Convolutive Transfer Function Generalized Sidelobe Canceler

Ronen Talmon, Israel Cohen, *Senior Member, IEEE*, and Sharon Gannot, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a convolutive transfer function generalized sidelobe canceler (CTF-GSC), which is an adaptive beamformer designed for multichannel speech enhancement in reverberant environments. Using a complete system representation in the short-time Fourier transform (STFT) domain, we formulate a constrained minimization problem of total output noise power subject to the constraint that the signal component of the output is the desired signal, up to some prespecified filter. Then, we employ the general sidelobe canceler (GSC) structure to transform the problem into an equivalent unconstrained form by decoupling the constraint and the minimization. The CTF-GSC is obtained by applying a convolutive transfer function (CTF) approximation on the GSC scheme, which is a more accurate and a less restrictive than a multiplicative transfer function (MTF) approximation. Experimental results demonstrate that the proposed beamformer outperforms the transfer function GSC (TF-GSC) in reverberant environments and achieves both improved noise reduction and reduced speech distortion.**

***Index Terms*—Adaptive signal processing, array signal processing, beamforming, generalized sidelobe canceler (GSC), linearly constrained minimum variance (LCMV), microphone arrays, minimum variance distortionless response (MVDR), noise reduction, speech enhancement.**

I. Introduction

THE problem of speech enhancement and noise reduction using a sensor array has been an active area of research for many years. In reverberant environments, the signal acquired by the microphone array is distorted by the room impulse response and usually contaminated by noise. Beamforming techniques, which aim at recovering the desired source signal from the reverberant and noisy output of the sensors, have attracted most of the research efforts. The basic idea of such beamformers is to incorporate spatial and spectral information in order to form a beam and point it to a desired direction. As a result, signals from this look direction are reinforced, while signals from all the other directions are attenuated. Several criteria can be applied in the design of a beamformer, among them the most common are the linearly constrained minimum variance (LCMV), and the minimum variance distortionless response (MVDR), where the latter can be considered as a special case of the former [1] (MVDR beamformer is an LCMV beamformer with just a single look direction constraint).

Frost proposed an MVDR adaptive beamformer [2], which reduces the background noise by applying a constrained minimization on the total output power under a look direction constraint. Griffiths and Jim [3] introduced the generalized sidelobe canceler (GSC), which transforms the MVDR beamformer into an unconstrained form by decoupling the constraint and the output minimization. Therefore, the GSC beamformer is equivalent to the MVDR beamformer [4], but it is characterized by two advantages. First, the unconstrained algorithm is computationally more efficient than the constrained algorithm. Second, it can be implemented using a standard normalized least mean square (NLMS) adaptive scheme [5]. The GSC structure comprises of three blocks. The first block is a fixed beamformer, which is designed to satisfy the constraint. The second block is a blocking matrix, which blocks the desired signal and produces noise-only reference signals. The third block is an unconstrained adaptive algorithm (e.g., the least mean square (LMS) algorithm) that aims at canceling the residual noise at the fixed beamformer output (the first block) given the noise-only reference signals at the output of the blocking matrix (the second block).

Both of the above beamformers are designed to steer the beam towards a single direction of the desired source location, while minimizing the response in all other directions. The main drawback is that a single direction of arrival cannot be determined in reverberant environments since reflections from different directions are also captured by the sensor array. Thus, in order to handle reverberant environments, beamformers often require estimates of the acoustic impulse response (AIR) of the desired source in order to model the propagation of the speech signal more accurately than simply as delay and attenuation. In practice, the desired source AIRs are unknown and difficult to acquire. Thus, suboptimal solutions, aiming at noise reduction rather than estimation of the desired source signal, were developed using estimates of the relative transfer functions (RTFs), which represent the coupling between pairs of sensors with respect to the desired source [6]–[10]. Among these solutions is the time domain MVDR beamforming technique, recently proposed by Chen *et al.* [10]. Theoretically, this approach yields an optimal solution for the problem at hand. In practice, due to computational complexity considerations, the RTFs are modeled as short filters, while RTFs in typical reverberant rooms are long. This degrades the noise reduction and increases the speech distortion.

Gannot *et al.* [6] have proposed a transfer function GSC (TF-GSC), which exploits the RTFs in the time–frequency domain. The TF-GSC is an efficient solution, with low computational

Manuscript received October 07, 2008; revised March 22, 2009. Current version published July 31, 2009. This work was supported by the Israel Science Foundation under Grant 1085/05. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nakatani Tomohiro.

R. Talmon and I. Cohen are with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: ronenta2@techunix.technion.ac.il; icohen@ee.technion.ac.il).

S. Gannot is with the School of Engineering, Bar-Ilan University, Ramat-Gan, 52900, Israel (e-mail: gannot@eng.biu.ac.il).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2020891

requirements, that achieves improved enhancement of multichannel noisy speech signals compared to the original GSC. One of the main challenges in this approach is the RTF representation and estimation, since the duration of the relative impulse response in reverberant environments may reach several thousand taps, as previously noted. In order to deal with this challenge, the TF-GSC method is based on a multiplicative transfer function (MTF) approximation [11], which is a common approach in time–frequency domain solutions. The MTF approximation enables to replace the linear convolution in the time domain with a scalar multiplication in each subband in the short-time Fourier transform (STFT) domain.[1] Clearly, this approximation becomes more accurate when the length of the time frame increases, relative to the length of the impulse response. However, long time frames may increase the estimation variance, increase the computational complexity and latency and restrict the ability to track temporal changes in the RTF [11]. Hence, in practice, short frames are used resulting in inaccurate representation of the RTF in reverberant environments. The TF-GSC scheme, which is based on inaccurate RTF representation imposed by the MTF approximation, does not meet the MVDR criterion, since both the power minimization and the constraint of the TF-GSC are formed inaccurately, resulting in degraded performance, especially in highly reverberant environments. It is worthwhile noting that several studies avoid this problem of RTF representation in signal cancellation methods. For example, in [12] the authors proposed a three-block structure similar to the GSC, where the blocking matrix is modified to operate adaptively, circumventing an explicit use of the RTF. A comprehensive comparison between these different methods for addressing the leakage problem is presented in [13].

Recently, we introduced an RTF identification method that relies on a convolutive transfer function (CTF) approximation in the STFT domain [14]. The CTF approximation enables to independently adjust the length of the analysis window and the lengths of filters in subbands, yielding advantageous RTF identification method compared to that which relies on the MTF approximation [15]. The representation of long responses using short time frames has major advantages. On the one hand, in reverberant environments the RTF convey large amount of reverberations, thus representing the RTF as a long filter is more suitable. On the other hand, since the input signal used for the RTF identification is of finite length to enable tracking of time variations, using short time frames increases the number of observation in each subband, which may improve the estimation [16]. Furthermore, when identifying the RTF [14], [15], power spectral density (PSD) terms of the speech signal in each time frame are used. Thus, we implicitly assume that the speech is stationary in each time frame. Clearly, this assumption becomes more accurate and the speech signal is better represented as the time frames become shorter. We also point out that by using longer time frames or longer subband filters, we pay with higher latency and computationally less efficient processing. For further details regarding system representation and identification under the MTF and CTF models and for an analysis of the influence of time frame length, we refer the readers to [11] and [16].

We focus our work on the commonly used STFT representation of signals and linear convolutions in the time–frequency domain [17]. This particular choice was encouraged by recent studies on system representation in the STFT domain [11], [16]. However, alternative transformations and filter-banks can be considered for representing a linear convolution in the time–frequency domain [18], and the presented solution can easily be adjusted to support them. Among these schemes, the *block partitioned convolution* [19] is of particular interest since it takes similar shape and enables the same desired property as the CTF approximation: to decouple the length of the time frame (or the frequency resolution) and the length of the convolved filter. The block partitioned convolution implies that the MTF approximation is employed in some block frequency-domain representation, while the CTF model is derived by taking the band-to-band filter of the complete linear convolution representation in the STFT domain.

In this paper, we propose a convolutive TF-GSC, which can be utilized in reverberant environments. Using a complete system representation in the STFT domain without any approximation [16], [20], [21], we formulate a constrained minimization problem [22], and employ the TF-GSC structure to transform the problem into an equivalent unconstrained form. The minimization and constraint are decoupled into two parallel processing branches, by uniquely decomposing the linear space into a speech component subspace and a noise component subspace in the STFT domain. We present explicit expressions for building each of the three GSC building blocks, using only the RTFs between the microphones with respect to the speech source.

Implementing the proposed GSC scheme under a complete system representation in the STFT domain is inefficient due to the high computational complexity requirements of the model. Thus, we apply approximations in order to obtain practical solutions. We show that by applying the MTF approximation on the GSC general structure in the STFT domain, the building blocks terms are reduced exactly to the building blocks of the TF-GSC. Hence, we get a twofold result. First, we prove that the TF-GSC, which was developed entirely under the MTF model, is equivalent to the MVDR beamformer represented in the STFT domain under the MTF approach. Second, the proposed framework provides tools to evaluate the inaccuracy involved in applying such an approximation.

We present a practical solution by applying the CTF approximation on the GSC complete structure. The CTF approximation enables representation of long impulse responses using short time frames, and thus, it does not suffer from the MTF model limitations and becomes especially advantageous in reverberant environments. By applying the CTF approximation on the explicit expressions of the GSC building blocks, we obtain an efficient solution, which merely requires estimates of the RTFs under the CTF model [14]. Experimental results demonstrate that the proposed beamformer outperforms the TF-GSC method in reverberant environments as long as the SNR is sufficiently high. In particular, we show that in a reverberant environment the new approach achieves both improved noise reduction and reduced speech distortion at the beamformer output.

[1]Zero padding of the STFT analysis window is required in order to avoid circular convolution resulting from the scalar multiplication, obtaining larger frequency resolution than the length of each time frame. Alternatively, one can consider such an analysis window shape with values close to zero at its right side to create equivalent effect.

This paper is organized as follows. In Section II, we formulate the problem in the STFT domain. In Section III, we develop an MVDR beamformer as a constrained optimization problem. In Section IV, we transform the problem into an unconstrained form using the GSC structure. In Section V, we introduce approximate representations and propose a practical beamformer based on the CTF model. Finally, in Section VI, we present experimental results that demonstrate the advantage of the proposed beamformer.

