Signal Enhancement Using Beamforming and Nonstationarity with Applications to Speech

Sharon Gannot, *Student Member, IEEE*, David Burshtein, *Senior Member, IEEE*, and Ehud Weinstein, *Fellow, IEEE*

***Abstract*—We consider a sensor array located in an enclosure, where arbitrary transfer functions (TFs) relate the source signal and the sensors. The array is used for enhancing a signal contaminated by interference. Constrained minimum power adaptive beamforming, which has been suggested by Frost and, in particular, the generalized sidelobe canceler (GSC) version, which has been developed by Griffiths and Jim, are the most widely used beamforming techniques. These methods rely on the assumption that the received signals are simple delayed versions of the source signal. The good interference suppression attained under this assumption is severely impaired in complicated acoustic environments, where arbitrary TFs may be encountered. In this paper, we consider the arbitrary TF case. We propose a GSC solution, which is adapted to the general TF case. We derive a suboptimal algorithm that can be implemented by estimating the TFs ratios, instead of estimating the TFs. The TF ratios are estimated by exploiting the nonstationarity characteristics of the desired signal. The algorithm is applied to the problem of speech enhancement in a reverberating room. The discussion is supported by an experimental study using speech and noise signals recorded in an actual room acoustics environment.**

***Index Terms*—Beamforming, nonstationarity, speech enhancement.**

I. Introduction

SIGNAL quality might significantly deteriorate in the presence of interference, especially when the signal is also subject to reverberation. Multisensor-based enhancement algorithms typically incorporate both spatial and spectral information. Hence, they have the potential to improve on single sensor solutions that utilize only spectral information. In particular, when the desired signal is speech, single microphone solutions are known to be limited in their performance. Beamforming methods have therefore attracted a great deal of interest in the past three decades. Applications of beamforming to the speech enhancement problem have also emerged recently.

Constrained minimum power adaptive beamforming, which has been suggested by Frost [1], deals with the problem of a broadband signal received by an array, where pure delay relates each pair of source and sensor. Each sensor signal is processed by a tap delay line after applying a proper time delay compensation. The algorithm is capable of satisfying some desired frequency response in the look direction while minimizing the output noise power by using constrained minimization of the total output power. This minimization is realized by adjusting the taps of the filters under the desired constraint. Frost suggested a constrained LMS-type algorithm. Griffiths and Jim [2] reconsidered Frost's algorithm and introduced the generalized sidelobe canceler (GSC) solution. The GSC algorithm is comprised of three building blocks. The first is a fixed beamformer, which satisfies the desired constraint. The second is a blocking matrix, which produces noise-only reference signals by blocking the desired signal (e.g., by subtracting pairs of time-aligned signals). The third is an unconstrained LMS-type algorithm that attempts to cancel the noise in the fixed beamformer output. In [2], it is shown that Frost algorithm can be viewed as a special case of the GSC. The main drawback of the GSC algorithm is its delay-only propagation assumption.

Van Veen and Buckley [3] summarized various methods for spatial filtering, including the GSC, and introduced a wider range of possible constraints on the beam pattern. Cox *et al.* [4] suggested constraint of the norm of the adaptive canceler coefficients in order to solve the superdirectivity problem, i.e., its sensitivity to steering errors. In particular, they have suggested to update Frost's (or the Griffiths and Jim) algorithm by applying a quadratic constraint on the norm of the noise canceler coefficients. This constraint, which can limit the superdirectivity, is added to the usual linear constraints.

Some authors have recently suggested using the GSC for speech enhancement in a reverberating environment. Hoshuyama *et al.* [5]–[7] used a three-block structure similar to the GSC. However, the blocking matrix has been modified to operate adaptively. In order to limit the leakage of the desired signal, which is responsible for distortion in the output signal, a quadratic constraint is imposed on the norm of the noise canceler coefficients. Alternatively, use of the leaky LMS algorithm has been suggested.

Nordholm *et al.* [8] used a GSC solution in which the blocking matrix is realized by spatial highpass filtering, thus yielding improved noise-only reference signals. Meyer and Sydow [9] have suggested to construct the noise reference signals by steering the lobes of a multibeam beamformer toward the noise and desired signal directions separately.

Widrow and Stearns [10] have proposed a dual structure beamformer. The *master* beamformer adapts its coefficients to minimize the output power while maintaining the beam-pattern toward a predetermined *pilot* signal from the desired direction. Those coefficients are continuously copied to a *slave* beamformer that is used to enhance the speech signal. Dahl *et al.* [11] have extended this solution by proposing a dual

Manuscript received March 28, 2000; revised April 30, 2001. The associate editor coordinating the review of this paper and approving it for publication was Dr. Alex C. Kot.

S. Gannot is with the Department of Electrical Engineering (SISTA), Katholieke Universiteit Leuven, Leuven, Belgium (e-mail: Sharon.Gannot@esat.kuleuven.ac.be).

D. Burshtein and E. Weinstein are with the Department of Electrical Engineering—Systems, Tel-Aviv University, Tel-Aviv, Israel (e-mail: burstyn@eng.tau.ac.il; udi@eng.tau.ac.il).

Publisher Item Identifier S 1053-587X(01)05874-3.

1053–587X/01$10.00 © 2001 IEEE

beamformer that attempts to cancel both noise and jammer signals (e.g., loudspeaker). The pilot signal is constructed by offline recordings of the jammer and desired signal in the actual acoustic environment during a calibration phase. Thus, both echo cancellation and noise suppression are achieved simultaneously.

Other solutions utilize a beamformer type algorithm, followed by a postprocessor. Zelinski [12] suggested a Wiener filter, followed by further noise reduction in a postprocessing configuration. Meyer and Simmer [13] addressed the problem of high coherence between the microphone signals at low frequencies, as indicated by Dal-Degan and Prati [14]. They have suggested the use of a spectral subtraction algorithm in the low-frequency band and Wiener filtering in the high-frequency band. Fischer and Kammeyer [15] suggested to further split the microphone array into differentially equispaced subarrays. This structure has been further analyzed by Marro *et al.* [16]. Bitzer *et al.* [17] analyzed the performance of the GSC solution and showed its dependence on the noise field. They showed that the noise reduction might be infinitely large when the noise source is directional. However, in the more practical situation of a reverberant enclosure, when the noise field can be regarded as diffused, the performance degrades severely. Bitzer *et al.* [18] suggested a GSC with fixed Wiener filters in the noise canceling block and further postfilters at the GSC output. An improved performance in the lower frequency range is achieved. In [19], it is shown that the Wiener filters can be computed in advance by utilizing prior knowledge of the noise field.

Jan and Flanagan [20] suggested a matched filter beamforming (MFBF) instead of the conventional delay and sum beamformer (DSBF). The MFBF configuration realizes signal alignment by convolving the microphone signals with the (estimated) acoustic transfer function (TF). Rabinkin *et al.* [21] proved that the performance of MFBF is superior to DSBF, provided that the room acoustics TF is not too complicated. They have also suggested truncation of the estimated acoustic TFs to ensure reliable estimates.

Grenier *et al.* [22]–[29] have proposed GSC-based enhancement algorithms. In [29], the case where general TFs relate the source and microphones was considered. A subspace tracking solution [30] has been proposed. The resulting TFs are constrained to the array manifold under the assumption of an FIR model and small displacements of the talker. The fixed beamformer block of the GSC is realized using MFBF.

In this paper, we consider a sensor array located in an enclosure, where general TFs relate the source signal and the sensors. The array is used for enhancing a signal contaminated by interference. We propose a GSC solution, which is adapted to the general TF case. The TFs are estimated by exploiting the nonstationarity characteristics of the desired signal. The algorithm is applied to the problem of speech enhancement in a reverberating room. The discussion is supported by an experimental study using speech and noise signals recorded in an actual room acoustics environment. The outcome consists of the assessment of sound sonograms, signal-to-noise ratio (SNR) enhancement, and informal subjective listening tests. The paper is organized as follows. In Section II, we formulate the problem of beamforming in a general TF environment in the frequency domain. The constrained power minimization is presented in Section III, where both Frost's algorithm [1] and the Griffiths and Jim [2] interpretation are derived in the frequency domain. This derivation motivates the intuitive structure suggested by other authors for the beamforming problem in reverberant environments. We then show that a suboptimal algorithm can be implemented by estimating the TF ratios instead of estimating the actual TFs. In Section IV, we address the problem of estimating the TF ratios by extending the nonstationarity principle, which was suggested by Shalvi and Weinstein [31]. An application of the suggested algorithm to the speech enhancement problem is presented in Section V. Section VI concludes the paper.

II. Problem Formulation

Consider an array of sensors in a noisy and reverberant environment. The received signal is comprised of two components. The first is some nonstationary (e.g., speech) signal. The second is some stationary interference signal. Our goal is to reconstruct the nonstationary signal component from the received signals. We use the following notation.

$z_m(t)$	mth sensor signal;
$s(t)$	desired signal source;
$n_m(t)$	interference signal of the mth sensor comprised of some directional noise component and some ambient noise component;
$a_m(t)$	time-varying TFs from the desired speech source to the mth sensor.

We have

$$z_m(t) = a_m(t) * s(t) + n_m(t); \qquad m = 1, \ldots, M \tag{1}$$

where $*$ denotes convolution. Suppose that the analysis frame duration T is chosen such that the signal may be considered stationary over the analysis frame. Typically, the TFs are changing slowly in time so that they may also be considered stationary over the analysis frame. Multiplying both sides of (1) by a rectangular window function $w(t)$ [$w(t) = 1$ over the analysis frame $w(t) = 0$ otherwise] and applying the discrete time Fourier transform (DTFT) operator yields

$$Z_m(t, e^{j\omega}) \approx A_m(e^{j\omega}) S(t, e^{j\omega}) + N_m(t, e^{j\omega}) \qquad m = 1, \ldots, M. \tag{2}$$

The approximation is justified for T sufficiently large. $Z_m(t, e^{j\omega})$, $S(t, e^{j\omega})$ and $N_m(t, e^{j\omega})$ are the short time Fourier transforms (STFTs) of the respective signals. $A_m(e^{j\omega})$ is the TF of the mth sensor. Note that we have assumed that the TFs are time invariant.

