Multichannel Eigenspace Beamforming in a Reverberant Noisy Environment with Multiple Interfering Speech Signals

Shmulik Markovich, Sharon Gannot, *Senior Member, IEEE*, and
Israel Cohen, *Senior Member, IEEE*

Abstract

In many practical environments we wish to extract several desired speech signals, which are contaminated by non-stationary and stationary interfering signals. The desired signals may also be subject to distortion imposed by the acoustic Room Impulse Responses (RIRs). In this paper, a Linearly Constrained Minimum Variance (LCMV) beamformer is designed for extracting the desired signals from multi-microphone measurements. The beamformer satisfies two sets of linear constraints. One set is dedicated to maintaining the desired signals, while the other set is chosen to mitigate both the stationary and non-stationary interferences. Unlike classical beamformers, which approximate the RIRs as delay-only filters, we take into account the entire RIR [or its respective Acoustic Transfer Function (ATF)]. The LCMV beamformer is then reformulated in a Generalized Sidelobe Canceler (GSC) structure, consisting of a Fixed Beamformer (FBF), Blocking Matrix (BM) and Adaptive Noise Canceler (ANC). It is shown that for spatially-white noise field, the beamformer reduces to a FBF, satisfying the constraint sets, without power minimization. It is shown that the application of the adaptive ANC contributes to interference reduction, but only when the constraint sets are not completely satisfied. We show that Relative Transfer Functions (RTFs), which relate the speech sources and the microphones, and a basis for the interference subspace suffice for constructing the beamformer. The RTFs are estimated by applying the Generalized Eigenvalue Decomposition (GEVD) procedure to the Power Spectrum Density (PSD) matrices of the received signals and the stationary noise. A basis for the interference subspace is estimated by collecting eigenvectors, calculated in segments where non-stationary interfering sources are active and the desired sources are inactive. The rank of the basis is then reduced by the application the Orthogonal Triangular Decomposition (QRD). This procedure relaxes the common requirement for non-overlapping activity periods of the interference sources. A comprehensive experimental study in both simulated and real environments demonstrates the performance of the proposed beamformer.

Index Terms

Array signal processing, Speech Enhancement, Subspace methods, Interference cancellation.

EDICs terms - SPE-ENHA, AUD-LMAP, AUD-SSEN

S. Markovich and S. Gannot are with the School of Engineering, Bar-Ilan University, Ramat-Gan, 52900, Israel (e-mail: gannot@eng.biu.ac.il). I. Cohen is with the Department of Electrical Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel (e-mail: shmuelm@techunix.technion.ac.il; icohen@ee.technion.ac.il).

I. Introduction

Speech enhancement techniques, utilizing microphone arrays, have attracted the attention of many researchers for the last thirty years, especially in hands-free communication tasks. Usually, the received speech signals are contaminated by interfering sources, such as competing speakers and noise sources, and also distorted by the reverberating environment. Whereas single microphone algorithms might show satisfactory results in noise reduction, they are rendered useless in competing speaker mitigation task, as they lack the spatial information, or the statistical diversity used by multi-microphone algorithms. Here we address the problem of extracting several desired sources in a reverberant environment containing both non-stationary (competing speakers) and stationary interferences.

Two families of microphone array algorithms can be defined, namely, the Blind Source Separation (BSS) family and the beamforming family. BSS aims at separating all the involved sources, regardless of their attribution to the desired or interfering sources [1]. On the other hand, the beamforming family of algorithms, concentrate on enhancing the sum of the desired sources while treating all other signals as interfering sources. The BSS family of algorithms exploit the independence of the involved sources. Independent Component Analysis (ICA) algorithms [2], [3] are commonly applied for solving the BSS problem. The ICA algorithms are distinguished by the way the source independence is imposed. Commonly used techniques include *second-order statistics* [4], *high-order statistics* [5], and *Information theoretic* based measures [6]. BSS methods can also be used in reverberant environments, but they tend to get very complex (for time domain approaches [7]) or have an inherent problem of *permutation and gain ambiguity* [8] (for frequency domain algorithms [3]).

Our proposed algorithm belongs to the beamformers family of algorithms. The term beamforming refers to the design of a spatio-temporal filter. Broadband arrays comprise a set of filters, applied to each received microphone signal, followed by a summation operation. The main objective of the beamformer is to extract a desired signal, impinging on the array from a specific position, out of noisy measurements thereof. The simplest structure is the *delay-and-sum* beamformer, which first compensates for the relative delay between distinct microphone signals and then sums the steered signal to form a single output. This beamformer, which is still widely used, can be very effective in mitigating noncoherent, i.e., spatially white, noise sources, provided that the number of microphones is relatively high. However, if the noise source is coherent, the Noise Reduction (NR) is strongly dependent on the direction of arrival of the noise signal. Consequently, the performance of the delay-and-sum beamformer in reverberant environments is often insufficient. Jan and Flanagan [9] extended the delay-and-sum concept by introducing the so called *filter-and-sum* beamformer. This structure, designed for multipath environments, namely reverberant enclosures, replaces the simpler delay compensator with a matched filter. The array beam-pattern can generally be designed to have a specified response. This can be done by properly setting the values of the multichannel filters weights. Statistically optimal beamformers are designed based on the statistical properties of the desired and interference signals. In general, they aim at enhancing the desired signals, while rejecting the interfering signals. Several criteria can be applied in the design of the beamformer, e.g., Maximum Signal to Noise Ratio (MSNR), minimum mean-squared error (MMSE), Minimum Variance Distortionless Response (MVDR) and LCMV. A summary of several design criteria can be found in [10], [11]. Cox et al. [12] introduced an improved adaptive beamformer that maintains a set of linear constraints as well as a quadratic inequality constraint.

In [13] a Multichannel Wiener Filter (MWF) technique has been proposed that produces a Minimum Mean Squared Error (MMSE) estimate of the desired speech component in one of the microphone signals, hence simultaneously performing noise reduction and limiting speech distortion. In addition, the MWF is able to take speech distortion into account in its optimization criterion, resulting in the Speech Distortion Weighted Multichannel Wiener Filter (SDW-MWF) [14]. In a MVDR beamformer [15], [16], the power of the output signal is minimized under the constraint that signals arriving from the assumed direction of the desired speech source are processed without distortion. A widely studied adaptive implementation of this beamformer is the GSC [17]. The standard GSC consists of a spatial pre-processor, i.e. a FBF and a BM, combined with a multichannel ANC. The FBF provides a spatial focus on the speech source, creating a so-called speech reference; the BM steers nulls in the direction of the speech source, creating so-called noise references; and the multichannel ANC eliminates the noise components in the speech reference that are correlated with the noise references. Several researchers (e.g. Er and Cantoni [18]) have proposed modifications to the MVDR for dealing with multiple linear constraints, denoted LCMV. Their work was motivated by the desire to apply further control to the array/beamformer beam-pattern, beyond that of a steer-direction gain constraint. Hence, the LCMV can be applied for constructing a beam-pattern satisfying certain constraints for a set of directions, while minimizing the array response in all other directions. Breed and Strauss [19] proved that the LCMV extension has also an equivalent GSC structure, which decouples the constraining and the minimization operations. The GSC structure was reformulated in the frequency domain, and extended to deal with the more complicated general ATFs case by Affes and Grenier [20] and later by Gannot et al. [21]. The latter frequency-domain version, which takes into account the reverberant nature of the enclosure, was nicknamed the Transfer Function Generalized Sidelobe Canceler (TF-GSC).

Several beamforming algorithms based on subspace methods were developed. Ephraim and Van Trees [22] considered the single microphone scenario. The Eigenvalue Decomposition (EVD) of the noisy speech correlation matrix is used to determine the signal and noise subspaces. Each of the eigenvalues of the signal subspaces is then processed to obtain the minimum distorted speech signal under a permissible level of residual noise at the output. Hu and Loizou [23] extended this method to deal with the colored noise case by using the GEVD rather than the EVD as in the white noise case. Gazor et al. [24] propose to use a beamformer based on the MVDR criterion and implemented as a GSC to enhance a narrowband signal contaminated by additive noise and received by multiple sensors. Under the assumption that the Direction of Arrival (DOA) entirely determines the transfer function relating the source and the microphones, it is shown that determining the signal subspace suffices for the construction of the algorithm. An efficient DOA tracking system, based on the Projection Approximation Subspace Tracking (deflation) (PASTd) algorithm [25] is derived. An extension to the wide-band case is presented by the same authors [26]. However the demand for a delay-only impulse response is still not relaxed. Affes and Grenier [20] apply the PASTd algorithm to enhance speech signal contaminated by spatially white noise, where arbitrary ATFs relate the speaker and the microphone array. The algorithm proves to be efficient in a simplified trading-room scenario, where the Direct to Reverberant Ratio (DRR) is relatively high and the reverberation time relatively low. Doclo and Moonen [27] extend the structure to deal with the more complicated colored noise case by using the Generalized Singular Value Decomposition (GSVD) of the received data matrix. Warsitz et al. [28] propose to replace the BM in [21]. They use a new BM based on the GEVD of the received microphone data, providing an indirect estimation of the ATFs relating the desired speaker and the microphones.

Affes et al. [29] extend the structure presented in [24] to deal with the multi-source case. The constructed multi-source GSC, which enables multiple target tracking, is based on the PASTd algorithm and on constraining the estimated steering vector to the array manifold. Asano et al. [30] address the problem of enhancing multiple speech sources in a non-reverberant environment. The Multiple Signal Classification (MUSIC) method, proposed by Schmidt [31], is utilized to estimate the number of sources and their respective steering vectors. The noise components are reduced by manipulating the generalized eigenvalues of the data matrix. Based on the subspace estimator, a LCMV beamformer is constructed. The LCMV constraints set consists of two subsets: one for maintaining the desired sources and the second for mitigating the interference sources. Benesty et al. [32] also address beamforming structures for multiple input signals. In their contribution, derived in the time-domain, the microphone array is treated as a Multiple Input Multiple Output (MIMO) system. In their experimental study, it is assumed that the filters relating the sources and the microphones are a priori known, or alternatively, that the sources are not active simultaneously. Reuven et al. [33] deal with the scenario in which one desired source and one competing speech source coexist in noisy and reverberant environment. The resulting algorithm, denoted Dual source Transfer Function Generalized Sidelobe Canceler (DTF-GSC) is tailored to the specific problem of two sources and cannot be easily generalized to the multiple desired and interference sources.

In this paper, we propose a novel beamforming technique, aiming at the extraction of multiple desired speech sources, while attenuating several interfering sources (both stationary and non-stationary) in a reverberant environment. The resulting LCMV beamformer is first reformulated in a GSC structure. It is shown that in the spatially-white sensor noise case only the FBF branch is active. The ANC branch contributes to the interference reduction only when the constraints set is not accurately estimated. We derive a practical method for estimating all components of the eigenspace-based beamformer.

We first show that the desired signals' RTFs (defined as the ratio between ATFs which relate the speech sources and the microphones) and a basis of the interference subspace suffice for the construction of the beamformer. The RTFs of the desired signals are estimated by applying the GEVD procedure to the received signals' PSD matrix and the stationary noise PSD matrix. A basis spanning the interference subspace is estimated by collecting eigenvectors, calculated in segments in which the non-stationary signals are active and the desired signals are inactive. A novel method, based on the QRD, of reducing the rank of interference subspace is derived. This procedure relaxes the common requirement for non-overlapping activity periods of the interference signals.

The structure of the paper is as follows. In Sec. II the problem of extracting multiple desired sources contaminated by multiple interference in a reverberant environment is introduced. In Sec. III the multiple constrained LCMV beamformer is presented and stated in a GSC structure. In Sec. IV we describe a novel method for estimating the interferences' subspace as well as a GEVD based method for estimating the RTFs of the desired sources. The entire algorithm is summarized in Sec. V. In Sec. VI we present a typical test scenario, discuss some implementation considerations of the algorithm, and show experimental results for both a simulated room and a real conference room scenarios. We draw some conclusions and summarize our work in Sec. VII.

II. Problem Formulation

Consider the general problem of extracting K desired sources, contaminated by N_s stationary interfering sources and N_{ns} non-stationary sources. The signals are received by M sensors arranged in an arbitrary array. Each of the involved signals undergo filtering by the RIR before being picked up by the microphones. The reverberation effect can be modeled by a Finite Impulse Response (FIR) filter operating on the sources. The signal received by the mth sensor is given by:

$$z_m(n) = \sum_{i=1}^{K} s_i^d(n) * h_{im}^d(n) + \sum_{i=1}^{N_s} s_i^s(n) * h_{im}^s(n) + \sum_{i=1}^{N_{ns}} s_i^{ns}(n) * h_{im}^{ns}(n) + v_m(n) \tag{1}$$

where $s_1^d(n), \ldots, s_K^d(n)$, $s_1^s(n), \ldots, s_{N_s}^s(n)$ and $s_1^{ns}(n), \ldots, s_{N_{ns}}^{ns}(n)$ are the desired sources, the stationary and non-stationary interfering sources in the room, respectively. We define $h_{im}^d(n)$, $h_{im}^s(n)$ and $h_{im}^{ns}(n)$ to be the Linear Time Invariant (LTI) RIRs relating the desired sources, the interfering sources, and each sensor m, respectively. $v_m(n)$ is the sensor noise. $z_m(n)$ is transformed into the Short Time Fourier Transform (STFT) domain with a rectangular window of length N_{DFT}, yielding:

$$z_m(\ell, k) = \sum_{i=1}^{K} s_i^d(\ell, k) h_{im}^d(\ell, k) + \sum_{i=1}^{N_s} s_i^s(\ell, k) h_{im}^s(\ell, k) + \sum_{i=1}^{N_{ns}} s_i^{ns}(\ell, k) h_{im}^{ns}(\ell, k) + v_m(\ell, k) \tag{2}$$

where ℓ is the frame number and k is the frequency index. The assumption that the window length is much larger then the RIR length ensures the Multiplicative Transfer Function (MTF) approximation [34] validness.