II. Problem Formulation

The problem considered in this work is an array of M microphones in a noisy and reverberant environment, where we have a single speech source located inside the enclosure. The output of the mth microphone is given by

$$\begin{aligned} y_m(n) &= a_m(n) * s(n) + u_m(n), m = 1, 2, \ldots, M \\ &\triangleq d_m(n) + u_m(n) \end{aligned} \tag{1}$$

where * denotes convolution, $s(n)$ represents a (nonstationary) speech source, $a_m(n)$ represents the acoustic room impulse response between the speech source and the mth microphone and $d_m(n)$ and $u_m(n)$ are the speech and noise components received at the mth microphone. In this paper, it is assumed that the noise signals $u_m(n), m = 1, 2, \ldots, M$ are stationary and uncorrelated with the speech source. Alternative representation of (1) can be written with respect to the speech component at the first microphone

$$y_m(n) = h_m(n) * d_1(n) + u_m(n), m = 1, 2, \ldots, M \tag{2}$$

where $h_m(n)$ represents the *relative* impulse response between the mth microphone and the first microphone with respect to the speech source location, which satisfies $a_m(n) = h_m(n) * a_1(n)$. It is worthwhile noting that $h_m(n)$ is generally of infinite length since it represents the impulse response of the ratio between a couple of room transfer functions. However, the energy of the relative impulse response decays rapidly and therefore the assumption that the support of $h_m(n)$ is finite is practically not very restrictive.

The signals can be divided into overlapping time frames and analyzed using the short time Fourier transform. Let N_y denote the number of time frames of the observed signals $y_m(n)$, N denote the length of each time frame, and L denote the framing step. According to [16], [20], [21] a filter convolution in the time domain is transformed into a sum of N cross-band filter convolutions in the STFT domain. The cross-band filters are used for canceling the aliasing caused by sampling in each frequency subband [23]. Hence, we can represent (2) in the STFT domain

$$\begin{aligned} y_m(p,k) &= d_m(p,k) + u_m(p,k) \\ &= \sum_{k'=0}^{N-1} \sum_{p'} h_m(p', k', k) d_1(p - p', k') + u_m(p,k) \\ &= \sum_{k'=0}^{N-1} h_m(p, k', k) * d_1(p, k') + u_m(p,k) \end{aligned} \tag{3}$$

where p is the time frame index, k and k' are the frequency subband indices, and $h_m(p, k', k)$ is the cross-band filter between frequency band k' and k of length N_h. Let $\mathbf{d}_m(k)$, $\mathbf{u}_m(k)$ and $\mathbf{y}_m(k)$ denote column stack vectors of length N_y comprised of the STFT samples at subband k of the signals $d_m(n)$, $u_m(n)$ and $y_m(n)$, respectively, and let $\mathbf{H}_m(k', k)$ denote the convolution matrix of the cross-band filter $h_m(p, k', k)$ of size $N_y \times N_y$. Then, (3) can be written in matrix representation

$$\mathbf{y}_m(k) = \sum_{k'=0}^{N-1} \mathbf{H}_m(k', k) \mathbf{d}_1(k') + \mathbf{u}_m(k). \tag{4}$$

In this derivation, we assume that $h_m(n)$ is time-invariant during the entire signal length. Thus, in order to track time variations the signal may be approximately divided into intervals in which the filter is assumed to be fixed.

The objective of this work is estimation of $d_1(n)$ given observations from the microphone array $\{y_m(n)\}_{m=1}^M$, i.e., we aim at estimating an undistorted and noisy-free version of the speech component received at the first microphone.

III. Minimum Variance Distortionless Response Beamformer

An estimator of $d_1(n)$ can be obtained by passing the measurements $\{y_m(n)\}_{m=1}^M$ through M finite-impulse response (FIR) filters

$$\begin{aligned} \hat{d}_1(n) &= g_1^*(n) * y_1(n) + \cdots + g_M^*(n) * y_M(n) \\ &= \sum_{m=1}^{M} g_m^*(n) * y_m(n) \end{aligned} \tag{5}$$

where * denotes complex conjugate.[2] Similarly to (3), we can write (5) in the STFT domain and get

$$\hat{d}_1(p,k) = \sum_{m=1}^{M} \sum_{k'=0}^{N-1} g_m^*(p, k', k) * y_m(p, k') \tag{6}$$

or in a compact matrix form

$$\hat{\mathbf{d}}_1(k) = \mathbf{G}^H(k) \mathbf{y} \tag{7}$$

where ^{H} represents complex conjugate transpose, $\mathbf{y}$ is a column stack vector of the STFT samples of the observed signals from all sensors in all subbands, and $\mathbf{G}(k)$ represents a concatenation of convolution matrices of the cross-band filters $g_m(p, k', k)$ (see Appendix A for the derivation and exact notations). Using (7), we can define the estimation error

$$\begin{aligned} \mathbf{e}(k) &\triangleq \hat{\mathbf{d}}_1(k) - \mathbf{d}_1(k) \\ &= \mathbf{G}^H(k) \mathbf{y} - \mathbf{d}_1(k). \end{aligned} \tag{8}$$

Since the observed signals consist of a speech component and a noise component we have $\mathbf{y} = \mathbf{d} + \mathbf{u}$, where $\mathbf{d}$ and $\mathbf{u}$ are defined

[2]As will be shown later in this paper, it is more convenient for the following derivations to define the estimator convolution using complex conjugate, yielding a positive spectrum of the estimated signal.

similarly to $\mathbf{y}$. Thus, we can write the estimation error as a sum of two terms

$$\mathbf{e}(k) = \mathbf{e}_d(k) + \mathbf{e}_u(k) \tag{9}$$

where

$$\mathbf{e}_d(k) \triangleq \mathbf{G}^H(k)\mathbf{d} - \mathbf{d}_1(k) \tag{10}$$

represents the *speech distortion*, and

$$\mathbf{e}_u(k) \triangleq \mathbf{G}^H(k)\mathbf{u} \tag{11}$$

represents the *residual noise*.

Now, from (11), the MSE of each subband in the STFT domain associated with the residual noise is

$$\begin{aligned} J_u(\mathbf{G}(k)) &\triangleq \mathrm{tr}\left[E\left\{\mathbf{e}_u(k)\mathbf{e}_u^H(k)\right\}\right] \\ &= \mathrm{tr}[\mathbf{G}^H(k)\mathbf{R}_u\mathbf{G}(k)] \end{aligned} \tag{12}$$

where $\mathrm{tr}[\cdot]$ is a matrix trace and

$$\mathbf{R}_u \triangleq E\{\mathbf{u}\mathbf{u}^H\}.$$

Thus, we can state the noise reduction problem subject to zero speech distortion as follows:

$$\begin{aligned} &\mathbf{G}_{\mathrm{opt}}(k) = \arg\min_{\mathbf{G}(k)} J_u(\mathbf{G}(k)) \\ &\text{s.t.} \quad \mathbf{e}_d(k) = 0; \quad \forall k = 0, 1, \ldots, N-1. \end{aligned} \tag{13}$$

By using the relative impulse responses, which represent the coupling between the speech components at each microphone, the zero speech distortion constraint can be written explicitly (see Appendix B), and we can rewrite the optimization problem as

$$\begin{aligned} &\mathbf{G}_{\mathrm{opt}}(k) = \arg\min_{\mathbf{G}(k)} J_u(\mathbf{G}(k)) \\ &\text{s.t. } \mathbf{H}^H\mathbf{G}(k) = \mathbf{W}(k); \quad \forall k = 0, 1, \ldots, N-1 \end{aligned} \tag{14}$$

where $\mathbf{H}$ is defined as $\mathbf{H} = [\mathbf{I}\ \mathbf{H}_2^T \ldots \mathbf{H}_M^T]^T$, $\mathbf{H}_m$ is a matrix consisting of the N^2 convolution matrices $\mathbf{H}_m(k', k)$, and $\mathbf{W}^T(k)$ is a constant matrix given by

$$\mathbf{W}^T(k) = \left[\underbrace{\mathbf{0}\ldots\mathbf{0}}_{(k-1)N_y}\ \mathbf{I}_{N_y}\ \underbrace{\mathbf{0}\ldots\mathbf{0}}_{(N-k)N_y}\right]$$

where $\mathbf{0}$ is a vector of zeros and $\mathbf{I}_{N_y}$ is a unit matrix.

Solving the above optimization problem requires estimates of the RTFs of all sensors in the STFT domain $\hat{\mathbf{H}}$ and estimates of the noise signals PSDs $\hat{\mathbf{R}}_u$. A geometric illustration of the optimal beamformer is shown in Fig. 1.

The constraint in (14), posed on the matrix $\mathbf{G}(k)$, can be broken into N_y constraints on each of its columns $\mathbf{g}^i(k) = \mathbf{G}(k)\mathbf{e}^i_{N_y}$, where $\mathbf{e}^i_{N_y}$ is a unit vector of length N_y.[3] Thus, each constraint takes the following form

$$\mathbf{H}\mathbf{G}(k)\mathbf{e}^i = \mathbf{w}^i(k), \quad \forall i = 1, \ldots, N_y \tag{15}$$

where $\mathbf{w}^i(k)$ is the ith column of the matrix $\mathbf{W}(k)$, i.e., it is a vector of length NN_y and equals $\mathbf{w}^i(k) = \mathbf{e}^{(k-1)N_y+i}, \forall i = 1, \ldots, N_y$. Now, (14) can be solved using N_y Lagrange multipliers.

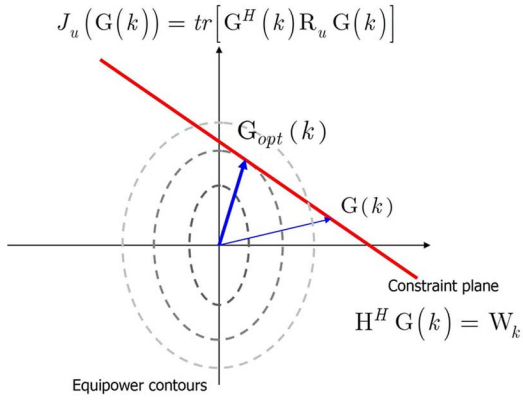

Fig. 1. Geometric interpretation of the optimal beamformer.

IV. Generalized Sidelobe Canceler

Based on the generalized sidelobe canceler (GSC) structure [6], we transform the constrained optimization problem into an unconstrained form. The minimization and the constraint are decoupled into two parallel processing branches, yielding an MVDR-equivalent beamformer carried out in the STFT domain.