The vector formulation of the equation set (2) is

$$\mathbf{Z}(t, e^{j\omega}) = \mathbf{A}(e^{j\omega}) S(t, e^{j\omega}) + \mathbf{N}(t, e^{j\omega}) \tag{3}$$

where

$$\mathbf{Z}^T(t, e^{j\omega}) = \left[Z_1(t, e^{j\omega}) \;\; Z_2(t, e^{j\omega}) \;\; \cdots \;\; Z_M(t, e^{j\omega}) \right]$$
$$\mathbf{A}^T(e^{j\omega}) = \left[A_1(e^{j\omega}) \;\; A_2(e^{j\omega}) \;\; \cdots \;\; A_M(e^{j\omega}) \right]$$
$$\mathbf{N}^T(t, e^{j\omega}) = \left[N_1(t, e^{j\omega}) \;\; N_2(t, e^{j\omega}) \;\; \cdots \;\; N_M(t, e^{j\omega}) \right].$$

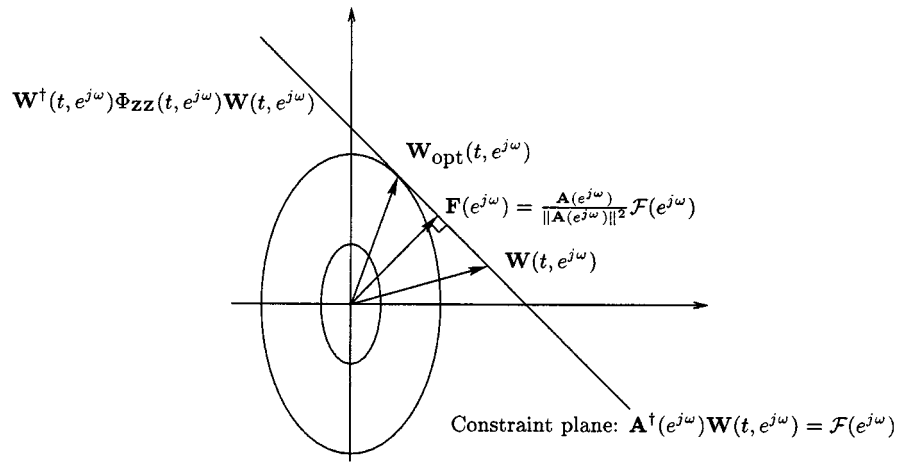

Fig. 1. Constrained minimization.

III. Constrained Output Power Minimization

In [1], a beamforming algorithm was proposed under the assumption that the TF from the desired signal source to each sensor includes only gain and delay values. In this section, we consider the general case of arbitrary TFs. By following the derivation of [1] in the frequency domain, we derive a beamforming algorithm for the general TF case. First, we obtain a closed-form, linearly constrained, minimum variance beamformer. Then, we derive an adaptive solution. The outcome will be a constrained LMS-type algorithm. We proceed, following the footsteps of Griffiths and Jim [2], and formulate an unconstrained adaptive solution. We will initially assume that the TFs are known. Later, in Section IV, we deal with the problem of estimating the TFs.

A. Frequency Domain Frost Algorithm

1) Optimal Solution: Let $W^*(t, e^{j\omega}); m = 1, \ldots, M$ be a set of M filters

$$\mathbf{W}^{\dagger}(t, e^{j\omega}) = \left[W_1^*(t, e^{j\omega}) \;\; W_2^*(t, e^{j\omega}) \;\; \cdots \;\; W_M^*(t, e^{j\omega})\right]$$

where * denotes conjugation, and † denotes conjugation transpose. A beamformer is realized by filtering each sensor output by $W^*(t, e^{j\omega})\, m = 1, \ldots, M$ and summing the outputs

$$\begin{aligned} Y(t, e^{j\omega}) &= \mathbf{W}^{\dagger}(t, e^{j\omega})\mathbf{Z}(t, e^{j\omega}) \\ &= \mathbf{W}^{\dagger}(t, e^{j\omega})\mathbf{A}(e^{j\omega})S(t, e^{j\omega}) \\ &\quad + \mathbf{W}^{\dagger}(t, e^{j\omega})\mathbf{N}(t, e^{j\omega}) \\ &\triangleq Y_s(t, e^{j\omega}) + Y_n(t, e^{j\omega}) \end{aligned} \tag{4}$$

where $Y_s(t, e^{j\omega})$ is the desired signal part, and $Y_n(t, e^{j\omega})$ is the noise part. The output power of the beamformer is

$$\begin{aligned} &E\left\{Y(t, e^{j\omega})Y^*(t, e^{j\omega})\right\} \\ &\quad = E\left\{\mathbf{W}^{\dagger}(t, e^{j\omega})\mathbf{Z}(t, e^{j\omega})\mathbf{Z}^{\dagger}(t, e^{j\omega})\mathbf{W}(t, e^{j\omega})\right\} \\ &\quad = \mathbf{W}^{\dagger}(t, e^{j\omega})\mathbf{\Phi_{ZZ}}(t, e^{j\omega})\mathbf{W}(t, e^{j\omega}) \end{aligned}$$

where $\mathbf{\Phi_{ZZ}}(t, e^{j\omega}) \triangleq E\{\mathbf{Z}(t, e^{j\omega})\mathbf{Z}^{\dagger}(t, e^{j\omega})\}$. We want to minimize the output power subject to the following constraint on $Y_s(t, e^{j\omega})$

$$\begin{aligned} Y_s(t, e^{j\omega}) &= \mathbf{W}^{\dagger}(t, e^{j\omega})\mathbf{A}(e^{j\omega})S(t, e^{j\omega}) \\ &= \mathcal{F}^*(t, e^{j\omega})S(t, e^{j\omega}) \end{aligned}$$

where $\mathcal{F}^*(t, e^{j\omega})$ is some prespecified filter (usually a simple delay). We thus have the following minimization problem:

$$\begin{aligned} &\min_{\mathbf{W}} \left\{\mathbf{W}^{\dagger}(t, e^{j\omega})\mathbf{\Phi_{ZZ}}(t, e^{j\omega})\mathbf{W}(t, e^{j\omega})\right\} \\ &\text{subject to } \mathbf{W}^{\dagger}(t, e^{j\omega})\mathbf{A}(e^{j\omega}) = \mathcal{F}^*(t, e^{j\omega}). \end{aligned} \tag{5}$$

The minimization (5) is demonstrated in Fig. 1. The point where the equipower contours are tangent to the constraint plane is the optimum vector of beamforming filters. The perpendicular $\mathbf{F}(e^{j\omega})$ from the origin to the constraint plane will be calculated in Section III-A2.

To solve (5), we first define the following complex Lagrange functional:

$$\begin{aligned} \mathcal{L}(\mathbf{W}) &= \mathbf{W}^{\dagger}(t, e^{j\omega})\mathbf{\Phi_{ZZ}}(t, e^{j\omega})\mathbf{W}(t, e^{j\omega}) \\ &\quad + \lambda\left[\mathbf{W}^{\dagger}(t, e^{j\omega})\mathbf{A}(e^{j\omega}) - \mathcal{F}^*(t, e^{j\omega})\right] \\ &\quad + \lambda^*\left[\mathbf{A}^{\dagger}(t, e^{j\omega})\mathbf{W}(e^{j\omega}) - \mathcal{F}(t, e^{j\omega})\right] \end{aligned}$$

where λ is a Lagrange multiplier. Setting the derivative with respect to $\mathbf{W}^*$ to 0 (e.g., [32]) yields

$$\nabla_{\mathbf{W}^*}\mathcal{L}(W) = \mathbf{\Phi_{ZZ}}(t, e^{j\omega})\mathbf{W}(t, e^{j\omega}) + \lambda\mathbf{A}(e^{j\omega}) = 0.$$

Now, recalling the constraint in (5), we obtain the following set of optimal filters:

$$\begin{aligned} \mathbf{W}^{\mathrm{opt}}(t, e^{j\omega}) = &\left[\mathbf{A}^{\dagger}(e^{j\omega})\mathbf{\Phi}_{\mathbf{ZZ}}^{-1}(t, e^{j\omega})\mathbf{A}(e^{j\omega})\right]^{-1} \\ &\cdot \mathbf{\Phi}_{\mathbf{ZZ}}^{-1}(t, e^{j\omega})\mathbf{A}(e^{j\omega})\mathcal{F}(e^{j\omega}). \end{aligned}$$

This closed-form solution is difficult to implement and does not have the ability to track changes in the environment. Therefore, an adaptive solution should be more useful.

$$\mathbf{W}(t=0, e^{j\omega}) = \mathbf{F}(e^{j\omega})$$
$$\mathbf{W}(t+1, e^{j\omega}) =$$
$$P(e^{j\omega})\left[\mathbf{W}(t, e^{j\omega}) - \mu\mathbf{Z}(t, e^{j\omega})Y^*(t, e^{j\omega})\right] + \mathbf{F}(e^{j\omega})$$
$$t = 0, 1, \ldots$$
$$(P(e^{j\omega}) \text{ and } \mathbf{F}(e^{j\omega}) \text{ are defined by (6) and (7)}).$$

Fig. 2. Frequency domain frost algorithm.

2) Adaptive Solution: Consider the following steepest descent, adaptive algorithm:

$$\begin{aligned}\mathbf{W}(t+1, e^{j\omega}) &= \mathbf{W}(t, e^{j\omega}) - \mu\nabla_{\mathbf{W}^*}\mathcal{L}(e^{j\omega}) \\ &= \mathbf{W}(t, e^{j\omega}) - \mu\left[\Phi_{\mathbf{ZZ}}(t, e^{j\omega})\mathbf{W}(t, e^{j\omega}) + \lambda\mathbf{A}(e^{j\omega})\right].\end{aligned}$$

Imposing our constraint on $\mathbf{W}(t+1, e^{j\omega})$ yields

$$\begin{aligned}\mathcal{F}(e^{j\omega}) &= \mathbf{A}^\dagger(e^{j\omega})\mathbf{W}(t+1, e^{j\omega}) \\ &= \mathbf{A}^\dagger(e^{j\omega})\mathbf{W}(t, e^{j\omega}) \\ &\quad - \mu\mathbf{A}^\dagger(e^{j\omega})\Phi_{\mathbf{ZZ}}(t, e^{j\omega})\mathbf{W}(t, e^{j\omega}) \\ &\quad - \mu\mathbf{A}^\dagger(e^{j\omega})\mathbf{A}(e^{j\omega})\lambda.\end{aligned}$$

Solving for the Lagrange multiplier and applying further rearrangement of terms yields

$$\begin{aligned}\mathbf{W}(t+1, e^{j\omega}) &= P(e^{j\omega})\mathbf{W}(t, e^{j\omega}) \\ &\quad - \mu P(e^{j\omega})\Phi_{\mathbf{ZZ}}(t, e^{j\omega})\mathbf{W}(t, e^{j\omega}) + \mathbf{F}(e^{j\omega})\end{aligned}$$

where

$$P(e^{j\omega}) = I - \frac{\mathbf{A}(e^{j\omega})\mathbf{A}^\dagger(e^{j\omega})}{||\mathbf{A}(e^{j\omega})||^2} \tag{6}$$

and

$$\mathbf{F}(e^{j\omega}) = \frac{\mathbf{A}(e^{j\omega})}{||\mathbf{A}(e^{j\omega})||^2}\mathcal{F}(e^{j\omega}). \tag{7}$$

Further simplification can be achieved by replacing $\Phi_{\mathbf{ZZ}}(t, e^{j\omega})$ by its instantaneous estimator $\mathbf{Z}(t, e^{j\omega})\mathbf{Z}^\dagger(t, e^{j\omega})$ and recalling (4). We thus obtain

$$\begin{aligned}\mathbf{W}(t+1, e^{j\omega}) &= P(e^{j\omega})\left[\mathbf{W}(t, e^{j\omega}) - \mu\mathbf{Z}(t, e^{j\omega})Y^*(t, e^{j\omega})\right] + \mathbf{F}(e^{j\omega}).\end{aligned}$$

The algorithm is summarized in Fig. 2.

