EURASIP Journal on Applied Signal Processing 2005:18, 2930–2937

An Improved Array Steering Vector Estimation Method and Its Application in Speech Enhancement

Zhu Liang Yu

Center for Signal Processing, Nanyang Technological University, Singapore 639798
Email: ezlyu@ntu.edu.sg

Meng Hwa Er

School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798
Email: emher@ntu.edu.sg

Received 30 April 2004; Revised 22 February 2005

We propose a robust microphone array for speech enhancement and noise suppression. To overcome target signal cancellation problem of conventional beamformes caused by array imperfections or reverberation effects, the proposed method adopts arbitrary transfer function relating each microphone and target speech signal as array channel model. This is achieved in two ways. First, we propose a method to estimate the array steering vector (ASV) by means of exploiting the nonstationarity of speech signal to combat stationary noise and interference. Next, with the estimated ASV, a robust matched-filter-(MF-) array-based generalized sidelobe canceller (MF-GSC) is constructed to enhance the speech signal and suppress noise/interference. In addition, it also has the capability to reduce the reverberation effects of the acoustic enclosure. Numerical results show that the proposed method demonstrates high performance even in adverse environments.

Keywords and phrases: adaptive microphone array, matched-filter array, generalized sidelobe canceller, noise suppression, speech enhancement, nonstationarity.

1. INTRODUCTION

Speech enhancement and noise suppression have received increasing interest in speech-related applications in adverse environments. Conventional single-channel speech enhancement methods, such as spectral subtraction [1], do not provide sufficient improvement to the speech signal, especially when noise is strong. In the past few decades, microphone array has been proposed as a promising technique for speech enhancement. It uses the signals captured by multimicrophones, which are distributed at different positions, to exploit the spatial-temporal information of the target signal, interferences, and noise for the purpose of improving the signal-to-noise ratio (SNR) by suppressing background noise and interferences.

Beamforming is the key technique in microphone array for speech enhancement and noise suppression. Many beamforming methods [2, 3, 4, 5, 6] have been proposed in literature. Among them, the most famous algorithm for wideband beamforming is the constrained minimum power adaptive beamformer proposed by Frost [2], also called Frost beamformer. It is capable of satisfying certain desired frequency response in the looking direction while minimizing the output noise power through constrained minimization of the total output power. Griffiths and Jim [3] reconstructed the Frost beamformer into the generalized sidelobe canceller (GSC). It transforms the constrained optimization problem in Frost beamformer into an unconstrained one and, consequently, improves the convergence performance. To improve the robustness of Frost beamformer or GSC, numerous methods [5, 6, 7, 8, 9, 10, 11] have been proposed to combat array imperfections, such as steering error, sensor location error, array channel mismatch, and so forth.

The methods mentioned above assume that the target signal propagates through known direct path and the geometry of the array is also known. However, in applications such as speech acquisition in adverse acoustic environments, the source signal propagates not merely along direct path. There are also unknown multipath and reverberation effects. Moreover, in some applications, the array geometry is unknown or changing, such as the microphone array mounted on human body. In such cases, the target signal is often cancelled to some extent in conventional adaptive beamforming approaches. This problem is especially serious for microphone array in strong reverberant environments. The performance significantly degrades due to the reverberation effects.

Concerning the existence of array imperfections and reverberation effects, a new channel model was adopted in adaptive array processing. The impulse response (IR) of the channel relating target source and each array element is modeled as an arbitrary linear filter, which conveys all the effects of array imperfections as well as reverberation. Since these IRs are unknown, the identification of IRs is necessary to avoid target signal cancellation. A straightforward solution is to use a training signal [12, 13]. However, it has limited applications because the reestimation of these IRs is inconvenient when the environment changes or signal source moves. Another potential solution is the blind channel identification (BCI) technique ([14, 15] and references therein), which is not so successful as in wireless communications because the length of IR is large in acoustic applications.

In some cases, for example, speech enhancement [16], the phase response of the target signal is not important. Moreover, the human auditory system is capable of tolerating distortion to some extent in speech signal. With these relaxations, in frequency domain, the IR identification problem can be simplified as array steering vector (ASV) estimation. The method in [12] was proposed to estimate ASV by computing the principal component extracted from the covariance matrix of the received array snapshots, in the case where the power of the target signal is far greater than the power of noise and interferences [17]. However, these methods are sensitive to noise and interferences, so that the estimation requires high SNR. In [18], the nonstationarity of the target signal is exploited to achieve accurate estimate of system transfer function against stationary noise and interferences. This idea is used to estimate frequency response ratio (RR) for wideband beamforming [19]. Unfortunately, the RR method has a drawback that its estimation error increases when the reference channel has low response or null at specific frequency bins.

In this paper, we propose an improved ASV estimation method by exploiting multichannel signals and nonstationarity of speech based on the idea in [18, 19]. Compared with [19], this paper differs in three aspects. Firstly, it is proved that the nonstationarity of the reference signal weakens if signal is corrupted by stationary noise. Since the error variance of estimated ASV depends on the nonstationarity of the reference signal [18], the error variance increases if the SNR of the reference signal decreases. Therefore, high SNR of reference signal is appreciated. Secondly, in this paper, a new reference signal, which exploits multichannel signals, is used. The multichannel signals are linearly weighted and summed up to produce an output signal, which is used as a reference signal in estimate of TF [18, 19]. A method to estimate optimal weight is also proposed. The SNR of the new reference signal is improved. Consequently, the accuracy of the estimated ASV improves. Thirdly, a normalized ASV vector is used to construct an extended GSC for speech enhancement and noise suppression. Such extended GSC can greatly improve the robustness of the beamformer as well as the performance of signal enhancement. Moreover, it also reduces the reverberation effects in the output signal.

This paper is organized as follows. The system model is reviewed in Section 2. In Section 3, an improved method for ASV estimation exploiting the nonstationarity of speech signal is derived. An extended GSC is then proposed in Section 4. It takes the advantage of the estimated ASV to combat reverberation effects as well as array imperfections. Some numerical results are shown in Section 5 to evaluate the performance of the proposed method. In Section 6, a brief conclusion is given.

