# A Blind Channel Identification-Based Two-Stage Approach to Separation and Dereverberation of Speech Signals in a Reverberant Environment

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

***Abstract*—Blind separation of independent speech sources from their convolutive mixtures in a reverberant acoustic environment is a difficult problem and the state-of-the-art blind source separation techniques are still unsatisfactory. The challenge lies in the coexistence of spatial interference from competing sources and temporal echoes due to room reverberation in the observed mixtures. Focusing only on optimizing the signal-to-interference ratio is inadequate for most if not all speech processing systems. In this paper, we deduce that spatial interference and temporal echoes can be separated and an $M \times N$ MIMO system will be converted into $M$ SIMO systems that are free of spatial interference. Furthermore we show that the channel matrices of these SIMO systems are irreducible if the channels from the same source in the MIMO system do not share common zeros. Thereafter we can apply the Bezout theorem to remove reverberation in those SIMO systems. Such a two-stage procedure leads to a novel sequential source separation and speech dereverberation algorithm based on blind multichannel identification. Simulations with measurements obtained in the varechoic chamber at Bell Labs demonstrate the success and robustness of the proposed algorithm in highly reverberant acoustic environments.**

***Index Terms*—Bezout theorem, blind channel identification (BCI), blind source separation (BSS), independent component analysis (ICA), multiple-input multiple-output (MIMO) systems, single-input multiple-output (SIMO) systems, speech dereverberation.**

## I. Introduction

SOURCE separation techniques aim to extract independent signals from their linear mixtures captured by a number of sensors. In many cases, *a priori* knowledge about the characteristics of the source signals and the way in which they are mixed together is either inaccessible or very expensive to acquire. Consequently, the separation is carried out only on the basis of the mixtures with the assumption of mutual statistical independence among the source signals and is hence called a "blind" method. The task of blind source separation (BSS) is typically accomplished by independent component analysis (ICA) algorithms that assume mutually independent source signals. However, source signals distorted by arbitrary filters still are independent of each other. Thereafter deconvolution needs to be performed to mitigate the linear distortion and reconstruct the involved source signals. Recently blind source separation and deconvolution has become an increasingly active area of research because of a variety of its applications, e.g., biomedical signal analysis and processing [1], image enhancement [2], acoustic and speech processing [3], multiple-antenna wireless communications [4], etc.

In the BSS problem, the mixing procedure is generally delineated with a multiple-input multiple-output (MIMO) mathematical model. Such a model is either memoryless or with memory, being referred to as instantaneous and convolutive mixtures, respectively. The former was predominantly the focus of early work on BSS for its relative simplicity [5], [6]. But convolutive mixtures are more realistic and recently have gained much more attention [7]. A prevailing approach is to transform a computationally intensive convolutive BSS problem in the time domain into multiple independent instantaneous BSS problems in the frequency domain [8]. However, a fundamental problem of permutation ambiguity arises in frequency-domain BSS algorithms for convolutive mixtures and limits their separation performance [9]. This problem is less prominent when the mixing channels have only few taps in their impulse responses as encountered in wireless communications. But in a reverberant acoustic environment, the length of the mixing channels can be very long (filter lengths in thousands of taps are not uncommon) and solving the permutation ambiguity problem is very challenging [10]. In this paper, we will examine the problem of blind separation and dereverberation of speech signals in a reverberant environment from a different perspective and propose a blind channel identification (BCI)-based two-stage algorithm.

Separating independent, competing speech signals in a reverberant environment is well known as the *cocktail party phenomenon*. Although research in cognitive psychology is yet to produce thorough understanding about how humans concentrate their attention on a speaker of interest in a noisy cocktail party and block out other interfering conversations in the room, traditional BSS algorithms treat the MIMO acoustic system as a *black box* and are determined to recover the original speech source signals with no intention to shed light on the inside of the box. As a result, such characteristics of the room acoustics as the locations of independent speech sources are not explicitly provided in the solutions of traditional BSS algorithms. For each speech source, the solution is only a *monaural* signal. Recently, the need for attaining the spatial

Manuscript received April 12, 2004; revised September 8, 2004. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Rainer Martin.

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

J. Benesty is with the Université du Québec, Montréal, QC, H5A 1K6, Canada (e-mail: benesty@emt.inrs.ca).

Digital Object Identifier 10.1109/TSA.2005.851941

perceptibility of separated speech signals has emerged in stereo or multichannel speech processing systems and pleasingly efforts have been made to meet it [11]. In this so-called single-input multiple-output (SIMO) based BSS algorithm, a number of independent component analyzers are constructed to estimate distinct source observations corresponding to different microphones. It intends to separate speech components of the mixture at each microphone. Therefore the solution for each source is a set of the SIMO outputs. This work is interesting and inspirational. However, one can easily determine the component of a microphone signal corresponding to a specified speech source after its monaural signal has been successfully separated from the mixtures. It is not clear that the SIMO-based BSS algorithm would be more attractive for producing better voice quality, not even to mention the overwhelming amount of computational complexity that it further causes to prevent all ICA's from adapting in the same manner. Therefore, we attempt to take a different strategy to tackle this problem. Instead of estimating the source speech signals directly, we would like to blindly identify the unknown MIMO system first, and then extract the desired speech signals with perfect separation and dereverberation. Since the MIMO system is decomposed into a number of SIMO systems which will be blindly identified at different time, the proposed source separation algorithm will not have the annoying permutation ambiguity problem.

In a MIMO acoustic system, the speech mixtures contain both speech echoes due to reverberation by room surfaces and interference from other co-existing sources. To recover the source signals, not only interference but also echoes need to be removed. In this paper, we will show that echoes and interference can be completely separated by converting an $M \times N$ MIMO system into $M$ interference-free SIMO systems. The channel matrices of these SIMO systems will be irreducible if the channels from the same input in the MIMO system do not share common zeros. For irreducible SIMO systems, dereverberation can be performed by using the Bezout theorem. If co-prime channels are not true for all inputs, we will deduce what is the best possible solution for just partial dereverberation. This discussion leads to the proposal of a sequential source separation and speech dereverberation algorithm based on blind multichannel identification. Simulation results show that this algorithm performs well at low noise levels (for achieving a reliable estimation of channel impulse responses with blind channel identification algorithms) with high signal-to-interference ratio (SIR) and low speech distortion. The idea of separating spatial interference and temporal echoes was first proposed by the authors in an earlier study about MIMO equalization for wireless communications [12]. In this paper, we will see that it can be successfully applied in acoustic environments.

This paper is organized as follows. Section II delineates the MIMO signal model and briefly reviews traditional approaches to the problem of blind source separation and speech dereverberation. In Section III, we demonstrate how to blindly identify a MIMO system. Section IV explains how to derive $M$ independent SIMO systems from a MIMO system with $M$ speech sources such that each SIMO system is free of interference from other sources. In Section V, we show how to perform dereverberation for a SIMO system using the Bezout theorem. Section VI evaluates the proposed approach by simulations. Finally, we give our conclusions in Section VII.

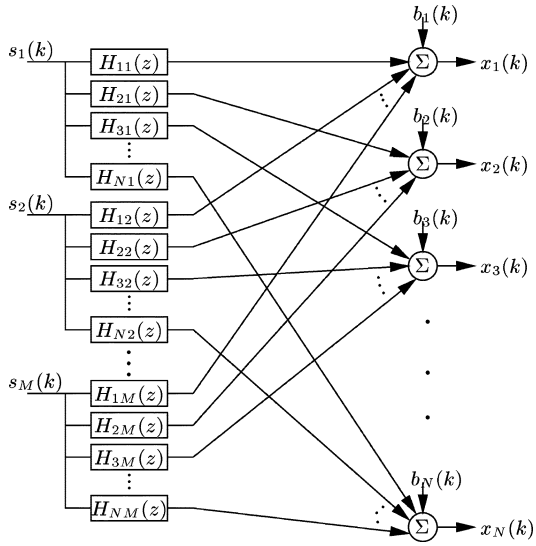



Fig. 1. Illustration of a MIMO FIR acoustic system having $M$ speech sources and $N$ microphones.

## II. Signal Model and Problem Formulation

### A. MIMO System

Suppose that we have $M$ independent speech sources and $N$ microphones with $M < N$ in a room, which is mathematically described by an $M \times N$ MIMO FIR system as shown in Fig. 1. At the $n$th microphone and at the $k$th sample time, we have

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

where $(\cdot)^T$ denotes the transpose of a matrix or a vector

$$\mathbf{h}_{nm} = [h_{nm,0} \quad h_{nm,1} \quad \cdots \quad h_{nm,L_h-1}]^T, \quad n = 1, 2, \cdots, N, \quad m = 1, 2, \cdots, M$$

is the impulse response (of length $L_h$, $\forall m, n$) between source $m$ and microphone $n$

$$\mathbf{s}_m(k, L_h) = [s_m(k) \quad s_m(k-1) \quad \cdots \quad s_m(k-L_h+1)]^T$$

is a vector containing the last $L_h$ samples of the $m$th source signal $s_m$, and $b_n(k)$ is a zero-mean additive white Gaussian noise (AWGN) with variance $\sigma_b^2$, $\forall n$.

Using the $z$ transform, the signal model of the MIMO system (1) is expressed as

$$X_n(z) = \sum_{m=1}^{M} H_{nm}(z) S_m(z) + B_n(z), \quad n = 1, 2, \cdots, N \tag{2}$$

where

$$H_{nm}(z) = \sum_{l=0}^{L_h-1} h_{nm,l} z^{-l}. \tag{3}$$

### B. Traditional Blind Source Separation and Speech Dereverberation Approaches

In BSS methods, the *a priori* knowledge about neither the channel impulse responses $h_{nm}$ nor the source signals $s_m(k)$ is assumed. Only the mutual independence among the source signals is utilized to separate them from the observations of their mixtures $x_n(k)$.