B. Generalized Sidelobe Canceler (GSC) Interpretation

In [2], Griffiths and Jim considered the case where each TF is a delay element (with some gain). Griffiths and Jim obtained an unconstrained adaptive enhancement algorithm, using the same constrained, minimum output power criterion used by Frost [1]. The unconstrained algorithm is computationally more efficient than the constrained algorithm. Furthermore, the unconstrained algorithm is based on the well behaved NLMS scheme. In Section III-A2, we obtained an adaptive algorithm for the case where each TF is represented by an arbitrary linear time-invariant system by tracing the derivation of Frost in the frequency domain. We now repeat the arguments of Griffiths and Jim for our case (arbitrary TFs) and derive an unconstrained adaptive enhancement algorithm.

Consider the null space of $\mathbf{A}(e^{j\omega})$, which is defined by

$$\mathcal{N}(e^{j\omega}) \triangleq \left\{\mathbf{W} \,|\, \mathbf{A}^\dagger(e^{j\omega})\mathbf{W} = 0\right\}.$$

The constraint hyperplane

$$\Lambda(e^{j\omega})) \triangleq \left\{\mathbf{W} \,|\, \mathbf{A}^\dagger(e^{j\omega})\mathbf{W} = \mathcal{F}(e^{j\omega})\right\}$$

is parallel to $\mathcal{N}(e^{j\omega})$. In addition to that, let

$$\mathcal{R}(e^{j\omega}) \triangleq \left\{\kappa\mathbf{A}(e^{j\omega}) \,|\, \text{for any real } \kappa\right\}$$

be the column space. By the fundamental theorem of linear algebra (e.g., [33]) $\mathcal{R}(e^{j\omega}) \perp \mathcal{N}(e^{j\omega})$. In particular, $\mathbf{F}(e^{j\omega})$ is perpendicular to $\mathcal{N}(e^{j\omega})$ since

$$\mathbf{F}(e^{j\omega}) = \frac{\mathcal{F}(e^{j\omega})}{||\mathbf{A}(e^{j\omega})||^2}\mathbf{A}(e^{j\omega}) \in \mathcal{R}(e^{j\omega}).$$

Furthermore

$$\begin{aligned}&\mathbf{A}^\dagger(e^{j\omega})\mathbf{F}(e^{j\omega}) \\ &= \mathbf{A}^\dagger(e^{j\omega})\mathbf{A}(e^{j\omega})\left(\mathbf{A}^\dagger(e^{j\omega})\mathbf{A}(e^{j\omega})\right)^{-1}\mathcal{F}(e^{j\omega}) = \mathcal{F}(e^{j\omega}).\end{aligned}$$

Thus, $\mathbf{F}(e^{j\omega}) \in \Lambda(e^{j\omega})$ and $\mathbf{F}(e^{j\omega}) \perp \Lambda(e^{j\omega})$. Hence, $\mathbf{F}(e^{j\omega})$ is the perpendicular from the origin to the constraint hyperplane $\Lambda(e^{j\omega})$. The matrix $P(e^{j\omega})$, which is defined in (6), is the *projection matrix* to the null space of $\mathbf{A}(e^{j\omega})$, $\mathcal{N}(e^{j\omega})$.

Now, a vector in linear space can be uniquely split into a sum of two vectors in mutually orthogonal subspaces (e.g., [33]). Hence

$$\mathbf{W}(t, e^{j\omega}) = \mathbf{W}_0(t, e^{j\omega}) - \mathbf{V}(t, e^{j\omega}) \tag{8}$$

where $\mathbf{W}_0(t, e^{j\omega}) \in \mathcal{R}(e^{j\omega})$, and $-\mathbf{V}(t, e^{j\omega}) \in \mathcal{N}(e^{j\omega})$. By the definition of $\mathcal{N}(e^{j\omega})$

$$\mathbf{V}(t, e^{j\omega}) = \mathcal{H}(e^{j\omega})\mathbf{G}(t, e^{j\omega}) \tag{9}$$

where $\mathcal{H}(e^{j\omega})$ is some $M \times (M-1)$ matrix, such that the columns of $\mathcal{H}(e^{j\omega})$ span the null space of $\mathbf{A}(e^{j\omega})$, i.e.,

$$\mathbf{A}^\dagger(e^{j\omega})\mathcal{H}(e^{j\omega}) = 0 \qquad \text{rank}\left\{\mathcal{H}(e^{j\omega})\right\} = M - 1. \tag{10}$$

The vector $\mathbf{G}(t, e^{j\omega})$ is an $(M-1) \times 1$ vector of adjustable filters. By the geometrical interpretation of Frost's algorithm

$$\mathbf{W}_0(t, e^{j\omega}) = \mathbf{F}(e^{j\omega}) = \frac{\mathbf{A}(e^{j\omega})}{||\mathbf{A}(e^{j\omega})||^2}\mathcal{F}(e^{j\omega}). \tag{11}$$

[Recall that $\mathbf{F}(e^{j\omega})$ is the perpendicular from the origin to the constraint hyperplane $\Lambda(e^{j\omega})$.] Now, using (4), (8), and (9) we get

$$Y(t, e^{j\omega}) = Y_{\text{FBF}}(t, e^{j\omega}) - Y_{\text{NC}}(t, e^{j\omega}) \tag{12}$$

where

$$\begin{aligned}Y_{\text{FBF}}(t, e^{j\omega}) &= \mathbf{W}_0^\dagger(t, e^{j\omega})\mathbf{Z}(t, e^{j\omega}) \\ Y_{\text{NC}}(t, e^{j\omega}) &= \mathbf{G}^\dagger(t, e^{j\omega})\mathcal{H}^\dagger(e^{j\omega})\mathbf{Z}(t, e^{j\omega}).\end{aligned} \tag{13}$$

The output of the constrained beamformer is a difference of two terms, both operating on the input signal $\mathbf{Z}(t, e^{j\omega})$. The first term $Y_{\mathrm{FBF}}(t, e^{j\omega})$ utilizes only fixed components (which depend on the TFs); therefore, it can be viewed as a fixed beamformer (FBF). We now examine the second term $Y_{\mathrm{NC}}(t, e^{j\omega})$. Note that

$$\begin{aligned}\mathbf{U}(t, e^{j\omega}) &= \mathcal{H}^{\dagger}(e^{j\omega})\mathbf{Z}(t, e^{j\omega}) \\ &= \mathcal{H}^{\dagger}(e^{j\omega})\left[\mathbf{A}(e^{j\omega})S(t, e^{j\omega}) + \mathbf{N}(t, e^{j\omega})\right] \\ &= \mathcal{H}^{\dagger}(e^{j\omega})\mathbf{N}(t, e^{j\omega}). \end{aligned} \tag{14}$$

The last transition is due to (10). $\mathbf{U}(t, e^{j\omega})$ are *reference noise* signals. Hence, the signal dependent component of $Y_{\mathrm{NC}}(t, e^{j\omega})$ is completely eliminated (blocked) by $\mathcal{H}^{\dagger}(e^{j\omega})$ so that $Y_{\mathrm{NC}}(t, e^{j\omega})$ is a pure noise term. The noise term of $Y_{\mathrm{FBF}}(t, e^{j\omega})$ can be reduced by properly adjusting the filters $\mathbf{G}(t, e^{j\omega})$, using the minimum output power criterion. This adjustment problem is in fact the classical multichannel noise cancellation problem. An adaptive LMS solution to the problem was proposed by Widrow [34].

The GSC solution is comprised of three components:

1) fixed beamformer (FBF);
2) blocking matrix (BM) that constructs the noise reference signals;
3) multichannel noise canceler (NC).

We now discuss each of these components in details.

1) Fixed Beamformer (FBF): By (3), (11), and (13), we have

$$\begin{aligned}Y_{\mathrm{FBF}}(t, e^{j\omega}) &= \mathcal{F}^{*}(e^{j\omega})S(t, e^{j\omega}) \\ &\quad + \frac{\mathcal{F}^{*}(e^{j\omega})}{\|\mathbf{A}(e^{j\omega})\|^2}\mathbf{A}^{\dagger}(e^{j\omega})\mathbf{N}(t, e^{j\omega}).\end{aligned}$$

The first term on the right-hand side is the signal term. The second is the noise term. Note that by setting $\mathcal{F}^{*}(e^{j\omega}) = e^{-j\omega\tau}$ (i.e., a delay), the signal component of $Y_{\mathrm{FBF}}(t, e^{j\omega})$ is an undistorted, delayed version of the desired signal.