2. SYSTEM MODEL

Notations used in this paper are defined before we formulate the problem and develop the algorithm. For example, a, $\mathbf{a}$, and $\mathbf{A}$ denote scalar, vector, and matrix, respectively. The operators $E\{\cdot\}$, $(\cdot)^*$, $(\cdot)^T$, $(\cdot)^H$, $\star$, and $\|\cdot\|$ stand for mathematical expectation, complex conjugate, transpose, Hermitian transpose, linear convolution, and Euclidean norm, respectively.

The microphone array system with M sensors is studied in this paper. The target speech signal $s(k)$ propagates through the ith channel with an impulse response (IR) $h_i(k)$, $i = 1, 2, \ldots, M$, and is corrupted by additive noise $n_i(k)$. The noise $n_i(k)$ may include environment noise, sensor noise and interferences if there are any. The sensor received signal $x_i(k)$ of the ith channel is then expressed as

$$x_i(k) = h_i(k) \star s(k) + n_i(k), \quad i = 1, 2, \ldots, M. \tag{1}$$

Splitting the received signal $x_i(k)$ in (1) into frames with suitable length N, and taking fast Fourier transform (FFT) on the signal of mth frame, it yields

$$\mathbf{x}(m, \omega) = s(m, \omega)\mathbf{h}(\omega) + \mathbf{n}(m, \omega), \tag{2}$$

where

$$\begin{aligned}\mathbf{x}(m, \omega) &= \left[x_1(m, \omega)x_2(m, \omega) \cdots x_M(m, \omega)\right]^T,\\ \mathbf{n}(m, \omega) &= \left[n_1(m, \omega)n_2(m, \omega) \cdots n_M(m, \omega)\right]^T,\\ \mathbf{h}(\omega) &= \left[h_1(\omega)h_2(\omega) \cdots h_M(\omega)\right]^T,\end{aligned} \tag{3}$$

where $\omega = 0, 1, \ldots, N/2$ denotes the frequency index. The transformed signals $x_i(m, \omega)$, $n_i(m, \omega)$, $s(m, \omega)$ and transfer function (TF) $h_i(\omega)$ are the Fourier transform of $x_i(k)$, $n_i(k)$, $s(k)$, and $h_i(k)$, respectively. In this paper, we call the transfer function vector $\mathbf{h}(\omega)$ an extended array steering vector (ASV).

Unlike the pure delay channel model used in conventional array processing, the IRs $\{h_i(n)\}$ in (1) are arbitrary linear filters which convey the effects of reverberation as well as the array imperfections. Considering this extended model in frequency domain, the robust beamforming can be achieved by using the estimated ASV $\mathbf{h}(\omega)$ instead of the nominal one, which is constructed based on pure delay channel model and perfect array channel assumption, in adaptive array processing.

3. BLIND ARRAY STEERING VECTOR IDENTIFICATION

The following assumptions are made in this paper.

(AS1) $\mathbf{h}(\omega)$ is fixed or slowly changing compared with the variation of the target signal in time domain.

(AS2) The target speech signal is uncorrelated with the interference and background noise.

(AS3) The target speech signal is quasi-nonstationary. The interference and background noise are stationary.

To estimate $\mathbf{h}(\omega)$, an intermediate signal $u(m,\omega)$ is firstly formed by combining the array received signals as

$$u(m,\omega) = \mathbf{d}^H(\omega)\mathbf{x}(m,\omega) = b(\omega)s(m,\omega) + \bar{n}(m,\omega), \quad (4)$$

where

$$\begin{aligned} \mathbf{d}(\omega) &= \left[d_1(\omega)\cdots d_M(\omega)\right]^T, \\ b(\omega) &= \mathbf{d}^H(\omega)\mathbf{h}(\omega), \\ \bar{n}(m,\omega) &= \mathbf{d}^H(\omega)\mathbf{n}(m,\omega), \end{aligned} \quad (5)$$

where $\mathbf{d}(\omega)$ is the weight vector to form the intermediate signal $u(m,\omega)$; $b(\omega)$ is the overall response relating the source signal and intermediate signal. With suitable weight $\mathbf{d}(\omega)$, the response $b(\omega)$ remains nonzero unless all the channel responses $\{h_i(\omega)\}$ at this frequency bin are zero. Moreover, the SNR of $u(m,\omega)$ is improved if suitable weight $\mathbf{d}(\omega)$ is selected.

Substituting (4) into (2), we obtain

$$\mathbf{x}(m,\omega) = \bar{\mathbf{h}}(\omega)u(m,\omega) + \tilde{\mathbf{n}}(m,\omega), \quad (6)$$

where

$$\bar{\mathbf{h}}(\omega) = \frac{\mathbf{h}(\omega)}{b(\omega)}, \qquad \tilde{\mathbf{n}}(m,\omega) = \mathbf{n}(m,\omega) - \frac{\mathbf{h}(\omega)}{b(\omega)}\bar{n}(m,\omega). \quad (7)$$

From (6), it can be found that the intermediate signal $u(m,\omega)$ is related to each array received signal with the ASV $\mathbf{h}(\omega)$ up to a multiplicative scale $1/b(\omega)$. In this paper, the normalized ASV vector is later used in the beamformer. The multiplicative scale $1/b(\omega)$ is eliminated by the normalization procedure if $b(\omega)$ is not zero. Therefore, in this paper, the vector $\bar{\mathbf{h}}(\omega)$ is estimated instead of $\mathbf{h}(\omega)$.