Consider the null space of $\mathbf{H}$, defined by[4]

$$\mathcal{N} \triangleq \{\mathbf{g} | \mathbf{H}^H\mathbf{g} = \mathbf{0}\} \tag{16}$$

and the constraints hyperplanes, defined as

$$\Lambda^i_k \triangleq \{\mathbf{g} | \mathbf{H}^H\mathbf{g} = \mathbf{w}^i(k)\}, \quad i \in 1, \ldots, N_y. \tag{17}$$

Thus, we get N_y hyperplanes, parallel to the null space $\mathcal{N}$. It is worthwhile noting that such a null space exists due to $\mathbf{H}$ dimensions. Let $\mathcal{R}$ denote the range of $\mathbf{H}$, i.e.,

$$\mathcal{R} \triangleq \{\mathbf{H}\boldsymbol{\kappa} | \boldsymbol{\kappa} \in \mathbb{C}^{NN_y}\}. \tag{18}$$

Using the fundamental theorem of Linear Algebra [24], we have $\mathcal{N} \perp \mathcal{R}$. Thus, each vector $\mathbf{g}(k)$ in the linear space can be uniquely split into a sum of two vectors in mutually orthogonal subspaces, as follows:

$$\mathbf{g}(k) = \mathbf{g}_0(k) - \mathbf{v}(k) \tag{19}$$

[3]In the following, we neglect the unit vector length notation for simplicity.

[4]This derivation is in higher dimensions than the null space definition presented in [6] since our presentation involves dependency between time frames in each subband.

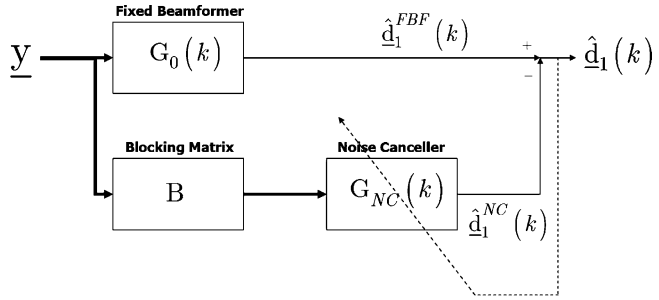

Fig. 2. GSC structure in the STFT domain.

where $\mathbf{g}_0(k) \in \mathcal{R}$ and $-\mathbf{v}(k) \in \mathcal{N}$. Consequently, a beamformer filter $\mathbf{G}(k)$ in the STFT domain can be written as

$$\mathbf{G}(k) = \mathbf{G}_0(k) - \mathbf{V}(k) \tag{20}$$

where $\mathbf{G}_0(k)$ columns satisfy $\mathbf{g}_0^i(k) \in \mathcal{R}$ and $-\mathbf{V}(k)$ columns satisfy $-\mathbf{v}^i(k) \in \mathcal{N}$. According to the definition of $\mathcal{N}$, we can write $\mathbf{V}(k)$ as

$$\mathbf{V}(k) = \mathbf{B}\mathbf{G}_{\mathrm{NC}}(k) \tag{21}$$

where $\mathbf{G}_{\mathrm{NC}}(k)$ is designated as a *noise canceler* matrix of size $(M-1)NN_y \times N_y$ and $\mathbf{B}$ is designated as a *blocking matrix* of size $MNN_y \times (M-1)NN_y$, whose columns are in the null space $\mathcal{N}$. Hence, $\mathbf{B}$ satisfies

$$\mathbf{H}^H\mathbf{B} = \mathbf{O} \tag{22}$$

where $\mathbf{O}$ is a matrix of zeros.

Thus, using (20), the estimator defined in (7) can be written as

$$\begin{aligned}\hat{\mathbf{d}}_1(k) &= \mathbf{G}^H(k)\mathbf{y} \\ &= \hat{\mathbf{d}}_1^{\mathrm{FBF}} - \hat{\mathbf{d}}_1^{\mathrm{NC}}\end{aligned} \tag{23}$$

where

$$\hat{\mathbf{d}}_1^{\mathrm{FBF}} \triangleq \mathbf{G}_0^H(k)\mathbf{y} \tag{24}$$

$$\begin{aligned}\hat{\mathbf{d}}_1^{\mathrm{NC}} &\triangleq \mathbf{V}^H(k)\mathbf{y} \\ &= \mathbf{G}_{\mathrm{NC}}^H(k)\mathbf{B}^H\mathbf{y}.\end{aligned} \tag{25}$$

In (23), we obtain the generalized sidelobe canceler structure, illustrated in Fig. 2. Thus, the solution comprises three blocks. The first is a fixed beamformer $\mathbf{G}_0(k)$, which satisfies the constraints and hence steers the beam towards the desired direction, i.e., the speech component at the first microphone is kept undistorted. The second is a blocking matrix $\mathbf{B}$, that blocks the desired signal and produces noise-only reference signals. The third is a noise canceler adjustable filter $\mathbf{G}_{\mathrm{NC}}(k)$, that is designed to cancel the coherent noise in the fixed beamformer output, and is built using an unconstrained adaptive LMS algorithm.

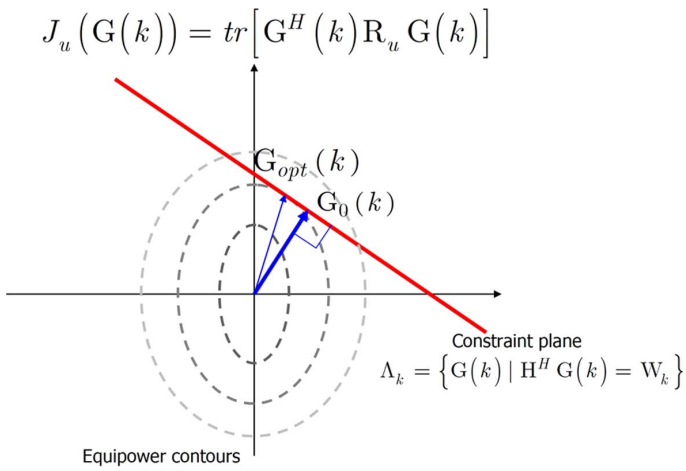

Fig. 3. Geometric interpretation of the fixed beamformer.

A. Fixed Beamformer

By setting the fixed beamformer $\mathbf{G}_0(k)$ to be

$$\mathbf{G}_0(k) = \mathbf{H}(\mathbf{H}^H\mathbf{H})^{-1}\mathbf{W}(k) \tag{26}$$

it satifies

$$\mathbf{H}^H\mathbf{G}_0(k) = \mathbf{W}(k). \tag{27}$$

According to (17) and (18), we obtain that each column of $\mathbf{G}_0(k)$ is in both the constraint hyperplane $\mathbf{g}_0^i(k) \in \Lambda_k^i$ and the range $\mathbf{g}_0^i(k) \in \mathcal{R}$. Since Λ_k^i is parallel to $\mathcal{N}$ and $\mathcal{R}$ is orthogonal to $\mathcal{N}$, each column of $\mathbf{G}_0(k)$ is the perpendicular from the origin to the corresponding constraint hyperplane. Fig. 3 shows an illustration of the fixed beamformer.

Now, substituting (26) into (24), yields

$$\begin{aligned}\hat{\mathbf{d}}_1^{\mathrm{FBF}}(k) &= \mathbf{W}^T(k)(\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H\mathbf{y} \\ &= \mathbf{W}^T(k)(\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H\mathbf{d} \\ &\quad + \mathbf{W}^T(k)(\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H\mathbf{u}.\end{aligned} \tag{28}$$

Using the coupling between the microphone signals, we have

$$\mathbf{d} = \mathbf{H}\mathbf{d}_1. \tag{29}$$

Substituting (29) into (28), yields

$$\begin{aligned}\hat{\mathbf{d}}_1^{\mathrm{FBF}}(k) &= \mathbf{W}^T(k)\mathbf{d}_1 + \mathbf{W}^T(k)(\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H\mathbf{u} \\ &= \mathbf{d}_1(k) + \mathbf{W}^T(k)(\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H\mathbf{u}.\end{aligned} \tag{30}$$

Thus, the output of the proposed fixed beamformer $\mathbf{G}_0(k)$ comprises of the desired undistorted signal, i.e., the speech component at the first microphone, and an additive noise, which is a mixture of the noise components received at the microphones.

B. Blocking Matrix

Define the blocking matrix as

$$\mathbf{B} = \begin{bmatrix} -\mathbf{H}_2^H & -\mathbf{H}_3^H & \cdots & -\mathbf{H}_M^H \\ \mathbf{I} & 0 & \cdots & 0 \\ 0 & \mathbf{I} & & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{I} \end{bmatrix}. \tag{31}$$

Using simple arithmetic, we obtain that $\mathbf{B}$ satisfies $\mathbf{H}^H\mathbf{B} = \mathbf{O}$. From (25), using $\mathbf{y} = \mathbf{d} + \mathbf{u}$ and (29), the noise canceler output is

$$\hat{\mathbf{d}}_1^{\mathrm{NC}} = \mathbf{G}_{\mathrm{NC}}^H(k)\mathbf{B}^H\mathbf{H}\mathbf{d}_1 + \mathbf{G}_{\mathrm{NC}}^H(k)\mathbf{B}^H\mathbf{u} \tag{32}$$

and by using the proposed blocking matrix from (31), we obtain

$$\hat{\mathbf{d}}_1^{\mathrm{NC}} = \mathbf{G}_{\mathrm{NC}}^H(k)\mathbf{B}^H\mathbf{u}. \tag{33}$$

Hence, the desired signal is blocked and the noise canceler output is comprised of reference noise-only signals.

C. Noise Canceler

Our goal is to minimize the power of the output signal, i.e.,

$$\min_{\mathbf{G}_{\mathrm{NC}}(k)} \left\| \hat{\mathbf{d}}_1^{\mathrm{FBF}}(k) - \hat{\mathbf{d}}_1^{\mathrm{NC}}(k) \right\|^2 \tag{34}$$

which is obtained by adjusting the noise canceler filter $\mathbf{G}_{\mathrm{NC}}(k)$. From (33), we know that the noise canceler receives noise-only signals (the output of the blocking matrix), and that it is designed to cancel out the coherent noise at the output of the fixed beamformer (30). This adjustment problem is the classical multichannel noise cancellation problem, that can be solved by using the Wiener filter. In this paper, we implement the noise canceler adaptively using the NLMS algorithm. Subsequently, we take advantage of the fact that speech signals are better represented in the STFT domain than in the time domain. The condition number of the autocorrelation matrix of the STFT samples of the speech signal is closer to 1, yielding an improved convergence rate of adaptive algorithms in the STFT domain than in the time domain [18], [23], [25].

V. Proposed Beamformer

Implementing the GSC scheme as described above is insufficient due to large dimensionality of the system when perfectly represented in the STFT domain. Thus, we propose approximate representations that reduce the model complexity, while maintaining satisfactory performance.

We focus our work on convolution approximations in the STFT domain. These approximations can be applied on the matrix $\mathbf{H}$, which represents the convolution with the RTF and describes the coupling between the speech components. Applying such an approximation on the GSC framework influence both processing branches and has a major impact on the beamformer output. At the upper branch, the approximated fixed beamformer does not meet the constraint, thus, speech distortion is introduced into the system. At the lower branch, the approximated blocking matrix does not belong to the null space, and as a result does not block the speech signal completely. Clearly, the amount of leakage has a major influence on the quality of the beamformer output. In case of significant leakage from the blocking matrix, speech traces are left at the output of the noise canceler and subtracted from the fixed beamformer output, causing distortion at the beamformer output.

