 (MINT) methods, the LCMV1 and GSC can also perform perfect speech dereverberation, but their interference suppression performance is limited. The reason behind this has been explained earlier on. Briefly, it is because LCMV1 and GSC did not use the channel information from the interferers to the microphones. Third, 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 for all the methods except MVDR. For example, for $L_h = 64$, when $L_g = 189$, the LCMV1 yields an output SIR of 17.99 dB. The corresponding SIR gain is 27 dB. But when L_g is reduced to 50, the output SIR is only -8 dB, and the SIR gain is only 1 dB. Therefore, if we want a reasonable amount of interference suppression, the filter g_1 should be set to a long enough value. However, the length of this filter is upper bounded, as we explained in Section III. In addition, we see from Table I that IS distances obtained by the LS, LCMV1, LCMV2 (MINT), and GSC methods are close to zero, indicating that these techniques have accomplished good speech dereverberation. This coincides with the theoretical analysis made throughout the text. Finally, it is remarkable to see that, in terms of interference rejection, the MVDR method is very robust to the changes of both L_h and L_g. When L_g is small, this method can even achieve more interference suppression than the other four approaches. However, the IS measures with this method are very large. Therefore, we may have to use dereverberation techniques in order to further reduce speech distortion.

In the previous experiments, we assumed that the impulse responses were known *a priori*. In real applications, it is very difficult if not impossible to know the true impulse responses. Therefore, we have to estimate such information based on the data observed at microphones. In our application scenario, the source signals are generally not accessible, so the estimation of channel impulse responses have to be done in a blind manner. However, blind identification of a MIMO system is a very difficult problem and no effective solution is available thus far, particularly for acoustic applications. Fortunately, in natural communication environments, not all the sources are active at the same time. In many time periods, the observation signal is occupied by a single source exclusively. If we can detect those periods, the MIMO identification problem can be converted to SIMO identification problem in each time period. This is assumed to be the case in our study and the channel impulse responses are estimated using the techniques developed in [35]. After the estimation of channel impulse responses, we can recover the desired source signals by beamforming. The results for this experiment are shown in Table II where we studied two situations. While in the first one, we assume that we know the length of the true impulse responses during blind channel identification, in the second case, the length of the modelling filter, i.e., $L_{\hat{h}}$, during blind channel identification is set to less than L_h. Evidently, the second case is more realistic since in reality the real impulse responses can be very long, but we cannot use a very long modeling filter due to many practical limitations.

Comparing Tables II and I, one can see when $L_{\hat{h}} = L_h$, all the techniques suffer some but not significant performance degradation. However, if $L_{\hat{h}}$ is less than L_h, which is true in most real applications, the LS and LCMV2 (MINT) suffer significant performance degradation in both interference suppression and speech dereverberation. The reason may be explained as follows. In our case, we truncated the impulse response to either 64 or 128 points. Due to the strong reverberation, the tail of the truncated impulse responses consists of significant energy. As a result, dramatic errors were introduced during channel identification when decreasing $L_{\hat{h}}$. This in turn degrades the performance of beamforming. However, comparing with the LS and LCMV2 (MINT), we see that the LCMV1 and GSC suffer some but not serious deterioration. We also noticed a very interesting property of the MVDR approach from Table II that its performance does not deteriorate much as $L_{\hat{h}}$ decreases. This robust feature is due to the fact that MVDR poses less constraints than the other studied methods. However, as we noticed before, MVDR suffers dramatic signal distortion, as indicated by its large IS distances. So further dereverberation techniques may have to be considered after the MVDR processing.

V. Conclusion

Microphone array beamforming is a very challenging problem. Most existing algorithms exhibit very limited performance in real acoustic environments. The reasons behind this are multiple. The main one is due to the reverberation effect, which has not been fully taken into account in the current techniques. In this paper, we developed a general framework for microphone array beamforming, in which beamforming is treated as a MIMO signal processing problem. Under this general framework, we analyzed the lower and upper bounds for the length of the beamforming filter, which in turn show the performance bounds of beamforming in terms of speech dereverberation and interference suppression. We addressed the connection between beamforming and the 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 are for the different beamforming techniques to work. As expected, both the theoretical analysis and experimental results showed that the impulse responses from the desired sources as well as the interferers have to be employed in order to achieve good interference suppression and speech dereverberation. In practice, however, the true impulse responses are in general not accessible. Therefore, we have to estimate them via either blind or nonblind techniques. If the estimated channel impulse responses approximate the true ones, beamforming may not suffer much performance degradation. However, if the channel estimates are inaccurate, the performance of interference suppression and speech dereverberation may deteriorate dramatically. The degree of performance degradation varies from algorithm to algorithm. Therefore, great care has to be taken when we select beamforming algorithms.

