Multichannel Post-Filtering in Nonstationary Noise Environments

Israel Cohen, *Senior Member, IEEE*

***Abstract*—In this paper, we present a multichannel post-filtering approach for minimizing the log-spectral amplitude distortion in nonstationary noise environments. The beamformer is realistically assumed to have a steering error, a blocking matrix that is unable to block all of the desired signal components, and a noise canceller that is adapted to the pseudo-stationary noise but not modified during transient interferences. A mild assumption is made that a desired signal component is stronger at the beamformer output than at any reference noise signal, and a noise component is strongest at one of the reference signals. The ratio between the transient power at the beamformer output and the transient power at the reference noise signals is used to indicate whether such a transient is desired or interfering. Based on a Gaussian statistical model and combined with an appropriate spectral enhancement technique, we derive estimators for the signal presence probability, the noise power spectral density, and the clean signal. The proposed method is tested in various nonstationary noise environments. Compared with single-channel post-filtering, a significantly reduced level of nonstationary noise is achieved without further distorting the desired signal components.**

***Index Terms*—Acoustic noise measurement, adaptive signal processing, array signal processing, signal detection, spectral analysis, speech enhancement.**

I. Introduction

Multichannel systems are often used for high-quality hands-free communication in reverberant and noisy environments [1]. Compared with single channel systems, a substantial gain in performance is obtainable due to the spatial filtering capability to suppress interfering signals coming from undesired directions. However, in cases of spatially incoherent noise fields, beamforming alone does not provide sufficient noise reduction, and post-filtering is normally required [2], [3].

Multichannel post-filtering, generalized to an arbitrary number of sensors, was first introduced by Zelinski [4], [5]. Accordingly, the output of a delay-and-sum beamformer is post-filtered using an adaptive Wiener filtering in the time domain, based on the auto and cross spectral densities of the sensor signals. However, Zelinski's approach overestimates the noise power density and, therefore, is not optimal in the Wiener sense [6]. A modified post-filtering version was suggested by Simmer and Wasiljeff, which employs the power spectral density of the beamformer output, rather than the average of the power spectral densities of individual sensor signals [6]. The underlying assumption is that noise components at different sensors are mutually uncorrelated. Unfortunately, in a diffuse noise field, where the low-frequency noise components are coherent, the noise reduction performance severely deteriorates.

To overcome this problem, Fischer *et al.* [7]–[9] proposed a noise reduction system, which is based on the generalized sidelobe canceller (GSC). The GSC reasonably suppresses the coherent noise components, whereas a Wiener filter in the look direction is designed to suppress the spatially incoherent noise components. Bitzer *et al.* analyzed the performance of the GSC and adaptive post-filtering techniques in various noise fields [10], [11]. They showed that in a diffuse noise field, neither the GSC nor the adaptive post-filtering performs well at low frequencies. Therefore, at the output of a GSC with standard Wiener post-filtering, they used a second post-filter to reduce the spatially correlated noise components [12], [13]. Le Bouquin-Jeannès *et al.* suggested the modification of the cross power spectrum estimation and the Wiener post-filtering to take the presence of some correlated noise components into account [14]. The cross power spectrum of the noise signals is averaged during pauses in the desired signal. Subsequently, it is subtracted from the cross power spectrum of the sensor signals, which is calculated during signal presence. Meyer and Simmer [15] proposed to combine a delay-and-sum beamformer with Wiener filtering and spectral subtraction. The Wiener filtering is applied in the high-frequency band for the suppression of low-coherence noise components, whereas the spectral subtraction is used in the low-frequency band for high-coherence noise reduction. Mamhoudi [16] and Mamhoudi and Drygajlo [17] considered a nonlinear coherence filtering in the wavelet domain to improve the performance of the Wiener post-filtering. Instead of the conventional coherence between the individual sensor signals, they used the coherence between the output and the input of the beamformer sensor signals, which is assumed to be low, even for correlated noise components. Fischer and Kameyer [18] suggested the application of Wiener filtering to the output of a broadband beamformer, which is built up by several harmonically nested subarrays. They showed that the resulting noise-reduction system performance is nearly independent of the correlation properties of the noise field. This structure has been further analyzed by Marro *et al.* [2]. McCowan *et al.* used a near-field super-directive beamforming and investigated the effect of a Wiener post-filter on speech recognition performance [19]. They showed that in the case of nearfield sources and diffuse noise conditions, improved recognition performance can be achieved compared with conventional adaptive beamformers. A theoretical analysis of Wiener multichannel post-filtering is presented in [3].

Manuscript received April 18, 2002; revised July 15, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Inbar Fijalkow.

The author is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Technion City, Haifa 32000, Israel (e-mail: icohen@ee.technion.ac.il).

Digital Object Identifier 10.1109/TSP.2004.826166

A major drawback of existing multichannel post-filtering techniques is that highly nonstationary noise components are not dealt with. The time variation of the interfering signals is assumed to be sufficiently slow, such that the post-filter can track and adapt to the changes in the noise statistics. Unfortunately, transient interferences are often much too brief and abrupt for the above post-filtering methods. Furthermore, Wiener filtering minimizes the mean-square error (MSE) distortion of the signal estimate, which is essentially not the optimal criterion for enhancing noisy speech. A more appropriate distortion measure for speech-enhancement systems is based on the MSE of the spectral, or log-spectral, amplitude [20], [21].

In this paper, we present a multichannel post-filtering approach to minimize the log-spectral amplitude distortion in nonstationary noise environments. Presumably, a desired signal component is stronger at the beamformer output than at any reference noise signal, and a noise component is strongest at one of the reference signals. Hence, the ratio between the transient power at beamformer output and the transient power at the reference signals indicates whether such a transient is desired or interfering. Based on a Gaussian statistical model [20] and an appropriate decision-directed *a priori* SNR estimate [22], we derive an estimator for the signal presence probability. This estimator controls the rate of recursive averaging for obtaining a noise spectrum estimate by the *minima controlled recursive averaging* (MCRA) approach [22], [23]. Subsequently, spectral enhancement of the beamformer output is achieved by applying an optimal gain function, which minimizes the MSE of the log-spectra. The performance of the proposed post-filtering approach is evaluated under nonstationary noise conditions using objective quality measures, a subjective study of speech spectrograms, and informal listening tests. We show that single-channel post-filtering is inefficient at attenuating highly nonstationary noise components since it lacks the ability to differentiate such components from the desired source components. By contrast, the proposed multichannel post-filtering approach achieves a significantly reduced level of background noise, whether stationary or not, without distorting the signal components further.

The paper is organized as follows. In Section II, we review the linearly constrained adaptive beamformer and derive relations in the power-spectral domain between the beamformer output, the reference noise signals, the desired source signal, and the input transient interferences. In Section III, the problem of signal detection in the time-frequency plane is addressed. Signal components are discriminated from transient noise components based on the transient power ratio between the beamformer output and the reference signals. In Section IV, we introduce an estimator for the time-varying spectrum of the beamformer output noise and describe the multichannel post-filtering approach. Finally, in Section V, we evaluate the proposed method and present experimental results, which validate its effectiveness.

II. Linearly Constrained Adaptive Beamforming

Let $x(t)$ denote a desired source signal, and let signal vectors $\mathbf{d}_s(t)$ and $\mathbf{d}_t(t)$ denote multichannel uncorrelated interfering signals at the output of M sensors. The vector $\mathbf{d}_s(t)$ represents pseudo-stationary interferences, and $\mathbf{d}_t(t)$ represents undesired transient components. The observed signal at the ith sensor is given by

$$z_i(t) = a_i(t) * x(t) + d_{is}(t) + d_{it}(t), \quad i = 1, \ldots, M \tag{1}$$

where $a_i(t)$ is the impulse response of the ith sensor to the desired source, $*$ denotes convolution, and d_{is} and d_{it} are the interference signals corresponding to the ith sensor. The observed signals are divided in time into overlapping frames by the application of a window function and analyzed using the short-time Fourier transform (STFT). Assuming time-invariant transfer functions [24], we have in the time-frequency domain

$$\mathbf{Z}(k,\ell) = \mathbf{A}(k)X(k,\ell) + \mathbf{D}_s(k,\ell) + \mathbf{D}_t(k,\ell) \tag{2}$$

where k represents the frequency bin index, ℓ the frame index, and

$$\begin{aligned}\mathbf{Z}(k,\ell) &\triangleq [Z_1(k,\ell) \quad Z_2(k,\ell) \quad \cdots \quad Z_M(k,\ell)]^T\\ \mathbf{A}(k) &\triangleq [A_1(k) \quad A_2(k) \quad \cdots \quad A_M(k)]^T\\ \mathbf{D}_s(k,\ell) &\triangleq [D_{1s}(k,\ell) \quad D_{2s}(k,\ell) \quad \cdots \quad D_{Ms}(k,\ell)]^T\\ \mathbf{D}_t(k,\ell) &\triangleq [D_{1t}(k,\ell) \quad D_{2t}(k,\ell) \quad \cdots \quad D_{Mt}(k,\ell)]^T.\end{aligned}$$

We note that in [24], transient interferences are not dealt with. The interfering signals are assumed to be stationary, and signal enhancement is based on the nonstationarity of the desired source signal. In our case, we have to include a mechanism that discriminates interfering transients from desired signal components.

