A Minimum Distortion Noise Reduction Algorithm With Multiple Microphones

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

***Abstract*—The problem of noise reduction using multiple microphones has long been an active area of research. Over the past few decades, most efforts have been devoted to beamforming techniques, which aim at recovering the desired source signal from the outputs of an array of microphones. In order to work reasonably well in reverberant environments, this approach often requires such knowledge as the direction of arrival (DOA) or even the room impulse responses, which are difficult to acquire reliably in practice. In addition, beamforming has to compromise its noise reduction performance in order to achieve speech dereverberation at the same time. This paper presents a new multichannel algorithm for noise reduction, which formulates the problem as one of estimating the speech component observed at one microphone using the observations from all the available microphones. This new approach explicitly uses the idea of spatial–temporal prediction and achieves noise reduction in two steps. The first step is to determine a set of inter-sensor optimal spatial–temporal prediction transformations. These transformations are then exploited in the second step to form an optimal noise-reduction filter. In comparison with traditional beamforming techniques, this new method has many appealing properties: it does not require DOA information or any knowledge of either the reverberation condition or the channel impulse responses; the multiple microphones do not have to be arranged into a specific array geometry; it works the same for both the far-field and near-field cases; and, most importantly, it can produce very good and robust noise reduction with minimum speech distortion in practical environments. Furthermore, with this new approach, it is possible to apply postprocessing filtering for additional noise reduction when a specified level of speech distortion is allowed.**

***Index Terms*—Beamforming, generalized sidelobe canceller (GSC), linearly constrained minimum variance (LCMV), microphone arrays, minimum-mean-square error (MMSE), minimum variance distortionless response (MVDR), noise reduction, speech enhancement.**

I. Introduction

ACOUSTIC noise is ubiquitous and can have a profound impact on human-to-human and human-to-machine communications, including modifying the characteristics of the speech signal, degrading speech quality and intelligibility, and affecting the listener's perception and a machine's processing of recorded speech. In order to make voice communication feasible, natural, and comfortable in the presence of noise regardless of the noise level, it is desirable to develop digital signal processing techniques to "clean up" the noise-corrupted signal before it is stored, analyzed, transmitted, or played out. This problem is often referred to as either noise reduction or speech enhancement. It has been an active research area since the spectral-subtraction technique was invented in the middle 1960s [1]–[3]. Over the past few decades, researchers and engineers have approached this challenging problem by exploiting different facets of the properties of speech and noise signals, and a large number of algorithms have been developed. By and large, the developed solutions can be categorized into two broad classes depending on the number of microphone channels: single-channel and multichannel techniques.

In the single-channel situation, the observed microphone signal is modeled as a superposition of the clean speech and noise. An estimate of the clean speech is obtained by passing the noisy speech through a linear (time-varying) filter/transformation. Since speech and noise normally have very different characteristics, the filter/transformation can be designed to significantly attenuate the noise level without dramatically distorting the speech signal. The representative algorithms in this group include Wiener filters [3]–[7], subspace methods [8], statistical estimators [9]–[11], and speech-model-based approaches [12]–[15]. The single-channel techniques have many appealing properties. For example, they can be integrated into most existing communication devices without requiring architectural changes, and they are in general economic to implement. However, with this class of techniques, speech distortion is unavoidable and the amount of speech distortion is in general proportional to the amount of noise reduction [16]. So, the more the noise is reduced, the more the speech is distorted.

In order to control the amount of speech distortion while achieving noise reduction, tremendous attention has been paid to the use of multiple microphones. In this scenario, each microphone output can be modeled as the source speech signal convolved with the corresponding acoustic channel impulse response and then corrupted by background noise. The noise-reduction problem is typically formulated as one of estimating the source signal from the multiple microphone observations. The most straightforward approach to the problem is the delay-and-sum beamformer [17]. The basic underlying idea can be described as synchronizing-and-adding. If we assume that the acoustic channels are free of reverberation, the signal components across all sensors can be synchronized by delaying (or advancing) each microphone output by a proper amount of time. When these aligned signals are weighted and summed together, the signal components will be combined coherently and hence

Manuscript received August 6, 2007; revised November 22, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Israel Cohen.

J. Chen is with Bell Labs, Alcatel-Lucent, Murray Hill, NJ 07974 USA (e-mail: jingdong@research.bell-labs.com).

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

Y. Huang is with WeVoice, Inc., Bridgewater, NJ 08807 USA.

Digital Object Identifier 10.1109/TASL.2007.914969

1558-7916/$25.00 © 2008 IEEE

reinforced. In contrast, the noise signals are added up incoherently (in power) due to their random nature. This results in a gain factor for the signal-to-noise ratio (SNR).

Because phase delay is frequency dependent, the delay-and-sum idea is good only for narrowband signals. For broadband speech, the directivity pattern of a delay-and-sum beamformer would not be the same across a broad frequency band. If we use such a beamformer, when the steering direction is different from the source incident angle, the source signal will be low-pass filtered. In addition, noise coming from a direction different from the beamformer's look direction will not be uniformly attenuated over its entire spectrum. This "spectral tilt" results in a disturbing artifact in the array output [18]. One way to overcome this problem is to perform narrowband decomposition and design narrowband beamformers independently at each frequency. This structure is equivalent to applying a finite-duration impulse response (FIR) filter to each microphone output and then summing the filtered signals together. Therefore, this method is often referred to as filter-and-sum beamforming, which was first introduced by Frost [19].

Traditionally, the filter coefficients for a filter-and-sum beamformer are determined based on a prespecified beampattern and hence are independent of the signal characteristics and room reverberation condition. This so-called fixed beamforming method performs reasonably well in anechoic situations where the speech component observed at each microphone is purely a delayed and attenuated copy of the source signal. However, its performance (in terms of noise reduction and speech distortion) degrades significantly in practical acoustic environments where reverberation is inevitable. One way to improve noise-reduction performance in the presence of reverberation is to compute the filter coefficients in an adaptive way based on the room propagation condition. For example, if we know (or can estimate) the signal incident angle, we can optimize the filter coefficients and steer the beamformer's look direction such that the desired signal is passed through without attenuation while the signal contributions from all other directions are minimized [21]. This so-called minimum variance distortionless response (MVDR) or Capon method can dramatically improve the beamformer's noise-reduction performance. However, the speech distortion with this method is also substantial in real acoustic environments [22]. In order to minimize speech distortion, more sophisticated adaptive algorithms such as linearly constrained minimum variance (LCMV) [19]–[29], generalized sidelobe canceller (GSC) [25], [30], [31], and multiple-input/output inverse theorem (MINT) [32] were developed. These approaches use the acoustic channel impulse responses from the desired sources to the multiple microphones to determine the beamforming filter coefficients. They can achieve high performance when the channel impulse responses are known *a priori* (or can be estimated accurately) and the background noise level is low. However, the performance is very sensitive to the measurement error of channel impulse responses and a small amount of measurement error can lead to significant performance degradation.

Note that the single-channel and beamforming techniques formulate the noise-reduction problem in a very different way. Specifically, the former expresses the problem as one of estimating the speech component (speech source filtered by the room impulse response) in the microphone observation, while the latter formulates the problem as one of estimating the original source signal. So, unlike the single-channel methods, which exclusively focus on noise reduction, beamforming actually tries to solve both speech dereverberation and noise reduction at the same time. However, speech dereverberation alone is a very difficult task, and there have not been any good, practical solutions so far. If we consider both dereverberation and noise reduction at the same time, this would only make the problem more complicated.

Recently, much efforts have been made to reformulate the beamforming problem so that noise reduction can be achieved without performing speech dereverberation [33]–[36]. Similar to the single-channel techniques, this new formulation focuses on estimating the speech component observed at one microphone using observations from an array of microphones, so it can be viewed as a generalization of the single-channel noise reduction to the multichannel case. Among the recently developed multichannel noise-reduction approaches, the so-called transfer function GSC (TF-GSC) [33], [36] is of particular interest. This approach approximates the linear convolution in the discrete-Fourier-transform (DFT) domain using the circular convolution. It then explicitly exploits the channel diversity through the so-called relative transfer function (RTF) to estimate the short-time speech spectrum and achieves noise reduction. However, the estimation of RTF, which has to rely on the nonstationarity of the source signal [37]–[39], is not a trivial problem. So further research efforts are indispensable to explore new signal models and develop new algorithms.

In this paper, we develop a new noise-reduction approach. Similarly to the single-channel and recently formulated multichannel techniques, we put aside speech dereverberation and formulate the problem as one of estimating the speech component observed at one of the multiple microphones. This new approach achieves noise reduction in two steps. The first step is to determine a set of inter-sensor optimal spatial–temporal prediction transformations, which takes into account not only the channel diversity, but also the source self-correlation information. These optimal transformations are then used in the second step to form an optimal noise-reduction filter under the constraint of no speech distortion. It will be shown that our approach has many appealing properties over beamforming techniques, including but not limited to the following: 1) it does not require array geometry information; 2) there is no need to estimate either the DOA or the room impulse responses; 3) it works the same for both the far-field and near-field cases; and, 4) it can produce very good and robust noise reduction with practically minimum speech distortion.

