12 Separation and Dereverberation of Speech Signals with Multiple Microphones

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

[1] Bell Laboratories, Lucent Technologies
Murray Hill, NJ 07974, USA
E-mail: {arden, jingdong}@researchbell-labs.com
[2] Université du Québec, INRS-EMT
Montréal, QC H5A 1K6, Canada
E-mail: benesty@inrs-emt.uquebec.ca

Abstract. Speech enhancement was not and should not be examined solely with the tool of time-frequency analysis. Approaching this problem from different perspectives or incorporating other knowledges helps to expand the number of options open to us when developing a speech enhancement system. Using multiple microphones at different locations makes it possible to develop more sophisticated source separation and dereverberation technologies for speech enhancement, which enable man-made systems to extract a speech signal of interest in a noisy environment with competing speech and/or noise sources. This phenomenon is referred to as the cocktail party effect demonstrated by human beings and many other creatures with few efforts. However, separating and dereverberating speech signals is a very difficult problem in reverberant environments and the state-of-the-art algorithms 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 microphone signals. Focusing only on optimizing the signal-to-interference ratio is inadequate for most speech processing systems where source separation and speech dereverberation are two fully-integrated problems. In this chapter, we study these two problems in a unified framework. We deduce that spatial interference and temporal reverberation can be separated and a SIMO system with the speech signal of interest as input is extracted from the MIMO system. Furthermore, this interference-free SIMO system is dereverberated using the MINT theorem. 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 verified the proposed algorithm.

12.1 Introduction

Speech enhancement is essential for tremendous applications of speech processing and communications since we are living in a natural environment where noise or disturbance is perpetual and ubiquitous. Speech signals can seldom be recorded in pure form and in most cases they are immersed in acoustic ambient noise or reverberation.

In order to develop an effective approach to extracting a desired speech signal from their corrupt observations, we need to understand how distortions are introduced. From a statistical viewpoint, there are only two sources of distortion. One is uncorrelated or even independent noise or competing speech, and the other is correlated reverberation or echo. While many single-channel algorithms and techniques have had varying success in noise reduction as explained in previous chapters, speech enhancement in the sense of separation and dereverberation would be very difficult if not impossible to accomplish using only one microphone in distance. A listener has the ability of choosing to focus on a specific speaker in a room where several people are talking concurrently and where noise sources might meanwhile exist. This phenomenon is referred to as the *cocktail party effect* or *attentional selectivity* [1]. This effect is mainly attributed to the fact that we have two ears and our perception of speech is based on binaural hearing, which can be easily demonstrated by observing the difference in understanding between using both ears and with either ear covered when listening in a cocktail-party-like environment. This suggests the use of two or more microphones, i.e., microphone arrays, in the development of prospect speech separation and dereverberation algorithms and systems.

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 multispeaker environment, sound sources are different in location and statistics in addition to spectrum, which leads to two different categories of speech separation method using multiple microphones: beamforming and blind source separation (BSS).

Beamforming is a form of spatial filtering that enhances the signal from "look direction" and attenuates signals that propagate from directions other than the "look direction" [2]. Therefore a beamformer can not only separate multiple sound sources but also suppress reverberation for the speech source of interest. However, its performance is limited by a number of factors in practice. Beamforming relies on the knowledge of the speaker's position, which is seldom available. While the position of the speaker can be estimated by analysis of the microphone outputs, errors are inevitable particularly when the room is considerably reverberant [3]. Furthermore, current microphone array technologies including beamforming originated from array signal processing. But compared to classical sensor array processing with antenna arrays [4], the basic conditions are significantly different in acoustics: speech is a baseband signal and the localization and recording take place in the nearfield with respect to the microphone array.

Alternatively BSS methods tackle this problem by taking advantage of the difference in statistics among multiple sound sources under investigation [5]. BSS that is typically accomplished by independent component analysis (ICA) algorithms [6] assumes mutually independent sound sources. The mixing procedure is typically delineated with a multiple-input multiple-output

(MIMO) mathematical model, which is either memoryless or with memory, being referred to as instantaneous and convolutive mixtures, respectively. An ICA processes microphone signals with a de-mixing system whose outputs are estimates of the separated source signals satisfying the independent assumption. 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 [7], higher (than two) order statistics [8], and information-theory-based measures [9]. BSS methods allow for near-field sources and reverberant acoustic environments. But in reverberant environments, they are either very complex (for time-domain approaches [10]) or have an inherent problem of so-called permutation inconsistency [11] (for frequency-domain algorithms [12]). Moreover, it is not true nevertheless that current BSS methods work for arbitrary source positions. When sources are at the 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 original speech signals.

Speech dereverberation remains a challenging problem even after three decades of continuous research. While the number of employed microphone signals is a common way to classify current speech dereverberation methods, another insightful approach is based on whether the channel impulse responses need to be known or estimated. In the case of a single microphone with the corresponding acoustic channel impulse response not being able to access anyway, either cepstral-domain processing techniques were suggested to separate speech from reverberation [13], [14] or characteristics of speech (usually in statistical forms) could be exploited with the attempt to recover the energy envelope of the original speech [15]. But they achieved only moderate successes because of a very large variety of applications. If the acoustic channel impulse responses are known, speech dereverberation can be performed by inverting those impulse responses. It is well known that the impulse response of a single acoustic channel needs to a minimum-phase sequence for stable and causal *exact* inversion [16]. Otherwise the inverting filter would either be IIR (noncausal and with a long delay) for exact inversion or just produce an LS (least-squares) solution. However, using multiple microphones, we can carry out perfect speech dereverberation with causal FIR filters even for non-minimum-phase channels. The principle is widely known as the MINT (multichannel inverse) theorem [17].

In this chapter, we will investigate the problem of speech enhancement by separating the speech of interest from concurrent interference (speech and/or noise) sources and by mitigating distortion due to room reverberation from a novel perspective within a unified framework using multiple microphones. In a MIMO acoustic system, microphone outputs are convolu-

tive mixtures containing both reverberant speech and competing interference. We will show that the reverberant speech and interference can be completely separated given the blindly estimated channel impulse responses from interference sources. It is assumed that the number of microphones would be greater than the number of speech and interference sources. Then by choosing different combinations of microphone outputs, we obtain a number of diversely distorted speech signals, which composes a single-input multiple-output (SIMO) system. For such a system, we can again blindly identify its channel impulse responses and then apply the MINT theorem to remove reverberation. Therefore, the speech enhancement algorithm that will be developed here is a two-step procedure dealing with interference and reverberation sequentially. As a result, we are able to mimic the cocktail party effect with man-made machines.

This chapter is organized as follows. Section 12.2 introduces the MIMO signal model, formulates the problem of speech separation and dereverberation, and explains all assumptions that will be made. In Section 12.3, we brief the technique of blindly identifying a SIMO system. Section 12.4 explains how to separate reverberant speech and interference. A SIMO system with the speech of interest as the input will be extracted from the MIMO system. Since the SIMO is free of interference, we can again blindly identify its impulse responses and perform exact dereverberation using the MINT theorem, which will be illustrated in Section 12.5. Section 12.6 evaluates the developed algorithm by simulations and Section 12.7 draws the conclusions.

12.2 Signal Model and Problem Formulation

We consider an acoustic environment where there are one speech source of interest, $M-1$ concurrent sound sources, and N microphones with $M < N$. The speech source and $M-1$ other sound sources are mutually independent. Those competing sound sources can be speech or noise, and are regarded as interference. Such a system is mathematically described by an $M \times N$ MIMO FIR model as shown in Fig. 12.1. Without loss of generality, we label the speech source of interest as the first. At the n-th microphone and at the k-th sample time, we have:

$$\begin{aligned} x_n(k) &= \sum_{m=1}^{M} \mathbf{h}_{nm}^T \mathbf{s}_m(k, L_h) + b_n(k), \\ & k = 1, 2, \cdots, K, \quad n = 1, 2, \cdots, N, \end{aligned} \tag{12.1}$$

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

$$\begin{aligned} \mathbf{h}_{nm} &= \left[\begin{array}{cccc} h_{nm,0} & h_{nm,1} & \cdots & h_{nm,L_h-1} \end{array} \right]^T, \\ & n = 1, 2, \cdots, N, \quad m = 1, 2, \cdots, M, \end{aligned}$$

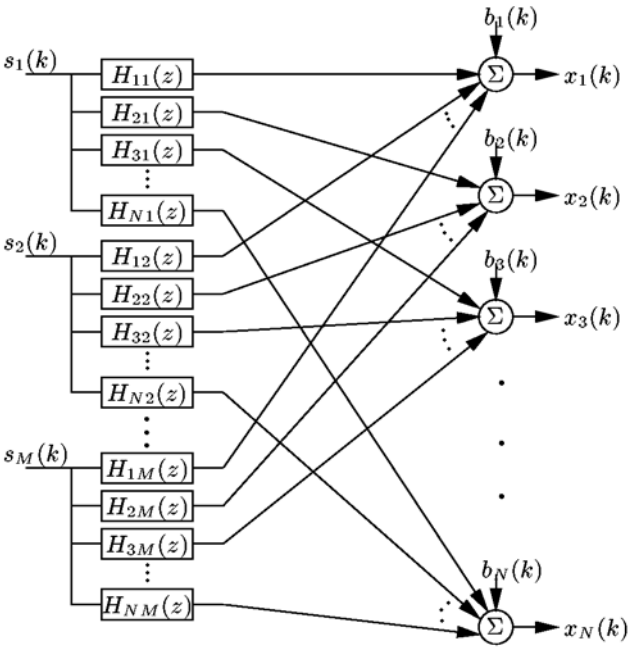

Fig. 12.1. Illustration of a MIMO FIR acoustic system having M sound sources and N microphones.