Fig. 1 shows a generalized sidelobe canceller structure for a linearly constrained adaptive beamformer [25], [26], which is also utilizable in case $\mathbf{A}(k)$—the transfer function from the desired source to the sensor array—is arbitrary [24]. The beamformer comprises three parts:

1) a fixed beamformer $\mathbf{W}(k)$ that is proportional to the transfer function ratios $A_1^{-1}(k)\mathbf{A}(k)$;
2) a blocking matrix $\mathbf{B}(k)$, which takes into account the assumed propagation path and constructs the reference noise signals $\{U_i(k,\ell) : 2 \le i \le M\}$;
3) a multichannel adaptive noise canceller $\{H_i(k,\ell) : 2 \le i \le M\}$, which eliminates the stationary noise that leaks through the sidelobes of the fixed beamformer.

We assume that the noise canceller is adapted only to the stationary noise. It is not modified during transient interferences, which are characterized by brief and abrupt variations. Furthermore, we assume that the desired source is distributed and that steering error might occur. Accordingly, some desired signal components may pass through the blocking matrix.

The reference noise signals

$$\mathbf{U}(k,\ell) = [U_2(k,\ell) \ U_3(k,\ell) \ \cdots \ U_M(k,\ell)]^T$$

are generated by applying the blocking matrix to the observed signal vector:

$$\begin{aligned}\mathbf{U}(k,\ell) &= \mathbf{B}^H(k)\mathbf{Z}(k,\ell)\\ &= \mathbf{B}^H(k)\left[\mathbf{A}(k)X(k,\ell) + \mathbf{D}_s(k,\ell) + \mathbf{D}_t(k,\ell)\right].\end{aligned} \tag{3}$$

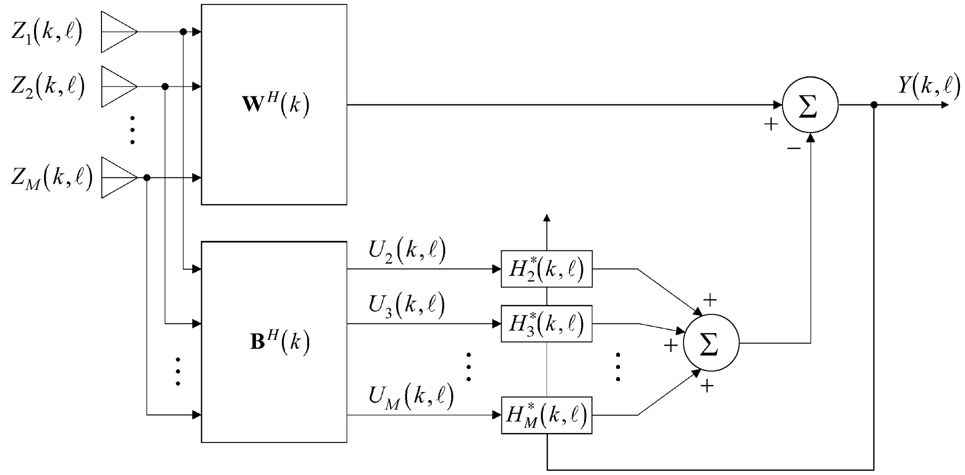

Fig. 1. Block diagram of the Griffiths–Jim adaptive beamformer.

The reference signals are emphasized by the adaptive noise canceller and subtracted from the output of the fixed beamformer, yielding

$$Y(k,\ell) = \left[\mathbf{W}^H(k) - \mathbf{H}^H(k,\ell)\mathbf{B}^H(k)\right]\mathbf{Z}(k,\ell) \tag{4}$$

where $\mathbf{H}(k,\ell) = [H_2(k,\ell) \quad H_3(k,\ell) \quad \cdots \quad H_M(k,\ell)]^T$. The optimal solution for the filters $\mathbf{H}(k,\ell)$ is obtained by minimizing the output power of the stationary noise [27]. Let $\mathbf{\Phi}_{\mathbf{D}_s\mathbf{D}_s}(k,\ell) = E\{\mathbf{D}_s(k,\ell)\mathbf{D}_s^H(k,\ell)\}$ denote the power-spectral density (PSD) matrix of the input stationary noise. Then, the power of the stationary noise at the beamformer output is minimized by solving the unconstrained optimization problem:

$$\begin{aligned}\min_{\mathbf{H}}\Big\{&[\mathbf{W}(k) - \mathbf{B}(k)\mathbf{H}(k,\ell)]^H\,\mathbf{\Phi}_{\mathbf{D}_s\mathbf{D}_s}(k,\ell)\\ &\times[\mathbf{W}(k) - \mathbf{B}(k)\mathbf{H}(k,\ell)]\Big\}.\end{aligned} \tag{5}$$

The multichannel Wiener solution is given by [28]

$$\begin{aligned}\mathbf{H}(k,\ell) = &\left[\mathbf{B}^H(k)\mathbf{\Phi}_{\mathbf{D}_s\mathbf{D}_s}(k,\ell)\mathbf{B}(k)\right]^{-1}\\ &\times\mathbf{B}^H(k)\mathbf{\Phi}_{\mathbf{D}_s\mathbf{D}_s}(k,\ell)\mathbf{W}(k).\end{aligned} \tag{6}$$

If we assume that the stationary, as well as transient, noise fields are homogeneous, then the PSD matrices of the input noise signals are related to the corresponding spatial coherence matrices $\mathbf{\Gamma}_s(k,\ell)$ and $\mathbf{\Gamma}_t(k,\ell)$ by

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{D}_s\mathbf{D}_s}(k,\ell) &= \lambda_s(k,\ell)\mathbf{\Gamma}_s(k,\ell)\\ \mathbf{\Phi}_{\mathbf{D}_t\mathbf{D}_t}(k,\ell) &= \lambda_t(k,\ell)\mathbf{\Gamma}_t(k,\ell)\end{aligned}$$

where $\lambda_s(k,\ell)$ and $\lambda_t(k,\ell)$ represent the input noise power at a single sensor. The input PSD-matrix is therefore given by

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{ZZ}}(k,\ell) = \lambda_x(k,\ell)\mathbf{A}(k)\mathbf{A}^H(k) &+ \lambda_s(k,\ell)\mathbf{\Gamma}_s(k,\ell)\\ &+\lambda_t(k,\ell)\mathbf{\Gamma}_t(k,\ell)\end{aligned} \tag{7}$$

where $\lambda_x(k,\ell) \triangleq E\{|X(k,\ell)|^2\}$ is the PSD of the desired source signal. Using (3) and (4), the PSD matrix of the reference signals and the PSD of the beamformer output are obtained by

$$\mathbf{\Phi}_{\mathbf{UU}}(k,\ell) = \mathbf{B}^H(k)\mathbf{\Phi}_{\mathbf{ZZ}}(k,\ell)\mathbf{B}(k) \tag{8}$$

$$\begin{aligned}\phi_{YY}(k,\ell) = &\,[\mathbf{W}(k) - \mathbf{B}(k)\mathbf{H}(k,\ell)]^H\,\mathbf{\Phi}_{\mathbf{ZZ}}(k,\ell)\\ &\times[\mathbf{W}(k) - \mathbf{B}(k)\mathbf{H}(k,\ell)].\end{aligned} \tag{9}$$

Substituting (7) into (8) and (9), we have the following linear relation between the PSDs of the beamformer output, the reference signals, the desired source signal, and the input interferences:

$$\begin{bmatrix}\phi_{YY}(k,\ell)\\ \phi_{U_2U_2}(k,\ell)\\ \vdots\\ \phi_{U_MU_M}(k,\ell)\end{bmatrix} = \begin{bmatrix}C_{11}(k,\ell) & C_{12}(k,\ell) & C_{13}(k,\ell)\\ \vdots & \vdots & \vdots\\ C_{M1}(k,\ell) & C_{M2}(k,\ell) & C_{M3}(k,\ell)\end{bmatrix} \times\begin{bmatrix}\lambda_x(k,\ell)\\ \lambda_s(k,\ell)\\ \lambda_t(k,\ell)\end{bmatrix} \tag{10}$$

where

$$\begin{aligned}[C_{11} \quad C_{12} \quad C_{13}] &= [\mathbf{W} - \mathbf{BH}]^H[\mathbf{AA}^H \quad \mathbf{\Gamma}_s \quad \mathbf{\Gamma}_t]\\ &\quad\times(\mathbf{I}_3 \otimes [\mathbf{W} - \mathbf{BH}])\end{aligned} \tag{11}$$

$$[C_{21} \quad \cdots \quad C_{M1}] = \mathrm{diag}\{\mathbf{B}^H\mathbf{AA}^H\mathbf{B}\} \tag{12}$$

$$[C_{22} \quad \cdots \quad C_{M2}] = \mathrm{diag}\{\mathbf{B}^H\mathbf{\Gamma}_s\mathbf{B}\} \tag{13}$$

$$[C_{23} \quad \cdots \quad C_{M3}] = \mathrm{diag}\{\mathbf{B}^H\mathbf{\Gamma}_t\mathbf{B}\} \tag{14}$$

where $\mathbf{I}_3$ is a 3-by-3 identity matrix, $\otimes$ denotes Kronecker product, and $\mathrm{diag}\{\cdot\}$ represents a row vector constructed from the diagonal of a square matrix.

The beamformer is designed to maximize the ratio of the signal power to that of the interference plus noise, which is known as the *signal-to-interference-plus-noise ratio* (SINR). The blocking matrix performs a projection of the observed signals onto the $(M-1)$-dimensional subspace orthogonal to the look direction. Therefore, the desired signal component is expected to be significantly stronger at the beamformer output than at any reference noise signal, i.e., $C_{11}(k,\ell) \gg \max\{C_{i1}(k,\ell)|2 \leq i \leq M\}$. On the other hand, the pseudo-stationary interference is strongest at one of the reference signals since the sidelobe canceller $(\mathbf{H})$ adaptively minimizes its power at the beamformer output. Hence, $C_{12}(k,\ell) \leq \max\{C_{i2}(k,\ell)|2 \leq i \leq M\}$. Furthermore, the *transient beam-to-reference ratio* (TBRR), which is defined by the ratio between the transient power at beamformer output and the transient power at the reference signals, is expected to be lower in case of undesired transient components compared with that associated with the desired source components. Accordingly

$$\frac{C_{13}(k,\ell)}{\max\{C_{i3}(k,\ell)|2 \le i \le M\}} < \frac{C_{11}(k,\ell)}{\max\{C_{i1}(k,\ell)|2 \le i \le M\}}. \tag{15}$$

Our objective is to detect desired source components at the beamformer output and to differentiate them from the transient interfering components based on the TBRR.