A. MTF Approximation

Now we apply the MTF approximation on the GSC scheme in the STFT domain. Under this approximation, a convolution in the time domain becomes a scalar multiplication in the STFT domain (as previously mentioned, with a proper zero padding of the STFT analysis window). Thus, the cross-band filters are neglected and the band-to-band filter is approximated as a single coefficient. Accordingly, the convolution matrix of the cross-band filters of the RTF between the speech component at the mth microphone and the speech component at the first microphone are neglected $\mathbf{H}_m(k', k) = 0, \forall k' \neq k$ and the convolution matrix of the band-to-band filter is approximated by

$$\tilde{\mathbf{H}}_m(k, k) = h_m(k)\mathbf{I}. \tag{35}$$

Thus, the matrix $\mathbf{H}_m$ under the MTF model is reduced to a diagonal matrix

$$\tilde{\mathbf{H}}_m = \begin{bmatrix} h_m(0)\mathbf{I} & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & h_m(1)\mathbf{I} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{O} \\ \mathbf{O} & \cdots & \mathbf{O} & h_m(N-1)\mathbf{I} \end{bmatrix}. \tag{36}$$

By substituting (36) into (26), we obtain the approximated fixed beamformer

$$\tilde{\mathbf{G}}_0(k) \triangleq \mathbf{F}\mathbf{W}(k) \tag{37}$$

where $\mathbf{F} = [\mathbf{F}_1 \cdots \mathbf{F}_M]^T$ and

$$\mathbf{F}_m = \begin{bmatrix} f_m(0)\mathbf{I} & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & f_m(1)\mathbf{I} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{O} \\ \mathbf{O} & \cdots & \mathbf{O} & f_m(N-1)\mathbf{I} \end{bmatrix}$$

where

$$f_m(k) = \frac{h_m(k)}{\sum_{m'=1}^{M} |h_{m'}(k)|^2}.$$

Substituting (37) into (24), reduces the fixed beamformer output based on the MTF approximation to

$$\left\{ \hat{\mathbf{d}}_1^{\mathrm{FBF}}(k) \right\}_{\mathrm{MTF}} = \frac{\sum_{m=1}^{M} h_m^*(k)\mathbf{y}_m(k)}{\sum_{m'=1}^{M} |h_{m'}(k)|^2} \tag{38}$$

which is the output of the fixed beamformer using the TF-GSC method [6].

Let $\mathbf{z}_m, \forall m = 2, \ldots, M$ denote the mth output channel of the blocking matrix. Thus, it can be written as

$$\mathbf{z} = \mathbf{B}^H\mathbf{y} \tag{39}$$

where $\mathbf{z} = [\mathbf{z}_2 \ \dots \ \mathbf{z}_M]^T$. Using the proposed blocking matrix from (31), we obtain

$$\mathbf{z}_m = -\mathbf{H}_m \mathbf{y}_1 + \mathbf{y}_m. \tag{40}$$

Now, substituting the MTF approximation of the RTF from (36) into (40), yields at each subband

$$\mathbf{z}_m(k) = \mathbf{y}_m(k) - h_m(k)\mathbf{y}_1(k), \quad \forall m = 2, \dots, M \tag{41}$$

which is the output of the blocking matrix using the TF-GSC method (written in matrix form for all time frames).

Building the noise canceler under the MTF approximation requires a single multiplicative coefficient for each frequency subband, and it is built adaptively using the NLMS algorithm.

By applying the MTF approximation to the general GSC scheme in the STFT domain, we obtain the same algorithm (the TF-GSC) which was derived under the MTF model (used for representing the signals in the STFT domain and for the decoupling of the problem into two orthogonal subspaces). This is an important result, which explains the good results achieved by the TF-GSC method. However, since in practice short time frames are used, the MTF approximation becomes inaccurate when a long response representation is required, e.g., in reverberant environments.

B. CTF Approximation

In order to obtain a better implementable representation of a convolution in the STFT domain, we apply the CTF approximation on the suggested GSC scheme. Under this approximation a convolution in the time domain becomes a convolution in the STFT domain with the band-to-band filter at each frequency subband. Accordingly, the cross-band filters are neglected, i.e., the convolution matrices satisfy $\mathbf{H}_m(k', k) = 0, \forall k' \neq k$, and hence the matrix $\mathbf{H}_m$ becomes sparser relative to the complete representation and is reduced to a block-diagonal matrix

$$\tilde{\mathbf{H}}_m = \begin{bmatrix} \mathbf{H}_m(0,0) & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \mathbf{H}_m(1,1) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{O} \\ \mathbf{O} & \cdots & \mathbf{O} & \mathbf{H}_m(N-1,N-1) \end{bmatrix}. \tag{42}$$

As noted before, the form of the CTF model resembles the structure of the block partitioned convolution [19], which can be obtained by employing the MTF approximation in some block frequency domain representation.

Substituting (42) into (26), yields the approximated fixed beamformer

$$\tilde{\mathbf{G}}_0(k) \triangleq \mathbf{F}\mathbf{W}(k) \tag{43}$$

where $\mathbf{F} = [\mathbf{F}_1 \ \cdots \ \mathbf{F}_M]^T$ and

$$\mathbf{F}_m = \begin{bmatrix} \mathbf{F}_m(0) & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \mathbf{F}_m(1) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{O} \\ \mathbf{O} & \cdots & \mathbf{O} & \mathbf{F}_m(N-1) \end{bmatrix}$$

where

$$\mathbf{F}_m(k) = \mathbf{H}_m(k,k) \left[\sum_{m'=1}^{M} \mathbf{H}_{m'}^H(k,k)\mathbf{H}_{m'}(k,k) \right]^{-1}.$$

By substituting (43) into (24), we get a reduced fixed beamformer, which output is given by

$$\{\hat{\mathbf{d}}_1^{\mathrm{FBF}}(k)\}_{\mathrm{CTF}} = \mathbf{K}(k) \sum_{m=1}^{M} \mathbf{H}_m^H(k,k)\mathbf{y}_m(k) \tag{44}$$

where

$$\mathbf{K}(k) = \left[\sum_{m=1}^{M} \mathbf{H}_m^H(k,k)\mathbf{H}_m(k,k) \right]^{-1}. \tag{45}$$

Similarly to (41), the blocking matrix output based on the CTF model is

$$\mathbf{z}_m(k) = \mathbf{y}_m(k) - \mathbf{H}_m(k,k)\mathbf{y}_1(k), \quad \forall m = 2, \dots, M. \tag{46}$$

Under the CTF approximation, the noise canceler is reduced to a band-to-band filter at each subband, and is adaptively implemented using the NLMS algorithm.

Thus, we obtain a new practical beamforming technique, derived from the GSC scheme in the STFT domain, under the CTF approximation. Based on this approximation, we obtain flexibility in adjusting the RTF length and time frames length, resulting in a good STFT representation of the signals in reverberant environments. Moreover, this representation is compact enough, enabling the employment of a feasible solution.

C. RTF Identification Based on the CTF Model

In order to build the GSC blocks derived under the CTF approximation, estimates of the RTFs relating the speech components at all the sensors, namely $\mathbf{H}_m, \quad m = 1, \dots, M$ are required. In the following we briefly review a recently proposed RTF identification method using the CTF model [14]. Let $\mathbf{Y}_1(k)$ be an $N_y \times N_h$ Toeplitz matrix constructed from the STFT coefficients of $y_1(n)$ in the kth subband. Similarly, let $\mathbf{U}_1(k)$ be an $N_y \times N_h$ Toeplitz matrix constructed from the STFT coefficients of $u_1(n)$. From (3), based on the CTF approximation (neglecting the cross-band filters), we have for each microphone measurement

$$\mathbf{y}_m(k) = \mathbf{Y}_1(k)\mathbf{h}_m(k,k) + \mathbf{v}_m(k) \tag{47}$$

where $\mathbf{h}_m(k,k)$ is the band-to-band filter of the relative impulse response given by

$$\mathbf{h}_m(k,k) = [h_m(0,k,k), \ldots, h_m(N_h-1,k,k)]^T$$

and

$$\mathbf{v}_m(k) = \mathbf{u}_m(k) - \mathbf{U}_1(k)\mathbf{h}_m(k,k). \tag{48}$$

By taking expectation of the cross multiplication of the STFT samples of the two observed signals $y_m(p,k)$ and $y_1^*(p,k)$, we obtain from (47)

$$\Phi_{m,1}^y(k) = \Psi_{1,1}^y(k)\mathbf{h}_m(k,k) + \Phi_{m,1}^v(k) \tag{49}$$

where $\Psi_{1,1}^y(k)$ is an $N_y \times N_h$ matrix and its (p,l)th term is

$$\left[\Psi_{1,1}^y(k)\right]_{p,l} = E\{y_1(p-l,k)y_1^*(p,k)\} \triangleq \psi_{1,1}^y(p,l,k) \tag{50}$$

and $\Phi_{m,1}^y(k)$ and $\Phi_{m,1}^v(k)$ are vectors of length N_y, given as

$$\Phi_{m,1}^y(k) = \left[\phi_{m,1}^y(0,k) \quad \cdots \quad \phi_{m,1}^y(N_y-1,k)\right]^T \tag{51}$$

$$\Phi_{m,1}^v(k) = \left[\phi_{m,1}^v(0,k) \quad \cdots \quad \phi_{m,1}^v(N_y-1,k)\right]^T \tag{52}$$

where $E\{\cdot\}$ denotes mathematical expectation, $\phi_{m,1}^y(p,k)$ denotes the cross PSD between the signals $y_m(n)$, and $y_1(n)$, $\phi_{m,1}^v(p,k)$ denotes the cross PSD between the signals $v_m(n)$ and $y_1(n)$ and $\psi_{1,1}^y(p,l,k)$ denotes the cross PSD between the signal $y_1(n)$ and its delayed version $y_1'(n) \triangleq y_1(n-lL)$, all at time frame p and frequency k. Since the speech signal $s(n)$ is uncorrelated with the noise signal $u(n)$, by taking mathematical expectation of the cross multiplication of the STFT samples $v_m(p,k)$ and $y_1^*(p,k)$, we get from (48)

$$\Phi_{m,1}^v(k) = \Phi_{m,1}^u(k) - \Psi_{1,1}^u(k)\mathbf{h}_m(k,k) \tag{53}$$

where $\Phi_{m,1}^u(k)$ is a vector of length N_y, given as

$$\Phi_{m,1}^u(k) = \left[\phi_{m,1}^u(k) \quad \cdots \quad \phi_{m,1}^u(k)\right]^T \tag{54}$$

and $\Psi_{1,1}^u(k)$ is an $N_y \times N_h$ matrix and its (p,l)th term is given by

$$\left[\Psi_{1,1}^u(k)\right]_{p,l} = E\{u_1(p-l,k)u_1^*(p,k)\} \triangleq \psi_{1,1}^u(l,k) \tag{55}$$

where $\phi_{m,1}^u(k)$ denotes the cross PSD between the signals $u_m(n)$ and $u_1(n)$, and $\psi_{1,1}^u(l,k)$ denotes the cross PSD between the signal $u_1(n)$ and its delayed version $u_1'(n) \triangleq u_1(n-lL)$, both at frequency bin k. It is worth noting that since the noise signals are stationary during our observation interval, the noise PSD terms are independent of the time frame index.