In the general form, traditional BSS algorithms construct a set of de-mixing filters and apply them to the microphone signals. The output of the de-mixing system are regarded as estimates of the separated signals, which are presumably independent. Existing BSS methods differ in the way that the dependence of the separated speech signals is defined, or equivalently, the employed criteria for optimizing the de-mixing filters. Accordingly, BSS methods can be broadly dichotomized into the class of second-order statistics (SOS) algorithms and the class of higher-order statistics (HOS) algorithms. To minimize estimation variance, computing an HOS measure demands a large number of observations, which leads to an increase in computational complexity. However, the assumption of mutual independence alone is not sufficient to solve the problem using only SOS and hence speech's nonstationary nature is exploited.

With a de-mixing system reinforcing the assumption of mutual independence, the speech signals are separated inherently up to an arbitrary filter and permutation. Permutation inconsistency is a challenging problem in a frequency-domain approach and will apparently impair the separation performance. Even if permutation ambiguity could be somehow overcome, the arbitrary filter itself implies undesirable distortion and consequently speech quality can not be predicted. Currently the research on this problem is in the direction of how to incorporate the distortion of separated speech into the cost function while adapting the de-mixing filters, e.g., the minimal distortion principle in [13]. However, the convergence would be sensitive to the relative weights of the two components, i.e., mutual independence and speech distortion, in the cost function and the overall performance is limited. New ideas are necessary for solving the problem of blind source separation and speech dereverberation.

## III. Blind Identification of a MIMO System

In this paper, we intend to separate competing speech sources by first blindly identifying the MIMO FIR system. Blind MIMO identification is difficult even for communication systems with short channel impulse responses. It becomes dramatically complicated when an acoustic system is the target as the case studied in this paper. Trying to solve it all at once involves a huge number of parameters to estimate and the current research in this area remains at the stage of feasibility investigations. Moreover, scaling and permutation ambiguities are similar to what have been observed in the BSS problem. Therefore we propose to decompose the problem into several subproblems in which SIMO systems are blindly identified. We assume that from time to time each speaker occupies at least one exclusive interval alone and when they start talking simultaneously the room acoustics have not significantly varied. Then in each single-talk interval a SIMO system will be blindly identified and its channel impulse responses will be saved for later use in source separation and speech deconvolution during double or multiple talk periods. The speech source detection technique that distinguishes single and multiple talk is an interesting and important issue, but is beyond the scope of this paper. The reader who is interested in this topic can read a recently published paper on this problem [14] and references therein.

Blind identification of a SIMO system can be achieved with only the SOS of system outputs as long as the following two conditions are met [15]:

1) polynomials formed from the channel impulse responses are co-prime, i.e., the channel transfer functions do not share any common zeros;
2) autocorrelation matrix of the source signal is of full rank, making the SIMO system fully excited.

In an earlier study [16], we developed a number of adaptive algorithms for blind identification of a SIMO system in the time domain, including multichannel LMS (MCLMS) and multichannel Newton methods. The idea of adaptive blind SIMO identification was later implemented in the frequency domain for computational efficiency and fast convergence [17]. This so-called unconstrained normalized multichannel frequency-domain LMS (UNMCFLMS) algorithm was shown to perform well with an acoustic system and will be employed in this paper.

## IV. Separating Spatial Interference and Temporal Echoes

In this section, we will explain how to separate spatial interference from other co-existing sources and temporal echoes due to the reflection by room surfaces. From the signal processing perspective, this separation is achieved by converting an $M \times N$ MIMO system into $M$ interference-free SIMO systems. The development begins with an example of the simplest $2 \times 3$ MIMO system and then extends to a general $M \times N$ case.

### A. Example: Conversion of a 2 × 3 MIMO System to Two SIMO Systems

For a $2 \times 3$ MIMO system, the spatial interference can be cancelled by using two output signals at a time. For instance, we can remove the interference in $X_1(z)$ and $X_2(z)$ caused by $S_2(z)$ (from the perspective of source 1) as follows:

$$\begin{aligned} &X_1(z)H_{22}(z) - X_2(z)H_{12}(z) \\ &\quad = \left[H_{11}(z)H_{22}(z) - H_{21}(z)H_{12}(z)\right] S_1(z) \\ &\qquad + \left[H_{22}(z)B_1(z) - H_{12}(z)B_2(z)\right]. \end{aligned} \tag{4}$$

Similarly, the interference caused by $S_1(z)$ (from the perspective of source 2) in these two outputs can also be cancelled. Therefore, by selecting different pairs from the three outputs, we could obtain six interference-free signals and then could construct two separate single-input three-output systems with respect to two distinct inputs, respectively. This procedure is visualized in Fig. 2 and will be described in a more systematic way in the following.

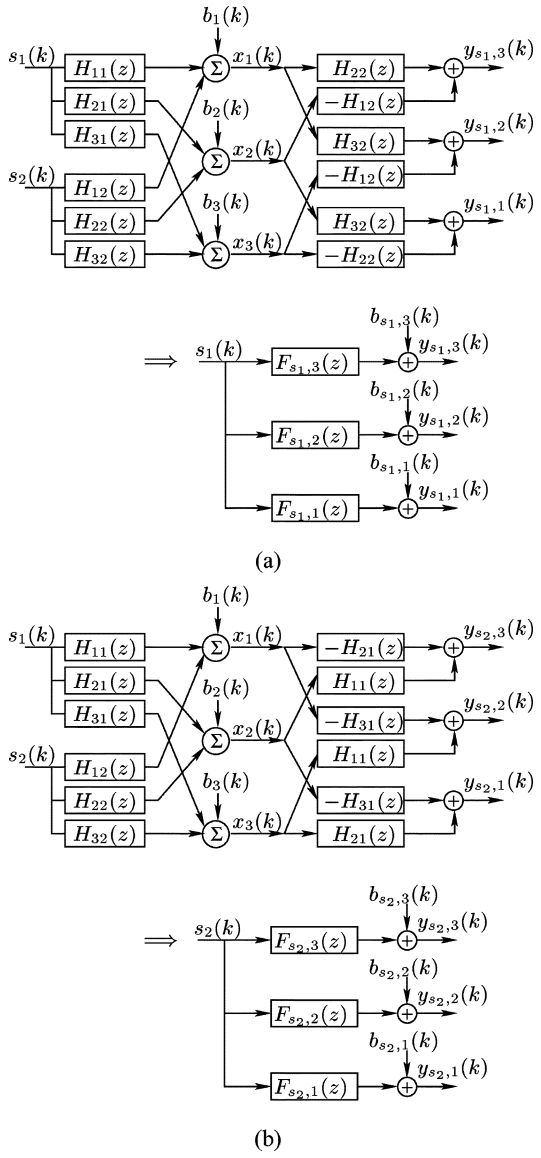



Fig. 2. Illustration of the conversion from a 2 × 3 MIMO system to two interference-free SIMO systems with respect to (a) $s_1(k)$ and (b) $s_2(k)$.

Let us consider the following equation:

$$\begin{aligned} Y_{s_1,p}(z) &= H_{s_1,p1}(z)X_1(z) + H_{s_1,p2}(z)X_2(z) \\ &\quad + H_{s_1,p3}(z)X_3(z) \\ &= \sum_{q=1}^{3} H_{s_1,pq}(z)X_q(z), \quad p = 1,2,3 \end{aligned} \tag{5}$$

where $H_{s_1,pp}(z) = 0$, $\forall p$. This means that (5) considers only two microphone signals for each $p$. The objective is to find the polynomials $H_{s_1,pq}(z), p,q = 1,2,3, p \neq q$, in such a way that

$$Y_{s_1,p}(z) = F_{s_1,p}(z)S_1(z) + B_{s_1,p}(z), \quad p = 1,2,3 \tag{6}$$

which represents a SIMO system where $s_1$ is the source signal, $y_{s_1,p}$, $p = 1, 2, 3$, are the received microphone signals, $f_{s_1,p}$ are the corresponding acoustic paths, and $b_{s_1,p}$ is the noise at microphone $p$. Using (2) in (5), we deduce that

$$\begin{aligned} Y_{s_1,1}(z) &= [H_{s_1,12}(z)H_{21}(z) + H_{s_1,13}(z)H_{31}(z)]\,S_1(z) \\ &\quad + [H_{s_1,12}(z)H_{22}(z) + H_{s_1,13}(z)H_{32}(z)]\,S_2(z) \\ &\quad + H_{s_1,12}(z)B_2(z) + H_{s_1,13}(z)B_3(z) \end{aligned} \tag{7}$$

$$\begin{aligned} Y_{s_1,2}(z) &= [H_{s_1,21}(z)H_{11}(z) + H_{s_1,23}(z)H_{31}(z)]\,S_1(z) \\ &\quad + [H_{s_1,21}(z)H_{12}(z) + H_{s_1,23}(z)H_{32}(z)]\,S_2(z) \\ &\quad + H_{s_1,21}(z)B_1(z) + H_{s_1,23}(z)B_3(z) \end{aligned} \tag{8}$$

$$\begin{aligned} Y_{s_1,3}(z) &= [H_{s_1,31}(z)H_{11}(z) + H_{s_1,32}(z)H_{21}(z)]\,S_1(z) \\ &\quad + [H_{s_1,31}(z)H_{12}(z) + H_{s_1,32}(z)H_{22}(z)]\,S_2(z) \\ &\quad + H_{s_1,31}(z)B_1(z) + H_{s_1,32}(z)B_2(z). \end{aligned} \tag{9}$$

As shown in Fig. 2, one possibility is to choose

$$\begin{aligned} H_{s_1,12}(z) &= H_{32}(z), \quad H_{s_1,13}(z) = -H_{22}(z), \\ H_{s_1,21}(z) &= H_{32}(z), \quad H_{s_1,23}(z) = -H_{12}(z), \\ H_{s_1,31}(z) &= H_{22}(z), \quad H_{s_1,32}(z) = -H_{12}(z). \end{aligned} \tag{10}$$

In this case, we find that

$$\begin{aligned} F_{s_1,1}(z) &= H_{32}(z)H_{21}(z) - H_{22}(z)H_{31}(z), \\ F_{s_1,2}(z) &= H_{32}(z)H_{11}(z) - H_{12}(z)H_{31}(z), \\ F_{s_1,3}(z) &= H_{22}(z)H_{11}(z) - H_{12}(z)H_{21}(z) \end{aligned} \tag{11}$$

and

$$\begin{aligned} B_{s_1,1}(z) &= H_{32}(z)B_2(z) - H_{22}(z)B_3(z), \\ B_{s_1,2}(z) &= H_{32}(z)B_1(z) - H_{12}(z)B_3(z), \\ B_{s_1,3}(z) &= H_{22}(z)B_1(z) - H_{12}(z)B_2(z). \end{aligned} \tag{12}$$

Since $\deg[H_{nm}(z)] = L_h - 1$, where $\deg[\cdot]$ is the degree of a polynomial, therefore $\deg[F_{s_1,p}(z)] \leq 2L_h - 2$. We can see from (11) that polynomials $F_{s_1,1}(z)$, $F_{s_1,2}(z)$, and $F_{s_1,3}(z)$ share common zeros if $H_{12}(z)$, $H_{22}(z)$, and $H_{32}(z)$ [or if $H_{11}(z)$, $H_{21}(z)$, and $H_{31}(z)$] share common zeros.