III. Detection of Source Signals in Nonstationary Noise

In this section, we address the problem of signal detection in the time-frequency plane and discrimination between desired and undesired transient components. First, we detect transients at the beamformer output. Then, if there are no simultaneous transients at the reference signals, we determine that these transients are likely source components. In that case, a cautious enhancement would be involved. On the other hand, if a simultaneous transient at one of the reference signals is detected, then the TBRR would determine the extent to which such a transient is suppressed or preserved.

A. Detection of Transients at the Beamformer Output

Let $\mathcal{S}$ be a smoothing operator in the power spectral domain, and let $\mathcal{M}$ denote a single-channel estimator for the PSD of the background pseudo-stationary noise. For example, a causal $\mathcal{S}$ may be defined by recursively averaging past spectral power values of the noisy measurement:

$$\mathcal{S}Y(k,\ell) = \alpha_s \cdot \mathcal{S}Y(k,\ell-1) + (1-\alpha_s)\sum_{i=-w}^{w} b_i\,|Y(k-i,\ell)|^2 \tag{16}$$

where α_s $(0 \le \alpha_s \le 1)$ is a forgetting factor for the smoothing in time, and b is a normalized window function $(\sum_{i=-w}^{w} b_i = 1)$ that determines the order of smoothing in frequency. A useful estimator $\mathcal{M}$, particularly under low SNR and nonstationary noise conditions, can be obtained by the *minima controlled recursive averaging* approach [22], [23].

As with the Welch's spectrum estimation technique [29], the smoothing operator allows one to trade a reduction in spectral resolution for a reduction in variance. However, the retained resolution should be consistent with the spectral and temporal structure one wants to reveal. In the case of speech signals, a good compromise between smoothing the noise and tracking the speech signal is obtained with a time-frequency smoothing window of about 150 ms by 60 Hz [23]. A spectrogram corresponding to 32-ms frames and 75% overlap is therefore typically smoothed using a forgetting factor $\alpha_s = 0.9$ and a frequency window $b = [0.25\ \ 0.5\ \ 0.25]$.

For a given signal, we define its local nonstationarity (LNS) by the local ratio between the total and pseudo-stationary spectral power:

$$\Lambda(Y(k,\ell)) = \frac{\mathcal{S}Y(k,\ell)}{\mathcal{M}Y(k,\ell)}. \tag{17}$$

The LNS is a statistic of Y, fluctuating about one in the absence of transients, and expectedly well above one in the neighborhood of time-frequency bins that contain transients.

Let three hypotheses H_{0s}, H_{0t}, and H_1 indicate, respectively, absence of transients, presence of an interfering transient, and presence of a desired source transient at the beamformer output (the pseudo-stationary interference is present in any case). Let Λ_0 denote a threshold value of the LNS for the detection of transients at the beamformer output (i.e., accept $H_1 \cup H_{0t}$ if $\Lambda(Y) > \Lambda_0$ and accept H_{0s} otherwise). Then, the false alarm and detection probabilities are, respectively, defined by

$$\begin{aligned} P_{f,Y} &= \mathcal{P}(\Lambda(Y) > \Lambda_0 | H_{0s}) \\ &= \mathcal{P}(\mathcal{S}Y(k,\ell) > \Lambda_0 \cdot \mathcal{M}Y(k,\ell) | H_{0s}) \end{aligned} \tag{18}$$

$$\begin{aligned} P_{d,Y} &= \mathcal{P}(\Lambda(Y) > \Lambda_0 | H_1 \cup H_{0t}) \\ &= \mathcal{P}(\mathcal{S}Y(k,\ell) > \Lambda_0 \cdot \mathcal{M}Y(k,\ell) | H_1 \cup H_{0t}). \end{aligned} \tag{19}$$

Since $\mathcal{S}Y(k,\ell)$ is approximately chi-square distributed with μ degrees of freedom (see Appendix A)

$$F_{\mathcal{S}Y(k,\ell)}(x) \approx F_{\chi^2;\mu}\left(\frac{\mu x}{\phi_{YY}(k,\ell)}\right)$$

we have (see Appendix B) that for a specified false alarm probability $P_{f,Y}$, the required threshold value is

$$\Lambda_0 = \frac{1}{\mu} F_{\chi^2;\mu}^{-1}(1 - P_{f,Y}) \tag{20}$$

and the detection probability is

$$P_{d,Y} = 1 - F_{\chi^2;\mu}\left[\frac{1}{1+\xi_Y} F_{\chi^2;\mu}^{-1}(1 - P_{f,Y})\right] \tag{21}$$

where

$$\xi_Y \triangleq \frac{C_{11}\lambda_x + C_{13}\lambda_t}{C_{12}\lambda_s} \tag{22}$$

represents the ratio between the transient and pseudo-stationary power at the beamformer output. Fig. 2 shows the receiver operating characteristic (ROC) curve for detection of transients at the beamformer output, with the false alarm probability as parameter, and μ set to 32 [this value of μ is obtained for a smoothing $\mathcal{S}$ of the form (16), with $\alpha_s = 0.9$, and $b = [0.25\ \ 0.5\ \ 0.25]$]. Suppose that we require a false alarm probability no larger than $P_{f,Y} = 10^{-2}$, and suppose that transients at the beamformer output are defined by $\xi_Y \ge 2$. Then, the detection probability obtained using a detector $\Lambda(Y) > \Lambda_0 = 1.67$ is $P_{d,Y} = 0.98$.

B. Detection of Transients at the Reference Noise Signals

Given that a transient was detected at the beamformer output, its modification rule depends on the presence of a simultaneous transient at one of the reference signals. Let

$$\Lambda(\mathbf{U}(k,\ell)) = \max_{2 \le i \le M}\left\{\frac{\mathcal{S}U_i(k,\ell)}{\mathcal{M}U_i(k,\ell)}\right\} \tag{23}$$

denote the LNS of the reference signals, and let Λ_1 be a corresponding threshold value for detecting transients. Then, the

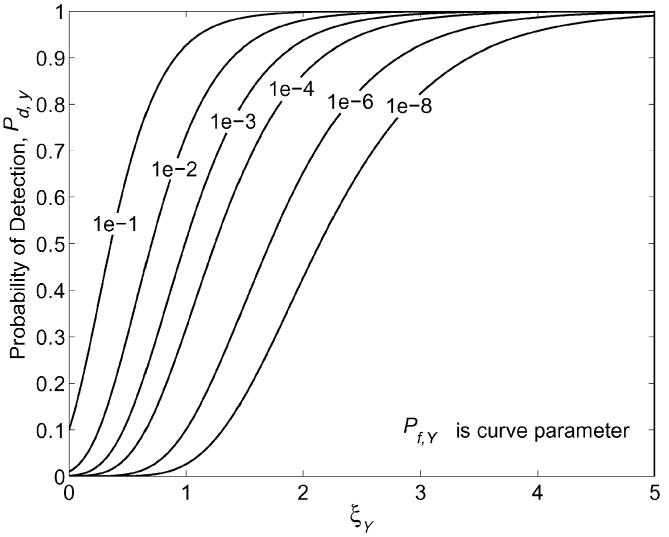

Fig. 2. Receiver operating characteristic curve for detection of transients at the beamformer output ($\mu = 32$).

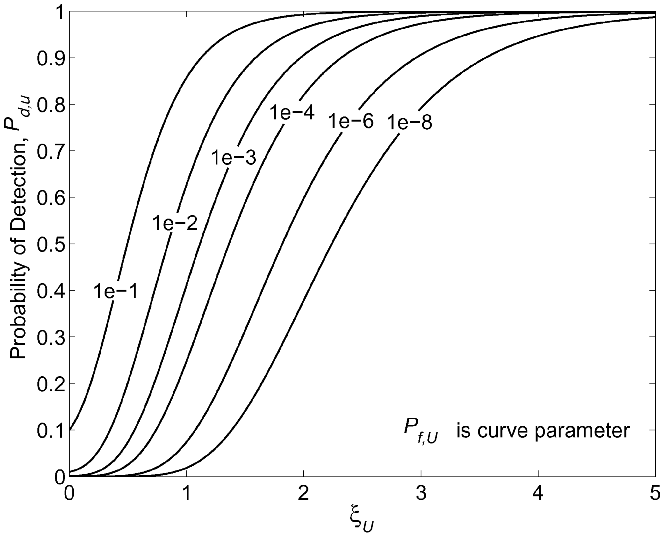

Fig. 3. Receiver operating characteristic curve for detection of transients at the reference noise signals, using $M = 4$ sensors ($\mu = 32$).

false alarm and detection probabilities are, respectively, defined by

$$P_{f,\mathbf{U}} = \mathcal{P}\left(\Lambda\left(\mathbf{U}(k,\ell)\right) > \Lambda_1 | H_{0s}\right) \quad (24)$$

$$P_{d,\mathbf{U}} = \mathcal{P}\left(\Lambda\left(\mathbf{U}(k,\ell)\right) > \Lambda_1 | H_1 \cup H_{0t}\right). \quad (25)$$

