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

Shmulik Markovich

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

Final Paper Thesis

As Partial Fulfillment of the Requirements for
the Degree Master of Science in Electrical Engineering

Shmulik Markovich

Submitted to the Senate of the Technion - Israel Institute of Technology

Av 5768 Haifa August 2008

The Final Paper Thesis was Done under the Supervision of Dr. Sharon Gannot from the School of Engineering at Bar-Ilan University and Prof. Israel Cohen from the Department of Electrical Engineering at the Technion.

Acknowledgement

I wish to express my deep gratitude and appreciation to my supervisors Dr. Sharon Gannot and Prof. Israel Cohen for their guidance and dedicated supervision. Thank for your professional support, for your encouragement to perfection, and for many valuable suggestions throughout all the stages of this research. I would also like to thank Dr. Emanuel Habets for many fruitful discussions.

Special thanks to my beloved Liran who encouraged and supported me through the whole way, to my mother Orit, my late father Yaakov, and to Eli and Orly Golan.

The Generous Financial Help of the Technion is Gratefully Acknowledged.

Contents

List of Figures

List of Tables

List of Algorithms

Abstract

In many practical environments we wish to extract several desired speech signals, which are contaminated by both non-stationary interfering signals (such as a competing talkers), as well as by stationary noise. Furthermore, the received signals are often subject to distortion imposed by the RIR of the acoustic environment.

Typical examples for this problem include the conference call scenario with multiple participants; a hands-free cellular phone conversation in a car environment, when several speaking passengers interfere with the desired speaker; and the Cocktail Party scenario, in which desired conversation blend with many simultaneous conversations.

In this thesis multi-microphone measurements are utilized to perform the task of the desired speakers extraction, by designing the array beam-pattern to satisfy a set of multiple linear constraints. One subset of the constraints is dedicated to maintain the desired signals and the second subset is chosen to mitigate both the stationary and non-stationary interference signals. Unlike classical beamformers, in which the RIRs are approximated by a delay-only filter, we take into account the entire RIR [or its respective ATF].

Firstly, we show that the RTFs, defined as the ratio between ATFs relating the speech sources and the microphones, suffice for the construction of the beamformer. Secondly, the null subspace, comprised of all interfering signals, is estimated by using the union of all estimated eigenvectors, relaxing the commonly used demand that the interference signals' activity periods do not overlap. Finally, the Generalized Eigenvalue Decomposition (GEVD) procedure is applied to the received signals' Power Spectrum Density (PSD) matrix and the interference-only PSD matrix (obtained by the second stage) for estimating the RTF of the desired signals.

It is shown that an application of the adaptive RNC to the output of the beamformer enables further reduction of the residual interference signals, caused by inaccuracies in the subspace estimation, and hence increases the robustness of the proposed method.

A comprehensive experimental study, consisting of both simulated and real environments,

proves the applicability of the proposed algorithm to the multiple source extraction task. Furthermore, it is shown that the proposed algorithm outperforms the TF-GSC algorithm, in the task of enhancing one desired speech signal contaminated by several interference signals.

Notation

x	scalar
$\boldsymbol{x}$	column vector
x_i	the ith element of the vector $\boldsymbol{x}$
$\boldsymbol{A}$	matrix
A_{ij}	the (i,j) element of the matrix $\boldsymbol{A}$
$\boldsymbol{A}^{-1}$	matrix inverse
$(\cdot)^T$	transpose operation
$(\cdot)^*$	conjugate operation
$(\cdot)^\dagger$	transpose-conjugate operation
$\mathrm{diag}\{\boldsymbol{x}\}$	diagonal matrix with the vector $\boldsymbol{x}$ on its diagonal
$(\cdot)^{\frac{1}{2}}$	for diagonal matrices, a diagonal matrix with the square root of the diagonal
$\|\cdot\|$	Euclidian norm operation
$\boldsymbol{I}$	identity matrix
$\mathrm{E}\,(\cdot)$	expectation operation
$x(\ell,k)$	time-frequency coefficient

Abbreviations

BSS Blind Source Separation

LTI Linear Time Invariant

PSD Power Spectrum Density

MV Minimum Variance

DS Delay and Sum

LCMV Linearly Constrained Minimum Variance

MVDR Minimum Variance Distortionless Response

GSC Generalized Sidelobe Canceler

ATF Acoustic Transfer Function

RTF Relative Transfer Function

EVD Eigenvalue Decomposition

GEVD Generalized Eigenvalue Decomposition

ICA Independent Component Analysis

NR Noise Reduction

MMSE Minimum Mean Squared Error

MSNR Maximum Signal to Noise Ratio

GSVD Generalized Singular Value Decomposition

ANC Adaptive Noise Canceler

LMS Least Mean Squares

NLMS Normalized Least Mean Squares

FBF Fixed Beamformer

D-GSC Delay Generalized Sidelobe Canceler

TF-GSC Transfer Function Generalized Sidelobe Canceler

DTF-GSC Dual Transfer Function Generalized Sidelobe Canceler

RIR Room Impulse Response

RNC Residual Noise Canceler

MTF Multiplicative Transfer Function

FIR Finite Impulse Response

STFT Short Time Fourier Transform

BM Blocking Matrix

OMLSA Optimally Modified Log Spectral Amplitude

SIR Signal to Interference Ratio

EDC Energy Decay Curve

QR Orthogonal Triangular Decomposition

DFT Discrete Fourier Transform

BF Beamformer

VAD Voice Activity Detector

MWF Multichannel Wiener Filter

SDW-MWF Speech Distortion Weighted Multichannel Wiener Filter

MIMO Multiple Input Multiple Output

DOA Direction of Arrival

PASTd Projection Approximation Subspace Tracking (deflation)