Now suppose that $C_2(z) = \gcd[H_{12}(z), H_{22}(z), H_{32}(z)]$, where $\gcd[\cdot]$ denotes the greatest common divisor of the polynomials involved. We have

$$H_{n2}(z) = C_2(z)H'_{n2}(z), \quad n = 1,2,3. \tag{13}$$

It is clear that the signal $s_2$ in (5) can be canceled by using the polynomials $H'_{n2}(z)$ [instead of $H_{n2}(z)$ as given in (10)], so that the SIMO system represented by (6) will change to

$$Y'_{s_1,p}(z) = F'_{s_1,p}(z)S_1(z) + B'_{s_1,p}(z), \quad p = 1,2,3 \tag{14}$$

where

$$\begin{aligned} F'_{s_1,p}(z)C_2(z) &= F_{s_1,p}(z), \\ B'_{s_1,p}(z)C_2(z) &= B_{s_1,p}(z). \end{aligned}$$

It is worth noticing that $\deg[F'_{s_1,p}(z)] \leq \deg[F_{s_1,p}(z)]$ and that polynomials $F'_{s_1,1}(z)$, $F'_{s_1,2}(z)$, and $F'_{s_1,3}(z)$ share common zeros if and only if $H_{11}(z)$, $H_{21}(z)$, and $H_{31}(z)$ share common zeros.

The second SIMO system corresponding to the source $s_2$ can be derived in a similar way. Indeed, we can find the output signals

$$Y_{s_2,p}(z) = F_{s_2,p}(z)S_2(z) + B_{s_2,p}(z), \quad p = 1,2,3 \tag{15}$$

by making $F_{s_2,p}(z) = F_{s_1,p}(z)$ ($p = 1, 2, 3$) where the noise is

$$\begin{aligned} B_{s_2,1}(z) &= -H_{31}(z)B_2(z) + H_{21}(z)B_3(z) \\ B_{s_2,2}(z) &= -H_{31}(z)B_1(z) + H_{11}(z)B_3(z) \\ B_{s_2,3}(z) &= -H_{21}(z)B_1(z) + H_{11}(z)B_2(z). \end{aligned}$$

This means that the two SIMO systems [for $s_1$ and $s_2$, represented by (6) and (15)] have identical channels but the noise at the microphones is different.

Now let's see what we can do if $H_{n1}(z)$ ($n = 1, 2, 3$) share common zeros. Suppose that $C_1(z)$ is the greatest common divisor of $H_{11}(z)$, $H_{21}(z)$, and $H_{31}(z)$. Then we have

$$H_{n1}(z) = C_1(z)H'_{n1}(z), \quad n = 1,2,3 \tag{16}$$

and the SIMO system of (15) becomes

$$Y'_{s_2,p}(z) = F'_{s_2,p}(z)S_2(z) + B'_{s_2,p}(z), \quad p = 1,2,3 \tag{17}$$

where

$$\begin{aligned} F'_{s_2,p}(z)C_1(z) &= F_{s_2,p}(z), \\ B'_{s_2,p}(z)C_1(z) &= B_{s_2,p}(z). \end{aligned}$$

We see that

$$\begin{aligned} &\gcd\left[F'_{s_2,1}(z), F'_{s_2,2}(z), F'_{s_2,3}(z)\right] \\ &\quad = \gcd\left[H_{12}(z), H_{22}(z), H_{32}(z)\right] \\ &\quad = C_2(z) \end{aligned}$$

and in general $F'_{s_1,p}(z) \neq F'_{s_2,p}(z)$.

### B. Generalization

The approach to separating spatial interference and temporal echoes explained in the previous subsection on a simple example will be generalized here to an $(M, N)$ MIMO system ($M < N$). We begin with writing (2) into a vector/matrix form

$$\vec{\mathbf{X}}(z) = \mathbf{H}(z)\vec{\mathbf{S}}(z) + \vec{\mathbf{B}}(z) \tag{18}$$

where

$$\begin{aligned} \vec{\mathbf{X}}(z) &= \left[X_1(z) \quad X_2(z) \quad \cdots \quad X_N(z)\right]^T, \\ \mathbf{H}(z) &= \begin{bmatrix} H_{11}(z) & H_{12}(z) & \cdots & H_{1M}(z) \\ H_{21}(z) & H_{22}(z) & \cdots & H_{2M}(z) \\ \vdots & \vdots & \vdots & \vdots \\ H_{N1}(z) & H_{N2}(z) & \cdots & H_{NM}(z) \end{bmatrix}, \\ \vec{\mathbf{S}}(z) &= \left[S_1(z) \quad S_2(z) \quad \cdots \quad S_M(z)\right]^T, \\ \vec{\mathbf{B}}(z) &= \left[B_1(z) \quad B_2(z) \quad \cdots \quad B_N(z)\right]^T. \end{aligned}$$

If $C_m(z) = \gcd[H_{1m}(z), H_{2m}(z), \cdots, H_{Nm}(z)]$ ($m = 1, 2, \cdots, M$), then $H_{nm}(z) = C_m(z)H'_{nm}(z)$ and the channel matrix $\mathbf{H}(z)$ can be rewritten as

$$\mathbf{H}(z) = \mathbf{H}'(z)\mathbf{C}(z) \tag{19}$$

where $\mathbf{H}'(z)$ is an $N \times M$ matrix containing the elements $H'_{nm}(z)$ and $\mathbf{C}(z)$ is an $M \times M$ diagonal matrix with $C_m(z)$ as its nonzero components.

Let us choose $M$ from $N$ microphone outputs and we have $P = C_N^M$ different ways of doing so. For the $p$th ($p = 1, 2, \cdots, P$) combination, we denote the index of the $M$ selected microphone signals as $p_m$, $m = 1, 2, \cdots, M$, and get an $(M, M)$ MIMO subsystem.

Consider the following equations:

$$\vec{\mathbf{Y}}_p(z) = \mathbf{H}_{s,p}(z)\vec{\mathbf{X}}_p(z), \quad p = 1,2,\cdots,P \tag{20}$$

where

$$\begin{aligned} \vec{\mathbf{Y}}_p(z) &= \left[Y_{s_1,p}(z) \quad Y_{s_2,p}(z) \quad \cdots \quad Y_{s_M,p}(z)\right]^T, \\ \mathbf{H}_{s,p}(z) &= \begin{bmatrix} H_{s_1,p1}(z) & H_{s_1,p2}(z) & \cdots & H_{s_1,pM}(z) \\ H_{s_2,p1}(z) & H_{s_2,p2}(z) & \cdots & H_{s_2,pM}(z) \\ \vdots & \vdots & \vdots & \vdots \\ H_{s_M,p1}(z) & H_{s_M,p2}(z) & \cdots & H_{s_M,pM}(z) \end{bmatrix}, \\ \vec{\mathbf{X}}_p(z) &= \left[X_{p_1}(z) \quad X_{p_2}(z) \quad \cdots \quad X_{p_M}(z)\right]^T. \end{aligned}$$

Let $\mathbf{H}_p(z)$ be the $M \times M$ matrix obtained from the system's channel matrix $\mathbf{H}(z)$ by keeping its rows corresponding to the $M$ selected microphone signals. Then similar to (18), we have

$$\vec{\mathbf{X}}_p(z) = \mathbf{H}_p(z)\vec{\mathbf{S}}(z) + \vec{\mathbf{B}}_p(z) \tag{21}$$

where

$$\vec{\mathbf{B}}_p(z) = \left[B_{p_1}(z) \quad B_{p_2}(z) \quad \cdots \quad B_{p_M}(z)\right]^T.$$

Substituting (21) into (20) yields

$$\vec{\mathbf{Y}}_p(z) = \mathbf{H}_{s,p}(z)\mathbf{H}_p(z)\vec{\mathbf{S}}(z) + \mathbf{H}_{s,p}(z)\vec{\mathbf{B}}_p(z). \tag{22}$$

In order to remove the spatial interference, the objective here is to find the matrix $\mathbf{H}_{s,p}(z)$ whose components are linear combinations of $H_{nm}(z)$ such that the product

$$\mathbf{\Phi}_p(z) \triangleq \mathbf{H}_{s,p}(z)\mathbf{H}_p(z) \tag{23}$$

would be a diagonal matrix. Consequently, we have

$$\begin{aligned} Y_{s_m,p}(z) = F_{s_m,p}(z)S_m(z) + B_{s_m,p}(z), \\ m = 1,2,\cdots,M, \quad p = 1,2,\cdots,P. \end{aligned} \tag{24}$$

In the above, we showed that spatial interference and temporal echoes are separable by converting an $(M, N)$ MIMO acoustic system into $M$ interference-free SIMO systems. Although source separation has been achieved, the obtained multiple interference-free speech signals would sound possibly more reverberant due to the prolonged impulse response of the equivalent channels. In this section, we will illustrate how these annoying temporal echoes can be perfectly removed

and the original speech signal can be recovered from a SIMO system.If $\mathbf{C}_p(z)$ [obbtained from $\mathbf{C}(z)$ in a similar way as $\mathbf{H}_p(z)$ is constructed] is not equal to the identity matrix, then $\mathbf{H}_p(z) = \mathbf{H}'_p(z)\mathbf{C}(z)$, where $\mathbf{H}'_p(z)$ has full column normal rank in acoustic environments as we assume in this paper[1] (i.e., $\mathrm{nrank}[\mathbf{H}'_p(z)] = M$, see [18] for a definition of normal rank), and the interference-free signals are determined as

$$\vec{\mathbf{Y}}'_p(z) = \mathbf{H}'_{s,p}(z)\mathbf{H}'_p(z)\mathbf{C}_p(z)\vec{\mathbf{S}}(z) + \mathbf{H}'_{s,p}(z)\vec{\mathbf{W}}_p(z) \quad (25)$$

and

$$Y'_{s_m,p}(z) = F'_{s_m,p}(z)S_m(z) + W'_{s_m,p}(z). \quad (26)$$

Obviously a good choice for $\mathbf{H}'_{s,p}(z)$ to make the product $\mathbf{H}'_{s,p}(z)\mathbf{H}'_p(z)$ a diagonal matrix is the adjoint of matrix $\mathbf{H}'_p(z)$, i.e., the $(i,j)$th element of $\mathbf{H}'_{s,p}(z)$ is the $(j,i)$th cofactor of $\mathbf{H}'_p(z)$. Consequently, the polynomial $F'_{s_m,p}(z)$ would be the determinant of $\mathbf{H}'_p(z)$.