II. Problem Description

The problem considered in this paper is illustrated in Fig. 1, where we have a speech source in the sound field and use N microphones to collect signals from their field of view. The output of the nth microphone is given by

$$\begin{aligned} y_n(k) &= s(k) * g_n + v_n(k) \\ &= x_n(k) + v_n(k), \qquad n = 1, 2, \ldots, N \end{aligned} \tag{1}$$

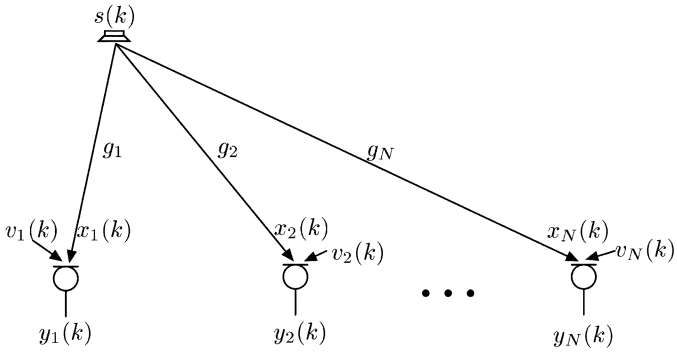


Fig. 1. Illustration of a multiple-microphone system.

where $*$ denotes convolution, $s(k)$ is the source signal, g_n represents the acoustic channel impulse response from the source to microphone n, and $x_n(k), v_n(k)$, and $y_n(k)$ are, respectively, the speech, the background noise, and their composite observed at the nth microphone. It is assumed that both $x_n(k)$ and $v_n(k)$ are zero-mean random processes that are mutually uncorrelated with each other. It is also assumed that the noise signals $v_n(k), n = 1, 2, \ldots, N$ are not completely coherent.

In traditional beamforming-based techniques, the problem is formulated as one of estimating the source signal $s(k)$ from the observed noisy signals $y_n(k)$. This would involve two subtasks, i.e., speech dereverberation and noise reduction. In this paper, similar to some recently developed multichannel noise-reduction techniques [33]–[36], we put aside speech dereverberation and focus exclusively on noise reduction. So, the problem considered here can be described as one of estimating the speech component observed at one microphone from the noisy signals received at all N microphones. Let us assume that we want to estimate the speech signal at the mth $(1 \leq m \leq N)$ microphone. Then, the objective of this paper is to estimate $x_m(k)$, given $y_n(k), n = 1, 2, \ldots, N$.

Putting the signal model (1) into vector/matrix form, we have

$$\begin{aligned}\mathbf{y}_n(k) &= \mathbf{G}_n\mathbf{s}(k) + \mathbf{v}_n(k) \\ &= \mathbf{x}_n(k) + \mathbf{v}_n(k), \qquad n = 1, 2, \ldots, N \end{aligned} \tag{2}$$

where, as shown in (3a)–(3c) at the bottom of the page, $\mathbf{G}_n$ is the channel (Sylvester) matrix of size $L \times L_s, L_s = L + L_g - 1, L_g$ is the length of the channel impulse responses, $[\cdot]^T$ denotes the transpose of a vector or a matrix, and $\mathbf{v}_n(k)$ and $\mathbf{x}_n(k)$ are defined similarly to $\mathbf{y}_n(k)$. With this vector/matrix form of the signal model, the noise-reduction problem considered in this paper can be described as one of estimating the speech signal vector $\mathbf{x}_m(k)$, given the observed signal vectors $\mathbf{y}_n(k)$, $n = 1, 2, \ldots, N$.

III. Sample-by-Sample-Based MMSE Estimator Using Multiple Microphones

In this section, we derive a minimum-mean-square-error (MMSE) estimator of $x_m(k)$.

A. MMSE Estimator

With the signal model given in (1), an estimate of the speech component $x_m(k)$ can be obtained by passing the N observed signals through N temporal filters, i.e.,

$$\begin{aligned}\hat{x}_m(k) &= \mathbf{h}_{1m}^T\mathbf{y}_1(k) + \mathbf{h}_{2m}^T\mathbf{y}_2(k) + \cdots + \mathbf{h}_{Nm}^T\mathbf{y}_N(k) \\ &= \sum_{n=1}^{N}\mathbf{h}_{nm}^T\mathbf{y}_n(k) \end{aligned} \tag{4}$$

where

$$\mathbf{h}_{nm} = [h_{nm,0} \quad h_{nm,1} \quad \cdots \quad h_{nm,L-1}]^T, \qquad n = 1, 2, \ldots, N$$

are the N FIR filters of length L, and $\mathbf{y}_n(k), n = 1, 2, \ldots, N$, are the observation signal vectors as defined in (3a). The corresponding error signal obtained by this estimation is written as

$$\begin{aligned}e_m(k) &\triangleq \hat{x}_m(k) - x_m(k) \\ &= \sum_{n=1}^{N}\mathbf{h}_{nm}^T\mathbf{y}_n(k) - x_m(k). \end{aligned} \tag{5}$$

Substituting (2) into (5), we can decompose the above error signal into the following form:

$$e_m(k) = e_{x,m}(k) + e_{v,m}(k) \tag{6}$$

where

$$e_{x,m}(k) \triangleq \sum_{n=1}^{N}\mathbf{h}_{nm}^T\mathbf{x}_n(k) - x_m(k) \tag{7}$$

and

$$e_{v,m}(k) \triangleq \sum_{n=1}^{N}\mathbf{h}_{nm}^T\mathbf{v}_n(k). \tag{8}$$

The term $e_{x,m}(k)$ quantifies how much the speech sample $x_m(k)$ is distorted due to the filtering operation. The larger the mean-square value of $e_{x,m}(k)$, the higher the speech

$$\mathbf{y}_n(k) = [y_n(k) \quad y_n(k-1) \quad \cdots \quad y_n(k-L+1)]^T \tag{3a}$$

$$\mathbf{s}(k) = [s(k) \quad s(k-1) \quad \cdots \quad s(k-L_s+1)]^T \tag{3b}$$

$$\mathbf{G}_n = \begin{bmatrix} g_{n,0} & g_{n,1} & \cdots & \cdots & \cdots & g_{n,L_g-1} & 0 & 0 & \cdots & 0 \\ 0 & g_{n,0} & g_{n,1} & \cdots & \cdots & \cdots & g_{n,L_g-1} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & g_{n,0} & g_{n,1} & \cdots & \cdots & \cdots & \cdots & g_{n,L_g-1} \end{bmatrix} \tag{3c}$$

distortion. In comparison, the term $e_{v,m}(k)$ tells how much the noise is reduced. The smaller the mean-square value of $e_{v,m}(k)$, the higher the noise reduction. So, ideally, noise reduction is a problem of finding an optimal set of the filters $\mathbf{h}_{nm}(n=1,2,\ldots,N)$ such that the mean-square error (MSE) corresponding to the residual noise is minimized while keeping the speech distortion $e_{x,m}(k)$ close to 0.

From (8), we can write the MSE associated with the residual noise as

$$\begin{aligned} J_{v,m}(\mathbf{h}_m) &= E\left[e_{v,m}^2(k)\right] \\ &= \mathbf{h}_m^T\mathbf{R}_{vv}\mathbf{h}_m \end{aligned} \tag{9}$$

where $E[\cdot]$ denotes mathematical expectation

$$\mathbf{h}_m = \begin{bmatrix}\mathbf{h}_{1m}^T & \mathbf{h}_{2m}^T & \cdots & \mathbf{h}_{Nm}^T\end{bmatrix}^T \tag{10}$$

$$\mathbf{R}_{vv} = E[\mathbf{v}(k)\mathbf{v}^T(k)] \tag{11}$$

is the noise correlation matrix, and

$$\mathbf{v}(k) = \begin{bmatrix}\mathbf{v}_1^T(k) & \mathbf{v}_2^T(k) & \cdots & \mathbf{v}_N^T(k)\end{bmatrix}^T. \tag{12}$$

Now, the noise-reduction problem can be mathematically formulated as follows:

$$\mathbf{h}_{m,\mathrm{o}} = \arg\min_{\mathbf{h}_m} J_{v,m}(\mathbf{h}_m) \quad \text{subject to} \quad e_{x,m}(k) = 0. \tag{13}$$

The solution to (13) depends on the number of microphones. We have two cases: $N=1$ and $N\geq 2$.

Case 1: $N=1$*:* In this case, we have $m=N=1$. If the current speech sample $x_1(k)$ cannot be completely predicted from its past samples (which is generally true in practice), we can easily check that the solution to (13) is

$$\mathbf{h}_{1,\mathrm{o}} = \mathbf{u}_1 \tag{14}$$

where

$$\mathbf{u}_1 = \begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix}^T \tag{15}$$

is a unit vector of length L. With this degenerate filter, there will be no noise reduction. So, in the single-channel scenario, if we want to keep the speech undistorted, there will be no noise reduction. However, if we still want to achieve some noise reduction, we need to loosen the constraint to allow some speech distortion. Indeed, this is almost the *de facto* standard practice in the existing single-channel noise-reduction techniques, where noise reduction is achieved by trading off speech distortion [8], [16]

Case 1: $N\geq 2$*:* In the single-channel situation, there is a fundamental compromise between noise reduction and speech distortion. However, if we use multiple microphones, we can take advantage of the redundancy among the microphones to achieve noise reduction without introducing any speech distortion.