Assuming that $\{\mathcal{S}U_i(k,\ell)/\mathcal{M}U_i(k,\ell)\}_{i=2}^{M}$ are statistically independent, we obtain (see Appendix C) that for a specified false alarm probability $P_{f,\mathbf{U}}$, the required threshold value is

$$\Lambda_1 = \frac{1}{\mu} F^{-1}_{\chi^2;\mu}\left[(1-P_{f,\mathbf{U}})^{\frac{1}{M-1}}\right] \quad (26)$$

and the detection probability of a transient at one of the reference signals satisfies

$$\begin{aligned} P_{d,\mathbf{U}} &\approx 1 - \prod_{i=2}^{M} F_{\chi^2;\mu}\left(\frac{1}{1+\xi_{U_i}} F^{-1}_{\chi^2;\mu}\left[(1-P_{f,\mathbf{U}})^{\frac{1}{M-1}}\right]\right) \\ &\geq 1 - (1-P_{f,\mathbf{U}})^{\frac{M-2}{M-1}} \\ &\quad \cdot F_{\chi^2;\mu}\left(\frac{1}{1+\xi_{\mathbf{U}}} F^{-1}_{\chi^2;\mu}\left[(1-P_{f,\mathbf{U}})^{\frac{1}{M-1}}\right]\right) \end{aligned} \quad (27)$$

where $\xi_{U_i} \triangleq ((C_{i1}\lambda_x + C_{i3}\lambda_t)/C_{i2}\lambda_s)$ denotes the ratio of transient to pseudo-stationary power at the ith reference signal, and $\xi_{\mathbf{U}} \triangleq \max\{\xi_{U_i} | 2 \leq i \leq M\}$. Equality in (27) is derived when all ξ_{U_i} but one are identically zero. Fig. 3 shows the ROC curve for detection of transients at the reference noise signals, with the false alarm probability as a parameter. Four sensors are used, and μ is set to 32. Suppose that we require a false alarm probability no larger than $P_{f,\mathbf{U}} = 10^{-2}$, and suppose that transients at the reference outputs are defined by $\xi_{\mathbf{U}} \geq 2$. Then, the detection probability obtained using a detector $\Lambda(\mathbf{U}) > \Lambda_1 = 1.81$ is $P_{d,\mathbf{U}} = 0.96$.

C. TBRR

The TBRR is a useful statistic to determine the origin of a transient once it is detected simultaneously at the beamformer output and at one of the reference noise signals [30]. Since the operator $\mathcal{S}$ gives a measure of local spectral power, and $\mathcal{M}$ estimates the background pseudo-stationary power, then their difference yields a measure of the local transient power.[1] We define the TBRR by

$$\Omega(Y,\mathbf{U}) = \frac{\mathcal{S}Y - \mathcal{M}Y}{\max\limits_{2\leq i\leq M}\{\mathcal{S}U_i - \mathcal{M}U_i\}}. \quad (28)$$

Then, given that H_1 or H_{0t} is true, we have

$$\begin{aligned} \Omega(Y,\mathbf{U})|_{H_1\cup H_{0t}} &\approx \frac{\phi_{YY}(k,\ell) - C_{12}(k,\ell)\lambda_s(k,\ell)}{\max\limits_{2\leq i\leq M}\{\phi_{U_iU_i}(k,\ell) - C_{i2}(k,\ell)\lambda_s(k,\ell)\}} \\ &= \frac{C_{11}(k,\ell)\lambda_x(k,\ell) + C_{13}(k,\ell)\lambda_t(k,\ell)}{\max\limits_{2\leq i\leq M}\{C_{i1}(k)\lambda_x(k,\ell) + C_{i3}(k,\ell)\lambda_t(k,\ell)\}}. \end{aligned} \quad (29)$$

Transient signal components are relatively strong at the beamformer output, whereas transient noise components are relatively strong at one of the reference signals. Hence, we expect $\Omega(Y,\mathbf{U})$ to be large for signal transients and small for noise transients. Let Ω_0 denote a threshold value of the TBRR for the decision between signal and noise (i.e., accept H_1 only if $\Omega(Y,\mathbf{U}) > \Omega_0$), where the false alarm and detection probabilities are defined by

$$P_{f,\Omega} = \mathcal{P}\left\{\Omega(Y,\mathbf{U}) > \Omega_0 | H_{0t}\right\} \quad (30)$$

$$P_{d,\Omega} = \mathcal{P}\left\{\Omega(Y,\mathbf{U}) > \Omega_0 | H_1\right\}. \quad (31)$$

Then, by (15), we can choose a threshold $\Omega_0(k,\ell)$ such that

$$\begin{aligned} \Omega(Y,\mathbf{U})|_{H_{0t}} &\approx \frac{C_{13}(k,\ell)}{\max\limits_{2\leq i\leq M}\{C_{i3}(k,\ell)\}} < \Omega_0(k,\ell) \\ &< \frac{C_{11}(k,\ell)}{\max\limits_{2\leq i\leq M}\{C_{i1}(k)\}} \approx \Omega(Y,\mathbf{U})|_{H_1} \end{aligned} \quad (32)$$

which implies $P_{f,\Omega} \to 0$ and $P_{d,\Omega} \to 1$.

[1] Recall that transient components are assumed to be uncorrelated with pseudo-stationary noise components.

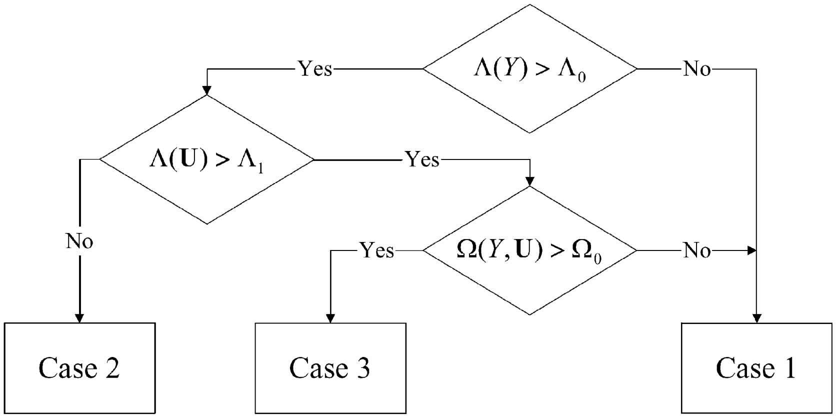

Fig. 4. Block diagram for detection of desired source components at the beamformer output.

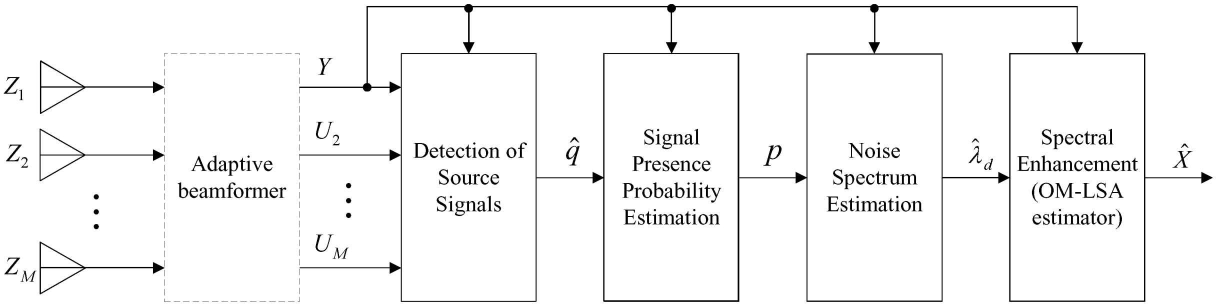
Z1
Z2
ZM
Adaptive beamformer
Y
U2
UM
Detection of Source Signals
q̂
Signal Presence Probability Estimation
p
Noise Spectrum Estimation
λ̂d
Spectral Enhancement (OM-LSA estimator)
X̂

Fig. 5. Block diagram of the multichannel post-filtering.

The ratio

$$Q \triangleq \frac{C_{11}}{C_{13}} \cdot \frac{\max\limits_{2\le i\le M}\{C_{i3}\}}{\max\limits_{2\le i\le M}\{C_{i1}\}} \tag{33}$$

defines the *transient discrimination quality* (TDQ) of the beamformer. It follows that discrimination between transient noise and desired signal components is possible when $Q \geq 1$. However, in practice, we obtained good performance $P_{f,\Omega} \rightarrow 0$, $P_{d,\Omega} \rightarrow 1$ when $Q \geq 3$.

Fig. 4 summarizes a block diagram for the detection of desired source components at the beamformer output. The detection is carried out in the time-frequency plane for each frame and frequency bin. Case 1 is reached when no transients have been detected at the beamformer output or when the TBRR is lower than the threshold Ω_0. In this case, presumably no desirable transients are present at the beamformer output, and consequently, strong noise suppression is applicable. Considering Case 2, a transient has been detected at the beamformer output but not at any reference signal. This case indicates that the transient is likely a desirable source component, and a cautious noise suppression would therefore be involved. Finally, Case 3 is determined when transients are simultaneously detected at the beamformer output and at a reference signal, and conjunctionally, the value of the TBRR is above Ω_0. In this case, the larger the TBRR is, the higher the likelihood that a transient originates from a desired source.