DRR Direct to Reverberant Ratio

MUSIC Multiple Signal Classification

NSR Noise-dominant Subspace Reduction

SNR Signal to Noise Ratio

CSS Coherent Subspaces

Chapter 1

Introduction

Speech enhancement techniques, utilizing microphone arrays, have attracted the attention of many researchers for the last twenty 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.

In this thesis 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 BSS family and the beamforming family. BSS aims at separating all the involved sources, regardless of their attribution to the desired or interfering sources [2]. 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. Since the BSS family of algorithms is not focal point of this thesis, it is only shortly introduced in Section 1.1. The beamforming family of algorithms is surveyed in Section 1.2. In Section 1.3 the structure of the thesis is presented.

1.1 BSS Algorithms

The BSS family of algorithms exploit the independence of the involved sources. Independent Component Analysis (ICA) algorithms [3, 4] 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* [5], *high-order statistics* [6], and

Information theoretic based measures [7]. BSS methods can also be used in reverberant environments, but they tend to get very complex (for time domain approaches [8]) or have an inherent problem of *permutation and gain ambiguity* [9] (for frequency domain algorithms [4]).

1.2 Beamforming 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 [10] extended the delay and sum concept by introducing the 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 [11, 12]. Cox et al. [13] introduced an improved adaptive beamformer that maintains a set of linear constraints as well as a quadratic inequality constraint.

In [14] 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) [15].

In a MVDR beamformer [16, 17], 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 [18]. The standard GSC consists of a spatial pre-processor, i.e. a Fixed Beamformer (FBF) and a Blocking Matrix (BM), combined with a multichannel Adaptive Noise Canceler (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 [19]) 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 steer-direction gain constraints. 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 [20] proved that the LCMV extension has also an equivalent GSC [18] structure, which decouples the constraining and the minimization operations. The GSC structure was rederived in the frequency domain, and extended to deal with the more complicated general ATFs case by Affes and Grenier [21] and later by Gannot et al. [22]. The latter frequency-domain version, which takes into account the reverberant nature of the enclosure, was nicknamed the TF-GSC. Though related, these contributions differ in their channel identification. Affes and Grenier require an a-priori calibration of the propagated energy at each frequency in order to obtain the ATF by exploring the eigen-space of the correlation matrix. Gannot et al. deals with arbitrary sources and microphones location, and takes advantage of the non-stationarity of the speech signals opposed to the slowly varying ATF to estimate the RTF.

Several beamforming algorithms based on subspace methods were developed. Ephraim and Van Trees [23] 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 [24] 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. [25] 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 [26] is derived. An extension to the wide-band case is presented by the same authors [27]. However the demand for a delay-only impulse response is still not relaxed. Affes and Grenier [21] 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 [28] 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. [29] propose to replace the BM in [22]. 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 to the microphones.

Affes et al. [30] extend the structure presented in [25] 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. [1] 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-TF-GSC is tailored to the specific problem of two sources and cannot be easily generalized to the multiple desired and interference sources.

In the next chapter we formulate the problem and elaborate on two of the algorithms coping the speech extraction. The first is the TF-GSC algorithm, and the second is the algorithm presented by Asano et al. [1].

In this contribution 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. We derive a practical method for estimating all components of the eigenspace-based beamformer. We first show that the RTFs, defined as the ratio between ATFs relating the speech sources and the microphones, is a sufficient quantity for the construction of the beamformer. We relax the commonly used demand that the interference signals' activity periods do not overlap and estimate the null subspace, comprised of all interfering signals. For the final estimation of the RTFs of the desired signals, the GEVD procedure is applied to the estimated PSD matrix of the noisy microphone signals and the PSD matrix of the interference-only components (obtained by the second stage).

1.3 Thesis Structure

The structure of the thesis is as follows. In Chapter 2 the problem of extracting multiple desired sources contaminated by multiple interference in reverberant environment is introduced. Two recent algorithms that deal with a subset of the general problem are then explored. The drawbacks of these algorithms are discussed, motivating the new research. In Chapter 3 we present a novel method for source extraction based on the LCMV beamformer. The various components of the beamformer are estimated using eigenspace analysis. In Chapter 4 the proposed algorithm is evaluated in different scenarios. For the single desired source scenario the proposed method is further compared with the TF-GSC algorithm . In Chapter 5 the thesis is concluded and future research topics are proposed.