Once again, by exploiting the fact that the speech signal $s(n)$ and the noise signal $u(n)$ are uncorrelated, we obtain $\Psi_{1,1}^y(k) = \Psi_{1,1}^s(k) + \Psi_{1,1}^u(k)$, where $\Psi_{1,1}^s(k)$ is defined similarly to (50). Thus, from (49) and (53), we have

$$\Phi_{m,1}^y(k) = \Psi_{1,1}^s(k)\mathbf{h}_{k,k} + \Phi_{m,1}^u(k). \tag{56}$$

Now, rewriting (56) in terms of the PSD estimates, we obtain

$$\hat{\Phi}(k) = \hat{\Psi}(k)\mathbf{h}_m(k,k) + \mathbf{e}(k) \tag{57}$$

where $\mathbf{e}(k)$ denotes the PSD estimation error, and

$$\hat{\Phi}(k) \triangleq \hat{\Phi}_{m,1}^y(k) - \hat{\Phi}_{m,1}^u(k) \tag{58}$$

$$\hat{\Psi}(k) \triangleq \hat{\Psi}_{1,1}^s(k) = \hat{\Psi}_{1,1}^y(k) - \hat{\Psi}_{1,1}^u(k). \tag{59}$$

A weighted least square (WLS) solution to (57) is of the form[5]:

$$\hat{\mathbf{h}}_m(k,k) = (\hat{\Psi}^H(k)\Gamma(k)\hat{\Psi}(k))^{-1}\hat{\Psi}^H(k)\Gamma(k)\hat{\Phi}(k) \tag{60}$$

where $\Gamma(k)$ is the weight matrix (see [14] for more details regarding the proper choice of weights). This yields an RTF identification estimator carried out in the STFT domain using the CTF approximation. This estimator requires estimates of the PSD terms $\phi_{m,1}^y(p,k)$, $\phi_{m,1}^u(k)$, $\psi_{1,1}^y(p,l,k)$ and $\psi_{1,1}^u(l,k)$. We can estimate $\hat{\phi}_{m,1}^y(p,k)$ and $\hat{\psi}_{1,1}^y(p,l,k)$ directly from the measurements, while, the stationary noise signals PSDs $\hat{\psi}_{1,1}^u(l,k)$ and $\hat{\phi}_{m,1}^u(k)$ can be obtained from silent periods (where the speech signal is absent).

VI. Simulation Results

In this section, we evaluate the performance of the proposed MVDR beamformer implemented in a GSC scheme using the CTF approximation in various environments, and compare it with the TF-GSC method, which was formulated in this paper as an MVDR beamformer in the GSC structure under the MTF model.

In the following experiments we use Habets' simulator [26] for simulating acoustic impulse responses, based on Allen and Berkley's image method [27]. The responses are generated for a rectangular room, 4 m wide by 7 m long by 2.75 m high. We locate a linear five microphone array at the center of the room, at (2,3.5,1.375). The microphone array topology consists of five microphones in a horizontal straight line with (3, 5, 7, 9) cm spacings. In order to improve the RTFs identification, the primary microphone (designated here as the "first" microphone) was set to be the microphone positioned at the middle of the array in order to minimize the distance between the reference microphones and the primary microphone. As the distance between the microphones becomes shorter, the amount of similar components in the acoustic transfer functions (between the source and the microphones) increases. Since these components may be canceled in the transfer function ratio, the RTFs may become simpler and easier to estimate. A speech source at (2, 5.5, 1.375) is 2 m distant from the primary microphone[6]. We simulate a noise source in two locations: relatively near the microphone array at (1.5, 4, 1.375) and relatively far from the microphone array at (1, 6, 1.375). As the noise source is moved further away from the microphone array, a more diffused noise

[5]Assuming $(\Psi^H(k)\Gamma(k)\Psi(k))$ is not singular. Otherwise, a regularization in needed.

[6]Creating a far-end field regime.

field is created at the array, which might be more difficult to handle and more suitable for simulating realistic scenarios.

The signals are sampled at 8 kHz. The speech source signal is a recorded speech from the TIMIT database [28] and the noise source signal is a computer generated white zero mean Gaussian noise with variance that varies to control the SNR level. The microphones measurements are generated by convolving the source signals with the corresponding simulated impulse responses. We use a short period of noise-only signals at the beginning of each experiment for estimating the noise signals PSDs and for adjusting the noise canceler adaptively. In practice, the noise PSDs can be evaluated adaptively using the Minimum Statistics [29], MCRA [30], or IMCRA [31] methods, and the noise canceler can be updated online using a voice activity detector (VAD). The STFT is implemented using Hamming windows of length $N = 512$ with 50% overlap (i.e., $L = 256$).

The relative impulse response is infinite but in both beamformers it is approximated as an FIR. Under the MTF approximation, the RTF length is limited by the length of the time frame, whereas under the CTF approximation the RTF length can be set as desired. The RTF represents the ratio between the room transfer functions of the speech source and a couple of relatively close microphones. Consequently, as mentioned before, the two transfer functions may contain similar components, which cancel each other in the ratio. Thus, the effective length of the RTF may be shorter in practice than the length of the room transfer function, despite the fact that it contains one room transfer function and an inverse of another. In the following experiments we set the estimated RTF length to be 1/8 of the room reverberation time T_{60}. This particular ratio was set since empirical tests produced satisfactory results [14]. Let Q denote the length of the relative impulse response. Hence for example, in reverberation time $T_{60} = 0.5$ s, using the above sampling frequency and ratio, we obtain that the relative impulse response consists of $Q = 500$ taps. Now, the length of the RTF under the CTF model is set according to the linear convolution complete representation in the STFT domain [16], which is given by

$$N_h = \left\lceil \frac{Q+N-1}{L} \right\rceil + \left\lceil \frac{N}{L} \right\rceil - 1 \tag{61}$$

where the first two terms on the right-hand side represent the length of the causal and non-causal parts of the band-to-band filters. Consequently, to represent a relative impulse response of length $Q = 500$, we use band-to-band filter of length $N_h = 5$ in this case. In order to compensate for both the band-to-band filter non causal coefficients and the non causal part of the RTFs, we introduce an artificial delay of length $(\lceil N/L \rceil - 1)L + \lceil Q/2 \rceil$ into the system.

We set the time frame weights matrix to be the unit matrix $\Gamma(k) = \mathbf{I}$ when applying the estimator (60), i.e., each time frame has the same weight. For evaluating an RTF $h_m(n)$ identification performance, we use a measure of the signal blocking factor (SBF) [14], [15] defined by

$$\mathrm{SBF} = 10\log_{10}\frac{E\left\{d_1^2(n)\right\}}{E\left\{r_m^2(n)\right\}}$$

TABLE I
BLOCKING ABILITY (SBF) IN dB

	mic. 2	mic. 3	mic. 4	mic. 5
MTF	17	12	14	8
CTF	22	17	18	13

where $E\{d_1^2(n)\}$ is the power of the speech component received at the primary sensor, and $E\{r_m^2(n)\}$ is the energy contained in the leakage signal $r_m(n) = h_m(n) * d_1(n) - \hat{h}_m(n) * d_1(n)$, where $\hat{h}_m(n)$ is the RTF estimate. The leakage signal represents the difference between the reverberated speech at the reference sensor and its estimate given the speech at the primary sensor. It is worthwhile noting that the RTF identification is implemented as batch processing, by solving (60) for each subband. Thus, the overall algorithm implemented here is not completely adaptive. An adaptive version of the RTF identification using the CTF approximation is not available and needs to be developed. Such a development requires additional derivation and testing, which extends the scope of this paper and hence left for future work. In [25], adaptive system identification is proposed under a different approximation, and in [32], an online version of the TF-GSC is proposed. These studies show the potential of the proposed algorithm to become completely adaptive.

In the first experiment, we assume that the RTFs are known, and the simulated room reverberation time is set to $T_{60} = 0.5$ s. Table I presents the SBF achieved using the known RTFs, under both MTF and CTF approximations, at each of the reference microphones in the array. We observe that by using the RTFs under the CTF approximation, we obtain better blocking ability than by using the RTFs under the MTF approximation. For more details on the RTF identification in various setups using different parameters, we refer the reader to [14].

Fig. 4(a)–(d) shows the waveform of the speech component received by the primary microphone, the noisy measurement with SNR level of 0 dB at the primary microphone and the enhanced speech at the output of the TF-GSC and the proposed methods (audio files are available online [13]). It is obvious that the enhanced signal achieved by the proposed method is less noisy than the enhanced signal achieved by the TF-GSC method.

