Chapter 10
Extraction of Desired Speech Signals in Multiple-Speaker Reverberant Noisy Environments

Shmulik Markovich, Sharon Gannot, and Israel Cohen

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 response (RIR). In this chapter, 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)]. We show that the relative transfer functions (RTFs), which relate the speech sources and the microphones, and a basis for the interference subspace suffice for constructing the beamformer. Additionally, in the case of one desired speech signal, we compare the proposed LCMV beamformer and the minimum variance distortionless response (MVDR) beamformer. These algorithms differ in their treatment of the interference sources. A comprehensive experimental study in both simulated and real environments demonstrates the performance of the proposed beamformer. Particularly, it is shown that the LCMV beamformer outperforms the MVDR beamformer provided that the acoustic environment is time-invariant.

Shmulik Markovich and Sharon Gannot
Bar-Ilan University, Israel, e-mail: shmulik.markovich@gmail.com,gannot@eng.biu.ac.il

Israel Cohen
Technion–Israel Institute of Technology, Israel, e-mail: icohen@ee.technion.ac.il

I. Cohen et al. (Eds.): Speech Processing in Modern Communication, STSP 3, pp. 255–279.
springerlink.com

10.1 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, by exploiting their statistical independence, 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.

We will focus on 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 [2] 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), MVDR, and LCMV. A summary of several design criteria can be found in [3, 4]. Cox et al. [5] introduced an improved adaptive beamformer that maintains a set of linear constraints as well as a quadratic inequality constraint.

In [6] a multichannel Wiener filter (MWF) technique has been proposed that produces an 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 (SDW)-MWF [7]. In an MVDR beamformer [8, 9], 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 generalized sidelobe canceler (GSC) [10]. Several researchers (e.g. Er and Cantoni [11]) 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 [12] 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 [13] and later by Gannot et al. [14]. The latter frequency-domain version, which takes into account the reverberant nature of the enclosure, was nicknamed the transfer function GSC (TF-GSC).

Several beamforming algorithms based on subspace methods were developed. Gazor et al. [15] 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 [16] is derived. An extension to the wide-band case is presented by the same authors [17]. However the demand for a delay-only impulse response is still not relaxed. Affes and Grenier [13] 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 [18] 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. [19] propose to replace the blocking matrix (BM) in [14]. They use a new BM based on the generalized eigenvalue decomposition (GEVD) of the received microphone data, providing an indirect estimation of the ATFs relating the desired speaker and the microphones.

Affes et al. [20] extend the structure presented in [15] 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. [21] address the problem of enhancing multiple speech sources in a non-reverberant environment. The multiple signal classification (MUSIC) method, proposed by Schmidt [22], 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, an 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. [23] 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. [24] 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 TF-GSC (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 chapter, we present a novel beamforming technique, aiming at the extraction of multiple desired speech sources, while attenuating several interfering sources by using an LCMV beamformer (both stationary and non-stationary) in a reverberant environment. 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' power spectral density (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 orthogonal triangular decomposition (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 chapter is as follows. In Section 10.2 the problem of extracting multiple desired sources contaminated by multiple interference in a reverberant environment is introduced. In Section 10.3 the multiple constrained LCMV beamformer is presented. In Section 10.4 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 Section 10.5. In Section 10.6 we present typical test scenarios, discuss some implementation considerations of the algorithm, and show experimental results for both a simulated room and a real conference room scenarios. We address both the problem of extracting multiple desired sources as well as single desired source. In the later case, we compare the performance of the novel beamformer with the TF-GSC. We draw some conclusions and summarize our work in Section 10.7.

10.2 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

$$z_m(n) = \sum_{i=1}^{K} s_i^d(n) * h_{im}^d(n) + \sum_{i=1}^{N_s} s_i^s(n) * h_{im}^s(n) + \sum_{i=1}^{N_{ns}} s_i^{ns}(n) * h_{im}^{ns}(n) + v_m(n), \tag{10.1}$$

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:

$$\begin{aligned} z_m(\ell,k) = & \sum_{i=1}^{K} s_i^d(\ell,k) h_{im}^d(\ell,k) + \qquad (10.2) \\ & \sum_{i=1}^{N_s} s_i^s(\ell,k) h_{im}^s(\ell,k) + \sum_{i=1}^{N_{ns}} s_i^{ns}(\ell,k) h_{im}^{ns}(\ell,k) + v_m(\ell,k), \end{aligned}$$

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 [25] validness.

The received signals in (10.2) can be formulated in a vector notation:

$$\begin{aligned}\mathbf{z}(\ell,k) &= H^d(\ell,k)\mathbf{s}^d(\ell,k) + H^s(\ell,k)\mathbf{s}^s(\ell,k) + H^{ns}(\ell,k)\mathbf{s}^{ns}(\ell,k) + \mathbf{v}(\ell,k)\\ &= H(\ell,k)\mathbf{s}(\ell,k) + \mathbf{v}(\ell,k),\end{aligned} \tag{10.3}$$

where

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

$$\begin{aligned}H^d(\ell,k) &\triangleq \left[\, \mathbf{h}_1^d(\ell,k) \,\ldots\, \mathbf{h}_K^d(\ell,k) \,\right]\\ H^s(\ell,k) &\triangleq \left[\, \mathbf{h}_1^s(\ell,k) \,\ldots\, \mathbf{h}_{N_s}^s(\ell,k) \,\right]\\ H^{ns}(\ell,k) &\triangleq \left[\, \mathbf{h}_1^{ns}(\ell,k) \,\ldots\, \mathbf{h}_{N_{ns}}^{ns}(\ell,k) \,\right]\\ H^i(\ell,k) &\triangleq \left[\, H^s(\ell,k)\ H^{ns}(\ell,k) \,\right]\\ H(\ell,k) &\triangleq \left[\, H^d(\ell,k)\ H^s(\ell,k)\ H^{ns}(\ell,k) \,\right]\end{aligned}$$