The received signals in (2) can be formulated in a vector notation:

$$\begin{aligned} \boldsymbol{z}(\ell,k) &= \boldsymbol{H}^d(\ell,k)\boldsymbol{s}^d(\ell,k) + \boldsymbol{H}^s(\ell,k)\boldsymbol{s}^s(\ell,k) + \boldsymbol{H}^{ns}(\ell,k)\boldsymbol{s}^{ns}(\ell,k) + \boldsymbol{v}(\ell,k) \\ &= \boldsymbol{H}(\ell,k)\boldsymbol{s}(\ell,k) + \boldsymbol{v}(\ell,k) \end{aligned} \tag{3}$$

where

$$\begin{aligned} \boldsymbol{z}(\ell,k) &\triangleq \begin{bmatrix} z_1(\ell,k) & \dots & z_M(\ell,k) \end{bmatrix}^T \\ \boldsymbol{v}(\ell,k) &\triangleq \begin{bmatrix} v_1(\ell,k) & \dots & v_M(\ell,k) \end{bmatrix}^T \\ \boldsymbol{h}_i^d(\ell,k) &\triangleq \begin{bmatrix} h_{i1}^d(\ell,k) & \dots & h_{iM}^d(\ell,k) \end{bmatrix}^T \qquad i = 1,\dots,K \\ \boldsymbol{h}_i^s(\ell,k) &\triangleq \begin{bmatrix} h_{i1}^s(\ell,k) & \dots & h_{iM}^s(\ell,k) \end{bmatrix}^T \qquad i = 1,\dots,N_s \\ \boldsymbol{h}_i^{ns}(\ell,k) &\triangleq \begin{bmatrix} h_{i1}^{ns}(\ell,k) & \dots & h_{iM}^{ns}(\ell,k) \end{bmatrix}^T \qquad i = 1,\dots,N_{ns} \end{aligned}$$

$$\begin{aligned} \boldsymbol{H}^d(\ell,k) &\triangleq \begin{bmatrix} \boldsymbol{h}_1^d(\ell,k) & \dots & \boldsymbol{h}_K^d(\ell,k) \end{bmatrix} \\ \boldsymbol{H}^s(\ell,k) &\triangleq \begin{bmatrix} \boldsymbol{h}_1^s(\ell,k) & \dots & \boldsymbol{h}_{N_s}^s(\ell,k) \end{bmatrix} \\ \boldsymbol{H}^{ns}(\ell,k) &\triangleq \begin{bmatrix} \boldsymbol{h}_1^{ns}(\ell,k) & \dots & \boldsymbol{h}_{N_{ns}}^{ns}(\ell,k) \end{bmatrix} \\ \boldsymbol{H}^i(\ell,k) &\triangleq \begin{bmatrix} \boldsymbol{H}^s(\ell,k) & \boldsymbol{H}^{ns}(\ell,k) \end{bmatrix} \\ \boldsymbol{H}(\ell,k) &\triangleq \begin{bmatrix} \boldsymbol{H}^d(\ell,k) & \boldsymbol{H}^s(\ell,k) & \boldsymbol{H}^{ns}(\ell,k) \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \boldsymbol{s}^d(\ell,k) &\triangleq \begin{bmatrix} s_1^d(\ell,k) & \dots & s_K^d(\ell,k) \end{bmatrix}^T \\ \boldsymbol{s}^s(\ell,k) &\triangleq \begin{bmatrix} s_1^s(\ell,k) & \dots & s_{N_s}^s(\ell,k) \end{bmatrix}^T \\ \boldsymbol{s}^{ns}(\ell,k) &\triangleq \begin{bmatrix} s_1^{ns}(\ell,k) & \dots & s_{N_{ns}}^{ns}(\ell,k) \end{bmatrix}^T \\ \boldsymbol{s}(\ell,k) &\triangleq \begin{bmatrix} (\boldsymbol{s}^d(\ell,k))^T & (\boldsymbol{s}^s(\ell,k))^T & (\boldsymbol{s}^{ns}(\ell,k))^T \end{bmatrix}^T . \end{aligned}$$

Assuming the desired speech signals, the interference and the noise signals to be uncorrelated, the received signals' correlation matrix is given by:

$$\begin{gathered} \boldsymbol{\Phi}_{zz}(\ell,k) = \boldsymbol{H}^d(\ell,k)\boldsymbol{\Lambda}^d(\ell,k)\big(\boldsymbol{H}^d(\ell,k)\big)^\dagger + \\ \boldsymbol{H}^s(\ell,k)\boldsymbol{\Lambda}^s(\ell,k)\big(\boldsymbol{H}^s(\ell,k)\big)^\dagger + \boldsymbol{H}^{ns}(\ell,k)\boldsymbol{\Lambda}^{ns}(\ell,k)\big(\boldsymbol{H}^{ns}(\ell,k)\big)^\dagger + \boldsymbol{\Phi}_{vv}(\ell,k) \end{gathered} \tag{4}$$

where

$$\begin{aligned}\boldsymbol{\Lambda}^d(\ell,k) &\triangleq \text{diag}\left(\begin{bmatrix} (\sigma_1^d(\ell,k))^2 & \dots & (\sigma_K^d(\ell,k))^2 \end{bmatrix}\right)\\ \boldsymbol{\Lambda}^s(\ell,k) &\triangleq \text{diag}\left(\begin{bmatrix} (\sigma_1^s(\ell,k))^2 & \dots & (\sigma_{N_s}^s(\ell,k))^2 \end{bmatrix}\right)\\ \boldsymbol{\Lambda}^{ns}(\ell,k) &\triangleq \text{diag}\left(\begin{bmatrix} (\sigma_1^{ns}(\ell,k))^2 & \dots & (\sigma_{N_{ns}}^{ns}(\ell,k))^2 \end{bmatrix}\right).\end{aligned}$$

where $(\bullet)^\dagger$ is the conjugate-transpose operation, and diag $(\bullet)$ is a square matrix with the vector in brackets on its main diagonal. $\boldsymbol{\Phi}_{vv}(\ell,k)$ is the sensor noise correlation matrix usually assumed to be spatially-white, i.e. $\boldsymbol{\Phi}_{vv}(\ell,k) = \sigma_v^2 \boldsymbol{I}_{M\times M}$ where $\boldsymbol{I}_{M\times M}$ is the identity matrix.

III. Proposed Method

In this section the proposed algorithm is derived. First, the LCMV beamformer is introduced and reformulated in a GSC structure[1]. In the following subsections we define a set of constraints used for extracting the desired sources and mitigating the interference sources. Then we replace the constraints set by an equivalent set which can be more easily estimated. Finally, we relax our constraint for extracting the exact input signals, as transmitted by the sources, and replace it by the extraction of the desired speech components at an arbitrarily chosen microphone. The outcome of the latter, a modified constraints set, will constitute a feasible system.

A. The LCMV Beamformer and the GSC Formulation

A beamformer is a system realized by processing each of the sensor signals $z_m(k,\ell)$ by the filters $w_m^*(\ell,k)$ and summing the outputs. The beamformer output $y(\ell,k)$ is given by

$$y(\ell,k) = \boldsymbol{w}^\dagger(\ell,k)\boldsymbol{z}(\ell,k) \tag{5}$$

where

$$\boldsymbol{w}(\ell,k) = \begin{bmatrix} w_1(\ell,k),\dots,w_M(\ell,k) \end{bmatrix}^T. \tag{6}$$

The filters are set to satisfy the LCMV criterion with multiple constraints:

$$\boldsymbol{w}(\ell,k) = \underset{\boldsymbol{w}}{\text{argmin}}\{\boldsymbol{w}^\dagger(\ell,k)\boldsymbol{\Phi}_{zz}(\ell,k)\boldsymbol{w}(\ell,k)\} \text{ subject to } \boldsymbol{C}^\dagger(\ell,k)\boldsymbol{w}(\ell,k) = \boldsymbol{g}(\ell,k) \tag{7}$$

where

$$\boldsymbol{C}^\dagger(\ell,k)\boldsymbol{w}(\ell,k) = \boldsymbol{g}(\ell,k) \tag{8}$$

is the constraints set. The well-known solution to (7) is given by [10]:

$$\boldsymbol{w}(\ell,k) = \boldsymbol{\Phi}_{zz}^{-1}(\ell,k)\boldsymbol{C}(\ell,k)\left(\boldsymbol{C}^\dagger(\ell,k)\boldsymbol{\Phi}_{zz}^{-1}(\ell,k)\boldsymbol{C}(\ell,k)\right)^{-1}\boldsymbol{g}(\ell,k) \tag{9}$$

[1] The authors wish to express their gratitude to Dr. Emanuël Habets for the fruitful discussions and for his assistance in clarifying the GSC formulation.

The LCMV can be implemented using the GSC formulation [19]. In this structure the filter set $\boldsymbol{w}(\ell,k)$ can be split to two orthogonal components [10], one in the constraint plane and the other in the orthogonal subspace:

$$\begin{aligned}\boldsymbol{w}(\ell,k) &= \boldsymbol{w}_0(\ell,k) - \boldsymbol{w}^n(\ell,k)\\ &= \boldsymbol{w}_0(\ell,k) - \boldsymbol{B}^\dagger(\ell,k)\boldsymbol{q}(\ell,k)\end{aligned} \tag{10}$$

where $\boldsymbol{B}(\ell,k)$ is the projection matrix to the "null" subspace, denoted BM, i.e. $\boldsymbol{BC}(\ell,k)=0$. $\boldsymbol{w}_0(\ell,k)$ is the FBF satisfying the constraints set, $\boldsymbol{w}^n(\ell,k)$ is orthogonal to $\boldsymbol{w}_0(\ell,k)$, and $\boldsymbol{q}(\ell,k)$ is a set of ANC filters adjusted to obtain the (unconstrained) minimization. In the original GSC structure the filters $\boldsymbol{q}(\ell,k)$ are calculated adaptively using the Least Mean Squares (LMS) algorithm.

Using [10] the FBF is given by:

$$\boldsymbol{w}_0(\ell,k) = \boldsymbol{C}(\ell,k)\big(\boldsymbol{C}^\dagger(\ell,k)\boldsymbol{C}(\ell,k)\big)^{-1}\boldsymbol{g}(\ell,k). \tag{11}$$

The BM can be determined as the projection matrix to the null subspace of the column-space of $\boldsymbol{C}$:

$$\boldsymbol{B}(\ell,k) = \boldsymbol{I}_{M\times M} - \boldsymbol{C}(\ell,k)\big(\boldsymbol{C}^\dagger(\ell,k)\boldsymbol{C}(\ell,k)\big)^{-1}\boldsymbol{C}^\dagger(\ell,k) \tag{12}$$

and a closed-form (Wiener) solution for $\boldsymbol{q}(\ell,k)$ is:

$$\boldsymbol{q}(\ell,k) = \big(\boldsymbol{B}(\ell,k)\boldsymbol{\Phi}_{zz}(\ell,k)\boldsymbol{B}^\dagger(\ell,k)\big)^{-1}\boldsymbol{B}(\ell,k)\boldsymbol{\Phi}_{zz}(\ell,k)\boldsymbol{w}_0(\ell,k). \tag{13}$$

A block diagram of the GSC structure is depicted in Fig. 1. The GSC comprises three blocks. The FBF is responsible for the alignment of the desired sources and the BM blocks the directional signals. The output of the BM, denoted $\mathbf{u}(\ell,k)$ is then processed by the ANC filters $\boldsymbol{q}(\ell,k)$ for further reduction of the residual interference signals at the output. More details regarding each block of the GSC blocks will be given in the subsequent subsections for the various definitions of the constraints set.

B. The constraints set

We start with the straightforward approach, in which the beam-pattern is constrained to cancel out all interfering sources while maintaining all desired sources (for each frequency bin). Note, that unlike the DTF-GSC approach [33], the stationary noise sources are treated similarly to the interference (non-stationary) sources. We therefore define the following constraints. For each desired source $\{s_i^d\}_{i=1}^K$ we apply the constraint:

$$\big(\boldsymbol{h}_i^d(\ell,k)\big)^\dagger\boldsymbol{w}(\ell,k) = 1,\ i=1,\dots,K. \tag{14}$$

For each interfering source, both stationary and non-stationary, $\{s_i^s\}_{i=1}^{N_s}$ and $\{s_j^{ns}\}_{j=1}^{N_{ns}}$, we apply:

$$\big(\boldsymbol{h}_i^s(\ell,k)\big)^\dagger\boldsymbol{w}(\ell,k) = 0, \tag{15}$$

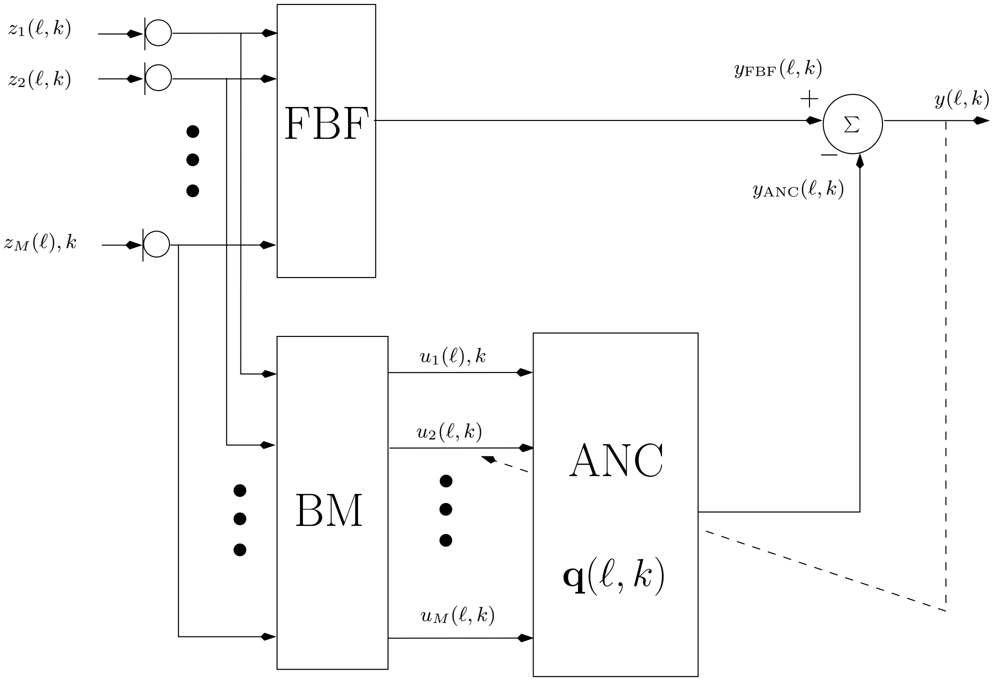


Fig. 1. The proposed LCMV beamformer reformulated in a GSC structure.

and

$$\left(\boldsymbol{h}_j^{ns}(\ell, k)\right)^{\dagger}\boldsymbol{w}(\ell, k) = 0. \tag{16}$$

Define $N \triangleq K + N_s + N_{ns}$ the total number of signals in the environment (including the desired sources, stationary interference signals, and the non-stationary interference signals). Assuming the column-space of $\boldsymbol{H}(\ell, k)$ is linearly independent (i.e. the ATFs are independent), it is obvious that for the solution in (9) to exist we require that the number of microphones will be greater or equal the number of constraints, namely $M \geq N$. It is also understood that whenever the constraints contradict each other, the desired signal constraints will be preferred.