Chapter 2

Background

In this chapter the problem of multiple sources extraction is mathematically stated. Two earlier attempts providing partial solution of the general problem are explored. The problem is formulated in Section 2.1. The TF-GSC [22], aiming at the enhancement of one desired source in noisy and reverberant environment is explored in Section 2.2. A subspace method, proposed by Asano et al. [1], for extracting multiple sources in mildly reverberant environment is presented in Section 2.3. We conclude this chapter in Section 2.4 by discussing the limitations of these methods, motivating the derivation of the novel algorithm, presented in Chapter 3.

2.1 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:

$$z_m(n) = \sum_{i=1}^{K} s_i^d(n) * h_{im}^d(n) + \sum_{i=1}^{N_s} s_i^s(n) * h_{im}^s(n) + \sum_{i=1}^{N_{ns}} s_i^{ns}(n) * h_{im}^{ns}(n) + v_m(n) \qquad (2.1)$$

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 a spatially white noise with

zero mean and variance σ_v^2 . $z_m(n)$ is transformed into the Short Time Fourier Transform (STFT) domain with a rectangular window of length N_{DFT}, yielding:

$$z_m(\ell,k) = \sum_{i=1}^{K} s_i^d(\ell,k) h_{im}^d(\ell,k) + \sum_{i=1}^{N_s} s_i^s(\ell,k) h_{im}^s(\ell,k) + \sum_{i=1}^{N_{ns}} s_i^{ns}(\ell,k) h_{im}^{ns}(\ell,k) + v_m(\ell,k) \quad (2.2)$$

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 (2.2) can be formulated in vector notation:

$$\begin{aligned} \boldsymbol{z}(\ell,k) &= \boldsymbol{H}^d(\ell,k)\boldsymbol{s}^d(\ell,k) + \boldsymbol{H}^s(\ell,k)\boldsymbol{s}^s(\ell,k) + \boldsymbol{H}^{ns}(\ell,k)\boldsymbol{s}^{ns}(\ell,k) + \boldsymbol{v}(\ell,k) \\ &= \boldsymbol{H}(\ell,k)\boldsymbol{s}(\ell,k) + \boldsymbol{v}(\ell,k) \end{aligned} \quad (2.3)$$

where

$$\begin{aligned}
\boldsymbol{z}(\ell,k) &\triangleq \left[\, z_1(\ell,k) \quad \dots \quad z_M(\ell,k) \,\right]^T \\
\boldsymbol{v}(\ell,k) &\triangleq \left[\, v_1(\ell,k) \quad \dots \quad v_M(\ell,k) \,\right]^T \\
\boldsymbol{h}_i^d(\ell,k) &\triangleq \left[\, h_{i1}^d(\ell,k) \quad \dots \quad h_{iM}^d(\ell,k) \,\right]^T \qquad i = 1,\dots,K \\
\boldsymbol{h}_i^s(\ell,k) &\triangleq \left[\, h_{i1}^s(\ell,k) \quad \dots \quad h_{iM}^s(\ell,k) \,\right]^T \qquad i = 1,\dots,N_s \\
\boldsymbol{h}_i^{ns}(\ell,k) &\triangleq \left[\, h_{i1}^{ns}(\ell,k) \quad \dots \quad h_{iM}^{ns}(\ell,k) \,\right]^T \qquad i = 1,\dots,N_{ns} \\
\boldsymbol{H}^d(\ell,k) &\triangleq \left[\, \boldsymbol{h}_1^d(\ell,k) \quad \dots \quad \boldsymbol{h}_K^d(\ell,k) \,\right] \\
\boldsymbol{H}^s(\ell,k) &\triangleq \left[\, \boldsymbol{h}_1^s(\ell,k) \quad \dots \quad \boldsymbol{h}_{N_s}^s(\ell,k) \,\right] \\
\boldsymbol{H}^{ns}(\ell,k) &\triangleq \left[\, \boldsymbol{h}_1^{ns}(\ell,k) \quad \dots \quad \boldsymbol{h}_{N_{ns}}^{ns}(\ell,k) \,\right] \\
\boldsymbol{H}^i(\ell,k) &\triangleq \left[\, \boldsymbol{H}^s(\ell,k) \quad \boldsymbol{H}^{ns}(\ell,k) \,\right] \\
\boldsymbol{H}(\ell,k) &\triangleq \left[\, \boldsymbol{H}^d(\ell,k) \quad \boldsymbol{H}^s(\ell,k) \quad \boldsymbol{H}^{ns}(\ell,k) \,\right] \\
\boldsymbol{s}^d(\ell,k) &\triangleq \left[\, s_1^d(\ell,k) \quad \dots \quad s_K^d(\ell,k) \,\right]^T \\
\boldsymbol{s}^s(\ell,k) &\triangleq \left[\, s_1^s(\ell,k) \quad \dots \quad s_{N_s}^s(\ell,k) \,\right]^T \\
\boldsymbol{s}^{ns}(\ell,k) &\triangleq \left[\, s_1^{ns}(\ell,k) \quad \dots \quad s_{N_{ns}}^{ns}(\ell,k) \,\right]^T \\
\boldsymbol{s}(\ell,k) &\triangleq \left[\, (\boldsymbol{s}^d(\ell,k))^T \quad (\boldsymbol{s}^s(\ell,k))^T \quad (\boldsymbol{s}^{ns}(\ell,k))^T \,\right]^T .
\end{aligned}$$