Taking the cross power spectrum density (PSD) between $\mathbf{x}(m,\omega)$ and $u(m,\omega)$, we have

$$R_{\mathbf{x}u}(m,\omega) = \bar{\mathbf{h}}(\omega)\sigma_u^2(m,\omega) + R_{\tilde{\mathbf{n}}u}(m,\omega), \quad (8)$$

where

$$\begin{aligned} R_{\mathbf{x}u}(m,\omega) &= E\{\mathbf{x}(m,\omega)u^*(m,\omega)\}, \\ \sigma_u^2(m,\omega) &= E\{u(m,\omega)u^*(m,\omega)\}, \\ R_{\tilde{\mathbf{n}}u}(m,\omega) &= E\{\tilde{\mathbf{n}}(m,\omega)u^*(m,\omega)\}. \end{aligned} \quad (9)$$

Based on assumption AS2, the cross-PSD between the target signal and interference/noise is zero. Hence, $R_{\tilde{\mathbf{n}}u}(m,\omega)$ only contains the components of the PSD or cross-PSD between the interferences and background noise. Moreover, according to assumption AS3, $R_{\tilde{\mathbf{n}}u}(m,\omega)$ is almost independent of frame index m due to the stationarity of interferences and noise. In other words, we can assume that $R_{\tilde{\mathbf{n}}u}(m,\omega)$ is time invariant, that is, $R_{\tilde{\mathbf{n}}u}(m,\omega) = R_{\tilde{\mathbf{n}}u}(\omega)$. Since the target signal is nonstationary, the PSD $\sigma_u^2(m,\omega)$ and cross-PSD $R_{\mathbf{x}u}(m,\omega)$ are time variant, and their estimates $\hat{\sigma}_u^2(m,\omega)$ and $\hat{R}_{\mathbf{x}u}(m,\omega)$ vary frame by frame. Substituting the estimates $\hat{\sigma}_u^2(m,\omega)$ and $\hat{R}_{\mathbf{x}u}(m,\omega)$ in (8), we have

$$\begin{aligned} &\hat{R}_{\mathbf{x}u}(m,\omega) \\ &\quad = \bar{\mathbf{h}}(\omega)\hat{\sigma}_u^2(m,\omega) + R_{\tilde{\mathbf{n}}u}(\omega) + \boldsymbol{\varepsilon}(m,\omega), \quad m = 1,\ldots,L, \end{aligned} \quad (10)$$

where $\boldsymbol{\varepsilon}(m,\omega)$ is a zero-mean estimation error vector. Concentrating the equations in (10), we obtain

$$\begin{aligned} \mathbf{b} &\triangleq \begin{bmatrix} \hat{R}_{\mathbf{x}u}(1,\omega) \\ \vdots \\ \hat{R}_{\mathbf{x}u}(L,\omega) \end{bmatrix} \\ &= \begin{bmatrix} \begin{bmatrix} \hat{\sigma}_u^2(1,\omega)\mathbf{I} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \hat{\sigma}_u^2(L,\omega)\mathbf{I} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{h}}(\omega) \\ R_{\tilde{\mathbf{n}}u}(\omega) \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}(1,\omega) \\ \vdots \\ \boldsymbol{\varepsilon}(L,\omega) \end{bmatrix} \\ &\triangleq \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\xi}, \end{aligned} \quad (11)$$

where $\mathbf{I}$ is an identity matrix. The weighted least-square (WLS) estimate of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}}(\mathbf{b} - \mathbf{A}\boldsymbol{\theta})^H\mathbf{W}(\mathbf{b} - \mathbf{A}\boldsymbol{\theta}) = \left(\mathbf{A}^H\mathbf{W}\mathbf{A}\right)^{-1}\mathbf{A}^H\mathbf{W}\mathbf{b}, \quad (12)$$

where $\mathbf{W}$ is a positive Hermitian matrix.

Similar to the RR method in [19], the proposed method is an extension of [18] to multichannel applications. However, it differs from the RR method in the reference signal. In the proposed method, the reference signal is a linear combination of the multichannel signals, while the RR method is a special case of the proposed method when $\mathbf{d}(\omega)$ has one nonzero entry, for example, $[0\cdots010\cdots0]^T$, which means it only uses the signal of one selected channel.

Following similar analysis in [18], it shows that the proposed method produces unbiased estimate of $\boldsymbol{\theta}$. Moreover, the estimation error variance of each element of ASV increases when the nonstationarity of the reference signal reduces. In [18], the signal nonstationarity is indicated by

$$k(\omega) \triangleq \langle\sigma_u^2(m,\omega)\rangle\left\langle\frac{1}{\sigma_u^2(m,\omega)}\right\rangle, \quad (13)$$

where the operator $\langle\cdot\rangle$ is temporal average defined by [18]

$$\left\langle\sigma_u^2(m,\omega)\right\rangle \triangleq \sum_{m=1}^{L}\alpha_m\sigma_u^2(m,\omega), \tag{14}$$

where α_m is positive weight.

Large value of $k(\omega)$ is appreciated to produce estimate with low error variance. For a given nonstationary signal corrupted by stationary noise, here we show how $k(\omega)$ is affected by SNR. According to AS2, the spectrum $\sigma_u^2(m,\omega)$ can be expressed as

$$\sigma_u^2(m,\omega) = \left|b(\omega)\right|^2\sigma_s^2(m,\omega) + \sigma_{\bar{n}}^2(\omega), \tag{15}$$

where $\sigma_s^2(m,\omega)$ and $\sigma_{\bar{n}}^2(\omega)$ are the spectra of speech signal $s(m,\omega)$ and noise $\bar{n}(\omega)$, respectively. We define the SNR $\rho(m,\omega)$ of mth interval as

$$\rho(m,\omega) = \frac{\left|b(\omega)\right|^2\sigma_s^2(m,\omega)}{\sigma_{\bar{n}}^2(\omega)}. \tag{16}$$

The nonstationarity indicator $k(\omega)$ in (13) can be expressed in terms of SNR $\rho(m,\omega)$ as

$$k(\rho,\omega) = \left\langle\rho(m,\omega)+1\right\rangle\left\langle\frac{1}{\rho(m,\omega)+1}\right\rangle. \tag{17}$$