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

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

is a vector containing the last L_h samples of the m-th source signal s_m, and $b_n(k)$ is zero-mean additive white Gaussian noise (AWGN).

Using the z transform, the signal model of the MIMO system (12.1) is expressed as

$$X_n(z) = \sum_{m=1}^{M} H_{nm}(z) S_m(z) + B_n(z), \quad n = 1, 2, \cdots, N, \tag{12.2}$$

where

$$H_{nm}(z) = \sum_{l=0}^{L_h-1} h_{nm,l} z^{-l}. \tag{12.3}$$

In this system, we assume no *a priori* knowledge about the original speech signal $s_1(k)$, the interference signals $s_m(k)$ $(m = 2, \cdots, M)$, or the channel impulse responses h_{nm}. All that we have are microphone outputs $x_n(k)$

$(n = 1, 2, \cdots, N)$. But the speech enhancement algorithm that will be developed needs to know the channel impulse responses from interference sources to each microphone h_{nm} $(m = 2, \cdots, M, \forall n)$. Therefore we have to estimate them blindly. While blind MIMO identification is practically appealing, it still is a theoretically open problem and the current research in this area remains at the stage of feasibility investigation. This problem is already very difficult for communication systems with short channel impulse responses. Then for an acoustic system where the filter length of a channel impulse response in thousands of samples is not uncommon, blind MIMO identification seems formidable. Therefore we propose to decompose the problem into several subproblems in which SIMO systems are blindly identified. Presumably interference sources are motionless or move very slowly. Consequently their corresponding channel impulse responses change very slowly in time. It is further assumed that from time to time each interference source occupies at least one exclusive interval alone. Then during every single-talk interval, a SIMO system is blindly identified and its channel impulse responses are saved for later speech separation and dereverberation when all sources voice out simultaneously. Although these assumptions make the developed algorithm less flexible, they still are reasonable and stand in many practical scenarios. Apparently a sound source detection algorithm that distinguishes single and multiple talk is necessary and also interesting, but it is beyond the scope of this study.

In this chapter, we suppose that noise comes from one single point source or multiple point sources, and additive, dispersive noise is negligible, i.e., $b_n(k) = 0, \forall n, k$. Therefore, the blind SIMO identification system could yield accurate estimates of channel impulse responses and we can assure satisfactory performance for subsequent speech separation and dereverberation.

12.3 Blind Identification of a SIMO System

As assumed in the previous section, from time to time an interference source $s_m(k)$ $(m = 2, \cdots, M)$ would alone occupy an exclusive interval, during which the MIMO system becomes a SIMO system and the corresponding channel impulse responses will be blindly estimated. In this section, we will briefly review the technique of blind SIMO identification and its adaptive implementations. In order to have a concise presentation and to keep consistent with the conventional notation used in the literature of blind SIMO identification, we omit the subscript indicating the source index m in and also *only* in this section, which we believe would cause no ambiguity if the reader could pay slightly more attention.

For a SIMO system, we have the following expression for microphone signals:

$$x_n(k) = h_n * s(k) + b_n(k), \quad n = 1, 2, \cdots, N, \tag{12.4}$$

where the symbol $*$ denotes the linear convolution operator and $b_n(k)$ can be neglected by assumption as explained in the previous section. In a vector/matrix form, such a signal model (12.4) becomes:

$$\mathbf{x}_n(k) = \mathbf{H}_n \cdot \mathbf{s}(k), \tag{12.5}$$

where

$$\begin{aligned}
\mathbf{x}_n(k) &= \left[\, x_n(k) \;\; x_n(k-1) \;\; \cdots \;\; x_n(k-L_h+1) \,\right]^T, \\
\mathbf{H}_n &= \begin{bmatrix} h_{n,0} & h_{n,1} & \cdots & h_{n,L_h-1} & 0 & \cdots & 0 \\ 0 & h_{n,0} & \cdots & h_{n,L_h-2} & h_{n,L_h-1} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & h_{n,0} & h_{n,1} & \cdots & h_{n,L_h-1} \end{bmatrix}, \\
\mathbf{s}(k) &= \left[\, s(k) \;\; s(k-1) \;\; \cdots \;\; s(k-L_h+1) \;\; \cdots \;\; s(k-2L_h+2) \,\right]^T.
\end{aligned}$$

In order to ensure that the SIMO system can be blindly identified, the following two conditions (one on the channel diversity and the other on the input source signal) need to be met and are normally assumed in earlier studies as well as in this chapter [18]:

1. The polynomials formed from $\mathbf{h}_n = \left[\, h_{n,0}\, h_{n,1} \cdots h_{n,L_h-1} \,\right]^T$, $n = 1, 2, \cdots, N$, are co-prime, i.e., the channel transfer functions $H_n(z)$ do not share any common zeros;
2. The autocorrelation matrix $\mathbf{R}_{ss} = E\left\{\mathbf{s}(k)\mathbf{s}^T(k)\right\}$ of the source signal is of full rank (such that the SIMO system can be fully excited from a perspective of system identification), where $E\{\cdot\}$ denotes mathematical expectation.

The idea of blind SIMO identification using only second order statistics of the outputs was first proposed by Tong *et al.* [19] and now there are many different ways to explain the principle. We present here the one that we usually use in our research. It can be shown that the vector of channel impulse responses lies in the null space of a cross-correlation like matrix [20]:

$$\mathbf{R}_x \mathbf{h} = \mathbf{0}, \tag{12.6}$$

where

$$\begin{aligned}
\mathbf{R}_x &= \begin{bmatrix} \sum_{n\neq 1} \mathbf{R}_{x_n x_n} & -\mathbf{R}_{x_2 x_1} & \cdots & -\mathbf{R}_{x_N x_1} \\ -\mathbf{R}_{x_1 x_2} & \sum_{n\neq 2} \mathbf{R}_{x_n x_n} & \cdots & -\mathbf{R}_{x_N x_2} \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{R}_{x_1 x_N} & -\mathbf{R}_{x_2 x_N} & \cdots & \sum_{n\neq N} \mathbf{R}_{x_n x_n} \end{bmatrix}, \\
\mathbf{R}_{x_i x_j} &= E\left\{\mathbf{x}_i(k)\mathbf{x}_j^T(k)\right\}, \quad i, j = 1, 2, \cdots, N, \\
\mathbf{h} &= \left[\, \mathbf{h}_1^T \;\; \mathbf{h}_2^T \;\; \cdots \;\; \mathbf{h}_N^T \,\right]^T.
\end{aligned}$$

If the SIMO system is blindly identifiable, Matrix $\mathbf{R}_x$ is rank deficient by 1 (in the absence of noise) and channel impulse responses can be uniquely determined from $\mathbf{R}_x$, which contains only the second-order statistics of the system outputs. When additive noise is present, $\mathbf{h}$ would be the eigenvector of $\mathbf{R}_x$ corresponding to its smallest eigenvalue.