References

[1] M. Branstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer, 2001.

[2] J. Benesty and Y. Huang, Eds., *Adaptive Signal Processing: Applications to Real-World Problems*. Berlin, Germany: Springer, 2003.

[3] S. A. Schelkunoff, "A mathematical theory of linear arrays," *Bell Syst. Tech. J.*, vol. 22, pp. 80–107, Jan. 1943.

[4] B. Widrow, P. Mantey, L. Griffiths, and B. Goode, "Adaptive antenna systems," *Proc. IEEE*, vol. 55, no. 12, pp. 2143–2159, Dec. 1967.

[5] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[6] W. F. Gabriel, "Adaptive arrays—An introduction," *Proc. IEEE*, vol. 64, no. 2, pp. 239–272, Feb. 1976.

[7] R. A. Monzingo and T. W. Miller, *Introduction to Adaptive Arrays*. New York: Wiley, 1980.

[8] T. J. Shan and T. Kailath, "Adaptive beamforming for coherence signals and interference," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 3, pp. 527–536, Jun. 1985.

[9] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-35, no. 10, pp. 1365–1376, Oct. 1987.

[10] D. B. Ward, R. C. Williamson, and R. A. Kennedy, "Broadband microphone arrays for speech acquisition," *Acoust. Australia*, vol. 26, pp. 17–20, Apr. 1998.

[11] J. L. Flanagan, D. A. Berkeley, G. W. Elko, J. E. West, and M. M. Sondhi, "Autodirective microphone systems," *Acustica*, vol. 73, pp. 58–71, Feb. 1991.

[12] W. Kellermann, "A self-steering digital microphone array," in *Proc. IEEE ICASSP*, 1991, vol. 5, pp. 3581–3584.

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

[14] C. W. Jim, "A comparison of two LMS constrained optimal array structures," *Proc. IEEE*, vol. 65, no. 12, pp. 1730–1731, Dec. 1977.

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

[16] J. L. Flanagan, J. D. Johnson, R. Zahn, and G. W. Elko, "Computer-steered microhone arrays for sound transduction in large rooms," *J. Acoust. Soc. Amer.*, vol. 75, pp. 1508–1518, Nov. 1985.

[17] M. M. Sondhi and G. W. Elko, "Adaptive optimization of microphone arrays under a nonlinear constriant," in *Proc. IEEE ICASSP*, 1986, pp. 19.9.1–19.9.4.

[18] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[19] S. Affes, S. Gazor, and Y. Grenier, "Robust adaptive beamforming via LMS-like target tracking," in *Proc. IEEE ICASSP*, Apr. 1994, pp. 19–22.

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

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

[22] W. Herbordt and W. Kellermann, "Adaptive beamforming for audio signal acquisition," in *Adaptive Signal Processing: Applications to Real-World Problems*, J. Benesty and Y. Huang, Eds. Berlin, Germany: Springer, 2003, ch. 6, pp. 155–194.

[23] K. M. Buckley, "Broad-band beamforming and the generalized sidelobe canceller," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 5, pp. 1322–1323, Oct. 1986.

[24] S. Werner, J. A. Apolinário, and M. L. R. de Campos, "On the equivalence of RLS implementations of LCMV and GSC processors," *IEEE Signal Process. Lett.*, vol. 10, no. 12, pp. 356–359, Dec. 2003.

[25] B. R. Breed and J. Strauss, "A short proof of the equivalence of LCMV and GSC beamforming," *IEEE Signal Process. Lett.*, vol. 9, no. 6, pp. 168–169, Jun. 2004.