Assuming the desired speech signals, the interference and the noise signals to be uncorrelated,

we have from 2.3:

$$\begin{aligned}\boldsymbol{\Phi}_{zz}(\ell,k) &= \boldsymbol{H}^{d}(\ell,k)\boldsymbol{\Lambda}^{d}(\ell,k)\big(\boldsymbol{H}^{d}(\ell,k)\big)^{\dagger}+ \\ &\boldsymbol{H}^{ns}(\ell,k)\boldsymbol{\Lambda}^{ns}(\ell,k)\big(\boldsymbol{H}^{ns}(\ell,k)\big)^{\dagger}+\boldsymbol{H}^{s}(\ell,k)\boldsymbol{\Lambda}^{s}(\ell,k)\big(\boldsymbol{H}^{s}(\ell,k)\big)^{\dagger}+\boldsymbol{\Phi}_{vv}(\ell,k) \\ &\triangleq \boldsymbol{H}(\ell,k)\boldsymbol{\Lambda}(\ell,k)\boldsymbol{H}^{\dagger}(\ell,k)+\boldsymbol{\Phi}_{vv}(\ell,k)\end{aligned} \tag{2.4}$$

where

$$\begin{aligned}\boldsymbol{\Lambda}^{d}(\ell,k) &\triangleq \operatorname{diag}\left(\left[\begin{array}{ccc}(\sigma_{1}^{d}(\ell,k))^{2} & \dots & (\sigma_{K}^{d}(\ell,k))^{2}\end{array}\right]\right) \\ \boldsymbol{\Lambda}^{s}(\ell,k) &\triangleq \operatorname{diag}\left(\left[\begin{array}{ccc}(\sigma_{1}^{s}(\ell,k))^{2} & \dots & (\sigma_{N_s}^{s}(\ell,k))^{2}\end{array}\right]\right) \\ \boldsymbol{\Lambda}^{ns}(\ell,k) &\triangleq \operatorname{diag}\left(\left[\begin{array}{ccc}(\sigma_{1}^{ns}(\ell,k))^{2} & \dots & (\sigma_{N_{ns}}^{ns}(\ell,k))^{2}\end{array}\right]\right) \\ \boldsymbol{\Lambda}(\ell,k) &\triangleq \operatorname{blkdiag}\left(\begin{array}{ccc}\boldsymbol{\Lambda}^{d}(\ell,k) & \boldsymbol{\Lambda}^{s}(\ell,k) & \boldsymbol{\Lambda}^{ns}(\ell,k)\end{array}\right).\end{aligned}$$

$(\bullet)^{\dagger}$ is the conjugate-transpose operation, $\operatorname{diag}(\bullet)$ is a square matrix with the vector in brackets on its main diagonal, and $\operatorname{blkdiag}(\bullet)$ is a block diagonal matrix with the matrices in brackets on its main diagonal. Further define the PSD of the stationary component of the output signal as:

$$\boldsymbol{\Phi}_{zz}^{s}(\ell,k) \triangleq \boldsymbol{H}^{s}(\ell,k)\boldsymbol{\Lambda}^{s}(\ell,k)\big(\boldsymbol{H}^{s}(\ell,k)\big)^{\dagger}+\boldsymbol{\Phi}_{vv}(\ell,k). \tag{2.5}$$

It is usually assumed that $\boldsymbol{\Phi}_{vv}(\ell,k)=\sigma_{v}^{2}\boldsymbol{I}_{M\times M}$ where $\boldsymbol{I}_{M\times M}$ is the identity matrix, i.e. the noise field is assumed to be non-coherent, spatially-white.

A beamformer is constructed by applying a set of filters $w^{*}(\ell,k)$ to each microphone signal and summing up all the signals:

$$y(\ell,k)=\boldsymbol{w}^{\dagger}(\ell,k)\boldsymbol{z}(\ell,k) \tag{2.6}$$

where $y(\ell,k)$ is the beamformer output at the STFT representation, and $\boldsymbol{w}(\ell,k)$ is the beamformer's weights at time frame ℓ and frequency bin k.

2.2 TF-GSC

An approach for signal enhancement based on the desired signal non-stationarity was proposed by Gannot et al. [22]. This approach, aiming at enhancing a single desired source, can be exploited for extracting several desired sources by activating K beamformers in parallel, one for each desired source. Without loss of generality we discuss the enhancement of

the first desired speaker. The beamformer, $\boldsymbol{w}(\ell,k)$, is determined by solving the following minimization problem:

$$\underset{\boldsymbol{w}}{\operatorname{argmin}} \left\{ \boldsymbol{w}^{\dagger}(\ell,k)\boldsymbol{\Phi}_{zz}(\ell,k)\boldsymbol{w}^{\dagger}(\ell,k) \right\} \text{subject to } \boldsymbol{w}^{\dagger}(\ell,k)\boldsymbol{h}_1^d(\ell,k) = 1\,. \tag{2.7}$$

Under this criterion the signal component at the output is equal to the desired signal $s_1^d(\ell,k)$. It is shown, that if the constraint is relaxed, such that the desired signal component at the output is given by $s_1^d(\ell,k)h_{i1}^d(\ell,k)$, i.e. the desired signal as received by the first microphone, the RTF

$$\tilde{\boldsymbol{h}}_1^d(\ell,k) \triangleq \frac{\boldsymbol{h}_i^d(\ell,k)}{h_{11}^d(\ell,k)} \tag{2.8}$$

suffices for implementing the MVDR beamformer.