To develop an adaptive implementation, a simple way is to take advantage of the cross relations among the outputs [21]. By following the fact that

$$x_i * h_j = s * h_i * h_j = x_j * h_i, \quad i, j = 1, 2, \cdots, N, \ i \neq j, \tag{12.7}$$

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

$$\mathbf{x}_i^T(k)\mathbf{h}_j = \mathbf{x}_j^T(k)\mathbf{h}_i, \quad i, j = 1, 2, \cdots, N, \ i \neq j. \tag{12.8}$$

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:

$$e_{ij}(k+1) = \mathbf{x}_i^T(k+1)\hat{\mathbf{h}}_j(k) - \mathbf{x}_j^T(k+1)\hat{\mathbf{h}}_i(k), \quad i, j = 1, 2, \cdots, N, \tag{12.9}$$

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

$$\hat{\mathbf{h}}(k) = \left[\begin{array}{cccc} \hat{\mathbf{h}}_1^T(k) & \hat{\mathbf{h}}_2^T(k) & \cdots & \hat{\mathbf{h}}_N^T(k) \end{array} \right]^T,$$

leading to the normalized error signal

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

Accordingly, the cost function is formulated as:

$$J(k+1) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \epsilon_{ij}^2(k+1), \tag{12.10}$$

and the update equation of the multichannel LMS (MCLMS) algorithm is deduced as follows [21]:

$$\hat{\mathbf{h}}(k+1) = \hat{\mathbf{h}}(k) - \mu \nabla J(k+1), \tag{12.11}$$

where μ is a small positive step size,

$$\nabla J(k+1) = \frac{\partial J(k+1)}{\partial \hat{\mathbf{h}}(k)} = \frac{2\left[\tilde{\mathbf{R}}_x(k+1)\hat{\mathbf{h}}(k) - J(k+1)\hat{\mathbf{h}}(k)\right]}{\|\hat{\mathbf{h}}(k)\|^2}, \tag{12.12}$$

and

$$\tilde{\mathbf{R}}_x(k) = \left[\begin{array}{cccc} \sum_{n\neq 1} \tilde{\mathbf{R}}_{x_n x_n}(k) & -\tilde{\mathbf{R}}_{x_2 x_1}(k) & \cdots & -\tilde{\mathbf{R}}_{x_M x_1}(k) \\ -\tilde{\mathbf{R}}_{x_1 x_2}(k) & \sum_{n\neq 2} \tilde{\mathbf{R}}_{x_n x_n}(k) & \cdots & -\tilde{\mathbf{R}}_{x_N x_2}(k) \\ \vdots & \vdots & \ddots & \vdots \\ -\tilde{\mathbf{R}}_{x_1 x_N}(k) & -\tilde{\mathbf{R}}_{x_2 x_N}(k) & \cdots & \sum_{n\neq N} \tilde{\mathbf{R}}_{x_n x_n}(k) \end{array} \right],$$

$$\tilde{\mathbf{R}}_{x_i x_j}(k) = \mathbf{x}_i(k)\mathbf{x}_j^T(k), \quad i, j = 1, 2, \cdots, N.$$

The idea of adaptive blind SIMO identification could be implemented in the frequency domain for computational efficiency and fast convergence [22]. The 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 chapter.

12.4 Separating Reverberant Speech and Concurrent Interference

In this section, we will explain how to extract reverberant speech from concurrent interfering sound sources. It is supposed that channel impulse responses corresponding to the interfering sound (speech or noise) sources have been blindly identified using the method developed in the previous section. The knowledge of these channel impulse responses is being used here to convert the $M \times N$ MIMO system into a SIMO system with the speech signal of interest as the sole input. The development begins with an example of the simplest 2×3 MIMO system and then extends to a general $M \times N$ case with $M < N$.

12.4.1 Example: Removing Interference Signals in a 2×3 MIMO Acoustic System

For a 2×3 MIMO acoustic system, interference signals can be cancelled by using two microphone outputs at a time. For instance, we can cancel the interference in $X_1(z)$ and $X_2(z)$ caused by $S_2(z)$ (note that Source 1 is supposed to be the speech source of interest) as follows:

$$\begin{aligned} X_1(z)H_{22}(z) - X_2(z)H_{12}(z) = \\ \left[H_{11}(z)H_{22}(z) - H_{21}(z)H_{12}(z)\right] S_1(z) + \\ \left[H_{22}(z)B_1(z) - H_{12}(z)B_2(z)\right], \end{aligned} \quad (12.13)$$

where channel impulse responses $H_{12}(z)$ and $H_{22}(z)$ corresponding to Source 2 were blindly estimated ahead of time. Similarly, we can select different pair of microphone signals and obtain distinctive interference-free though distorted observations of $s_1(k)$. This procedure is visualized in Fig. 12.2 and will be described in a more systematic way in the following.

Let us consider the following equation:

$$\begin{aligned} Y_p(z) &= H_{s_1,p1}(z)X_1(z) + H_{s_1,p2}(z)X_2(z) + H_{s_1,p3}(z)X_3(z) \\ &= \sum_{q=1}^{3} H_{s_1,pq}(z)X_q(z), \quad p = 1, 2, 3, \end{aligned} \quad (12.14)$$

where $H_{s_1,pp}(z) = 0, \ \forall p$. This means that (12.14) considers only two microphone signals for each p. The objective is to find the polynomials $H_{s_1,pq}(z)$,

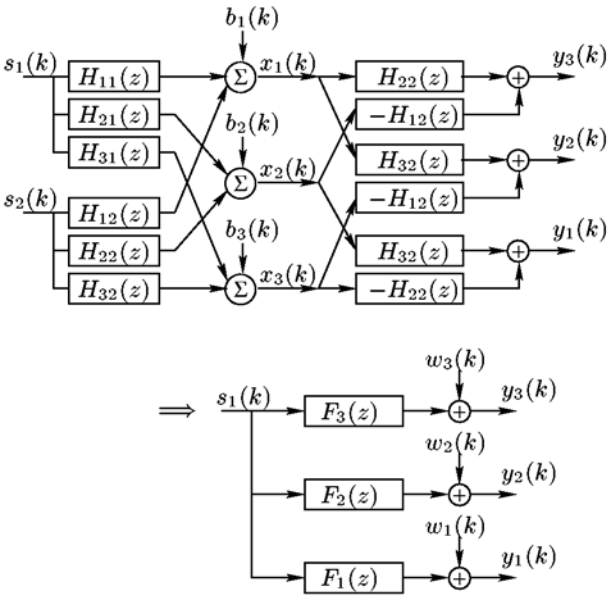

Fig. 12.2. Illustration of removing interference signals from a 2×3 MIMO acoustic system. Source 1 is the speech source of interest and Source 2 is an interference source as supposed.

$p, q = 1, 2, 3$, $p \neq q$, in such a way that:

$$Y_p(z) = F_p(z)S_1(z) + W_p(z), \quad p = 1, 2, 3, \tag{12.15}$$

which represents a SIMO system where s_1 is the source signal, $y_p(k), p = 1, 2, 3$, are the observed microphone signals, f_p are the corresponding acoustic channel impulse responses, and $w_p(k)$ is the noise at microphone p. If no dispersive noise is assumed, i.e., $b_n(k) = 0, \forall n, k$, then the noise component $w_p(k)$ is zero too.

Using (12.2) in (12.14), we deduce that:

$$\begin{aligned} Y_1(z) = {} & [H_{s_1,12}(z)H_{21}(z) + H_{s_1,13}(z)H_{31}(z)]\, S_1(z) + \\ & [H_{s_1,12}(z)H_{22}(z) + H_{s_1,13}(z)H_{32}(z)]\, S_2(z) + \\ & H_{s_1,12}(z)B_2(z) + H_{s_1,13}(z)B_3(z), \end{aligned} \tag{12.16}$$

$$\begin{aligned} Y_2(z) = {} & [H_{s_1,21}(z)H_{11}(z) + H_{s_1,23}(z)H_{31}(z)]\, S_1(z) + \\ & [H_{s_1,21}(z)H_{12}(z) + H_{s_1,23}(z)H_{32}(z)]\, S_2(z) + \\ & H_{s_1,21}(z)B_1(z) + H_{s_1,23}(z)B_3(z), \end{aligned} \tag{12.17}$$

$$\begin{aligned} Y_3(z) = {} & [H_{s_1,31}(z)H_{11}(z) + H_{s_1,32}(z)H_{21}(z)]\, S_1(z) + \\ & [H_{s_1,31}(z)H_{12}(z) + H_{s_1,32}(z)H_{22}(z)]\, S_2(z) + \\ & H_{s_1,31}(z)B_1(z) + H_{s_1,32}(z)B_2(z). \end{aligned} \tag{12.18}$$