Let us assume that we can find N spatial–temporal prediction matrices, $\mathbf{W}_{nm}(n=1,2,\ldots,N)$, such that

$$\mathbf{x}_n(k) = \mathbf{W}_{nm}\mathbf{x}_m(k),\ n=1,2,\ldots,N. \tag{16}$$

Apparently, for $n=m$, we have $\mathbf{W}_{mm}=\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. We will discuss later how to determine an optimal estimate of the matrix $\mathbf{W}_{nm}$ for $n\neq m$; but for now, we assume that $\mathbf{W}_{nm}$ are known. Substituting (16) into (7), we obtain

$$e_{x,m}(k) = \mathbf{x}_m^T(k)[\mathbf{W}_m\mathbf{h}_m - \mathbf{u}_1] \tag{17}$$

where

$$\mathbf{W}_m = \begin{bmatrix}\mathbf{W}_{1m}^T & \mathbf{W}_{2m}^T & \cdots & \mathbf{W}_{Nm}^T\end{bmatrix}. \tag{18}$$

With this expression of the speech distortion, we can rewrite the constrained estimation problem (13) in the following form:

$$\mathbf{h}_{m,\mathrm{o}} = \min_{\mathbf{h}_m} J_{v,m}(\mathbf{h}_m) \quad \text{subject to} \quad \mathbf{W}_m\mathbf{h}_m = \mathbf{u}_1. \tag{19}$$

If we use a Lagrange multiplier to adjoin the constraint to the cost function, the estimation problem in (19) can be written as

$$\mathbf{h}_{m,\mathrm{o}} = \arg\min_{\mathbf{h}_m} \mathcal{L}(\mathbf{h}_m,\boldsymbol{\lambda}) \tag{20}$$

where

$$\begin{aligned} \mathcal{L}(\mathbf{h}_m,\boldsymbol{\lambda}) &= J_{v,m}(\mathbf{h}_m) + \boldsymbol{\lambda}^T(\mathbf{W}_m\mathbf{h}_m - \mathbf{u}_1) \\ &= \mathbf{h}_m^T\mathbf{R}_{vv}\mathbf{h}_m + \boldsymbol{\lambda}^T(\mathbf{W}_m\mathbf{h}_m - \mathbf{u}_1) \end{aligned}$$

and vector $\boldsymbol{\lambda}$ is the Lagrange multiplier. Evaluating the gradient of $\mathcal{L}(\mathbf{h}_m,\boldsymbol{\lambda})$ with respect to $\mathbf{h}_m$ and equating the result to zero produces

$$\frac{\partial}{\partial\mathbf{h}_m}\mathcal{L}(\mathbf{h}_m,\boldsymbol{\lambda}) = 2\mathbf{R}_{vv}\mathbf{h}_m + \mathbf{W}_m^T\boldsymbol{\lambda} = \mathbf{0}. \tag{21}$$

From (21) and using the constraint, we find the solution to (20) (assuming that the noise signals at the microphones are not completely coherent so that the noise covariance matrix $\mathbf{R}_{vv}$ is full rank):

$$\mathbf{h}_{m,\mathrm{o}} = \mathbf{R}_{vv}^{-1}\mathbf{W}_m^T\left[\mathbf{W}_m\mathbf{R}_{vv}^{-1}\mathbf{W}_m^T\right]^{-1}\mathbf{u}_1. \tag{22}$$

We see that, in order to compute the optimal filter $\mathbf{h}_{m,\mathrm{o}}$, we need to know the two matrices $\mathbf{R}_{vv}$ and $\mathbf{W}_m$. The noise correlation matrix $\mathbf{R}_{vv}$ can be estimated during periods where speech is absent. In the next subsection, we will elaborate on this and discuss how to determine the $\mathbf{W}_m$ matrix.

B. Estimation of the $\mathbf{W}_m$ *Matrix*

From (16), we can construct the following MSE cost function:

$$J(\mathbf{W}_{nm}) = E\{[\mathbf{x}_n(k) - \mathbf{W}_{nm}\mathbf{x}_m(k)]^T[\mathbf{x}_n(k) - \mathbf{W}_{nm}\mathbf{x}_m(k)]\}. \tag{23}$$

Differentiating $J(\mathbf{W}_{nm})$ with respect to $\mathbf{W}_{nm}$ and equating the result to zero, we can obtain an optimal estimate of the $\mathbf{W}_{nm}$ matrix:

$$\mathbf{W}_{nm,\mathrm{o}} = \mathbf{R}_{x_n x_m}\mathbf{R}_{x_m x_m}^{-1} \tag{24}$$

where $\mathbf{R}_{x_n x_m} = E\{\mathbf{x}_n(k)\mathbf{x}_m^T(k)\}$ and $\mathbf{R}_{x_m x_m} = E\{\mathbf{x}_m(k)\mathbf{x}_m^T(k)\}$ are, respectively, the cross-correlation and correlation matrices of the speech signals.

Using the signal model given in (2), we can easily see that

$$\mathbf{R}_{x_n x_m} = \mathbf{G}_n\mathbf{R}_{ss}\mathbf{G}_m^T \tag{25}$$
$$\mathbf{R}_{x_m x_m} = \mathbf{G}_m\mathbf{R}_{ss}\mathbf{G}_m^T \tag{26}$$

where $\mathbf{R}_{ss} = E[\mathbf{s}(k)\mathbf{s}^T(k)]$ is the source correlation matrix. Substituting (25) and (26) into (24), we obtain

$$\mathbf{W}_{nm,\mathrm{o}} = \mathbf{G}_n\mathbf{R}_{ss}\mathbf{G}_m^T\left[\mathbf{G}_m\mathbf{R}_{ss}\mathbf{G}_m^T\right]^{-1}. \tag{27}$$

If the source signal $s(k)$ is white, then

$$\mathbf{R}_{ss} = \sigma_s^2\cdot\mathbf{I} \tag{28}$$

where σ_s^2 is the variance of the source signal. The optimal prediction matrix becomes

$$\mathbf{W}_{nm,\mathrm{o}} = \mathbf{G}_n\mathbf{G}_m^T\left[\mathbf{G}_m\mathbf{G}_m^T\right]^{-1} \tag{29}$$

which depends solely on the channel information. In this particular case, the $\mathbf{W}_{nm,\mathrm{o}}$ matrix can be viewed as the time-domain counterpart of the RTF, so the MMSE estimator given in (22) is equivalent to the TF-GSC approach [33]. However, in practical applications, speech signal is not white. Then, $\mathbf{W}_{nm,\mathrm{o}}$ depends not only on the channel impulse responses, but also on the source correlation matrix. This indicates that the developed MMSE estimator exploits both the spatial and temporal prediction information for noise reduction.

In real applications, the signals $\mathbf{x}_n(k)$ and $\mathbf{x}_m(k)$ are not observable, so the direct computation of $\mathbf{W}_{nm,\mathrm{o}}$ seems difficult. However, using the relation $\mathbf{x}_n(k) = \mathbf{y}_n(k) - \mathbf{v}_n(k)$ and the fact that the noise and speech are uncorrelated, we can verify that

$$\mathbf{R}_{x_n x_m} = \mathbf{R}_{y_n y_m} - \mathbf{R}_{v_n v_m} \tag{30}$$

and

$$\mathbf{R}_{x_m x_m} = \mathbf{R}_{y_m y_m} - \mathbf{R}_{v_m v_m} \tag{31}$$

where $\mathbf{R}_{y_n y_m}$ and $\mathbf{R}_{v_n v_m}$ are defined similarly to $\mathbf{R}_{x_n x_m}$, and $\mathbf{R}_{y_m y_m}$ and $\mathbf{R}_{v_m v_m}$ are defined similarly to $\mathbf{R}_{x_m x_m}$. As a result

$$\mathbf{W}_{nm,\mathrm{o}} = (\mathbf{R}_{y_n y_m} - \mathbf{R}_{v_n v_m})(\mathbf{R}_{y_m y_m} - \mathbf{R}_{v_m v_m})^{-1}. \tag{32}$$

Now the optimal filter matrix depends only on the second-order statistics of the noise and noisy signals. The statistics of the noisy signals can be directly computed from the observed signals. We assume that the noise is stationary or at least slowly-varying so that its characteristics stay the same from a silence period [i.e., when $x_n(k) = 0$] to the following period when speech is active. In this case, if we use a voice activity detector (VAD), the noise characteristics can be estimated during silence periods.

Using either (24) or (32), we can obtain an optimal estimate of the $\mathbf{W}_m$ matrix, i.e., $\mathbf{W}_{m,\mathrm{o}}$. Substituting $\mathbf{W}_{m,\mathrm{o}}$ into (22), the optimal transformation $\mathbf{h}_{m,\mathrm{o}}$ can be rewritten as

$$\mathbf{h}_{m,\mathrm{o}} = \mathbf{R}_{vv}^{-1}\mathbf{W}_{m,\mathrm{o}}^T\left[\mathbf{W}_{m,\mathrm{o}}\mathbf{R}_{vv}^{-1}\mathbf{W}_{m,\mathrm{o}}^T\right]^{-1}\mathbf{u}_1. \tag{33}$$

If $\mathbf{x}_n(k) = \mathbf{W}_{nm,\mathrm{o}}\mathbf{x}_m(k)$, applying $\mathbf{h}_{m,\mathrm{o}}$ to filter the observed signals can reduce noise without introducing any speech distortion. In practice, however, we in general do not have exactly $\mathbf{x}_n(k) = \mathbf{W}_{nm,\mathrm{o}}\mathbf{x}_m(k)$, so that some speech distortion is expected. However, for long filters, we can approach this equality so that the distortion can be kept very low.