This minimization can be efficiently implemented by constructing a GSC structure as depicted in Fig. 2.1. The GSC solution is comprised of three components: A FBF responsible

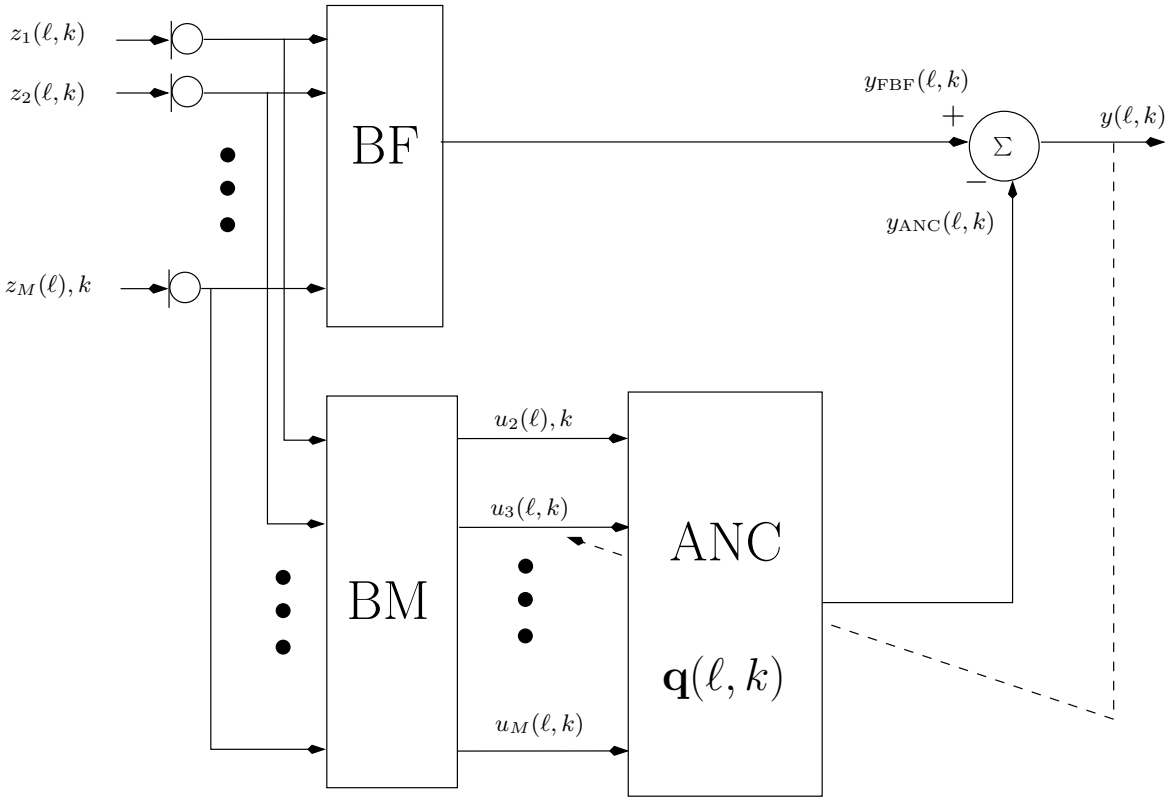


Figure 2.1: GSC solution for the general ATFs case (TF-GSC)

of aligning the desired signal component, a BM which blocks the desired signal and constructs noise reference signals (comprised of all stationary and transient noise components as

well as all competing speakers), and a multichannel ANC which cancels out all interference components from the FBF output by using the reference signals.

It is shown in [22] that the FBF can be implemented by:

$$\boldsymbol{w}_0(\ell,k) = \frac{\tilde{\boldsymbol{h}}_1^d(\ell,k)}{\|\tilde{\boldsymbol{h}}_1^d(\ell,k)\|^2} \tag{2.9}$$

and that

$$\boldsymbol{B}(\ell,k) = \begin{bmatrix} -(h_{12}^d(\ell,k))^* & -(h_{13}^d(\ell,k))^* & \dots & -(h_{1M}^d(\ell,k))^* \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & & \dots & \ddots \\ 0 & 0 & \dots & 1 \end{bmatrix}. \tag{2.10}$$

is a proper BM. The ANC filters $\boldsymbol{q}(\ell,k)$ are adjusted to minimize the power at the output, $y(\ell,k)$, exactly as in the classical Widrow problem. The filters are usually constrained to an FIR structure for stabilizing the update algorithm.