In order to compare the performance of the competing algorithms, we use three measures. The first is the signal-to-noise ratio (SNR) defined by

$$\mathrm{SNR} \triangleq 10\log_{10}\frac{\sum_{n\in T_s} d_1^2(n)}{\sum_{n\in T_s}(\hat{d}_1(n) - d_1(n))^2}$$

where T_s denotes periods in the observation interval where the speech signal is present. The second measure is the segmental signal-to-noise ratio (SegSNR), defined by

$$\mathrm{SegSNR} \triangleq \frac{1}{|\mathcal{L}|} \times \sum_{l\in\mathcal{L}} \mathcal{T}\left(10\log_{10}\frac{\sum_{n=0}^{N-1} d_1^2(n+lL)}{\sum_{n=0}^{N-1}[d_1(n+lL) - \hat{d}_1(n+lL)]^2}\right)$$

where $\mathcal{L}$ represents the set of frames which contain speech, $|\mathcal{L}|$ denotes the number of elements in $\mathcal{L}$, and $\mathcal{T}$ confines the SNR at each frame to a perceptually meaningful range between -10 and

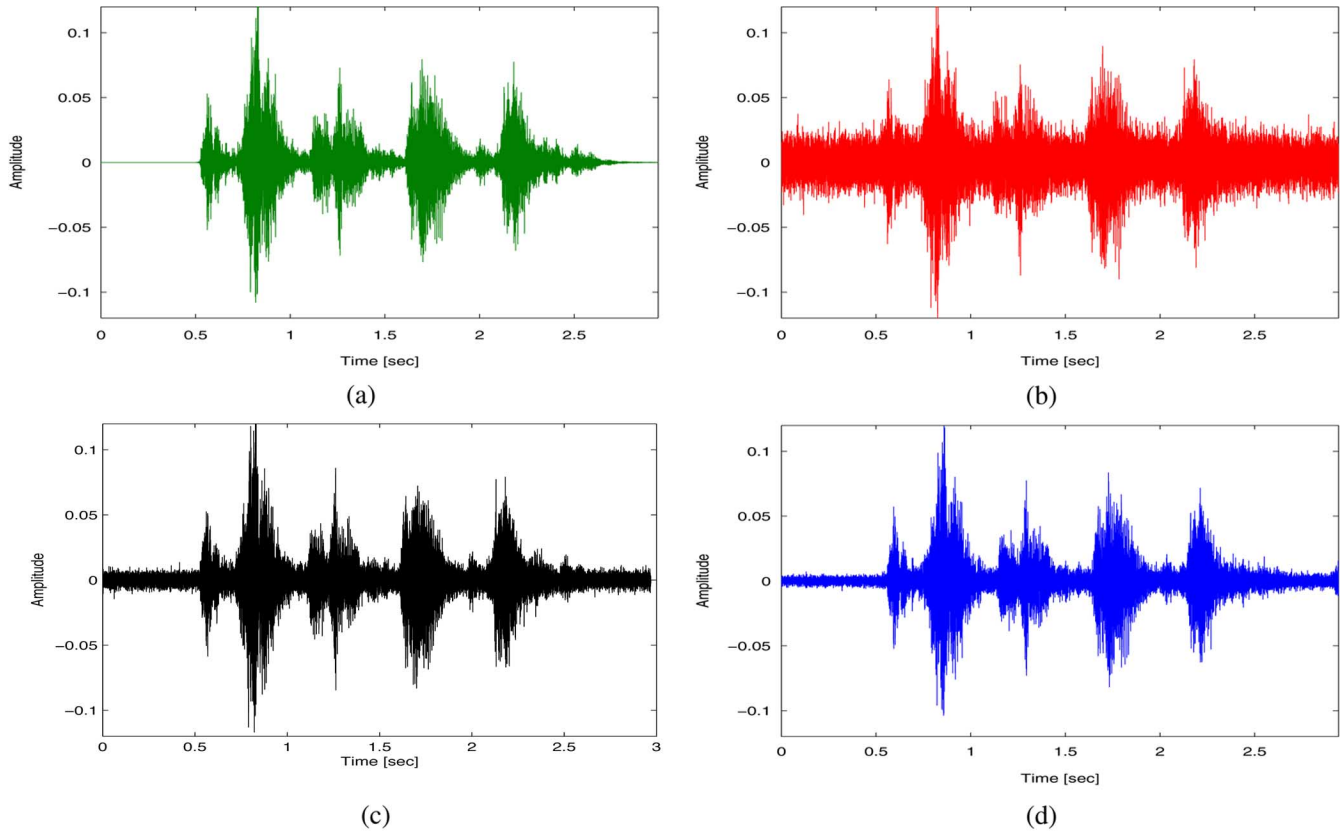

Fig. 4. Signal waveforms under known RTFs scenario. (a) Reverberant speech source received at the first microphone. (b) Noisy signal received at the first microphone, SNR $= 0$ dB. (c) Enhanced signal obtained at the TF-GSC output, SNR $= 5.8$ dB. (d) Enhanced signal obtained at the CTF-GSC output, SNR $= 11.7$ dB.

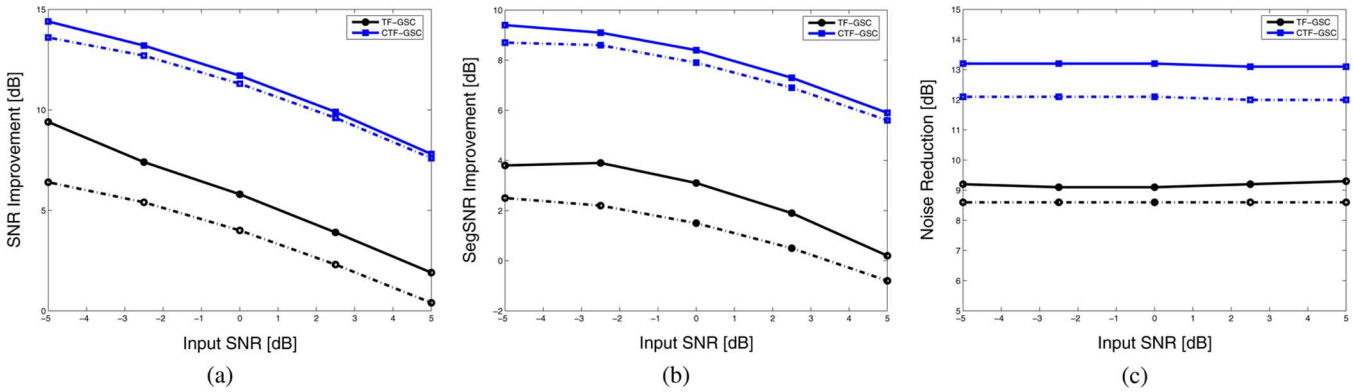

Fig. 5. Results obtained under known RTF scenario with room reverberation time set to 0.5 s. Curves obtained when the noise source is positioned relatively near the array are plotted in solid line. Curves obtained when the noise source is positioned relatively far away from the array are plotted in dashed line. (a) SNR improvement. (b) SegSNR improvement. (c) Noise reduction.

35 dB. The third measure is the noise reduction (NR), defined by

$$\mathrm{NR} \triangleq 10\log_{10}\frac{\sum_{n\in T_n} y_1^2(n)}{\sum_{n\in T_n} \hat{d}_1^2(n)}$$

where T_n denotes periods where the speech signal is absent. Thus, the NR is the ratio between the variance of the noise at the input of the system and the variance of the residual noise at the output of the system, which indeed may give a sense of noise reduction.

Fig. 5 shows the SNR improvement, the SegSNR improvement and the noise reduction obtained by the TF-GSC and the proposed algorithm in various input SNR levels. From Fig. 5(a) and (b), we observe that both the SNR improvement and the SegSNR improvement achieved by the proposed method is higher than the improvement achieved by the TF-GSC method. In addition, Fig. 5(c) shows that the proposed technique obtains better noise reduction than the TF-GSC method. Furthermore, the noise reduction obtained by both algorithms is at a fixed level during this experiment and is independent of the input SNR level. We can also observe that a constant difference between the SNR improvement (and the SegSNR improvement)

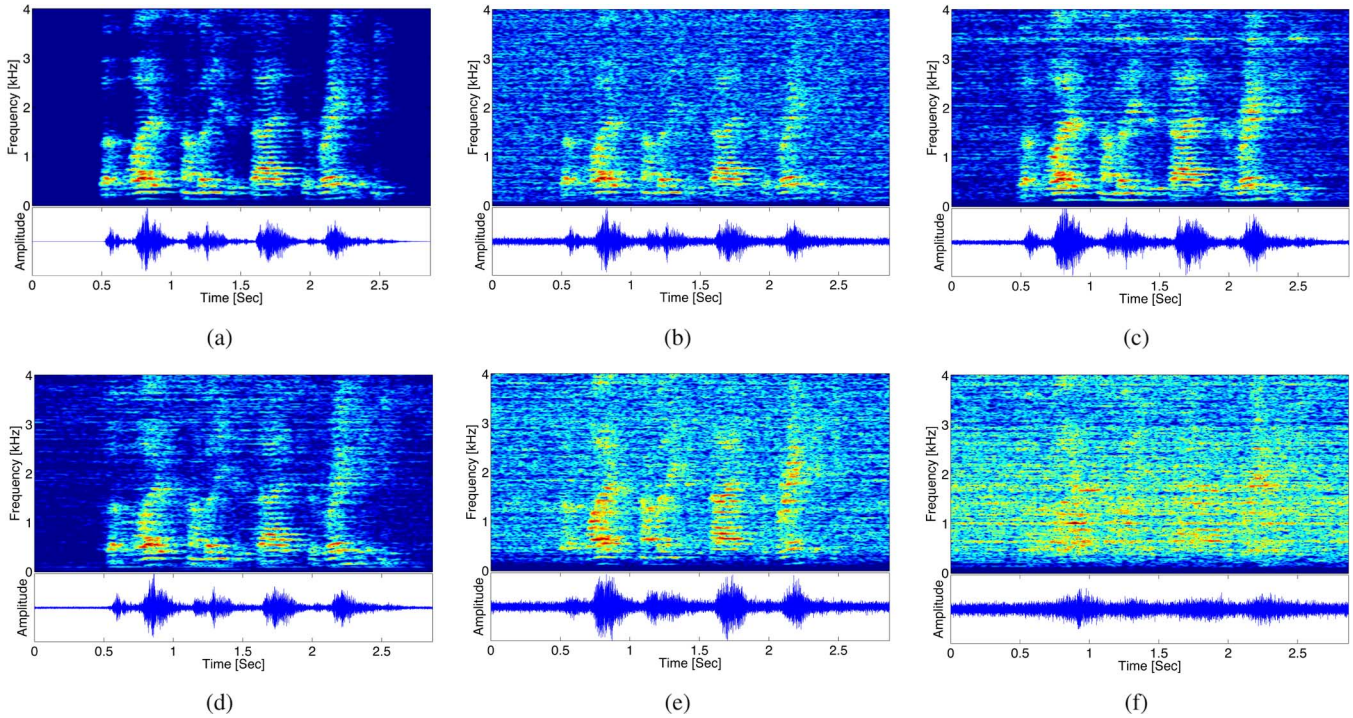

Fig. 6. Signal waveforms and spectrograms under identified RTF scenario. (a) Reverberant speech source received at the first microphone. (b) Noisy signal received at the first microphone, SNR $= 5$ dB. (c) Enhanced signal obtained at the TF-GSC output, SNR $= 6.1$ dB. (d) Enhanced signal obtained at the CTF-GSC output, SNR $= 9.2$ dB. (e) Reference noise signal at the output of the TF-GSC blocking matrix. (f) Reference noise signal at the end of the CTF-GSC blocking matrix.

achieved by the competing methods is maintained. Thus, since the RTFs are known and are not influenced by changes of the input SNR, we may conclude that both competing algorithms depend on the input SNR only through the RTFs identification. Another observation drawn from Fig. 5 is the competing methods performance under the two different locations of the noise source. It shows that both methods obtain better results when the noise source is located near the microphone array, since in this case, the created noise field at the array is less diffused and is easier to attenuate.