Summarizing, we have a constraint matrix:

$$\boldsymbol{C}(\ell, k) \triangleq \boldsymbol{H}(\ell, k) \tag{17}$$

and a desired response vector:

$$\boldsymbol{g} \triangleq \left[\ \underbrace{1 \ldots 1}_{K}\ \ \underbrace{0 \ldots 0}_{N-K}\ \right]^T. \tag{18}$$

Under these definitions, and using (3) and (11), the FBF output is given by:

$$\begin{aligned}
y_{\text{FBF}}(\ell,k) &= \boldsymbol{w}_0^\dagger(\ell,k)\boldsymbol{z}(\ell,k) = \\
&\boldsymbol{g}^\dagger\left(\boldsymbol{C}^\dagger(\ell,k)\boldsymbol{C}(\ell,k)\right)^{-1}\boldsymbol{C}^\dagger(\ell,k)\left(\boldsymbol{H}(\ell,k)\boldsymbol{s}(\ell,k)+\boldsymbol{v}(\ell,k)\right) = \\
&\boldsymbol{g}^\dagger\boldsymbol{s}(\ell,k)+\boldsymbol{g}^\dagger\left(\boldsymbol{H}^\dagger(\ell,k)\boldsymbol{H}(\ell,k)\right)^{-1}\boldsymbol{H}^\dagger(\ell,k)\boldsymbol{v}(\ell,k) = \\
&\sum_{i=1}^{K} s_i^d(\ell,k)+\boldsymbol{g}^\dagger\left(\boldsymbol{H}^\dagger(\ell,k)\boldsymbol{H}(\ell,k)\right)^{-1}\boldsymbol{H}^\dagger(\ell,k)\boldsymbol{v}(\ell,k). \qquad (19)
\end{aligned}$$

Using (13) and (4) the ANC filters are given by:

$$\begin{aligned}
\boldsymbol{q}(\ell,k) &= \left(\boldsymbol{B}(\ell,k)\boldsymbol{\Phi}_{zz}(\ell,k)\boldsymbol{B}^\dagger(\ell,k)\right)^{-1}\boldsymbol{B}(\ell,k)\boldsymbol{\Phi}_{zz}(\ell,k)\boldsymbol{w}_0(\ell,k) \qquad (20)\\
&= \left(\boldsymbol{B}(\ell,k)\boldsymbol{\Phi}_{zz}(\ell,k)\boldsymbol{B}^\dagger(\ell,k)\right)^{-1}\boldsymbol{B}(\ell,k)\left(\boldsymbol{H}(\ell,k)\boldsymbol{\Lambda}(\ell,k)\boldsymbol{H}^\dagger(\ell,k)+\boldsymbol{\Phi}_{vv}(\ell,k)\right)\boldsymbol{w}_0(\ell,k).
\end{aligned}$$

Hence, using $\boldsymbol{B}(\ell,k)\boldsymbol{C}(\ell,k)=\boldsymbol{B}(\ell,k)\boldsymbol{H}(\ell,k)=0$ we have:

$$\boldsymbol{q}(\ell,k) = \left(\boldsymbol{B}(\ell,k)\boldsymbol{\Phi}_{zz}(\ell,k)\boldsymbol{B}^\dagger(\ell,k)\right)^{-1}\boldsymbol{B}(\ell,k)\boldsymbol{\Phi}_{vv}(\ell,k)\boldsymbol{w}_0(\ell,k). \qquad (21)$$

Now, using (11) we have

$$\begin{aligned}
&\boldsymbol{q}(\ell,k) = \qquad (22)\\
&\left(\boldsymbol{B}(\ell,k)\boldsymbol{\Phi}_{zz}(\ell,k)\boldsymbol{B}^\dagger(\ell,k)\right)^{-1}\boldsymbol{B}(\ell,k)\boldsymbol{\Phi}_{vv}(\ell,k)\boldsymbol{C}(\ell,k)\left(\boldsymbol{C}^\dagger(\ell,k)\boldsymbol{C}(\ell,k)\right)^{-1}\boldsymbol{g}
\end{aligned}$$

For the spatially-white sensor noise case, $\boldsymbol{\Phi}_{vv}(\ell,k)=\sigma_v^2\boldsymbol{I}_{M\times M}$, the ANC filters simplifies to:

$$\boldsymbol{q}(\ell,k) = \sigma_v^2\left(\boldsymbol{B}(\ell,k)\boldsymbol{\Phi}_{zz}(\ell,k)\boldsymbol{B}^\dagger(\ell,k)\right)^{-1}\boldsymbol{B}(\ell,k)\boldsymbol{C}(\ell,k)\left(\boldsymbol{C}^\dagger(\ell,k)\boldsymbol{C}(\ell,k)\right)^{-1}\boldsymbol{g}. \qquad (23)$$

Using once more the projection identity $\boldsymbol{B}(\ell,k)\boldsymbol{C}(\ell,k)=0$ we finally conclude that $\boldsymbol{q}(\ell,k)=0$. Hence, the lower branch of the GSC beamformer has no contribution to the output signal in this case, and the LCMV simplifies to the FBF beamformer, i.e. no minimization of the output power is performed.

The LCMV beamformer output is therefore given by (19). It comprises a sum of two terms: the first is the sum of all the desired sources and the second is the response of the array to the sensor noise.

C. Equivalent constraints set

The matrix $\boldsymbol{C}(\ell,k)$ in (17) comprises the ATFs relating the sources and the microphones $\boldsymbol{h}_i^d(\ell,k)$, $\boldsymbol{h}_i^s(\ell,k)$ and $\boldsymbol{h}_i^{ns}(\ell,k)$. Hence, the solution given in (11) requires an estimate of the various filters. Obtaining such estimates might be a cumbersome task in practical scenarios, where it is usually required that the sources are not active simultaneously (see e.g. [32]). We will show now that the actual ATFs of the interfering sources can be replaced by the basis vectors spanning the same interference subspace, without sacrificing the accuracy of the solution.

Let

$$N_i \triangleq N_s + N_{ns} \qquad (24)$$

be the number of interferences, both stationary and non-stationary, in the environment. For conciseness we assume that the ATFs of the interfering sources are linearly independent at each frequency bin, and define $\boldsymbol{E} \triangleq [\boldsymbol{e}_1 \ \ldots \ \boldsymbol{e}_{N_i}]$ to be any basis[2] that spans the column space of the interfering sources $\boldsymbol{H}^i(\ell,k) = [\boldsymbol{H}^s(\ell,k) \ \boldsymbol{H}^{ns}(\ell,k)]$. Hence, the following identity holds:

$$\boldsymbol{H}^i(\ell,k) = \boldsymbol{E}(\ell,k)\boldsymbol{\Theta}(\ell,k) \tag{25}$$

where $\boldsymbol{\Theta}_{N_i \times N_i}(\ell,k)$ is comprised of the projection coefficients of the original ATFs on the basis vectors. When the ATFs associated with the interference signals are linearly independent, $\boldsymbol{\Theta}_{N_i \times N_i}(\ell,k)$ is an invertible matrix.

Define

$$\tilde{\boldsymbol{\Theta}}(\ell,k) \triangleq \begin{bmatrix} \boldsymbol{I}_{K\times K} & \mathrm{O}_{K\times N_i} \\ \mathrm{O}_{N_i\times K} & \boldsymbol{\Theta}(\ell,k) \end{bmatrix}_{N\times N} \tag{26}$$

where $\boldsymbol{I}_{K\times K}$ is a $K \times K$ identity matrix. Multiplication by $(\tilde{\boldsymbol{\Theta}}^\dagger(\ell,k))^{-1}$ of both sides of the original constraints set in (8), with the definitions (17)-(18), yields:

$$(\tilde{\boldsymbol{\Theta}}^\dagger(\ell,k))^{-1}\boldsymbol{C}^\dagger(\ell,k)\boldsymbol{w}(\ell,k) = (\tilde{\boldsymbol{\Theta}}^\dagger(\ell,k))^{-1}\boldsymbol{g}. \tag{27}$$

Starting with the left-hand-side of (27) we have:

$$\begin{aligned}
&(\tilde{\boldsymbol{\Theta}}^\dagger(\ell,k))^{-1}\boldsymbol{C}^\dagger(\ell,k)\boldsymbol{w}(\ell,k) \\
&= \begin{bmatrix} \boldsymbol{I}_{K\times K} & \mathrm{O}_{K\times N_i} \\ \mathrm{O}_{N_i\times K} & (\boldsymbol{\Theta}^\dagger(\ell,k))^{-1} \end{bmatrix} \begin{bmatrix} (\boldsymbol{H}^d(\ell,k))^\dagger \\ (\boldsymbol{H}^i(\ell,k))^\dagger \end{bmatrix} \boldsymbol{w}(\ell,k) \\
&= \begin{bmatrix} (\boldsymbol{H}^d(\ell,k))^\dagger \\ (\boldsymbol{\Theta}^{-1}(\ell,k))^\dagger(\boldsymbol{H}^i(\ell,k))^\dagger \end{bmatrix} \boldsymbol{w}(\ell,k) \\
&= \begin{bmatrix} (\boldsymbol{H}^d(\ell,k))^\dagger \\ (\boldsymbol{H}^i(\ell,k)\boldsymbol{\Theta}^{-1}(\ell,k))^\dagger \end{bmatrix} \boldsymbol{w}(\ell,k) \\
&= \begin{bmatrix} (\boldsymbol{H}^d(\ell,k)) & \boldsymbol{E}(\ell,k) \end{bmatrix}^\dagger \boldsymbol{w}(\ell,k) \\
&\triangleq \dot{\boldsymbol{C}}^\dagger(\ell,k)\boldsymbol{w}(\ell,k)
\end{aligned}$$

where the equivalent constraint matrix is defined as

$$\dot{\boldsymbol{C}}(\ell,k) \triangleq \begin{bmatrix} \boldsymbol{H}^d(\ell,k) & \boldsymbol{E}(\ell,k) \end{bmatrix}. \tag{28}$$

[2] If this linear independency assumption does not hold, the rank of the basis can be smaller than N_i in several frequency bins. In this contribution we assume the interference subspace to be full rank.

For the right-hand-side of (27) we have:

$$
\begin{aligned}
&(\tilde{\boldsymbol{\Theta}}^{\dagger}(\ell,k))^{-1}\boldsymbol{g} = \\
&\begin{bmatrix} \boldsymbol{I}_{K\times K} & \mathbf{O}_{K\times N_i} \\ \mathbf{O}_{N_i\times K} & (\boldsymbol{\Theta}^{\dagger}(\ell,k))^{-1} \end{bmatrix}\boldsymbol{g} = \\
&= \begin{bmatrix} (\underbrace{1\ \ldots\ 1}_{K})\boldsymbol{I}_{K\times K} & (\underbrace{0\ \ldots\ 0}_{N-K})(\boldsymbol{\Theta}(\ell,k))^{-1} \end{bmatrix}^{\dagger} = \boldsymbol{g}.
\end{aligned}
$$

Hence, it is shown that $\boldsymbol{w}(\ell,k)$ that satisfies the original constraints set $\boldsymbol{C}^{\dagger}(\ell,k)\boldsymbol{w}(\ell,k) = \boldsymbol{g}$ also satisfies the equivalent constraints set

$$\dot{\boldsymbol{C}}^{\dagger}(\ell,k)\boldsymbol{w}(\ell,k) = \boldsymbol{g}. \tag{29}$$

Since the constraint is satisfied by the FBF branch, and since the original LCMV beamformer and the LCMV beamformer with the equivalent constraints set are derived similarly, it is also guaranteed that $\boldsymbol{q}(\ell,k)$ in the later structure becomes zero for the spatially-white sensor noise case.

D. *Modified constraints set*

Both the original and equivalent constraints sets in (17) and (28) respectively, require estimates of the desired sources ATFs $\boldsymbol{H}^{d}(\ell,k)$. Estimating these ATFs might be a cumbersome task, due to the large order of the respective RIRs. In the current section we relax our demand for a distortionless beamformer [as depicted in the definition of g in (18)] and replace it by constraining the output signal to be comprised of the desired speech components at an arbitrarily chosen microphone.