$$\begin{aligned}\mathbf{s}^d(\ell,k) &\triangleq \left[\, s_1^d(\ell,k) \,\ldots\, s_K^d(\ell,k) \,\right]^T\\ \mathbf{s}^s(\ell,k) &\triangleq \left[\, s_1^s(\ell,k) \,\ldots\, s_{N_s}^s(\ell,k) \,\right]^T\\ \mathbf{s}^{ns}(\ell,k) &\triangleq \left[\, s_1^{ns}(\ell,k) \,\ldots\, s_{N_{ns}}^{ns}(\ell,k) \,\right]^T\\ \mathbf{s}(\ell,k) &\triangleq \left[\, (\mathbf{s}^d(\ell,k))^T\ (\mathbf{s}^s(\ell,k))^T\ (\mathbf{s}^{ns}(\ell,k))^T \,\right]^T.\end{aligned}$$

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

$$\begin{aligned}\Phi_{zz}(\ell,k) &= H^d(\ell,k)\Lambda^d(\ell,k)\big(H^d(\ell,k)\big)^\dagger +\\ & H^s(\ell,k)\Lambda^s(\ell,k)\big(H^s(\ell,k)\big)^\dagger + H^{ns}(\ell,k)\Lambda^{ns}(\ell,k)\big(H^{ns}(\ell,k)\big)^\dagger + \Phi_{vv}(\ell,k),\end{aligned} \tag{10.4}$$

where

$$\begin{aligned}\Lambda^d(\ell,k) &\triangleq \operatorname{diag}\left(\left[\, (\sigma_1^d(\ell,k))^2 \,\ldots\, (\sigma_K^d(\ell,k))^2 \,\right]\right),\\ \Lambda^s(\ell,k) &\triangleq \operatorname{diag}\left(\left[\, (\sigma_1^s(\ell,k))^2 \,\ldots\, (\sigma_{N_s}^s(\ell,k))^2 \,\right]\right),\\ \Lambda^{ns}(\ell,k) &\triangleq \operatorname{diag}\left(\left[\, (\sigma_1^{ns}(\ell,k))^2 \,\ldots\, (\sigma_{N_{ns}}^{ns}(\ell,k))^2 \,\right]\right).\end{aligned}$$

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

In the special case of a single desired source, i.e. $K = 1$, the following definition applies: $H(\ell,k) \triangleq \left[\mathbf{h}_1^d(\ell,k)\ H^s(\ell,k)\ H^{ns}(\ell,k)\right]$ and $\mathbf{s}(\ell,k) \triangleq \left[s_1^d(\ell,k)\ (\mathbf{s}^s(\ell,k))^T\ (\mathbf{s}^{ns}(\ell,k))^T\right]^T$.

10.3 Proposed Method

In this section the proposed algorithm is derived. In the following subsections we adopt the LCMV structure and 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. In the case of single desired source and multiple interference signals, the MVDR strategy can be adopted instead of the derived LCMV strategy. Hence, in this case, both beamformers are presented.

10.3.1 The LCMV and MVDR Beamformers

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

$$y(\ell,k) = \mathbf{w}^\dagger(\ell,k)\mathbf{z}(\ell,k), \tag{10.5}$$

where

$$\mathbf{w}(\ell,k) = \left[w_1(\ell,k),\ldots,w_M(\ell,k)\right]^T. \tag{10.6}$$

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

$$\mathbf{w}(\ell,k) = \underset{\mathbf{w}}{\mathrm{argmin}}\{\mathbf{w}^\dagger(\ell,k)\Phi_{zz}(\ell,k)\mathbf{w}(\ell,k)\}$$
$$\text{subject to } C^\dagger(\ell,k)\mathbf{w}(\ell,k) = \mathbf{g}(\ell,k), \tag{10.7}$$

where

$$C^\dagger(\ell,k)\mathbf{w}(\ell,k) = \mathbf{g}(\ell,k) \tag{10.8}$$

is the constraints set. The well-known solution to (10.7) is given by [3]:

$$\mathbf{w}(\ell,k) = \Phi_{zz}^{-1}(\ell,k)C(\ell,k)\left(C^{\dagger}(\ell,k)\Phi_{zz}^{-1}(\ell,k)C(\ell,k)\right)^{-1}\mathbf{g}(\ell,k). \tag{10.9}$$

Projecting (10.9) to the column space of the constraints matrix yields a beamformer which satisfies the constraint set but not necessarily minimizes the noise variance at the output. This beamformer is given by [3]

$$\mathbf{w}_0(\ell,k) = C(\ell,k)\left(C^{\dagger}(\ell,k)C(\ell,k)\right)^{-1}\mathbf{g}(\ell,k). \tag{10.10}$$

It is shown in [26] that in the case of spatially-white sensor noise, i.e. $\Phi_{vv}(\ell,k) = \sigma_v^2 I_{M\times M}$, and when the constraint set is accurately known, both beamformers defined by (10.9) and (10.10) are equivalent.

Two paradigms can be adopted in the design of a beamformer which is aimed at enhancing a single desired signal contaminated by both noise and interference. These paradigms differ in their treatment of the interference (competing speech and/or directional noise), which is manifested by the definition of the constraints set, namely $C(\ell,k)$ and $\mathbf{g}(\ell,k)$.