Unfortunately, we usually do not have access to the actual TFs ($A_m(e^{j\omega})$; $m = 1, \ldots, M$). Later, we show how we can estimate the TFs ratio

$$H_m(e^{j\omega}) = \frac{A_m(e^{j\omega})}{A_1(e^{j\omega})}; \qquad m = 1, \ldots, M. \tag{15}$$

Let

$$\mathbf{H}^{T}(e^{j\omega}) = \left[1 \;\; \frac{A_2(e^{j\omega})}{A_1(e^{j\omega})} \;\; \cdots \;\; \frac{A_M(e^{j\omega})}{A_1(e^{j\omega})}\right] = \frac{\mathbf{A}^{T}(e^{j\omega})}{A_1(e^{j\omega})}.$$

If in (11), the actual TFs are replaced by the TFs ratios, then

$$\mathbf{W}_0(t, e^{j\omega}) = \frac{\mathbf{H}(e^{j\omega})}{\|\mathbf{H}(e^{j\omega})\|^2}\mathcal{F}(e^{j\omega}). \tag{16}$$

By (3) and (13), we have

$$\begin{aligned}Y_{\mathrm{FBF}}(t, e^{j\omega}) &= A_1(e^{j\omega})\mathcal{F}^{*}(e^{j\omega})S(t, e^{j\omega}) \\ &\quad + \frac{\mathcal{F}^{*}(e^{j\omega})}{\|\mathbf{H}(e^{j\omega})\|^2}\mathbf{H}^{\dagger}(e^{j\omega})\mathbf{N}(t, e^{j\omega}).\end{aligned}$$

Thus, when $\mathbf{W}_0(t, e^{j\omega})$ is given by (16), the signal term of $Y_{\mathrm{FBF}}(t, e^{j\omega})$ is the desired signal distorted only by the first TF $A_1(e^{j\omega})$. Now, suppose that

$$\mathbf{W}_0(t, e^{j\omega}) = \mathbf{H}(e^{j\omega})\mathcal{F}(e^{j\omega}). \tag{17}$$

In this case, $\mathbf{W}_0(t, e^{j\omega})$ is comprised of the cascade of $\mathbf{H}(e^{j\omega})$, which is a filter matched to the TFs ratio, and $\mathcal{F}(e^{j\omega})$. The new $\mathbf{W}_0(t, e^{j\omega})$ can be derived from (16) under the assumption that $\|\mathbf{H}(e^{j\omega})\|^2$ is constant. In fact, Grenier *et al.* [29] argue that this assumption can be verified empirically. The FBF term of the output is now given by

$$\begin{aligned}Y_{\mathrm{FBF}}(t, e^{j\omega}) &= \frac{\|\mathbf{A}(e^{j\omega})\|^2}{A_1^{*}(e^{j\omega})}\mathcal{F}^{*}(e^{j\omega})S(t, e^{j\omega}) \\ &\quad + \mathcal{F}^{*}(e^{j\omega})\mathbf{H}^{\dagger}(e^{j\omega})\mathbf{N}(t, e^{j\omega}).\end{aligned} \tag{18}$$

The signal component of $Y_{\mathrm{FBF}}(t, e^{j\omega})$ is now distorted. Hence, only a suboptimal solution is achieved. Note, however, that all the sensor outputs are added together coherently [this can be seen from the term $\|\mathbf{A}(e^{j\omega})\|^2$].

2) Blocking Matrix (BM): Consider the following $M \times (M-1)$ matrix $\mathcal{H}(e^{j\omega})$:

$$\mathcal{H}(e^{j\omega}) = \begin{bmatrix} -\dfrac{A_2^{*}(e^{j\omega})}{A_1^{*}(e^{j\omega})} & -\dfrac{A_3^{*}(e^{j\omega})}{A_1^{*}(e^{j\omega})} & \cdots & -\dfrac{A_M^{*}(e^{j\omega})}{A_1^{*}(e^{j\omega})} \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & \cdots & \ddots & \\ 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{19}$$

It can be easily verified that this matrix satisfies (10) and is, hence, a proper blocking matrix that may be used for generating the reference noise signals $\mathbf{U}(t, e^{j\omega})$. By (14), we have

$$\begin{aligned}U_m(e^{j\omega}) &= Z_m(t, e^{j\omega}) - \frac{A_m(e^{j\omega})}{A_1(e^{j\omega})}Z_1(t, e^{j\omega}) \\ &\quad m = 2, \ldots, M.\end{aligned} \tag{20}$$

Thus, the knowledge of the TFs ratios $H_m(e^{j\omega}) = A_m(e^{j\omega})/A_1(e^{j\omega})$ is sufficient to implement the sidelobe canceler.

3) Noise Canceler: By the GSC derivation, we have constructed two signals. The first is $Y_{\mathrm{FBF}}(t, e^{j\omega})$, which contains both a desired speech term and a residual noise term. The second signal is $Y_{\mathrm{NC}}(t, e^{j\omega})$. $Y_{\mathrm{NC}}(t, e^{j\omega})$ consists of an adaptive set of filters $\mathbf{G}(t, e^{j\omega})$ that are applied to the noise-only signals $\mathbf{U}(t, e^{j\omega})$.

Recall that our goal is to minimize the output power under a constraint on the response at the desired direction. By setting $\mathbf{W}_0(t, e^{j\omega})$ according to (11), the constraint is satisfied. Hence, minimization of the output power is achieved by adjusting the filters $\mathbf{G}(t, e^{j\omega})$. This is an unconstrained minimization, exactly as in Widrow's classical problem [34]. We can implement

it by using the multichannel Wiener filter. Recalling (12), our goal is to set $\mathbf{G}(t, e^{j\omega})$ to minimize

$$E\left\{||Y_{\mathrm{FBF}}(t, e^{j\omega}) - \mathbf{G}^{\dagger}(t, e^{j\omega})\mathbf{U}(t, e^{j\omega})||^2\right\}.$$

Let

$$\Phi_{\mathbf{U}Y}(t, e^{j\omega}) = E\left\{\mathbf{U}(t, e^{j\omega})Y^{*}_{\mathrm{FBF}}(t, e^{j\omega})\right\}$$
$$\Phi_{\mathbf{UU}}(t, e^{j\omega}) = E\left\{\mathbf{U}(t, e^{j\omega})\mathbf{U}^{\dagger}(t, e^{j\omega})\right\}.$$

Then, the multichannel Wiener filter is given by [19], [35]

$$\mathbf{G}(t, e^{j\omega}) = \Phi^{-1}_{\mathbf{UU}}(t, e^{j\omega})\Phi_{\mathbf{U}Y}(t, e^{j\omega}). \tag{21}$$

In order to be able to track changes, we process the signals by segments. The following frequency domain LMS algorithm is used. Let the residual signal be

$$Y(t, e^{j\omega}) = Y_{\mathrm{FBF}}(t, e^{j\omega}) - \mathbf{G}^{\dagger}(t, e^{j\omega})\mathbf{U}(t, e^{j\omega}).$$

Note that the residual signal is also the output of the enhancement algorithm. By the orthogonality principle, the error is orthogonal to the measurements. Thus

$$E\left\{\mathbf{U}(t, e^{j\omega})Y^{*}(t, e^{j\omega})\right\} = 0. \tag{22}$$

Following the standard Widrow procedure, the solution is

$$\mathbf{G}(t+1, e^{j\omega}) = \mathbf{G}(t, e^{j\omega}) + \mu\mathbf{U}(t, e^{j\omega})Y^{*}(t, e^{j\omega}).$$

Usually, a more stable solution is achieved by using the normalized LMS (NLMS) algorithm, in which each frequency is normalized separately, yielding

$$G_m(t+1, e^{j\omega}) = G_m(t, e^{j\omega}) + \mu\frac{U_m(t, e^{j\omega})Y^{*}(t, e^{j\omega})}{P_{\mathrm{est}}(t, e^{j\omega})}$$
$$m = 2, \ldots, M$$

where

$$P_{\mathrm{est}}(t, e^{j\omega}) = \rho P_{\mathrm{est}}(t-1, e^{j\omega}) + (1-\rho)\sum_{m}\left|Z_m(t, e^{j\omega})\right|^2 \tag{23}$$

ρ is a forgetting factor (typically $0.8 < \rho < 1$). Another possibility is to calculate P_{est} using the power of the noise reference signals. However, in that case, an energy detector is required so that $\mathbf{G}(t, e^{j\omega})$ is updated only when there is no active signal. If on the other hand, we calculate $P_{\mathrm{est}}(t, e^{j\omega})$ using the input sensor signals, as indicated in (23); then, an energy detector may be avoided. This is due to the fact that the adaptation term becomes relatively small during periods of active input signal.

We assume that the noncasual TFs ratios $\mathbf{h}_m$ and the noise canceling filters $\mathbf{g}_m$ are both FIRs:

$$\mathbf{h}_m^T = [h_m(-q_L), \ldots, h_m(q_R)]$$
$$\mathbf{g}_m^T = [g_m(-K_L), \ldots, g_m(K_R)] \tag{24}$$

(both $\mathbf{h}_m$ and $\mathbf{g}_m$ are functions of time; however, for notational simplicity, we omit this dependence). Note that the TFs might have zeros outside the unit circle. Thus, to ensure stability of the TFs ratios, we do not impose them to be causal. When $A_1(e^{j\omega})$ contains zeros that are close to the unit circle, the noise reference signals $U_m(e^{j\omega})$ at the corresponding frequencies might assume very large values [recall (20)]. This may result in sharp peaks in the reconstructed spectrum. This problem is partially overcome by constraining the impulse response of $\mathbf{h}_m$ to an FIR structure. It is also possible to constrain the maximal value of the estimated $|H_m(e^{j\omega})|$ to be lower than some threshold.

In order to fulfill the FIR structure constraint (24), the filters update is now given by

$$\tilde{G}_m(t+1, e^{j\omega}) = G_m(t, e^{j\omega}) + \mu\frac{U_m(t, e^{j\omega})Y^{*}(t, e^{j\omega})}{P_{\mathrm{est}}(t, e^{j\omega})}$$
$$G_m(t+1, e^{j\omega}) \overset{\mathrm{FIR}}{\longleftarrow} \tilde{G}_m(t+1, e^{j\omega}) \tag{25}$$

for $m = 2, \ldots, M$. The operator $\overset{\mathrm{FIR}}{\longleftarrow}$ includes the following three stages. First, we transform $\tilde{G}_m(t+1, e^{j\omega})$ to the time domain. Second, we truncate the resulting impulse response to the interval $[-K_L, K_R]$ (i.e., we impose an FIR constraint). Third, we transform back to the frequency domain.

Note that the various filtering operations (multiplications in the transform domain) are realized using the overlap and save method [36].

The new algorithm can be regarded as an extension of the Griffiths and Jim algorithm for the general TF case. Figs. 3 and 4 summarize our suggested solution. The ratios of the TFs are assumed to be known at this stage.