In the second experiment, the RTFs are unknown and should be estimated from the measurements. To make a fair comparison we implemented an improved version of the TF-GSC technique. The original version of the TF-GSC, proposed in [6], is based on the *nonstationarity* RTF identification method [33], which assumes the presence of nonstationary source and stationary uncorrelated noise. We improve this algorithm by replacing the RTF identification method with a method adapted to speech signals [15] which also takes advantage of silent periods. Thus, both the improved version of the TF-GSC and the proposed method require knowledge of speech presence probabilities, which can be obtained using a VAD.[7]

Fig. 6(a)–(f) shows the waveform and the spectrogram of the speech component received by the primary microphone, the primary microphone noisy measurement at SNR level of 5 dB, the enhanced speech obtained by the TF-GSC and the proposed methods, and a reference noise signal obtained at the output of the blocking matrix in both methods (audio files are available online[8]). We clearly observe that the enhanced signal obtained by the proposed algorithm is less noisy than the enhanced signal obtained by the TF-GSC technique. In addition, we can observe a significant speech distortion at the output of the TF-GSC method (e.g., from the waveforms in the range of 1.5–2 s), whereas the output of the proposed method seems undistorted. Examining the reference noise signals in Fig. 6(e)–(f) may suggest an explanation. It shows that the reference noise signal obtained at the output of the blocking matrix in the proposed method has less components of speech than the reference noise signal obtained at the output of the blocking matrix of the TF-GSC method. These speech components are leaked into the output of the noise canceler and then are subtracted from the fixed beamformer output, inducing a distortion.

Fig. 7 summarizes the SNR improvement, the SegSNR improvement and the noise reduction obtained by the competing algorithms. It shows that the SNR improvement (and the SegSNR improvement) achieved by the proposed method is higher than the SNR improvement (and the SegSNR improvement) achieved by the TF-GSC. In addition, the proposed algorithm obtains better noise reduction. Furthermore, we notice that as the input SNR level increases, the difference between the SNR improvement achieved by the competing methods increases. In particular, at higher input SNR levels the proposed method based on the CTF approximation becomes more advantageous. Since the CTF model is associated with larger model complexity than the MTF model, as the input

[7]As mentioned above, in these experiments we use a short period of noise-only signal in an *a priori* known location rather than using a VAD. The silent time frames are used for estimation the noise PSDs and the time frames that contain speech are used for identifying the RTFs.

[8][Online]. Available: http://www.ee.technion.ac.il/people/IsraelCohen/Publications/CTF-GSC-audio-files/waves.pdf

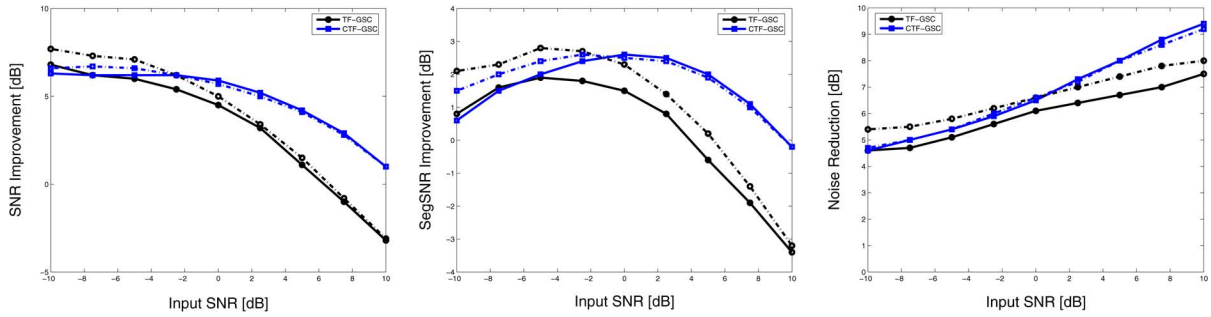

Fig. 7. Results obtained under identified RTF scenario with room reverberation time set to 0.5 s. Curves obtained when the noise source is positioned relatively near the array are plotted in solid line. Curves obtained when the noise source is positioned relatively far away from the array are plotted in dashed line. (a) SNR improvement. (b) SegSNR improvement. (c) Noise reduction.

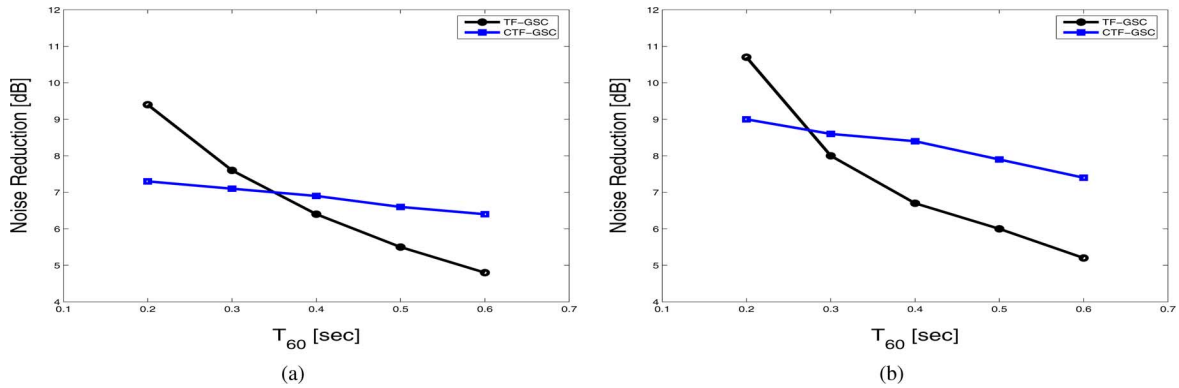

Fig. 8. Noise reduction obtained in various room reverberation times relying on identified RTF. The noise source is positioned relatively far away from the array. (a) Input SNR 0 dB. (b) Input SNR 5 dB.

SNR level increases and the data becomes more reliable, larger number of parameters can be accurately estimated [14], [16]. Similar trends can be observed in the SegSNR improvement and in the noise reduction findings, where the difference in these measures between the competing methods increases in favor of the proposed method as the input SNR increases. We can also observe that in this experiment the competing methods obtain better results when the noise source is located relatively far away from the microphone array, unlike the results shown in Fig. 5. The input SNR is defined as the ratio between the energy of the speech component and the noise component at the primary microphone. Now, since the acoustic room impulse response between the noise source and the array conveys more energy as the noise source becomes further away, the noise source power is decreased in order to maintain a certain input SNR level. Consequently, the RTF identification improves when the noise source is moved further away from the array and its power is decreased.

In the third experiment, we explore the proposed method performance in various room reverberation times. As in the previous experiment, the RTFs are unknown and should be estimated from the measurements and the remote noise source is simulated (located further away from the microphone array). As previously mentioned, since the input SNR is defined as the ratio between the variance of the reverberant speech component and the noise component at the primary microphone, it is significantly influenced by changes of the room reverberation time. In order to circumvent these changes, we redefine the input SNR as the ratio between the variances of the *clean sources* (hence, the input SNR and the reverberation time are independent) and maintain it on a fixed level during this experiment. In addition, we use the noise reduction measure to evaluate the performance since both the SNR and the SegSNR measures might be influenced by the room reverberation time.

Fig. 8 shows the noise reduction obtained by the proposed method and the TF-GSC in various reverberation times. It shows that as the reverberation time increases, the noise reduction obtained by both competing methods decreases. In addition, we observe that in shorter reverberation times the TF-GSC achieves better noise reduction, whereas in longer reverberation times the proposed method performs better. Since the effective length of the RTF increases with the reverberation time, the CTF model becomes more appropriate for the RTF representation. As mentioned in previous sections, under the MTF model the RTF length is bounded by the length of the time frame (which should be relatively short to obtain larger number of frames for reducing the estimation variance, and to validate the assumption that the speech is stationary in each time frame). However, under the CTF model, long RTFs can be represented using short time frames by using longer CTF filters. It is worthwhile noting that when using the proposed method, the CTF filter length increases [according to (61)] and more variables are estimated as the reverberation time increases, whereas, the number of variables is unchanged when using the

TF-GSC (i.e., larger part of the RTF is truncated). We can also observe that both competing methods perform better in higher input SNR levels [Fig. 8(b)] than in lower input SNR levels [Fig. 8(a)]. In addition, the proposed method becomes advantageous over the TF-GSC in shorter reverberation time when the input SNR is higher. In Fig. 8(b), the intersection point between the curves is at reverberation time of approximately 280 ms, and in Fig. 8(a) the intersection point between the curves is at reverberation time of approximately 350 ms. As already stated, the CTF model is associated with a greater model complexity than the MTF model [16]; thus, the CTF-GSC becomes more advantageous when the input SNR is higher and the data is more reliable.

VII. Conclusion

We have proposed an MVDR beamformer based on a new approach for signal and system representation in the STFT domain. The proposed algorithm is implemented using the GSC scheme, yielding an unconstrained minimization problem which can be solved efficiently. Unlike other classical methods, which rely on the multiplicative model for linear convolution representation (the so-called MTF approximation), our method is based on a convolutive model (the CTF approximation). The CTF approximation, which was shown to be more accurate and less restrictive, enables representations of long transfer functions with short time frames. This property may be especially useful in reverberant environments, where acoustic room impulse responses are long. We demonstrated the performance of the proposed method and compared it with the TF-GSC in reverberant environments. When the input SNR is sufficiently high, the CTF approximation and proposed method enable improved SNR and better noise reduction. The improved experimental results imply that the CTF approximation may be beneficially utilized also in other beamforming methods.

An adaptive version of the proposed solution is a topic for future reserach. The proposed RTF identification is based on batch processing, and therefore, an adaptive version of the RTF identification is required. In addition, online estimation of speech presence probabilities and noise PSDs should be incorporated into the system, in order to fully enjoy the advantages of the proposed GSC scheme.