With the same nonstationary speech signal $s(m,\omega)$ and different levels of noise, $\sigma_{\bar{n}_1}^2(\omega)$ and $\sigma_{\bar{n}_2}^2(\omega)$, where $\sigma_{\bar{n}_1}^2(\omega) \leq \sigma_{\bar{n}_2}^2(\omega)$, we have

$$\begin{aligned}\rho_1(m,\omega) &= \frac{\left|b(\omega)\right|^2\sigma_s^2(m,\omega)}{\sigma_{\bar{n}_1}^2(\omega)} \geq \rho_2(m,\omega)\\ &= \frac{\left|b(\omega)\right|^2\sigma_s^2(m,\omega)}{\sigma_{\bar{n}_2}^2(\omega)}.\end{aligned} \tag{18}$$

Defining $\beta = (\sigma_{\bar{n}_1}^2(\omega)/\sigma_{\bar{n}_2}^2(\omega)) \leq 1$, we have

$$\rho_2(m,\omega) = \beta\rho_1(m,\omega). \tag{19}$$

Comparing the values of $k(\rho_1,\omega)$ and $k(\rho_2,\omega)$, we have

$$\begin{aligned}
&k(\rho_1,\omega) - k(\rho_2,\omega)\\
&\quad= \left\langle\rho_1(m,\omega)+1\right\rangle\left\langle\frac{1}{\rho_1(m,\omega)+1}\right\rangle\\
&\qquad- \left\langle\rho_2(m,\omega)+1\right\rangle\left\langle\frac{1}{\rho_2(m,\omega)+1}\right\rangle\\
&\quad= \sum_{i=1}^{L}\sum_{j=1}^{L}\alpha_i\alpha_j\frac{\rho_1(i,\omega)+1}{\rho_1(j,\omega)+1} - \sum_{i=1}^{L}\sum_{j=1}^{L}\alpha_i\alpha_j\frac{\rho_2(i,\omega)+1}{\rho_2(j,\omega)+1}\\
&\quad= \sum_{i=2}^{L}\sum_{j=1}^{i-1}\alpha_i\alpha_j\left(\frac{\rho_1(i,\omega)+1}{\rho_1(j,\omega)+1} + \frac{\rho_1(j,\omega)+1}{\rho_1(i,\omega)+1}\right.\\
&\qquad\qquad\left. - \frac{\rho_2(i,\omega)+1}{\rho_2(j,\omega)+1} - \frac{\rho_2(j,\omega)+1}{\rho_2(i,\omega)+1}\right)\\
&\quad= \sum_{i=2}^{L}\sum_{j=1}^{i-1}\alpha_i\alpha_j(1-\beta)\\
&\qquad\times\frac{\left(\rho_1(i,\omega)-\rho_1(j,\omega)\right)^2\left(\beta\rho_1(i,\omega)+\beta\rho_1(j,\omega)+\beta+1\right)}{\left(\rho_1(j,\omega)+1\right)\left(\beta\rho_1(j,\omega)+1\right)\left(\rho_1(i,\omega)+1\right)\left(\beta\rho_1(i,\omega)+1\right)}\\
&\quad\geq 0.
\end{aligned} \tag{20}$$

Therefore, we have $k(\rho_1,\omega) \geq k(\rho_2,\omega)$ if $\rho_1(m,\omega) \geq \rho_2(m,\omega)$. It can be explained that, with the existence of stronger stationary noise, the nonstationarity of the signal $u(m,\omega)$ is weaker. If noise is dominant ($\rho(m,\omega) \ll 1$), signal $u(m,\omega)$ becomes almost stationary ($k(\rho,\omega) \approx 1$). It results in estimate with infinite variance ([18, equation (28)]). In such case, the proposed method cannot work. To decrease the estimation error, high SNR of the reference signal is appreciated.

The RR method in [19] uses one selected channel as the reference which is the filtered version of the speech signal by the channel impulse response. In the frequency bin where either the channel has low response or the target signal has low power, the resulting low SNR causes significant estimation error. With multichannel signals available, it is possible to combine them to produce a reference signal with higher SNR. Therefore, we optimize $\mathbf{d}(\omega)$ to maximize the SNR of the reference signal at frequency ω:

$$\mathbf{d}_{\text{opt}}(m,\omega) = \arg\max_{\mathbf{d}(\omega)}\frac{\mathbf{d}^H(\omega)C_s(m,\omega)\mathbf{d}(\omega)}{\mathbf{d}^H(\omega)C_n(\omega)\mathbf{d}(\omega)}, \tag{21}$$

where

$$\begin{aligned}C_s(m,\omega) &= E\left\{s(m,\omega)\mathbf{h}(\omega)s(m,\omega)\mathbf{h}^H(\omega)\right\}\\ &= \sigma_s^2(m,\omega)\mathbf{h}(\omega)\mathbf{h}^H(\omega),\\ C_n(\omega) &= E\left\{\mathbf{n}(\omega)\mathbf{n}^H(\omega)\right\}\end{aligned} \tag{22}$$

are the covariance matrices of the distorted speech signal $s(m,\omega)\mathbf{h}(\omega)$ and the noise $\mathbf{n}(\omega)$. A speech signal is a quasistationary signal in a short period of time. Its covariance matrix can be estimated using temporal averaging [20]. The noise covariance matrix $C_n(\omega)$ is estimated during the speech pause in the case where a robust speech detection algorithm [21] is used. Because of the independence of speech and noise, the covariance matrix of speech signal can be estimated as

$$C_s(m,\omega) = C_x(m,\omega) - C_n(\omega), \tag{23}$$

where $C_x(m,\omega) = E\{\mathbf{x}(m,\omega)\mathbf{x}(m,\omega)^H\}$, which can be estimated when speech signal is active. With the estimated matrices of $C_n(\omega)$ and $C_s(m,\omega)$, $\mathbf{d}_{\text{opt}}(m,\omega)$ is given as the eigenvector of matrix $C_n^{-1}(\omega)C_s(m,\omega)$ corresponding to the largest eigenvalue [22]. The matrix $C_n^{-1}(\omega)C_s(m,\omega)$ only differs in a varying scale $\sigma_s^2(m,\omega)$ at a different time. Since the multiplicative nonzero scale does not change the eigenvectors of a matrix, we can consider that the estimated $\mathbf{d}_{\text{opt}}(m,\omega)$ is time invariant, that is, $\mathbf{d}_{\text{opt}}(\omega) = \mathbf{d}_{\text{opt}}(m,\omega)$. In practice, the estimate of $\mathbf{d}_{\text{opt}}(\omega)$ can be performed at long intervals if the signal environment does not change. The obtained $\mathbf{d}_{\text{opt}}(\omega)$ is