IV. System Identification Using Nonstationarity

Thus far, we assumed that the TFs ratio vector $\mathbf{H}(e^{j\omega})$ is known. In practice, however, $\mathbf{H}(e^{j\omega})$ are not known and should be estimated. Rearranging terms in (20), we have

$$Z_m(t, e^{j\omega}) = H_m(e^{j\omega})Z_1(t, e^{j\omega}) + U_m(t, e^{j\omega}). \tag{26}$$

We have assumed that the TFs ratios are slowly changing in time compared to the time variations of the desired signal. We further assume that the statistics of the noise signal is slowly changing compared with the statistics of the desired signal. Consider some analysis interval during which both the TFs and the noise signal are assumed to be stationary. We divide that analysis interval into frames such that the desired signal may be considered stationary during each frame. Consider the kth frame. By (26), we have

$$\Phi^{(k)}_{z_m z_1}(e^{j\omega}) = H_m(e^{j\omega})\Phi^{(k)}_{z_1 z_1}(e^{j\omega}) + \Phi_{u_m z_1}(e^{j\omega})$$
$$k = 1, \ldots, K \tag{27}$$

where K is the number of frames used. $\Phi^{(k)}_{z_i z_j}(e^{j\omega})$ is the cross-PSD between z_i and z_j during the kth frame. $\Phi_{u_m z_1}(e^{j\omega})$ is the cross-PSD between u_m and z_1. Now, (2) and (20) imply that

$$U_m(t, e^{j\omega}) = N_m(t, e^{j\omega}) - H_m(e^{j\omega})N_1(t, e^{j\omega}) \tag{28}$$
$$Z_1(t, e^{j\omega}) = A_1(e^{j\omega})S(t, e^{j\omega}) + N_1(t, e^{j\omega}). \tag{29}$$

Since $N_m(t, e^{j\omega})$, $m = 1, \ldots, M$ are assumed stationary over the analysis interval and since $S(t, e^{j\omega})$ is independent of $N_m(t, e^{j\omega})$, it follows that $\Phi_{u_m z_1}(e^{j\omega})$ is independent of the frame index k.

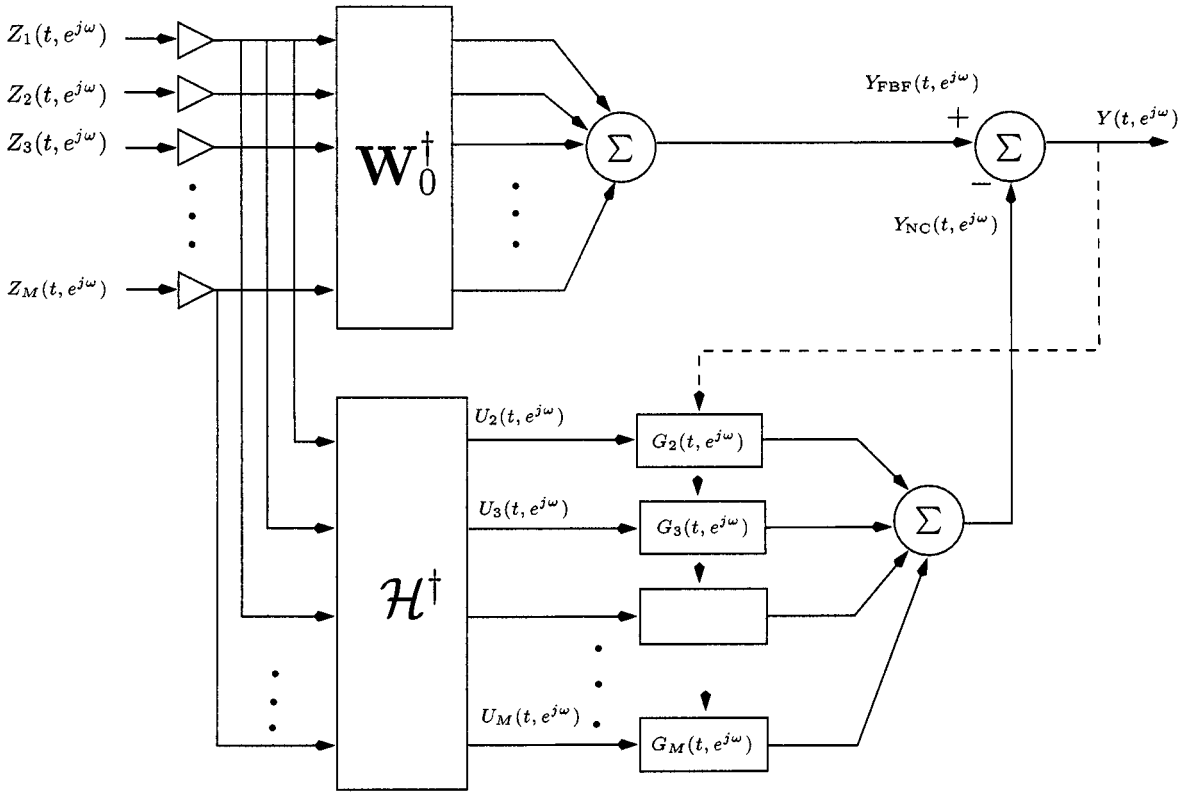

Fig. 3. Linearly constrained adaptive beamformer.

1) TF-s ratios: $\mathbf{H}(e^{j\omega}) = \frac{\mathbf{A}(e^{j\omega})}{A_1(e^{j\omega})}$
2) Fixed beamformer:
$Y_{\mathrm{FBF}}(t, e^{j\omega}) = \mathbf{W}_0^{\dagger}(e^{j\omega})\mathbf{Z}(t, e^{j\omega})$
3) Noise reference signals:
$\mathbf{U}(t, e^{j\omega}) = \mathcal{H}^{\dagger}(e^{j\omega})\mathbf{Z}(t, e^{j\omega})$
4) Output signal:
$Y(t, e^{j\omega}) = Y_{\mathrm{FBF}}(t, e^{j\omega}) - \mathbf{G}^{\dagger}(t, e^{j\omega})\mathbf{U}(t, e^{j\omega})$
5) Filters update, for $m = 1, \ldots, M-1$:
$\tilde{G}_m(t+1, e^{j\omega}) = G_m(t, e^{j\omega}) + \mu \frac{U_m(t, e^{j\omega}) Y^*(t, e^{j\omega})}{P_{\mathrm{est}}(t, e^{j\omega})}$
$G_m(t+1, e^{j\omega}) \xleftarrow{\mathrm{FIR}} \tilde{G}_m(t+1, e^{j\omega})$
where,
$P_{\mathrm{est}}(t, e^{j\omega}) = \rho P_{\mathrm{est}}(t-1, e^{j\omega}) + (1-\rho)\sum_m |Z_m(t, e^{j\omega})|^2$
6) keep only non-aliased samples.
(note: $\mathbf{W}_0(e^{j\omega})$ is defined in (16).
$\mathcal{H}(e^{j\omega})$ is defined in (19)).

Fig. 4. Suggested algorithm.

Let $\hat{\Phi}_{z_1 z_1}^{(k)}(e^{j\omega})$, $\hat{\Phi}_{z_m z_1}^{(k)}(e^{j\omega})$ and $\hat{\Phi}_{u_m z_1}^{(k)}(e^{j\omega})$ be estimates of $\Phi_{z_1 z_1}^{(k)}(e^{j\omega})$, $\Phi_{z_m z_1}^{(k)}(e^{j\omega})$ and $\Phi_{u_m z_1}(e^{j\omega})$, respectively. The estimates are obtained by replacing expectations with averages. Note that (27) also holds for the estimated values. Let $\varepsilon_m^{(k)}(e^{j\omega}) = \hat{\Phi}_{u_m z_1}^{(k)}(e^{j\omega}) - \Phi_{u_m z_1}(e^{j\omega})$ denote the estimation error of the cross-PSD between z_1 and u_m in the kth frame. We then have

$$\hat{\Phi}_{z_m z_1}^{(k)}(e^{j\omega}) = H_m(e^{j\omega})\hat{\Phi}_{z_1 z_1}^{(k)}(e^{j\omega}) + \Phi_{u_m z_1}(e^{j\omega}) + \varepsilon_m^{(k)}(e^{j\omega}).$$

If the noise reference signals $U_m(t, e^{j\omega})$, $m = 2, \ldots, M$ were uncorrelated with $Z_1(t, e^{j\omega})$, then the standard system identification estimate $H_m(e^{j\omega}) = \hat{\Phi}_{z_m z_1}(e^{j\omega})/\hat{\Phi}_{z_1 z_1}(e^{j\omega})$ could be used to obtain an unbiased estimate of $H_m(e^{j\omega})$. Unfortunately, by (28) and (29), $U_m(t, e^{j\omega})$ and $Z_1(t, e^{j\omega})$ are, in general, correlated. Hence, in [31], it is proposed that we obtain an unbiased estimate of $H_m(e^{j\omega})$ by applying least squares to the following set of overdetermined equations

$$\begin{bmatrix} \hat{\Phi}_{z_m z_1}^{(1)}(e^{j\omega}) \\ \hat{\Phi}_{z_m z_1}^{(2)}(e^{j\omega}) \\ \vdots \\ \hat{\Phi}_{z_m z_1}^{(K)}(e^{j\omega}) \end{bmatrix} = \begin{bmatrix} \hat{\Phi}_{z_1 z_1}^{(1)}(e^{j\omega}) & 1 \\ \hat{\Phi}_{z_1 z_1}^{(2)}(e^{j\omega}) & 1 \\ \vdots & \\ \hat{\Phi}_{z_1 z_1}^{(K)}(e^{j\omega}) & 1 \end{bmatrix} \begin{bmatrix} H_m(e^{j\omega}) \\ \Phi_{u_m z_1}(e^{j\omega}) \end{bmatrix} + \begin{bmatrix} \varepsilon_m^{(1)}(e^{j\omega}) \\ \varepsilon_m^{(2)}(e^{j\omega}) \\ \vdots \\ \varepsilon_m^{(K)}(e^{j\omega}) \end{bmatrix} \tag{30}$$

(a separate set of equations is used for $m = 2, \ldots, M$). The solution to (30) is given by

$$H_m(e^{j\omega}) = \frac{\left\langle \hat{\Phi}_{z_1 z_1}(e^{j\omega})\hat{\Phi}_{z_m z_1}(e^{j\omega}) \right\rangle - \left\langle \hat{\Phi}_{z_1 z_1}(e^{j\omega}) \right\rangle \left\langle \hat{\Phi}_{z_m z_1}(e^{j\omega}) \right\rangle}{\left\langle \hat{\Phi}_{z_1 z_1}^2(e^{j\omega}) \right\rangle - \left\langle \hat{\Phi}_{z_1 z_1}(e^{j\omega}) \right\rangle^2} \tag{31}$$

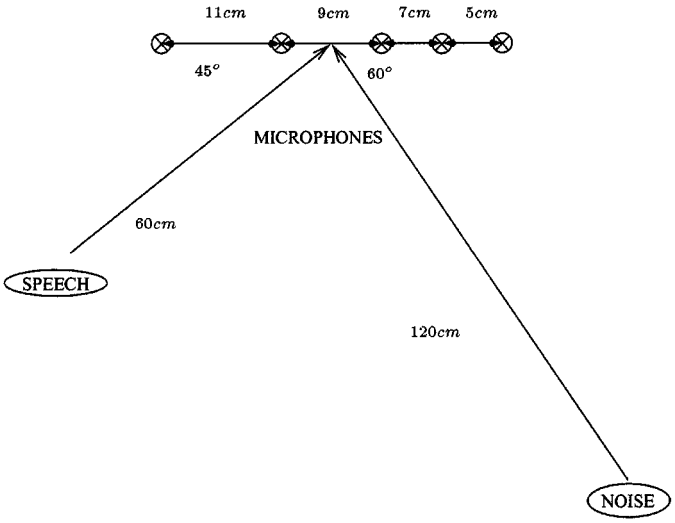

Fig. 5. Test scenario: Array of five microphones in a noisy conference room.

where for a given set of K values $\beta^{(k)}(e^{j\omega})$, we define the average operation

$$\left\langle \beta(e^{j\omega}) \right\rangle \triangleq \frac{1}{K} \sum_{k=1}^{K} \beta^{(k)}(e^{j\omega}).$$

Special attention should be given when choosing the frame length. On the one hand, it should be longer than the correlation length of $z_m(t)$, which must be longer than the length of the filter $a_m(t)$. On the other hand, it should be short enough for the quasistationarity assumption to hold.

