Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Signal Processing 86 (2006) 1278–1295

SIGNAL PROCESSING

www.elsevier.com/locate/sigpro

Identification of acoustic MIMO systems: Challenges and opportunities

Yiteng Huang[a,*], Jacob Benesty[b], Jingdong Chen[a]

[a]*Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974, USA*
[b]*Université du Québec, INRS-EMT, Montréal, Québec, H5A 1K6, Canada*

Received 1 January 2005; received in revised form 24 February 2005; accepted 15 June 2005
Available online 19 October 2005

Abstract

A systematic overview of acoustic MIMO identification algorithms is presented in this paper, which explains why we believe that the mother of all challenges in acoustic signal processing is how to accurately estimate room acoustic impulse responses in real time. From non-blind to blind methods with respect to both single-channel and multichannel acoustic systems, we scan the state of the art in acoustic channel identification technologies and outline fundamental challenges that still are waiting for breakthroughs. A number of acoustic signal processing problems are briefly reviewed and their connections to acoustic MIMO identification are clarified. Several successful real-time systems based on acoustic MIMO identification are discussed to confirm the value of this technique.

Keywords: Acoustic MIMO systems; Blind channel identification; Frequency-domain adaptive filters; Source separation; Speech dereverberation

1. Introduction

From analog to digital signals, from narrowband to broadband speech, from wireline to wireless terminals, and from circuit-switched to packet-switched networks, there have been tremendous advances in voice telecommunication technology ever since Alexander Graham Bell invented the telephone in 1876. However, conversation and collaboration between people over long distance who use today's audio communication technology are still unnatural and even clumsy. The distraction of holding a superfluous device such as a close-talk microphone and the lack of sensibility of remote speaking environments lead to diminished interaction and productivity, and eventually cause customer dissatisfaction. It is no longer a luxury but truly a rational demand to create a lifelike voice communication mode that gives the involved people the impression of being in the same acoustic environment, which is referred to as "*immersive experience*" in the multimedia communication literature [1]. To achieve this goal, the acoustic interface at the transmitting end needs to acquire high-fidelity speech and sound while synchronously recording the source location information, while still allowing the users to move freely without holding or wearing a microphone. Consequently, problems that must be addressed include, but are

*Corresponding author.
E-mail address: arden@research.bell-labs.com (Y. Huang).

0165-1684/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2005.06.023

not limited to, (multichannel) echo cancellation, (blind) source separation, source localization, noise reduction, and speech dereverberation. At the receiving end, a sound field is rendered such that the local acoustic environment is masked, but the remote (or virtual) environment is reproduced (or constructed) in the human's perception. All of these problems are under active research, wherein acoustic signal processing plays a central role.

Acoustic waves studied in multimedia communications are simply pressure disturbances propagating in the air. They carry information of the sound source and their energy is radiated spherically from the origin, i.e., the location of the sound source. The governing law of physics in this radiation process is the natural fall-off of the signal level as a function of distance from the origin. As a rule of thumb, the loss is 6 dB for doubling the distance. This phenomenon makes distant acquisition of a speech signal vulnerable to interference from other concurrent speech sources and ambient noise. Moreover, in an enclosure, acoustic waves are reflected possibly many times by boundaries before they reach a microphone, leading to distortion observed in microphone signals. Acoustic signal processing helps extract and interpret the information in these distorted microphone signals. There is no doubt that acoustic signal processing problems in these scenarios are difficult, and we believe that the difficulties are essentially attributable to the deficiency of knowledge by an acoustic signal processing algorithm about the characteristics of its surrounding acoustic environment. In other words, the core difficulty of most of acoustic signal processing problems is associated with our incapability of dynamically identifying an acoustic system.

An immersive acoustic communication system by its nature involves multiple sound sources (including loudspeakers) and multiple microphones [2], fitting well into the multiple-input multiple-output (MIMO) structure. Over the last several years, the MIMO model has been extensively investigated in wireless communications since a multiple-antenna system holds the promise of much higher spectral efficiencies for wireless channels [3,4]. Although wireless and acoustic channels have many things in common (e.g., time varying, frequency selective), MIMO acoustic systems are substantially different from that of wireless communications. As opposed to communication receivers, the human ear has an extremely wide dynamic range and is much more sensitive to weak tails of the channel impulse responses. As a result, the length of acoustic channel impulse response models is significantly greater than that in wireless communications. Filter lengths of thousands of samples are not uncommon in MIMO acoustic systems while wireless impulse responses consist of usually not more than a few of taps. In addition, since communication systems can use a pilot signal for channel identification, such a problem has never been an obstacle to developing practical wireless MIMO communication systems. But, in acoustics, we seek techniques for human talkers. The source signals in acoustic systems are random and their statistics are considerably different from that in wireless communications. The speech signal is neither stationary nor white, and does not form a known set of signal alphabet as in wireless communications, making the identification of a MIMO system apparently much more challenging.

The methods for identifying a MIMO system can be broadly dichotomized into two classes: the class of non-blind methods and the class of blind methods, depending on the availability of the knowledge of source signals. While non-blind system identification algorithms with known source signals are not easy to develop (multichannel echo cancellation is a good example), blind MIMO identification is much more difficult although not completely insolvable. But for many acoustic signal processing problems, particularly in the design of an immersive acoustic interface, source signals are inaccessible in practice and a blind system identification algorithm has to be developed. Clearly, the challenges are great, but so are opportunities after the identification of an acoustic system. In this paper, we would like to shed some light onto the question: what do we expect if the identification of a MIMO acoustic system can be surmounted? We will provide an overview of the technology for identifying acoustic MIMO systems and explain the impact of any possible progress in this technology on the development of other acoustic signal processing applications in the future.

The rest of the paper is organized as follows. Section 2 briefly presents the mathematical models for describing an acoustic system and Section 3 summarizes both non-blind and blind channel identification algorithms. In Section 4, we explain how to approach many other acoustic signal processing problems after the MIMO system is identified. Five typical applications are discussed, namely, acoustic echo cancellation, time delay

estimation, cross-talk cancellation, source separation, and speech dereverberation. In Section 5, we review several successful real-time, channel-identification-based acoustic signal processing systems that we developed in the past. Finally, we draw our conclusions in Section 6.

2. Signal models for MIMO acoustics

System modeling is fundamental to signal-processing and control theories, and so is to acoustic applications. Creating a mathematical representation of the acoustic environment helps us to gain a better understanding of what is going on and enables better visualization of the main elements of an acoustic signal processing problem. Certainly it also forms a basis for discussion of various acoustic problems using the same convenient language of mathematics used in the rest of this paper.

Typically an acoustic environment is abstracted into a linear system merely because a linear model is simple enough to allow comprehensive analysis. An acoustic channel is expressed by an FIR filter. Four models for describing acoustic systems are depicted as follows:

I. *Single-input single-output* (*SISO*) *system* [Fig. 1(a)]. The output signal is given by

$$x(k) = h * s(k) + b(k), \tag{1}$$

where h is the channel impulse response, the symbol $*$ denotes the linear convolution operator, $s(k)$ is the source signal at time k, and $b(k)$ is the additive noise at the output. The channel can be time invariant or time varying, depending on the application. In vector/matrix form, the SISO signal model (1) is written as:

$$x(k) = \mathbf{h}^{\mathrm{T}}\mathbf{s}(k) + b(k), \tag{2}$$

where

$$\mathbf{h} = [h_0 \;\; h_1 \;\; \cdots \;\; h_{L-1}]^{\mathrm{T}},$$

$$\mathbf{s}(k) = [s(k) \;\; s(k-1) \;\; \cdots \;\; s(k-L+1)]^{\mathrm{T}},$$

$(\cdot)^{\mathrm{T}}$ denotes the transpose of a matrix or a vector, and L is the channel length.

II. *Single-input multiple-output* (*SIMO*) *system* [Fig. 1(b)]. In this system, there are N outputs from the same sound source as input and the nth output is expressed as:

$$x_n(k) = \mathbf{h}_n^{\mathrm{T}}\mathbf{s}(k) + b_n(k), \quad n = 1, 2, \ldots, N, \tag{3}$$

where $x_n(k)$, $\mathbf{h}_n$, and $b_n(k)$ are defined in a similar way to those in (2), and L is the length of the longest channel impulse response in this SIMO system.

III. *Multiple-input single-output* (*MISO*) *system* [Fig. 1(c)]. There are M sound sources but only one output:

$$\begin{aligned} x(k) &= \sum_{m=1}^{M} \mathbf{h}_m^{\mathrm{T}}\mathbf{s}_m(k) + b(k), \\ &= \mathbf{h}^{\mathrm{T}}\mathbf{s}(k) + b(k), \end{aligned} \tag{4}$$

where

$$\mathbf{h}_m = [h_{m,0} \;\; h_{m,1} \;\; \cdots \;\; h_{m,L-1}]^{\mathrm{T}},$$

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

$$\mathbf{h} = [\mathbf{h}_1^{\mathrm{T}} \;\; \mathbf{h}_2^{\mathrm{T}} \;\; \cdots \;\; \mathbf{h}_M^{\mathrm{T}}]^{\mathrm{T}},$$

$$\mathbf{s}(k) = [\mathbf{s}_1^{\mathrm{T}}(k) \;\; \mathbf{s}_2^{\mathrm{T}}(k) \;\; \cdots \;\; \mathbf{s}_M^{\mathrm{T}}(k)]^{\mathrm{T}}.$$

IV. *Multiple-input multiple-output* (*MIMO*) *system* [Fig. 1(d)]. A MIMO system with M inputs and N outputs is referred to as an $M \times N$ system. At time k, we have

$$\mathbf{x}(k) = \mathbf{H}\mathbf{s}(k) + \mathbf{b}(k), \tag{5}$$

where

$$\mathbf{x}(k) = [x_1(k) \;\; x_2(k) \;\; \cdots \;\; x_N(k)]^{\mathrm{T}},$$

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}_{11}^{\mathrm{T}} & \mathbf{h}_{12}^{\mathrm{T}} & \cdots & \mathbf{h}_{1M}^{\mathrm{T}} \\ \mathbf{h}_{21}^{\mathrm{T}} & \mathbf{h}_{22}^{\mathrm{T}} & \cdots & \mathbf{h}_{2M}^{\mathrm{T}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{h}_{N1}^{\mathrm{T}} & \mathbf{h}_{N2}^{\mathrm{T}} & \cdots & \mathbf{h}_{NM}^{\mathrm{T}} \end{bmatrix}_{N \times ML},$$

$$\mathbf{h}_{nm} = [h_{nm,0} \;\; h_{nm,1} \;\; \cdots \;\; h_{nm,L-1}]^{\mathrm{T}},$$

$$\mathbf{b}(k) = [b_1(k) \;\; b_2(k) \;\; \cdots \;\; b_N(k)]^{\mathrm{T}},$$

h_{nm} $(n = 1, 2, \ldots, N, \;\; m = 1, 2, \ldots, M)$ is the impulse response of the channel from input m to output n, and $\mathbf{s}(k)$ is defined similarly to that in (4).