As shown in Fig. 12.2, one possibility is to choose:

$$\begin{aligned} &H_{s_1,12}(z) = H_{32}(z), \;\; H_{s_1,13}(z) = -H_{22}(z), \\ &H_{s_1,21}(z) = H_{32}(z), \;\; H_{s_1,23}(z) = -H_{12}(z), \\ &H_{s_1,31}(z) = H_{22}(z), \;\; H_{s_1,32}(z) = -H_{12}(z). \end{aligned} \tag{12.19}$$

In this case, we find that:

$$\begin{aligned} F_1(z) &= H_{32}(z)H_{21}(z) - H_{22}(z)H_{31}(z), \\ F_2(z) &= H_{32}(z)H_{11}(z) - H_{12}(z)H_{31}(z), \\ F_3(z) &= H_{22}(z)H_{11}(z) - H_{12}(z)H_{21}(z), \end{aligned} \tag{12.20}$$

and

$$\begin{aligned} W_1(z) &= H_{32}(z)B_2(z) - H_{22}(z)B_3(z), \\ W_2(z) &= H_{32}(z)B_1(z) - H_{12}(z)B_3(z), \\ W_3(z) &= H_{22}(z)B_1(z) - H_{12}(z)B_2(z). \end{aligned} \tag{12.21}$$

Since $\deg\left[H_{nm}(z)\right] = L_h - 1$, where $\deg[\cdot]$ is the degree of a polynomial, therefore $\deg\left[F_p(z)\right] \leq 2L_h - 2$. We can see from (12.20) that polynomials $F_1(z)$, $F_2(z)$, and $F_3(z)$ 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 $C_2(z) = \gcd\left[H_{12}(z), H_{22}(z), H_{32}(z)\right]$, where $\gcd[\cdot]$ denotes the greatest common divisor of the polynomials involved. We have:

$$H_{n2}(z) = C_2(z)H'_{n2}(z), \quad n = 1, 2, 3. \tag{12.22}$$

It is clear that the signal s_2 in (12.14) can be cancelled by using the polynomials $H'_{n2}(z)$ [instead of $H_{n2}(z)$ as given in (12.19)], so that the SIMO system represented by (12.15) will change to:

$$Y'_p(z) = F'_p(z)S_1(z) + W'_p(z), \quad p = 1, 2, 3, \tag{12.23}$$

where

$$F'_p(z)C_2(z) = F_p(z), \quad W'_p(z)C_2(z) = W_p(z).$$

It should be pointed out that

$$\deg\left[F'_p(z)\right] \leq \deg\left[F_p(z)\right]$$

and that polynomials $F'_1(z)$, $F'_2(z)$, and $F'_3(z)$ share common zeros if and only if $H_{11}(z)$, $H_{21}(z)$, and $H_{31}(z)$ share common zeros.

12.4.2 Generalization

The approach to extracting reverberant speech from interference signals explained in the previous subsection on a simple example will be generalized

here to an (M, N) MIMO acoustic system with $M < N$. We begin with writing (12.2) into a vector/matrix form

$$\vec{\mathbf{X}}(z) = \mathbf{H}(z)\vec{\mathbf{S}}(z) + \vec{\mathbf{B}}(z), \tag{12.24}$$

where

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

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

$$\mathbf{H}(z) = \mathbf{H}'(z)\mathbf{C}(z), \tag{12.25}$$

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 $M \times M$ MIMO sub-system.

Consider the following equations:

$$Y_p(z) = \vec{\mathbf{H}}^T_{s_1,p}(z)\vec{\mathbf{X}}_p(z), \quad p = 1, 2, \cdots, P, \tag{12.26}$$

where

$$\begin{aligned}
\vec{\mathbf{H}}_{s_1,p}(z) &= \left[\begin{array}{cccc} H_{s_1,p1}(z) & H_{s_1,p2}(z) & \cdots & H_{s_1,pM}(z) \end{array} \right]^T, \\
\vec{\mathbf{X}}_p(z) &= \left[\begin{array}{cccc} X_{p_1}(z) & X_{p_2}(z) & \cdots & X_{p_M}(z) \end{array} \right]^T.
\end{aligned}$$

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 (12.24), we have

$$\vec{\mathbf{X}}_p(z) = \mathbf{H}_p(z)\vec{\mathbf{S}}(z) + \vec{\mathbf{B}}_p(z), \tag{12.27}$$

where

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

Substituting (12.27) into (12.26) yields

$$Y_p(z) = \vec{\mathbf{H}}^T_{s_1,p}(z)\mathbf{H}_p(z)\vec{\mathbf{S}}(z) + \vec{\mathbf{H}}^T_{s_1,p}(z)\vec{\mathbf{B}}_p(z). \tag{12.28}$$

In order to remove the interference from other competing speech or noise sources, the objective here is to find the vector $\vec{\mathbf{H}}_{s_1,p}(z)$ whose components are linear combinations of $H_{nm}(z)$ $(m = 2, 3, \cdots, M,\ n = 1, 2, \cdots N)$ such that

$$\boldsymbol{\phi}_p^T(z) \triangleq \vec{\mathbf{H}}_{s_1,p}^T(z)\mathbf{H}_p(z) = \left[\ F_p(z)\ \ 0\ \ \cdots\ \ 0\ \right]. \tag{12.29}$$

Consequently, we have

$$Y_p(z) = F_p(z)S_1(z) + W_p(z), \tag{12.30}$$

where

$$W_p(z) = \vec{\mathbf{H}}_{s_1,p}^T(z)\vec{\mathbf{B}}_p(z).$$

If $\mathbf{C}_p(z)$ [obtained from $\mathbf{C}(z)$ in a similar way as $\mathbf{H}_p(z)$ is constructed] is not equal to the identity matrix, then $\mathbf{H}_p(z) = \mathbf{H}'_p(z)\mathbf{C}_p(z)$, where $\mathbf{H}'_p(z)$ has full column normal rank in acoustic environments as we assume in this chapter[1] (i.e. nrank $\left[\mathbf{H}'_p(z)\right] = M$, see [23] for a definition of normal rank), and the interference-free observations of $s_1(k)$ are determined as follows

$$Y'_p(z) = \vec{\mathbf{H}}'^{\,T}_{s_1,p}(z)\mathbf{H}'_p(z)\mathbf{C}_p(z)\mathbf{S}(z) + \vec{\mathbf{H}}'^{\,T}_{s_1,p}(z)\vec{\mathbf{B}}_p(z). \tag{12.31}$$

The filter vector $\vec{\mathbf{H}}'_{s_1,p}(z)$ is chosen in a way such that

$$Y'_p(z) = F'_p(z)S_1(z) + W'_p(z). \tag{12.32}$$

Obviously a good choice is to let the i-th element of $\vec{\mathbf{H}}'_{s_1,p}(z)$ be the $(i,1)$-th cofactor[2] of $\mathbf{H}'_p(z)$. Consequently, the polynomial $F'_p(z)$ would be the determinant of $\mathbf{H}'_p(z)$. Note that the $(i,1)$-th cofactor of $\mathbf{H}'_p(z)$ is only a linear combination of $H_{nm}(z)$ or $H'_{nm}(z)$ $(m = 2, 3, \cdots, M,\ n = 1, 2, \cdots N)$. Therefore even though the channel impulse responses corresponding to the speech signal of interest $s_1(k)$ are not known or at least have not yet been blindly identified, we still are able to separate the reverberant speech from concurrent interference.

Since

$$\begin{aligned} F'_p(z) &= \vec{\mathbf{H}}'^{\,T}_{s,p}(z)\vec{\mathbf{H}}'_{p,:1}(z) \\ &= \sum_{q=1}^{M} H'_{s_1,pq}(z)H_{p,q1}(z), \end{aligned} \tag{12.33}$$

[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.

[2] The (i,j)-th cofactor c_{ij} of a matrix $\mathbf{A}$ is a signed version of $\mathbf{A}$'s minor d_{ij}:

$$c_{ij} \triangleq (-1)^{i+j} d_{ij},$$

where the minor d_{ij} is the determinant of a reduced matrix that is formed by omitting the i-th row and j-th column of the matrix A.

where $\vec{\mathbf{H}}_{p,:1}(z)$ is the first column vector of $\mathbf{H}_p(z)$ and $H'_{s_1,pq}(z)$ $(q = 1, 2, \cdots, M)$ are co-prime. It is clear that the polynomials $F'_p(z)$ $(p = 1, 2, \cdots, P)$ share common zeros if and only if the polynomials $H_{n1}(z)$ $(n = 1, 2, \cdots, N)$ share common zeros. Therefore, if the channels with respect to any one input are co-prime for the (M, N) MIMO acoustic system, we can remove interference from the reverberant speech of interest and obtain a SIMO system whose C_N^M channels are also co-prime.