Define a modified vector of desired responses:

$$\tilde{\boldsymbol{g}}(\ell,k) = \Big[\ \underbrace{(h^{d}_{11}(\ell,k))^{*}\ \ldots\ (h^{d}_{K1}(\ell,k))^{*}}_{K}\ \ \underbrace{0\ \ldots\ 0}_{N-K}\ \Big]^{T}$$

where microphone $\#1$ was arbitrarily chosen as the reference microphone. The modified FBF satisfying the modified response $\dot{\boldsymbol{C}}^{\dagger}(\ell,k)\tilde{\boldsymbol{w}}(\ell,k) = \tilde{\boldsymbol{g}}(\ell,k)$ is then given by

$$\tilde{\boldsymbol{w}}_{0}(\ell,k) \triangleq \dot{\boldsymbol{C}}(\ell,k)\big(\dot{\boldsymbol{C}}^{\dagger}(\ell,k)\dot{\boldsymbol{C}}(\ell,k)\big)^{-1}\tilde{\boldsymbol{g}}(\ell,k). \tag{30}$$

Indeed, using the equivalence between the column subspaces of $\dot{\boldsymbol{C}}(\ell,k)$ and $\boldsymbol{H}(\ell,k)$, the FBF output is now given by:

$$
\begin{aligned}
&y_{\text{FBF}}(\ell,k) = \tilde{\boldsymbol{w}}_{0}^{\dagger}(\ell,k)\boldsymbol{z}(\ell,k) = \\
&\quad \tilde{\boldsymbol{g}}^{\dagger}(\ell,k)\big(\dot{\boldsymbol{C}}^{\dagger}(\ell,k)\dot{\boldsymbol{C}}(\ell,k)\big)^{-1}\dot{\boldsymbol{C}}^{\dagger}(\ell,k)\left(\boldsymbol{H}(\ell,k)\boldsymbol{s}(\ell,k) + \boldsymbol{v}(\ell,k)\right) = \\
&\quad \tilde{\boldsymbol{g}}^{\dagger}(\ell,k)\boldsymbol{s}(\ell,k) + \tilde{\boldsymbol{g}}^{\dagger}(\ell,k)\big(\dot{\boldsymbol{C}}^{\dagger}(\ell,k)\dot{\boldsymbol{C}}(\ell,k)\big)^{-1}\dot{\boldsymbol{C}}^{\dagger}(\ell,k)\boldsymbol{v}(\ell,k) = \\
&\quad \sum_{i=1}^{K} h^{d}_{i1}(\ell,k)s^{d}_{i}(\ell,k) + \tilde{\boldsymbol{g}}^{\dagger}(\ell,k)\big(\dot{\boldsymbol{C}}^{\dagger}(\ell,k)\dot{\boldsymbol{C}}(\ell,k)\big)^{-1}\dot{\boldsymbol{C}}^{\dagger}(\ell,k)\boldsymbol{v}(\ell,k)
\end{aligned} \tag{31}
$$

as expected from the modified constraint response. As mentioned before, estimating the desired signal ATFs is a cumbersome task. Nevertheless, in Sec. IV we will show that a practical method for estimating the RTF can be derived. We will therefore reformulate in the sequel the constraints set in terms of the RTFs.

It is easily verified that the modified desired response is related to the original desired response (18) by:

$$\tilde{\boldsymbol{g}}(\ell,k)=\tilde{\boldsymbol{\Psi}}^{\dagger}(\ell,k)\boldsymbol{g}$$

where:

$$\boldsymbol{\Psi}(\ell,k)=\text{diag}\left(\begin{bmatrix} h_{11}^{d}(\ell,k) & \ldots & h_{K1}^{d}(\ell,k)\end{bmatrix}\right)$$

and

$$\tilde{\boldsymbol{\Psi}}(\ell,k)=\begin{bmatrix}\boldsymbol{\Psi}(\ell,k) & \mathbf{O}_{K\times N_i}\\ \mathbf{O}_{N_i\times K} & \boldsymbol{I}_{N_i\times N_i}\end{bmatrix}.$$

Now, a beamformer having the modified beam-pattern should satisfy the modified constraints set:

$$\dot{\boldsymbol{C}}^{\dagger}(\ell,k)\tilde{\boldsymbol{w}}(\ell,k)=\tilde{\boldsymbol{g}}(\ell,k)=\tilde{\boldsymbol{\Psi}}^{\dagger}(\ell,k)\boldsymbol{g}.$$

Hence,

$$(\tilde{\boldsymbol{\Psi}}^{-1}(\ell,k))^{\dagger}\dot{\boldsymbol{C}}^{\dagger}(\ell,k)\tilde{\boldsymbol{w}}(\ell,k)=\boldsymbol{g}.$$

Define

$$\tilde{\boldsymbol{C}}(\ell,k)\triangleq\dot{\boldsymbol{C}}(\ell,k)\tilde{\boldsymbol{\Psi}}^{-1}(\ell,k)=\begin{bmatrix}\tilde{\boldsymbol{H}}^{d}(\ell,k) & \boldsymbol{E}(\ell,k)\end{bmatrix} \tag{32}$$

where

$$\tilde{\boldsymbol{H}}^{d}(\ell,k)\triangleq\begin{bmatrix}\tilde{\boldsymbol{h}}_{1}^{d}(\ell,k) & \ldots & \tilde{\boldsymbol{h}}_{K}^{d}(\ell,k)\end{bmatrix} \tag{33}$$

with

$$\tilde{\boldsymbol{h}}_{i}^{d}(\ell,k)\triangleq\frac{\boldsymbol{h}_{i}^{d}(\ell,k)}{h_{i1}^{d}(\ell,k)} \tag{34}$$

defined as the RTF with respect to microphone $\#1$.

Finally, the modified FBF is given by:

$$\tilde{\boldsymbol{w}}_{0}(\ell,k)\triangleq\tilde{\boldsymbol{C}}(\ell,k)\big(\tilde{\boldsymbol{C}}(\ell,k)^{\dagger}\tilde{\boldsymbol{C}}(\ell,k)\big)^{-1}\boldsymbol{g} \tag{35}$$

and its corresponding output is therefore given by:

$$\begin{aligned} y_{\text{FBF}}(\ell,k)&=\tilde{\boldsymbol{w}}_{0}^{\dagger}(\ell,k)\boldsymbol{z}(\ell,k)=\\ &\sum_{i=1}^{K}s_{i}^{d}(\ell,k)h_{i1}^{d}(\ell,k)+\boldsymbol{g}^{\dagger}\big(\tilde{\boldsymbol{C}}^{\dagger}(\ell,k)\tilde{\boldsymbol{C}}(\ell,k)\big)^{-1}\tilde{\boldsymbol{C}}^{\dagger}(\ell,k)\boldsymbol{v}(\ell,k).\end{aligned} \tag{36}$$

The modified beamformer output therefore comprises the sum of the desired sources as measured at the reference microphone (arbitrarily chosen as microphone $\#1$) and the sensor noise contribution. It is easily verified that $\tilde{\boldsymbol{B}}(\ell,k)$, the projection matrix to the modified constraint matrix $\tilde{\boldsymbol{C}}(\ell,k)$, also satisfies $\tilde{\boldsymbol{B}}(\ell,k)\boldsymbol{H}(\ell,k)=0$ (see similar arguments in [33]) and hence the ANC branch becomes zero for the spatially-white sensor noise, yielding $y(\ell,k)=y_{\text{FBF}}(\ell,k)$.

E. Residual Noise Cancellation

It was shown in the previous subsection that the proposed LCMV beamformer can be formulated in a GSC structure. Note a profound difference between the proposed method and the algorithms presented in [21] and [33]. While the purpose of the ANC in both the TF-GSC and DTF-GSC structures is to eliminate the stationary-directional noise source passing through the BM, in the proposed structure all directional signals, including the stationary directional noise signal, are treated by the FBF branch and the ANC does not contribute to the interference cancellation, when the sensor noise is spatially-white.

However, in non-ideal scenarios the ANC branch has a significant contribution to the overall performance of the proposed beamformer. The proposed method requires an estimate of the RTFs relating each of the desired sources and the microphones, and a basis that spans the ATFs relating each of the interfering source and the microphones. As these quantities are not known, we use instead estimates thereof. The estimation procedure will be discussed in Sec. IV. In case no estimation errors occur, the BM outputs consist of solely the sensor noise. When the sensor noise is spatially white, the ANC filters converge to 0, as discussed in III-B.

Due to inevitable estimation errors, the constraints set is not exactly satisfied, resulting in leakage of residual interference signals (as well as residual desired sources) to the beamformer output, as well as desired signal distortion. These residual signals do not exhibit spatial-whiteness anymore, therefore enabling the ANC filters to contribute to the noise and interference cancellation.

The adaptation rule of the ANC filters $\boldsymbol{q}(\ell, k)$ is derived in [21] and is presented in Alg. 1. We note however, that as both the desired sources and the interference sources are expected to leak through the BM, mis-convergence of the filters can be avoided by adapting $\mathbf{q}(\ell, k)$ only when the desired sources are inactive. This necessitates the application of an activity detector for the desired sources.

A comparison between the TF-GSC algorithm and the proposed method in the single desired source scenario can be found in [35].

IV. Estimation of the Constraints Matrix

In the previous sections we have shown that knowledge of the RTFs related to the desired sources and a basis that spans the subspace of the interfering sources suffice for implementing the beamforming algorithm. This section is dedicated to the estimation procedure necessary to acquire this knowledge. We start by making some restrictive assumptions regarding the activity of the sources. First, we assume that there are time segments for which none of the non-stationary sources is active. These segments are used for estimating the stationary noise PSD. Second, we assume that there are time segments in which all the desired sources are inactive. These segments are used for estimating the interfering sources subspace (with arbitrary activity pattern). Third, we assume that for every desired source, there is at least one time segment when it is the only non-stationary source active. These segments are used for estimating the RTFs of the desired sources. These assumptions, although restrictive, can be met in realistic scenarios, for which double talk only rarely occurs. A possible way to extract the activity information can be a video signal acquired in parallel to the sound acquisition. In this contribution it is however assumed that the number of desired sources and their activity pattern is available.

In the rest of this section we discuss the subspace estimation procedure. The RTF estimation procedure can be regarded, in this aspect, as a multi-source, colored-noise extension of the single source subspace estimation method proposed by Affes and Grenier [20]. We further assume that the various filters are slowly time-varying filters, i.e $\boldsymbol{H}(\ell, k) \approx \boldsymbol{H}(k)$.

A. *Interferences Subspace Estimation*

Let $\ell = \ell_1, \ldots, \ell_{N_{seg}}$, be a set of N_{seg} frames for which all desired sources are inactive. For every segment we estimate the subspace spanned by the active interferences (both stationary and non-stationary). Let $\hat{\boldsymbol{\Phi}}_{zz}(\ell_i, k)$ be a PSD estimate at the interference-only frame ℓ_i. Using the EVD we have $\hat{\boldsymbol{\Phi}}_{zz}(\ell_i, k) = \boldsymbol{E}_i(k)\boldsymbol{\Lambda}_i(k)\boldsymbol{E}_i^{\dagger}(k)$. Interference-only segments consist of both directional interference and noise components and spatially-white sensor noise. Hence, the larger eigenvalues can be attributed to the coherent signals while the lower eigenvalues to the spatially-white signals.

Define two values $\Delta\text{EV}_{\text{TH}}(k)$ and MEV_{TH}. All eigenvectors corresponding to eigenvalues that are more than $\Delta\text{EV}_{\text{TH}}$ below the largest eigenvalue or not higher than MEV_{TH} above the lowest eigenvalue, are regarded as sensor noise eigenvectors and are therefore discarded from the interference signal subspace. Assuming that the number of sensors is larger than the number of directional sources, the lowest eigenvalue level will correspond to the sensor noise variance σ_v^2. The procedure is demonstrated in Fig. 2 for the 11 microphone test scenario presented in the sequel in Sec. VI. A segment which comprises three directional sources (one stationary and two non-stationary interferences) is analyzed using the EVD by 11 microphone array (i.e. the dimensions of the multi-sensor correlation matrix is 11×11). The eigenvalue level as a function of the frequency bin is depicted in the Figure. The blue line depicts MEV_{TH} threshold and the dark green frequency-dependent line depicts the threshold $\text{EV}_{\text{TH}}(k)$. All eigenvalues that do not meet the thresholds, depicted as gray lines in the Figure, are discarded from the interference signal subspace. The number of the remaining eigenvalues as a function of the frequency bin, that are used for the interference subspace, is depicted in Fig. 3. It can be seen from the Figure that in most frequency bins the algorithm correctly identified the three directional sources. Most of the erroneous reading are found in the lower frequency band, where the directivity of the array is low, and in the upper frequency band, where the signals' power is low. The use of two thresholds is shown to increase the robustness of the procedure.