[26] J. E. Greenberg and P. M. Zurek, "Adaptive beamformer performance in reverberation," in *Proc. IEEE ASSP Workshop on Applications Signal Process. Audio Acoust.*, 1991, pp. 101–102.

[27] J. Bitzer, K. U. Simmer, and K.-D. Kammeyer, "Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement," in *Proc. IEEE ICASSP*, 1999, pp. 2965–2968.

[28] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE ICASSP*, 1988, pp. 2578–2581.

[29] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. Speech Audio Process.*, vol. 6, pp. 240–259, May 1998.

[30] I. Cohen, "Analysis of two-channel generalized sidelobe canceller (GSC) with post-filtering," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 684–699, Nov. 2003.

[31] J. Benesty, S. Makino, and J. E. Chen*, Speech Enhancement*. Berlin, Germany: Springer, 2005.

[32] Y. Huang, J. Benesty, and J. Chen*, Acoustic MIMO Signal Process.*. Boston, MA: Springer, 2006.

[33] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[34] F. Furuya, "Noise reduction and dereverberation using correlation matrix based on the multiple-input/output inverse-filtering theorem (MINT)," in *Proc. Int. Workshop on Hands-Free Speech Commun.*, 2001, pp. 59–62.

[35] 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, no. 5, pp. 882–895, Sep. 2005.

[36] M. Buck, T. Haulick, and H.-J. Pfleiderer, "Self-calibrating microphone arrays for speech signal acquisition: A systematic approach," *Signal Process.*, to be published.

[37] A. Härmä, "Acoustic measurement data from the varechoic chamber," Agere Systems, Nov. 2001, Tech. Memo.

[38] W. C. Ward, G. W. Elko, R. A. Kubli, and W. C. McDougld, "The new varechoic chamber at AT&T Bell Labs," in *Proc. Wallance Clement Sabine Centennial Symp.*, 1994, pp. 343–346.

[39] M. Z. Ikram and D. R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environments," in *Proc IEEE ICASSP*, 2000, pp. 1041–1044.

[40] F. Itakura and S. Saito, "A statistical method for esimation fo speech spectral density and formant frequencies," *Electron. Commun. Jpn.*, vol. 53A, pp. 36–43, 1970.

[41] L. R. Rabiner and B. H. Juang*, Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[42] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements*, Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[43] G. Chen, S. N. Koh, and I. Y. Soon, "Enhanced Itakura measure incorporating masking properties of human auditory system," *Signal Process.*, vol. 83, pp. 1445–1456, Jul. 2003.

Jacob Benesty (M'92–SM'04) was born in Marrakech, Morocco, in 1963. He received the M.S. degree in microwaves from Pierre and Marie Curie University, Paris, France, in 1987 and the Ph.D. degree in control and signal processing from Orsay University, Paris, in April 1991. During his Ph.D. program (from November 1989 to April 1991), he worked on adaptive filters and fast algorithms at the Centre National d'Etudes des Télécommunications (CNET), Paris.

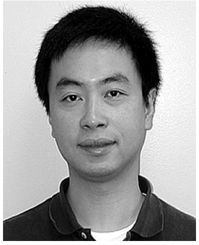
From January 1994 to July 1995, he worked at Telecom Paris on multichannel adaptive filters and acoustic echo cancellation. From October 1995 to May 2003, he was first a Consultant and then a Member of the Technical Staff at Bell Laboratories, Murray Hill, NJ. In May 2003, he joined the Université du Québec INRS-EMT, Montréal, QC, Canada, as an Associate Professor. His research interests are in acoustic signal processing and multimedia communications. He coauthored the books *Advances in Network and Acoustic Echo Cancellation* (Berlin, Germany: Springer-Verlag, 2001) and *Acoustic MIMO Signal Processing* (Berlin, Germany: Springer-Verlag, 2006). He is also a coeditor/coauthor of the books *Speech Enhancement* (Berlin, Germany: Springer-Verlag, 2005), *Audio Signal Processing for Next-Generation Multimedia Communication Systems* (Boston, MA: Kluwer, 2004), *Adaptive Signal Processing: Applications to Real-World Problems* (Berlin, Germany: Springer-Verlag, 2003), and *Acoustic Signal Processing for Telecommunication* (Boston, MA: Kluwer, 2000).