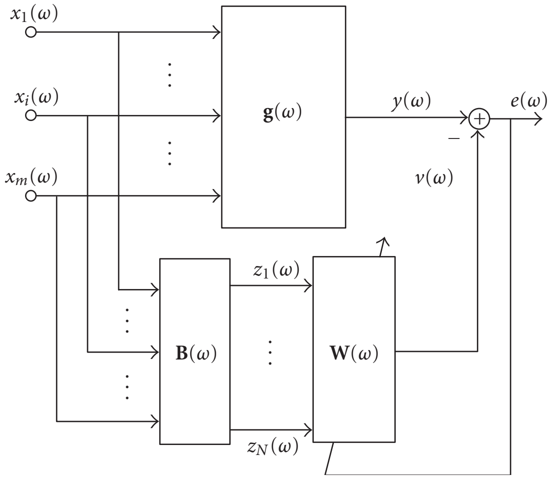

Figure 1: Schematic of MF-GSC in frequency domain.

then used to produce the reference signal $u(m,\omega)$ with higher SNR for ASV identification.

The estimated vector $\bar{\mathbf{h}}(\omega)$ has a multiplicative scale $1/b(\omega)$. This scale varies in different frequency bins, which introduces additional magnitude response perturbation in the output of beamformer if $\bar{\mathbf{h}}(\omega)$ is directly applied without any modification. In this paper, we propose a simple skill to normalize $\bar{\mathbf{h}}(\omega)$. The resulting normalized ASV cannot only eliminate the effect of multiplicative scale $1/b(\omega)$, but can also reduce the reverberation effects of the acoustic enclosure [23]. The new ASV $\tilde{\mathbf{h}}(\omega)$ is obtained by normalizing the vector $\bar{\mathbf{h}}(\omega)$ to its lth entry $\bar{h}_l(\omega)$ which has the largest norm, $|\bar{h}_l(\omega)| = max\{|\bar{h}_1(\omega)|, |\bar{h}_2(\omega)|, \ldots, |\bar{h}_M(\omega)|\}$:

$$\tilde{\mathbf{h}}(\omega) = \left[\frac{\bar{h}_1(\omega)}{\bar{h}_l(\omega)} \cdots \frac{\bar{h}_M(\omega)}{\bar{h}_l(\omega)}\right]^T = \left[\tilde{h}_1(\omega) \cdots \tilde{h}_M(\omega)\right]^T, \tag{24}$$

where $\tilde{h}_l(\omega) \equiv 1$. The new vector $\tilde{\mathbf{h}}(\omega)$ is unique. We still call $\tilde{\mathbf{h}}(\omega)$ an ASV. This normalization is carried out in each frequency bin independently. Therefore, the index l is frequency dependent.

4. MATCHED-FILTER-ARRAY-BASED GSC (MF-GSC) IN FREQUENCY DOMAIN

When microphone array works in acoustic enclosure with multipath and reverberant effects, the conventional beamformer based on pure delay model becomes inefficient. One severe problem is the target signal cancellation. In such case, a new robust beamformer should be designed. In this section, we propose an MF-array-based [13, 24, 25] GSC [3] exploiting the estimated ASV. The resulting array not only combats reverberant effects, but also suppresses the environment noise effectively.

The schematic of MF-GSC in frequency domain is shown in Figure 1. The array observed signal is firstly transformed into frequency domain. Next, signal at each frequency bin is processed by MF-GSC. The output signal at each frequency bin is finally transformed back to time domain to produce the enhanced output signal using overlap-and-save method [26].

Conventional GSC has three major parts, including fixed beamformer $\mathbf{g}(\omega)$, blocking filter $\mathbf{B}(\omega)$, and multichannel adaptive filter $\mathbf{w}(\omega)$. To utilize the estimated ASV $\tilde{h}(\omega)$, the conventional GSC [3] is modified. In the following context, the modification and its effects on system performance are presented in detail.

4.1. Fixed beamformer

The fixed beamformer is modified into a multiple-input MF array with transfer function $\mathbf{g}(\omega) = \mathbf{h}(\omega)/\|\mathbf{h}(\omega)\|^2$, which coherently sums all the multipath signals to achieve maximum SNR and dereverberate the target signal [13, 24, 25]. Since the true $\mathbf{h}(\omega)$ cannot be obtained, we use the estimated ASV vector $\tilde{\mathbf{h}}(\omega)$ instead. The output of the fixed beamformer (matched-filter array) is

$$y(m,\omega) = \frac{\tilde{\mathbf{h}}^H(\omega)\mathbf{x}(m,\omega)}{\|\tilde{\mathbf{h}}(\omega)\|^2}. \tag{25}$$

The overall response of GSC to the target signal is theoretically determined by the response of the fixed beamformer. From (25), the speech component $\hat{s}(m,\omega)$ in the output signal of fixed beamformer is

$$\hat{s}(m,\omega) = h_l(\omega)s(m,\omega), \tag{26}$$

where $h_l(\omega)$ is the element of $\mathbf{h}(\omega)$ which has the largest norm. The reverberation is caused by the ripples of room response. As the location of the microphone in the room changes, the position of the ripple also changes. Therefore, if there are sufficient microphones and they have enough dispersiveness, the largest response at each frequency bin can be used to reduce the ripple, that is, reverberation effects [23].