Clearly the MIMO system is a more general model and all other three systems can be treated as special examples of a MIMO system. But the difference among these models that looks trivial here will lead to great divergence in difficulty and complexity of their identification algorithms as will be shown in the following sections.

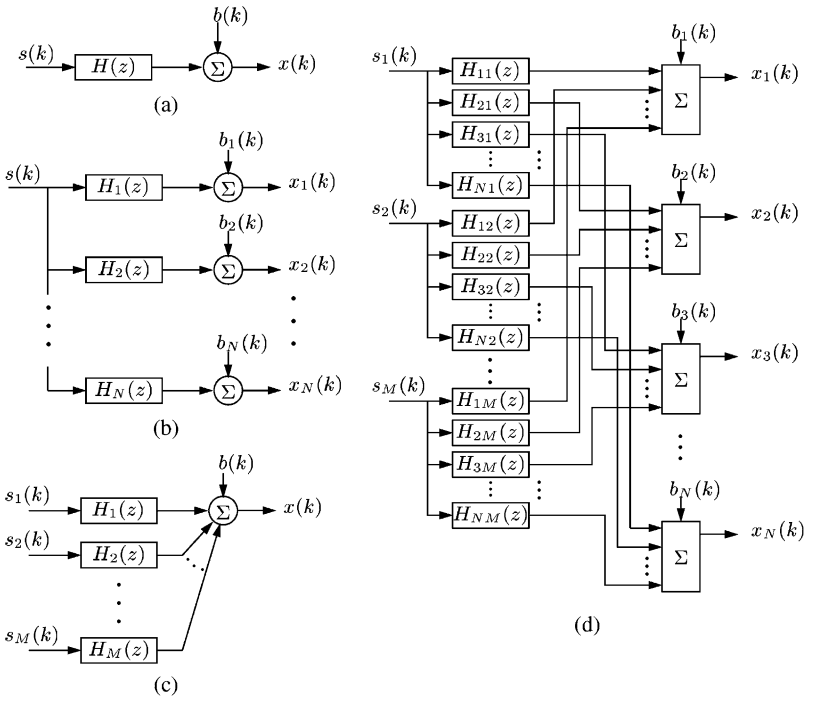

Fig. 1. Four mathematical models for describing acoustic systems: (a) single-input single-output (SISO), (b) single-input multiple-output (SIMO), (c) multiple-input single-output (MISO), and (d) multiple-input multiple-output (MIMO) systems.

3. Channel identification

In this section, we will briefly review channel identification technologies. We begin with traditional non-blind methods and then discuss more advanced blind algorithms.

3.1. Non-blind methods

System identification given the reference signal(s) is one of the oldest problems in signal processing. Its theory has been well developed but the adaptive implementation technique continues progressing. In the following, we will describe methods for identifying single-input (SISO and SIMO) and multiple-input (MISO and MIMO) systems separately.

I. *Single-input systems*
Visibly, a SIMO system can be decomposed into M SISO systems and the identification of each SISO system is independent of each other. For a SISO system (2), we assume that the channel is time-invariant during the process of identification and the channel length L is known. Then we define the error signal at time k as follows;

$$e(k) \triangleq x(k) - \hat{x}(k) = x(k) - \hat{\mathbf{h}}^{\mathrm{T}}\mathbf{s}(k), \tag{6}$$

where

$$\hat{\mathbf{h}} = [\hat{h}_0 \;\; \hat{h}_1 \;\; \cdots \;\; \hat{h}_{L-1}]^{\mathrm{T}}$$

is the model filter. The mean-square error criterion with respect to the model filter is given by

$$J(\hat{\mathbf{h}}) \triangleq E\{e^2(k)\}, \tag{7}$$

where $E\{\cdot\}$ denotes mathematical expectation. The minimization of (7) leads to the well-known Wiener–Hopf equation [5]:

$$\mathbf{R}_{ss}\hat{\mathbf{h}} = \mathbf{r}_{sx}, \tag{8}$$

where $\mathbf{R}_{ss} = E\{\mathbf{s}(k)\mathbf{s}^{\mathrm{T}}(k)\}$ is the covariance matrix of the input signal of size $L \times L$, and $\mathbf{r}_{sx} = E\{\mathbf{s}(k)x(k)\}$ is the cross-correlation vector of size $L \times 1$ between the input and output. From (8), we see that if $\mathbf{R}_{ss}$ is non-singular and therefore invertible (which is usually true for speech signals), the impulse response of a SISO (or SIMO) system can be easily determined as $\hat{\mathbf{h}} = \mathbf{R}_{ss}^{-1}\mathbf{r}_{sx}$.

II. *Multiple-input systems*
The identification of a $M \times N$ MIMO system can also be decomposed into N independent MISO identification, but this decomposition is

not as obvious as that for single-input systems. Here we will develop the identification with respect to a MIMO system and then build its connection with that for MISO systems.

For the $M \times N$ MIMO system given in (5), we define the error signal at time k as

$$\mathbf{e}(k) \triangleq \mathbf{x}(k) - \hat{\mathbf{x}}(k) = \mathbf{x}(k) - \hat{\mathbf{H}}\mathbf{s}(k), \tag{9}$$

where $\hat{\mathbf{H}}$ is the channel matrix for the model system defined similarly to $\mathbf{H}$ in (5). Having written the error signal, we now define the mean-square error criterion with respect to the model system:

$$J(\hat{\mathbf{H}}) \triangleq E\{\mathbf{e}^{\mathrm{T}}(k)\mathbf{e}(k)\}. \tag{10}$$

Taking the derivative of (10) with respect to $\hat{\mathbf{H}}$

$$\frac{\partial J(\hat{\mathbf{H}})}{\partial \hat{\mathbf{H}}} = -2E\{\mathbf{e}(k)\mathbf{s}^{\mathrm{T}}(k)\} \tag{11}$$

and equating the result to zero produces the multichannel Wiener–Hopf equations:

$$\hat{\mathbf{H}}\mathbf{R}_{ss} = \mathbf{R}_{xs}, \tag{12}$$

where $\mathbf{R}_{ss} = E\{\mathbf{s}(k)\mathbf{s}^{\mathrm{T}}(k)\}$ is of size $ML \times ML$ and $\mathbf{R}_{xs} = E\{\mathbf{x}(k)\mathbf{s}^{\mathrm{T}}(k)\}$ is of size $N \times ML$.

It can easily be seen that the multichannel Wiener–Hopf (12) can be written as N independent Wiener–Hopf equations, each one corresponding to a MISO system:

$$\hat{\mathbf{h}}_{n:}^{\mathrm{T}}\mathbf{R}_{ss} = \mathbf{r}_{x_n s}^{\mathrm{T}}, \quad n = 1, 2, \ldots, N, \tag{13}$$

where $\hat{\mathbf{h}}_{n:}^{\mathrm{T}}$ is the nth row of matrix $\hat{\mathbf{H}}$ and $\mathbf{r}_{x_n s} = E\{x_n(k)\mathbf{s}(k)\}$ is the cross-correlation vector between the nth output and the input (which is also the nth row of matrix $\mathbf{R}_{xs}$). This result implies that minimizing $J(\hat{\mathbf{H}})$ is equivalent to minimizing each $E\{e_n^2(k)\}$ independently. This means that the identification of the MISO subsystem at one MIMO output is completely independent of the others and we can decompose the identification of a $M \times N$ MIMO system into N independent tasks of MISO identification.

As we explained previously for single-input systems, the input signal's autocorrelation matrix $\mathbf{R}_{ss}$ needs to be non-singular so that a MISO or MIMO system can be identified. While for single-input acoustic systems this condition can be met in most practical cases, input signals are sometimes correlated or even perfectly coherent in a MISO or MIMO system. For example in a multichannel echo cancellation system, the loudspeaker signals are obtained from a common source in the transmitting room:

$$s_n(k) = g_n * u(k), \quad n = 1, 2, \ldots, N, \tag{14}$$

where g_n is the impulse response between the source $u(k)$ and the loudspeaker signal $s_n(k)$. It can be shown that

$$\mathbf{s}(k) = \mathbf{G}_{\mathrm{c}}\mathbf{u}(k), \tag{15}$$

where

$$\mathbf{G}_{\mathrm{c}} = [\mathbf{G}_{\mathrm{c},1}^{\mathrm{T}} \ \mathbf{G}_{\mathrm{c},2}^{\mathrm{T}} \ \cdots \ \mathbf{G}_{\mathrm{c},M}^{\mathrm{T}}]_{ML\times(2L-1)}^{\mathrm{T}},$$

$$\mathbf{G}_{\mathrm{c},m} = \begin{bmatrix} g_{m,0} & g_{m,1} & \cdots & g_{m,L-1} & 0 & \cdots & 0 \\ 0 & g_{m,0} & \cdots & g_{m,L-2} & g_{m,L-1} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & g_{m,0} & g_{m,1} & \cdots & g_{m,L-1} \end{bmatrix}_{L\times(2L-1)},$$

$$m = 1, 2, \ldots, M,$$

are convolutive matrices (subscript c), and

$$\mathbf{u}(k) = [u(k) \ u(k-1) \ \cdots \ u(k-L+1) \ \cdots \ u(k-2L+2)]_{(2L-1)\times 1}^{\mathrm{T}}.$$

Therefore, the autocorrelation matrix of the loudspeaker signals is given by

$$\mathbf{R}_{ss} = E\{\mathbf{s}(k)\mathbf{s}^{\mathrm{T}}(k)\} = \mathbf{G}_{\mathrm{c}}E\{\mathbf{u}(k)\mathbf{u}^{\mathrm{T}}(k)\}\mathbf{G}_{\mathrm{c}}^{\mathrm{T}} = \mathbf{G}_{\mathrm{c}}\mathbf{R}_{uu}\mathbf{G}_{\mathrm{c}}^{\mathrm{T}}, \tag{16}$$

and its rank is bounded by

$$\mathrm{rank}(\mathbf{R}_{ss}) \leqslant \mathrm{rank}(\mathbf{R}_{uu}) \leqslant 2L - 1. \tag{17}$$

As a result, $\mathbf{R}_{ss}$ is singular and the multiple-input system cannot be uniquely identified. Since the dimension of $\mathbf{R}_{ss}$'s null space depends on the number of inputs and is equal to $ML-(2L-1) = (M-2)L+1$, this problem (referred

to as the non-uniqueness problem in multichannel echo cancellation algorithms) becomes worse as M increases.