IV. Multichannel Post-Filtering

In this section, we address the problem of estimating the time-varying spectrum of the beamformer output noise and present the multichannel post-filtering approach. Fig. 5 describes the block diagram of the proposed multichannel post-filtering. Desired source components are detected at the beamformer output, and an estimate $\hat{q}(k,\ell)$ for the *a priori* signal absence probability is produced. Based on a Gaussian statistical model [20] and a decision-directed estimator for the *a priori* SNR under signal presence uncertainty [22], we derive an estimator $p(k,\ell) \triangleq \mathcal{P}(H_1|Y,\mathbf{U})$ for the signal presence probability. This estimator controls the components that are introduced as noise into the PSD estimator. Finally, spectral enhancement of the beamformer output is achieved by applying an *optimally modified log-spectral amplitude* (OM-LSA) gain function [22]. This gain minimizes the mean-square error of the log-spectra under signal presence uncertainty.

Referring to Fig. 4, Cases 1 and 2 imply presumable signal absence and presence, respectively. Therefore, we set $\hat{q}(k,\ell)$ to 1 in Case 1 and to 0 in Case 2. However, when transients are simultaneously detected in both the beamformer output and one of the reference signals, and the TBRR is larger than Ω_0 (Case 3), then the *a priori* signal absence probability decreases as the TBRR increases. For simplicity, we assume that the *a priori* signal absence probability linearly decreases in the region $\Omega(Y,\mathbf{U}) \in [\Omega_0, 3\Omega_0]$. That is

$$\hat{q}_\Omega(k,\ell) = \begin{cases} 1, & \text{if } \Omega(Y,\mathbf{U}) \leq \Omega_0 \\ \frac{3\Omega_0 - \Omega(Y,\mathbf{U})}{2\Omega_0}, & \text{if } \Omega_0 < \Omega(Y,\mathbf{U}) < 3\Omega_0 \\ 0, & \text{otherwise.} \end{cases} \tag{34}$$

On the other hand, since the TBRR is based on smoothed spectra, we can further improve the noise reduction by evaluating the *a posteriori* SNR at the beamformer output with respect to the pseudo-stationary noise $\gamma_s(k,\ell) \triangleq |Y(k,\ell)|^2/\mathcal{M}Y(k,\ell)$ [23]. Specifically, for $\Omega(Y,\mathbf{U}) \geq 3\Omega_0$, the *a priori* signal absence probability is determined according to

$$\hat{q}_{\gamma_s}(k,\ell) = \begin{cases} 1, & \text{if } \gamma_s(k,\ell) \leq 1 \\ \frac{\gamma_0 - \gamma_s(k,\ell)}{\gamma_0 - 1}, & \text{if } 1 < \gamma_s(k,\ell) < \gamma_0 \\ 0, & \text{otherwise} \end{cases} \tag{35}$$

where γ_0 denotes a constant satisfying

$$\mathcal{P}\left(\gamma_s(k,\ell) \geq \gamma_0 | H_{0s}\right) < \epsilon \tag{36}$$

for a certain significance level ϵ (typically, we use $\epsilon = 0.01$ and $\gamma_0 = -\log(\epsilon) = 4.6$) [23]. Indeed, from (36), we have that when the *a posteriori* SNR is larger than γ_0, either H_1 or H_{0t} is true (H_{0s} is very unlikely). On the other hand, $\Omega(Y,\mathbf{U})$ discriminates between desired source components (H_1) and noise transients (H_{0t}). Therefore, combining the conditions on γ_s and $\Omega(Y,\mathbf{U})$, and assuming smooth bilinear transition from signal absence to presence in the regions $\gamma_s \in [E\{\gamma_s|H_{0s}\}, \gamma_0] = [1,\gamma_0]$ and $\Omega(Y,\mathbf{U}) \in [\Omega_0, 3\Omega_0]$, the *a priori* signal absence probability is given by

$$\hat{q}(k,\ell) = \begin{cases} 1, & \text{if } \gamma_s(k,\ell) \leq 1 \\ & \quad \text{or } \Omega(Y,\mathbf{U}) \leq \Omega_0 \\ 0, & \text{if } \gamma_s(k,\ell) \geq \gamma_0 \\ & \quad \text{and } \Omega(Y,\mathbf{U}) \geq 3\Omega_0 \\ \max\left\{ \frac{\gamma_0 - \gamma_s(k,\ell)}{\gamma_0 - 1}, \right. & \\ \qquad \left. \frac{3\Omega_0 - \Omega(Y,\mathbf{U})}{2\Omega_0} \right\}, & \text{otherwise.} \end{cases} \tag{37}$$

Under the assumed statistical model, the signal presence probability for $\hat{q}(k,\ell) < 1$ is obtained by [20]

$$p(k,\ell) = \left[1 + \frac{\hat{q}(k,\ell)}{1 - \hat{q}(k,\ell)} \left(1 + \xi(k,\ell)\right) \exp\left(-v(k,\ell)\right) \right]^{-1} \tag{38}$$

where $\xi(k,\ell) \triangleq E\{|X(k,\ell)|^2\}/\lambda_d(k,\ell)$ is the *a priori* SNR, $\lambda_d(k,\ell)$ is the noise PSD at the beamformer output, $v(k,\ell) \triangleq \gamma(k,\ell)\xi(k,\ell)/(1+\xi(k,\ell))$, and $\gamma(k,\ell) \triangleq |Y(k,\ell)|^2/\lambda_d(k,\ell)$ is the *a posteriori* SNR. In case of $\hat{q}(k,\ell) = 1$, the signal presence probability $p(k,\ell)$ reduces to 0.

To evaluate (38), we need to estimate the *a priori* SNR $\xi(k,\ell)$ and the noise PSD at the beamformer output $\lambda_d(k,\ell)$. The *a priori* SNR is estimated by [22]

$$\hat{\xi}(k,\ell) = \alpha G_{H_1}^2(k,\ell-1)\gamma(k,\ell-1) + (1-\alpha)\max\{\gamma(k,\ell)-1, 0\} \tag{39}$$

where α is a weighting factor that controls the tradeoff between noise reduction and signal distortion, and

$$G_{H_1}(k,\ell) \triangleq \frac{\xi(k,\ell)}{1+\xi(k,\ell)} \exp\left(\frac{1}{2} \int_{v(k,\ell)}^{\infty} \frac{e^{-t}}{t} dt \right) \tag{40}$$

is the spectral gain function of the *log-spectral amplitude* (LSA) estimator when signal is surely present[2] [21]. The noise PSD at the beamformer output is estimated by the MCRA approach [23]. That is, past spectral power values of the noisy measurement are recursively averaged using a time-varying frequency-dependent smoothing parameter

$$\hat{\lambda}_d(k,\ell+1) = \tilde{\alpha}_d(k,\ell)\hat{\lambda}_d(k,\ell) + \beta \cdot [1 - \tilde{\alpha}_d(k,\ell)] \, |Y(k,\ell)|^2 \tag{41}$$

where $\tilde{\alpha}_d(k,\ell)$ is the smoothing parameter $(0 < \tilde{\alpha}_d(k,\ell) < 1)$, and β $(\beta \geq 1)$ is a factor that compensates the bias when the signal is absent. The smoothing parameter is determined by the signal presence probability $p(k,\ell)$ and a constant α_d $(0 < \alpha_d < 1)$ that represents its minimal value

$$\tilde{\alpha}_d(k,\ell) \triangleq \alpha_d + (1-\alpha_d)\, p(k,\ell). \tag{42}$$

When a signal is present, $\tilde{\alpha}_d$ is close to 1, thus preventing the noise estimate from increasing as a result of signal components. As the probability of signal presence decreases, the smoothing parameter gets smaller, facilitating a faster update of the noise estimate. The value of α_d compromises between the tracking rate (response rate to abrupt changes in the noise statistics) and the variance of the noise estimate. Typically, in case of high levels of nonstationary noise, a good compromise is obtained with $\alpha_d = 0.85$ [23].

The final step of the multichannel post-filtering is spectral enhancement of the beamformer output by applying the OM-LSA gain function. Specifically, the clean signal STFT is estimated by

$$\hat{X}(k,\ell) = G(k,\ell)Y(k,\ell) \tag{43}$$

where

$$G(k,\ell) = \{G_{H_1}(k,\ell)\}^{p(k,\ell)} \cdot G_{min}^{1-p(k,\ell)} \tag{44}$$

is the OM-LSA gain function, and G_{min} denotes a lower bound constraint for the gain when signal is absent. The implementation of the multichannel post-filtering algorithm is summarized in Fig. 6. Typical values of the respective parameters, for a sampling rate of 8 kHz, are given in Table I.

V. Experimental Results

To validate the usefulness of the proposed post-filtering approach under nonstationary noise conditions, we compare its performance to a single-channel post-filtering in various car environments. Specifically, multichannel speech signals are degraded by interfering speakers and various car noise types. Then, beamforming is applied to the noisy signals, followed by either single-channel or multichannel post-filtering. The performance evaluation includes objective quality measures, as well as a subjective study of speech spectrograms and informal listening tests.