C. Particular Case

To enable a better understanding of the optimal filter given in (22), let us study a special case where we have an equispaced linear array with N microphones, and the noise signals $v_n(k)(n = 1, 2, \ldots, N)$ are white Gaussian random processes with zero mean and variance of σ_v^2 and are uncorrelated with each other. Let us choose the first microphone as the reference and estimate the speech component at this microphone (i.e., $m = 1$). In this situation, we have $\mathbf{R}_{vv} = \sigma_v^2\mathbf{I}_{NL\times NL}$, and the optimal filter $\mathbf{h}_{1,\mathrm{o}}$ becomes

$$\begin{aligned}\mathbf{h}_{1,\mathrm{o}} &= \mathbf{W}_{1,\mathrm{o}}^T\left[\mathbf{W}_{1,\mathrm{o}}\mathbf{W}_{1,\mathrm{o}}^T\right]^{-1}\mathbf{u}_1\\ &= \begin{bmatrix}\mathbf{h}_{11,\mathrm{o}}^T & \mathbf{h}_{21,\mathrm{o}}^T & \cdots & \mathbf{h}_{N1,\mathrm{o}}^T\end{bmatrix}^T.\end{aligned} \tag{34}$$

Substituting (24) into (34), we find that

$$\begin{aligned}\mathbf{h}_{n1,\mathrm{o}} = \mathbf{R}_{x_n x_1}\left[\sum_{n=1}^{N}\mathbf{R}_{x_n x_1}^T\mathbf{R}_{x_n x_1}\right]^{-1}\mathbf{R}_{x_1 x_1}\mathbf{u}_1,\\ n = 1, 2, \ldots, N.\end{aligned} \tag{35}$$

Now let us assume that the application environment is free of reverberation and the sound source is located in the far field. In this case, if we neglect the propagation attenuation, the speech component received at the nth microphone can be written as

$$\begin{aligned}\mathbf{x}_n(k) &= \begin{bmatrix}x_n(k)\\ x_n(k-1)\\ \vdots\\ x_n(k-L+1)\end{bmatrix}\\ &= \begin{bmatrix}s[k-t+(n-1)\tau]\\ s[k-t+(n-1)\tau-1]\\ \vdots\\ s[k-t+(n-1)\tau-L+1]\end{bmatrix}\end{aligned} \tag{36}$$

where t is the propagation time (in samples) from the unknown source $s(k)$ to the reference microphone, and τ is the relative delay (in samples) between adjacent microphones. In this situation, the cross-correlation matrix $\mathbf{R}_{x_n x_1}$ can be expressed as

shown in (37) at the bottom of the page, where $\tau_{n1} = (n-1)\tau$, $r_{ss}(\tau_{n1}+i) = E\{s(k)s[k+(n-1)\tau+i]\}$ is the correlation coefficient of the source signal, and $-L+1 \leq i \leq L-1$. Now if we further assume that the source signal is a white Gaussian random process with zero mean and variance of σ_s^2, the correlation matrix $\mathbf{R}_{x_n x_1}$ can be simplified as

$$\mathbf{R}_{x_n x_1} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \sigma_s^2 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_s^2 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \cdots & \cdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_s^2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & \sigma_s^2 & 0 & \cdots & 0 \end{bmatrix} \Bigg\}\tau_{n1} . \tag{38}$$

Substituting (38) into (35), we readily derive that

$$\mathbf{h}_{n1,\mathrm{o}} = [\underbrace{0 \quad \cdots \quad 0}_{\tau_{n1}} \quad 1/N \quad 0 \quad \cdots \quad 0]^T \tag{39}$$

which is a unit impulse filter. So, in the ideal propagation situation and when both the source signal and noise are white Gaussian random processes, the solution is indeed a delay-and-sum beamformer. If the source signal is not a white random process, the optimal filter $\mathbf{h}_{n1,\mathrm{o}}$ is no longer a unit impulse filter; but the two filters $\mathbf{h}_{n1,\mathrm{o}}$ and $\mathbf{h}_{11,\mathrm{o}}$ satisfy $h_{n1,\tau_{n1}+l} = h_{11,l}(0 \leq l \leq L-\tau_{n1}-1)$. In other words, the filter $\mathbf{h}_{n1,\mathrm{o}}$ is a shifted (by τ_{n1}) version of the filter $\mathbf{h}_{11,\mathrm{o}}$. Therefore, if the application environment is free of reverberation, the optimal filter given in (22) can be viewed as a particular case of the filter-and-sum beamformer. With reverberation, however, the developed MMSE estimator differs significantly from conventional beamforming techniques, which will be furthered discussed in the following sections.

IV. Block-Based MMSE Estimator Using Multiple Microphones

In the previous section, we developed an MMSE estimator that estimates only one speech sample at a time. In many applications, it is also desirable to estimate a frame of speech from a given frame of noisy observations. Now we consider the signal model given in (2). An estimate of the speech vector $\mathbf{x}_m(k)$ can be obtained through the following linear transformation:

$$\mathbf{x}_m(k) = \sum_{n=1}^{N} \mathbf{H}_{nm}\mathbf{y}_n(k) \tag{40}$$

where $\mathbf{H}_{nm}$ is a matrix of size $L \times L$. The error signal vector obtained by this estimation is then written as

$$\begin{aligned} \mathbf{e}_m(k) &= \hat{\mathbf{x}}_m(k) - \mathbf{x}_m(k) \\ &= \sum_{n=1}^{N} \mathbf{H}_{nm}\mathbf{y}_n(k) - \mathbf{x}_m(k). \end{aligned} \tag{41}$$

Substituting (2) into (41) gives

$$\mathbf{e}_m(k) = \mathbf{e}_{x,m}(k) + \mathbf{e}_{v,m}(k) \tag{42}$$

where

$$\mathbf{e}_{x,m}(k) \triangleq \sum_{n=1}^{N} \mathbf{H}_{nm}\mathbf{x}_n(k) - \mathbf{x}_m(k) \tag{43}$$

represents the speech distortion due to the linear transformation and

$$\mathbf{e}_{v,m}(k) \triangleq \sum_{n=1}^{N} \mathbf{H}_{nm}\mathbf{v}_n(k) \tag{44}$$

is the residual noise. It is immediately clear that the objective of noise reduction is to find an optimal set of the matrices $\mathbf{H}_{nm}(n = 1,2,\ldots,N)$ such that the MSE of $\mathbf{e}_{v,m}(k)$ is minimized while keeping $\mathbf{e}_{x,m}(k)$ as close to $\mathbf{0}$ as possible.

Inspecting (44), we can write the MSE of the residual noise $\mathbf{e}_{v,m}(k)$ as

$$\begin{aligned} J_v(\mathbf{H}_m) &= \mathrm{tr}\left\{E\left[\mathbf{e}_{v,m}(k)\mathbf{e}_{v,m}^T(k)\right]\right\} \\ &= \mathrm{tr}\left(\mathbf{H}_m\mathbf{R}_{vv}\mathbf{H}_m^T\right) \end{aligned} \tag{45}$$

where

$$\mathbf{H}_m = [\mathbf{H}_{1m} \quad \mathbf{H}_{2m} \quad \cdots \quad \mathbf{H}_{Nm}].$$

Again, we assume that we can find N filter matrices, $\mathbf{W}_{nm}, n = 1,2,\ldots,N$ so that (16) is satisfied. Substituting (16) into (43), we obtain

$$\begin{aligned} \mathbf{e}_{x,m}(k) &= \sum_{n=1}^{N} \mathbf{H}_{nm}\mathbf{W}_{nm}\mathbf{x}_m(k) - \mathbf{x}_m(k) \\ &= \left[\mathbf{H}_m\mathbf{W}_m^T - \mathbf{I}\right]\mathbf{x}_m(k) \end{aligned} \tag{46}$$

where $\mathbf{W}_m$ is composed of $\mathbf{W}_{nm}, n = 1,2,\ldots,N$, as defined in (18).

$$\mathbf{R}_{x_n x_1} = \begin{bmatrix} r_{ss}(\tau_{n1}) & r_{ss}(\tau_{n1}+1) & \cdots & r_{ss}(\tau_{n1}+L-1) \\ r_{ss}(\tau_{n1}-1) & r_{ss}(\tau_{n1}) & \cdots & r_{ss}(\tau_{n1}+L-2) \\ \vdots & \vdots & \ddots & \vdots \\ r_{ss}(\tau_{n1}-L+1) & r_{ss}(\tau_{n1}-L+2) & \cdots & r_{ss}(\tau_{n1}) \end{bmatrix} \tag{37}$$

Now the noise-reduction problem can be formulated as one of estimating the optimal transformation $\mathbf{H}_m$ to minimize $J_v(\mathbf{H}_m)$ with the constraint that $\mathbf{H}_m\mathbf{W}_m^T = \mathbf{I}$. Mathematically, this estimation problem is written as

$$\mathbf{H}_{m,\mathrm{o}} = \arg\min_{\mathbf{H}_m} J_v(\mathbf{H}_m) \quad \text{subject to } \mathbf{H}_m\mathbf{W}_m^T = \mathbf{I}. \tag{47}$$

In order to adjoin the constraint to the cost function, we break the constraint on the right-hand side of (47) into the following form:

$$\mathbf{H}_m\mathbf{W}_m^T\mathbf{u}_l = \mathbf{u}_l, \quad l = 1, 2, \ldots, L \tag{48}$$

where

$$\mathbf{u}_l = [\underbrace{0 \quad \cdots \quad 0}_{l-1} \quad 1 \quad \underbrace{0 \quad \cdots \quad 0}_{L-l}]^T \tag{49}$$

is a unit vector. Now using the Lagrange method, we can rewrite the constrained optimization problem in (47) as

$$\mathbf{H}_{m,\mathrm{o}} = \arg\min_{\mathbf{H}_m} \mathcal{L}(\mathbf{H}_m, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_L) \tag{50}$$

where

$$\begin{aligned}\mathcal{L}(\mathbf{H}_m, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_L) &= J_v(\mathbf{H}_m) + \sum_{l=1}^{L} \boldsymbol{\lambda}_l^T \left[\mathbf{H}_m\mathbf{W}_m^T\mathbf{u}_l - \mathbf{u}_l\right] \\ &= \mathrm{tr}\left(\mathbf{H}_m\mathbf{R}_{vv}\mathbf{H}_m^T\right) + \sum_{l=1}^{L} \boldsymbol{\lambda}_l^T \left[\mathbf{H}_m\mathbf{W}_m^T\mathbf{u}_l - \mathbf{u}_l\right]\end{aligned} \tag{51}$$

and vectors $\boldsymbol{\lambda}_l (l = 1, 2, \ldots, L)$ are the Lagrange multipliers.