Since

$$F'_{s_m,p}(z) = \sum_{q=1}^{M} H'_{s_m,pq}(z)H_{p_q m}(z)$$

and $H'_{s_m,pq}(z)$ $(q = 1,2,\cdots,M)$ are co-prime, the polynomials $F'_{s_m,p}(z)$ $(p = 1,2,\cdots,P)$ share common zeros if and only if the polynomials $H_{nm}(z)$ $(n = 1,2,\cdots,N)$ share common zeros. Therefore, if the channels with respect to any one input are co-prime for an $(M,N)$ MIMO system, we can convert it into $M$ interference-free SIMO systems whose $C_N^M$ channels are also co-prime, i.e., their channel matrices are irreducible.

Also, it can easily be checked that $\deg[F'_{s_m,p}(z)] \le M(L_h - 1)$. As a result, the length of the FIR filter $f'_{s_m,p}$ would be

$$L_f \le M(L_h - 1) + 1. \quad (27)$$

## V. Speech Dereverberation for SIMO Systems

### A. Principle

For the SIMO system with respect to source $s_m$ $(m = 1,2,\cdots,M)$, we consider the polynomials $G_{s_m,p}(z)$ $(p = 1,2,\cdots,P)$ and the equation

$$\begin{aligned}\widehat{S}_m(z) &= \sum_{p=1}^{P} G_{s_m,p}(z)Y'_{s_m,p}(z) \\ &= \left[\sum_{p=1}^{P} F'_{s_m,p}(z)G_{s_m,p}(z)\right] S_m(z) \\ &\quad + \sum_{p=1}^{P} G_{s_m,p}(z)B'_{s_m,p}(z). \end{aligned} \quad (28)$$

[1]For a square matrix $(M \times M)$, the normal rank is full if and only if the determinant, which is a polynomial in $z$, is not identically zero for all $z$. In this case, the rank is less than $M$ only at a finite number of points in the $z$ plane.

The polynomials $G_{s_m,p}(z)$ should be found in such a way that $\widehat{S}_m(z) = S_m(z)$ in the absence of noise by using the Bezout theorem which is mathematically expressed as follows:

$$\begin{aligned}&\gcd\left[F'_{s_m,1}(z), F'_{s_m,2}(z), \cdots, F'_{s_m,P}(z)\right] = 1 \\ &\quad \Leftrightarrow \exists G_{s_m,1}(z), G_{s_m,2}(z), \cdots, G_{s_m,P}(z): \\ &\qquad \sum_{p=1}^{P} F'_{s_m,p}(z)G_{s_m,p}(z) = 1. \end{aligned} \quad (29)$$

In other words, if the polynomials $F'_{s_m,p}(z)$ $(p = 1,2,\cdots,P)$ have no common zeros (which is equivalent to saying that the polynomials $H_{nm}(z)$, $n = 1,2,\cdots,N$, don't share any common zeros), it is possible to perfectly equalize (in the noiseless case) each one of the $M$ SIMO systems. The idea of using the Bezout theorem for dereverberation of an acoustic SIMO system was first proposed in [19] in the context of room acoustics, where the method is more widely referred to as the MINT theory. It relieves the constraint on a single-channel acoustic system for perfect dereverberation that the channel impulse response must be a minimum-phase polynomial.

If the channels of the SIMO system share common zeros, i.e.,

$$C'_{s_m}(z) = \gcd\left[F'_{s_m,1}(z), F'_{s_m,2}(z), \cdots, F'_{s_m,P}(z)\right] \ne 1 \quad (30)$$

then we have

$$F'_{s_m,p}(z) = C'_{s_m}(z)F''_{s_m,p}(z), \quad p = 1,2,\cdots,P \quad (31)$$

and the polynomials $G_{s_m,p}(z)$ can be found such that

$$\sum_{p=1}^{P} F''_{s_m,p}(z)G_{s_m,p}(z) = 1. \quad (32)$$

In this case, (28) becomes

$$\widehat{S}_m(z) = C'_{s_m}(z)S_m(z) + \sum_{p=1}^{P} G_{s_m,p}(z)B'_{s_m,p}(z). \quad (33)$$

We see that by using the Bezout theorem, the $m$th SIMO system can be equalized up to the polynomial $C'_{s_m}(z)$. So when there are common zeros, the Bezout theorem can only partially dereverberate the speech signal. For complete dereverberation, we have to add another stage to the process by examining $C'_{s_m}(z)$. If $C'_{s_m}(z)$ is minimum phase (i.e., the zeros are inside the unit circle), its inversion is stable and a complete dereverberation still can be attained

$$\begin{aligned}\widehat{\widehat{S}}_m(z) &= \frac{1}{C'_{s_m}(z)}\widehat{S}_m(z) \\ &= S_m(z) + \frac{1}{C'_{s_m}(z)}\sum_{p=1}^{P} G_{s_m,p}(z)B'_{s_m,p}(z). \end{aligned} \quad (34)$$

Otherwise, a least squares solution is derived to at best minimize the effect of $C'_{s_m}(z)$ in (33).

To find the dereverberation filters, we write the Bezout (29) in the time domain as

$$\begin{aligned}\mathbf{F}'^{\mathrm{c}}_{s_m}\mathbf{g}_{s_m} &= \sum_{p=1}^{P} \mathbf{F}'^{\mathrm{c}}_{s_m,p}\mathbf{g}_{s_m,p} \\ &= \mathbf{e}_1, \quad m = 1,2,\cdots,M \end{aligned} \quad (35)$$

where

$$\mathbf{F}_{s_m}^{\prime c} = \left[\mathbf{F}_{s_m,1}^{\prime c} \quad \mathbf{F}_{s_m,2}^{\prime c} \quad \cdots \quad \mathbf{F}_{s_m,P}^{\prime c}\right],$$
$$\mathbf{g}_{s_m} = \left[\mathbf{g}_{s_m,1}^{T} \quad \mathbf{g}_{s_m,2}^{T} \quad \cdots \quad \mathbf{g}_{s_m,P}^{T}\right]^{T},$$
$$\mathbf{g}_{s_m,p} = \left[g_{s_m,p,0} \quad g_{s_m,p,1} \quad \cdots \quad g_{s_m,p,L_g-1}\right]^{T},$$
$$m = 1,2,\cdots,M, \quad p = 1,2,\cdots,P$$

$L_g$ is the length of the FIR filter $g_{s_m,p}$

$$\mathbf{F}_{s_m,p}^{\prime c} = \begin{bmatrix} f'_{s_m,p,0} & 0 & \cdots & 0 \\ f'_{s_m,p,1} & f'_{s_m,p,0} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ f'_{s_m,p,L_f-1} & \cdots & \cdots & \vdots \\ 0 & f'_{s_m,p,L_f-1} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & f'_{s_m,p,L_f-1} \end{bmatrix}$$

is an $(L_f + L_g - 1) \times L_g$ matrix, $L_f$ is the length of the FIR filter $f'_{s_m,p}$, and

$$\mathbf{e}_1 = [1 \quad 0 \quad \cdots \quad 0]^T$$

is an $(L_f + L_g - 1) \times 1$ vector. In order to have a unique solution for (35), $L_g$ must be chosen in such a way that $\mathbf{F}_{s_m}^{\prime c}$ is a square matrix. In this case, we have

$$L_g = \frac{L_f - 1}{P - 1}. \tag{36}$$

Using (27), the length of the dereverberation filter is bounded by

$$L_g \leq \frac{M(L_h - 1)}{P - 1}. \tag{37}$$

### B. Least-Squares Implementation

It is now clear that by using the Bezout theorem the $M$ SIMO system can be perfected dereverberated in the noiseless case as long as their channel impulse responses share no common zeros. In addition, we derived what is the minimum length $L_g$ of the dereverberation filters, as given in (37). Although finding the shortest dereverberation filters involves the lowest computational complexity and leads to the most cost effective implementation, the performance may not be the best due to noise in practice. Moreover, the smallest $L_g$ may not be even possible since (36) does not guarantee an integer solution. Therefore, we choose a larger $L_g$ than necessary in our implementation and solve (35) for $\mathbf{g}_{s_m}$ in the least squares sense

$$\mathbf{g}_{s_m,\mathrm{LS}} = \mathbf{F}_{s_m}^{\prime c\dagger}\mathbf{e}_1 \tag{38}$$

where

$$\mathbf{F}_{s_m}^{\prime c\dagger} = \left(\mathbf{F}_{s_m}^{\prime cT}\mathbf{F}_{s_m}^{\prime c}\right)^{-1}\mathbf{F}_{s_m}^{\prime cT}$$

is the pseudo-inverse of the matrix $\mathbf{F}_{s_m}^{\prime c}$. If a decision delay $d$ is taken into account, then the dereverberation filters turn out to be

$$\mathbf{g}_{s_m,\mathrm{LS}} = \mathbf{F}_{s_m}^{\prime c\dagger}\mathbf{e}_d \tag{39}$$