We denote the eigenvectors that passed the thresholds as $\hat{\boldsymbol{E}}_i(k)$, and their corresponding eigenvalues as $\hat{\boldsymbol{\Lambda}}_i(k)$. This procedure is repeated for each segment ℓ_i; $i = 1, 2, \ldots, N_{seg}$. These

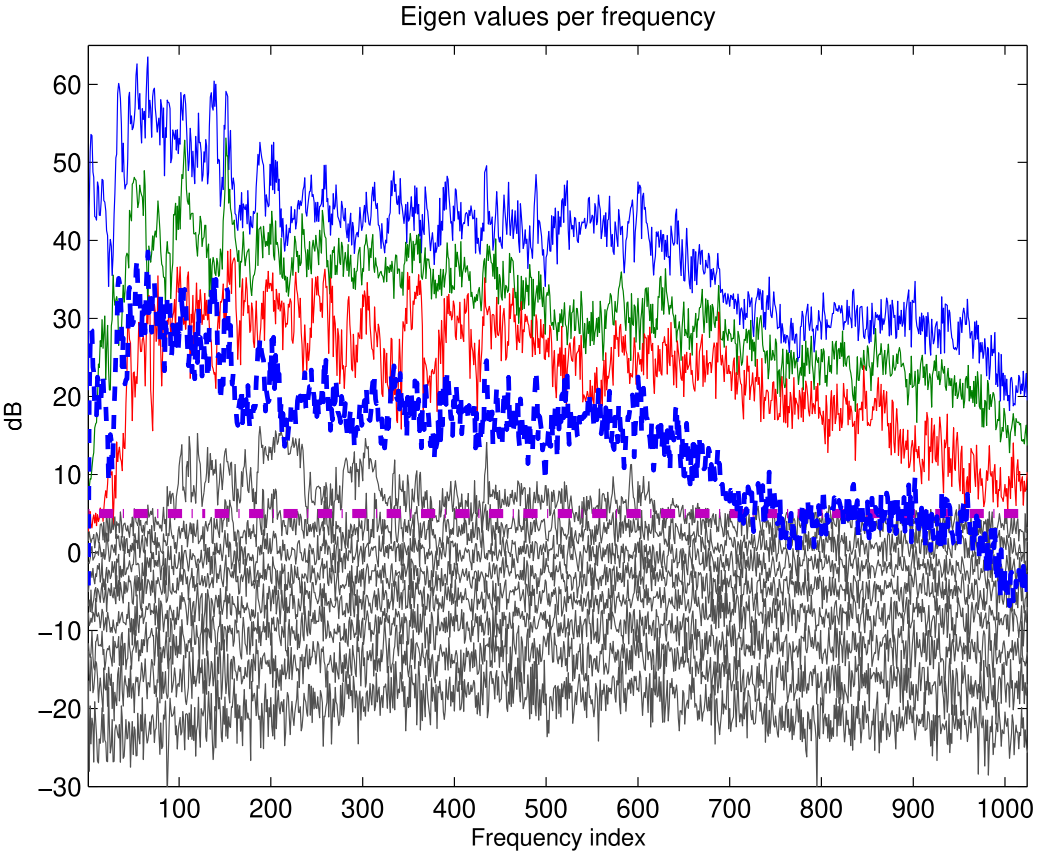


Fig. 2. Eigenvalues of an interference-only segments as a function of the frequency bin (solid thin colors). Eigenvalues that do not meet the thresholds MEV_{TH} (dashed-dotted thick pink) and $\text{EV}_{\text{TH}}(k)$ (dashed thick blue) are depicted in gray and discarded from the interference signal subspace.

vectors should span the basis of the entire interference subspace:

$$\boldsymbol{H}^i(\ell,k) = \boldsymbol{E}(\ell,k)\boldsymbol{\Theta}(\ell,k)$$

defined in (25). To guarantee that the eigenvectors $i = 1, 2, \ldots, N_{seg}$ that are common to more than one segment are not counted more than once they should be collected by the union operator:

$$\hat{\boldsymbol{E}}(k) \triangleq \bigcup_{i=1}^{N_{seg}} \hat{\boldsymbol{E}}_i(k) \tag{37}$$

where $\hat{\boldsymbol{E}}(k)$ is an estimate for the interference subspace basis $\boldsymbol{E}(\ell,k)$ assumed to be time-invariant in the observation period. Unfortunately, due to arbitrary activity of sources and estimation errors, eigenvectors that correspond to the same source can be manifested as a different eigenvector in each segment. These differences can unnecessarily inflate the number of estimated interference sources. Erroneous rank estimation is one of causes to the well-known desired signal

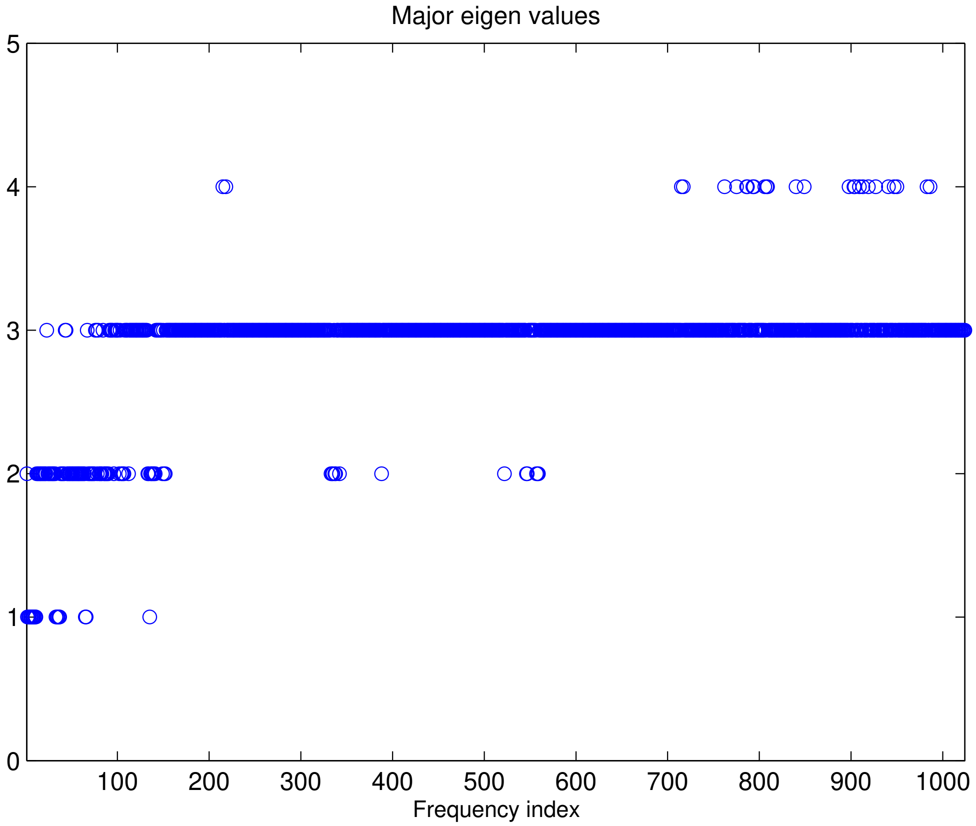


Fig. 3. The number of major eigenvalues, as a function of the frequency bin, that are used for constructing the interference subspace.

cancelation phenomenon in beamformer structures, since desired signal components may be included in the null subspace. The union operator can be implemented in many ways. Here we chose to use the QRD.

Consider the following QRD of the subspace spanned by the major eigenvectors (weighted in respect to their eigenvalues) obtained by the previous procedure:

$$\left[\begin{array}{ccc} \hat{\boldsymbol{E}}_1(k)\hat{\boldsymbol{\Lambda}}_1^{\frac{1}{2}}(k) & \ldots & \hat{\boldsymbol{E}}_{N_{seg}}(k)\hat{\boldsymbol{\Lambda}}_{N_{seg}}^{\frac{1}{2}}(k) \end{array}\right]\boldsymbol{P}(k)=\boldsymbol{Q}(k)\boldsymbol{R}(k) \tag{38}$$

where $\boldsymbol{Q}(k)$ is a unitary matrix, $\boldsymbol{R}(k)$ is an upper triangular matrix with decreasing diagonal absolute values, $\boldsymbol{P}(k)$ is a permutation matrix and $(\cdot)^{\frac{1}{2}}$ is a square root operation performed on each of the diagonal elements.

All vectors in $\boldsymbol{Q}(k)$ that correspond to values on the diagonal of $\boldsymbol{R}(k)$ that are lower than $\Delta\mathrm{U}_{\mathrm{TH}}$ below their largest value, or less then $\mathrm{MU}_{\mathrm{TH}}$ above their lowest value are not counted as basis vectors of the directional interference subspace. The collection of all vectors passing the designated thresholds, constitutes $\hat{\boldsymbol{E}}(k)$, the estimate of the interference subspace basis.

The reduction of the interference subspace rank using the QRD is further demonstrated in Table. I. Consider three segments for which one stationary and two non-stationary sources may be active (see detailed description of the test scenario in the sequel). We do not require any particular activity pattern for these sources during the considered three segments. In the first segment only one eigenvector passed the thresholds, in the second segment two eigenvectors passed the thresholds, and in the third segment three major eigenvectors were identified. In the columns of Table I associated with e_i, $1 \leq i \leq 3$ we depict the absolute value of the inner product between the normalized ATFs of each of the interference signals and the estimated eigenvector. The rotation of the eigenvectors from segment to segment is manifested by the different projections. This phenomenon can be attributed to the non-stationarity of the sources (in particular the sources can change their activity state across segments) and to estimation errors.

TABLE I

PROJECTION COEFFICIENTS OF THE INTERFERENCES' ATFS ON THE ESTIMATED BASIS AT DIFFERENT TIME SEGMENTS AND THE CORRESPONDING BASIS OBTAINED BY THE QRD BASED UNION PROCEDURE.

Interfering Source	Segment 1		Segment 2			Segment 3				QRD			
	e_1	$e\mathcal{N}$	e_1	e_2	$e\mathcal{N}$	e_1	e_2	e_3	$e\mathcal{N}$	e_1	e_2	e_3	$e\mathcal{N}$
h_1^{ns}	0.40	0.92	0.39	0.57	0.72	0.84	0.53	0.07	0.02	0.92	0.35	0.18	0.02
h_2^{ns}	0.20	0.98	0.58	0.81	0.03	0.62	0.74	0.26	0.02	0.76	0.65	0.07	0.01
h_1^{s}	1.00	0.02	0.91	0.41	0.02	0.65	0.32	0.69	0.02	0.31	0.55	0.77	0.02

Define a subspace $\mathcal{E}$ spanned by the identified eigenvectors. The value $e\mathcal{N}$ depicts the norm of the projection of the normalized ATF, associated with the row, and the null subspace orthogonal to $\mathcal{E}$. Low level of $e\mathcal{N}$ indicates that the ATF in the corresponding row can be modeled by the basis. Therefore, it is evident that only h_1^s can be modeled by the basis identified in the first segment, both h_1^s and h_2^{ns} can be modeled in the second segment, and all three ATFs, i.e. h_1^s, h_1^{ns} and h_2^{ns}, are modeled by the basis estimated in the third segment. Note, however, that as can be deduced from the different projections, the identified eigenvectors are different in each segment. Hence, without any subspace reduction procedure, six eigenvectors would have been identified, unnecessarily inflating the interference subspace rank. The last column of Table I depicts the basis obtained by the QRD. The reduced subspace, comprised of only three eigenvectors, can model all interference ATFs, as evident from the low level of $e\mathcal{N}$ associated with all ATFs. This reduced basis is in general different from the eigenvectors identified in each of the three segments, but still spans the interference subspace (consisting of the three designated sources).

The novel procedure relaxes the widely-used requirement for non-overlapping activity periods of the distinct interference sources. Moreover, since several segments are collected, the procedure tends to be more robust than methods that rely on PSD estimates obtained by only one segment.

B. Desired Sources RTF Estimation

Consider time frames for which only the stationary sources are active and estimate the corresponding PSD matrix:

$$\hat{\boldsymbol{\Phi}}^{s}_{zz}(\ell,k) \approx \boldsymbol{H}^{s}(\ell,k)\boldsymbol{\Lambda}^{s}(\ell,k)\big(\boldsymbol{H}^{s}(\ell,k)\big)^{\dagger} + \sigma_v^2 \boldsymbol{I}_{M\times M}. \tag{39}$$

Assume that there exists a segment ℓ_i during which the only active non-stationary signal is the ith desired source $i = 1, 2, \ldots, K$. The corresponding PSD matrix will then satisfy:

$$\hat{\boldsymbol{\Phi}}^{d,i}_{zz}(\ell_i,k) \approx (\sigma_i^d(\ell_i,k))^2 \boldsymbol{h}_i^d(\ell_i,k)\big(\boldsymbol{h}_i^d(\ell_i,k)\big)^{\dagger} + \hat{\boldsymbol{\Phi}}^{s}_{zz}(\ell,k). \tag{40}$$

Now, applying the GEVD to $\hat{\boldsymbol{\Phi}}^{d,i}_{zz}(\ell_i,k)$ and the stationary-noise PSD matrix $\hat{\boldsymbol{\Phi}}^{s}_{zz}(\ell,k)$ we have:

$$\hat{\boldsymbol{\Phi}}^{d,i}_{zz}(\ell_i,k)\boldsymbol{f}_i(k) = \lambda_i(k)\hat{\boldsymbol{\Phi}}^{s}_{zz}(\ell,k)\boldsymbol{f}_i(k) \tag{41}$$

The generalized eigenvectors corresponding to the generalized eigenvalues with values other than 1 span the desired sources subspace. Since we assumed that only source i is active in segment ℓ_i, this eigenvector corresponds to a scaled version of the source ATF. To prove this relation for the single eigenvector case, let $\lambda_i(k)$ correspond the largest eigenvalue at segment ℓ_i and $\boldsymbol{f}_i(k)$ its corresponding eigenvector. Substituting $\hat{\boldsymbol{\Phi}}^{d,i}_{zz}(\ell_i,k)$ as defined in (40) in the left-hand side of (41) yields

$$(\sigma_i^d(\ell_i,k))^2 \boldsymbol{h}_i^d(\ell_i,k)\big(\boldsymbol{h}_i^d(\ell_i,k)\big)^{\dagger}\boldsymbol{f}_i(k) + \hat{\boldsymbol{\Phi}}^{s}_{zz}(\ell,k)\boldsymbol{f}_i(k) = \lambda_i(k)\hat{\boldsymbol{\Phi}}^{s}_{zz}(\ell,k)\boldsymbol{f}_i(k)$$

therefore

$$(\sigma_i^d(\ell_i,k))^2 \boldsymbol{h}_i^d(\ell_i,k)\underbrace{\big(\boldsymbol{h}_i^d(\ell_i,k)\big)^{\dagger}\boldsymbol{f}_i(k)}_{\text{scalar}} = \big(\lambda_i(k)-1\big)\hat{\boldsymbol{\Phi}}^{s}_{zz}(\ell,k)\boldsymbol{f}_i(k)$$

and finally,

$$\boldsymbol{h}_i^d(\ell_i,k) = \underbrace{\frac{\lambda_i(k)-1}{(\sigma_i^d(\ell_i,k))^2\big(\boldsymbol{h}_i^d(\ell_i,k)\big)^{\dagger}\boldsymbol{f}_i(k)}}_{\text{scalar}}\hat{\boldsymbol{\Phi}}^{s}_{zz}(\ell,k)\boldsymbol{f}_i(k) \therefore$$

Hence, the desired signal ATF $\boldsymbol{h}_i^d(\ell_i,k)$ is a scaled and rotated version of the eigenvector $\boldsymbol{f}_i(k)$ (with eigenvalue other than 1). As we are interested in the RTFs rather than the entire ATFs the scaling ambiguity can be resolved by the following normalization:

$$\hat{\tilde{\boldsymbol{h}}}_i^d(\ell,k) \triangleq \frac{\boldsymbol{\Phi}^{s}_{zz}(\ell,k)\boldsymbol{f}_i(k)}{\big(\boldsymbol{\Phi}^{s}_{zz}(\ell,k)\boldsymbol{f}_i(k)\big)_1} \tag{42}$$

where $(\cdot)_1$ is the first component of the vector corresponding to the reference microphone (arbitrarily chosen to be the first microphone).