3.2. Blind multichannel identification

The innovative idea of identifying a system without reference signals was first proposed by Sato in [6]. Early studies of blind channel identification and equalization focused primarily on higher (than second) order statistics (HOS) based methods. Because HOS cannot be accurately computed from a small number of observations, slow convergence is the critical drawback of all existing HOS methods. In addition, a cost function based on the HOS is barely concave and an HOS algorithm can be misled to a local minimum by corrupting noise in the observations. Therefore, after it was recognized that the problem can be solved in the light of only second-order statistics of system outputs [7], the focus of the blind channel identification research has shifted to SOS methods. Here we will briefly review only SOS algorithms that are applicable to real-time acoustic signal processing systems.

Using SOS to blindly identify a system requires that the number of outputs would be greater than the number of inputs. As a result, our discussion is with respect to only SIMO and MIMO systems.

I. *SIMO systems*

A SIMO system can be blindly identified using only SOS of the system's output if the following two conditions (one on the channel diversity and the other on the input signals) are met [8]:

1. The polynomials formed from $\mathbf{h}_n$ $(n = 1,2,\ldots,N)$ are co-prime, i.e., the channel transfer functions $H_n(z)$ do not share any common zeros;
2. The autocorrelation matrix $\mathbf{R}_{ss} = E\{\mathbf{s}(k)\mathbf{s}^{\mathrm{T}}(k)\}$ of the input signal is of full rank (such that the SIMO system can be fully excited).

There are many ways to approach the principle of blind multichannel identification. Presented here is one that we used in our own research. For a SIMO system as described in Section 2–II, the vector of channel impulse responses lies in the null space of the cross-correlation-like matrix of system outputs [9]

$$\mathbf{R}_{x+}\mathbf{h} = \mathbf{0}, \tag{18}$$

where

$$\mathbf{R}_{x+} = \begin{bmatrix} \sum_{n\neq 1}\mathbf{R}_{x_n x_n} & -\mathbf{R}_{x_2 x_1} & \cdots & -\mathbf{R}_{x_N x_1} \\ -\mathbf{R}_{x_1 x_2} & \sum_{n\neq 2}\mathbf{R}_{x_n x_n} & \cdots & -\mathbf{R}_{x_N x_2} \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{R}_{x_1 x_N} & -\mathbf{R}_{x_2 x_N} & \cdots & \sum_{n\neq N}\mathbf{R}_{x_n x_n} \end{bmatrix},$$

$$\mathbf{h} = [\mathbf{h}_1^{\mathrm{T}} \ \mathbf{h}_2^{\mathrm{T}} \ \cdots \ \mathbf{h}_N^{\mathrm{T}}]^{\mathrm{T}},$$

$$\mathbf{R}_{x_i x_j} = E\{\mathbf{x}_i(k)\mathbf{x}_j^{\mathrm{T}}(k)\}, \quad i,j = 1,2,\ldots,N, \text{ and}$$

$$\mathbf{x}_n(k) = [x_n(k) \ x_n(k-1) \ \cdots \ x_n(k-L+1)]^{\mathrm{T}}, \quad n = 1,2,\ldots,N.$$

If the SIMO system is blindly identifiable, the matrix $\mathbf{R}_{x+}$ is rank deficient by 1 (in the absence of noise) and the channel impulse responses can be uniquely determined. When additive noise is present, $\mathbf{h}$ would be the eigenvector of $\mathbf{R}_{x+}$ corresponding to its smallest eigenvalue (here we assume that the noise is incoherent or uncorrelated and weaker than source signals). Note that the estimated channel impulse response vector is aligned to the true one, but up to a scale.

A simple way to develop an adaptive implementation is to take advantage of the cross relations among the outputs [10]. By following the fact that

$$x_i * h_j = s * h_i * h_j = x_j * h_i, \quad i,j = 1,2,\ldots,N, \ i\neq j, \tag{19}$$

we have, in the absence of noise, the following cross relation at time k:

$$\mathbf{x}_i^{\mathrm{T}}(k)\mathbf{h}_j = \mathbf{x}_j^{\mathrm{T}}(k)\mathbf{h}_i, \quad i,j = 1,2,\ldots,N, \ i\neq j. \tag{20}$$

When noise is present and/or the estimate of channel impulse responses is deviated from the true value, an a priori error signal is produced:

$$e_{ij}(k+1) = \mathbf{x}_i^{\mathrm{T}}(k+1)\hat{\mathbf{h}}_j(k) - \mathbf{x}_j^{\mathrm{T}}(k+1)\hat{\mathbf{h}}_i(k), \quad i,j = 1,2,\ldots,N, \tag{21}$$

where $\hat{\mathbf{h}}_i(k)$ is the model filter for the ith channel at time k. In order to avoid the trivial estimate of all zero elements, a unit-norm constraint is imposed on

$$\hat{\mathbf{h}}(k) = [\hat{\mathbf{h}}_1^{\mathrm{T}}(\mathbf{k}) \ \hat{\mathbf{h}}_2^{\mathrm{T}}(\mathbf{k}) \ \cdots \ \hat{\mathbf{h}}_N^{\mathrm{T}}(\mathbf{k})]^{\mathrm{T}},$$

leading to the normalized error signal

$$\varepsilon_{ij}(k+1) = e_{ij}(k+1)/\|\hat{\mathbf{h}}(k)\|.$$

Accordingly, the cost function is formulated as:

$$J(k+1) \triangleq \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \varepsilon_{ij}^2(k+1), \tag{22}$$

and the update equation of the multichannel LMS (MCLMS) algorithm is deduced as follows [10]:

$$\hat{\mathbf{h}}(k+1) = \hat{\mathbf{h}}(k) - \mu \nabla J(k+1), \tag{23}$$

where μ is a small positive step size,

$$\nabla J(k+1) = \frac{\partial J(k+1)}{\partial \hat{\mathbf{h}}(k)} = \frac{2[\tilde{\mathbf{R}}_{x+}(k+1)\hat{\mathbf{h}}(k) - J(k+1)\hat{\mathbf{h}}(k)]}{\|\hat{\mathbf{h}}(k)\|^2}, \tag{24}$$

$$\tilde{\mathbf{R}}_{x+}(k) = \begin{bmatrix} \sum_{n\neq 1} \tilde{\mathbf{R}}_{x_n x_n}(k) & -\tilde{\mathbf{R}}_{x_2 x_1}(k) & \cdots & -\tilde{\mathbf{R}}_{x_M x_1}(k) \\ -\tilde{\mathbf{R}}_{x_1 x_2}(k) & \sum_{n\neq 2} \tilde{\mathbf{R}}_{x_n x_n}(k) & \cdots & -\tilde{\mathbf{R}}_{x_N x_2}(k) \\ \vdots & \vdots & \ddots & \vdots \\ -\tilde{\mathbf{R}}_{x_1 x_N}(k) & -\tilde{\mathbf{R}}_{x_2 x_N}(k) & \cdots & \sum_{n\neq N} \tilde{\mathbf{R}}_{x_n x_n}(k) \end{bmatrix},$$

and $\tilde{\mathbf{R}}_{x_i x_j}(k) = \mathbf{x}_i(k)\mathbf{x}_j^{\mathrm{T}}(k)$, $i,j = 1,2,\ldots,N$, is an instantaneous estimate of the correlation.

II. *MIMO systems*

Blind identification of a MIMO FIR system with $N > M$ is not just more complicated. Actually it is much more difficult or might be unfeasible. We will review the fundamentals and comment on the state-of-the-art techniques beginning with a close examination of the SOS of the MIMO system outputs given by (5):

$$\mathbf{R}_{xx}(\kappa) \triangleq E\{\mathbf{x}(k)\mathbf{x}^{\mathrm{T}}(k-\kappa)\} = \mathbf{H}\mathbf{R}_{ss}(\kappa)\mathbf{H}^{\mathrm{T}} + \mathbf{R}_{bb}(\kappa), \tag{25}$$

where $\kappa \geqslant 0$ is a delay,

$$\mathbf{R}_{ss}(\kappa) \triangleq E\{\mathbf{s}(k)\mathbf{s}^{\mathrm{T}}(k-\kappa)\}_{ML\times ML}, \text{ and}$$

$$\mathbf{R}_{bb}(\kappa) \triangleq E\{\mathbf{b}(k)\mathbf{b}^{\mathrm{T}}(k-\kappa)\}_{N\times N}.$$

Assuming that the inputs are uncorrelated with each other and with the noise as well, and the noise signals are white and spatially uncorrelated, we then get

$$\mathbf{R}_{s_i s_j}(\kappa) \triangleq E\{\mathbf{s}(k)\mathbf{s}^{\mathrm{T}}(k-\kappa)\} = \mathbf{0}, \quad \text{if } i \neq j \ (i,j = 1,2,\ldots,M) \text{ or } \kappa \geqslant L, \tag{26}$$

$$\mathbf{R}_{bb}(\kappa) = \delta(\kappa)\sigma_b^2 \mathbf{I}_{N\times N}, \tag{27}$$

where $\delta(\kappa)$ is the delta function, σ_b^2 is the noise power, and $\mathbf{I}$ is the identity matrix.