Also, it can easily be check that $\deg[F_p(z)] \leq M(L_h - 1)$ (or $\deg\left[F'_p(z)\right] \leq M(L_h - 1)$). As a result, the length of the FIR filter f_p (or f'_p) would be

$$L_f \leq M(L_h - 1) + 1. \tag{12.34}$$

12.5 Speech Dereverberation

In the above, we showed that reverberant speech and competing interference could be separated given that the channel impulse responses corresponding to interference sources have been blindly identified. After this processing, we obtained a SIMO system with the speech signal of interest as the input. 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 to perfectly remove those annoying reverberation and how to recover the original speech signal from the SIMO system. Here the assumption that the channel impulse responses $H_{nm}(z)$, $\forall m$ $(n = 1, 2, \cdots, N)$ are co-prime (i.e., the MIMO system is irreducible) needs to be employed to blindly identify the SIMO system first and then to perform speech dereverberation by using the MINT theorem. Therefore, the outputs of the SIMO system are given by (12.30).

12.5.1 Principle

For the SIMO system with respect to the source of interest s_1, we intend to apply the MINT theorem (also called the Bezout theorem in the mathematic literature). Let's consider the polynomials $G_p(z)$ $(p = 1, 2, \cdots, P)$ and the equation:

$$\begin{aligned}\widehat{S}_1(z) &= \sum_{p=1}^{P} G_p(z) Yp(z) \\ &= \left[\sum_{p=1}^{P} Fp(z) G_p(z)\right] S_1(z) + \sum_{p=1}^{P} G_p(z) Wp(z).\end{aligned} \tag{12.35}$$

The polynomials $G_p(z)$ should be found in such a way that $\widehat{S}_1(z) = S_1(z)$ in the absence of noise by using the Bezout theorem which is mathematically

expressed as follows:

$$\begin{aligned}&\gcd\left[F_1(z), F_2(z), \cdots, F_P(z)\right] = 1\\ &\Leftrightarrow \exists\ G_1(z), G_2(z), \cdots, G_P(z)\ :\ \sum_{p=1}^{P} F_p(z)G_p(z) = 1.\end{aligned} \tag{12.36}$$

In other words, if the polynomials $F_p(z)$ $(p = 1, 2, \cdots, P)$ have no common zeros (which is equivalent to saying that the MIMO system is irreducible), it is possible to perfectly equalize (in the noiseless case) the SIMO system. The MINT theorem relieves the constraint on a single-channel acoustic system for perfect dereverberation that the channel impulse response must be a minimum-phase polynomial.

To find the dereverberation filters $G_p(z)$, we need to know the channel impulse responses $F_p(z)$. Since the MIMO system's channel impulse responses $H_{nm}(z)$, $\forall m$, $(n = 1, 2, \cdots, N)$ do not share common zeros as assumed in this chapter, the channel impulse responses f_p are co-prime as well such that they can be blindly identified again using the adaptive algorithms presented in Section 12.3. Starting from this point, we suppose that f_p's are known and we make no difference between f_p and its estimate.

Let's write the Bezout equation (12.36) in the time domain as follows:

$$\mathbf{F}_{\mathrm{c}}\mathbf{g} = \sum_{p=1}^{P} \mathbf{F}_{\mathrm{c},p}\mathbf{g}_p = \mathbf{e}_1, \tag{12.37}$$

where

$$\begin{aligned}\mathbf{F}_{\mathrm{c}} &= \left[\ \mathbf{F}_{\mathrm{c},1}\ \ \mathbf{F}_{\mathrm{c},2}\ \ \cdots\ \ \mathbf{F}_{\mathrm{c},P}\ \right],\\ \mathbf{g} &= \left[\ \mathbf{g}_1^T\ \ \mathbf{g}_2^T\ \ \cdots\ \ \mathbf{g}_P^T\ \right]^T,\\ \mathbf{g}_p &= \left[\ g_{p,0}\ \ g_{p,1}\ \ \cdots\ \ g_{p,L_g-1}\ \right]^T,\\ &\quad p = 1, 2, \cdots, P,\end{aligned}$$

L_g is the length of the FIR filter g_p,

$$\mathbf{F}_{\mathrm{c},p} = \begin{bmatrix} f_{p,0} & 0 & \cdots & 0\\ f_{p,1} & f_{p,0} & \cdots & 0\\ \vdots & \ddots & \ddots & \vdots\\ f_{p,L_f-1} & \cdots & \cdots & \vdots\\ 0 & f_{p,L_f-1} & \ddots & \vdots\\ \vdots & \vdots & \ddots & \vdots\\ 0 & \cdots & 0 & f_{p,L_f-1}\end{bmatrix}$$

is an $(L_f + L_g - 1) \times L_g$ matrix, L_f is the length of the FIR filter f_p, and

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

is an $(L_f + L_g - 1) \times 1$ vector. In order to have a unique solution for (12.37), L_g must be chosen in such a way that $\mathbf{F}_{\rm c}$ is a square matrix. In this case, we have:

$$L_g = \frac{L_f - 1}{P - 1}. \tag{12.38}$$

Using (12.34), the length of the dereverberation filter is bounded by

$$L_g \leq \frac{M(L_h - 1)}{P - 1}. \tag{12.39}$$

12.5.2 The Least-Squares Implementation

It is now clear that by using the Bezout theorem the SIMO system with respect to the speech source of interest can be perfected dereverberated 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 (12.39). 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 and error in the estimates of the channel impulse responses. Moreover, the smallest L_g may not be even possible since (12.38) does not guarantee an integer solution. Therefore, we choose a larger L_g than necessary in our implementation and solve (12.37) for $\mathbf{g}$ in the least squares sense:

$$\mathbf{g}_{\rm LS} = \mathbf{F}_{\rm c}^{\dagger}\mathbf{e}_1, \tag{12.40}$$

where

$$\mathbf{F}_{\rm c}^{\dagger} = \left(\mathbf{F}_{\rm c}^T\mathbf{F}_{\rm c}\right)^{-1}\mathbf{F}_{\rm c}^T$$

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

$$\mathbf{g}_{\rm LS} = \mathbf{F}_{\rm c}^{\dagger}\mathbf{e}_d, \tag{12.41}$$

where

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

Performing speech dereverberation based on the MINT theorem is sensitive to errors in the estimated channel impulse responses. In our research, we found that the performance of speech dereverberation would vary with the value of the decision delay d when a blind method has some difficulties to accurately identify the channels. Since this still is an open research problem, in our simulations, we either choose a fixed delay or search for the delay that produces the best speech dereverberation performance in the neighborhood of a pre-specified decision delay.

12.6 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.

12.6.1 Performance Measures

Similar to what was adopted in our earlier study [22], we will use the normalized projection misalignment (NPM) to evaluate the performance of a BCI algorithm [24]. The NPM is defined as:

$$\text{NPM} \triangleq 20\log_{10}\left[\frac{\|\boldsymbol{\epsilon}\|}{\|\mathbf{h}\|}\right], \tag{12.42}$$

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 [25] 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 $M < N$. Moreover, the M sources are equally important in [25] while here the first source is the speech source of interest and is more important than others.

Since only the first speech source is what we are interested in extracting, the SIR would be defined in a way where a component contributed by $s_1(k)$ is treated as the signal and the rest as the interference. We first define the input SIR at microphone n as:

$$\text{SIR}_n^{\text{in}} \triangleq \frac{E\left\{[h_{n1} * s_1(k)]^2\right\}}{\sum_{i=2}^{M} E\left\{[h_{ni} * s_i(k)]^2\right\}},\ n = 1, 2, \cdots, N, \tag{12.43}$$

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

$$\text{SIR}^{\text{in}} \triangleq \frac{1}{N}\sum_{n=1}^{N}\text{SIR}_n^{\text{in}}. \tag{12.44}$$

The output SIR is defined using the same principle but the expression will be more complicated. For a concise presentation, we denote ϕ_{p,s_i} $(p =$

$1, 2, \cdots, P, i = 1, 2, \cdots, M$) as the impulse response of the equivalent channel from the i-th source $s_i(k)$ to the output $y_p(k)$ for the p-th $M \times M$ separation subsystem. From (12.28) and (12.29), we know that ϕ_{p,s_i} corresponds to the i-th element of $\boldsymbol{\phi}_p(z)$ and $\phi_{p,s_1} = f_p$. Then the average output SIR for the p-th subsystem is:

$$\text{SIR}_p^{\text{out}} \triangleq \frac{E\left\{[f_p * s_1(k)]^2\right\}}{\sum_{i=2}^{M} E\left\{[\phi_{p,s_i} * s_i(k)]^2\right\}}, \quad p = 1, 2, \cdots, P. \tag{12.45}$$