A linear array consisting of four microphones with 5-cm spacing is mounted in a car on the visor. Clean speech signals are recorded at a sampling rate of 8 kHz in the absence of background noise (standing car, silent environment). An interfering speaker and car noise signals are recorded while the car is moving at about 60 km/h, and windows are either closed or the window next to the driver is slightly open (about 5 cm). The input microphone signals are generated by mixing

[2] The advantage of $\hat{\xi}(k,\ell)$ over the "decision-directed" estimator of Ephraim and Malah [20], particularly for weak signal components and low input SNR, is discussed in [22].

```
Initialize variables at the first frame for all frequency bins $k$:

    $\mathcal{S}Y(k,0) = \mathcal{M}Y(k,0) = \hat{\lambda}_d(k,0) = |Y(k,0)|^2; \quad G_{H_1}(k,0) = \gamma(k,0) = 1.$

For all time frames $\ell$

    For all frequency bins $k$

        Compute the recursively averaged spectrum of the beamformer output $\mathcal{S}Y(k,\ell)$ using (16), and update the MCRA estimate of the background pseudo-stationary noise $\mathcal{M}Y(k,\ell)$ using [23].

        Compute the local non-stationarities of the beamformer output and reference signals, $\Lambda(Y)$ and $\Lambda(\mathbf{U})$, using (17) and (23), and compute the transient beam-to-reference ratio, $\Omega(Y,\mathbf{U})$, using (28).

        Using the block diagram in Fig. 4, determine which case applies to each frequency bin; Set the *a priori* signal absence probability $\hat{q}(k,\ell)$ to 1 in Case 1, and to 0 in Case 2, and compute its value according to (37) in Case 3.

        Compute the *a priori* SNR $\hat{\xi}(k,\ell)$ using (39), the conditional gain $G_{H_1}(k,\ell)$ using (40), and the signal presence probability $p(k,\ell)$ using (38).

        Compute the time-varying smoothing parameter $\tilde{\alpha}_d(k,\ell)$ using (42), and update the noise spectrum estimate $\hat{\lambda}_d(k,\ell+1)$ using (41).

        Compute the OM-LSA estimate of the clean signal, $\hat{X}(k,\ell)$, using (43) and (44).
```

Fig. 6. Multichannel post-filtering algorithm.

TABLE I
VALUES OF PARAMETERS USED IN THE IMPLEMENTATION OF THE PROPOSED MULTICHANNEL POST-FILTERING, FOR A SAMPLING RATE OF 8 kHz

$\Lambda_0 = 1.67$	$\Lambda_1 = 1.81$	$\Omega_0 = 1$	$\gamma_0 = 4.6$
$\alpha = 0.92$	$\alpha_s = 0.9$	$\alpha_d = 0.85$	$\beta = 1.47$
$b = \begin{bmatrix} 0.25 & 0.5 & 0.25 \end{bmatrix}$		$\mu = 32$	$G_{min} = -20$ dB

the speech and noise signals at various SNR levels in the range [−5, 10] dB.

An adaptive beamformer (specifically, the TF-GSC, proposed by Gannot *et al.* [24]) is applied to the noisy multichannel signals. The beamformer output is enhanced using the OM-LSA estimator [22] and is referred to as the single-channel post-filtering output. Alternatively, the beamformer output, which is enhanced using the procedure described in the previous section, is referred to as the multichannel post-filtering output. Three different objective quality measures are used in our evaluation. The first is segmental SNR, in decibels, defined by [31]

$$\begin{aligned}\text{SegSNR} &= \frac{1}{L}\sum_{\ell=0}^{L-1}\text{SNR}_\ell \\ &= \frac{10}{L}\sum_{\ell=0}^{L-1} \log_{10}\frac{\sum_{n=0}^{N-1} x^2\left(n+\frac{\ell N}{2}\right)}{\sum_{n=0}^{N-1}\left[x\left(n+\frac{\ell N}{2}\right)-\hat{x}\left(n+\frac{\ell N}{2}\right)\right]^2}\end{aligned} \tag{45}$$

where L represents the number of frames in the signal, and $N = 256$ is the number of samples per frame (corresponding to 32-ms frames and 50% overlap). The SNR at each frame SNR_ℓ is limited to perceptually meaningful range between 35 and -10 dB. This prevents the segmental SNR measure from being biased in either a positive or negative direction due to a few silence or unusually high SNR frames that do not contribute significantly to the overall speech quality [32], [33]. This measure takes into account both residual noise and speech distortion. The second quality measure is noise reduction (NR), in decibels, which is defined by

$$\text{NR} = \frac{10}{|\mathcal{L}'|}\sum_{\ell\in\mathcal{L}'}\log_{10}\frac{\sum_{n=0}^{N-1} z_1^2\left(n+\frac{\ell N}{2}\right)}{\sum_{n=0}^{N-1}\hat{x}^2\left(n+\frac{\ell N}{2}\right)} \tag{46}$$

where $\mathcal{L}'$ represents the set of frames that contain only noise, and $|\mathcal{L}'|$ its cardinality. The NR measure compares the noise level in the enhanced signal to the noise level recorded by the first microphone. The third quality measure is log-spectral distance (LSD), in decibels, which is defined by

$$\begin{aligned}\text{LSD} = \frac{10}{L}\sum_{\ell=0}^{L-1}\Bigg\{\frac{1}{\frac{N}{2}+1}\sum_{k=0}^{\frac{N}{2}}\Big[&\log_{10}\mathcal{A}X(k,\ell) \\ &-\log_{10}\mathcal{A}\hat{X}(k,\ell)\Big]^2\Bigg\}^{\frac{1}{2}}\end{aligned} \tag{47}$$

where $\mathcal{A}X(k,\ell) \triangleq \max\{|X(k,\ell)|^2,\delta\}$ is the spectral power, clipped such that the log-spectrum dynamic range is confined to about 50 dB (that is, $\delta = 10^{-50/10}\max_{k,\ell}\{|X(k,\ell)|^2\}$).

Fig. 7 shows experimental results of the average segmental SNR obtained for various noise types and at various noise levels. The segmental SNR is evaluated at the first microphone, the beamformer output, and the post-filtering outputs. A theoretical limit post-filtering, which is achievable by calculating the noise spectrum from the noise itself, is also considered. Results of the NR and LSD measures are presented in Figs. 8 and 9, respectively. It can be readily seen that beamforming alone does not provide sufficient noise reduction in a car environment, owing to its limited ability to reduce diffuse noise [24]. Furthermore, multichannel post-filtering is consistently better than single-channel post-filtering under all noise conditions. The improvement in

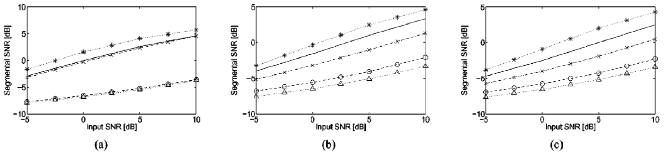

Fig. 7. Average segmental SNR at (△) microphone #1, (○) beamformer output, (×) single-channel post-filtering output, (solid line) multichannel post-filtering output, and (∗) theoretical limit post-filtering output for various car noise conditions. (a) Closed windows. (b) Open window. (c) Interfering speaker.

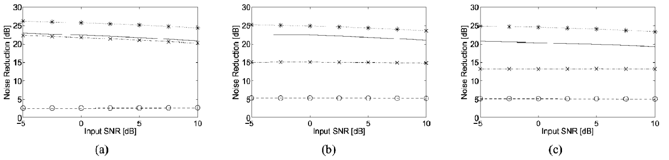

Fig. 8. Average noise reduction at (○) beamformer output, (×) single-channel post-filtering output, (solid line) multichannel post-filtering output, and (∗) theoretical limit post-filtering output for various car noise conditions. (a) Closed windows. (b) Open window. (c) Interfering speaker.

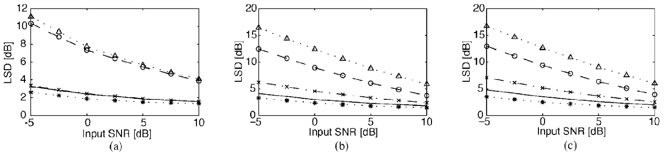

Fig. 9. Average log-spectral distance at (△) microphone #1, (○) beamformer output, (×) single-channel post-filtering output, (solid line) multichannel post-filtering output, and (∗) theoretical limit post-filtering output for various car noise conditions. (a) Closed windows. (b) Open window. (c) Interfering speaker.

performance of the former over the latter is expectedly high in nonstationary noise environments (specifically, open windows or interfering speaker), but is insignificant otherwise, since multichannel post-filtering reduces to single-channel in pseudo-stationary noise environments.

A subjective comparison between multichannel and single-channel post-filtering was conducted using speech spectrograms and validated by informal listening tests. Typical examples of speech spectrograms are presented in Fig. 10 for the case of nonstationary noise (interfering speaker, open window) at SNR $= -0.9$ dB. The beamformer output [see Fig. 10(c)] is clearly characterized by a high level of noise. Its enhancement using single-channel post-filtering well suppresses the pseudo-stationary noise but adversely retains the transient noise components. By contrast, the enhancement using multichannel post-filtering results in superior noise attenuation while preserving the desired source components.

Fig. 11 shows traces of the improvement in segmental SNR and LSD measures, gained by the multichannel post-filtering and theoretical limit, in comparison with a single-channel post-filtering. The traces are averaged out over a period of about 400 ms (25 frames of 32 ms each, with 50% overlap). The noise PSD at the beamformer output varies substantially due to the residual interfering components of speech, the blowing wind, and passing cars. The improvement in performance over the single-channel post-filtering is obtained when the noise spectrum fluctuates. In some instances, the increase in segmental SNR surpasses as much as 8 dB, and the decrease in LSD is greater than 6 dB. Clearly, a single-channel post-filter is inefficient at attenuating highly nonstationary noise components since it lacks the ability to differentiate such components from the speech components. On the other hand, the proposed multichannel post-filtering approach achieves a significantly reduced level of background noise, whether stationary or not, without further distorting speech components. This is verified by subjective informal listening tests.