Let us first examine the simplest MIMO system with memoryless channels and with *white* inputs of the same power σ_s^2. In this case, only $\mathbf{R}_{xx}(0)$ is not equal to $\mathbf{0}$ and we get

$$\mathbf{R}_{xx}(0) = \sigma_s^2 \mathbf{H}\mathbf{H}^{\mathrm{T}} + \sigma_b^2 \mathbf{I}. \tag{28}$$

By singular value decomposition (SVD), the matrix $\mathbf{H}$ of size $N \times M$ can be written as a product

$$\mathbf{H} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}, \tag{29}$$

where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices ($\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}_{N\times N}$, $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{I}_{M\times M}$), and $\mathbf{D} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_M)$ is an $N \times M$ diagonal matrix where $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_M > 0$ (here, $\mathbf{H}$ is assumed to be irreducible, i.e., $\mathbf{H}$ has full column rank). Using this decomposition, we obtain

$$\mathbf{H}\mathbf{H}^{\mathrm{T}} = \mathbf{U}\mathbf{D}\mathbf{D}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}} = \mathbf{U}_{1:M}\mathbf{D}_{\mathrm{c}}^2\mathbf{U}_{1:M}^{\mathrm{T}}, \tag{30}$$

where $\mathbf{U}_{1:M}$ is an $N \times M$ matrix collecting the first M orthonormal columns of $\mathbf{U}$ and $\mathbf{D}_{\mathrm{c}} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_M)$ is an $M \times M$ diagonal matrix by cropping $\mathbf{D}$.

The eigenvalue decomposition (EVD) of $\mathbf{R}_{xx}(0)$ is therefore expressed:

$$\mathbf{R}_{xx}(0) = \mathbf{U}\mathbf{D}_x^2\mathbf{U}^{\mathrm{T}}, \tag{31}$$

where $\mathbf{D}_x^2 = \mathrm{diag}(\lambda_{x,1}^2, \lambda_{x,2}^2, \ldots, \lambda_{x,N}^2)$ is a diagonal matrix with $\lambda_{x,1}^2 \geqslant \lambda_{x,2}^2 \geqslant \cdots \geqslant \lambda_{x,N}^2 > 0$ and,

from (28), (30), and (31),

$$\lambda_{x,n}^2 = \begin{cases} \sigma_s^2\lambda_n^2 + \sigma_b^2, & n = 1, 2, \ldots, M \\ \sigma_b^2, & n = M+1, \ldots, N. \end{cases} \quad (32)$$

Therefore, it is straightforward to determine the variance of the noise: the smallest eigenvalue of the received signal covariance matrix is equal to σ_b^2. Furthermore, the unitary matrix $\mathbf{U}$ in the EVD of $\mathbf{R}_{xx}(0)$ is the same as the left unitary matrix of the SVD of $\mathbf{H}$.

Define the following matrices:

$$\bar{\mathbf{R}}_{xx}(0) \triangleq (\mathbf{R}_{xx}(0) - \lambda_{x,N}^2\mathbf{I}_{N\times N}) = \mathbf{H}\mathbf{H}^{\mathrm{T}} = \mathbf{U}_{1:M}\mathbf{D}_{\mathrm{c}}^2\mathbf{U}_{1:M}^{\mathrm{T}}, \quad (33)$$

$$\bar{\mathbf{H}} \triangleq \mathbf{U}_{1:M}\mathbf{D}_{\mathrm{c}}. \quad (34)$$

Then we have

$$\bar{\mathbf{R}}_{xx}(0) = \bar{\mathbf{H}}\bar{\mathbf{H}}^{\mathrm{T}}. \quad (35)$$

It is easy to determine $\bar{\mathbf{H}}$ from $\bar{\mathbf{R}}_{xx}(0)$. But the fact that $\mathbf{R}_{xx}(0) = \bar{\mathbf{H}}\bar{\mathbf{H}}^{\mathrm{T}} = \mathbf{H}\mathbf{H}^{\mathrm{T}}$ does not imply that $\mathbf{H}$ is equal to $\mathbf{H}$. The only thing that we can say is that:

$$\bar{\mathbf{H}} = \mathbf{H}\mathbf{V}, \quad (36)$$

where $\mathbf{V}$ is a unitary matrix. Equating (36) and (34), we see that this unitary matrix is, in fact, the right unitary matrix of the SVD of $\mathbf{H}$.

Clearly from the above analysis, using SOS only, we are able to determine the left unitary matrix $\mathbf{U}_{1:M}$ and the diagonal matrix $\mathbf{D}_{\mathrm{c}}$, but not the right unitary matrix $\mathbf{V}$, of the SVD of $\mathbf{H}$. This means that $\mathbf{H}$ can be determined up to an $M \times M$ unitary matrix, which is not acceptable for the blind identification problem. The channel matrix $\mathbf{H}$ of a MIMO system that is claimed to be blindly identifiable needs to be determined up to only scaling and permutation. Consequently, a memoryless MIMO system with white, uncorrelated inputs of the same power is not blindly identifiable using SOS only. The question would then naturally arise as to what MIMO systems can be blindly identified.

It is already known that a memoryless, irreducible MIMO system with spatially white noise is blindly identifiable using SOS only in two circumstances:

- the input signals are uncorrelated but with *distinct* power [11]; or
- the input signals are quasi-stationary [12].

For blind identification of a MIMO system with memory (convolutive channels), a common approach is to transform the signal model into the frequency domain such that the MIMO system is decomposed into a number of memoryless MIMO systems at different frequency points whose blind identifications are independent. The motivation of this approach is to simplify the problem and reduce the complexity of its solution. However, this approach has a fundamental problem termed *permutation inconsistency*. This arises because independent blind MIMO identifications at different frequency points are only unique up to scaling and possibly different permutation. There are also attempts to solve this problem of blindly identifying a convolutive MIMO system in the time domain [13], leading to extremely complicated implementations with intensive computational complexity. Although some filtering processing can be carried out in the frequency domain, it can only mildly reduce the complexity since major operations are caused by manipulation of matrices with a very large dimension. Intuitively, time-domain (or full-band) approaches might have less permutation inconsistency problems. But it lacks rigorous proof and has not yet been well understood nor accepted whether time-domain approaches can inherently overcome this ambiguity in permutation. It is our belief that blind MIMO identification is still an unsolved challenge and we look forward to breakthroughs.

3.3. Frequency-domain adaptive implementations

Adaptive filters play an increasingly important role in channel identification since they can identify and track unknown and time-varying systems. Time-domain adaptive algorithms, like the classical least mean square (LMS) [14,15] and recursive least-squares (RLS) [5], have been well developed and widely used in various signal processing systems. Adaptive filtering in the frequency domain, ever since its first introduction by Dentino et al. [16] has progressed rapidly and has become an essential constituent of adaptive filter theory. Although the idea of implementing an adaptive filter in the frequency domain by taking advantage of the fast Fourier transform (FFT) is easy to understand, the derivation has been thought quite intricate until [17]. Since then, a number of frequency-domain

adaptive channel identification algorithms have been proposed for such applications as multichannel echo cancellation [18] and time delay estimation [19,20]. Due to space limitations, we cannot detail the discussion on this important topic but would like to refer the interested readers to the above references and the references therein.

4. Applications of acoustic MIMO identification

In this section we discuss five applications of acoustic MIMO identification and use them as examples to explain why acoustic MIMO identification techniques are essential for various acoustic signal processing problems.

4.1. Acoustic echo cancellation

Acoustic echo is produced by voice coupling between the earpiece or loudspeaker and microphone in handsets and hands-free devices [21]. In a long-distance communication system, acoustic echoes are exacerbated by inherent transmission delays, and would significantly lower the call quality. When the delay approaches a quarter of a second, most people find it difficult to carry on a normal conversation [22]. Full-duplex telecommunication was impossible until the birth of echo cancellation theory given by Bell Labs researchers in the 1960s [23–25].

In the center of any echo cancellation system sits an adaptive filter as shown in Fig. 2(a), attempting to dynamically identify the acoustic environment that causes echoes. As long as we can obtain an accurate mathematical representation of the echo path, it would be straightforward to generate a good estimate of the echo and subtract it from the microphone signal. Since reference signals are available in this problem, non-blind methods are readily usable. However, we still have a long way to go before reaching the point of designing a practical system achieving acceptable results in various acoustic environments. Acoustic echo cancellation is an important application of MIMO identification

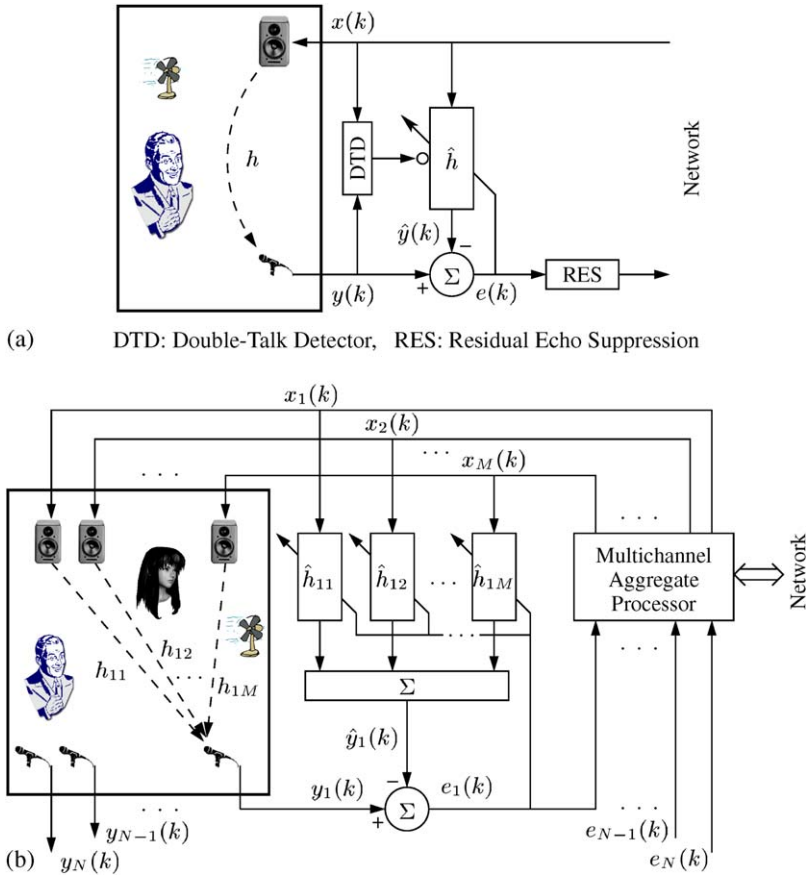

Fig. 2. Illustration of (a) single-channel and (b) multichannel acoustic echo cancellation systems.

and in turn has greatly promoted the development of adaptive filtering and system identification theories.