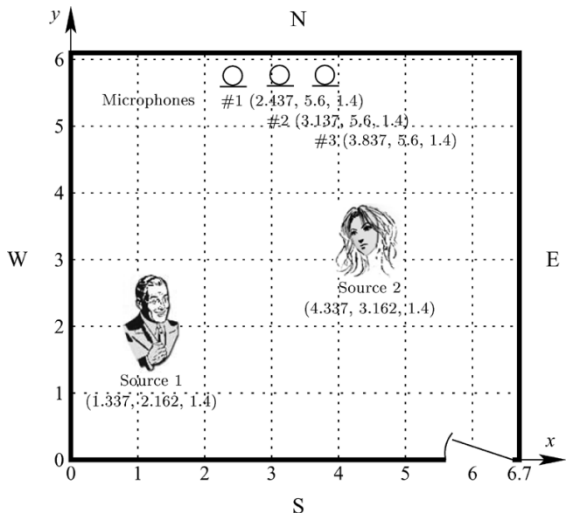



Fig. 3. Floor plan of the varechoic chamber at Bell Labs (coordinate values measured in meters).

where

$$\mathbf{e}_d = \left[\underbrace{0 \quad \cdots \quad 0}_{d} \quad 1 \quad \underbrace{0 \quad \cdots \quad 0}_{L_f+L_g-d-2}\right]^T.$$

## VI. Simulations

In this section, we will evaluate the performance of the proposed blind source separation and speech dereverberation algorithm via simulations in realistic acoustic environments.

### A. Performance Measures

Similar to what was adopted in our earlier study [17], we will use the normalized projection misalignment (NPM) to evaluate the performance of a BCI algorithm [20]. The NPM is defined as

$$\mathrm{NPM} \triangleq 20\log_{10}\left[\frac{\|\boldsymbol{\epsilon}\|}{\|\mathbf{h}\|}\right] \tag{40}$$

where

$$\boldsymbol{\epsilon} = \mathbf{h} - \frac{\mathbf{h}^T\hat{\mathbf{h}}}{\hat{\mathbf{h}}^T\hat{\mathbf{h}}}\hat{\mathbf{h}}$$

is the *projection misalignment* vector. By projecting $\mathbf{h}$ onto $\hat{\mathbf{h}}$ and defining a projection error, we take into account only the intrinsic misalignment of the channel estimate, disregarding an arbitrary gain factor.

To evaluate the performance of source separation and speech dereverberation, two measures, namely signal-to-interference ratio (SIR) and speech spectral distortion, are used in the simulations. For the SIR, we referred to the notion given in [10] but defined the measure in a different manner since their definition is applicable only for an $M \times M$ MIMO system. In this paper, our interest is in the more general $M \times N$ MIMO systems with $M < N$.

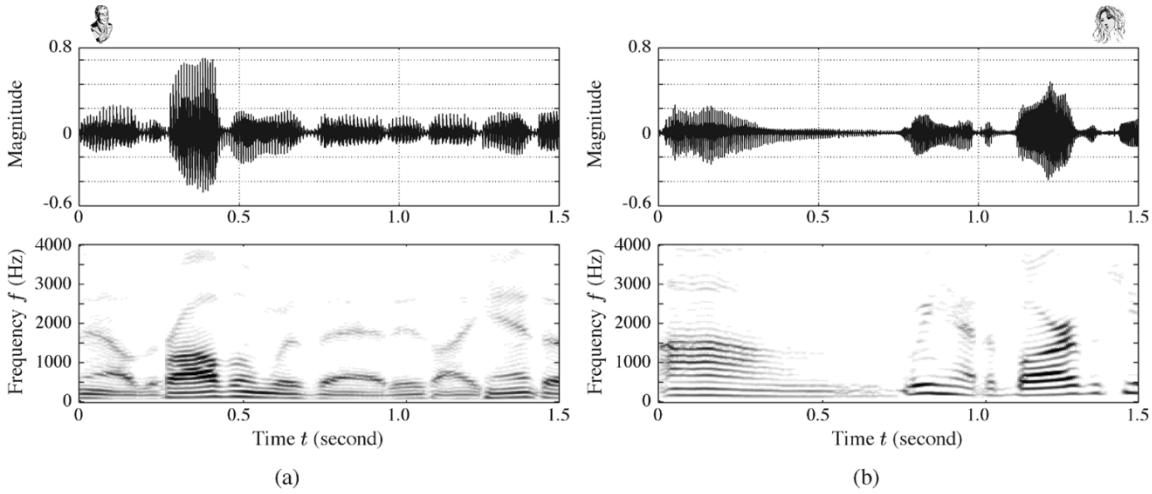



Fig. 4. Time sequence and spectrogram (30 Hz bandwidth) of the two speech source signals used in the simulations for the first 1.5 s. (a) $s_1(k)$ (male speaker) and (b) $s_2(k)$ (female speaker).

TABLE I
PERFORMANCE OF THE SOURCE SEPARATION AND SPEECH DEVERBERATION ALGORITHM BASED ON THE BATCH (SVD) AND ADAPTIVE FREQUENCY-DOMAIN BCI (UNMCFLMS) IMPLEMENTATIONS IN THE VARECHOIC CHAMBER AT BELL LABS WITH DIFFERENT PANEL CONFIGURATIONS

| BCI Methods | NPM (dB) | | $\mathrm{SIR}^{\mathrm{in}}$ | $\mathrm{SIR}^{\mathrm{out}}$ | After S.S. | | After S.D. | |
|---|---|---|---|---|---|---|---|---|
| | $\mathrm{SIMO}_{s_1}$ | $\mathrm{SIMO}_{s_2}$ | (dB) | (dB) | $d_{\mathrm{IS},s_1}^{\mathrm{SS}}$ | $d_{\mathrm{IS},s_2}^{\mathrm{SS}}$ | $d_{\mathrm{IS},s_1}^{\mathrm{SD}}$ | $d_{\mathrm{IS},s_2}^{\mathrm{SD}}$ |
| 89% panels open, $T_{60}=240$ ms, $L_h=256$ | | | | | | | | |
| Adaptive (UNMCFLMS) | -17.4716 | -17.8005 | 0.2306 | 52.0158 | 1.9476 | 2.1368 | 0.0476 | 0.0258 |
| Batch (2500 samples) | -43.5562 | -35.3735 | 0.2306 | 74.9704 | 2.1041 | 2.2104 | 0.0002 | 0.0006 |
| 75% panels open, $T_{60}=310$ ms, $L_h=256$ | | | | | | | | |
| Adaptive (UNMCFLMS) | -18.7371 | -18.0565 | 0.3911 | 52.8992 | 2.6878 | 2.8424 | 0.0307 | 0.0205 |
| Batch (2500 samples) | -50.5332 | -42.9942 | 0.3911 | 74.9660 | 2.8021 | 3.0904 | 0.0032 | 0.0031 |
| 30% panels open, $T_{60}=380$ ms, $L_h=512$ | | | | | | | | |
| Adaptive (UNMCFLMS) | -13.3664 | -11.5813 | 0.2493 | 44.7796 | 2.4084 | 3.9952 | 0.0755 | 0.1773 |
| Batch (3000 samples) | -38.7105 | -29.8682 | 0.2493 | 73.9322 | 2.8960 | 4.2943 | 0.0019 | 0.0044 |
| Panels all closed, $T_{60}=580$ ms, $L_h=512$ | | | | | | | | |
| Adaptive (UNMCFLMS) | -13.8181 | -14.4699 | 0.4745 | 44.9987 | 2.6414 | 4.8217 | 0.1592 | 0.1988 |
| Batch (3000 samples) | -50.5332 | -42.9942 | 0.4745 | 73.8734 | 2.6321 | 4.4309 | 0.0004 | 0.0007 |

NOTES: $\mathrm{SIMO}_{s_m}$ respresents the SIMO system corresponding to source $s_m$.
$T_{60}$ denotes 60-dB reverberation time in the 20-4000 Hz band.
S.S. and S.D. stand for source separation and speech dereverberation, respectively.

We first define the average input SIR at microphone $n$ as

$$\mathrm{SIR}_n^{\mathrm{in}} \triangleq \frac{1}{M}\sum_{m=1}^{M}\left(\frac{E\left\{[h_{nm}*s_m(k)]^2\right\}}{\sum_{i=1,i\neq m}^{M}E\left\{[h_{ni}*s_i(k)]^2\right\}}\right), \quad n=1,2,\cdots,N \tag{41}$$

where $*$ denotes linear convolution. Then the overall average input SIR is given by

$$\mathrm{SIR}^{\mathrm{in}} \triangleq \frac{1}{N}\sum_{n=1}^{N}\mathrm{SIR}_n^{\mathrm{in}}. \tag{42}$$

The output SIR is defined using the same principle but the expression will be more complicated. For a concise presentation, we denote $\phi_{p,ji}$ ($p=1,2,\cdots,P$, $i,j=1,2,\cdots,M$) as the impulse response of the equivalent channel from the $i$th input to the $j$th output for the $p$th $M\times M$ separation subsystem. From (22) and (23), we know that $\phi_{p,ji}$ corresponds to the $(j,i)$th element of $\mathbf{\Phi}_p(z)$ and $\psi_{p,mm}=f_{s_m,p}$. Then the average output SIR for the $p$th subsystem is:

$$\mathrm{SIR}_p^{\mathrm{out}} \triangleq \frac{\sum_{m=1}^{M}E\left\{[\phi_{p,ii}*s_i(k)]^2\right\}}{\sum_{j=1}^{M}\sum_{i=1,i\neq j}^{M}E\left\{[\phi_{p,ji}*s_i(k)]^2\right\}}, \quad p=1,2,\cdots,P. \tag{43}$$