The straightforward alternative is to apply a single constraint beamformer, usually referred to as MVDR beamformer, which was efficiently implemented by the TF-GSC [14], for the reverberant case. Another alternative suggests defining constraints for both the desired and the interference sources. Two recent contributions [24] and [26] adopt this alternative. It is shown in [27] that in static scenarios, well-designed nulls towards all interfering signals (as proposed by the LCMV structure) result in an improved undesired signal cancelation compared with the MVDR structure [14]. Naturally, while considering time-varying environments this advantage cannot be guaranteed.

10.3.2 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 [24], 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

$$\left(\mathbf{h}_i^d(\ell,k)\right)^{\dagger}\mathbf{w}(\ell,k) = 1,\ i = 1,\ldots,K. \tag{10.11}$$

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

$$\left(\mathbf{h}_i^s(\ell,k)\right)^{\dagger}\mathbf{w}(\ell,k) = 0, \tag{10.12}$$

and

$$\left(\mathbf{h}_j^{ns}(\ell,k)\right)^{\dagger}\mathbf{w}(\ell,k) = 0. \tag{10.13}$$

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 $H(\ell, k)$ is linearly independent (i.e. the ATFs are independent), it is obvious that for the solution in (10.10) 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

$$C(\ell, k) \triangleq H(\ell, k), \tag{10.14}$$

and a desired response vector

$$\mathbf{g} \triangleq \Big[\underbrace{1 \ \dots \ 1}_{K} \ \underbrace{0 \ \dots \ 0}_{N-K}\Big]^T . \tag{10.15}$$

Evaluating the beamformer (10.10) output for the input (10.3) and constraints set (10.8) gives:

$$\begin{aligned} y(\ell, k) &= \mathbf{w}_0^\dagger(\ell, k)\mathbf{z}(\ell, k) = \\ &\sum_{i=1}^{K} s_i^d(\ell, k) + \mathbf{g}^\dagger \left(H^\dagger(\ell, k) H(\ell, k)\right)^{-1} H^\dagger(\ell, k)\mathbf{v}(\ell, k). \end{aligned} \tag{10.16}$$

The output 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.

For the single desired sources scenario we get:

$$y(\ell, k) = s_1^d(\ell, k) + \mathbf{g}^\dagger \left(H^\dagger(\ell, k) H(\ell, k)\right)^{-1} H^\dagger(\ell, k)\mathbf{v}(\ell, k). \tag{10.17}$$

10.3.3 Equivalent Constraints Set

The matrix $C(\ell, k)$ in (10.14) comprises the ATFs relating the sources and the microphones $\mathbf{h}_i^d(\ell, k)$, $\mathbf{h}_i^s(\ell, k)$ and $\mathbf{h}_i^{ns}(\ell, k)$. Hence, the solution given in (10.10) 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. [23]). 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

$$N_i \triangleq N_s + N_{ns} \tag{10.18}$$

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 $E \triangleq [\mathbf{e}_1 \ \ldots \ \mathbf{e}_{N_i}]$ to be any basis[1] that spans the column space of the interfering sources $H^i(\ell,k) = [H^s(\ell,k) \ H^{ns}(\ell,k)]$. Hence, the following identity holds:

$$H^i(\ell,k) = E(\ell,k)\Theta(\ell,k), \tag{10.19}$$

where $\Theta_{N_i \times N_i}(\ell,k)$ 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, $\Theta_{N_i \times N_i}(\ell,k)$ is an invertible matrix.

Define

$$\tilde{\Theta}(\ell,k) \triangleq \begin{bmatrix} I_{K\times K} & \mathcal{O}_{K\times N_i} \\ \mathcal{O}_{N_i\times K} & \Theta(\ell,k) \end{bmatrix}_{N\times N}, \tag{10.20}$$

where $I_{K\times K}$ is a $K \times K$ identity matrix. Multiplication by $(\tilde{\Theta}^\dagger(\ell,k))^{-1}$ of both sides of the original constraints set in (10.8), with the definitions (10.14)–(10.15) and using the equality $\tilde{\Theta}^\dagger(\ell,k))^{-1}\mathbf{g} = \mathbf{g}$, yields an equivalent constraint set:

$$\dot{C}^\dagger(\ell,k)\mathbf{w}(\ell,k) = \mathbf{g}, \tag{10.21}$$

where the equivalent constraint matrix is

$$\dot{C}(\ell,k) = (\tilde{\Theta}^\dagger(\ell,k))^{-1}C^\dagger(\ell,k) = \left[H^d(\ell,k) \ E(\ell,k)\right]. \tag{10.22}$$

10.3.4 Modified Constraints Set

Both the original and equivalent constraints sets in (10.14) and (10.22) respectively, require estimates of the desired sources ATFs $H^d(\ell,k)$. 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 $\mathbf{g}$ in (10.15)] 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:

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

where microphone #1 was arbitrarily chosen as the reference microphone. The modified beamformer satisfying the modified response $\dot{C}^\dagger(\ell,k)\tilde{\mathbf{w}}(\ell,k) =$

[1] 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.

$\tilde{\mathbf{g}}(\ell,k)$ is then given by

$$\tilde{\mathbf{w}}_0(\ell,k) \triangleq \dot{C}(\ell,k)\big(\dot{C}^\dagger(\ell,k)\dot{C}(\ell,k)\big)^{-1}\tilde{\mathbf{g}}(\ell,k). \tag{10.23}$$

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

$$\begin{aligned} y(\ell,k) =&\tilde{\mathbf{w}}_0^\dagger(\ell,k)\mathbf{z}(\ell,k) = \\ &\sum_{i=1}^{K} h_{i1}^d(\ell,k)s_i^d(\ell,k) + \tilde{\mathbf{g}}^\dagger(\ell,k)\big(\dot{C}^\dagger(\ell,k)\dot{C}(\ell,k)\big)^{-1}\dot{C}^\dagger(\ell,k)\mathbf{v}(\ell,k), \end{aligned} \tag{10.24}$$

as expected from the modified constraint response.