The growth of bandwidth for data transmission makes the demand for immersive experience more reasonable in the future. As a result, multiple (at least stereo) audio channels have to be included as shown in Fig. 2(b) and multichannel (including stereo) echo cancellation (MEC) becomes imperative. Different from a traditional single-channel system, an MEC system has the non-uniqueness problem as explained before. Furthermore, for an MEC system, more impulse responses and more filter coefficients need to be determined. Therefore, a successful design requires that the adaptive filters converge faster to the true channel impulse responses. For a good survey and analysis of the MEC problem, the readers can refer to a widely cited paper [26] and the references therein.

4.2. Time delay estimation

Time delay estimation (TDE) is essential to many array and multichannel signal processing technologies. Relative time delay of arrival between two microphone signals might be the most important parameter, but at most is only one of many parameters of a multichannel acoustic system. Without directly identifying the surrounding acoustic environment, people have used a simplified model of the acoustic system, where only propagation attenuation and delay are considered, as shown in Fig. 3(a). With this model, the generalized cross-correlation (GCC) algorithm was developed [27] and is still widely used. But the GCC method cannot cope well with room reverberation since the open-space model is obviously unrealistic in a reverberant enclosure. If a realistic reverberant acoustic model is used and the multichannel system can be blindly identified, the relative time delays of arrival can be easily determined. This idea leads to the eigenvalue decomposition algorithm for two channels [28,29] and the multichannel LMS algorithm for multiple channels [10] for time delay estimation in reverberant environments. This idea is adopted more and more by researchers and engineers around the world.

4.3. Crosstalk cancellation

To deliver virtual sound to a single listener, either a headphone can be used or with loudspeakers a crosstalk cancellation system can let the listener enjoy a lifelike acoustic interface without wearing any cumbersome devices. A crosstalk cancellation system is illustrated in Fig. 4 [30]. The desired virtual sound effect would be obtained if $s_L(k)$ and $s_R(k)$ are delivered exactly to the listener's left and right ears, respectively. But due to the room acoustics, if those two signals are played out through two loudspeakers, the listener's left (right) ear will hear signals from $s_R(k)$ ($s_L(k)$ respectively) and the virtual sound effect would be impaired. The crosstalk cancellation system processes $s_L(k)$ and $s_R(k)$ with a group of g filters to get two loudspeaker signals $x_m(k)$ $(m = 1, 2)$ such that $y_L(k) = s_L(k)$ and $y_R(k) = s_R(k)$, i.e., the crosstalk signals are canceled.

Finding the filters g_{iL} and g_{iR} $(i = 1, 2)$ in a crosstalk cancellation system requires accurate knowledge of the room acoustic system with the four impulse responses from two loudspeaker to two ears being determined. Therefore acoustic MIMO identification technique plays an important role in the design of a crosstalk cancellation system. If the estimates of the channel impulse responses are

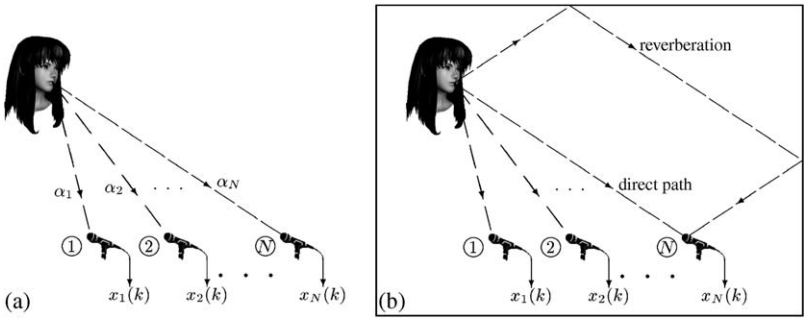

Fig. 3. Time delay estimation in two different acoustic environments: (a) open space with no room reverberation and (b) enclosure with considerable room reverberation.

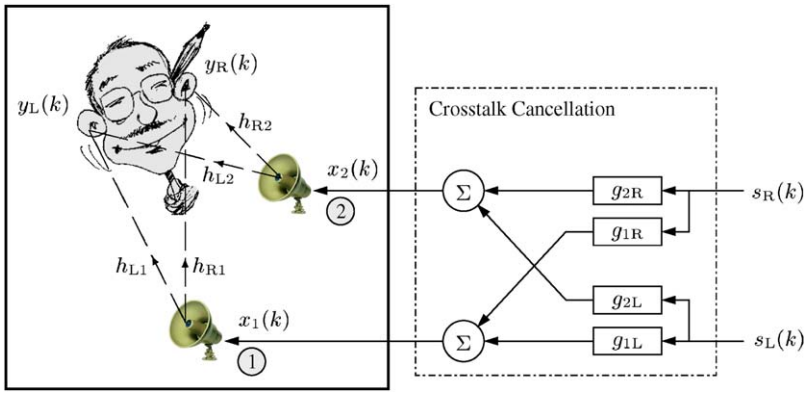

Fig. 4. Schematic diagram of a crosstalk cancellation system.

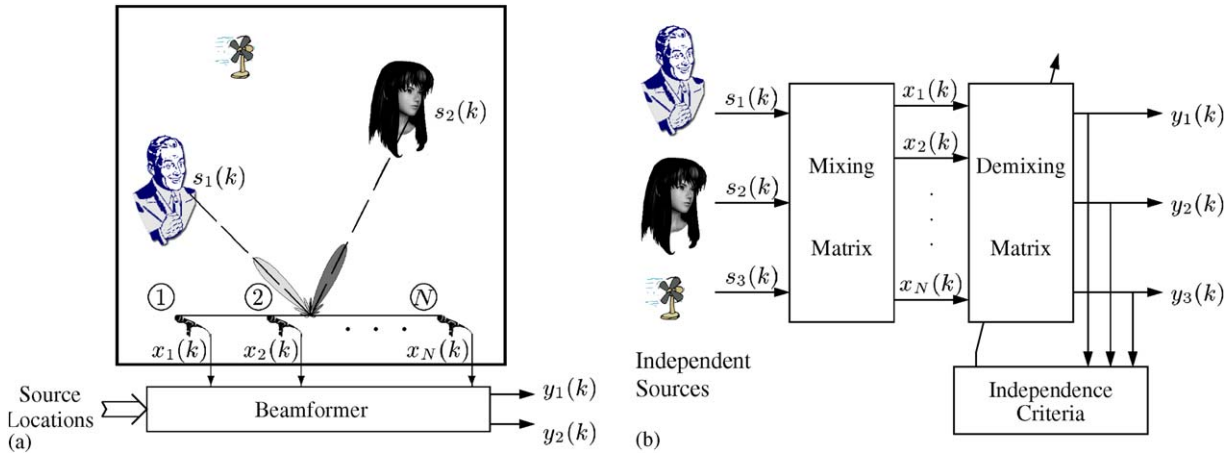

Fig. 5. Two classes of conventional techniques for source separation: (a) beamforming and (b) blind source separation.

inaccurate, the ideal virtual sound effect cannot be achieved. In addition, since acoustic impulse responses would change with the location of the listener, the performance of all crosstalk cancellation systems is critically dependent on whether the listener is in the fixed design position, the so-called "sweet-spot." The readers can refer to [31,32], and references therein for detailed discussions of the problem of acoustic crosstalk cancellation.

4.4. Source separation

Recently, source separation has received increasing attention since it can potentially be applied in a number of speech processing and communication systems. In the problem of source separation, we have multiple talkers and microphones, naturally forming an acoustic MIMO system. Since very little is known about the source signals, identifying such a system is extremely difficult as explained in Section 3.2. Alternatively, researchers have in the past tackled this problem without first trying to identify the acoustic MIMO system.

As a part of our daily experience, we know that distinguishing and even separating components of a mixture or collection depends on their distinctions. In a multi-talker environment, the sound sources are different in location and statistics in addition to spectrum, which leads to two different categories of source separation methods in the literature: beamforming and blind source separation (BSS).

Beamforming is a form of spatial filtering that enhances the signal from a specified "look direction" and attenuates signals that propagate from other directions [33], as shown in Fig. 5(a). Therefore a beamformer cannot only separate multiple sound sources in different directions with respect to the microphone array but also suppress reverberation. However, in practice its performance is limited by a number of factors. Beamforming

relies on knowledge of the speaker's position, which is seldom available. While the position of the speaker can be estimated by analyzing the microphone outputs, errors are inevitable, particularly when the room is considerably reverberant [34]. Furthermore, current microphone array technologies including beamforming originated from radar and sonar array signal processing. But compared to classical sensor array processing with antenna arrays [35], the basic conditions are significantly different in acoustics: speech is a baseband signal spanning almost three decades in frequency and the localization and recording take place in the nearfield with respect to the microphone array.