4.2. Blocking filter

The blocking filter[1] is used to suppress the target signal but passes the interference and noise as much as possible. In reverberant environment, since the signal components at different frequency bins have different response characteristics, the conventional blocking filter [3] cannot block the target signal efficiently. When the target signal leaks into the multichannel adaptive filter, it results in source signal cancellation. Therefore, the conventional blocking filter must also be modified to introduce temporal information in order to block all the components of the target signal. In this paper, we propose a simple blocking matrix design method, which is easy to implement and is able to suppress the target signal as well as its multipath signals.

In noiseless case,

$$x_i(m,\omega)h_j(\omega) - x_j(m,\omega)h_i(\omega) = 0, \quad i \neq j. \tag{27}$$

[1] It is also called blocking matrix in conventional GSC. In the following context, these two terms are exchangeable.

(27) indicates that the target signal is blocked for any $\{h_i(\omega)\}$. Since $\{h_i(\omega)\}$ are unknown, we use $\{\tilde{h}_i(\omega)\}$ instead because $\{\tilde{h}_i(\omega)\}$ are just the scaled version of $\{h_i(\omega)\}$. This does not affect the blocking filter. A blocking filter $\mathbf{B}(\omega)$ which is slightly different from the one in [19] is constructed as

$$\mathbf{B}(\omega) = \begin{bmatrix} \tilde{h}_l(\omega) & 0 & -\tilde{h}_1(\omega) & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & -\tilde{h}_{l-1}(\omega) & 0 & \cdots & 0 \\ 0 & \cdots & -\tilde{h}_{l+1}(\omega) & \tilde{h}_l(\omega) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & -\tilde{h}_M(\omega) & 0 & \cdots & \tilde{h}_l(\omega) \end{bmatrix}, \quad (28)$$

where the signal of lth channel is used as a reference signal to remove the target signal component in other $M-1$ channels. The output signal $\mathbf{z}(m,\omega)$ of the blocking filter is given by

$$\mathbf{z}(m,\omega) = \mathbf{B}(\omega)\mathbf{x}(m,\omega). \quad (29)$$

4.3. Multichannel adaptive filter

The output signal $v(m,\omega)$ of the multichannel adaptive filter is given by

$$v(m,\omega) = \mathbf{w}^H(\omega)\mathbf{z}(m,\omega), \quad (30)$$

where $\mathbf{w}(\omega)$ represents the adaptive filter coefficients in frequency domain. The output signal $e(m,\omega)$ of the MF-GSC is given by

$$e(m,\omega) = y(m,\omega) - v(m,\omega). \quad (31)$$

The optimal adaptive weight $\mathbf{w}(\omega)$ is obtained by solving the following optimization problem:

$$\min_{\mathbf{w}} E\{||y(m,\omega) - \mathbf{w}^H\mathbf{B}(\omega)\mathbf{x}(m,\omega)||^2\}. \quad (32)$$

The optimal solution $\mathbf{w}_{\text{opt}}(\omega)$ of (32) is easily obtained using well-known least-mean-square (LMS) method [27, 28]. In this paper, we use the leaky normalized least-mean-square (NLMS) method [28] instead of its robustness to small imperfection. The updating equation is expressed as

$$\begin{aligned} \mathbf{w}(m+1,\omega) &= \beta\mathbf{w}(m,\omega) + \rho_f \frac{e^*(m,\omega)\mathbf{z}(m,\omega)}{P_x(m,\omega)+\delta}, \\ P_x(m,\omega) &= \alpha P_x(m,\omega) + (1-\alpha)||\mathbf{x}(m,\omega)||^2, \end{aligned} \quad (33)$$

where $\beta(0 < \beta \le 1)$ is the leakage parameter, $\alpha(0 < \alpha < 1)$ is the forgetting factor, and $\rho_f(0 < \rho_f < 2)$ is the step size. δ is a small positive constant to avoid gradient amplification problem. When $\beta = 1$, (33) is similar to the NLMS algorithm. Since the weight $\mathbf{w}(m,\omega)$ should be updated only when there is no target signal, in (33), the power of signal $\mathbf{x}(m,\omega)$ is used instead of the power of $\mathbf{z}(m,\omega)$. With this modification, the multichannel noise canceller can always be on due to the fact that the adaptation term $\rho_f(e^*(m,\omega)\mathbf{z}(m,\omega)/P_x(m,\omega))$ is very small when target speech signal exists. Since the blocking filter and adaptive multichannel filter are not necessary causal filters, in this paper, noncausal FIR structure constraint [19] was used in the simulation. The coefficient of noncausal FIR filter has the form $[h(-L),\dots,h(R)]$, where L and R are half the filter length.

5. NUMERICAL STUDY

In this section, we evaluate the performance of the proposed method through simulation experiments on a microphone array system for speech enhancement and noise suppression. The microphone array consists of 5 elements, whose coordinates in x-axis are 1.24 m, 1.35 m, 1.44 m, 1.51 m, and 1.56 m, respectively. Coordinates in y-axis and z-axis are both 2.0 m and 1.5 m. It is placed in a small room with dimension $(x \times y \times z) = (2.8\,\text{m} \times 3.2\,\text{m} \times 2.2\,\text{m})$, wall reflection coefficient 0.6, and floor/celling reflection coefficient 0.4. The channel impulse response (IR) relating the speech source and each microphone is calculated by an image method [29] with a sampling rate 8 kHz. Simulation shows that the room is a reverberant environment. The channel IR has a long tail but decays quickly. A speech signal source is placed on the spot (0.4 m, 1.0 m, 1.5 m). Its interested frequency ranges from 150 Hz to 3.7 kHz. A point noise source recorded from a real conference room is placed on the spot (2.4 m, 1.0 m, 1.5 m).