For the single desired sources scenario the modified constraints set yields the following output:

$$y(\ell,k) = h_{i1}^d(\ell,k)s_1^d(\ell,k) + \tilde{\mathbf{g}}^\dagger(\ell,k)\big(\dot{C}^\dagger(\ell,k)\dot{C}(\ell,k)\big)^{-1}\dot{C}^\dagger(\ell,k)\mathbf{v}(\ell,k). \tag{10.25}$$

As mentioned before, estimating the desired signal ATFs is a cumbersome task. Nevertheless, in Section 10.4 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 (10.15) by

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

where

$$\Psi(\ell,k) = \operatorname{diag}\left(\left[\, h_{11}^d(\ell,k) \;\ldots\; h_{K1}^d(\ell,k)\,\right]\right),$$

and

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

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

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

Hence,

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

Define

$$\tilde{C}(\ell,k) \triangleq \dot{C}(\ell,k)\tilde{Psi}^{-1}(\ell,k) = \left[\, \tilde{H}^d(\ell,k) \; E(\ell,k)\,\right], \tag{10.26}$$

where

$$\tilde{H}^d(\ell,k) \triangleq \left[\tilde{\mathbf{h}}_1^d(\ell,k) \ldots \tilde{\mathbf{h}}_K^d(\ell,k)\right], \tag{10.27}$$

with

$$\tilde{\mathbf{h}}_i^d(\ell,k) \triangleq \frac{\mathbf{h}_i^d(\ell,k)}{h_{i1}^d(\ell,k)} \tag{10.28}$$

defined as the RTF with respect to microphone #1.

Finally, the modified beamformer is given by

$$\tilde{\mathbf{w}}_0(\ell,k) \triangleq \tilde{C}(\ell,k)\big(\tilde{C}(\ell,k)^\dagger \tilde{C}(\ell,k)\big)^{-1}\mathbf{g} \tag{10.29}$$

and its corresponding output is indeed given by

$$\begin{aligned} y(\ell,k) =&\tilde{\mathbf{w}}_0^\dagger(\ell,k)\mathbf{z}(\ell,k) = \\ &\sum_{i=1}^{K} s_i^d(\ell,k)h_{i1}^d(\ell,k) + \mathbf{g}^\dagger\left(\tilde{C}^\dagger(\ell,k)\tilde{C}(\ell,k)\right)^{-1}\tilde{C}^\dagger(\ell,k)\mathbf{v}(\ell,k). \end{aligned} \tag{10.30}$$

Therefore, the modified beamformer output comprises the sum of the desired sources as measured at the reference microphone (arbitrarily chosen as microphone #1) and the sensor noise contribution.

For the single desired sources scenario the modified beamformer output is reduced to

$$y(\ell,k) = s_1^d(\ell,k)h_{11}^d(\ell,k) + \mathbf{g}^\dagger\left(H^\dagger(\ell,k)H(\ell,k)\right)^{-1}H^\dagger(\ell,k)\mathbf{v}(\ell,k). \tag{10.31}$$

10.4 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 respect, as a multi-source, colored-noise, extension of the single source subspace estimation method proposed by Affes and Grenier [13]. We further assume that the various filters are slowly time-varying filters, i.e $H(\ell, k) \approx H(k)$. Due to inevitable estimation errors, the constraints set is not exactly satisfied, resulting in leakage of residual interference signals to the beamformer output, as well as desired signal distortion. This leakage reflects on the spatially white sensors noise assumption, and is dealt with in [26].

10.4.1 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 $\hat{\Phi}_{zz}(\ell_i, k)$ be a PSD estimate at the interference-only frame ℓ_i. Using the EVD we have $\hat{\Phi}_{zz}(\ell_i, k) = E_i(k)\Lambda_i(k)E_i^\dagger(k)$. 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 $\Delta\mathrm{EV}_{\mathrm{TH}}(k)$ and $\mathrm{MEV}_{\mathrm{TH}}$. All eigenvectors corresponding to eigenvalues that are more than $\Delta\mathrm{EV}_{\mathrm{TH}}$ below the largest eigenvalue or not higher than $\mathrm{MEV}_{\mathrm{TH}}$ 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. 10.1 for the 11 microphone test scenario presented in Section 10.6. 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 $\mathrm{MEV}_{\mathrm{TH}}$ threshold and the dark green frequency-dependent line depicts the threshold $\mathrm{EV}_{\mathrm{TH}}(k)$. All eigenvalues that do not meet the thresholds, depicted as gray lines in the figure, are discarded from the interference signal subspace. 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'