Alternatively BSS methods solve this problem by taking advantage of the difference in statistics among multiple sound sources under investigation. BSS that is typically accomplished by independent component analysis (ICA) algorithms assumes mutually independent sound sources [36,37]. An ICA processes microphone signals with a de-mixing system whose outputs are estimates of the separated source signals satisfying the independent assumption, as illustrated in Fig. 5(b). Existing ICA algorithms differ in the way the dependence of the separated source signals is defined, i.e., the employed criteria for minimization, which include second-order statistics [38], higher (than second) order statistics [39], and information-theory-based measures [40] (please refer to the book [41] and references therein for a more detailed discussion on various ICA methods). BSS methods allow for nearfield sources and reverberant acoustic environments. But in reverberant environments, they are either very complex (for time-domain approaches [42]) or have the inherent permutation inconsistency problem as encountered in the problem of blind MIMO identification [37,43–45] (for frequency-domain algorithms [46]). Similarly to blind MIMO identification, convolution operations in the time-domain BSS method for convolutive mixtures can be carried out in the frequency domain by using the FFT [47,48], but the complexity is still intensive. Moreover, current BSS methods do not work for arbitrary source positions. When sources are at positions such that the mixing matrix is singular, the de-mixing system (the inverse of the mixing matrix) does not exist and source separation cannot be attained. Finally, it should be noted that, in addition to the above drawbacks, independent but distorted source signals are valid solutions for BSS methods. Therefore deconvolution is usually needed to mitigate convolutive distortion and reconstruct the original speech signals.

From the above discussion, it is clear that existing source separation methods cannot fundamentally achieve satisfactory results because of the lack of effective algorithms to blindly identify an acoustic MIMO system. If a breakthrough in this area is achieved, source separation can be readily solved with generalized zero-forcing or minimum mean-square error (MMSE) equalizers [49], in which source signals are separated and equalized in one single step. This procedure is better understood by decomposing the processings of separation and equalization using the technique proposed in [50]. In that approach, multiple sources are separated by converting the $M \times N$ MIMO system into M interference-free SIMO systems. A simple example of this algorithm with respect to a 2×3 MIMO system is shown in Fig. 6. The generalization to an arbitrary $M \times N$ MIMO system with $M < N$ can be found in [50] as well. Apparently the separated signals using this approach are distorted and speech dereverberation is necessary, as will be discussed in the following section.

4.5. Speech dereverberation

Acoustic channels are rarely ideal and speech signals are linearly distorted by room reverberation before they reach microphones. The goal of speech dereverberation is to equalize the acoustic channels and to recover the original source speech signals. Even after three decades of continuous research, speech dereverberation remains a challenging problem. While there have been a number of ways to classify current speech dereverberation methods, we believe that an insightful approach is based on whether the channel impulse responses need to be known or estimated beforehand. If the channel impulse responses are not known or the acoustic system cannot be identified, either cepstral-domain processing techniques can be chosen [51] or the characteristics of speech (usually in statistical forms) can be exploited in an attempt to recover the energy envelope of the original speech [52]. As expected, these methods achieve only moderate success. For acoustic systems with known or accurately estimated channel impulse responses, there are three approaches to speech dereverberation, as illustrated in Fig. 7. The most straightforward is the direct inverse method. But it

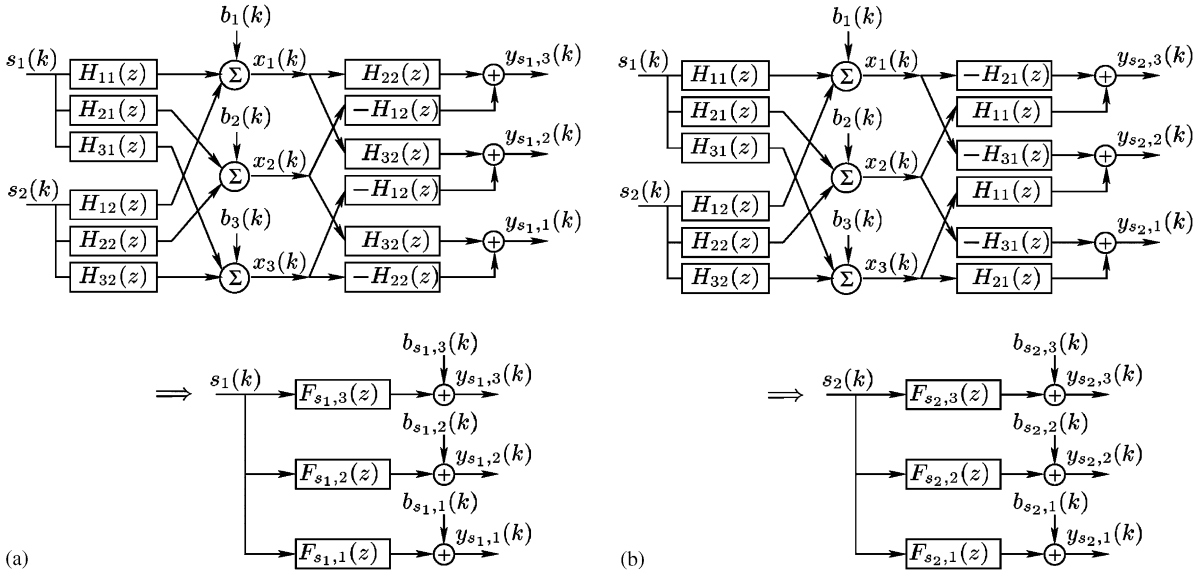

Fig. 6. Illustration of the conversion from a 2×3 MIMO system to two interference-free SIMO systems with respect to (a) $s_1(k)$ and (b) $s_2(k)$.

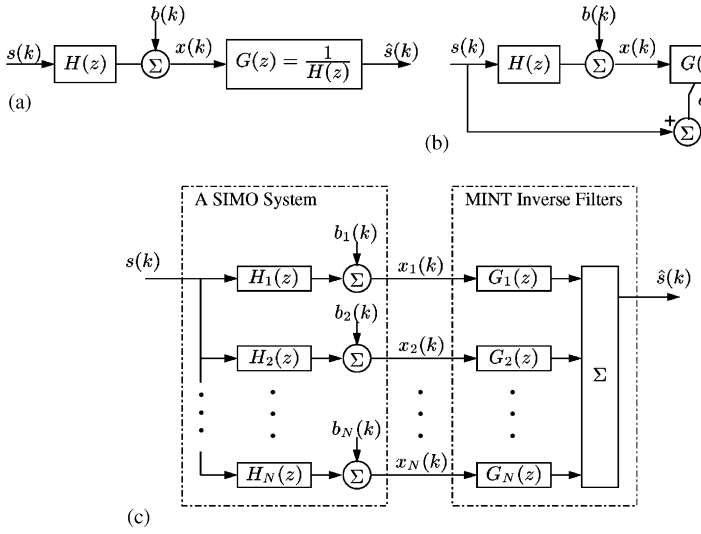

Fig. 7. Illustration of three widely-used approaches to speech dereverberation: (a) direct inverse, (b) least squares, and (c) the MINT method.

is well known that the impulse response of a single acoustic channel needs to be a minimum-phase sequence for *stable and causal* exact inversion [53]. Otherwise by using an all-pass filter, the resultant inverse filter g would be IIR which is non-causal and has a long delay. The second approach is the least squares (LS) method, which essentially equalizes the channel by inverting only those components whose zeros are inside the unit circle. In addition, in the process of determining the LS inverse filter, a reference signal needs to be accessible. Although the LS method has these constraints, it is quite useful in practice and has been widely employed in different systems. The third method is based on the MINT (multichannel inverse theorem) technique [54] for speech dereverberation with respect to a SIMO system [55]. As shown in Fig. 7(c), as long as the channel impulse responses h_n $(n = 1, 2, \ldots, N)$ are coprime (even though they may not be minimum phase), i.e., the SIMO system is irreducible, there exists a group of g filters to perfectly dereverberate the distorted speech signals. This feature is indeed

appealing and therefore it received a lot of attention immediately after it was proposed. However, this method is very sensitive to errors in the estimates of channel impulse responses and the computational complexity of determining the inverse filters is intensive. Further research needs to be carried out to overcome these drawbacks.

5. Successful real-time acoustic signal processing systems

In the above, we have discussed various channel identification techniques with respect to acoustic MIMO systems and have explained how they could positively impact on the research of a number of acoustic signal processing problems. In the following, we would like to present three successful real-time acoustic systems that we developed in our research, based on channel identification methods. Hopefully these systems can inspire the development of more advanced and practical acoustic signal processing algorithms in the future.

5.1. Stereo acoustic echo cancellation system for teleconferencing

Research on stereo acoustic echo cancellation (SAEC) can be traced back to the early 1990s as the need developed for multichannel audio communications, which at least involves stereo sound. However, SAEC is not a straightforward generalization of the monophonic acoustic echo cancellation principle and the non-uniqueness problem needs to be solved. A number of methods to decorrelate stereo loudspeaker signals without perceptible distortion were proposed but they are all unsatisfactory. A breakthrough was achieved with the introduction of a non-linear device into each loudspeaker signal path [26]. Among several non-linear transformations that were evaluated, the half-wave rectifier was suggested as the simplest, yet effective type [56]. In addition, computational complexity is another challenging problem in implementing a practical SAEC system. A real-time SAEC system was possible only by using specially designed digital signal processors (DSPs) until the development of efficient frequency-domain multichannel adaptive algorithms. The year of 1998 saw the world's first real-time SAEC system working between two conference rooms at Bell Labs [57], as illustrated in Fig. 8. This system was running on a Texas Instruments' TMS320C44 DSP and incorporated the non-linear transformation method and a two-channel subband fast recursive least squares (FRLS) algorithm. Meanwhile real-time DSP-based SAEC systems were also successfully developed by other research groups such as [58]. Later in 2000, a more efficient, more economical SAEC system based on Intel CPU's was developed using a two-channel frequency-domain adaptive algorithm [59]. These systems show great promises of wide deployment of SAEC technology in future audio communications, particularly teleconferencing.