Appendix A
Derivation of (7)

Writing (6) in matrix form yields

$$\hat{\mathbf{d}}_1(k) = \sum_{m=1}^{M} \sum_{k'=0}^{N-1} \mathbf{G}_m^H(k',k)\mathbf{y}_m(k') \tag{62}$$

where ^{H} represents complex conjugate transpose and $\mathbf{G}_m(k',k)$ is a convolution matrix of the cross-band filter $g_m(p,k',k)$ of size $N_y \times N_y$[9]. Let $\mathbf{y}_m$ be a concatenation of the $\mathbf{y}_m(k)$ from all subbands, i.e.,

$$\mathbf{y}_m = \left[\mathbf{y}_m^H(0)\mathbf{y}_m^H(1)\ldots\mathbf{y}_m^H(N-1)\right]^H$$

and let $\mathbf{G}_m(k)$ be a concatenation of all the convolution matrices of $\mathbf{G}_m(k',k); \quad \forall k'$ of size $N_y \times NN_y$, i.e.,

$$\mathbf{G}_m(k) = \left[\mathbf{G}_m^H(0,k)\mathbf{G}_m^H(1,k)\ldots\mathbf{G}_m^H(N-1,k)\right]^H.$$

Thus, we can rewrite (62) as

$$\hat{\mathbf{d}}_1(k) = \sum_{m=1}^{M} \mathbf{G}_m^H(k)\mathbf{y}_m \tag{63}$$

Now, by concatenating the filters and the STFT samples from all the microphones, the estimator (63) can be compactly expressed as

$$\hat{\mathbf{d}}_1(k) = \mathbf{G}^H(k)\mathbf{y} \tag{64}$$

where $\mathbf{y}$ is a vector of length MNN_y defined as

$$\mathbf{y} = \left[\mathbf{y}_1^H\mathbf{y}_2^H\ldots\mathbf{y}_M^H\right]^H$$

and $\mathbf{G}(k)$ is a matrix of size $MNN_y \times N_y$, defined as

$$\mathbf{G}(k) = \left[\mathbf{G}_1^H(k)\mathbf{G}_2^H(k)\ldots\mathbf{G}_M^H(k)\right]^H.$$

Appendix B
Derivation of (14)

From (4), we can write the coupling between the speech components at each microphone using the relative impulse response

$$\mathbf{d}_m(k) = \sum_{k'=0}^{N-1} \mathbf{H}_m(k',k)\mathbf{d}_1(k') \tag{65}$$

or in a compact form

$$\mathbf{d}_m = \mathbf{H}_m\mathbf{d}_1 \tag{66}$$

where $\mathbf{H}_m$ is a matrix of size $NN_y \times NN_y$ consisting of the N^2 convolution matrices $\mathbf{H}_m(k',k),\ \forall k',k$. It is worthwhile noting that $\mathbf{H}_1$ is a unit matrix $\mathbf{I}_{NN_y}$. Substituting (66) into (10) yields

$$\mathbf{e}_d(k) = \sum_{m=1}^{M} \mathbf{G}_m^H(k)\mathbf{H}_m\mathbf{d}_1 - \mathbf{W}^T(k)\mathbf{d}_1 \tag{67}$$

where $\mathbf{W}^T(k)$ is a constant matrix of size $N_y \times NN_y$ given by

$$\mathbf{W}^T(k) = \left[\underbrace{\mathbf{0}\ldots\mathbf{0}}_{(k-1)N_y}\ \mathbf{I}_{N_y}\ \underbrace{\mathbf{0}\ldots\mathbf{0}}_{(N-k)N_y}\right]$$

[9]We assumed that the number of time frames N_y is greater than the length of the cross-band filter $g_m(p,k',k)$.

where $\mathbf{0}$ is a vector of zeros and $\mathbf{I}_{N_y}$ is a unit matrix of size $N_y \times N_y$. Thus, from (67) we have

$$\mathbf{e}_d(k) = (\mathbf{G}^H(k)\mathbf{H} - \mathbf{W}^T(k))\mathbf{d}_1 \tag{68}$$

where $\mathbf{H}$ is a matrix of size $MNN_y \times NN_y$, defined as

$$\mathbf{H} = [\mathbf{I}\,\mathbf{H}_2^T \ldots \mathbf{H}_M^T]^T.$$

Now, from (68) we can write the constraint of zero speech distortion, i.e., $\mathbf{e}_d(k) = 0$, explicitly as

$$\mathbf{H}^H\mathbf{G}(k) = \mathbf{W}(k). \tag{69}$$

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive comments and helpful suggestions.

REFERENCES

[1] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 5, pp. 4–24, Apr. 1988.
[2] O. L. Frost III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Jan. 1972.
[3] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, Jan. 1982.
[4] B. R. Breed and J. Strauss, "A short proof of the equivalence of LCMV and GSC beamforming," *IEEE Signal Process. Lett.*, vol. 9, no. 6, pp. 168–169, Jun. 2002.
[5] S. Haykin, *Adaptive Filter Theory*. Upper Saddle RIver, New Jersey: Prentice-Hall, 2002.
[6] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.
[7] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 6, pp. 561–571, Nov. 2004.
[8] G. Reuven, S. Gannot, and I. Cohen, "Joint noise reduction and acoustic echo cancellation using the transfer-function generalized sidelobe canceller," *Special Iss. Speech Commun. Speech Enhancement*, vol. 49, pp. 623–635, Jul.–Aug. 2007.
[9] G. Reuven, S. Gannot, and I. Cohen, "Dual source transfer-function generalized sidelobe canceller," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 711–727, May 2008.
[10] J. Chen, J. Benesty, and Y. Huang, "A minimum distortion noise reduction algorithm with multiple microphones," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 481–493, Mar. 2008.
[11] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short time Fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, May 2007.
[12] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beam-former for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2677–2684, Oct. 1999.
[13] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer Handbook of Speech Processing*, J. Benesty, M. Mohan, and Y. Huang, Eds. New York: Springer, 2007, ch. 47, pt. H, pp. 945–978.
[14] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 546–555, May 2008.
[15] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Processs.*, vol. 12, no. 5, pp. 451–459, Sep. 2004.
[16] Y. Avargel and I. Cohen, "System identification in the short time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.
[17] *Springer Handbook of Speech Processing*, J. Benesty, M. Sondhi, and Y. Huang, Eds. New York: Springer, 2007.
[18] J. J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Process. Mag.*, vol. 9, no. 1, pp. 14–37, Jan. 1992.
[19] P. C. W. Sommen, "Partitioned frequency domain adaptive filters," in *Proc. Asilomar Conf. Siganls, Syst. Comput.*, 1989, pp. 676–681.
[20] M. Portnoff, "Time frequency representation of digital signals and systems based on short-time Fourier analysis," *IEEE Trans. Signal Process.*, vol. ASSP-28, no. 1, pp. 55–69, Feb. 1980.
[21] S. Farkash and S. Raz, "Linear systems in Gabor time–frequency space," *IEEE Trans. Signal Process.*, vol. 42, no. 3, pp. 611–617, Mar. 1994.
[22] J. Benesty, J. Chen, and J. Huang, "On microphone-array beamforming from a MIMO acoustic signal processing perspective," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1053–1065, Mar. 2007.
[23] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: Analysis, experiments and applications to acoustic echo cancellation," *IEEE Trans. Signal Process.*, vol. 40, no. 8, pp. 1862–1875, Aug. 1992.
[24] G. Strang, *Linear Algebra and Its Applications*. Orlando, FL: Harcourt Brace Jovanovich, 1988.
[25] Y. Avargel and I. Cohen, "Adaptive system identification in the short-time Fourier transform domain using cross-multiplicative transfer function approximation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 162–173, Jan. 2008.
[26] E. A. P. Habets, "Room impulse response (RIR) generator," Jul. 2006 [Online]. Available: http://home.tiscali.nl/ehabets/rir_generator.html
[27] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.
[28] J. S. Garofolo, *Getting Started With the DARPA TTMIT CD-ROM: An Acoustic-Phonetic Continuous Speech Database*. Gaithersburg, MD: National Inst. of Standards and Technol. (NIST), Feb. 1993.
[29] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.
[30] I. Cohen, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.
[31] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.
[32] I. Cohen, S. Gannot, and B. Berdugo, "An integrated real-time beam-forming and postfiltering system for non-stationary noise environments," *Special Iss. EURASIP J. Appl. Signal Process.: Signal Process. Acoust. Commun. Syst.*, pp. 1064–1073, Oct. 2003.
[33] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Trans. Signal Process.*, vol. 40, no. 8, pp. 2055–2063, Aug. 1996.

Ronen Talmon received the B.A. degree in mathematics and computer science from the Open University, Ra'anana, Israel, in 2005. He is currently pursuing the Ph.D. degree in electrical engineering at the Technion—Israel Institute of Technology, Haifa, Israel.

From 2000 to 2005, he was a Software Developer and Researcher in a technological unit of the Israeli Defense Forces. Since 2005, he has been a Teaching Assistant and a Project Supervisor with the Signal and Image Processing Lab (SIPL), Electrical Engineering Department, the Technion. His research interests are statistical signal processing, speech enhancement, system identification, and geometric methods for data analysis.

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

From 1990 to 1998, he was a Research Scientist with RAFAEL Research Laboratories, Haifa, Israel Ministry of Defense. From 1998 to 2001, he was a Postdoctoral Research Associate with the Computer Science Department, Yale University, New Haven, CT. In 2001, he joined the Electrical Engineering Department of the Technion, where he is currently an Associate Professor. His research interests are statistical signal processing, analysis and modeling of acoustic signals, speech enhancement, noise estimation, microphone arrays, source localization, blind source separation, system identification, and adaptive filtering. He served as Guest Editor of a special issue of the *EURASIP Journal on Advances in Signal Processing* on Advances in Multimicrophone Speech Processing and a special issue of the *EURASIP Speech Communication Journal* on Speech Enhancement. He is a coeditor of the Multichannel Speech Processing section of the *Springer Handbook of Speech Processing* (Springer, 2007), a coauthor of *Noise Reduction in Speech Processing* (Springer, 2009), and a Co-Chair of the 2010 International Workshop on Acoustic Echo and Noise Control.

Dr. Cohen received in 2005 and 2006 the Technion Excellent Lecturer awards and in 2009 the Muriel and David Jacknow Award for Excellence in Teaching. He served as an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and the IEEE SIGNAL PROCESSING LETTERS

Sharon Gannot (S'92–M'01–SM'06) received the B.Sc. degree (summa cum laude) from the Technion—Israel Institute of Technology, Haifa, Israel 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.

In 2001, he held a postdoctoral position in the Department of Electrical Engineering (SISTA), K.U. Leuven, Leuven, Belgium. From 2002 to 2003, he held a research and teaching position at the Faculty of Electrical Engineering, Technion-Israel Institute of Technology. Currently, he is a Senior Lecturer at the School of Engineering, Bar-Ilan University, Ramat-Gan, Israel. He is an Associate Editor of the *EURASIP Journal of Applied Signal Processing*, an Editor of a special issue on Advances in Multi-Microphone Speech Processing of the same journal, and a Guest Editor of the *ELSEVIER Speech Communication Journal*. His research interests include parameter estimation, statistical signal processing, and in particular speech processing using either single- or multi-microphone arrays.

Dr. Gannot is an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and a reviewer of many IEEE journals and conferences. He has been a member of the Technical and Steering Committee of the International Workshop on Acoustic Echo and Noise Control (IWAENC) since 2005 and is the General Co-Chair of IWAENC 2010 to be held in Tel-Aviv, Israel.