Finally, the overall average output SIR is found as:

$$\text{SIR}^{\text{out}} \triangleq \frac{1}{P}\sum_{p=1}^{P} \text{SIR}_p^{\text{out}}. \tag{12.46}$$

To assess the quality of dereverberated speech signals, we employed the Itakura-Saito (IS) distortion measure [26], which is the ratio of the residual energies produced by the original speech when inverse filtered using the LP coefficients derived from the 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:

$$d_{\text{IS},t} = \frac{\boldsymbol{\alpha}_t'^{T}\mathbf{R}_{tt}\boldsymbol{\alpha}_t'^{T}}{\boldsymbol{\alpha}_t^{T}\mathbf{R}_{tt}\boldsymbol{\alpha}_t^{T}} - 1. \tag{12.47}$$

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 $d_{\text{IS},t}$ over all frames. According to [27], 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 [28] 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. After source separation, the IS measure is obtained by averaging the result with respect to each one of the P SIMO outputs $y_p(k)$ and is denoted by $d_{\text{IS}}^{\text{SS}}$. After speech dereverberation, the final IS measure is denoted by $d_{\text{IS}}^{\text{SD}}$.

12.6.2 Experimental Setup

The simulations were conducted with the impulse responses measured in the varechoic chamber at Bell Labs [29]. A diagram of the floor plan layout is

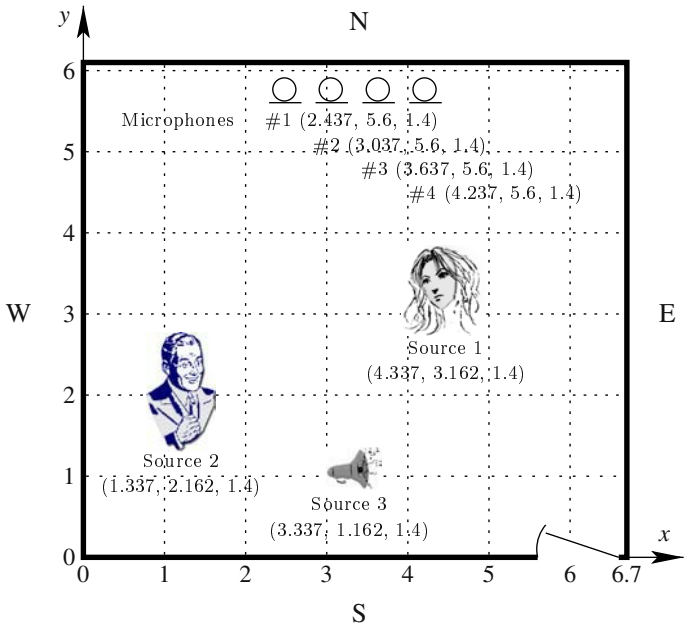

Fig. 12.3. Floor plan of the varechoic chamber at Bell Labs (coordinate values measured in meters).

shown in Fig. 12.3. 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 [30]. 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 can be simulated. In the database of channel impulse responses from [29], 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 omni-directional 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, four microphones and three speaker positions, which form a 3×4 MIMO system, were chosen and their locations are shown in Fig. 12.3. 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 $L_h = 512$. Among the three sources, the first female speaker's speech is the target for extraction. The other two sources include one male speaker and one noise source. The two speech sources are equally loud in volume while the noise source is 5 dB weaker than the speech sources. For the noise source, we tried two different kinds of noise. One is car noise and the other is babbling noise recorded in the New York Stock Exchange (NYSE). The time sequence and spectrogram (30 Hz bandwidth) of these source signals for the first 1.5 seconds are shown in Fig. 12.4. From the spectrograms, we can tell that car noise is a low-pass signal while the bandwidth of babbling noise is much wider. From the perspective of system identification, babbling noise is more favorable than car noise as an exciting source signal. 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.

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 [22]. For the batch method, the empirical spatial covariance matrix was obtained over the first 1500 samples of the microphone captures. For source separation and speech dereverberation, speech signals of duration 10 seconds were utilized to assess the performance. The decision delay d in (12.41) was fixed as $3L_h/2-1$ in the cases of employing a batch method for BCI while its best value was searched in the neighborhood of $3L_h/2-1$ when an adaptive BCI algorithm was utilized.

12.6.3 Experimental Results

Table 12.1 summarizes the results of 16 experiments with different combination of room acoustics, BCI method, and type of noise. Figures 12.5 and 12.6 visualizes what was observed in the experiment with 89% of panels open, the UNMCFLMS algorithm employed for BCI, and car noise used as the third source. Figure 12.7 shows the results for the experiment with all panels closed, the batch method employed for BCI, and babbling noise in the NYSE used as the third source.

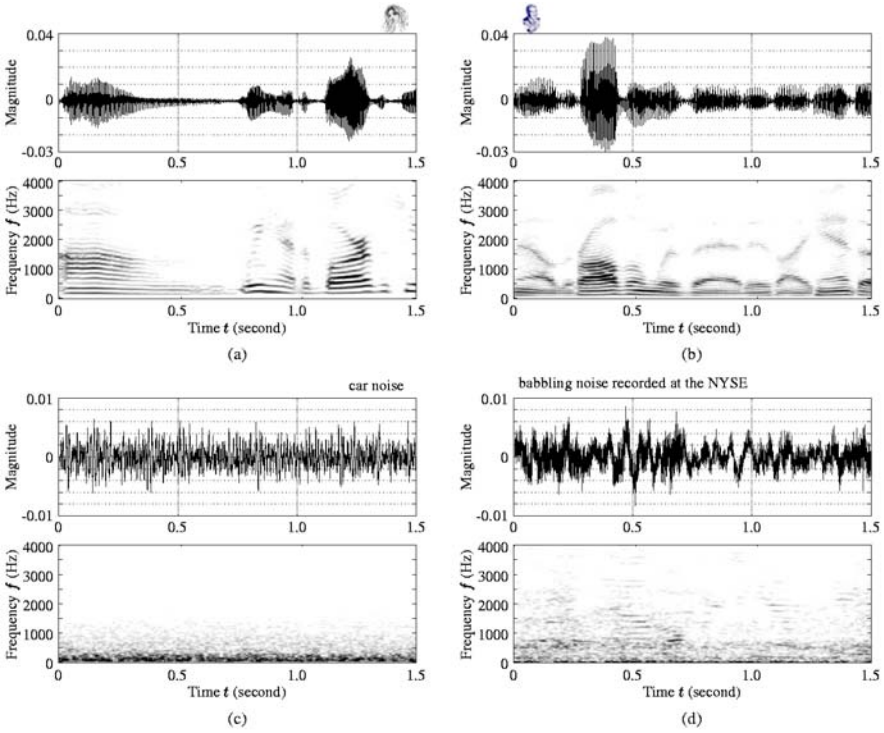

Fig. 12.4. Time sequence and spectrogram (30 Hz bandwidth) of the two speech source and two noise source signals used in the simulations for the first 1.5 seconds. (a) $s_1(k)$ (female speaker), (b) $s_2(k)$ (male speaker), (c) car noise, and (d) babbling noise recorded in the NYSE.