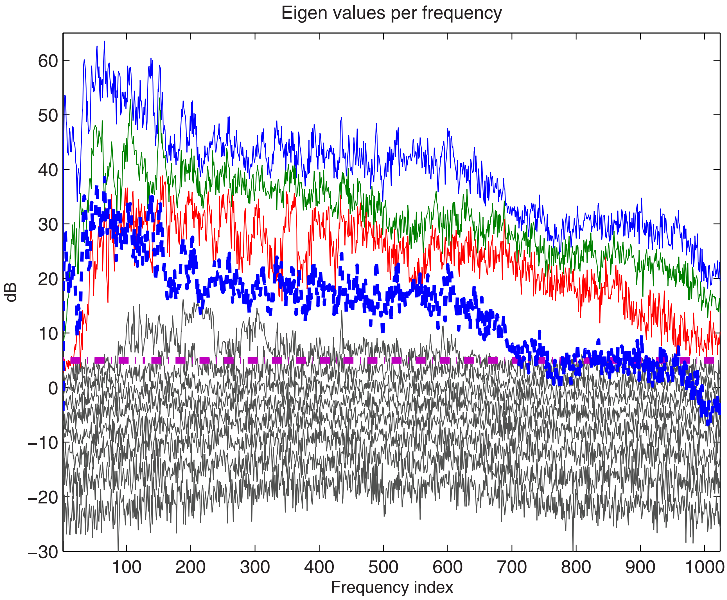


Fig. 10.1 Eigenvalues of an interference-only segment as a function of the frequency bin (solid thin lines). Eigenvalues that do not meet the thresholds $\mathrm{MEV}_{\mathrm{TH}}$ (thick black horizontal line) and $\mathrm{EV}_{\mathrm{TH}}(k)$ (thick black curve) are depicted in grey and discarded from the interference signal subspace.

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 $\hat{E}_i(k)$, and their corresponding eigenvalues as $\hat{\Lambda}_i(k)$. This procedure is repeated for each segment ℓ_i; $i = 1, 2, \ldots, N_{seg}$. These vectors should span the basis of the entire interference subspace:

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

defined in (10.19). 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:

$$\hat{E}(k) \triangleq \bigcup_{i=1}^{N_{seg}} \hat{E}_i(k), \tag{10.32}$$

where $\hat{E}(k)$ is an estimate for the interference subspace basis $E(\ell, k)$ 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 cancellation 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:

$$\left[\hat{E}_1(k)\hat{\Lambda}_1^{\frac{1}{2}}(k) \ldots \hat{E}_{N_{seg}}(k)\hat{\Lambda}_{N_{seg}}^{\frac{1}{2}}(k)\right] P(k) = Q(k)R(k), \tag{10.33}$$

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

All vectors in $Q(k)$ that correspond to values on the diagonal of $R(k)$ 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 $\hat{E}(k)$, the estimate of the interference subspace basis. 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.

10.4.2 Desired Sources RTF Estimation

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

$$\hat{\Phi}_{zz}^{s}(\ell, k) \approx H^s(\ell, k)\Lambda^s(\ell, k)\big(H^s(\ell, k)\big)^{\dagger} + \sigma_v^2 I_{M\times M}. \tag{10.34}$$

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

$$\hat{\Phi}_{zz}^{d,i}(\ell_i, k) \approx (\sigma_i^d(\ell_i, k))^2 \mathbf{h}_i^d(\ell_i, k)\big(\mathbf{h}_i^d(\ell_i, k)\big)^{\dagger} + \hat{\Phi}_{zz}^{s}(\ell, k). \tag{10.35}$$

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

$$\hat{\Phi}_{zz}^{d,i}(\ell_i, k)\mathbf{f}_i(k) = \lambda_i(k)\hat{\Phi}_{zz}^{s}(\ell, k)\mathbf{f}_i(k). \tag{10.36}$$

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 $\mathbf{f}_i(k)$ its corresponding eigenvector. Substituting $\hat{\Phi}^{d,i}_{zz}(\ell_i, k)$ as defined in (10.35) in the left-hand side of (10.36) yields

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

therefore

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

and finally,

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

Hence, the desired signal ATF $\mathbf{h}_i^d(\ell_i, k)$ is a scaled and rotated version of the eigenvector $\mathbf{f}_i(k)$ (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:

$$\hat{\tilde{\mathbf{h}}}_i^d(\ell, k) \triangleq \frac{\Phi^s_{zz}(\ell, k)\mathbf{f}_i(k)}{\big(\Phi^s_{zz}(\ell, k)\mathbf{f}_i(k)\big)_1}, \tag{10.37}$$

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. An alternative method for estimating the RTFs based on the non-stationarity of the speech is developed for single source scenario in [14], but can be used as well for the general scenario with multiple desired sources, provided that time frames for each the desired sources are not simultaneously active exist.

10.5 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

Algorithm 1 Summary of the proposed LCMV beamformer.

```
1) beamformer with modified constraints set:
   $y(\ell,k) \triangleq \tilde{\mathbf{w}}_0^\dagger(\ell,k)\mathbf{z}(\ell,k)$
   where
      $\tilde{\mathbf{w}}_0(\ell,k) \triangleq \tilde{C}(\ell,k)\big(\tilde{C}(\ell,k)^\dagger\tilde{C}(\ell,k)\big)^{-1}\mathbf{g}$
      $\tilde{C}(\ell,k) \triangleq \big[\,\tilde{H}^d(\ell,k)\ E(\ell,k)\,\big]$
      $\mathbf{g} \triangleq \Bigg[\underbrace{1\ \ldots\ 1}_{K}\,\underbrace{0\ \ldots\ 0}_{N-K}\Bigg]^T.$
   $\tilde{\mathbf{H}}^d(\ell,k)$ are the RTFs in respect to microphone #1.
2) Estimation:
   a) Estimate the stationary noise PSD using Welch method: $\hat{\Phi}^s_{zz}(\ell,k)$
   b) Estimate time-invariant desired sources RTFs $\tilde{H}^d(k) \triangleq \big[\,\tilde{\mathbf{h}}_1^d(k)\ldots\tilde{\mathbf{h}}_K^d(k)\,\big]$
      Using GEVD and normalization:
         i) $\hat{\Phi}^{d,i}_{zz}(\ell_i,k)\mathbf{f}_i(k) = \lambda_i\hat{\Phi}^s_{zz}(\ell,k)\mathbf{f}_i(k) \Rightarrow \mathbf{f}_i(k)$
         ii) $\hat{\tilde{\mathbf{h}}}_i^d(\ell,k) \triangleq \frac{\hat{\Phi}^s_{zz}(\ell,k)\mathbf{f}_i(k)}{\big(\hat{\Phi}^s_{zz}(\ell,k)\mathbf{f}_i(k)\big)_1}.$
   c) Interferences subspace:
      QRD factorization of eigen-spaces $\Big[\,E_1(k)\Lambda_1^{\frac{1}{2}}(k)\ \ldots\ E_{N_{seg}}(k)\Lambda_{N_{seg}}^{\frac{1}{2}}(k)\,\Big]$
      Where $\hat{\Phi}_{zz}(\ell_i,k) = E_i(k)\Lambda_i(k)E_i^\dagger(k)$ for time segment $\ell_i$.
```