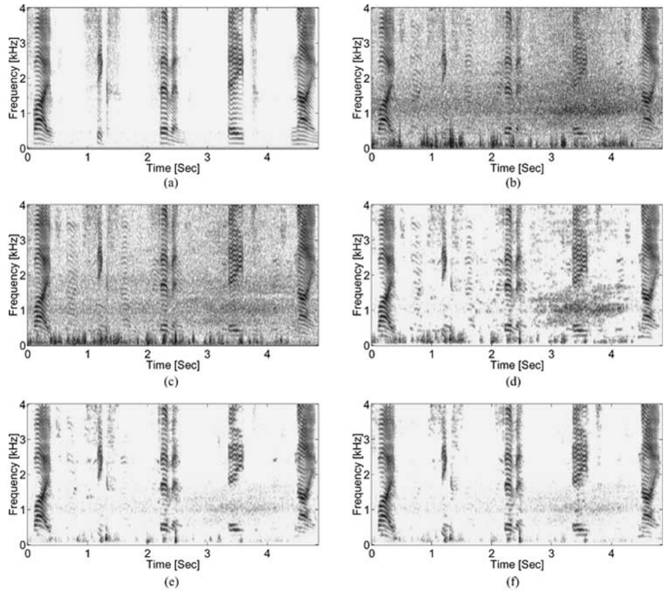

Fig. 10. Speech spectrograms. (a) Original clean speech signal at microphone #1: "five six seven eight nine." (b) Noisy signal at microphone #1 (car noise, open window, interfering speaker. SNR $= -0.9$ dB, SegSNR $= -6.2$ dB, LSD $= 15.4$ dB). (c) Beamformer output (SegSNR $= -5.3$ dB, NR $= 5.2$ dB, LSD $= 12.2$ dB). (d) Single-channel post-filtering output (SegSNR $= -3.8$ dB, NR $= 12.1$ dB, LSD $= 7.4$ dB). (e) Multichannel post-filtering output (SegSNR $= -1.3$ dB, NR $= 23.2$ dB, LSD $= 4.6$ dB). (f) Theoretical limit (SegSNR $= -0.4$ dB, NR $= 24.0$ dB, LSD $= 4.0$ dB).

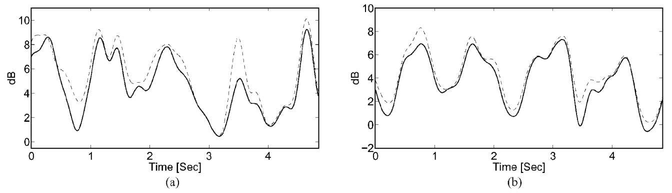

Fig. 11. Trace of the improvement over a single-channel post-filtering gained by the proposed multichannel post-filtering (solid) and theoretical limit (dashed). (a) Increase in segmental SNR. (b) Decrease in log-spectral distance.

VI. Conclusion

We have described a multichannel post-filtering approach for arbitrary beamformers that is particularly advantageous in nonstationary noise environments. The beamformer is realistically assumed to have a steering error, a blocking matrix that is unable to block all of the desired signal components, and a noise canceller that is adapted to the pseudo-stationary noise but not modified during transient interferences. Accordingly, the reference noise signals may include some desired signal components. Furthermore, transient noise components that leak through the sidelobes of the fixed beamformer may proceed to the beamformer primary output. A mild assumption is made with regard to the beamformer that a desired signal component is stronger at the beamformer output than at any reference noise signal, and a noise component is strongest at one of the reference signals. Consequently, transients are detected at the beamformer output and either suppressed or preserved based on the transient beam-to-reference ratio.

We derived an estimator for the signal presence probability that controls the rate of recursive averaging for obtaining a noise spectrum estimate. It also modifies the spectral gain function to obtain an estimate for the clean signal spectral amplitude. The proposed method was tested in various nonstationary car noise environments, and its performance was compared with a single-channel post-filtering approach. We showed that multichannel post-filtering is better than single-channel post-filtering, particularly under highly nonstationary noise conditions (such as noise resulting from blowing wind, passing cars, interfering speakers, etc.). While transient noise components are indistinguishable from desired source components if using state-of-the-art single-channel post-filtering, the enhancement of the beamformer output by multichannel post-filtering produces a significantly reduced level of residual transient noise without further distorting the desired signal components.

APPENDIX A
STATISTICS OF $\mathcal{S}Y(k,\ell)$

Successive spectral power values of the beamformer output $|Y(k,\ell)|^2$ are generally correlated, and there is no closed-form solution for the probability density function of $\mathcal{S}Y(k,\ell)$. However, (16) can be written as

$$\mathcal{S}Y(k,\ell) = (1-\alpha_s)\sum_{i=-w}^{w}\sum_{j=0}^{\infty} b_i \alpha_s^j \, |Y(k-i,\ell-j)|^2 . \quad (48)$$

Approximating $\mathcal{S}Y(k,\ell)$ as the sum of μ squared mutually independent normal variables [23], [34], its distribution function is given by

$$F_{\mathcal{S}Y(k,\ell)}(x) \approx F_{\chi^2;\mu}\left(\frac{\mu x}{\phi_{YY}(k,\ell)}\right) \quad (49)$$

where $F_{\chi^2;\mu}(x)$ denotes the standard chi-square distribution function, with μ degrees of freedom. Specifically, $F_{\mathcal{S}Y(k,\ell)}(x) = \Gamma((\mu/2),(\mu x/2\phi_{YY}(k,\ell)))u(x)/\Gamma(\mu/2)$, where $\Gamma(a) \triangleq \int_0^\infty t^{a-1}e^{-t}dt$ is the gamma function, $\Gamma(a,x) \triangleq \int_0^x t^{a-1}e^{-t}dt$ is the incomplete gamma function, and $u(x)$ is the unit step function (i.e., $u(x)=1$ for $x \geq 0$ and $u(x)=0$ otherwise). The equivalent degrees of freedom μ is determined by the smoothing parameter α_s, the window function b, and the spectral analysis parameters of the STFT (size and shape of the analysis window, and frame-update step). The value of μ can be estimated by generating a stationary white Gaussian noise $d(t)$, transforming it to the time-frequency domain, and substituting the sample mean and variance (over the entire time-frequency plane) into the expression $\hat{\mu} \approx 2E^2\{\mathcal{S}D(k,\ell)\}/\mathrm{var}\{\mathcal{S}D(k,\ell)\}$.

APPENDIX B
DETECTION OF TRANSIENTS AT THE BEAMFORMER OUTPUT

Substituting (49) into (18) and (19), we have

$$P_{f,Y} \approx 1 - F_{\chi^2;\mu}\left(\frac{\mu\Lambda_0 \cdot \mathcal{M}Y(k,\ell)}{\phi_{YY}(k,\ell)}\right)\Bigg|_{H_{0s}} \quad (50)$$

$$P_{d,Y} \approx 1 - F_{\chi^2;\mu}\left(\frac{\mu\Lambda_0 \cdot \mathcal{M}Y(k,\ell)}{\phi_{YY}(k,\ell)}\right)\Bigg|_{H_1 \cup H_{0t}} . \quad (51)$$

Equation (10) implies $\phi_{YY}|_{H_{0s}} = C_{12}\lambda_s$ and $\phi_{YY}|_{H_1 \cup H_{0t}} = C_{11}\lambda_x + C_{12}\lambda_s + C_{13}\lambda_t$. Then, by using the approximation $\mathcal{M}Y \approx \phi_{YY}|_{H_{0s}}$ (recall that $\mathcal{M}$ in an estimator for the PSD of the pseudo-stationary noise), we obtain

$$P_{f,Y} \approx 1 - F_{\chi^2;\mu}(\mu\Lambda_0) \quad (52)$$

$$P_{d,Y} \approx 1 - F_{\chi^2;\mu}\left(\frac{\mu\Lambda_0 C_{12}\lambda_s}{C_{11}\lambda_x + C_{12}\lambda_s + C_{13}\lambda_t}\right). \quad (53)$$

Consequently, the required threshold value for a specified $P_{f,Y}$ is

$$\Lambda_0 = \frac{1}{\mu}F^{-1}_{\chi^2;\mu}(1-P_{f,Y}). \quad (54)$$

Substituting this expression into (53), we have

$$P_{d,Y} = 1 - F_{\chi^2;\mu}\left[\frac{1}{1+\xi_Y}F^{-1}_{\chi^2;\mu}(1-P_{f,Y})\right] \quad (55)$$

where

$$\xi_Y \triangleq \frac{C_{11}\lambda_x + C_{13}\lambda_t}{C_{12}\lambda_s} \quad (56)$$

represents the ratio between the transient and pseudo-stationary power at the beamformer output.

APPENDIX C
DETECTION OF TRANSIENTS AT THE REFERENCE NOISE SIGNALS

Substituting (23) into (24) and (25), the false alarm and detection probabilities are, respectively, given by

$$P_{f,\mathbf{U}} = \mathcal{P}\left(\max_{2\leq i\leq M}\left\{\frac{\mathcal{S}U_i(k,\ell)}{\mathcal{M}U_i(k,\ell)}\right\} > \Lambda_1 | H_{0s}\right) \quad (57)$$

$$P_{d,\mathbf{U}} = \mathcal{P}\left(\max_{2\leq i\leq M}\left\{\frac{\mathcal{S}U_i(k,\ell)}{\mathcal{M}U_i(k,\ell)}\right\} > \Lambda_1 | H_1 \cup H_{0t}\right). \quad (58)$$

Using (49) and assuming that $\{\mathcal{S}U_i(k,\ell)/\mathcal{M}U_i(k,\ell)\}_{i=2}^M$ are statistically independent, we have

$$P_{f,\mathbf{U}} \approx 1 - \prod_{i=2}^{M} F_{\chi^2;\mu}\left(\frac{\mu\Lambda_1 \cdot \mathcal{M}U_i(k,\ell)}{\phi_{U_iU_i}(k,\ell)}\right)\Bigg|_{H_{0s}} \quad (59)$$

$$P_{d,\mathbf{U}} \approx 1 - \prod_{i=2}^{M} F_{\chi^2;\mu}\left(\frac{\mu\Lambda_1 \cdot \mathcal{M}U_i(k,\ell)}{\phi_{U_iU_i}(k,\ell)}\right)\Bigg|_{H_1 \cup H_{0t}} . \quad (60)$$