Let us first examine Table 12.1 and Fig. 12.5 for the accuracy of the channel impulse responses blindly estimated by the adaptive and batch BCI algorithms. It is clear that, 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 dependent on L_h. After it collects microphone outputs for only 1500 samples (equivalently 0.1875 second), the batch BCI method can produce a reliable channel estimate with less than -60 dB NPM for SIMO systems with long channels of length $L_h = 512$. However, performing SVD of a $N \cdot L_h \times N \cdot L_h$ 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 performance of the proposed blind source separation

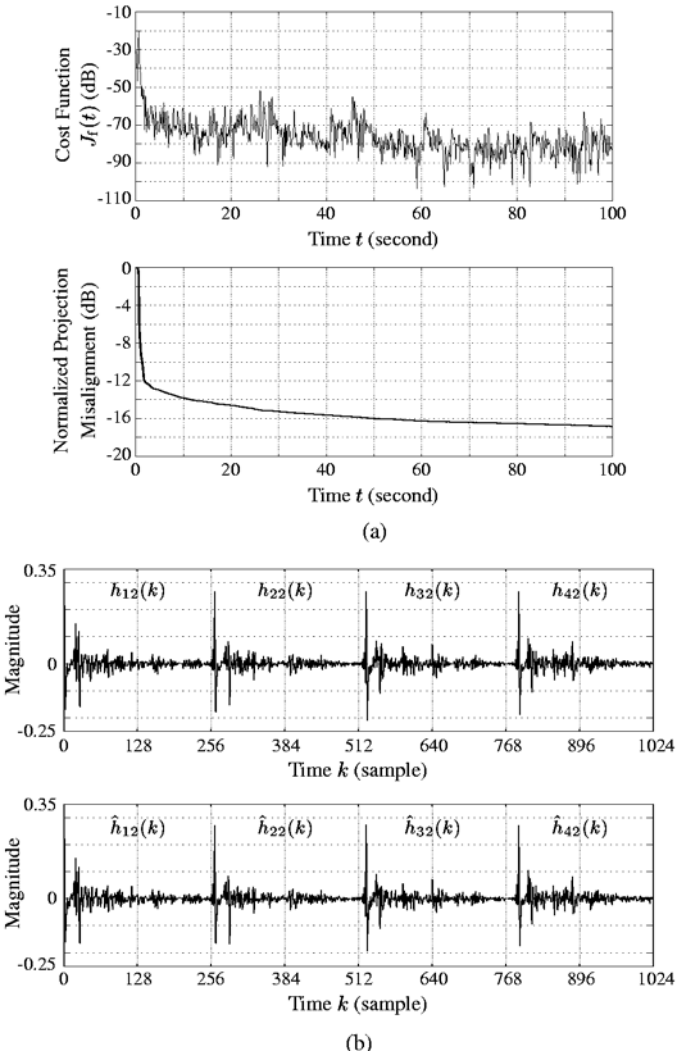

Fig. 12.5. Performance of the adaptive BCI (UNMCFLMS) algorithm with respect to the second source $s_2(k)$ in the varechoic chamber with 89% of panels open. (a) Running average (1000 samples) of the cost function and normalized projection misalignment, and (b) comparison of impulse responses between the actual channels and their estimates.

and speech dereverberation approach to speech enhancement with multiple microphones.

Figures 12.6 and 12.7 illustrate how the speech signal of interest is separated from other concurrent interference sources and how it is dereverberated.

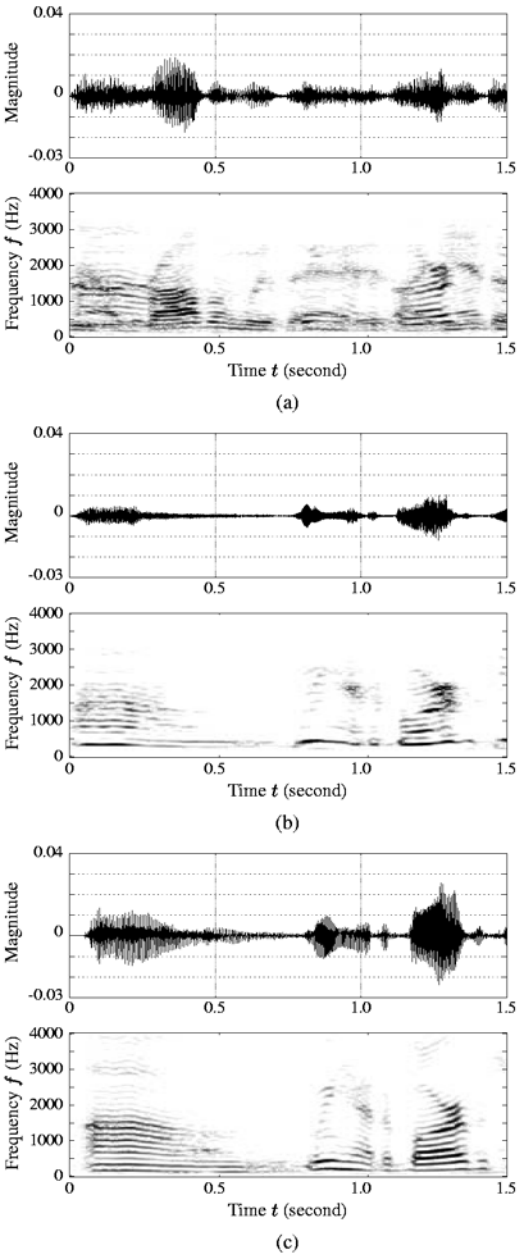

Fig. 12.6. Time sequence and spectrogram (30 Hz bandwidth) of (a) $x_1(k)$, (b) $y_1(k)$, and (c) $\hat{s}_1(k)$ for the experiment carried out in the varechoic chamber with 89% of panels open. In this experiment, $s_3(k)$ is car noise and the UNMCFLMS algorithm is used for BCI.

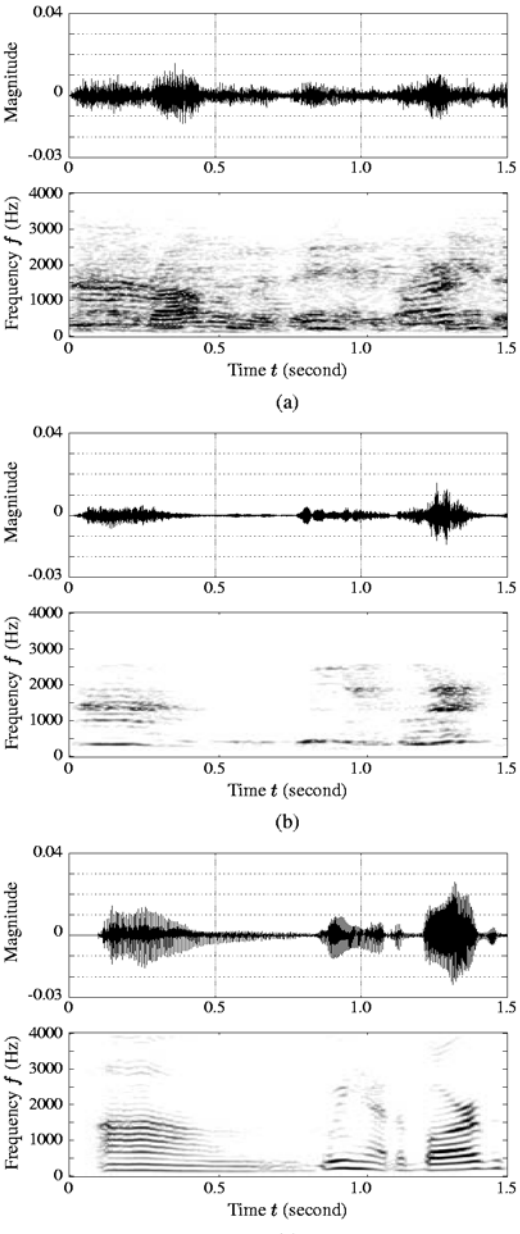

Fig. 12.7. Time sequence and spectrogram (30 Hz bandwidth) of (a) $x_1(k)$, (b) $y_1(k)$, and (c) $\hat{s}_1(k)$ for the experiment carried out in the varechoic chamber with all panels open. In this experiment, $s_3(k)$ is babbling noise recorded in the NYSE and the batch method is used for BCI.

Examining these figures together with the data in Table 12.1, we see that the output SIR's are very high (at least 41 dB) after source separation. Listening tests showed that these separated signals were certainly recognizable although they sounded more echoic as we expected. This can be justified by the spectrogram plots of $y_1(k)$ in Figs. 12.6(b) and 12.7(b). Apparently in periods of voiced speech on these narrow-band spectrograms, harmonics are vague, implying strong distortion which results in large IS measures (greater than at least 4.0). After dereverberation, the speech signal is satisfactorily recovered though delayed [clearly seen from time sequences of the recovered signal $\hat{s}_1(k)$ in these figures] with a relatively low IS measure. The accuracy of blindly identified channel impulse responses obviously has a great impact on the performance of the developed speech enhancement algorithm. But source separation and speech dereverberation are not equally affected by errors in the estimated channel impulse responses and the latter is more sensitive. When BCI is conducted with an adaptive algorithm, the NPM's are a lot lower than those obtained with a batch method. Although the final IS measures after speech dereverberation are significantly different particularly in more reverberant environments, their source separation performances in terms of output SIR are quite similar. Therefore it is our belief that using only SIR to evaluate a blind source separation (BSS) algorithm is inadequate if not misleading.