overlap & save procedure [28], avoiding any cyclic convolution effects. The PSD of the stationary interferences and the desired sources are estimated using the Welch method, with a Hamming window of length $D \times N_{\mathrm{DFT}}$ applied to each segment, and $(D-1) \times N_{\mathrm{DFT}}$ 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 operation. The interference subspace is estimated from a $L_{seg} \times N_{\mathrm{DFT}}$ 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 10.1. In cases where the sensor noise is not spatially-white or when the estimation of the constraint matrix is not accurate, the entire LCMV procedure (10.9) should be implemented. In these cases, the presented algorithm will be accompanied by an adaptive noise canceler (ANC) branch constituting a GSC structure, as presented in [26].

10.6 Experimental Study

In this section, we evaluate the performance of the proposed subspace beamformer. In case of one desired source we compare the presented algorithm with the TF-GSC algorithm [14].

Table 10.1 Parameters used by the subspace beamformer algorithm.

Parameter	Description	Value
General Parameters		
f_s	Sampling frequency	8KHz
	Desired signal to sensor noise ratio (determines σ_v^2)	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%
$\Delta\mathrm{EV}_{\mathrm{TH}}$	Eigenvectors corresponding to eigenvalues that are more than $\mathrm{EV}_{\mathrm{TH}}$ lower below the largest eigenvalue are discarded from the signal subspace	40dB
$\mathrm{MEV}_{\mathrm{TH}}$	Eigenvectors corresponding to eigenvalues not higher than $\mathrm{MEV}_{\mathrm{TH}}$ above the sensor noise are discarded from the signal subspace	5dB
$\Delta\mathrm{U}_{\mathrm{TH}}$	vectors of $\mathbf{Q}(k)$ corresponding to values of $\mathbf{R}(k)$ that are more than U_{TH} below the largest value on the diagonal of $\mathbf{R}(k)$	40dB
$\mathrm{MU}_{\mathrm{TH}}$	vectors of $\mathbf{Q}(k)$ corresponding to values of $\mathbf{R}(k)$ not higher than $\mathrm{MU}_{\mathrm{TH}}$ above the lowest value on the diagonal of $\mathbf{R}(k)$	5dB
Filters Lengths		
FL_r	Causal part of the beamformer filters	1000 taps
FL_l	Noncausal part of the beamformer filters	1000 taps

10.6.1 The Test Scenario

The proposed algorithm was tested both in simulated and real room environments in several test scenarios. In test scenario #1 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 [29] database were mixed.

In test scenarios #2-#4 the performance of the multi-constraints algorithm was compared to the TF-GSC algorithm [14] in a simulated room environment, using one desired speech source, one stationary speech-like noise drawn from NOISEX-92 [29] database, and various number of competing speakers (ranging from zero to two). For the simulated room scenario the image method [30] was used to generate the RIR. The implementation is described in [31]. 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 are depicted in Table 10.2 and Table 10.3 for scenario #1 and scenarios #2-#4, respectively.

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. During the first phase, the algorithm 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 performed the various estimations yielding $y(\ell, k)$, the actual algorithm output. In the second phase, the beamformer was *not* recalculated. Instead, the beamformer obtained in the first phase was applied to each of the unmixed sources.

Denote by $y_i^d(\ell, k)$; $i = 1, \ldots, K$, the desired signals components at the beamformer output, $y_i^{ns}(\ell, k)$; $i = 1, \ldots, N_{ns}$ the corresponding non-stationary interference components, $y_i^s(\ell, k)$; $i = 1, \ldots, N_s$ the stationary interference components, and $y^v(\ell, k)$ the sensor noise component at the beamformer output respectively.

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:

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

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

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

These quantities are compared with the corresponding beamformer outputs SIR:

$$\mathrm{SIR}^{ns}_{\mathrm{out},ij}[\mathrm{dB}] = 10\log_{10}\frac{\sum_\ell \sum_{k=0}^{N_{\mathrm{DFT}}-1}\left(y_i^d(\ell,k)\right)^2}{\sum_\ell \sum_{k=0}^{N_{\mathrm{DFT}}-1}\left(y_j^{ns}(\ell,k)\right)^2}$$

$$1 \le i \le K, 1 \le j \le N_{ns},$$

$$\mathrm{SIR}^{s}_{\mathrm{out},ij}[\mathrm{dB}] = 10\log_{10}\frac{\sum_\ell \sum_{k=0}^{N_{\mathrm{DFT}}-1}\left(y_i^d(\ell,k)\right)^2}{\sum_\ell \sum_{k=0}^{N_{\mathrm{DFT}}-1}\left(y_j^{s}(\ell,k)\right)^2}$$

$$1 \le i \le K, 1 \le j \le N_s.$$

For evaluating the distortion imposed on the desired sources we also calculated the squared error distortion (SED) and log spectral distance (LSD) distortion measures relating each desired source component $1 \le i \le K$ at the output, namely $y_i^d(\ell,k)$ and its corresponding component received by microphone #1, namely $s_i^d(\ell,k)h_{i1}^d$. Define the SED and the LSD distortion for each desired source $1 \le i \le K$:

$$\mathrm{SED}_{out,i}[\mathrm{dB}] = \tag{10.38}$$

$$10\log_{10}\frac{\sum_\ell \sum_{k=0}^{N_{\mathrm{DFT}}-1}\left(s_i^d(\ell,k)h_{i1}^d(\ell,k)\right)^2}{\sum_\ell \sum_{k=0}^{N_{\mathrm{DFT}}-1}\left(s_i^d(\ell,k)h_{i1}^d(\ell,k) - y_i^d(\ell,k)\right)^2},$$