Since the target speech is nonstationary, herein, we use average SNR $\varsigma(\omega)$ defined as

$$\varsigma(\omega) = 10\log_{10}\frac{\sum_{i=1}^{T}|s'(i,\omega)|^2}{\sum_{i=1}^{T}|n(i,\omega)|^2}, \quad (34)$$

where $n(i,\omega)$ and $s'(i,\omega)$ represent the noise and the distorted target speech signal by the acoustic channel impulse response at ωth frequency bin, respectively. T is the number of signal samples used to estimate the average SNR.

The array received signals were segmented into blocks of length 512. These data blocks were transformed into frequency domain by FFT. The system identification procedure utilized 13 segments. The length of each segment was 1024 samples. Speech detector [21] was used in the estimation of the covariance matrices of noise and speech signal, through which the optimal weight $\mathbf{d}(\omega)$ was obtained. Relative estimation error E_r is used as a performance evaluation criterion for ASV estimation, where E_r is defined as $E_r(\omega) = E_p(\omega)/\|\mathbf{h}(\omega)\|^2$, and $E_p(\omega)$ is the error between the true ASV and the projection misalignment vector [30] of the estimated ASV $\bar{\mathbf{h}}(\omega)$. It is defined as

$$\begin{aligned} E_p(\omega) &\triangleq \min_g ||\mathbf{h}(\omega) - g\bar{\mathbf{h}}(\omega)||^2 \\ &= \left\|\mathbf{h}(\omega) - \frac{\bar{\mathbf{h}}^H(\omega)\mathbf{h}(\omega)}{||\bar{\mathbf{h}}(\omega)||^2}\bar{\mathbf{h}}(\omega)\right\|^2. \end{aligned} \quad (35)$$

With such defined error $E_p(\omega)$, the effect of arbitrary nonzero multiplicative scale in the estimated ASV is eliminated.

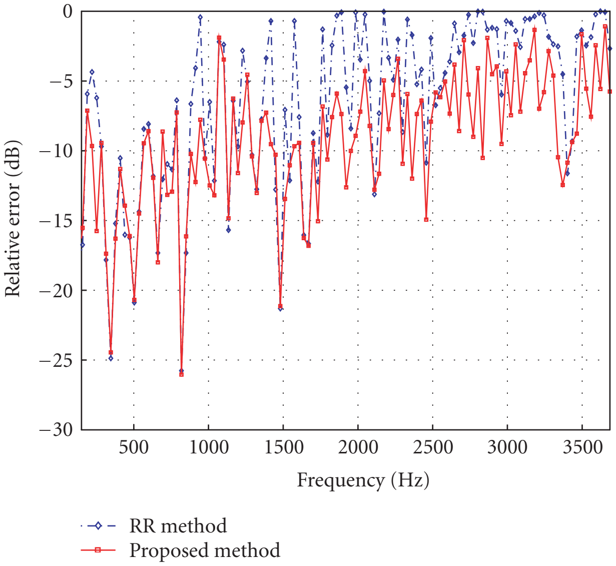

FIGURE 2: Comparison of the relative estimation error between the RR method and the proposed method at 0 dB input SNR.

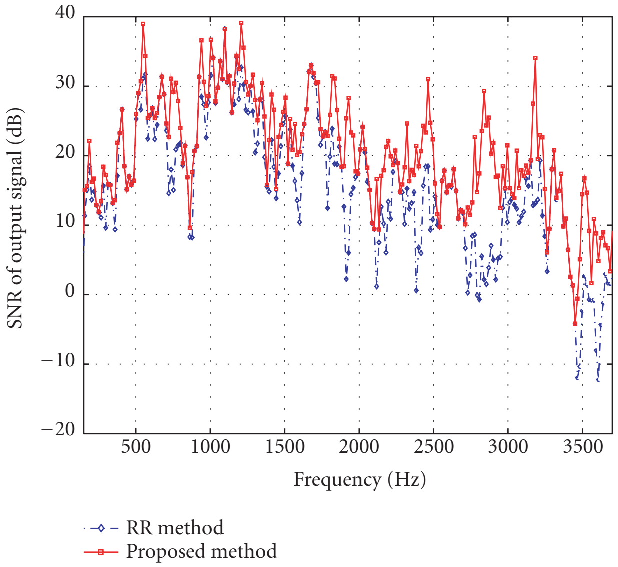

FIGURE 3: Comparison of the overall system output SNR between RR-GSC and MF-GSC at 0 dB input SNR.

In Figure 2, an example of the relative ASV estimation error is shown to compare the performance of the RR method with the proposed method. The experiment was carried out at 0 dB input SNR. It shows that, at most of the frequency bins, the proposed method has lower estimation error due to its SNR improvement of the reference signal.

The following experiments show the SNR improvement brought by the proposed MF-GSC compared with the GSC based on ASV estimated by the RR method (named RR-GSC). RR-GSC has similar implementation as the beamformer in [19]. In Figure 3, the overall system output SNR of MF-GSC and RR-GSC was compared. The SNR improvement of MF-GSC is higher than RR-GSC at most of the frequency bins, which can be explained by the improved accuracy of ASV estimation.

In Figure 4, we present the system performance under a different room reverberation time T_{60}. In these simulations, the parameters of the beamformers were fixed except the reverberation time of the acoustic enclosure. It is clear that the SNR improvement degrades for both methods if the acoustic enclosure has a longer reverberation time. However, for practical reverberation time, both methods still produce output signals with high SNR. On the other hand, with all the reverberation time under consideration, the output SNR of the proposed method is observed to be higher than that of RR-GSC.