In practice, $\tilde{\boldsymbol{h}}^d(\ell,k)$ is unknown and should be estimated. The estimation method presented in [22] is based on the nonstationarity of the desired speech signal. The analysis interval is split into frames, such that the desired signal may be considered stationary during each frame (quasistationarity assumption for speech signals), while $\tilde{\boldsymbol{h}}_1^d(\ell,k) \approx \tilde{\boldsymbol{h}}_1^d(k)$ is still considered time-invariant. Define $\phi_{z_i z_j}(\ell,k)$ as the cross PSD between z_i and z_j (ith and jth noisy signal observations, respectively) during the ℓth frame ($\ell = 1,\dots,L$). Further define $\phi_{u_m z_1}(\ell,k)$ the cross PSD between $u_m(n)$ and $z_1(n)$. Let $\hat{\phi}_{z_i z_j}(\ell,k)$ and $\hat{\Phi}_{u_m z_1}(\ell,k)$ represent the corresponding estimates. Further define The error term $\varepsilon_m(\ell,k) = \phi_{u_m z_1}(k) - \hat{\phi}_{u_m z_1}(k)$; $\ell = 1,\dots,L$. An unbiased estimate for $\tilde{h}_m^d(k)$ is obtained by applying the *least squares* criterion to the following set of over-determined equations

$$\begin{bmatrix} \hat{\phi}_{z_m z_1}(1,k) \\ \hat{\phi}_{z_m z_1}(2,k) \\ \vdots \\ \hat{\phi}_{z_m z_1}(L,k) \end{bmatrix} = \begin{bmatrix} \hat{\phi}_{z_1 z_1}(1,k) & 1 \\ \hat{\phi}_{z_1 z_1}(2,k) & 1 \\ \vdots & \\ \hat{\phi}_{z_1 z_1}(L,k) & 1 \end{bmatrix} \times \begin{bmatrix} \tilde{h}_{1m}^d(k) \\ \phi_{u_m z_1}(k) \end{bmatrix} + \begin{bmatrix} \varepsilon_m(1,k) \\ \varepsilon_m(2,k) \\ \vdots \\ \varepsilon_m(L,k) \end{bmatrix} \tag{2.11}$$

where a separate set of equations is used for each microphone signal ($m = 2,\dots,M$) and frequency index k.

A summery of the entire algorithm is given in Alg. 1.

Algorithm 1 Summary of the TF-GSC algorithm.

1) **RTFs**: $\tilde{\boldsymbol{h}}_1^d(k) = \frac{\boldsymbol{h}_1^d(k)}{h_{11}(k)}$
2) Construct a **blocking matrix**, $\boldsymbol{B}^\dagger(k)\boldsymbol{h}_1^d(k) = 0$.
3) FBF $\boldsymbol{w}_0(k) = \frac{\tilde{\boldsymbol{h}}_1^d(k)}{\|\tilde{\boldsymbol{h}}_1^d(k)\|^2}$.
   FBF output: $y_{\text{FBF}}(\ell,k) = \boldsymbol{w}_0^\dagger(k)\boldsymbol{z}(\ell,k)$.
4) **Noise reference signals**:
   $\boldsymbol{u}(\ell,k) = \boldsymbol{B}^\dagger(k)\boldsymbol{z}(\ell,k)$
5) **Output signal**: $y(\ell,k) = y_{\text{FBF}}(\ell,k) - \boldsymbol{q}^\dagger(\ell,k)\boldsymbol{u}(\ell,k)$.
6) **Filters update**. For $m = 1,\ldots,M-1$:
   $\tilde{q}_m(\ell+1,k) = q_m(\ell,k) + \mu\frac{u_m(\ell,k)y^*(\ell,k)}{p_{est}(\ell,k)}$
   $q_m(\ell+1,k) \xleftarrow{\text{FIR}} \tilde{q}_m(\ell+1,k)$
   where, $p_{est}(\ell,k) = \rho p_{est}(\ell-1,k) + (1-\rho)\|\boldsymbol{z}(\ell,k)\|^2$.
7) Keep only non-aliased samples, according to the
   overlap & save method.

2.3 Speech Enhancement Based on the Subspace Method (Asano et al. [1])

Asano et al. propose in [1] an algorithm for speech recognition that treats separately the directional and the non-coherent interference signals. This method can be adopted to the speech enhancement problem at hand. In this section the adopted algorithm is presented.

The algorithm derivation is based on two restrictive assumptions. First, the RIRs of directional signals are approximated by the first arrival. Second, it is assumed that the direct arrival of each of the sources is uncorrelated with its late arrivals. Hence, the latter can be attributed to a non-coherent noise component. Under the first assumption we have:

$$h_{im}^d(\ell,k) \approx e^{-j\frac{2\pi}{N_{\text{DFT}}}\tau_{im}} \tag{2.12}$$

where τ_{im} is the relative delay between source $i = 1,2,\ldots,N$ (desired, stationary and non-stationary interference signals) and microphone $m = 1,2,\ldots,M$. Due to the second assumption the non-coherent noise source cannot be regarded as white noise anymore, i.e. $\boldsymbol{\Phi}_{vv}(\ell,k) \neq \sigma_v^2\boldsymbol{I}_{M\times M}$.