V. Performance Evaluation

In this section, we apply the suggested algorithm to the speech enhancement problem and evaluate its performance. The scenario shown in Fig. 5 was studied. The enclosure is a conference room with dimensions 5 m × 4 m × 2.8 m. A linear array was placed on a table at the center of the room. Two loudspeakers were used: one for the speech source and the other for the noise source. The locations are marked in Fig. 5. The impulse response and frequency response between the speech source and the first microphone are depicted in Fig. 6. This response was obtained using a least squares fit between the input signal source and the received microphone signal (the response includes the loudspeaker). We note that in all our experiments, we used the actual recordings and did not use the estimated impulse responses. Let the *energy decay curve* (EDC) corresponding to some impulse response $a(t)$ be defined by [29]

$$\mathrm{EDC}(t) \triangleq \sum_{\tau=t}^{\infty} a^2(\tau).$$

The point where the EDC slope changes abruptly is called *total duration* (TD). The *clarity index* is defined by

$$C(a) \triangleq \frac{\mathrm{EDC}(t=0)}{\mathrm{EDC}(t=\mathrm{TD})}.$$

In Fig. 6, we also show the EDC of the impulse response between the speech source and the first microphone. The corresponding clarity index is 6.7 dB, which indicates a reverberated environment [29].

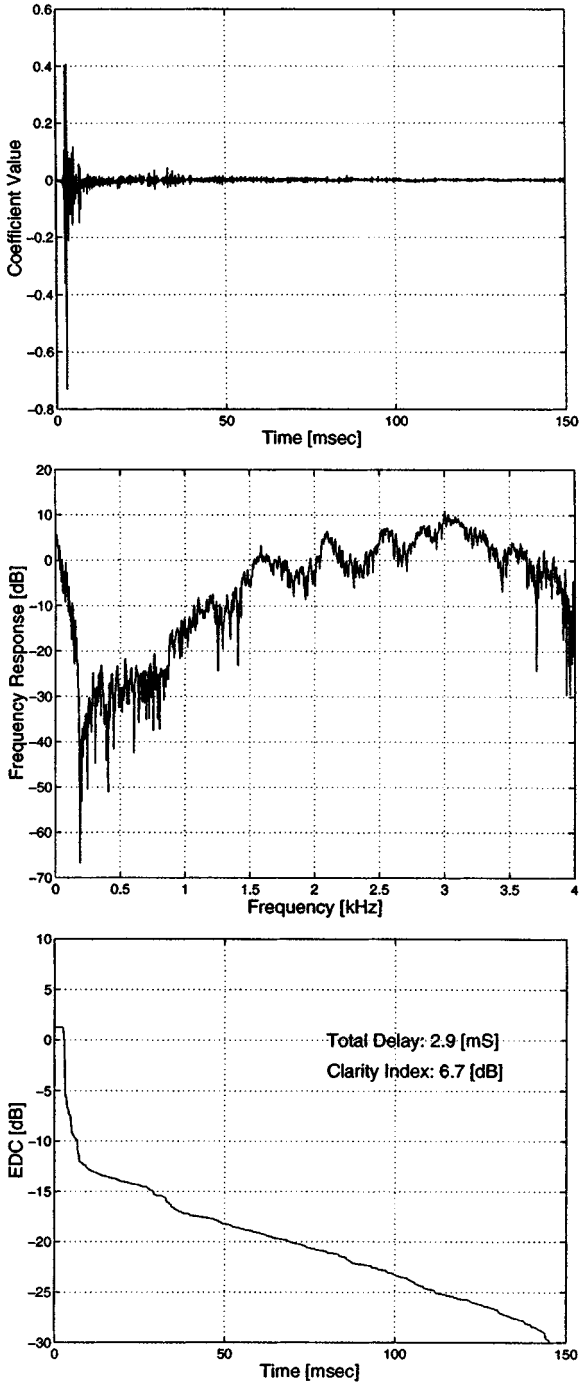

Fig. 6. (Top) Impulse response, (middle) frequency response, and (bottom) EDC of the TF between the speech loudspeaker and the first microphone.

The speech source was comprised of four TIMIT sentences with various gain levels. The input microphone signals were generated by mixing speech and noise components that were created separately at various SNR levels. We considered two test scenarios. The first was speech contaminated by a point noise source. The second was speech contaminated by a diffused noise source. In order to generate the speech component of the microphone signals, we transmitted the four sentences through a loudspeaker and recorded the resulting microphone signals. In

TABLE I
BLOCKING ABILITY FOR POINT SOURCE (TOP) AND DIFFUSED NOISE (BOTTOM) IN DECIBELS. FIVE MICROPHONES

Input SNR	TF-GSC SNR	D-GSC SNR
-4.5	-9.1	0.5
-1.5	-8.9	3.5
1.5	-7.8	5.8
4.5	-5.8	-2
7.5	-3.4	2.5
10.5	-0.6	5.0
13.5	2.3	8.1
16.5	5.3	10.8
20.4	8.2	14.0

Input SNR	TF-GSC SNR	D-GSC SNR
-4.6	-12.0	-9.2
-1.9	-10.7	-6.2
1.8	-8.7	-3.2
4.4	-6.2	0.1
7.4	-3.4	2.7
10.4	-0.5	5.8
13.4	2.3	8.2
16.5	5.3	11.2
19.4	8.2	14.2

order to generate the point noise source, we transmitted an actual recording of fan noise (lowpass spectrum) through another loudspeaker. The diffused noise source was generated by simulating an omnidirectional emittance of a white noise signal [14].

In our experiments, the noise canceler (NC) block was always active. As was noted earlier, this is due to the fact that in (23), we used the input signals, and not the noise reference signals, in order to calculate $P_{\text{est}}(t, e^{j\omega})$. Hence, a voice activity detector (VAD) was not necessary.

The blocking filters $\mathbf{h}_m$ were modeled by noncausal FIRs with 181 coefficients in the interval $[-90, 90]$. The canceling filters $\mathbf{g}_m$ were modeled by noncausal FIRs with 251 coefficients in the interval $[-125, 125]$. In order to implement the overlap and save procedure, segments with 512 samples were used. The system identification procedure utilized 13 segments. The length of each segment was 1000-samples (sampling rate was 8 kHz). We note that system identification was applied only during active speech periods. However, an accurate VAD is not necessary for this purpose. We have also implemented the standard (delay only) GSC algorithm. The algorithm was implemented in the time domain. In order to estimate the delays we used a cross correlation criterion that was also applied only during active speech periods. Essentially, there was no difference in the performance when using integer or fractional delays. The noise canceler filters were realized using the same length as in our implementation of the new suggested algorithm.

In Table I, we assess the ability of the blocking matrix (BM) to generate noise-only reference signals. For each input SNR value, we evaluated the SNR of the reference signals both for the standard (delays only) GSC (hereby designated as D-GSC) and for the new proposed algorithm (hereby designated as TF-GSC). A high SNR value indicates that there is a high leakage of speech to the noise reference, and hence, the resulting output is expected to be reverberated due to self cancellation. As can be seen, the quality of the noise reference produced by the TF-GSC algorithm is better than that produced by the D-GSC algorithm.

TABLE II
OUTPUT SNR AND NOISE REDUCTION (NR) FOR POINT SOURCE (TOP) AND FOR DIFFUSED SOURCE (BOTTOM) IN DECIBELS. FIVE MICROPHONES

In SNR	D-GSC SNR	TF-GSC SNR	D-GSC NR	TF-GSC NR
-6.4	-3.7	-0.8	6.5	6.8
-3.4	-2.4	3.3	6.3	8.4
-0.4	0.6	6.8	11.8	10.0
2.6	2.4	9.3	10.0	11.4
5.6	2.6	11.1	9.5	10.6
8.6	3.2	11.7	9.0	9.1
11.6	3.2	12.0	8.2	7.8
14.6	3.3	12.3	7.2	6.7
17.6	3.3	12.5	6.0	5.7

In SNR	D-GSC SNR	TF-GSC SNR	D-GSC NR	TF-GSC NR
-6.5	-3.6	-3.7	3.7	3.0
-3.5	-1.4	-0.6	3.7	3.2
-0.5	0.3	2.2	3.6	3.2
2.5	1.6	4.9	3.5	3.2
5.5	2.4	7.3	3.3	3.2
8.5	2.8	9.5	3.2	3.2
11.5	3.1	11.1	2.3	3.2
14.5	3.3	12.1	2.1	3.0
17.5	3.3	12.7	1.9	2.8

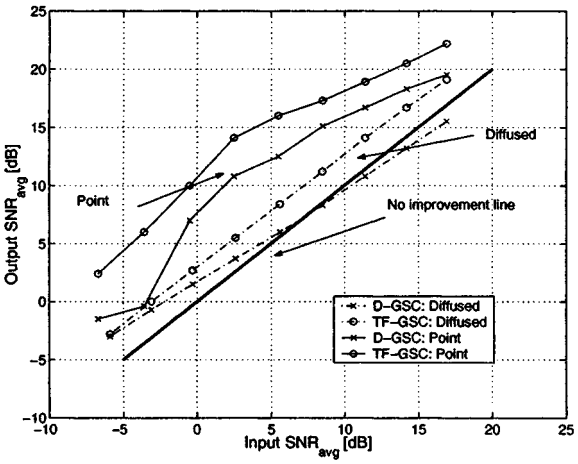

Fig. 7. Averaged SNR improvement for point noise source and diffused noise source.