If the noise covariance matrix $\mathbf{R}_{vv}$ is full rank, we find from (50) that

$$\mathbf{H}_{m,\mathrm{o}} = \left[\mathbf{W}_m\mathbf{R}_{vv}^{-1}\mathbf{W}_m^T\right]^{-1}\mathbf{W}_m\mathbf{R}_{vv}^{-1}. \tag{52}$$

If $\mathbf{x}_n(k) = \mathbf{W}_{nm,\mathrm{o}}\mathbf{x}_m(k)$ holds, this transformation can reduce noise without introducing any speech distortion. In practice, the condition of $\mathbf{x}_n(k) = \mathbf{W}_{nm,\mathrm{o}}\mathbf{x}_m(k)$ may not hold exactly so there will be some speech distortion. However, in general the distortion can be kept to a very low level so that it cannot be perceived by the human ear.

V. Experiments

We have developed, respectively in Sections III and IV, two multichannel algorithms for noise reduction. In this section, we will assess their performance in real acoustic environments. It can be easily checked that the optimal filter $\mathbf{h}_{m,\mathrm{o}}$ given in (22) is the transpose of the first row of the optimal transformation matrix $\mathbf{H}_{m,\mathrm{o}}$ given in (52). So, the two multichannel MMSE estimators are closely related to each other and, in general, they have similar performance. To make our presentation concise, we will only present the results obtained from the first estimator, i.e., the sample-by-sample version.

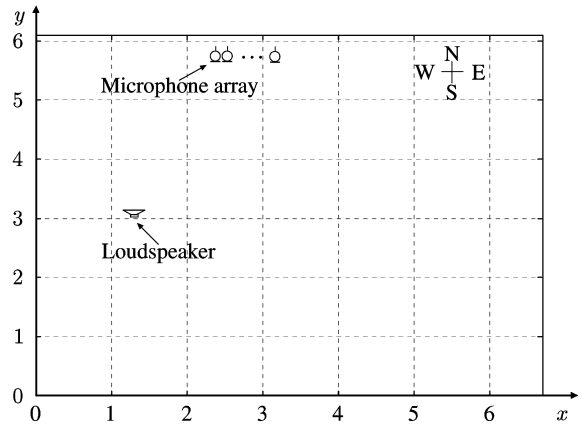


Fig. 2. Layout of the experimental setup in the varechoic chamber (coordinate values measured in meters). The sound source (a loudspeaker) is located at (1.337, 3.162, 1.600). The ten microphones of the linear array are located, respectively, at (x, 5.600, 1.400), where $x = 2.437 : 0.1 : 3.337$.

A. Experimental Setup

The experiments were conducted with the acoustic impulse responses measured in the varechoic chamber at Bell Labs. The chamber is a rectangular room, which measures 6.7 m long by 6.1 m wide by 2.9 m high ($x \times y \times z$) and is equipped with 368 electronically controlled panels. Each panel consists of two perforated sheets whose holes, if aligned, expose sound-absorbing material (fiberglass) behind, but if shifted to misalign, form a highly reflective surface. Each panel can be individually controlled so that the holes on a particular panel are either fully open (absorbing) or fully closed (reflective). As a result, a total of 2^{368} different room characteristics can be generated by varying the binary states of the 368 panels in different combinations. For a detailed description of the varechoic chamber and how the reverberation time is controlled, see [40] and [41].

The layout of the experimental setup is illustrated in Fig. 2, where a linear array of ten omni-directional microphones is mounted 1.4 m ($z = 1.400$) above the floor and parallel to the north wall at a distance of 0.5 m. The ten microphones are located, respectively, at (x, 5.600, 1.400), where $x = 2.437 : 0.1 : 3.337$. To simulate a sound source, we place a loudspeaker at (1.337, 3.162, 1.600), playing back a speech signal prerecorded from a female speaker. To make the experiments repeatable, we first measured the acoustic channel impulse responses from the source to the ten microphones (each impulse response is first measured at 48 kHz and then downsampled to 8 kHz). These measured impulse responses are then regarded as the true ones. During experiments, the microphone outputs are generated by convolving the source signal with the corresponding measured impulse responses, and noise is then added to the convolved results to control the SNR level.

In Section III, we showed that the developed multichannel noise-reduction algorithm degenerates to a delay-and-sum beamformer if both the source signal and noise are white random processes and the operating environment is free of reverberation. To verify this, we carried out an experiment. In

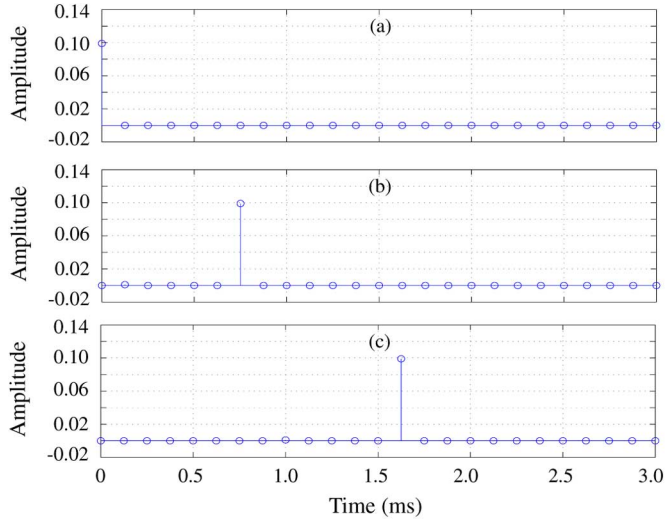


Fig. 3. Estimated filters $(\mathbf{h}_{n1})$ in an anechoic environment when both the source signal and noise are white Gaussian random processes and SNR $= 10$ dB. (a) $n = 1$. (b) $n = 5$. (c) $n = 10$.

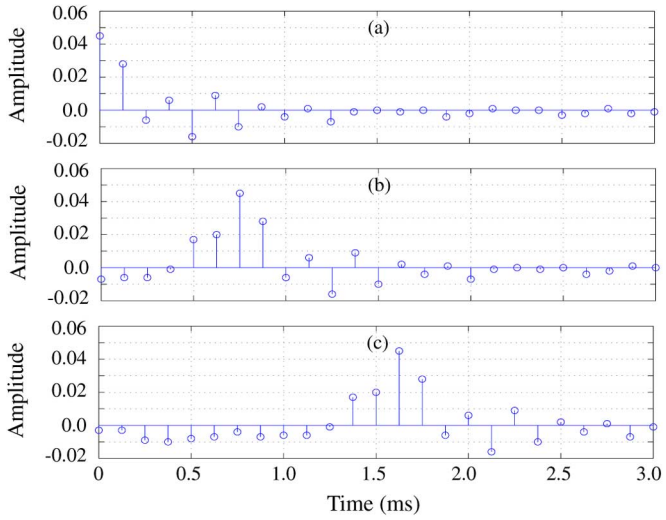


Fig. 4. Estimated filters $(\mathbf{h}_{n1})$ in an anechoic environment when the source is a speech signal, noise at each microphone is a Gaussian random process, and SNR $= 10$ dB. (a) $n = 1$. (b) $n = 5$. (c) $n = 10$.

order to simulate the anechoic situation, we take the impulse responses measured when 89% of the varechoic-chamber panels are open (the corresponding reverberation time $T_{60} = 240$ ms). We then keep only the direct path and set all the other taps into zero. It is seen from (22) that we need to specify the filter length L before estimating the optimal filter. For the anechoic situation, the determination of L is relatively easy, i.e., it only needs to be long enough to cover the maximal TDOA between the first and tenth microphone. In our setup, the maximal TDOA is approximately 3 ms, which corresponds to 24 sampling periods. So, we set the filter length L to 32, which is slightly larger than the maximal TDOA. The estimated filters for the first, the fifth, and the tenth microphones (i.e., $\mathbf{h}_{11}, \mathbf{h}_{51}$, and $\mathbf{h}_{101}$) are shown in Fig. 3. As clearly seen, each estimated filter has only one nonzero coefficient, whose location depends on the TDOA relative to the reference microphone. Therefore, the solution is indeed a delay-and-sum beamformer.