We repeat this estimation procedure for each desired source $i = 1, 2, \ldots, K$. The value of K is a design parameter of the algorithm.

V. Algorithm Summary

The entire algorithm is summarized in Alg. 1. The algorithm is implemented almost entirely in the STFT domain, using a rectangular analysis window of length N_{DFT}, and a shorter rectangular synthesis window, resulting in the *overlap & save* procedure [36], avoiding any cyclic convolution effects. The PSD of the stationary interferences and the desired sources is estimated using the Welch method, with a Hamming window of length $D \times N_{\text{DFT}}$ applied to each segment, and $(D-1) \times N_{\text{DFT}}$ overlap between segments. However, since only lower frequency resolution is required, we wrapped each segment to length N_{DFT} before the application of the Discrete Fourier Transform (DFT) operation. The interference subspace is estimated from a $L_{seg} \times N_{\text{DFT}}$ length segment. The overlap between segments is denoted OVRLP. The resulting beamformer estimate is tapered by a Hamming window resulting in a smooth filter in the coefficient range $[-FL_l, FL_r]$. The parameters used for the simulation are given in Table II.

VI. Experimental Study

A. The Test Scenario

The proposed algorithm was tested both in simulated and real room environments with five directional signals, namely two (male and female) desired speech sources, two (other male and female) speakers as competing speech signals, and a stationary speech-like noise drawn from NOISEX-92 [37] database. We used different set of signals for the simulated and real environments.

In the simulated room scenario the image method [38] was used to generate the RIR using the simulator in [39]. All the signals $i = 1, 2, \ldots, N$ were then convolved with the corresponding time-invariant RIRs. The microphone signals $z_m(\ell, k)$; $m = 1, 2, \ldots, M$ were finally obtained by summing up the contributions of all directional sources with an additional uncorrelated sensor noise. The level of all desired sources is equal. The desired signal to sensor noise ratio was set to 41dB (this ratio determines σ_v^2). The relative power between the desired sources and all interference sources is depicted in the simulation results in Tables III-IV.

In the real room scenario each of the signals was played by a loudspeaker located in a reverberant room (each signal was played by a different loudspeaker) and captured by an array of M microphones. The signals $\mathbf{z}(\ell, k)$ were finally constructed by summing up all recorded microphone signals with a gain related to the desired input Signal to Interference Ratio (SIR).

For evaluating the performance of the proposed algorithm, we applied the algorithms in two phases. In the first phase, the algorithm (consists of the LCMV beamformer and the ANC) was applied to an input signal, comprised of the sum of the desired speakers, the competing speakers, and the stationary noise (with gains in accordance with the respective SIR). In this phase, the algorithm was allowed to adapt yielding $y(\ell, k)$, the actual algorithm output. In the second phase, the beamformer was *not* updated. Instead, a copy of the coefficients, obtained in the first phase, was used as the weights. As the coefficients are time varying (due to the application of the

Algorithm 1 Summary of the proposed LCMV beamformer implemented as a GSC.

```
1) Output signal:
   $y(\ell,k) \triangleq y_{\text{FBF}}(\ell,k) - \mathbf{q}^{\dagger}(\ell,k)\mathbf{u}(\ell,k)$
2) FBF with modified constraints set :
   $y_{\text{FBF}}(\ell,k) \triangleq \tilde{\boldsymbol{w}}_0^{\dagger}(\ell,k)\boldsymbol{z}(\ell,k)$
   where
       $\tilde{\boldsymbol{w}}_0(\ell,k) \triangleq \tilde{\boldsymbol{C}}(\ell,k)\big(\tilde{\boldsymbol{C}}(\ell,k)^{\dagger}\tilde{\boldsymbol{C}}(\ell,k)\big)^{-1}\boldsymbol{g}$
       $\tilde{\boldsymbol{C}}(\ell,k) \triangleq \begin{bmatrix} \tilde{\boldsymbol{H}}^d(\ell,k) & \boldsymbol{E}(\ell,k) \end{bmatrix}$
       $\boldsymbol{g} \triangleq \begin{bmatrix} \underbrace{1 \ldots 1}_{K} & \underbrace{0 \ldots 0}_{N-K} \end{bmatrix}^T.$
   $\tilde{\boldsymbol{H}}^d(\ell,k)$ are the RTFs in respect to microphone #1.
3) Reference signals:
   $\boldsymbol{u}(\ell,k) \triangleq \boldsymbol{B}(\ell,k)\boldsymbol{z}(\ell,k)$
   where
       $\boldsymbol{B}(\ell,k) \triangleq \boldsymbol{I}_{M\times M} - \tilde{\boldsymbol{C}}(\ell,k)\big(\tilde{\boldsymbol{C}}^{\dagger}(\ell,k)\tilde{\boldsymbol{C}}(\ell,k)\big)^{-1}\tilde{\boldsymbol{C}}^{\dagger}(\ell,k).$
4) Update filters:
   $\tilde{\mathbf{q}}(\ell+1,k) = \mathbf{q}(\ell,k) + \mu_q \frac{\mathbf{u}(\ell,k)y^*(\ell,k)}{p_{\text{est}}(\ell,k)}$
   $\mathbf{q}(\ell+1,k) \xleftarrow{\text{FIR}} \tilde{\mathbf{q}}(\ell+1,k)$
   $p_{\text{est}}(\ell,k) = \alpha_p p_{\text{est}}(\ell-1,k) + (1-\alpha_p)\|\mathbf{u}(\ell,k)\|^2$
5) Estimation:
   a) Estimate the stationary noise PSD using Welch method: $\hat{\boldsymbol{\Phi}}^s_{zz}(\ell,k)$
   b) Estimate time-invariant desired sources RTFs $\tilde{\boldsymbol{H}}^d(k) \triangleq \begin{bmatrix} \tilde{\boldsymbol{h}}^d_1(k) \ldots \tilde{\boldsymbol{h}}^d_K(k) \end{bmatrix}$
      Using GEVD and normalization:
         i) $\hat{\boldsymbol{\Phi}}^{d,i}_{zz}(\ell_i,k)\boldsymbol{f}_i(k) = \lambda_i \hat{\boldsymbol{\Phi}}^s_{zz}(\ell,k)\boldsymbol{f}_i(k) \Rightarrow \boldsymbol{f}_i(k)$
         ii) $\hat{\tilde{\boldsymbol{h}}}^d_i(\ell,k) \triangleq \frac{\hat{\boldsymbol{\Phi}}^s_{zz}(\ell,k)\boldsymbol{f}_i(k)}{\big(\hat{\boldsymbol{\Phi}}^s_{zz}(\ell,k)\boldsymbol{f}_i(k)\big)_1}.$
   c) Interferences subspace:
      QRD factorization of eigen-spaces $\begin{bmatrix} \boldsymbol{E}_1(k)\boldsymbol{\Lambda}_1^{\frac{1}{2}}(k) & \ldots & \boldsymbol{E}_{N_{seg}}(k)\boldsymbol{\Lambda}_{N_{seg}}^{\frac{1}{2}}(k) \end{bmatrix}$
      Where $\hat{\boldsymbol{\Phi}}_{zz}(\ell_i,k) = \boldsymbol{E}_i(k)\boldsymbol{\Lambda}_i(k)\boldsymbol{E}_i^{\dagger}(k)$ for time segment $\ell_i$.
```

ANC), we used in each time instant the corresponding copy of the coefficients. The spatial filter was then applied to each of the unmixed sources.

Denote by $y^d_{\text{FBF},i}(\ell,k)$, $y^d_i(\ell,k)$; $i = 1,\ldots,K$, the desired signals components at the beamformer output and the total output (including the ANC), respectively, $y^{ns}_{\text{FBF},i}(\ell,k)$, $y^{ns}_i(\ell,k)$; $i = 1,\ldots,N_{ns}$ the corresponding non-stationary interference components, $y^s_{\text{FBF},i}(\ell,k)$, $y^s_i(\ell,k)$; $i = 1,\ldots,N_s$ the stationary interference components, and $y^v_{\text{FBF}}(\ell,k)$, $y^v(\ell,k)$ the sensor noise com-

TABLE II
PARAMETERS USED BY THE SUBSPACE BEAMFORMER ALGORITHM.

Parameter	Description	Value
General Parameters		
f_s	Sampling frequency	8KHz
determines σ_v^2	Desired signal to sensor noise ratio	41dB
PSD Estimation using Welch Method		
N_{DFT}	DFT length	2048
D	Frequency decimation factor	6
JF	Time offset between segments	2048
Interferences' Subspace Estimation		
L_{seg}	Number of DFT segments used for estimating a single interference subspace	24
OVRLP	The overlap between time segments that are used for interferences subspace estimation	50%
ΔEV$_{\text{TH}}$	Eigenvectors corresponding to eigenvalues that are more than EV$_{\text{TH}}$ lower below the largest eigenvalue are discarded from the signal subspace	40dB
MEV$_{\text{TH}}$	Eigenvectors corresponding to eigenvalues not higher than MEV$_{\text{TH}}$ above the sensor noise are discarded from the signal subspace	5dB
ΔU$_{\text{TH}}$	Vectors of $\boldsymbol{Q}(k)$ corresponding to values of $\boldsymbol{R}(k)$ that are more than U$_{\text{TH}}$ below the largest value on the diagonal of $\boldsymbol{R}(k)$	40dB
MU$_{\text{TH}}$	Vectors of $\boldsymbol{Q}(k)$ corresponding to values of $\boldsymbol{R}(k)$ not higher than MU$_{\text{TH}}$ above the lowest value on the diagonal of $\boldsymbol{R}(k)$	5dB
Filters Lengths		
FL_r	Causal part of the Beamformer (BF) filters	1000 taps
FL_l	Noncausal part of the BF filters	1000 taps
BL_r	Causal part of the BM filters	250 taps
BL_l	Noncausal part of the BM filters	250 taps
RL_r	Causal part of the ANC filters	500 taps
RL_l	Noncausal part of the ANC filters	500 taps
ANC Parameters		
μ_0	Normalized Least Mean Squares (NLMS) adaptation factor	0.18
ρ	Forgetting factor for the estimation of the normalization power $p_{est}(\ell, k)$	0.9

ponent at the beamformer and total output respectively. The entire test procedure is depicted in Fig. 4.