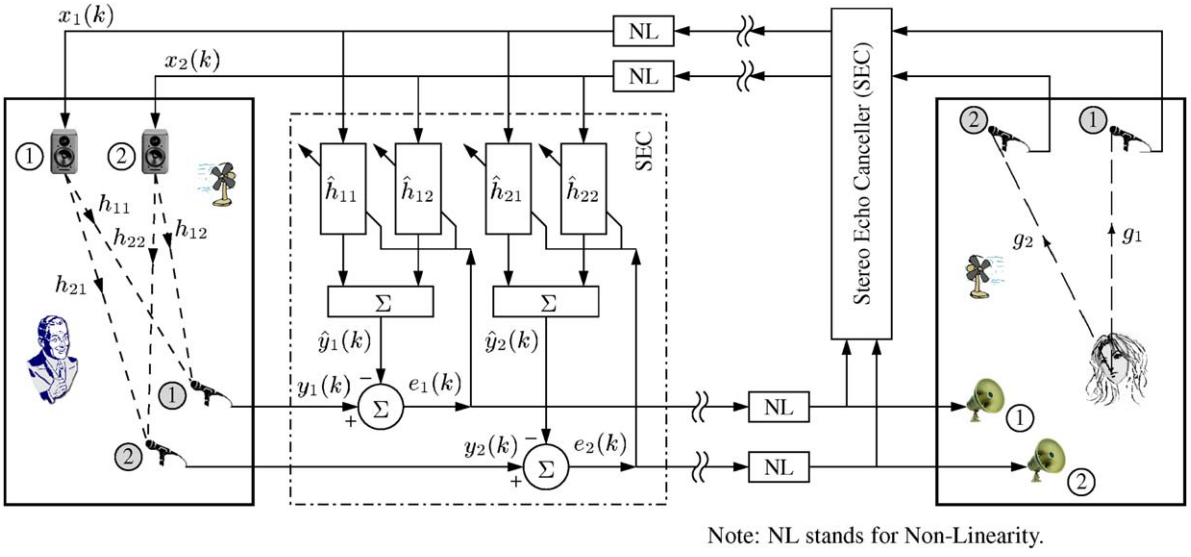

Fig. 8. The world's first real-time stereo acoustic echo cancellation system working between two conference rooms at Bell Labs.

5.2. Synthesized stereo audio bridge system for multi-party conferencing

Using multiple microphones and loudspeakers can ideally help an audio communication system deliver sound realism. But this requires new investments in audio hardware for such a sophisticated acoustic interface. Nowadays all personal computers (desktops or laptops) have at least one microphone and a pair of loudspeakers. They can be readily employed for lifelike multi-party conferencing with the support of a synthesized stereo audio bridge [60], as shown in Fig. 9. Although this system does not work well for the case in which there are multiple participants at one location, it is cheap and can improve the sound realism by presenting speech signals from different sites to the listener with different spatial cues and impressions.

5.3. Passive acoustic speaker tracking system for automatic camera steering in video conferencing

Recently, the technique of acoustic source localization and tracking has gained increasing attention since acoustic tracking systems have some advantages that vision-based systems do not possess: microphones can receive propagating sound omni-directionally and can function in dark or poor lighting conditions. Time-delay-estimation-based approaches have become the technique of choice since they are simple and can be implemented in real time with current digital systems. The difficulty of building a TDE-based acoustic source localization and tracking system lies in two areas: developing an accurate TDE algorithm that is robust to room reverberation and solving a group of non-linear equations of the estimated relative time delays of arrival for the source location. Without effective methods to solve these two problems, the size of microphone arrays has to be large to meet the performance requirement [61–63]. This situation changed after the invention of channel-identification-based TDE algorithms and a linear-correction least-squares source localization algorithm [64]. A real-time system for automatic camera steering in video conferencing using these techniques was successfully developed in 2000 [65] and Fig. 10 provides a schematic diagram of this system. A unique, small-size six-element microphone array makes this system portable and facilitates setup after being moved to a new room. The performance of this system in various acoustic environments justifies the effectiveness of applying channel identification techniques in time delay estimation.

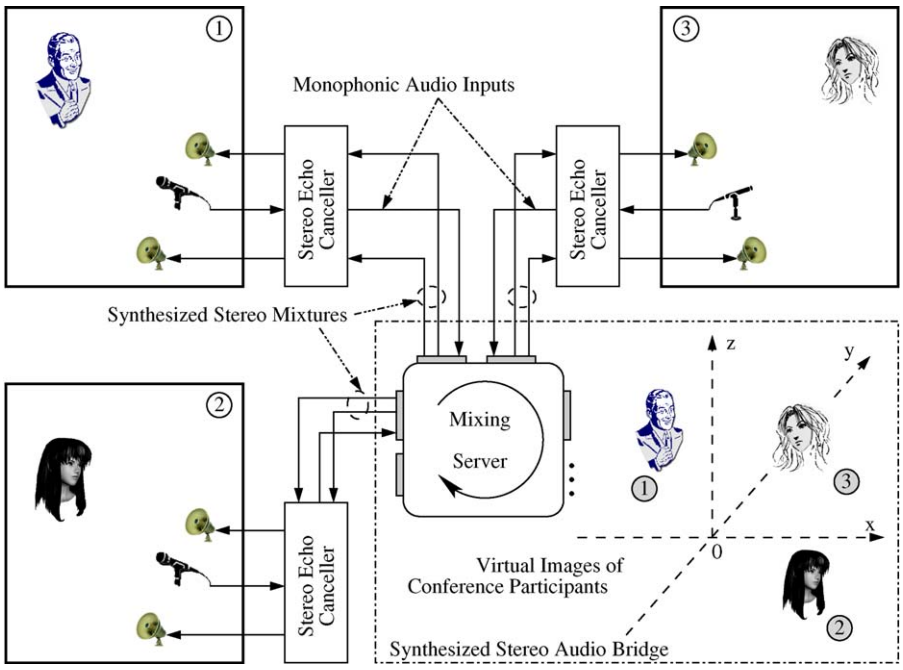

Fig. 9. The synthesized stereo audio bridge system combined with acoustic echo cancellation for multi-party conferencing developed at Bell Labs.

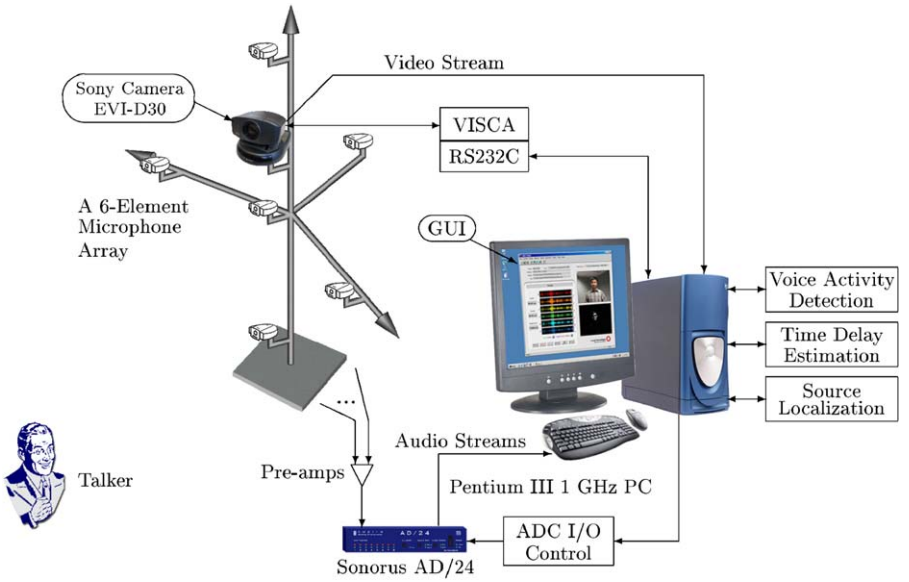

Fig. 10. The passive acoustic speaker tracking system for automatic camera steering in video conferencing developed at Bell Labs.

6. Conclusions

How to estimate accurately room acoustic impulse responses in real time is the mother of all challenges in acoustic signal processing and we have presented a systematic overview of acoustic MIMO identification algorithms in this paper. We have also discussed a wide range of technological problems that define MIMO acoustics or multichannel acoustic signal processing, including acoustic echo cancellation, time delay estimation, acoustic cross-talk cancellation, source separation, and speech dereverberation. In all these problems, system identification via either non-blind or blind methods plays a central role. This implies that any breakthroughs in acoustic MIMO identification would make a significant impact on the advancement of acoustic signal processing. Three real-time acoustic systems for different applications, all based on system identification, were given and their success confirmed our belief this approach.

Acknowledgements

We would like to thank Dennis R. Morgan for his constructive comments and suggestions that have improved the clarity of this paper.

References

[1] C. Kyriakakis, Fundamental and technological limitations of immersive audio systems, Proc. IEEE 86 (5) (May 1998) 941–951.
[2] C. Kyriakakis, P. Tsakalides, T. Holman, Surrounded by sound, IEEE Signal Process. Mag. 16 (1) (January 1999) 55–66.
[3] I.E. Telatar, Capacity of multi-antenna Gaussian channels, Technical Report, Bell Labs, 1995.
[4] G.J. Foschini, Layered space-time architecture for wireless communication in a fading environment using multi-element antennas, Bell Labs Tech. J. 1 (2) (1996) 41–59.
[5] S. Haykin, Adaptive Filter Theory, fourth ed., Prentice-Hall, Upper Saddle River, NJ, 2002.
[6] Y. Sato, A method of self-recovering equalization for multilevel amplitude-modulation, IEEE Trans. Comm. COM-23 (6) (June 1975) 679–682.
[7] L. Tong, G. Xu, T. Kailath, A new approach to blind identification and equalization of multipath channels, in: Proceedings of the 25th Asilomar Conference on Signals, Systems, and Computers, vol. 2, 1991, pp. 856–860.
[8] G. Xu, H. Liu, L. Tong, T. Kailath, A least-squares approach to blind channel identification, IEEE Trans. Signal Process. 43 (December 1995) 2982–2993.
[9] C. Avendano, J. Benesty, D.R. Morgan, A least squares component normalization approach to blind channel identification, in: Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing, vol. 4, 1999, pp. 1797–1800.
[10] Y. Huang, J. Benesty, Adaptive multi-channel least mean square and Newton algorithms for blind channel identification, Signal Process. 82 (August 2002) 1127–1138.
[11] Y. Hua, J.K. Tugnait, Blind identifiability of FIR-MIMO systems with colored input using second order