Also in anechoic environments, if the source signal is speech (or any signal that has some temporal correlation), the developed multichannel algorithm can take advantage of both the spatial redundancy among multiple microphones and the correlation among neighboring signal samples for better noise reduction. In the second experiment, we examine the optimal filter for speech sources. The experimental conditions are exactly the same as those of the previous experiment except that this time the source is a speech signal from a female speaker rather than a white Gaussian signal. The estimated optimal filters are plotted in Fig. 4. This time each filter is no longer a unit impulse response filter, and it is clearly seen that both $\mathbf{h}_{51}$ and $\mathbf{h}_{101}$ are a shifted version of $\mathbf{h}_{11}$. This confirms the analysis given in Section III. So if the propagation environment is free of reverberation and when the source signal is speech, the developed multichannel algorithm can be viewed as a particular case of the filter-and-sum (or Frost) beamformer. In more generic acoustic environments where there is reverberation and noise can be either white or colored, the developed multichannel algorithm is still a filter-and-sum structure; but differs significantly from the traditional filter-and-sum beamformer in many respects, as has been discussed in the previous sections.

We now begin to assess the noise-reduction performance of the multichannel algorithm. Without loss of generality, let us choose the first microphone as the reference microphone. Substituting the optimal filter into (4) and setting $m = 1$, we obtain the optimal speech estimate as

$$\hat{x}_1(k) = \sum_{n=1}^{N} \mathbf{h}_{nm,\mathrm{o}}^T \mathbf{y}_n(k) = \hat{x}_{1,\mathrm{nr}}(k) + \hat{v}_{1,\mathrm{nr}}(k)$$

where $\hat{x}_{1,\mathrm{nr}}(k) \triangleq \sum_{n=1}^{N} \mathbf{h}_{nm,\mathrm{o}}^T \mathbf{x}_n(k)$ and $\hat{v}_{1,\mathrm{nr}}(k) \triangleq \sum_{n=1}^{N} \mathbf{h}_{nm,\mathrm{o}}^T \mathbf{v}_n(k)$ are, respectively, the speech and residual noise filtered by the optimal filter. To assess the performance, we evaluate two criteria, namely the *a posteriori* SNR and the Itakura–Saito (IS) distance. The *a posteriori* SNR is defined as

$$\mathrm{SNR}_{\mathrm{o}} = \frac{E\left[\hat{x}_{1,\mathrm{nr}}^2(k)\right]}{E\left[\hat{v}_{1,\mathrm{nr}}^2(k)\right]}.$$

This measurement, when compared with the *a priori* SNR, tells us how much the noise is reduced. The IS distance is a speech-distortion measure. For a detailed description of the IS distance, we refer to [42] and [43]. Many studies have shown that the IS measure is highly correlated with subjective quality judgments and two speech signals would be perceptually nearly identical if the IS distance between them is less than 0.1. In this experiment, we compute the IS distance between $x_1(k)$ and $\hat{x}_{1,\mathrm{nr}}(k)$, which measures the degree of speech distortion due to the optimal filter.

As mentioned earlier, in order to estimate and use the optimal filter given in (22), we need to specify the filter length L. If there is no reverberation, it is relatively easy to determine L, i.e., it needs only to be long enough to cover the maximal TDOA between the reference and the other microphones. In the presence of reverberation, however, the determination of L would become more difficult and its value should, in theory, depend on the reverberation condition. Generally speaking, a longer filter has to

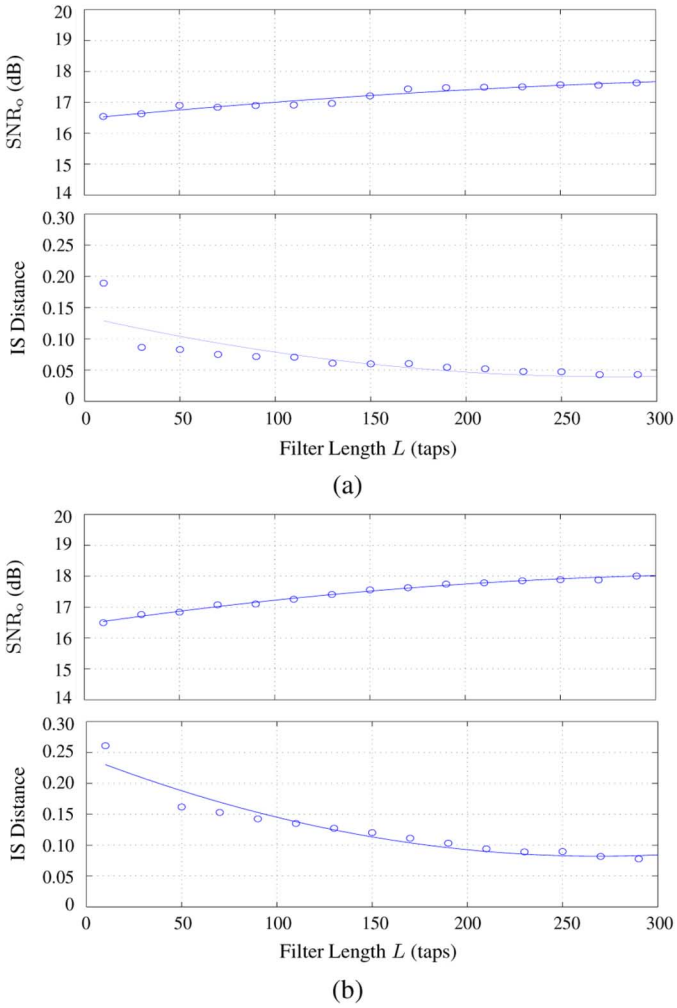


Fig. 5. The *a posteriori* SNR and IS distance, both as a function of the filter length L. (a) $T_{60} = 240$ ms and (b) $T_{60} = 580$ ms. The source is a speech signal from a female speaker; the background noise at each microphone is a computer-generated white Gaussian process, and SNR $= 10$ dB. The fitting curve is a second-order polynomial.

be used if the environment is more reverberant. The next experiment investigates the impact of the filter length on the algorithm performance. Here, to eliminate the effect due to noise estimation, we assume that the statistics of the noise signals are known *a priori*. We consider two cases. In the first case, 89% of the chamber panels are open. The corresponding reverberation time T_{60} is approximately 240 ms. The results are plotted in Fig. 5. One can see from Fig. 5(a) that the *a posteriori* SNR (in dB) increases with L. So, the longer the filter, the more the noise reduction. Contrary to SNR$_o$, the IS distance decreases with L. This is understandable, since as L increases, we will get a better prediction of $\mathbf{x}_n(k)$ from $\mathbf{x}_1(k)$. Consequently, as L increases, the algorithm achieves more noise reduction and causes less speech distortion. We also see from Fig. 5(a) that the *a posteriori* SNR (in dB) increases almost linearly with L. Unlike the SNR curve, the relationship between the IS distance and the filter length L is not linear. Instead, the curve first decreases quickly as the filter length increases, and then continues to decrease but at a slower rate. After $L = 250$, continuing to increase L does not seem to further decrease the IS distance. So, from a speech-distortion point of view, $L = 250$ is long enough for reasonably good performance.

Now we change the reverberation condition by opening 30% of the chamber panels and the corresponding reverberation time T_{60} is approximately 580 ms. The results are plotted in Fig. 5(b). Again, we see that the *a posteriori* SNR increases with L, whereas the IS distance decreases with L. Similar to the previous experiment, we see that after $L = 250$, further increasing L does not significantly reduce the IS distance. So we see again that $L = 250$ is long enough for reasonably good noise-reduction performance.

Comparing Figs. 5(a) and (b), one can see that with the same filter length L, the *a posteriori* SNRs in the two reverberation conditions are similar, which demonstrates the robustness of the proposed algorithm with respect to reverberation. However, the IS distance for $T_{60} = 580$ ms is much higher than that for $T_{60} = 240$ ms. This is, of course, understandable. As the environment becomes more reverberant, the prediction of $\mathbf{x}_n(k)$ from $\mathbf{x}_1(k)$ would become more difficult. However, for $L \geq 250$, we see that the IS distance in both conditions is less than 0.1, which is rather small, as this level of speech distortion is perceptually almost negligible.

Another important factor that would affect the algorithm performance is the number of microphones. The next experiment investigates the impact of the number of microphones on the noise-reduction performance. From the previous results, we see that good performance was achieved when the filter length L is 250 or longer. Note that when we increase the filter length, the computational complexity of the algorithm also grows. In addition, we also need more data to achieve a robust estimate of the covariance matrices. Therefore, the selection of filter length is basically a compromise between the noise-reduction performance and the complexity and robustness of the algorithm. In this experiment, we set $L = 250$. In addition, we assume again that the statistics of the noise signals are known *a priori*. The results are presented in Fig. 6.

If there is no reverberation, we can see, from Fig. 6(a) that the *a posteriori* SNR increases (in dB) linearly with the number of microphones. So the more the microphones, the higher the SNR. In the anechoic propagation situation, the signal observed at one microphone can, in principle, be perfectly predicted from the signal received at another microphone. So, there should be no speech distortion and the IS distance should be zero. However, we see that there is some minor speech distortion, and the IS distance grows with the number of microphones. This is because we use a square matrix in (16) to predict a frame of signal $x_n(k)$, i.e., $\mathbf{x}_n(k)$ from a frame of signal $\mathbf{x}_1(k)$. In our setup, most of the samples in $\mathbf{x}_n(k)$ can be perfectly predicted from $\mathbf{x}_1(k)$. However, there are a small number of samples at the end of the vector $\mathbf{x}_n(k)$ that cannot be predicted (the number depends the TDOA). It is this small unpredictable part that causes some speech distortion. Since we use a linear array, the TDOA between $x_n(k)$ and $x_1(k)$ increases with n. Therefore, the IS distance increases with N.