Finally, the overall average output SIR is found as

$$\mathrm{SIR}^{\mathrm{out}} \triangleq \frac{1}{P}\sum_{p=1}^{P}\mathrm{SIR}_p^{\mathrm{out}}. \tag{44}$$

To assess the quality of dereverberated speech signals, we employed the Itakura–Saito (IS) distortion measure [21], which is the ratio of the residual energies produced by the original speech when inverse filtered using the LP coefficients derived from the

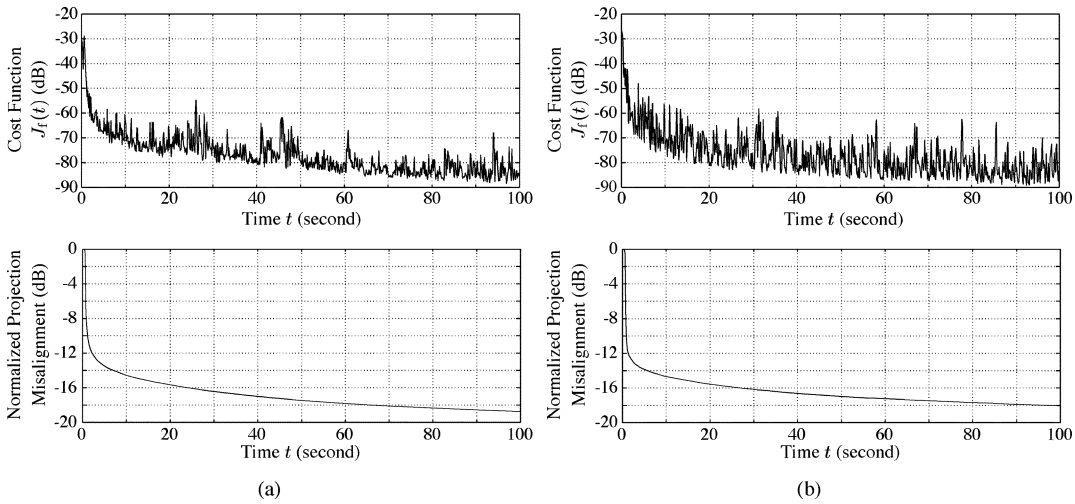



Fig. 5. Running average (1000 samples) of the cost function and normalized projection misalignment for blindly identifying the SIMO system corresponding to (a) source 1 and (b) source 2 with the UNMCFLMS algorithm in the varechoic chamber with 75% of panels open.

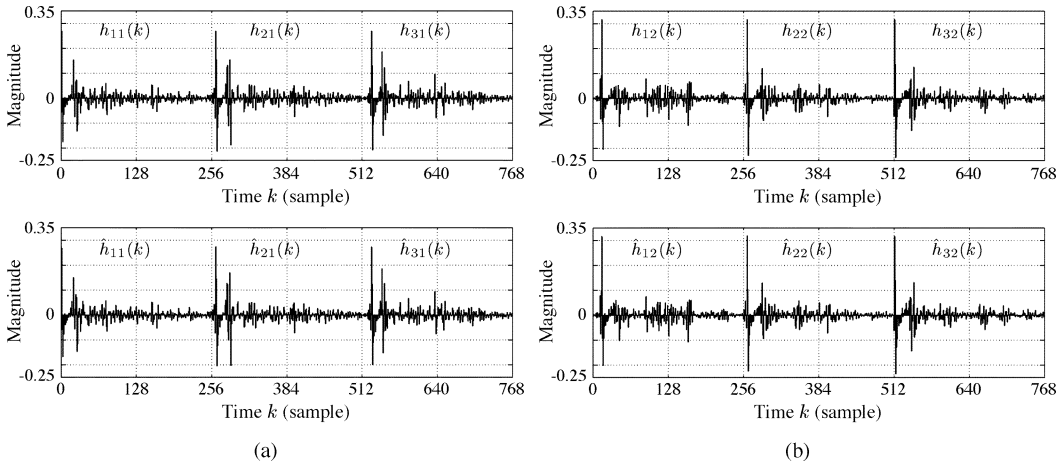



Fig. 6. Comparison of impulse responses between the actual channels and their estimates determined by using the UNMCFLMS algorithm in the varechoic chamber with 75% of panels open. Channels correspond to (a) source 1 and (b) source 2.

original and processed speech. Let $\boldsymbol{\alpha}_t$ and $\boldsymbol{\alpha}'_t$ be the LP coefficient vectors of an original speech signal frame $\mathbf{s}_t$ and the corresponding processed speech signal frame $\mathbf{s}'_t$ under examination, respectively. Denote $\mathbf{R}_{tt}$ as the Toeplitz autocorrelation matrix of the original speech signal. Then the IS measure is given as:

$$d_{\mathrm{IS},t} = \frac{\boldsymbol{\alpha}_t'^{T}\mathbf{R}_{tt}\boldsymbol{\alpha}_t'^{T}}{\boldsymbol{\alpha}_t^{T}\mathbf{R}_{tt}\boldsymbol{\alpha}_t^{T}} - 1. \tag{45}$$

Such a measure is calculated on a frame-by-frame basis. For the whole sequence of two speech signals, the mean IS measure is obtained by averaging $d_{\mathrm{IS},t}$ over all frames. According to [23], the IS measure exhibits a high correlation (0.59) with subjective judgments, suggesting that the IS distance is a good objective measure of speech quality. It was reported in [24] 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 speech codecs. Many experiments in speech recognition show that if the IS measure is less than about 0.1, the two spectra that we compare are perceptually nearly identical.

In our simulations, IS measures are calculated at different points (after source separation and after speech dereverberation) and with respect to every source. After source separation and for source $s_m$ $(m = 1, 2, \cdots, M)$, the IS measure is obtained by averaging the result of each one of $P$ SIMO outputs and is denoted by $d^{\mathrm{SS}}_{\mathrm{IS},s_m}$. After speech dereverberation, the final IS measure is denoted by $d^{\mathrm{SD}}_{\mathrm{IS},s_m}$.

### *B. Experimental Setup*

The simulations were conducted with the impulse responses measured in the varechoic chamber at Bell Labs [25]. A diagram of the floor plan layout is shown in Fig. 3. For convenience, positions in the floor plan are designated by $(x, y)$ coordinates with reference to the southwest corner and corresponding to meters along the (South, West) walls. The chamber measures $x = 6.7$ m wide by $y = 6.1$ m deep by $z = 2.9$ m high. It is a rectangular room with 368 electronically controlled panels that vary the acoustic absorption of the walls, floor, and ceiling [26]. 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. The panels are individually controlled so that the holes on one particular panel are either fully open (absorbing state) or fully closed (reflective state). Therefore, by varying the binary state of each panel in any combination, $2^{238}$ different room characteristics

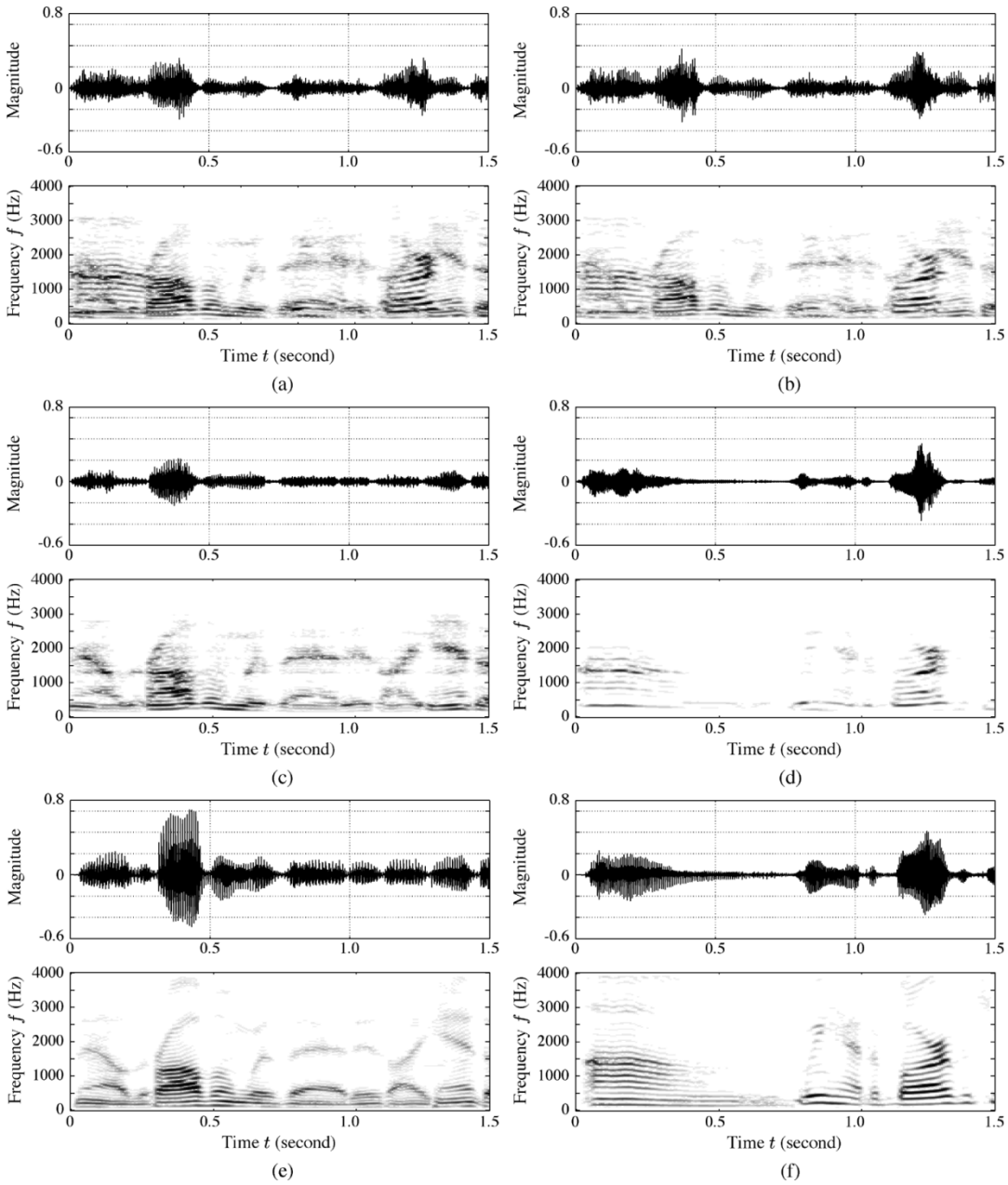



Fig. 7. Time sequence and spectrogram (30 Hz bandwidth) of (a) $x_1(k)$, (b) $x_2(k)$, (c) $y_{s1,1}(k)$, (d) $y_{s_2,1}(k)$, (e) $\hat{s}_1(k)$, and (f) $\hat{s}_2(k)$ for the experiment carried out in the varechoic chamber with 75% of panels open. This experiment used the UNMCFLMS algorithm for BCI.

can be simulated. In the database of channel impulse responses from [25], there are four panel configurations with 89%, 75%, 30%, and 0% of panels open, respectively corresponding to approximately 240, 310, 380, and 580 ms 60 dB reverberation time in the 20–4000 Hz band. All four configurations were used in this paper for evaluating performance of the proposed algorithm.