One quality measure used for evaluating the performance of the proposed algorithm is the improvement in the SIR level. Since, generally, there are several desired sources and interference sources we will use all pairs of SIR for quantifying the performance. The SIR of desired signal i relative to the non-stationary signal j as measured on microphone m_0 is defined as follows:

$$\text{SIR}^{ns}_{\text{in},ij}[\text{dB}] = 10\log_{10}\frac{\sum_\ell \sum_{k=0}^{N_{\text{DFT}}-1}\left(s_i^d(\ell,k)h^d_{im_0}(\ell,k)\right)^2}{\sum_\ell \sum_{k=0}^{N_{\text{DFT}}-1}\left(s_j^{ns}(\ell,k)h^{ns}_{jm_0}(\ell,k)\right)^2};\ 1 \le i \le K, 1 \le j \le N_{ns}.$$

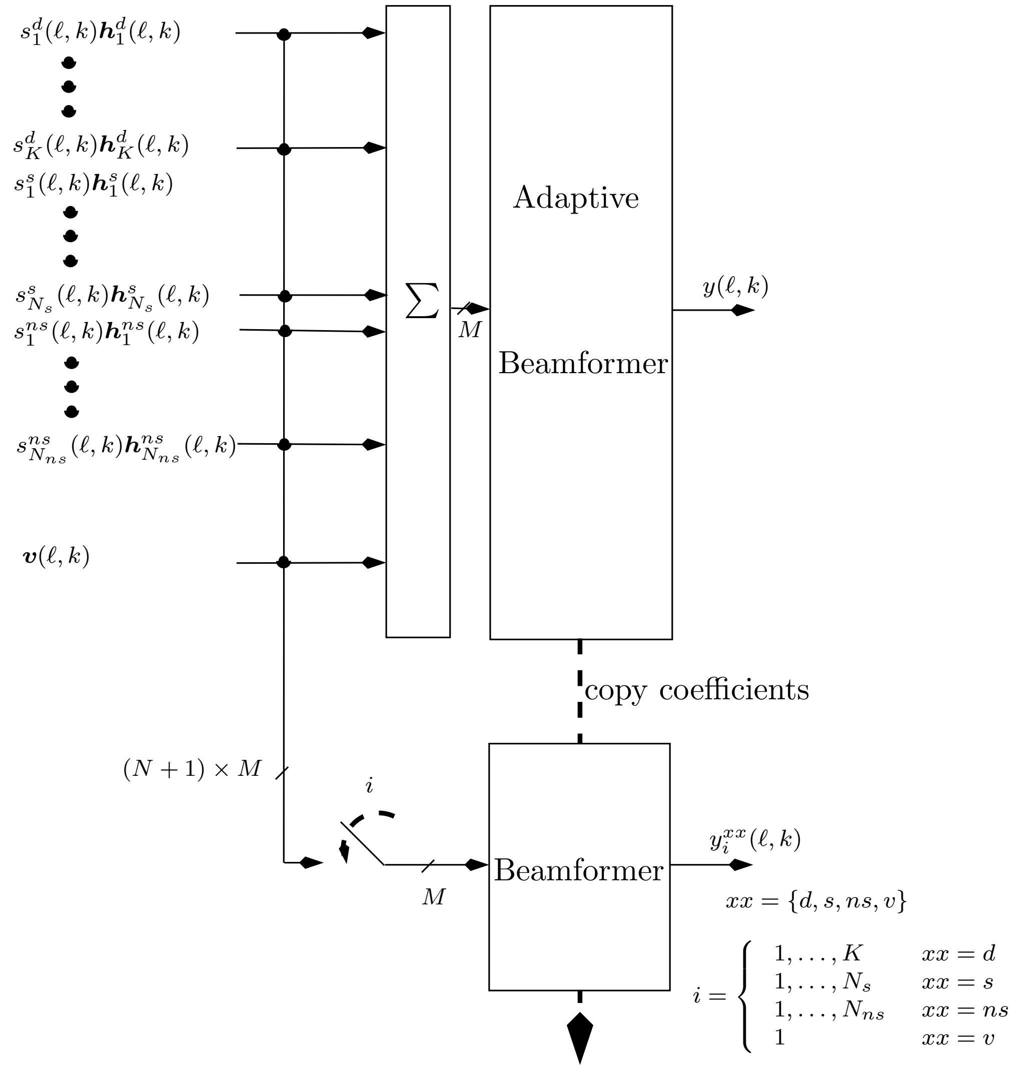

$s_1^d(\ell,k)\boldsymbol{h}_1^d(\ell,k)$
$s_K^d(\ell,k)\boldsymbol{h}_K^d(\ell,k)$
$s_1^s(\ell,k)\boldsymbol{h}_1^s(\ell,k)$
$s_{N_s}^s(\ell,k)\boldsymbol{h}_{N_s}^s(\ell,k)$
$s_1^{ns}(\ell,k)\boldsymbol{h}_1^{ns}(\ell,k)$
$s_{N_{ns}}^{ns}(\ell,k)\boldsymbol{h}_{N_{ns}}^{ns}(\ell,k)$
$\boldsymbol{v}(\ell,k)$
$\sum$
M
Adaptive
Beamformer
$y(\ell,k)$
copy coefficients
$(N+1)\times M$
i
M
Beamformer
$y_i^{xx}(\ell,k)$
$xx=\{d,s,ns,v\}$
$i=\begin{cases}1,\dots,K & xx=d\\ 1,\dots,N_s & xx=s\\ 1,\dots,N_{ns} & xx=ns\\ 1 & xx=v\end{cases}$

Fig. 4. Test procedure for evaluating the performance of the algorithm.

Similarly, the input SIR of the desired signal i relative to the stationary signal j :

$$\mathrm{SIR}_{\mathrm{in},ij}^{s}[\mathrm{dB}] = 10\log_{10}\frac{\sum_\ell\sum_{k=0}^{N_{\mathrm{DFT}}-1}\left(s_i^d(\ell,k)h_{im_0}^d(\ell,k)\right)^2}{\sum_\ell\sum_{k=0}^{N_{\mathrm{DFT}}-1}\left(s_j^s(\ell,k)h_{jm_0}^s(\ell,k)\right)^2};\ 1\le i\le K, 1\le j\le N_s.$$

These quantities are compared with the corresponding FBF and total outputs SIR:

$$\begin{aligned}
\mathrm{SIR}^{ns}_{\mathrm{FBF},ij}[\mathrm{dB}] &= 10\log_{10}\frac{\sum_\ell\sum_{k=0}^{N_{\mathrm{DFT}}-1}\left(y^d_{\mathrm{FBF},i}(\ell,k)\right)^2}{\sum_\ell\sum_{k=0}^{N_{\mathrm{DFT}}-1}\left(y^{ns}_{\mathrm{FBF},j}(\ell,k)\right)^2};\ 1\le i\le K, 1\le j\le N_{ns}\\
\mathrm{SIR}^{s}_{\mathrm{FBF},ij}[\mathrm{dB}] &= 10\log_{10}\frac{\sum_\ell\sum_{k=0}^{N_{\mathrm{DFT}}-1}\left(y^d_{\mathrm{FBF},i}(\ell,k)\right)^2}{\sum_\ell\sum_{k=0}^{N_{\mathrm{DFT}}-1}\left(y^{s}_{\mathrm{FBF},j}(\ell,k)\right)^2};\ 1\le i\le K, 1\le j\le N_{s}\\
\mathrm{SIR}^{ns}_{\mathrm{out},ij}[\mathrm{dB}] &= 10\log_{10}\frac{\sum_\ell\sum_{k=0}^{N_{\mathrm{DFT}}-1}\left(y^d_{i}(\ell,k)\right)^2}{\sum_\ell\sum_{k=0}^{N_{\mathrm{DFT}}-1}\left(y^{ns}_{j}(\ell,k)\right)^2};\ 1\le i\le K, 1\le j\le N_{ns}\\
\mathrm{SIR}^{s}_{\mathrm{out},ij}[\mathrm{dB}] &= 10\log_{10}\frac{\sum_\ell\sum_{k=0}^{N_{\mathrm{DFT}}-1}\left(y^d_{i}(\ell,k)\right)^2}{\sum_\ell\sum_{k=0}^{N_{\mathrm{DFT}}-1}\left(y^{s}_{j}(\ell,k)\right)^2};\ 1\le i\le K, 1\le j\le N_{s}.
\end{aligned}$$

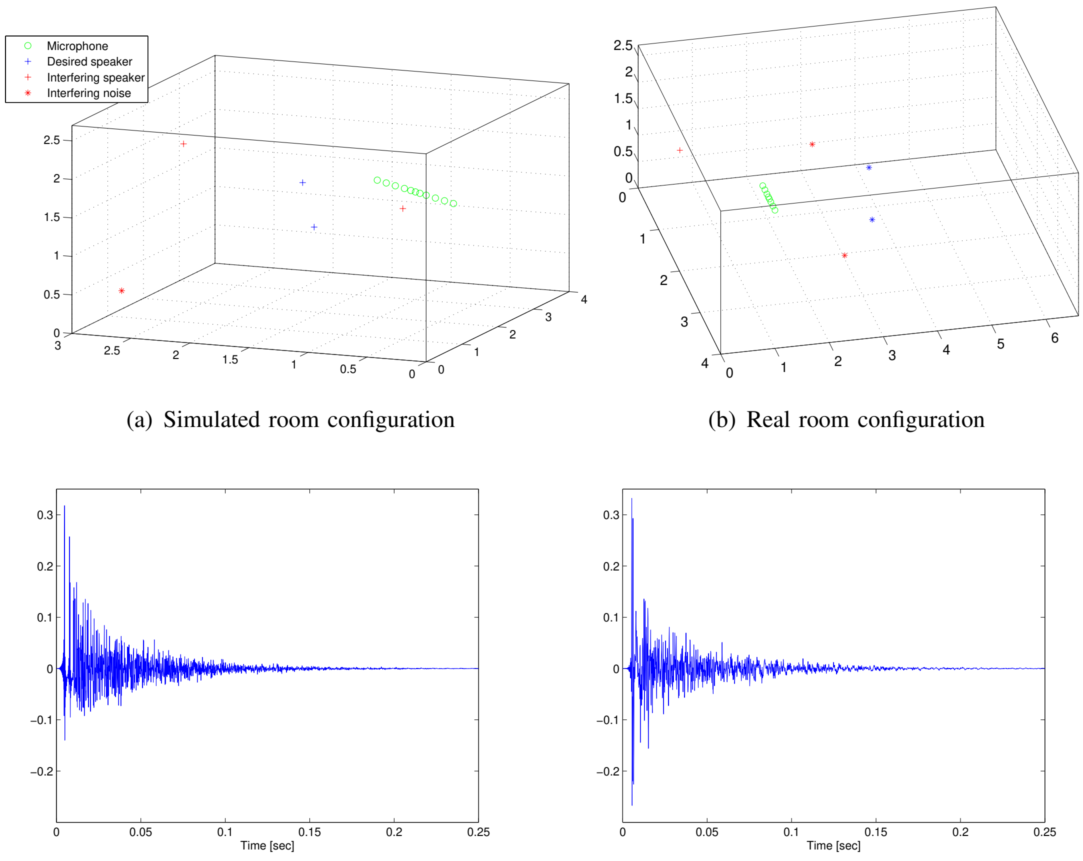
For evaluating the distortion imposed on the desired source we also calculated the Segmental Signal to Noise Ratio (SSNR) and Log Spectral Distortion (LSD) distortion measures relating each desired source component $1 \le i \le K$ at microphone #1, namely $s^d_i(\ell,k)h^d_{i1}$, and its corresponding component at the output, namely $y^d_i(\ell,k)$.

B. *Simulated Environment*

The algorithm was tested in the simulated room environment using reordered speech utterances, made in a quiet room [40]. The RIRs were simulated with a modified version [39] of Allen and Berkley's *image method* [38] with various reverberation levels ranging between 150–300mSec. The simulated environment was a $4m \times 3m \times 2.7m$ room. A nonuniform linear array consisting of 11 microphones with inter-microphone distances ranging from $5cm$ to $10cm$ was used for one set of experiments, and an 8 microphone subset of the same array was used for the second set of experiments. The microphone array and the various sources positions are depicted in Fig. 5(a). A typical RIR relating a source and one of the microphones is depicted in Fig. 5(c). The SIR improvements, as a function of the reverberation time T_{60}, obtained by the FBF branch and by the LCMV beamformer are depicted in Table III for the 8 microphone case and in Table IV for the 11 microphone case. The SSNR and the LSD distortion measures are also depicted for each source. Since the desired sources RTF are estimated when the competing speech signals are inactive, their relative power has no influence on the obtained performance, and is therefore kept fixed during the simulations.

The results in the Tables were obtained using the second phase of the test procedure described in Sec. VI-A. It is shown that on average the beamformer can gain approximately 11dB SIR improvement for the stationary interference in the 8 microphone case (15dB for 11 microphone case), and approximately 13dB SIR improvement for the non-stationary interference in the 8 microphone case (15dB for 11 microphone case).

The SSNR and LSD distortion measures depict that only low distortion is imposed on the desired sources. This result is subjectively verified by the assessment of the sonograms in Fig. 6. It can be easily verified that the interference signals are significantly attenuated while



(a) Simulated room configuration

(b) Real room configuration

(c) A typical simulated RIR

(d) A typical measured RIR

Fig. 5. Room configuration and the corresponding typical RIR for simulated and real scenarios.

TABLE III

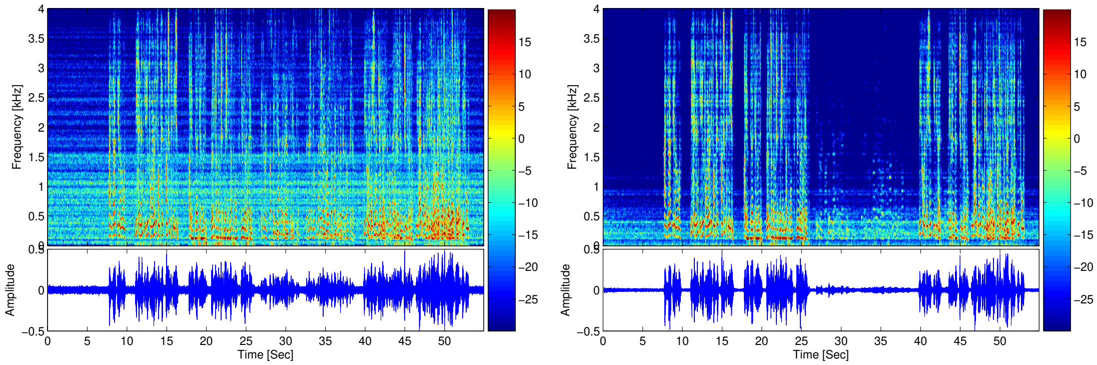
SIR IN dB FOR THE FBF OUTPUT AND THE TOTAL LCMV OUTPUT AND SPEECH DISTORTION MEASURES (SSNR AND LSD IN dB) BETWEEN THE DESIRED SOURCE COMPONENT RECEIVED BY MICROPHONE #1 AND RESPECTIVE COMPONENT AT THE LCMV OUTPUT. 8 MICROPHONE ARRAY, 2 DESIRED SPEAKERS, 2 INTERFERING SPEAKERS AND ONE STATIONARY NOISE WITH VARIOUS REVERBERATION LEVELS.