In a reverberant environment, we see from Fig. 6(b) that the *a posteriori* SNR also increases with the number of microphones. Similar to the previous experiment, the IS distance grows slightly as more microphones are used. The reason

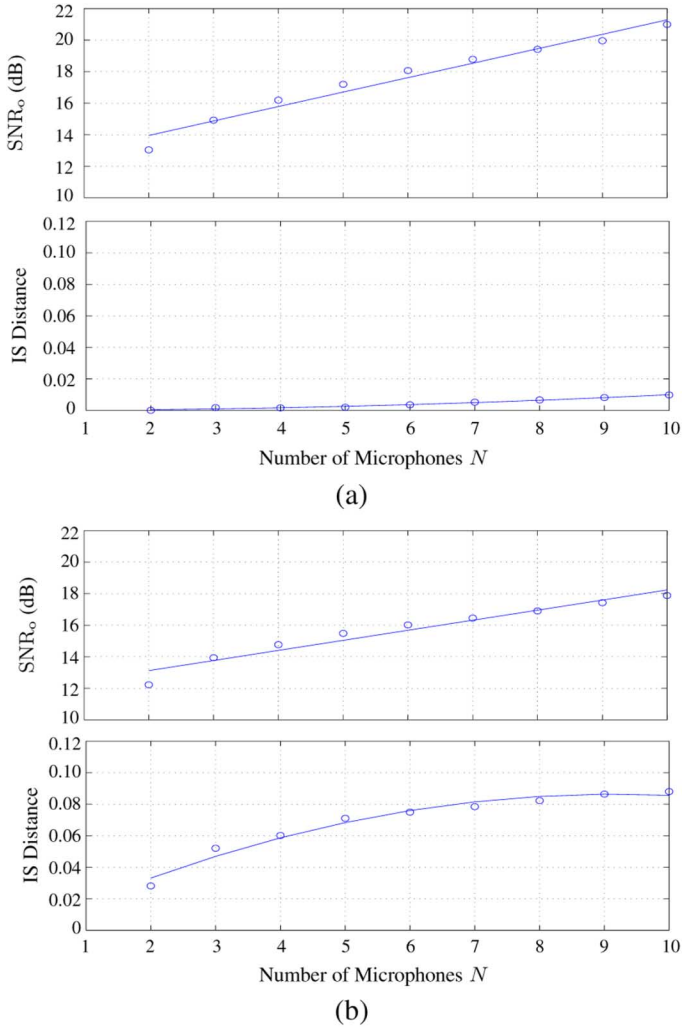


Fig. 6. The *a posteriori* SNR and IS distance, both as a function of the number of microphones N. (a) In a condition where there is no reverberation and (b) in a reverberation condition with $T_{60} = 380$ ms. The source is a speech signal from a female speaker, and the background noise at each microphone is a computer-generated white Gaussian process with SNR $= 10$ dB. The fitting curve is a second-order polynomial.

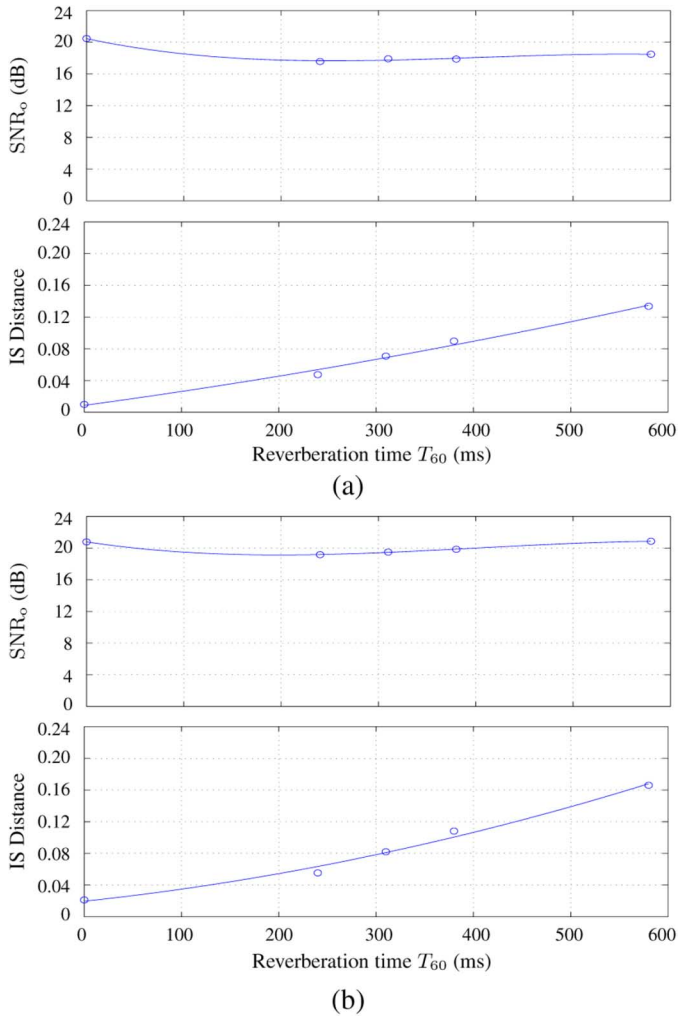


Fig. 7. Noise-reduction performance versus T_{60} (a) in white Gaussian noise and (b) in NYSE noise. $L = 250$ and SNR $= 10$ dB. The fitting curve is a second-order polynomial.

is also attributed to the imperfect prediction of $\mathbf{x}_n(k)$ from $\mathbf{x}_1(k)$. However, we see from that beyond seven microphones, the increase of IS distance with the number of microphone is negligible. In addition, the overall IS distance is very small (less than 0.1), so the resulting speech distortion is perceptually almost negligible.

The next experiment tests the robustness of the multichannel algorithm to reverberation. The parameters used are: $L = 250$, $N = 10$, and SNR $= 10$ dB. Compared with the previous experiments, this one does not assume to know the noise statistics. Instead, we developed a short-term energy-based VAD to distinguish speech-plus-noise from noise-only segments. The noise covariance matrix is then computed from the noise-only segments using a batch method and the optimal filter is subsequently estimated according to (33). We tested the algorithm in two noise conditions: computer generated white Gaussian noise and a noise signal recorded in a New York Stock Exchange (NYSE) room. (This is a nonstationary bubbling noise, which consists of sound from various sources such as speakers, telephone rings, electric fans, etc. It is recorded using a single microphone. However, for the outputs of a microphone array, we cut the whole recording into N segments, with each segment being added to one microphone.) The results are depicted in Fig. 7. We see that the *a posteriori* SNR in both situations does not vary much when the reverberation time is changed. This indeed demonstrates that the developed multichannel algorithm is very immune to reverberation. In contrast to SNR, we see that the IS distance grows with reverberation time. This result should not come as a surprise, since as the reverberation time T_{60} increases, it becomes more difficult to predict the speech observed at one microphone from that received at another microphone. As a result, a higher level of speech distortion is unavoidable.

In the final experiment, we compare the new multichannel noise-reduction approach with two widely used beamforming algorithms: a delay-and-sum and an LCMV. Here, we choose $L = 250$ and $N = 10$. The noise at each microphone is white Gaussian. To use the delay-and-sum beamformer, we need to know the TDOA information. In our experiment, the real room impulse responses have been measured so we computed the

TABLE I
PERFORMANCE OF NOISE REDUCTION AND SPEECH DISTORTION

			New Algorithm		Delay-And-Sum			LCMV		
SNR (dB)	Reverberation Condition	L_g (taps)	$\mathrm{SNR_o}$ (dB)	ISD	$\mathrm{SNR_o}$ (dB)	ISD	ISD*	$\mathrm{SNR_o}$ (dB)	ISD	ISD*
10	No reverberation	2048	20.46	0.009	20.00	0.000	0.000	19.99	0.000	0.000
	$T_{60} = 240$ ms	2048	17.57	0.047	16.87	0.225	0.699	−59.44	0.6232	0.000
	$T_{60} = 580$ ms	2048	17.88	0.089	14.89	0.292	0.999	−65.59	0.9854	0.000

Conditions: The source is a speech signal prerecorded from a female speaker; the background noise at each microphone is a computer-generated white Gaussian random process; L_g is the length of the room impulse responses; in the delay-and-sum method, the TDOA information is known *a priori*; the LCMV assumes that the room impulse responses are known *a priori* and the LCMV filter is constructed according to [22] [eq. (16)] with the filter length being given in [22] [eq. (25)]. ISD denotes the IS distance between $x_1(k)$ and $\hat{x}_{1,\mathrm{nr}}(k)$. ISD* denotes the IS distance between $s(k)$ and $\hat{x}_{1,\mathrm{nr}}(k)$. Note that the new algorithm does not provide an estimate of the source signal, so there is no ISD* for the new algorithm.

TDOAs by examining the direct paths of the room impulse responses. This is equivalent to saying that the TDOA information is known *a priori*. The LCMV algorithm in a reverberant room environment is given in [22]. Here, we assume that the room impulse responses are known *a priori* and we construct the LCMV filter according to [22] [(16)]. Note that in our experimental setup, there is only one source in the sound field, and the number of microphones is equal to ten. In this case, it is easily checked that the LCMV solution is the same as the multiple input/output inverse theorem (MINT). The connection between LCMV and MINT is explained in [22].