Equation (10) yields $\phi_{U_iU_i}|_{H_{0s}} = C_{i2}\lambda_s$ and $\phi_{U_iU_i}|_{H_1\cup H_{0t}} = C_{i1}\lambda_x + C_{i2}\lambda_s + C_{i3}\lambda_t$. Then, by using the approximation $\mathcal{M}U_i \approx \phi_{U_iU_i}|_{H_{0s}}$, we obtain

$$P_{f,\mathbf{U}} \approx 1 - F^{M-1}_{\chi^2;\mu}(\mu\,\Lambda_1) \tag{61}$$

$$P_{d,\mathbf{U}} \approx 1 - \prod_{i=2}^{M} F_{\chi^2;\mu}\left(\frac{\mu\Lambda_1 \cdot C_{i2}\lambda_s}{C_{i1}\lambda_x + C_{i2}\lambda_s + C_{i3}\lambda_t}\right). \tag{62}$$

Thus, for a specified false alarm probability $P_{f,\mathbf{U}}$, the threshold value is

$$\Lambda_1 = \frac{1}{\mu}F^{-1}_{\chi^2;\mu}\left[(1-P_{f,\mathbf{U}})^{\frac{1}{M-1}}\right]. \tag{63}$$

Substituting this expression into (62) and denoting by $\xi_{U_i} = (C_{i1}\lambda_x + C_{i3}\lambda_t)/C_{i2}\lambda_s$ the ratio of transient to pseudo-stationary power at the {i}th reference signal, we have

$$P_{d,\mathbf{U}} \approx 1 - \prod_{i=2}^{M} F_{\chi^2;\mu}\left(\frac{1}{1+\xi_{U_i}}F^{-1}_{\chi^2;\mu}\left[(1-P_{f,\mathbf{U}})^{\frac{1}{M-1}}\right]\right). \tag{64}$$

Since $F_{\chi^2;\mu}(x)$ is a monotone increasing distribution function, and $\xi_{U_i} \geq 0$, it follows that

$$F_{\chi^2;\mu}\left(\frac{1}{1+\xi_{U_i}}F^{-1}_{\chi^2;\mu}\left[(1-P_{f,\mathbf{U}})^{\frac{1}{M-1}}\right]\right) \leq (1-P_{f,\mathbf{U}})^{\frac{1}{M-1}}$$

for all $i \in \{2,\ldots,M\}$. In particular, applying this inequality to all indices $i \in \{2,\ldots,M\}$ besides the index (or one of the indices) that maximizes $\{\xi_{U_k}\}_{k=2}^{M}$ gives

$$\begin{aligned}P_{d,\mathbf{U}} \geq\ & 1 - (1-P_{f,\mathbf{U}})^{\frac{M-2}{M-1}} \\ & \cdot F_{\chi^2;\mu}\left(\frac{1}{1+\xi_{\mathbf{U}}}F^{-1}_{\chi^2;\mu}\left[(1-P_{f,\mathbf{U}})^{\frac{1}{M-1}}\right]\right)\end{aligned} \tag{65}$$

where $\xi_{\mathbf{U}} \triangleq \max\{\xi_{U_i}|2 \leq i \leq M\}$.

Acknowledgment

The author thanks Dr. B. Berdugo for helpful discussions, Dr. S. Gannot for making his adaptive beamforming code (TF-GSC) available, and the anonymous reviewers for their helpful comments.

References

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

[2] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with post-filtering," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 240–259, May 1998.

[3] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Analysis: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001, ch. 3, pp. 39–60.

[4] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. 13th IEEE Int. Conf. Acoust. Speech Signal Process.*, New York, Apr. 11–14, 1988, pp. 2578–2581.

[5] ——, "Noise reduction based on microphone array with LMS adaptive post-filtering," *Electron. Lett.*, vol. 26, no. 24, pp. 2036–2037, Nov. 1990.

[6] K. U. Simmer and A. Wasiljeff, "Adaptive microphone arrays for noise suppression in the frequency domain," in *Proc. 2nd Cost-229 Workshop Adaptive Algorithms Commun.*, Bordeaux, France, Sept. 30–Oct. 2 1992 [Online] Available: http://www.ant.uni-bremen.de/pub/speech, pp. 185–194.

[7] S. Fischer and K. U. Simmer, "An adaptive microphone array for hands-free communication," in *Proc. 4th Int. Workshop Acoust. Echo Noise Contr.*, Røros, Norway, June 21–23, 1995, pp. 44–47.

[8] ——, "Beamforming microphone arrays for speech acquisition in noisy environments," *Speech Commun.*, vol. 20, no. 3–4, pp. 215–227, Dec. 1996.

[9] K. U. Simmer, S. Fischer, and A. Wasiljeff, "Suppression of coherent and incoherent noise using a microphone array," *Ann. Télécommun.*, vol. 49, no. 7–8, pp. 439–446, July 1994.

[10] J. Bitzer, K. U. Simmer, and K.-D. Kammeyer, "Multichannel noise reduction-algorithms and theoretical limits," in *Proc. Eur. Signal Process. Conf.*, Rhodes, Greece, Sept. 8–11, 1998, pp. 105–108.

[11] ——, "Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement," in *Proc. 24th IEEE Int.. Conf. Acoust. Speech Signal Process.*, Phoenix, AZ, Mar. 15–19, 1999, pp. 2965–2968.

[12] ——, "Multi-microphone noise reduction by post-filter and superdirective beamformer," in *Proc. 6th Int. Workshop Acoust. Echo Noise Contr.*, Pocono Manor, PA, Sept. 27–30, 1999, pp. 100–103.

[13] ——, "Multi-microphone noise reduction techniques as front-end devices for speech recognition," *Speech Commun.*, vol. 34, no. 1-2, pp. 3–12, Apr. 2001.

[14] R. Le Bouquin-Jeannès, A. A. Azirani, and G. Faucon, "Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 484–487, Sept. 1997.

[15] J. Meyer and K. U. Simmer, "Multichannel speech enhancement in a car environment using Wiener filtering and spectral subtraction," in *Proc. 22th IEEE Int. Conf. Acoust. Speech Signal Process.*, Munich, Germany, Apr. 20–24, 1997, pp. 1167–1170.

[16] D. Mahmoudi, "A microphone array for speech enhancement using multiresolution wavelet transform," in *Proc. 5th Eur. Conf. Speech, Commun. Technol.*, Rhodes, Greece, Sept. 22–25, 1997, pp. 339–342.

[17] D. Mahmoudi and A. Drygajlo, "Combined Wiener and coherence filtering in wavelet domain for microphone array speech enhancement," in *Proc. 23th IEEE Int. Conf. Acoust. Speech Signal Process.*, Seattle, WA, May 12–15, 1998, pp. 385–388.

[18] S. Fischer and K.-D. Kammeyer, "Broadband beamforming with adaptive postfiltering for speech acquisition in noisy environments," in *Proc. 22th IEEE Int. Conf. Acoust. Speech Signal Process.*, Munich, Germany, Apr. 20–24, 1997, pp. 359–362.

[19] I. A. McCowan, C. Marro, and L. Mauuary, "Robust speech recognition using near-field superdirective beamforming with post-filtering," in *Proc. 25th IEEE Int. Conf. Acoust. Speech Signal Process.*, Istanbul, Turkey, June 5–9, 2000, pp. 1723–1726.

[20] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 1109–1121, Dec. 1984.

[21] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 443–445, Apr. 1985.

[22] I. Cohen and B. Berdugo, "Speech enhancement for nonstationary noise environments," *Signal Process.*, vol. 81, pp. 2403–2418, Nov. 2001.

[23] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Processing*, vol. 11, pp. 466–475, Sept. 2003.

[24] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Processing*, vol. 49, pp. 1614–1626, Aug. 2001.

[25] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propagat.*, vol. AP-30, pp. 27–34, Jan. 1982.

[26] C. W. Jim, "A comparison of two LMS constrained optimal array structures," *Proc. IEEE*, vol. 65, pp. 1730–1731, Dec. 1977.

[27] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1985.

[28] S. Nordholm, I. Claesson, and P. Eriksson, "The broadband Wiener solution for Griffiths-Jim beamformers," *IEEE Trans. Signal Processing*, vol. 40, pp. 474–478, Feb. 1992.

[29] P. D. Welch, "The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short modified periodograms," *IEEE Trans. Audio Electroacoust.*, vol. AU-15, pp. 70–73, June 1967.

[30] I. Cohen and B. Berdugo, "Microphone array post-filtering for nonstationary noise suppression," in *Proc. 27th IEEE Int. Conf. Acoust. Speech Signal Process.*, Orlando, FL, May 13–17, 2002, pp. 901–904.

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

[32] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, 2nd ed. New York: IEEE, 2000.

[33] P. E. Papamichalis, *Practical Approaches to Speech Coding*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[34] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 504–512, July 2001.

Israel Cohen (M'01–SM'03) received the B.Sc. (*Summa Cum Laude*), M.Sc., and Ph.D. degrees in electrical engineering in 1990, 1993, and 1998, respectively, all from the Technion—Israel Institute of Technology, Haifa, Israel.

From 1990 to 1998, he was a Research Scientist at RAFAEL Research Laboratories, Israel Ministry of Defense, Haifa. From 1998 to 2001, he was a Postdoctoral Research Associate at the Computer Science Department, Yale University, New Haven, CT. Since 2001, he has been a Senior Lecturer with the Electrical Engineering Department, Technion. His research interests are multichannel speech enhancement, image and multidimensional data processing, anomaly detection, wavelet theory, and applications.