T_{60}	Source	Input SIR			FBF SIR			Total SIR			SSNR	LSD
		s_1^{ns}	s_2^{ns}	s_1^{s}	s_1^{ns}	s_2^{ns}	s_1^{s}	s_1^{ns}	s_2^{ns}	s_1^{s}		
150ms	s_1^d	6.00	6.00	13.00	18.75	22.35	19.12	18.49	21.72	24.00	9.59	1.14
	s_2^d	6.00	6.00	13.00	18.74	22.34	19.11	18.68	21.88	24.19	10.18	1.65
200ms	s_1^d	6.00	6.00	13.00	18.06	20.62	19.54	18.29	21.30	24.66	7.20	1.54
	s_2^d	6.00	6.00	13.00	18.10	20.66	19.58	18.87	21.88	25.24	8.41	2.01
250ms	s_1^d	6.00	6.00	13.00	18.47	19.79	19.89	18.36	20.88	24.51	7.02	1.81
	s_2^d	6.00	6.00	13.00	18.48	19.80	19.90	19.43	21.95	25.58	7.74	2.37
300ms	s_1^d	6.00	6.00	13.00	17.56	17.63	19.46	18.27	19.3	23.64	6.94	2.18
	s_2^d	6.00	6.00	13.00	17.38	17.45	19.28	18.62	19.65	23.99	7.68	1.82

TABLE IV

SIR IN DB FOR THE FBF OUTPUT AND THE TOTAL LCMV OUTPUT AND SPEECH DISTORTION MEASURES (SSNR AND LSD IN DB) BETWEEN THE DESIRED SOURCE COMPONENT RECEIVED BY MICROPHONE #1 AND RESPECTIVE COMPONENT AT THE LCMV OUTPUT. 11 MICROPHONE ARRAY, 2 DESIRED SPEAKERS, 2 INTERFERING SPEAKERS AND ONE STATIONARY NOISE WITH VARIOUS REVERBERATION LEVELS.

T_{60}	Source	Input SIR			FBF SIR			Total SIR			SSNR	LSD
		s_1^{ns}	s_2^{ns}	s_1^s	s_1^{ns}	s_2^{ns}	s_1^s	s_1^{ns}	s_2^{ns}	s_1^s		
150ms	s_1^d	6.00	6.00	13.00	19.46	20.54	19.66	20.53	22.79	28.07	11.33	1.12
	s_2^d	6.00	6.00	13.00	19.46	22.34	19.11	20.39	22.98	27.93	13.41	1.13
200ms	s_1^d	6.00	6.00	13.00	18.39	18.43	20.02	18.97	20.91	26.20	9.51	1.39
	s_2^d	6.00	6.00	13.00	18.46	19.50	20.09	19.13	21.07	26.36	10.02	1.81
250ms	s_1^d	6.00	6.00	13.00	18.98	19.24	19.29	18.86	20.57	26.07	8.49	1.56
	s_2^d	6.00	6.00	13.00	18.73	18.99	19.04	19.19	20.90	26.40	8.04	1.83
300ms	s_1^d	6.00	6.00	13.00	19.18	17.96	18.93	19.53	19.79	26.21	7.78	1.86
	s_2^d	6.00	6.00	13.00	18.73	17.51	18.48	19.49	19.75	26.17	7.19	1.74



(a) Microphone #1 signal

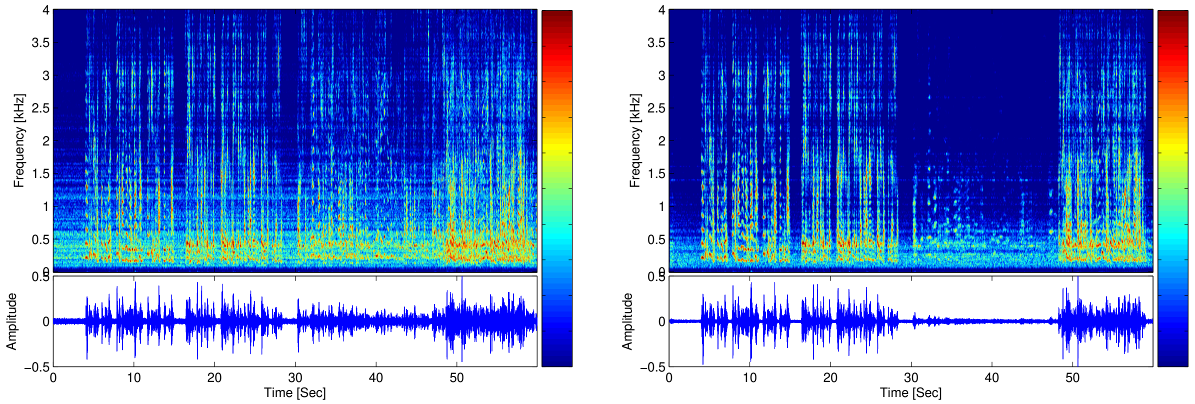
(b) Algorithm's output

Fig. 6. Sonograms and waveforms for the simulated room scenario depicting the algorithm's SIR improvement.

the desired sources remain almost undistorted. Speech samples demonstrating the performance of the proposed algorithm can be downloaded from [40].

C. Real Environment

In the real room environment we used as the directional signals four speakers drawn from the TIMIT [41] database and the speech-like noise described above. The performance was evaluated using real medium-size conference room equipped with furniture, book shelves, a large meeting table, chairs and other standard items. The room dimensions are $6.6m \times 4m \times 2.7m$. A linear nonuniform array consisting of 8 omni-directional microphones (AKG CK32) was used to pick up the sound signals (with the same configuration as in the simulated environment). The various sources were played separately from point loudspeakers (FOSTEX 6301BX). The



(a) Microphone #1 signal

(b) Algorithm's output

Fig. 7. Sonograms and waveforms for the real room scenario depicting the algorithm's SIR improvement.

algorithm's input was constructed by summing up all components contributions and additional, spatially white, computer-generated sensor noise signals. The source-microphone constellation is depicted in Fig. 5(b). The RIR and the respective reverberation time were estimated using the WinMLS2004 software (a product of Morset Sound Development). A typical RIR, having $T_{60} = 250$mSec, is depicted in Fig. 5(d).

In Fig. 7 sonograms of the input signal and the algorithm's output are depicted. The input SIR was $6dB$. A total SIR improvement of $15.28dB$ was obtained for the interfering speakers and $16.23dB$ for the stationary noise. The ANC contributed $1.32dB$ for the competing speakers, and $3.15dB$ for the stationary noise.

VII. Conclusions

We have addressed the problem of extracting several desired sources in a reverberant environment contaminated by both non-stationary (competing speakers) and stationary interferences. The LCMV beamformer (implemented as a GSC structure) was designed to satisfy a set of constraints for the desired and interference sources. A novel and practical method for estimating the interference subspace was presented. The ANC branch is identically zero for perfect estimate of the constraints set. However, for erroneous estimate of the constraint matrix the ANC branch significantly contributes to the interference reduction, while imposing only minor additional distortion on the desired signals.

Unlike common GSC structures, we chose to block all directional signals, including the stationary noise signals, by the BM. By treating the stationary sources as directional signals we obtained deeper nulls [35], which do not suffer from fluctuations caused by the adaptive process. In time varying environments, however, different adaptive forms may be used.

A two phase off-line procedure was applied. First, the test scene (comprising the desired and interference sources) was analyzed using few seconds of data for each source. Then, the BF was applied to the entire data. The proposed estimation methods assume that the RIRs are time-invariant. We note however, that this version of the algorithm can be applied for time-invariant scenarios. Recursive estimation methods for time-varying environments is a topic of ongoing research.

Experimental results for both simulated and real environments have demonstrated that the proposed method can be applied to extracting several desired sources from a combination of multiple sources in a complicated acoustic environment.

REFERENCES

[1] J. Cardoso, "Blind signal separation: Statistical principles," *Proc. of the IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

[2] P. Comon, "Independent component analysis: A new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[3] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, May 2000.

[4] L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlations," *Phys. Rev. Lett.*, vol. 72, no. 23, pp. 3634–3637, Jun. 1994.

[5] J. F. Cardoso, "Eigen-structure of the 4th-order cumulant tensor with application to the blind source separation problem," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2109–2112, May 1989.

[6] S. Amari, A. Chichocki, and H. H. Yang, *Unsupervised Adaptive Filtering*. New York: Wiley, 2000, vol. I, ch. Blind signal separation and extraction: Neural and information theoretic approaches, pp. 63–138.

[7] H. Wu and J. Principe, "A unifying criterion for blind source separation and decorrelation: Simultaneous diagonalization of correlation matrices," *Proc. IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, pp. 496–508, Sep. 1997.

[8] M. Z. Ikram and D. R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 1041–1044, Jun. 2000.

[9] E. Jan and J. Flanagan, "Microphone arrays for speech processing," *Int. Symposium on Signals, Systems, and Electronics (ISSSE)*, pp. 373–376, Oct. 1995.

[10] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[11] S. Gannot and I. Cohen, *Springer Handbook of Speech Processing*. Springer, 2007, ch. Adaptive Beamforming and Postfitering, pp. 199–228.

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

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

[14] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, 2004.

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

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

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

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

[19] B. R. Breed and J. Strauss, "A short proof of the equivalence of LCMV and GSC beamforming," *IEEE Signal Processing Lett.*, vol. 9, no. 6, pp. 168–169, Jun. 2002.

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

[21] S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” *Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

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

[23] Y. Hu and P. Loizou, “A generalized subspace approach for enhancing speech corrupted by colored noise,” *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, Jul. 2003.

[24] S. Gazor, S. Affes, and Y. Grenier, “Robust adaptive beamforming via target tracking,” *IEEE Trans. Signal Processing*, vol. 44, no. 6, pp. 1589–1593, Jun. 1996.

[25] B. Yang, “Projection approximation subspace tracking,” *IEEE Trans. Signal Processing*, vol. 43, no. 1, pp. 95–107, Jan. 1995.

[26] S. Gazor, S. Affes, and Y. Grenier, “Wideband multi-source beamforming with adaptive array location calibration and direction finding,” *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, pp. 1904–1907, May 1995.

[27] S. Doclo and M. Moonen, “Combined frequency-domain dereverberation and noise reduction technique for multi-microphone speech enhancement,” in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Darmstadt, Germany, Sep. 2001, pp. 31–34.

[28] E. Warsitz, A. Krueger, and R. Haeb-Umbach, “Speech enhancement with a new generalized eigenvector blocking matrix for application in generalized sidelobe canceler,” *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 73–76, Apr. 2008.

[29] S. Affes, S. Gazor, and Y. Grenier, “An algorithm for multi-source beamforming and multi-target tracking,” *IEEE Trans. Signal Processing*, vol. 44, no. 6, pp. 1512–1522, Jun. 1996.

[30] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, “Speech enhancement based on the subspace method,” *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 5, pp. 497–507, Sep. 2000.

[31] R. Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE Trans. Antennas Propagate.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

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

[33] G. Reuven, S. Gannot, and I. Cohen, “Dual-source transfer-function generalized sidelobe canceler,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 4, pp. 711–727, May 2008.

[34] Y. Avargel and I. Cohen, “System identification in the short-time Fourier transform domain with crossband filtering,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[35] S. Markovich, S. Gannot, and I. Cohen, “A comparison between alternative beamforming strategies for interference cancelation in noisy and reverberant environment,” in *the 25th convention of the Israeli Chapter of IEEE*, Eilat, Israel, Dec. 2008.

[36] J. J. Shynk, “Frequency-domain and multirate adaptive filtering,” *IEEE Signal Processing Magazine*, vol. 9, no. 1, pp. 14–37, Jan. 1992.

[37] A. Varga and H. J. M. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Communication*, vol. 12, no. 3, pp. 247–251, Jul. 1993.

[38] J. Allen and D. Berkley, “Image method for efficiently simulating small-room acoustics,” *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[39] E. Habets, “Room impulse response (RIR) generator,” `http://home.tiscali.nl/ehabets/rir_generator.html`, Jul. 2006.

[40] S. Gannot, “Audio sample files,” `http://www.biu.ac.il/˜gannot`, Sep. 2008.

[41] J. S. Garofolo, “Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database,” National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, Tech. Rep., 1988, (prototype as of December 1988).

Shmulik Markovich Received the B.Sc. (Cum Laude) and M.Sc. degrees in electrical engineering from the Technion – Israel Institute of Technology, Haifa, Israel, in 2002 and 2008 respectively. His research interests include statistical signal processing and speech enhancement using microphone arrays.

Sharon Gannot (S'92-M'01-SM'06) received his 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, Israel in 1995 and 2000 respectively, all in electrical engineering.

In the year 2001 he held a post-doctoral position at the department of Electrical Engineering (SISTA) at K.U.Leuven, Belgium. From 2002 to 2003 he held a research and teaching position at the Faculty of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel. Currently, he is a Senior Lecturer at the School of Engineering, Bar-Ilan University, Israel.

Dr. Gannot is an Associate Editor of the EURASIP Journal of Applied signal Processing, an Editor of two special issues on Multi-microphone Speech Processing of the same journal, a guest editor of ELSEVIER Speech Communication journal and a reviewer of many IEEE journals and conferences. Dr. Gannot is a member of the Technical and Steering committee of the International Workshop on Acoustic Echo and Noise Control (IWAENC) since 2005 and general co-chair of IWAENC 2010 to be held at Tel-Aviv, Israel. His research interests include parameter estimation, statistical signal processing and speech processing using either single- or multi-microphone arrays.

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 is a Co-Editor of the Multichannel Speech Processing section of the *Springer Handbook of Speech Processing*, and serves as Associate Editor of the IEEE SIGNAL PROCESSING LETTERS.

Dr. Cohen received in 2005 and 2006 the Technion Excellent Lecturer awards. He served as Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, and 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.

LIST OF TABLES

LIST OF FIGURES