A linear microphone array which consists of 22 omnidirectional microphones was employed in the measurement and the spacing between adjacent microphones is about 10 cm. The array was mounted 1.4 m above the floor and parallel to the North wall at a distance of 50 cm. A loudspeaker was placed at 31 different pre-specified positions to measure the impulse response to each microphone. In the simulations, three microphones and two speaker positions, which form a $2 \times 3$ MIMO system, were chosen and their locations are shown in Fig. 3. Signals were sampled at 8 kHz and the original impulse response measurements have 4096 samples. In the cases of 89% and 75% panels open, energy in reverberation decays quickly with arrival time and we cut impulse responses at $L_h = 256$. When 30% or none of planes are open, we set $L_h = 512$. In terms of the two speakers, one male and the other female, the time sequence and spectrogram (30 Hz bandwidth) of their speech for the first 1.5 s are shown in Fig. 4. Silent periods were manually removed from the speech signals to make the BCI methods converge faster due to the reduced nonstationarity in the inputs and to make the average IS measures more meaningful with respect to speech only. This implies that in practice a voice activity detector needs to be used. After having source signals and channel impulse responses, we calculated microphone outputs by convolution.

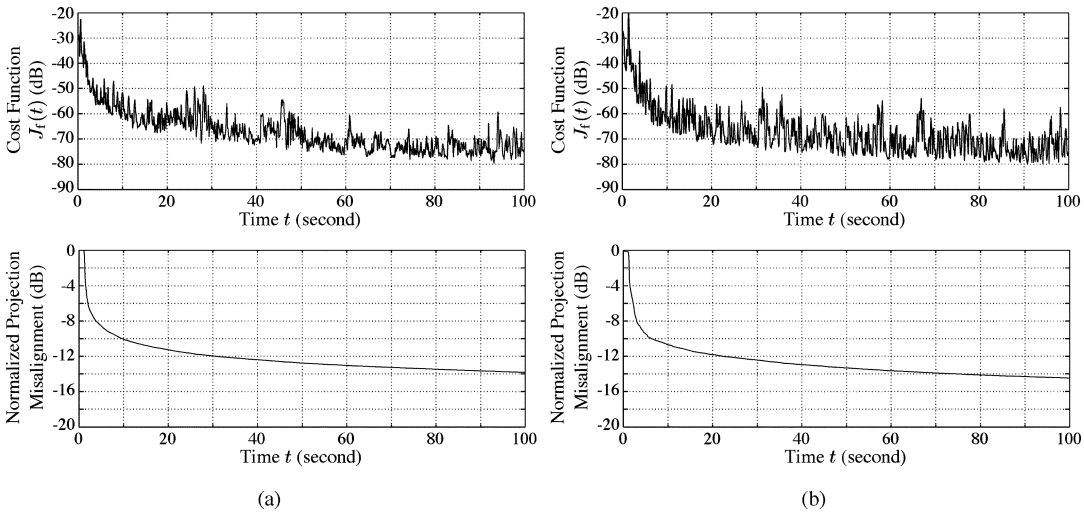



Fig. 8. Running average (1000 samples) of the cost function and normalized projection misalignment for blindly identifying the SIMO system corresponding to (a) source 1 and (b) source 2 with the UNMCFLMS algorithm in the varechoic chamber with all panels closed.

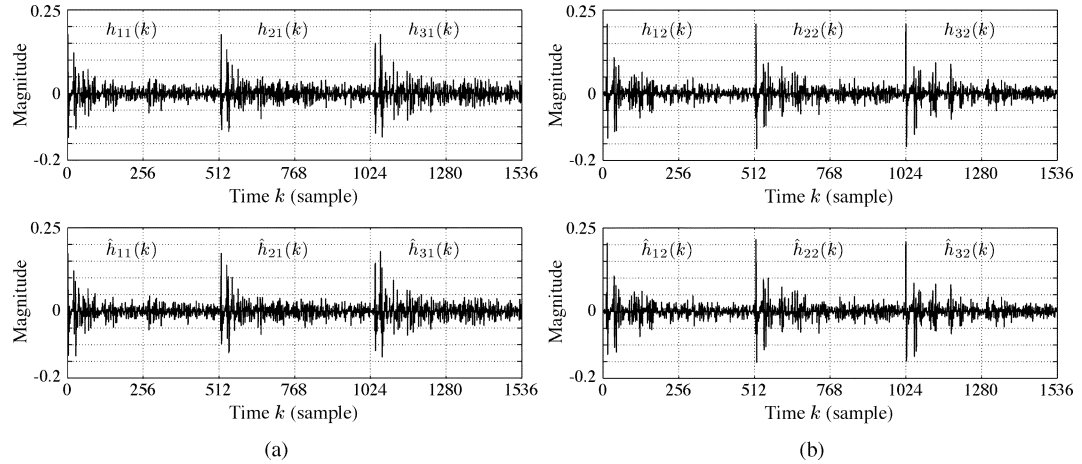



Fig. 9. Comparison of impulse responses between the actual channels and their estimates determined by using the UNMCFLMS algorithm in the varechoic chamber with all panels closed. Channels correspond to (a) source 1 and (b) source 2.

As we expected, the performance of the proposed source separation and speech dereverberation algorithm would be greatly affected by the accuracy of the blindly estimated channel impulse responses. In the simulations, both adaptive (the UNMCFLMS algorithm) and batch (the SVD-based algorithm) implementations were investigated [17]. For the batch method, the empirical spatial covariance matrix was obtained over the first 2500 and 3000 samples of the microphone captures for $L_h = 256$ and 512, respectively. In addition, additive noise was inserted at each microphone output at 75 dB signal-to-noise ratio (SNR). In experiments with the adaptive UNMCFLMS algorithm, no background noise was assumed. For source separation and speech dereverberation, speech signals of duration 10 s were utilized to assess the performance. The decision delay $d$ in (39) was fixed as $L_h - 1$.

### C. Experimental Results

Table I summarizes the experimental results for all four different room acoustics. Figs. 5–7 visualize what was observed in the experiment with 75% of panels open, and Figs. 8–10 with all panels closed.

Let us first examine the accuracy of the channel impulse responses blindly estimated by the adaptive and batch BCI algorithms. Comparing Figs. 5 and 8 reveals that the UNMCFLMS converges slower as $L_h$ increases. Given the same amount of microphone observations, the final projection misalignment error would be larger for the UNMCFLMS to identify a more reverberant SIMO system. Relatively, the batch method is more accurate and seems less sensitive to $L_h$. After it keeps collecting microphone outputs for only 0.375 s, the batch BCI method can produce a reliable channel estimate with less than $-29$ dB NPM for SIMO systems with long channels of length $L_h = 512$. However, performing SVD of a $3L_h \times 3L_h$ matrix in these simulations is too computationally intensive to be accomplished in real time by a commercial processor in the foreseeable near future. The reason why we carried out experiments with the batch BCI implementation and present here the results is to get an idea about what is the best possible per-