This holds both for the point noise source and for the diffused noise source. As can be seen, for low SNR inputs, the blocking ability of the D-GSC algorithm is poor. This is due to the fact that for such SNR values, the delay estimation routine of the D-GSC algorithm collapses.

In order to evaluate and compare the performance of the algorithms, we used three objective quality measures. The first is signal to noise ratio (SNR) defined by

$$\text{SNR} \triangleq \frac{\sum_{t\in T_s} z_{1,s}^2(t)}{\sum_{t\in T_s}(z_{1,s}(t) - Ky(t))^2}$$

where $z_{1,s}(t)$ is the signal component recorded by the first microphone. $y(t)$ is the algorithm output (reconstructed speech

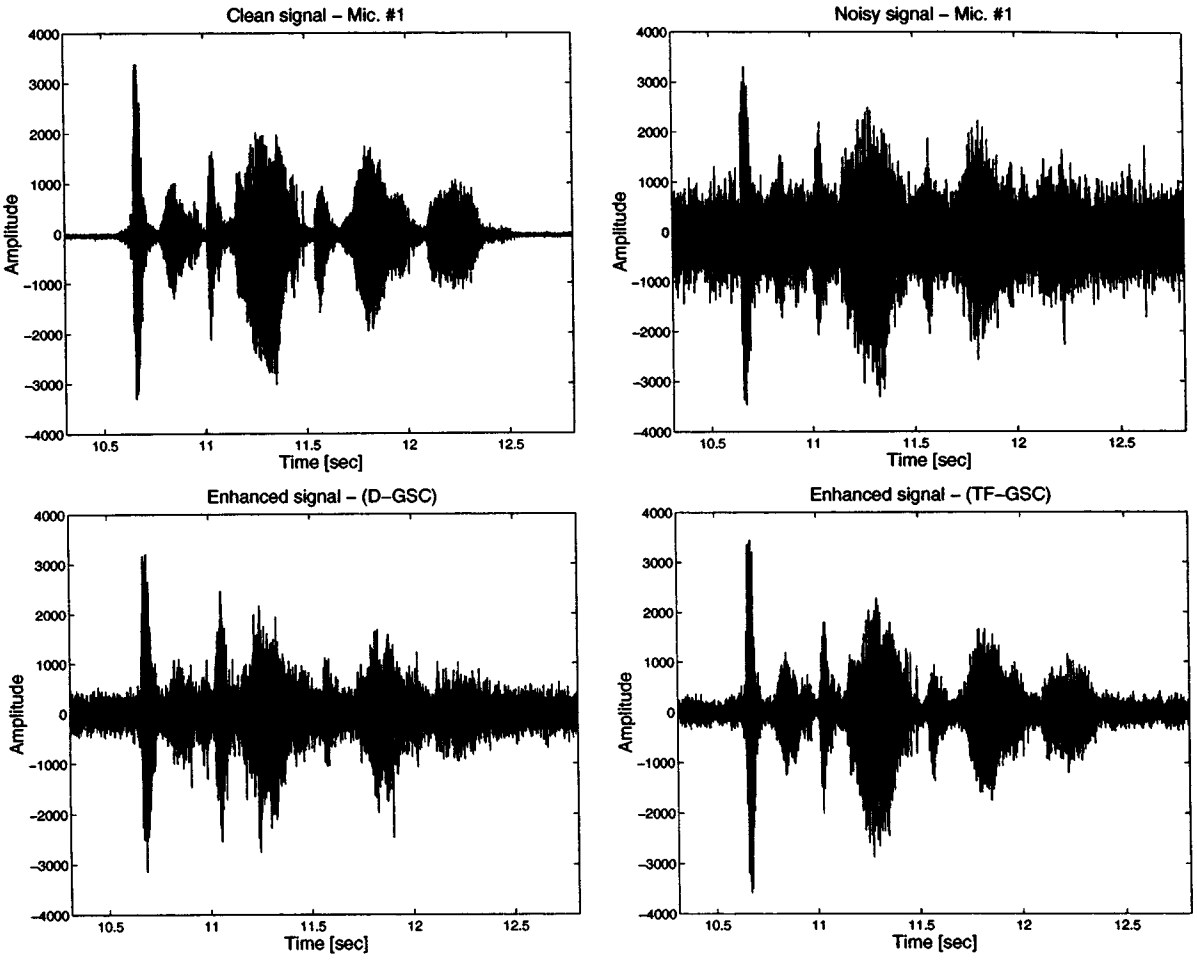

Fig. 8. Speech waveforms: Clean microphone #1. Noisy and enhanced (D-GSC, TF-GSC).

signal). T_s denotes periods in time where the speech signal is active. K is a gain factor that compensates for possible gain level variations of the signals. In addition to that, $z_{1,s}(t)$ and $y(t)$ are time aligned.

The second quality measure is noise reduction (NR), which is defined by

$$\text{NR} \triangleq \frac{\sum_{t\in T_n} (Ky(t))^2}{\sum_{t\in T_n} z_1^2(t)}$$

where T_n denotes periods in time where the speech signal is inactive. The quality measure NR compares the noise level in the reconstructed speech to the noise level recorded by the first microphone. Table II summarizes the SNR and NR values in decibels when using the D-GSC and TF-GSC algorithms. While the noise reduction ability of both algorithms is comparable, the SNR level achieved by the TF-GSC is much higher. These observations indicate that TF-GSC is characterized by a significantly lower speech distortion compared with D-GSC while keeping the same level of noise reduction. In the high input SNR region, although the algorithm degrades the SNR measure, it results in an overall enhanced output. This is due to the fact that it reduces the noise level. Finally, comparing our results for the two noise sources, it can be seen that the SNR and NR values of both algorithms are higher for the point noise source case (except for the SNR measure of D-GSC). This is due to the low coherence function in the diffused noise case, which degrades the performance of the noise canceling block of the algorithm [14].

The third quality measure is the averaged SNR, SNR_{avg}. Given some signal $x(t)$, SNR_{avg} is defined by

$$\text{SNR}_{\text{avg}} \triangleq \frac{\sum_{t\in T_s} x^2(t) - \sum_{t\in T_n} x^2(t)}{\sum_{t\in T_n} x^2(t)}. \tag{32}$$

This quality measure compares the signal energy in $x(t)$ to the noise energy. Fig. 7 shows SNR_{avg} in decibels both for D-GSC and TF-GSC for both noise types (point and diffused). As can be seen, TF-GSC yields higher values of SNR_{avg}. The SNR_{avg} values of both algorithms are higher for the point noise source case.

Fig. 8 shows the waveforms of the speech component recorded by the first microphone, the noisy speech at the first microphone, and the enhanced speech for both D-GSC and TF-GSC algorithms. The noise signal used was a point source at an SNR level of 0 dB. Fig. 9 shows sonograms of the same data. It can be seen that the TF-GSC algorithm produces an

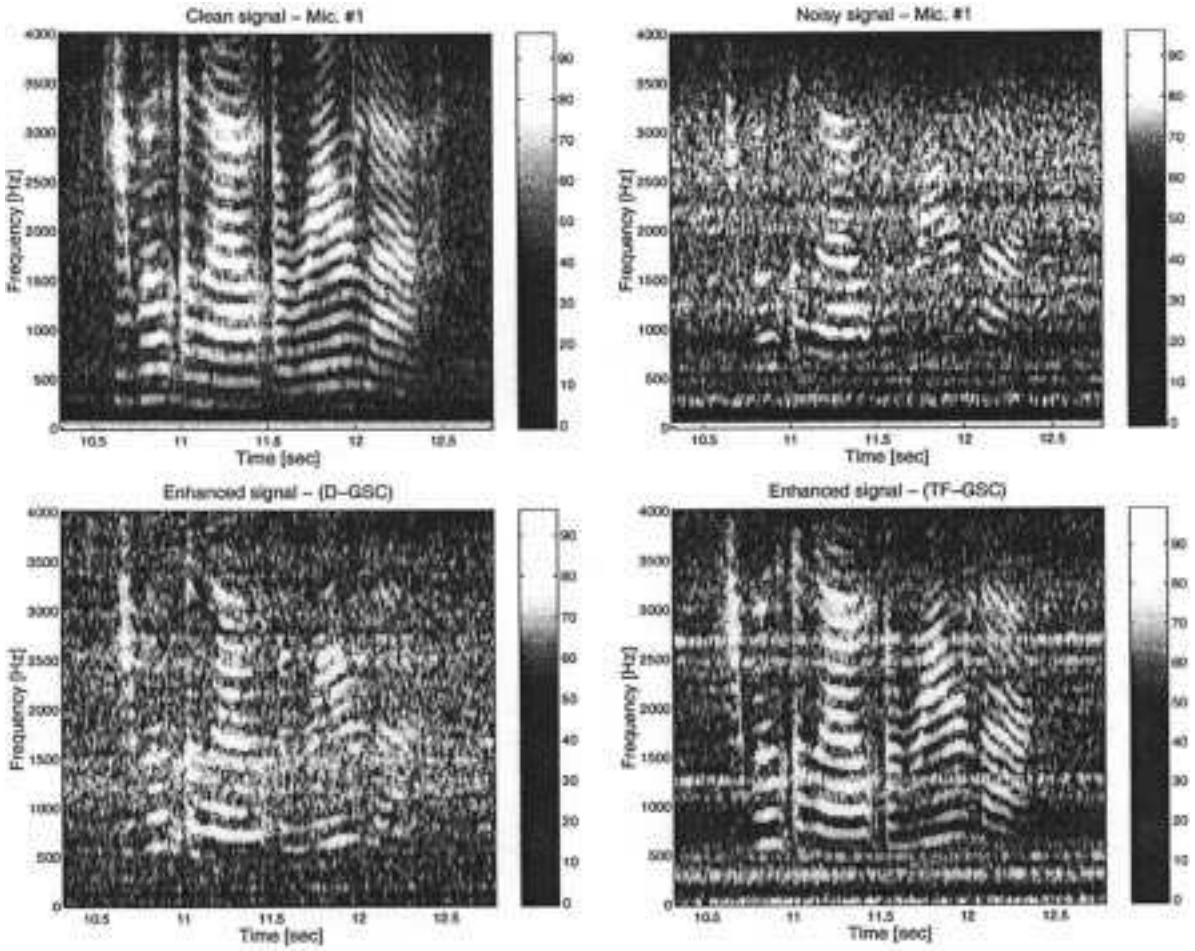

Fig. 9. Sonograms: Clean Microphone #1. Noisy and enhanced (D-GSC, TF-GSC).

enhanced speech signal with higher noise reduction and lower distortion. The residual noise that corresponds to the horizontal lines in the TF-GSC sonogram does not create an unnatural sound effect. Moreover, when adding a single channel speech enhancement post-processing device, this residual noise is significantly reduced. We note that when the SNR increases, these lines completely disappear.