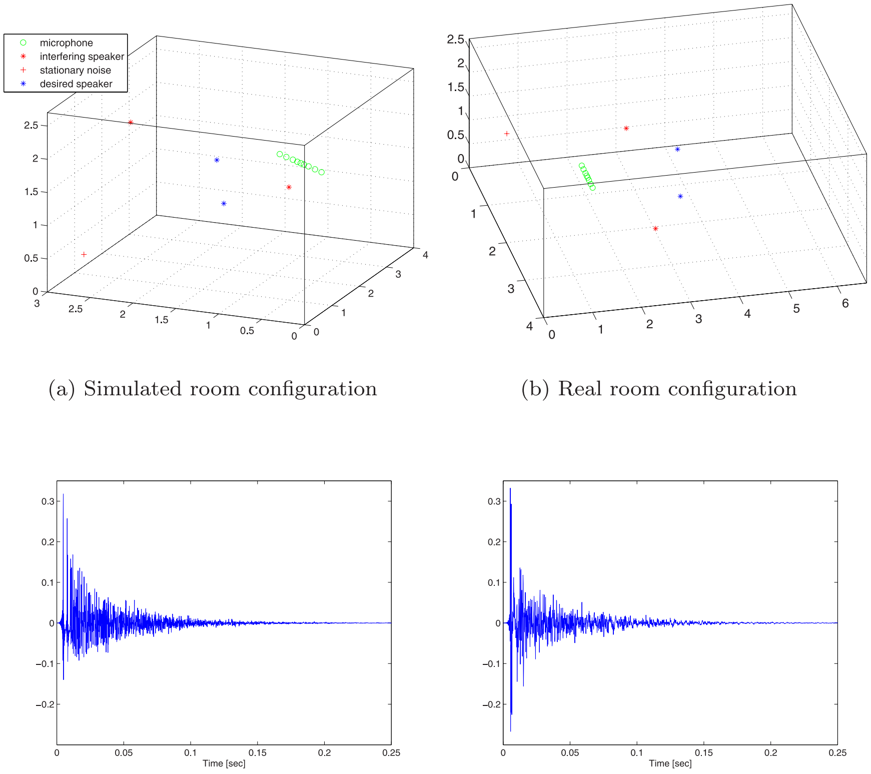
$$\mathrm{LSD}_{out,i} = \tag{10.39}$$

$$\frac{1}{L'}\sum_\ell \sqrt{\frac{1}{N_{\mathrm{DFT}}}\sum_{k=0}^{N_{\mathrm{DFT}}-1}\left[20\log_{10}|s_i^d(\ell,k)h_{i1}^d(\ell,k)| - 20\log_{10}|y_i^d(\ell,k)|\right]^2},$$

where L' is the number of speech active frames and $\{\ell \in \text{Speech Active}\}$. These figures-of-merit are also depicted in the Tables.

10.6.2 Simulated Environment

The RIRs were simulated with a modified version [31] of Allen and Berkley's *image method* [30] 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 $10cm$. The microphone array and the various sources positions are depicted in Fig. 2(a). A typical RIR relating a source and one of the microphones is depicted in Fig. 2(c). The SIR improvements, as a function of the reverberation time T_{60}, obtained by the LCMV beamformer for scenario 1 are depicted in Table 10.2. The SED and the LSD distortion measures are also depicted for each source. Since the desired sources RTFs are estimated when the competing speech signals are inactive, their relative



(a) Simulated room configuration

(b) Real room configuration

(c) A typical simulated RIR

(d) A typical measured RIR

Fig. 10.2 Room configuration and the corresponding typical RIR for simulated and real scenarios.

power has no influence on the obtained performance, and is therefore kept fixed during the simulations.

In Table 10.3 the multi-constraints algorithm and the TF-GSC are compared in terms of the objective quality measures, as explained above, for various number of interference sources. Since the TF-GSC contains an ANC branch, we compared it to a multi-constraint beamformer that also incorporates an ANC [27]. It is evident from Table 10.3 that the multi-constraint beamformer outperforms the TF-GSC algorithm in terms of SIR improvement, as well as distortion level measured by the SED and LSD values. The lower distortion of the multi-constraint beamformer can be attributed to the stable nature of the nulls in the beam-pattern as compared with the adaptive nulls of the TF-GSC structure. The results in the Tables were obtained using

Table 10.2 Test scenario #1: 11 microphone array, 2 desired speakers, 2 interfering speakers at $6dB$ SIR, and one stationary noise at $13dB$ SIR with various reverberation levels. SIR improvement in dB for the LCMV output and speech distortion measures (SED and LSD in dB) between the desired source component received by microphone #1 and respective component at the LCMV output.