Dr. Benesty received the 2001 Best Paper Award from the IEEE Signal Processing Society. He is a member of the editorial board of the *EURASIP Journal on Applied Signal Processing*. He was the Co-Chair of the 1999 International Workshop on Acoustic Echo and Noise Control.

Jingdong Chen (M'99) received the B.S. degree in electrical engineering and the M.S. degree in array signal processing from the Northwestern Polytechnic University, Xiaan, China, in 1993 and 1995, respectively, and the Ph.D. degree in pattern recognition and intelligence control from the Chinese Academy of Sciences, Beijing, in 1998. His Ph.D. research focused on speech recognition in noisy environments. He studied and proposed several techniques covering speech enhancement and HMM adaptation by signal transformation.

From 1998 to 1999, he was with ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan, where he conducted research on speech synthesis, speech analysis, as well as objective measurements for evaluating speech synthesis. He then joined the Griffith University, Brisbane, Australia, as a Research Fellow, where he engaged in research in robust speech recognition, signal processing, and discriminative feature representation. From 2000 to 2001, he was with the ATR Spoken Language Translation Research Laboratories, Kyoto, where he conducted research in robust speech recognition and speech enhancement. He joined Bell Laboratories, Murray Hill, NJ, as a Member of Technical Staff in July 2001. His current research interests include adaptive signal processing, speech enhancement, adaptive noise/echo cancellation, microphone array signal processing, signal separation, and source localization. He is a coauthor of the book *Acoustic MIMO Signal Processing* (Berlin, Germany: Springer-Verlag, 2006). He is also coeditor/coauthor of the book *Speech Enhancement* (Berlin, Germany: Springer-Verlag, 2005).

Dr. Chen is the recipient of 1998–1999 research grant from the Japan Key Technology Center and the 1996–1998 President's Award from the Chinese Academy of Sciences.



Yiteng (Arden) Huang (S'97–M'01) received the B.S. degree from the Tsinghua University, Beijing, China, in 1994, the M.S. and Ph.D. degrees from the Georgia Institute of Technology (Georgia Tech), Atlanta, in 1998 and 2001, respectively, all in electrical and computer engineering.

During his doctoral studies from 1998 to 2001, he was a Research Assistant with the Center of Signal and Image Processing, Georgia Tech, and was a Teaching Assistant with the School of Electrical and Computer Engineering, Georgia Tech. In the summers from 1998 to 2000, he worked with Bell Laboratories, Murray Hill, NJ and engaged in research on passive acoustic source localization with microphone arrays. Upon graduation, he joined Bell Laboratories as a Member of Technical Staff in March 2001. His current research interests are in multichannel acoustic signal processing, multimedia, and wireless communications. He is a coauthor of the book *Acoustic MIMO Signal Processing* (Berlin, Germany: Springer-Verlag, 2006). He is also coeditor/coauthor of the books *Audio Signal Processing for Next-Generation Multimedia Communication Systems* (Boston, MA: Kluwer, 2004) and *Adaptive Signal Processing: Applications to Real-World Problems* (Berlin, Germany: Springer-Verlag, 2003).

Dr. Huang is currently an Associated Editor of the IEEE SIGNAL PROCESSING LETTERS. He received the 2002 Young Author Best Paper Award from the IEEE Signal Processing Society, the 2000–2001 Outstanding Graduate Teaching Assistant Award from the School Electrical and Computer Engineering, Georgia Tech, the 2000 Outstanding Research Award from the Center of Signal and Image Processing, Georgia Tech, and the 1997–1998 Colonel Oscar P. Cleaver Outstanding Graduate Student Award from the School of Electrical and Computer Engineering, Georgia Tech.

Jacek Dmochowski was born in Gdansk, Poland, in December 1979. He received the B.Eng. degree in communications engineering and the M.A.Sc. degree in electrical engineering from Carleton University, Ottawa, ON, Canada, in 2003 and 2005, respectively.

His research interests are in the area of multichannel digital signal processing and include microphone array beamforming and source localization.

Mr. Dmochowski is the recipient of the Ontario Graduate Scholarship (2004–2006), and the National Science and Engineering Research Council Postgraduate Award at the Doctoral Level (2005–2007).