To further assess the output speech quality, we have conducted informal listening evaluations. All our listeners clearly indicated impressive noise reduction without any noticeable distortion for the TF-GSC algorithm. On the other hand, the D-GSC algorithm was classified as reverberated. This is due to self-cancellation, which is caused by leakage of the desired signal into the noise reference.

All our algorithms were implemented without a VAD in the noise canceling block. When a VAD is incorporated, there is no significant change in the performance of the TF-GSC algorithm. However, there is an improvement in the performance of the D-GSC, as noted in [37]. Even so, the quality of the enhanced speech produced by the TF-GSC algorithm is significantly higher than that produced by D-GSC.

In order to further improve the performance, we applied a single microphone speech enhancement algorithm [38] on the output of the multimicrophone speech enhancement algorithm (i.e., the single microphone algorithm was used in a postprocessing stage). Postprocessing yields a further improvement of about 10dB both in $\mathrm{SNR}_{\mathrm{avg}}$ and NR. It also results in a small improvement of about 1–2 dB in the SNR measure. Subjective listening evaluations confirm these improvements.

VI. Discussion

The suggested algorithm can be applied for enhancing an arbitrary nonstationary signal corrupted by stationary noise. An arbitrary TF and array geometry can be used. The use of TFs ratio rather than the TFs themselves (which is the counterpart of relative delay in delay-only arrays) improves the efficiency and robustness of the algorithm since shorter filters can be used. This might be due to pole-zero cancellation in the TFs ratio.

Although our algorithm was implemented in the frequency domain, it can also be implemented in the time domain. This applies both to the adaptive beamformer stage and to the system identification stage. Both versions of the algorithm yield comparable performance. However, the computational burden of the frequency domain algorithm is significantly smaller than that of the time domain version. In our (probably inefficient) MATLAB© implementation only three times real time was required (on our ultra 5 SUN workstation). There are two reasons for this. First, the system identification in the frequency domain involves only a 2×2 matrix inversion

for each frequency bin examined (in the time domain, it is required to invert a matrix whose order is the dimension of the desired filter). Second, the frequency domain system identification need not be implemented for frequency bands with too low-level speech signal components.

Although results are presented for the five-microphone case, the algorithm was also useful when a smaller number of microphones (e.g., two) were used.

In this paper, we have assumed that the noise is nonstationary. Sometimes, this assumption is not accurate (e.g., for a cocktail party noise). Nevertheless, whenever the noise is "more stationary" compared with the desired speech signal, the estimation method presented in Section IV is expected to be useful.

In order to use the proposed algorithm, one needs to re-estimate the TFs once the acoustic environment has changed. In order to reduce the computational complexity, recursive procedures [e.g., RLS methods for solving (30)] may be incorporated. This is left as a further research topic.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their comments that helped to improve the presentation of the paper.

REFERENCES

[1] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, pp. 926–935, Jan. 1972.

[2] 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.

[3] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoust., Speech, Signal Processing Mag.*, vol. 5, pp. 4–24, Apr. 1988.

[4] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-35, pp. 1365–1376, Oct. 1987.

[5] O. Hoshuyama and A. Sugiyama, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, Atlanta, GA, May 1996, pp. 925–928.

[6] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive microphone array with improved spatial selectivity and its evaluation in a real environment," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Munich, Germany, Apr. 1997, pp. 367–370.

[7] ——, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Processing*, vol. 47, pp. 2677–2684, Oct. 1999.

[8] S. Nordholm, I. Claesson, and B. Bengtsson, "Adaptive array noise suppression of handsfree speaker input in cars," *IEEE Trans. Veh. Technol.*, vol. 42, pp. 514–518, Nov. 1993.

[9] J. Meyer and C. Sydow, "Noise cancelling for microphone array," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Munich, Germany, Apr. 1997, pp. 211–214.

[10] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1985, ch. 13, p. 393.

[11] M. Dahl, I. Claesson, and S. Nordebo, "Simultaneous echo cancellation and car noise suppression employing a microphone array," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Munich, Germany, Apr. 1997, pp. 239–242.

[12] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1988, pp. 2578–2581.

[13] J. Meyer and K. U. Simmer, "Multichannel speech enhancement in a car environment using Wiener filtering and spectral subtraction," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Munich, Germany, Apr. 1997.

[14] N. Dal-Degan and C. Prati, "Acoustic noise analysis and speech enhancement techniques for mobile radio application," *Signal Process.*, vol. 15, no. 4, pp. 43–56, Jul. 1988.

[15] S. Fischer and K. D. Kammeyer, "Broadband beamforming with adaptive postfiltering for speech acquisition in noisy environment," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Munich, Germany, 1997, pp. 359–362.

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

[17] J. Bitzer, K. U. Simmer, and K. D. Kammeyer, "Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Phoenix, AZ, May 1999.

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

[19] J. Bitzer, K.-D. Kammeyer, and K. U. Simmer, "An alternative implementation of the superdirective beamformer," *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 1999.

[20] E. E. Jan and J. Flanagan, "Sound capture from spatial volumes: Matched-filter processing of microphone arrays having randomly-distributed sensors," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Atlanta, GA, 1996, pp. 917–920.

[21] D. Rabinkin, R. Renomeron, J. Flanagan, and D. F. Macomber, "Optimal truncation time for matched filter array processing," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Seattle, WA, May 1998, pp. 3269–3273.

[22] S. Affes, "Formation de voie adaptive en milieux réverbérants," Ph.D. dissertation, Ecole Nat. Supér. Télécommun., Paris, France, Oct. 1995.

[23] Y. Grenier, "A microphone array for car environments," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, San Francisco, CA, 1992, pp. 305–308.

[24] S. Affes, S. Gazor, and Y. Grenier, "Robust adaptive beamforming via LMS-like target tracking," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Adelaide, Australia, Apr. 1994, pp. 269–272.

[25] S. Gazor, S. Affes, and Y. Grenier, "Wideband multisource beamforming with adaptive array location calibration and direction finding," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1995, pp. 1904–1907.

[26] ——, "Robust adaptive beamforming via target tracking," *IEEE Trans. Signal Processing*, vol. 44, pp. 1589–1593, June 1996.

[27] S. Affes and Y. Grenier, "A source subspace tracking array of microphones for double talk situations," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Munich, Germany, Apr. 1996, pp. 269–272.

[28] S. Affes, S. Gazor, and Y. Grenier, "An algorithm for multisource beamforming and multitarget tracking," *IEEE Trans. Signal Processing*, vol. 44, pp. 1512–1522, Jun. 1996.

[29] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 425–437, Sept. 1997.

[30] B. Yang, "Projection approximation subspace tracking," *IEEE Trans. Signal Processing*, vol. 43, pp. 95–107, Jan. 1995.

[31] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Trans. Signal Processing*, vol. 44, pp. 2055–2063, Aug. 1996.

[32] D. H. Brandwood, "A complex gradient operator and its application in adaptive array theory," *Proc. Inst. Elect. Eng. F*, vol. 130, no. 1, pp. 11–16, Feb. 1983.

[33] G. Strang, *Linear Algebra and Its Application*, 2nd ed: Academic, 1980.

[34] B. Widrow *et al.*, "Adaptive noise cancelling: Principals and applications," *Proc. IEEE*, vol. 63, pp. 1692–1716, Dec. 1975.

[35] 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.

[36] R. E. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 99–102, Feb. 1980.

[37] D. van Compernolle, W. Ma, and F. Xie, "Speech recognition in noisy environments with the aid of microphone arrays," *Elsevier Speech Commun.*, vol. 9, pp. 433–442, 1990.

[38] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," in *Proc. 6th Euro. Conf. Speech Commun. Tech.*, vol. 6, Budapest, Hungary, Sept. 1999, pp. 2591–2594.

Sharon Gannot (S'95) received the B.Sc. degree (summa cum laude) from the Technion—Israeli Institute of Technology, Haifa, in 1986 and the M.Sc. (cum laude) and Ph.D. degrees from Tel-Aviv University, Tel-Aviv, Israel, in 1995 and 2000 respectively, all in electrical engineering.

Between 1986 and 1993, he was Head of the research and development section in the Israeli Defense Forces. Currently, he holds a post doctoral position at the Department of Electrical Engineering (SISTA), Katholieke Universiteit Leuven, Leuven, Belgium. His research interests include parameter estimation, statistical signal processing, and speech processing using either single or multimicrophone arrays.

David Burshtein (M'92–SM'99) received the B.Sc. and Ph.D. degrees in electrical engineering in 1982 and 1987, respectively, from Tel-Aviv University, Tel-Aviv, Israel.

From 1988 to 1989, he was a Research Staff Member with the Speech Recognition Group, IBM, T. J. Watson Research Center, Yorktown Heights, NY. In 1989, he joined the Department of Electrical Engineering—Systems, Tel-Aviv University. His research interests include information theory and speech and signal processing.

Ehud Weinstein (F'94) received the B.Sc. degree from the Technion—Israel Institute of Technology, Haifa, in 1975 and the Ph.D. degree from Yale University, New Haven, CT, in 1978, both in electrical engineering.

In 1980, he joined the Department of Electrical Engineering—Systems, Tel-Aviv University, Tel-Aviv, Israel, where he is a Professor. From 1989 to 1992, he served as the department chairman. He has been a research affiliate at the Research Laboratory of Electronics (RLE), Massachusetts Institute of Technology (MIT), Cambridge, since 1990, where he was on sabbatical from 1989 to 1990. He has also been an Adjunct Scientist and a Visiting Investigator at the Department of Applied Ocean Physics and Engineering, Woods Hole Oceanographic Institution, Woods Hole, MA, since 1978. He has published over 50 papers and holds several patents in the area of signal processing and digital communications. He was a co-founder and the Chairman and CEO of Libit Signal Processing Ltd., which was founded in 1994 and acquired by Texas Instruments (TI) in 1999.

Prof. Weinstein received several awards and special recognitions for his scientific contributions. He was awarded Senior Fellow of TI for his various scientific contributions and for his leadership role in establishing Libit and bringing its cable modem technology to market acceptance.