It should be pointed out here that it is not easy to fairly compare the above algorithms, as they aim at estimating different signal components. Specifically, the new algorithm is formulated to estimate the speech component received at one of the multiple microphones, while the beamforming techniques focus on estimating the source signal. The only condition for a fair comparison is when the environment is free of reverberation. Such a condition, however, is not very realistic. It is often more interesting to see a comparison in reverberation conditions. In order to make comparison results more meaningful, we evaluate three performance criteria for the beamforming techniques: the *a posterori* SNR, the IS distance (ISD) between the speech component observed at the reference microphone and that in the beamforming output, and the IS distance (ISD*) between the source signal and the speech component in the beamforming output.

The results of this experiment are shown in Table I. When there is no reverberation, one can see that all the algorithms yield similar performance. This should not come as a surprise. As a matter of fact, in an anechoic environment and if the background noise is white Gaussian, all of the algorithms will degenerate to the delay-and-sum structure, one way or another. Notice that when the environment is free of reverberation, the ISD and ISD* are the same since in this case the speech component at the reference microphone is just a delayed and attenuated version of the source signal.

In reverberant environments, we see that the delay-and-sum beamformer can still improve the SNR, where the degree of improvement depends on the reverberation condition. However, this method introduces significant speech distortion. The ISD* for the LCMV method is approximately zero, which indicates that the LCMV method has achieved perfect speech dereverberation. However, the SNR with this approach has been significantly degraded. The reason behind this can be explained as follows. When there is only one source in the sound field and if the room impulse responses are known *a priori*, the LCMV is the same as the MINT method, which basically achieves speech estimation by inverting the channel matrix. This inverse process may boost the background noise and hence causes SNR degradation. In comparison, the new multichannel algorithm achieves the highest SNR gain. Additionally, the resulting ISD shows that the speech distortion with this method is almost negligible.

VI. CONCLUSION

In this paper, we have focused on the noise-reduction problem using multiple microphones. We have formulated the problem as one of estimating the speech component received at one of the multiple microphones. We have developed two MMSE estimators, namely a sample-by-sample-based estimator and a block-based estimator. These two estimators are closely related to each other. Specifically, the optimal filter from the sample-by-sample-based method is the transpose of the first row of the optimal matrix in the block-based technique. Various experiments were carried out, and the results demonstrated that the developed techniques can achieve significant noise reduction while the resulting speech distortion is perceptually almost negligible. Compared with the traditional beamforming techniques, the developed algorithms have many appealing properties including but not limited to: they do not require the DOA information or any knowledge of either the reverberation condition or the channel impulse responses; the multiple microphones do not have to be arranged into a specific array geometry; they work the same for both the far-field and near-field cases; and they can produce very good and robust noise reduction with minimum speech distortion in practical environments.

REFERENCES

[1] M. R. Schroeder, "Apparatus for suppressing noise and distortion in communication signals," U.S. Patent No. 3 180 936, 1965, filed Dec. 1, 1960, issued Apr. 27.

[2] M. R. Schroeder, "Processing of communication signals to reduce effects of noise," U.S. Patent No. 3 403 224, 1968, filed May 28, 1965, issued Sep. 24.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, pp. 113–120, Apr. 1979.

[4] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[5] J. S. Lim, *Speech Enhancement*. Englewood Cliffs, NJ: Prentice-Hall, 1983.

[6] P. Vary, "Noise suppression by spectral magnitude estimation—mechanism and theoretical limits," *Signal Process.*, vol. 8, pp. 387–400, Jul. 1985.

[7] W. Etter and G. S. Moschytz, "Noise reduction by noise-adaptive spectral magnitude expansion," *J. Audio Eng. Soc.*, vol. 42, pp. 341–349, May 1994.

[8] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[10] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.

[11] P. J. Wolfe and S. J. Godsill, "Simple alternatives to the ephraim and malah suppression rule for speech enhancement," in *Proc. IEEE ICASSP*, 2001, pp. 496–499.

[12] K. K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. IEEE ICASSP*, 1987, pp. 177–180.

[13] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 373–385, Jul. 1998.

[14] Y. Ephraim, D. Malah, and B.-H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 12, pp. 1846–1856, Dec. 1989.

[15] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 19, pp. 1526–1555, Oct. 1992.

[16] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

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

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

[19] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972, III.

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

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

[22] J. Benesty, J. Chen, Y. Huang, and J. Dmochowski, "On microphone-array beamforming from a MIMO acoustic signal processing perspective," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1053–1065, Mar. 2007.

[23] M. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.

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

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

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

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

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

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

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

[31] S. Werner, J. A. Apolinário, and M. L. R. de Campos, "On the equivalence of RLS implementations of LCMV and GSC processors," *IEEE Signal Process. Lett.*, vol. 10, pp. 356–359, Dec. 2003.

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

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

[34] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep.. 2002.

[35] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 6, pp. 561–571, Nov. 2004.

[36] S. Gannot, D. Burshtein, and E. Weinstein, "Analysis of the power spectral deviation of the general transfer function GSC," *IEEE Trans. Signal Process.*, vol. 52, no. 4, pp. 1115–1121, Apr. 2004.

[37] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Trans. Signal Process.*, vol. 44, no. 8, pp. 2055–2063, Aug. 1996.

[38] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, Sep. 2004.

[39] I. Cohen, "Identification of speech source coupling between sensors in reverberant noisy environments," *IEEE Signal Process. Lett.*, vol. 11, no. 7, pp. 613–616, Jul. 2004.

[40] A. Härmä, "Acoustic measurement data from the varechoic chamber" Technical Memorandum, Agere Systems, Nov. 2001.

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

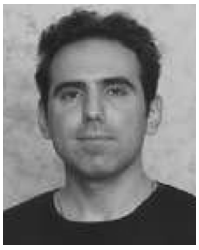
[42] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," *Electron. Commun. Japan*, vol. 53A, pp. 36–43, 1970.

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

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

From 1998 to 1999, he was with ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan, where he conducted research on speech synthesis, speech analysis, as well as objective measurements for evaluating speech synthesis. He then joined the Griffith University, Brisbane, Australia, as a Research Fellow, where he engaged in research in robust speech recognition, signal processing, and discriminative feature representation. From 2000 to 2001, he was with ATR Spoken Language Translation Research Laboratories, Kyoto, where he conducted research in robust speech recognition and speech enhancement. He joined Bell Laboratories, Murray Hill, NJ, as a Member of Technical Staff in July 2001. His current research interests include adaptive signal processing, speech enhancement, adaptive noise/echo cancellation, microphone array signal processing, signal separation, and source localization. He coauthored the book *Acoustic MIMO Signal Processing* (Springer-Verlag, 2006). He is a coeditor/coauthor of the book *Speech Enhancement* (Springer-Verlag, 2005) and a section editor of the reference *Springer Handbook of Speech Processing* (Springer-Verlag, 2007).

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



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

From January 1994 to July 1995, he worked at Telecom Paris University on multichannel adaptive filters and acoustic echo cancellation. From October 1995 to May 2003, he was first a Consultant and then a Member of the Technical Staff at Bell Laboratories, Murray Hill, NJ. In May 2003, he joined the University of Quebec, INRS-EMT, Montreal, QC, Canada, as an Associate Professor. His research interests are in signal processing, acoustic signal processing, and multimedia communications. He was a member of the editorial board of the *EURASIP Journal on Applied Signal Processing* and was the Co-Chair of the 1999 International Workshop on Acoustic Echo and Noise Control. He coauthored the books *Acoustic MIMO Signal Processing* (Springer-Verlag, 2006) and *Advances in Network and Acoustic Echo Cancellation* (Springer-Verlag, 2001). He is the Editor-in-Chief of the reference *Springer Handbook of Speech Processing* (Springer-Verlag, 2007). He is also a coeditor/coauthor of the books *Speech Enhancement* (Spinger-Verlag, 2005), *Audio Signal Processing for Next Generation Multimedia communication Systems* (Kluwer, 2004), *Adaptive Signal Processing: Applications to Real-World Problems* (Springer-Verlag, 2003), and *Acoustic Signal Processing for Telecommunication* (Kluwer, 2000).

Dr. Benesty received the 2001 Best Paper Award from the IEEE Signal Processing Society.

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

From March 2001 to January 2008, Dr. Huang was a Member of Technical Staff at Bell Laboratories, Murray Hill, NJ. In January 2008, he joined WeVoice, Inc., Bridgewater, NJ, and served as its CTO. His current research interests are in acoustic signal processing and multimedia communications. He is currently an Associated Editor of the *EURASIP Journal on Applied Signal Processing*. He served as a technical Co-Chair of the 2005 Joint Workshop on Hands-Free Speech Communication and Microphone Array and the 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. He is a coeditor/coauthor of the books *Springer Handbook of Speech Processing* (Springer-Verlag, 2007), *Acoustic MIMO Signal Processing* (Springer-Verlag, 2006), *Audio Signal Processing for Next-Generation Multimedia Communication Systems* (Kluwer, 2004) and *Adaptive Signal Processing: Applications to Real-World Problems* (Springer-Verlag, 2003).

Dr. Huang received the 2002 Young Author Best Paper Award from the IEEE Signal Processing Society, the 2000–2001 Outstanding Graduate Teaching Assistant Award from the School Electrical and Computer Engineering, Georgia Tech, the 2000 Outstanding Research Award from the Center of Signal and Image Processing, Georgia Tech, and the 1997–1998 Colonel Oscar P. Cleaver Outstanding Graduate Student Award from the School of Electrical and Computer Engineering, Georgia Tech. He served as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS from 2002 to 2005.