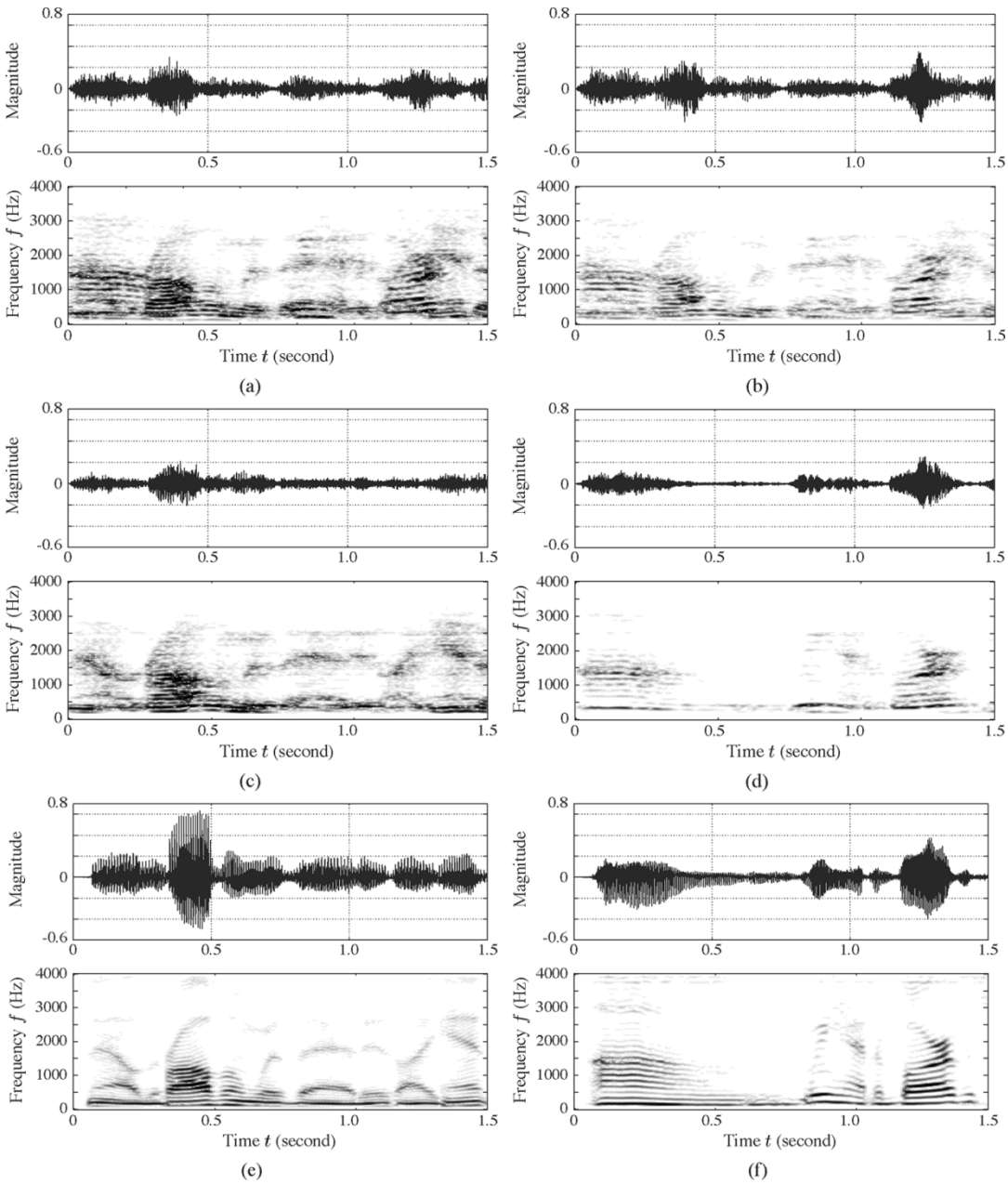



Fig. 10. Time sequence and spectrogram (30 Hz bandwidth) of (a) $x_1(k)$, (b) $x_2(k)$, (c) $y_{s1,1}(k)$, (d) $y_{s_2,1}(k)$, (e) $\hat{s}_1(k)$, and (f) $\hat{s}_2(k)$ for the experiment carried out in the varechoic chamber with all panels closed. This experiment used the UNMCFLMS algorithm for BCI.

formance of the proposed blind source separation and speech dereverberation approach.

Figs. 7 and 10 illustrate how spatial interference and temporal echoes are separated and how the two speech signals are finally recovered. Examining these figures together with the data in Table I, we see that the output SIR's are very high (at least 44 dB) after the conversion of the MIMO system into several SIMO systems. But meanwhile the separated signals sound more echoic and have more distortion, resulting in large IS measures (greater than 1.9) and vague harmonics in periods of voiced speech on the narrow-band spectrograms. After dereverberation, the speech signals are satisfactorily recovered though delayed [clearly seen from time sequences of the recovered signals $\hat{s}_1(k)$ and $\hat{s}_2(k)$ in these figures] with a very low IS measure (less than 0.2 even in the worst case). As explained before, speech with such an amount of distortion would not change its perceptual quality with respect to either humans or an speech recognition system. Therefore, the simulations show some promise of successful use of the proposed algorithm in prospect speech processing systems.

## VII. Conclusions

Room reverberation makes blind separation of speech sources from their convolutive mixtures a very difficult problem in a real reverberant environment. Existing blind source separation methods maximize solely the signal-to-interference ratio and possibly cause high distortion in their

separated signals, which is neither pleasing to a listener nor can be used in following speech processing systems. We demonstrated in this paper that spatial interference from competing sources and temporal echoes due to room reverberation can be perfectly separated by converting a MIMO system into several interference-free SIMO systems. The channel matrices of these SIMO systems are irreducible given that the channels from the same source in the MIMO system share no common zeros. For these SIMO systems, the original speech can be easily restored by using the Bezout theorem. If some channels share common zeros, we deduced what might be the best possible solution for speech dereverberation. This derivation led to the proposal of a novel sequential source separation and speech dereverberation algorithm. We conducted experiments using real impulse responses measured in the varechoic chamber at Bell Labs. The results demonstrated the success and robustness of the proposed algorithm in highly reverberant acoustic environments.

## REFERENCES


[1] S. Makeig, A. Bell, T.-P. Jung, and T. J. Sejnowski *et al.*, "Independent component analysis in electro-encephalographic data," in *Advances in Neural Information Processing Systems*, M. Mozer *et al.*, Eds. Cambridge, MA: MIT Press, 1996, pp. 145–151.

[2] A. Cichocki, W. Kasprzak, and S. Amari, "Neural network approach to blind separation and enhancement of images," *Signal Process.*, vol. I, pp. 579–582, Sep. 1996.

[3] F. Ehlers and H. G. Schuster, "Blind separation of convolutive mixtures and an application in automatic speech recognition in a noisy environment," *IEEE Trans. Signal Processing*, vol. 45, pp. 2608–2612, Oct. 1997.

[4] M. Torlak, L. K. Hansen, and G. Xu, "A fast blind source separation for digital wireless applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 6, 1998, pp. 3305–3308.

[5] P. Comon, "Independent component analysis: a new concept," *Signal Process.*, vol. 36, pp. 287–314, Apr. 1994.

[6] J.-F. Cardoso and P. Comon, "Independent component analysis, a survey of some algebraic methods," in *Proc. IEEE Int. Symp. Circuits and Systems*, vol. 2, 1996, pp. 93–96.

[7] K. Torkkola, "Blind separation of convolved sources based on information maximization," in *Proc. IEEE Workshop on Neural Networks for Signal Processing*, 1996, pp. 423–432.

[8] C. Servière, "Feasibility of source separation in frequency domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, 1998, pp. 2085–2088.

[9] L. Parra and C. Spence, "Convolutive blind separation of nonstationary sources," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 320–327, May 2000.

[10] M. Z. Ikram and D. R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 2000, pp. 1041–1044.

[11] T. Takatani, T. Nishikawa, H. Saruwatari, and K. Shikano, "SIMO-model-based independent component analysis for high-fidelity blind separation of acoustic signals," in *Proc. 4th Int. Symp. Independent Component Analysis and Blind Signal Separation*, 2003, pp. 993–998.

[12] Y. Huang, J. Benesty, and J. Chen, "Separating ISI and CCI in a two-step FIR Bezout equalizer for MIMO systems of frequency-selective channels," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2004.

[13] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation*, 2001, pp. 722–727.

[14] R. F. Brcich, A. M. Zoubir, and P. Pelin, "Detection of sources using bootstrap techniques," *IEEE Trans. Signal Processing*, vol. 50, no. 2, pp. 206–215, Nov. 2002.

[15] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Trans. Signal Processing*, vol. 43, pp. 2982–2993, Dec. 1995.

[16] Y. Huang and J. Benesty, "Adaptive multi-channel least mean square and Newton algorithms for blind channel identification," *Signal Process.*, vol. 82, pp. 1127–1138, Aug. 2002.

[17] ——, "A class of frequency-domain adaptive approaches to blind multi-channel identification," *IEEE Trans. Signal Processing*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[18] P. P. Vaidyanathan, *Multirate Systems and Filter Bank*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

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

[20] D. R. Morgan, J. Benesty, and M. M. Sondhi, "On the evaluation of estimated impulse responses," *IEEE Signal Processing Lett.*, vol. 5, no. 7, pp. 174–176, Jul. 1998.

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

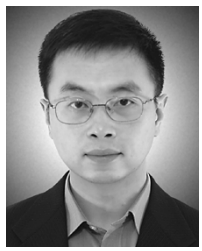
[22] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

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

[24] 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.

[25] A. Härmä, Acoustic Measurement Data From the Varechoic Chamber, Tech. Memo., Agere Systems, Nov. 2001.

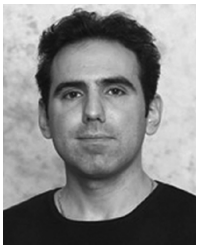
[26] W. C. Ward, G. W. Elko, R. A. Kubli, and W. C. McDougald, "The new varechoic chamber at AT&T Bell Labs," in *Proc. Wallance Clement Sabine Centennial Symp.*, 1994, pp. 343–346.


**Yiteng (Arden) Huang** (S'97–M'01) received the B.S. degree from Tsinghua University in 1994 and 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.

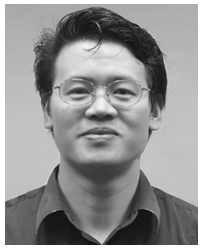
Dr. Huang is currently an associate editor of the *EURASIP Journal on Applied Signal Processing.* He served as an associate editor for the IEEE SIGNAL PROCESSING LETTERS from 2002 to 2005. He was a technical co-chair of the 2005 Joint Workshop on Hands-Free Speech Communication and Microphone Array. He is a co-editor/co-author 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). 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.



**Jacob Benesty** (M'92–SM'04) was born in Marrakech, Morocco, in 1963. He received the Masters degree in microwaves from Pierre and Marie Curie University, France, in 1987, and the Ph.D. degree in control and signal processing from Orsay University, France, in April 1991. During his Ph.D. studies (from November 1989 to April 1991), he worked on adaptive filters and fast algorithms.

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 is a member of the editorial board of the *Journal on Applied Signal Processing*.

Dr. Benesty received the 2001 Best Paper Award from the IEEE Signal Processing Society. He is currently an Associate Editor of the *EURASIP Journal on Applied Signal Processing*. He was the co-chair of the 1999 International Workshop on Acoustic Echo and Noise Control. He co-authored the book *Advances in Network and Acoustic Echo Cancellation* (Berlin, Germany: Springer-Verlag, 2001). He is also a co-editor/co-author of the books *Speech Enhancement* (Berlin: 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).



**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 in 1993 and 1995 respectively, and the Ph.D. degree in pattern recognition and intelligence control from the Chinese Academy of Sciences 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 ATR Spoken Language Translation Research Laboratories, Kyoto, where he conducted research in robust speech recognition and speech enhancement. He joined Bell Laboratories 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 co-editor/co-author 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.