statistics, IEEE Signal Process. Lett. 7 (12) (December 2000) 348–350.
[12] K. Rahbar, J.P. Reilly, J.H. Manton, Blind identification of MIMO FIR systems driven by quasistationary sources using second-order statistics: a frequency domain approach, IEEE Trans. Signal Process. 52 (2) (February 2004) 406–417.
[13] M. Kawamoto, Y. Inouye, Blind deconvolution of MIMO-FIR systems with colored inputs using second-order statistics, IEICE Trans. Fundamentals E86-A (3) (March 2003) 597–604.
[14] B. Widrow, M.E. Hoff, Jr., Adaptive switching circuits, in: IRE WESCON Convention Record, 1960, Pt. 4, pp. 96–104.
[15] B. Widrow, S. Stearns, Adaptive Signal Processing, Prentice-Hall, Englewood Cliffs, NJ, 1985.
[16] M. Dentino, J. McCool, B. Widrow, Adaptive filtering in the frequency domain, Proc. IEEE 66 (12) (December 1978) 1658–1659.
[17] J. Benesty, D.R. Morgan, Frequency-domain adaptive filtering revisited, generalization to the multi-channel case, and application to acoustic echo cancellation, in: Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing, vol. II, 2000, pp. 789–792.
[18] H. Buchner, J. Benesty, W. Kellermann, Multichannel frequency-domain adaptive filtering with application to multichannel acoustic echo cancellation, in: J. Benesty, Y. Huang (Eds.), Adaptive Signal Processing: Applications to Real-World Problems, Springer, Berlin, 2003 (Chapter 4).
[19] Y. Huang, J. Benesty, A class of frequency-domain adaptive approaches to blind multichannel identification, IEEE Trans. Signal Process. 51 (1) (January 2003) 11–24.
[20] Y. Huang, J. Benesty, Adaptive multichannel time delay estimation based on blind system identification for acoustic source localization, in: J. Benesty, Y. Huang (Eds.), Adaptive Signal Processing: Applications to Real-World Problems, Springer, Berlin, 2003 (Chapter 8).
[21] E. Hänsler, G. Schmidt, Acoustic Echo and Noise Control: A Practical Approach, Wiley, New York, 2004.
[22] S.L. Gay, J. Benesty, An introduction to acoustic echo and noise control, in: S.L. Gay, J. Benesty (Eds.), Acoustic Signal Processing for Telecommunication, Kluwer Academic, Boston, MA, 2000 (Chapter 1).
[23] M.M. Sondhi, A.J. Presti, A self-adaptive echo canceler, Bell Syst. Tech. J. 45 (1966) 1851–1854.
[24] F.K. Becker, H.R. Rudin, Application of automatic transversal filters to the problem of echo suppression, Bell Syst. Tech. J. 45 (1966) 1847–1850.
[25] M.M. Sondhi, An adaptive echo canceler, Bell Syst. Tech. J. 46 (March 1967) 497–511.
[26] J. Benesty, D.R. Morgan, M.M. Sondhi, A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation, IEEE Trans. Speech Audio Process. 6 (2) (March 1998) 156–165.
[27] C.H. Knapp, G.C. Carter, The generalized correlation method for estimation of time delay, IEEE Trans. Acoust. Speech Signal Process. ASSP-24 (4) (August 1976) 320–327.
[28] Y. Huang, J. Benesty, G.W. Elko, Adaptive eigenvalue decomposition algorithm for realtime acoustic source localization system, in: Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing, vol. 2, 1999, pp. 937–940.
[29] J. Benesty, Adaptive eigenvalue decomposition algorithm for passive acoustic source localization, J. Acoust. Soc. Am. 107 (January 2000) 384–391.
[30] B.S. Atal, M.R. Schroeder, Apparent sound source translator, U.S. Patent 3 236 949, February 1966.
[31] D.B. Ward, G.W. Elko, Virtual sound using loudspeakers: robust acoustic crosstalk cancellation, in: S.L. Gay, J. Benesty (Eds.), Acoustic Signal Processing for Telecommunication, Kluwer Academic, Boston, MA, 2000 (Chapter 14).
[32] D.B. Ward, Joint least squares optimization for robust acoustic crosstalk cancellation, IEEE Trans. Speech Audio Process. 8 (2) (February 2000) 211–215.
[33] B.D. Van Veen, K.M. Buckley, Beamforming: a versatile approach to spatial filtering, IEEE ASSP Mag. 5 (April 1988) 4–24.
[34] Y. Huang, J. Benesty, G.W. Elko, Source localization, in: Y. Huang, J. Benesty (Eds.), Audio Signal Processing for Next-Generation Multimedia Communication Systems, Kluwer Academic, Boston, MA, 2004 (Chapter 9).
[35] B. Widrow, P.E. Mantley, L.J. Griffiths, B.B. Goode, Adaptive antenna systems, Proc. IEEE 55 (12) (December 1967) 2143–2159.
[36] P. Comon, Independent component analysis: a new concept?, Signal Processing 36 (3) (April 1994) 287–314.
[37] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.
[38] L. Molgedey, H.G. Schuster, Separation of a mixture of independent signals using time delayed correlations, Phys. Rev. Lett. 72 (23) (June 1994) 3634–3637.
[39] J.-F. Cardoso, Eigenstructure of the 4th-order cumulant tensor with application to the blind source separation problem, in: Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing, 1989, pp. 2109–2112.
[40] S. Amari, A. Cichocki, H.H. Yang, Blind signal separation and extraction: neural and information-theoretic approaches, in: S. Haykin (Ed.), Unsupervised Adaptive Filtering, Blind Source Separation, vol. 1, Wiley, New York, 2000.
[41] A. Cichocki, S. Amari, Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications, Wiley, New York, 2002.
[42] B.S. Krongold, D.L. Jones, Blind source separation of nonstationary convolutively mixed signals, in: Proceedings of the IEEE SSAP, 2000, pp. 53–57.
[43] M.Z. Ikram, D.R. Morgan, Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment, in: Proceedings of the IEEE International Conference Acoustics, Speech, Signal Processing, 2000, pp. 1041–1044.
[44] S. Araki, R. Mukai, S. Makino, T. Nishikawa, H. Saruwatari, The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech, IEEE Trans. Speech Audio Process. 11 (2) (March 2003) 109–116.
[45] M.Z. Ikram, D.R. Morgan, Permutation inconsistency in blind speech separation: investigation and solutions, IEEE Trans. Speech Audio Process. 13 (1) (January 2005) 1–13.
[46] L. Parra, C. Spence, Convolutive blind separation of non-stationary sources, IEEE Trans. Speech Audio Process. 8 (3) (May 2000) 320–327.
[47] M. Joho, Blind signal separation of convolutive mixtures: a time-domain joint-diagonalization approach, in: Proceedings

of the International Symposium on Independent Component Analysis Blind Signal Separation, 2004, pp. 577–584.

[48] H. Buchner, R. Aichner, W. Kellermann, A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics, IEEE Trans. Speech Audio Process. 13 (1) (2005) 120–134.

[49] B.R. Petersen, D.D. Falconer, Suppression of adjacent-channel, co-channel, and intersymbol interference by equalizers and linear combiners, IEEE Trans. Comm. 42 (12) (December 1994) 3109–3118.

[50] Y. Huang, J. Benesty, J. Chen, Separating ISI and CCI in a two-step FIR Bezout equalizer for MIMO systems of frequency-selective channels, in: Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing, vol. IV, 2004, pp. 797–800.

[51] D. Bees, M. Blostein, P. Kabal, Reverberant speech enhancement using cepstral processing, in: Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing, vol. 2, 1991, pp. 977–980.

[52] T. Nakatani, M. Miyoshi, Blind dereverberation of single channel speech based on harmonic structure, in: Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing, vol. I, 2003, pp. 92–95.

[53] A.V. Oppenheim, R.W. Schafer, Discrete-Time Signal Processing, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[54] M. Miyoshi, Y. Kaneda, Inverse filtering of room acoustics, IEEE Trans. Acoust. Speech Signal Process. 36 (February 1988) 145–152.

[55] K. Furuya, Y. Kaneda, Two-channel blind deconvolution for non-minimum phase impulse responses, in: Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing, vol. 2, 1997, pp. 1315–1318.

[56] D.R. Morgan, J.L. Hall, J. Benesty, Investigation of several types of nonlinearities for use in stereo acoustic echo cancellation, IEEE Trans. Speech Audio Process. 9 (6) (September 2001) 686–696.

[57] P. Eneroth, S.L. Gay, T. Gänsler, J. Benesty, An implementation of a stereophonic acoustic echo canceller on a general purpose DSP, in: Proceedings of the ICSPAT, 1999.

[58] S. Shimauchi, S. Makino, Y. Haneda, A. Nakagawa, S. Sakauchi, A stereo echo canceller implemented using a stereo shaker and a duo-filter control system, in: Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing, vol. 2, 1999, pp. 857–860.

[59] T. Gänsler, J. Benesty, E.J. Diethorn, V. Fischer, Algorithm design of a stereophonic acoustic echo canceller system, in: Proceedings of the IEEE ASSP Workshop Applications Signal Processing Audio Acoustics, 2001, pp. 179–182.

[60] J. Benesty, D.R. Morgan, J.L. Hall, M.M. Sondhi, Synthesized stereo combined with acoustic echo cancellation for desktop conferencing, Bell Labs Tech. J. 3 (July–September 1998) 148–158.

[61] M. Brandstein, A Framework for Speech Source Localization Using Sensor Arrays, Ph.D. Thesis, Brown University, Providence, RI, 1995.

[62] D.V. Rabinkin, R.J. Ranomeron, J.C. French, J.L. Flanagan, M.H. Bianchi, A DSP implementation of source location using microphone arrays, in: Proceedings of the SPIE, vol. 2846, 1996, pp. 88–99.

[63] H. Wang, P. Chu, Voice source localization for automatic camera pointing system in videoconferencing, in: Proceedings of the IEEE Workshop Applications Signal Processing Audio Acoustics, 1997.

[64] Y. Huang, J. Benesty, G.W. Elko, R.M. Mersereau, Real-time passive source localization: a practical linear-correction least-squares approach, IEEE Trans. Speech Audio Process. 9 (8) (November 2001) 943–956.

[65] Y. Huang, J. Benesty, G.W. Elko, Passive acoustic source localization for video camera steering, in: Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing, vol. II, 2000, pp. 909–912.