As explained before, the perceptual quality of distorted speech whose IS distance from its original signal is lower than 0.1 would not change with respect to either humans or an speech recognition system. The proposed algorithm incorporating the batch BCI method can surely deliver an enhanced speech signal reaching this level of voice quality. But the implementation based on an adaptive BCI algorithm can do so only when the room reverberation is low with 89% panels open. In reverberant environments, the adaptive BCI algorithm cannot produce highly accurate estimates of channel impulse responses such that the IS measures are still more than (though slightly) 0.1. As a matter of fact, it is imperative while challenging to develop accurate adaptive BCI algorithms for acoustic applications in reverberant environments. It is appealing that the recovered speech signal can attain high perceptual quality with an IS measure lower than 0.1. But in most applications of speech processing, this is an excessive and unnecessary if not practical requirement. What we observed in these simulations nevertheless show some promise of successful use of the proposed algorithm in prospect speech processing systems.

Table 12.1 Performance of the source separation and speech dereverberation algorithm based on the batch (SVD) and adaptive frequency-domain (AF) BCI implementations in the varechoic chamber at Bell Labs with different panel configurations.

		NPM (dB)			SIR^{in}	SIR^{out}		
Noise	BCI	SIMO_{s_1}	SIMO_{s_2}	SIMO_{s_3}	(dB)	(dB)	$d_{\text{IS}}^{\text{SS}}$	$d_{\text{IS}}^{\text{SD}}$
		89% panels open, $T_{60} = 240$ ms, $L_h = 256$						
Car	AF	-18.6361	-16.8234	-18.8090	1.3668	46.6966	4.4508	0.0449
	SVD	-84.6206	-110.3696	-152.6868	1.3668	47.6157	4.5653	0.0090
		75% panels open, $T_{60} = 310$ ms, $L_h = 256$						
	AF	-17.9231	-18.9300	-21.4186	2.3715	48.3984	5.3169	0.2389
	SVD	-109.6788	-100.6108	-187.8868	2.3715	48.8862	5.8647	0.0087
		30% panels open, $T_{60} = 380$ ms, $L_h = 512$						
	AF	-12.1323	-13.0353	-11.9475	1.3344	41.8391	5.8099	0.2609
	SVD	-67.0139	-106.2407	-167.2407	1.3344	43.5094	7.4319	0.0335
		Panels all closed, $T_{60} = 580$ ms, $L_h = 512$						
	AF	-12.5600	-13.5057	-14.3649	2.1065	44.1663	9.0386	0.2108
	SVD	-83.2605	-103.3190	-160.8024	2.1065	43.6628	11.1346	0.0198
NYSE	AF	-18.6361	-16.8234	-20.7545	0.9445	44.7547	4.4056	0.0668
	SVD	-84.6255	-110.3696	-176.5423	0.9445	45.2597	4.5653	0.0086
		75% panels open, $T_{60} = 310$ ms, $L_h = 256$						
	AF	-17.9231	-18.9300	-23.7211	1.8695	45.1628	5.4774	0.1920
	SVD	-100.3681	-114.6819	-184.1510	1.8694	44.9935	5.8647	0.0092
		30% panels open, $T_{60} = 380$ ms, $L_h = 512$						
	AF	-12.1323	-13.0353	-12.5460	0.8362	40.2743	5.6497	0.3215
	SVD	-79.5856	-93.7725	-174.1163	0.8362	41.4932	7.4319	0.0395
		Panels all closed, $T_{60} = 580$ ms, $L_h = 512$						
	AF	-12.5600	-13.5057	-16.8997	1.7245	42.2751	9.5378	0.1441
	SVD	-72.9542	-107.9821	-127.0545	1.7245	41.8808	11.1346	0.0192

NOTES: SIMO_{s_m} represents the SIMO system corresponding to source s_m. T_{60} denotes 60-dB reverberation time in the 20-4000 Hz band.

12.7 Conclusions

Capturing a speech signal of interest among a number of competing sound sources in reverberant environments is difficult and a close-talking microphone is a common engineering solution to this problem. But in many speech communication systems, untethered voice access is demanded and speech enhancement in the sense of source separation and dereverberation must be performed. 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 chapter that spatial interference from competing sources and temporal echoes due to room reverberation can be perfectly separated and a SIMO system with the speech signal of interest as input is extracted from the MIMO system. The channel matrices of the interference-free SIMO system is irreducible given that the channels from the same source in the MIMO system share no common zeros. For such a SIMO system, the speech is then restored by using the MINT theorem. 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 promise of the proposed algorithm.

References

1. E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Am.*, vol. 25, pp. 975–979, Sept. 1953.
2. B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, pp. 4–24, Apr. 1988.
3. Y. Huang, J. Benesty, and G. W. Elko, "Source localization," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds., Boston, MA: Kluwer Academic, 2004.
4. B. Widrow, P. E. Mantley, L. J. Griffiths, and B. B. Goode, "Adaptive antenna systems," *Proc. of the IEEE*, vol. 55, pp. 2143–2159, Dec. 1967.
5. J. Herault, C. Jutten, and B. Ans, "Detection de grandeurs primitives dans un message composite par une architecture de calul neuromimetique un apprentissage non supervise," in *Proc. GRETSI*, 1985.
6. P. Comon, "Independent component analysis: a new concept?," *Signal Processing*, vol. 36, pp. 287–314, Apr. 1994.
7. L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlations," *Phys. Rev. Lett.*, vol. 72, no. 23, pp. 3634–3637, June 1994.
8. J.-F. Cardoso, "Eigenstructure of the 4th-order cumulant tensor with application to the blind source separation problem," in *Proc. IEEE ICASSP*, 1989, pp. 2109–2112.
9. S. Amari, A. Cichocki, and H. H. Yang, "Blind signal separation and extraction: neural and information-theoretic approaches," in *Unsupervised Adaptive Filtering, Volume 1: Blind Source Separation*, S. Haykin, Ed., New York: John Wiley & Sons, 2000.
10. H. Wee and J. Principe, "A criterion for BSS based on simultaneous diagonalization of time correlation matrices," in *Proc. IEEE Workshop NNSP*, 1997, pp. 496–508.
11. M. Z. Ikram and D. R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment," in *Proc. IEEE ICASSP*, 2000, vol. 2, pp. 1041–1044.

12. L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 320–327, May 2000.
13. D. Bees, M. Blostein, and P. Kabal, "Reverberant speech enhancement using cepstral processing," in *Proc. IEEE ICASSP*, 1991, vol. 2, pp. 977–980.
14. S. Subramaniam, A. P. Petropulu, and C. Wendt, "Cepstrum-based deconvolution for speech dereverberation," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 392–396, Sept. 1996.
15. T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech based on harmonic structure," in *Proc. IEEE ICASSP*, 2003, vol. I, pp. 92–95.
16. A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
17. M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 145–152, Feb. 1988.
18. 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.
19. L. Tong, G. Xu, and T. Kailath, "A new approach to blind identification and equalization of multipath channels," in *Proc. 25th Asilomar Conf. on Signals, Systems, and Computers*, 1991, vol. 2, pp. 856–860.
20. C. Avendano, J. Benesty, and D. R. Morgan, "A least squares component normalization approach to blind channel identification," in *Proc. IEEE ICASSP*, 1999, vol. 4, pp. 1797–1800.
21. Y. Huang and J. Benesty, "Adaptive multi-channel least mean square and Newton algorithms for blind channel identification," *Signal Processing*, vol. 82, pp. 1127–1138, Aug. 2002.
22. Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multi-channel identification," *IEEE Trans. Signal Processing*, vol. 51, pp. 11–24, Jan. 2003.
23. P. P. Vaidyanathan, *Multirate Systems and Filter Bank*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
24. D. R. Morgan, J. Benesty, and M. M. Sondhi, "On the evaluation of estimated impulse responses," *IEEE Signal Processing Lett.*, vol. 5, pp. 174–176, July 1998.
25. M. Z. Ikram and D. R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment," in *Proc. IEEE ICASSP*, 2000, vol. 2, pp. 1041–1044.
26. L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
27. S. R. Quackenbush, T. P. Barnwell, M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
28. G. Chen, S. N. Koh, I. Y. Soon, "Enhanced Itakura measure incorporating masking properties of human auditory system," *Elsevier Science Signal Processing*, vol. 83, pp. 1445–1456, July 2003.
29. A. Härmä, "Acoustic measurement data from the varechoic chamber," Technical Memorandum, Agere Systems, Nov. 2001.
30. 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 Symposium*, 1994, pp. 343–346.