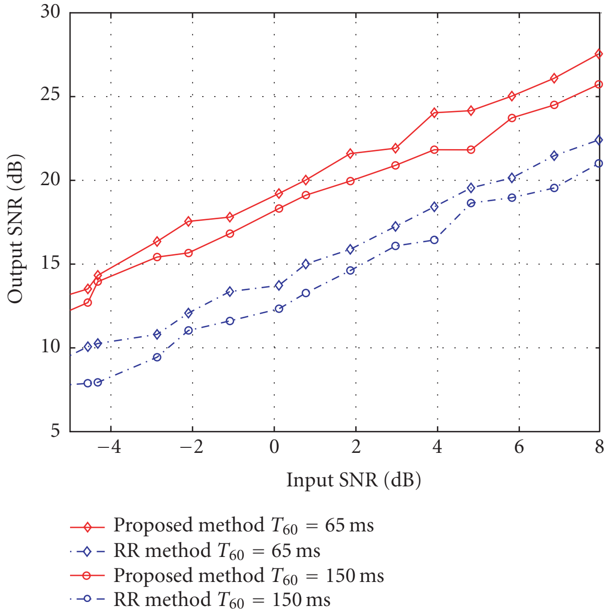

FIGURE 4: Comparison of the mean output SNR between RR-GSC and MF-GSC versus different input SNRs and different reverberation times.

6. CONCLUSION

A robust microphone array for speech enhancement and noise suppression is proposed in this paper. We present a method to improve estimation accuracy of the array steering vector (ASV) by utilizing the nonstationarity of speech signal and the stationarity of noise. The MF-GSC constructed by means of the above estimated ASV has higher SNR improvement than the RR-GSC method. The proposed method has the advantage that it can work in highly reverberant environments at low input SNR if the noise is stationary.

REFERENCES

[1] J. S. Lim, Ed., *Speech Enhancement*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1983.

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

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

[4] 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, 1988.

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

[6] M. H. Er and A. Cantoni, "Derivative constraints for broadband element space antenna array processors," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 31, no. 6, pp. 1378–1393, 1983.

[7] K. Buckley and L. Griffiths, "An adaptive generalized sidelobe canceller with derivative constraints," *IEEE Trans. Antennas Propagat.*, vol. 34, no. 3, pp. 311–319, 1986.

[8] M. H. Er and A. Cantoni, "An unconstrained partitioned realization for derivative constrained broad-band antenna array processors," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 6, pp. 1376–1379, 1986.

[9] G. L. Fudge, "Spatial blocking filter derivative constraints for the generalized sidelobe canceller and MUSIC," *IEEE Trans. Signal Processing*, vol. 44, no. 1, pp. 51–61, 1996.

[10] M. H. Er and A. Cantoni, "A new set of linear constraints for broad-band time domain element space processor," *IEEE Trans. Antennas Propagat.*, vol. 34, no. 3, pp. 320–329, 1986.

[11] K. M. Buckley, "Spatial/spectral filtering with linearly constrained minimum variance beamformers," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, no. 3, pp. 249–266, 1987.

[12] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 5, pp. 425–437, 1997.

[13] E. E. Jan, P. Svaizer, and J. L. Flanagan, "Matched-filter processing of microphone array for spatial volume selectivity," in *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS '95)*, vol. 2, pp. 1460–1463, Seattle, Wash, USA, April–May 1995.

[14] G. B. Giannakis, Y. Hua, P. Stoica, and L. Tong, Eds., *Signal Processing Advances in Wireless and Mobile Communications: Trends in Channel Estimation and Equalization*, vol. 1, Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001.

[15] Z. Ding and Y. Li, *Blind Equalization and Identification*, Marcel Dekker, New York, NY, USA, 2001.

[16] C. Y. Suen and R. D. Mori, *Computer Analysis and Perception*, CRC Press, Boca Raton, Fla, USA, 1982.

[17] Z. L. Yu and M. H. Er, "A generic upper bound of perturbation on subspace decomposition," submitted to IEEE Trans. Signal Processing.

[18] O. Shalvi and E. Weinstein, "System indentification using nonstationary signals," *IEEE Trans. Signal Processing*, vol. 44, no. 8, pp. 2055–2063, 1996.

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

[20] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1978.

[21] S. V. Gerven and F. Xie, "A comparative study of speech detection methods," in *Proc. 5th European Conference on Speech Communication and Technology (EUROSPEECH '97)*, vol. 3, pp. 1095–1098, Rhodes, Greece, September 1997.

[22] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.

[23] J. L. Flanagan and R. C. Lummis, "Signal processing to reduce multipath distortion in small rooms," *Journal of the Acoustical Society of America*, vol. 47, pp. 1475–1481, 1970.

[24] J. L. Flanagan, A. C. Surendran, and E. E. Jan, "Spatially selectivity sound capture for speech and audio processing," *Speech Communication*, vol. 13, no. 1-2, pp. 207–222, 1993.

[25] E. E. Jan and J. Flanagan, "Sound capture from spatial volumes: matched-filter processing of microphone arrays having randomly-distributed sensors," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '96)*, vol. 2, pp. 917–920, Atlanta, Ga, USA, May 1996.

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

[27] B. Widrow, J. Glover, J. McCool, et al., "Adaptive noise cancellating: Principles and applications," *Proc. IEEE*, vol. 63, no. 12, pp. 1692–1716, 1975.

[28] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, Englewood Cliffs, NJ, USA, 3rd edition, 1996.

[29] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

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

Zhu Liang Yu received the B.Eng. and M.Eng. degrees in electronic engineering from Nanjing University of Aeronautics and Astronautics, China, in 1995 and 1998, respectively. From 1998 to 2000, he was a Software Engineer of Shanghai Bell Co. Ltd. In 2000, he joined Center for Signal Processing, Nanyang Technological University, as a Research Engineer. Currently he is a Ph.D. candidate in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include array signal processing, acoustic signal processing, and adaptive signal processing.

Meng Hwa Er received the B.Eng. degree in electrical engineering with first-class honors from the National University of Singapore in 1981, and the Ph.D. degree in electrical and computer engineering from the University of Newcastle, Australia, in 1986. He joined the Nanyang Technological Institute/University in 1985 and was promoted to a Full Professor in 1996. He served as an Associate Editor of the IEEE Transactions on Signal Processing from 1997 to 1998 and is a Member of the Editorial Board of IEEE Signal Processing Magazine from 2005 to 2007. His research interests include array signal processing, satellite communications, computer vision, and optimization techniques.