The entire structure of the algorithm is depicted in Fig. 2.2 . At the first stage the received signal and the non-coherent signal PSD matrices, $\boldsymbol{\Phi}_{zz}(\ell,k)$ and $\boldsymbol{\Phi}_{vv}(\ell,k)$ are estimated. In this stage, denoted Coherent Subspaces (CSS), the PSD estimates are smoothed in the frequency domain to reduce the estimation variance.

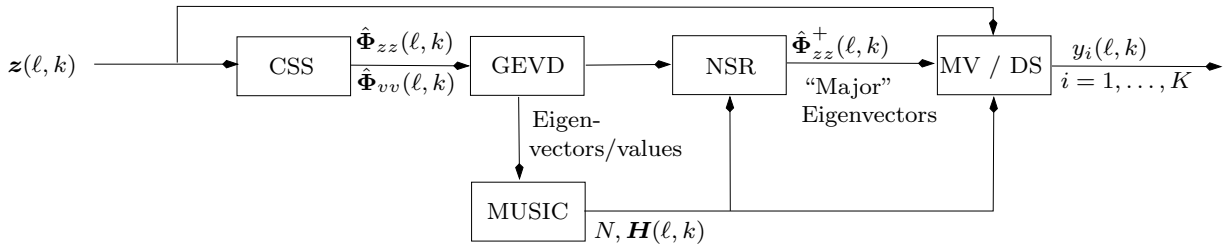


Figure 2.2: The structure of the speech enhancement based on subspace decomposition (Asano et al. [1])

Then, at the second stage, the GEVD of the PSD matrices is calculated, yielding a set of eigenvectors. The corresponding eigenvalues are inspected in the third step, where the MUSIC algorithm is used for determining the number and values of the eigenvectors corresponding to the largest eigenvalues. These eigenvectors can be attributed to the coherent sources, $\boldsymbol{h}_i^d(\ell,k), \boldsymbol{h}_i^s(\ell,k), \boldsymbol{h}_i^{ns}(\ell,k)$. The MUSIC algorithm can be applied due to the simplified model of the ATFs. When the reverberant energy becomes significant, further processing is required for determining the dominant eigenvectors. All weak eigenvalues and their corresponding eigenvectors are discarded in the fourth stage, denoted Noise-dominant Subspace Reduction (NSR), yielding the following PSD matrix:

$$\boldsymbol{\Phi}^{+}_{zz}(\ell,k) \approx \boldsymbol{H}(\ell,k)\boldsymbol{\Lambda}(\ell,k)\boldsymbol{H}^{\dagger}(\ell,k)$$

which is only comprised of coherent components.

Finally, K beamformers are constructed for extracting all the desired sources separately. the MVDR beamformer can be designed using:

$$\boldsymbol{w}_i^{\mathrm{MVDR}}(\ell,k) = \frac{(\boldsymbol{\Phi}^{+}_{zz}(\ell,k))^{-1}\boldsymbol{h}_i^d(\ell,k)}{(\boldsymbol{h}_i^d(\ell,k))^{\dagger}(\boldsymbol{\Phi}^{+}_{zz}(\ell,k))^{-1}\boldsymbol{h}_i^d(\ell,k)}$$

for $i = 1, 2, \ldots, K$. Alternatively a simple delay and sum beamformer can be used.

Several drawbacks can be encountered in the application of the method to the problem at hand. First, approximating the RIRs of the desired sources by their direct-path component becomes very inaccurate when the DRR reduces. Second, estimating $\boldsymbol{\Phi}_{vv}(\ell,k)$ might be a cumbersome task, due to the strong correlation between the direct-path and the late arrivals. Moreover, free decay periods of the reverberant tail, which are necessary for estimating $\boldsymbol{\Phi}_{vv}(\ell,k)$ cannot be easily identified.

These problems limit the applicability of the method in [1] to the multi-interference mitigation task in reverberant environments.

2.4 Discussion

We will briefly discuss now the drawbacks of the two presented algorithms [22, 1], limiting their ability to deal with the general source extraction problem.

A major drawback of the TF-GSC lies in the adaptation mechanism of its ANC. The input signal of the ANC is comprised of both nonstationary competing speech signals and stationary noise signals. The need of the ANC to adapt during both the stationary and nonstationary signals impose contradicting requirements on the adaptation rate. On the one hand, the adaptation factor μ should be high enough to allow for tracking of a fast varying signal, and on the other-hand it should be low enough to enable sufficient reduction of the stationary noise level. Such a requirement on the TF-GSC necessitates the use of an algorithm for distinguishing between the two types of nonstationary signals, namely, the desired and interference signals. Since the null towards the competing speech signals is adaptively established, insufficient amount of interference signal cancellation can be expected.

As will be shown in the experimental study in Sec. 4.4 the residual noise level of the TF-GSC structure is fluctuating over time (in accordance with the activity periods of the interference signal). As beamformer algorithms are very often followed by a postfilter [35, 36], and since postfilters are sensitive to nonstationary noise signals, it is crucial to maintain the residual noise level as stable as possible. The application of a postfilter is beyond the scope of this thesis.