T_{60}	Source	BF SIR imp.			SED	LSD
		s_1^{ns}	s_2^{ns}	s_1^{s}		
150ms	s_1^d	12.53	14.79	13.07	11.33	1.12
	s_2^d	12.39	14.98	12.93	13.41	1.13
200ms	s_1^d	10.97	12.91	11.20	9.51	1.39
	s_2^d	12.13	13.07	11.36	10.02	1.81
250ms	s_1^d	10.86	12.57	11.07	8.49	1.56
	s_2^d	11.19	12.90	11.40	8.04	1.83
300ms	s_1^d	11.53	11.79	11.21	7.78	1.86
	s_2^d	11.49	11.75	11.17	7.19	1.74

Table 10.3 Test scenario #2-#4: Simulated room environment with reverberation time $T_{60} = 300$mS, 11 microphones, one desired speaker, one stationary noise at $13dB$ SIR, and various number of interfering speakers at $6dB$ SIR. SIR improvement, SED and LSD in dB relative to microphone #1 as obtained by the TF-GSC [14] and the multi-constraint [26] beamformers.

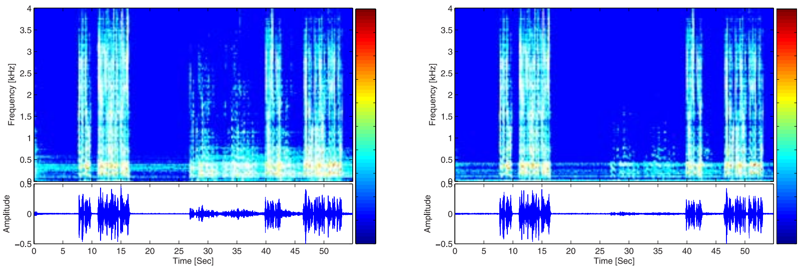
N_{ns}	TF-GSC					Multi-Constraint				
	SIR imp.			SED	LSD	SIR imp.			SED	LSD
	s_1^{ns}	s_2^{ns}	s_1^{s}			s_1^{ns}	s_2^{ns}	s_1^{s}		
0	–	–	15.62	6.97	2.73	–	–	27.77	14.66	1.31
1	9.54	–	13.77	6.31	2.75	21.01	–	23.95	12.72	1.35
2	7.86	10.01	10.13	7.06	2.77	17.58	20.70	17.70	11.75	1.39

the second phase of the test procedure. It is shown that for test scenario #1 the multi-constraint beamformer can gain an average value of 12.1dB SIR improvement for both stationary and non-stationary interferences.

The multi-constraints algorithm and the TF-GSC were also subjectively compared by informal listening tests and by the assessment of waveforms and sonograms. The outputs of the TF-GSC [14] the multi-constraint algorithm [26] algorithm for test scenario #3 (namely, one competing speaker) are depicted in Fig. 3(a) and Fig. 3(b) respectively. It is evident that the multi-constraint beamformer outperforms the TF-GSC beamformer especially in terms of the competing speaker cancellation. Speech samples demonstrating the performance of the proposed algorithm can be downloaded from [32].

10.6.3 Real Environment

In the real room environment we used as the directional signals four speakers drawn from the TIMIT [33] database and the speech-like noise described above. The performance was evaluated using real medium-size conference



(a) The TF-GSC [14] output.

(b) The LCMV [26] output.

Fig. 10.3 Test scenario #3: Sonograms depicting the difference between TF-GSC and LCMV.

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 various sources that were played separately from point loudspeakers (FOSTEX 6301BX). The algorithm's input was constructed by summing up all non-stationary components contributions with a 6dB SIR, the stationary noise with 13dB SIR and additional, spatially white, computer-generated sensor noise signals. The source-microphone constellation is depicted in Fig. 2(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. 2(d). A total SIR improvement of $15.28dB$ was obtained for the interfering speakers and $16.23dB$ for the stationary noise.

10.7 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 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. 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. We therefore note, 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 for extracting several desired sources from a combination of multiple sources in a complicated acoustic environment. In the case of one desired source, two alternative beamforming strategies for interference cancellation in noisy and reverberant environment were compared. The TF-GSC, which belongs to the MVDR family, applies a single constraint towards the desired signal, leaving the interference mitigation adaptive. Alternatively, the multi-constraint beamformer implicitly applies carefully designed nulls towards all interference signals. It is shown that for the time-invariant scenario the later design shows a significant advantage over the former beamformer design. It remains an open question what is the preferred strategy in slowly time-varying scenarios.

References

1. J. Cardoso, "Blind signal separation: Statistical principles," *Proc. of the IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.
2. E. Jan and J. Flanagan, "Microphone arrays for speech processing," *Int. Symposium on Signals, Systems, and Electronics (ISSSE)*, pp. 373–376, Oct. 1995.
3. 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.
4. S. Gannot and I. Cohen, *Springer Handbook of Speech Processing.* Springer, 2007, ch. Adaptive Beamforming and Postfitering, pp. 199–228.
5. 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.
6. 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.
7. 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.
8. J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.
9. O. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.
10. 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.
11. 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.
12. 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.
13. 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.

14. 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.
15. 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.
16. B. Yang, "Projection approximation subspace tracking," *IEEE Trans. Signal Processing*, vol. 43, no. 1, pp. 95–107, Jan. 1995.
17. 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.
18. 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.
19. 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.
20. 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.
21. 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.
22. R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagate.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.
23. 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.
24. 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.
25. 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.
26. S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *submitted to IEEE Transactions on Audio, Speech and Language Processing*, Jul. 2008.
27. ——, "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.
28. J. J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Processing Magazine*, vol. 9, no. 1, pp. 14–37, Jan. 1992.
29. 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.
30. 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.
31. E. Habets, "Room impulse response (RIR) generator," `http://home.tiscali.nl/ehabets/rir_generator.html`, Jul. 2006.
32. S. Gannot, "Audio sample files," `http://www.biu.ac.il/~gannot`, Sep. 2008.
33. 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).