Another drawback of the TF-GSC is its dependence on the FBF beampattern. If the interference signal level at the FBF output is significantly reduced while maintaining high level at the BM output, the ANC might increase the amount of interference leakage to the total output.

Several of the assumptions leading to the algorithm presented by Asano [1] cannot be met in highly reverberant environments. While it is reasonable to assume that the early arrivals of the signals are uncorrelated with the late arrivals, it is very difficult to obtain a clean estimate of the late arrivals PSD. In severe cases it is not guaranteed that the noise PSD matrix $\Phi_{nn}(k)$ contains any reverberant component. Hence, the presented algorithm can only be applied when the DRR is sufficiently high. These conditions can be met only either if the desired sources and the microphone are closely spaced or if the reverberation time is low.

To conclude: for the TF-GSC only limited amount of cancellation of nonstationary inter-

ference signals (such as competing speakers) is expected, and for the algorithm in [1] high performance in highly reverberant environment cannot be guaranteed. It is therefore the aim of this thesis to suggest a novel algorithms that might circumvent these limitations. Our proposed method is presented in Chapter .

Chapter 3

The Proposed method

In this chapter a novel method for multiple speakers extraction is presented. In Section 3.1 we introduce the proposed LCMV beamformer. In Section 3.1.1 a straightforward linear constraints set is introduced. This set necessitates the availably of the ATFs relating the sources and the microphones. Later, in Section 3.1.2 an equivalent constraints set, which replaces the actual ATFs of the interfering sources with any arbitrary basis spanning the same subspace, is derived. The demand for the exact knowledge of the desired sources ATFs is relaxed, in Section 3.1.3. It is shown that when the RTFs, relating the sources and the microphones, are used rather than the corresponding ATFs, a beamformer that extracts the desired sources, can still be designed. However, the new beamformer design is now aiming at extracting the signals as captured by the first microphone rather than the original signals.

Since the interferences subspace and the desired RTFs are usually not available and only approximate estimates are available, we implement in Section 3.2 a RNC branch in parallel to the proposed beamformer, which accounts for estimation errors in the interferences subspace. An estimation scheme based on the slow time-variation of the sources' ATFs is introduced in Section 3.3. The interferences subspace is estimated using a novel union operator in Section 3.3.1, and the desired sources RTFs are estimated in Section 3.3.2 using the GEVD of the received signals and stationary noise signal PSD matrices. The algorithm is summarized in Section 3.4.

3.1 The Beamformer

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 $y_{\mathrm{BF}}(\ell,k)$ is given by

$$y_{\mathrm{BF}}(\ell,k) = \boldsymbol{w}^{\dagger}(\ell,k)\boldsymbol{z}(\ell,k) \tag{3.1}$$

where

$$\boldsymbol{w}(\ell,k) = \left[\ w_1(\ell,k),\ldots,w_M(\ell,k)\ \right]^T. \tag{3.2}$$

The filters are set to satisfy the following set of linear constraints:

$$\boldsymbol{C}^{\dagger}(\ell,k)\boldsymbol{w}(\ell,k) = \boldsymbol{g}(\ell,k). \tag{3.3}$$

The well-known solution to the under-determined equation set is:

$$\boldsymbol{w}(\ell,k) \triangleq \boldsymbol{C}(\ell,k)\big(\boldsymbol{C}^{\dagger}(\ell,k)\boldsymbol{C}(\ell,k)\big)^{-1}\boldsymbol{g}(\ell,k) \tag{3.4}$$

In the following subsections we first define a set of constraints used for extracting the desired sources and mitigating the interference sources, then we replace the set by an equivalent set which is more easily implemented. Finally, we relax our constraint for extracting the exact input signals as transmitted by the sources and replace it by a demand for extracting the desired speech components at an arbitrarily chosen microphone. The outcome of the latter, a modified constraints set, will be a feasible system.

3.1.1 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 opposed to the Dual-Source-TF-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:

$$\big(\boldsymbol{h}_i^d(\ell,k)\big)^{\dagger}\boldsymbol{w}(\ell,k) = 1,\ i = 1,\ldots,K. \tag{3.5}$$

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:

$$\big(\boldsymbol{h}_i^s(\ell,k)\big)^{\dagger}\boldsymbol{w}(\ell,k) = 0, \tag{3.6}$$

and

$$\left(\boldsymbol{h}_j^{ns}(\ell,k)\right)^{\dagger}\boldsymbol{w}(\ell,k)=0. \tag{3.7}$$

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, it is obvious that for the solution in (3.4) 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:

$$\boldsymbol{C}(\ell,k) \triangleq \boldsymbol{H}(\ell,k) \tag{3.8}$$

and a desired response vector:

$$\boldsymbol{g} \triangleq \left[\underbrace{1\ \ldots\ 1}_{K}\ \ \underbrace{0\ \ldots\ 0}_{N-K}\right]^T. \tag{3.9}$$

Under these definitions, and using (3.4), (2.3) and (3.1), the beamformer output is given by:

$$\begin{aligned}
y_{\mathrm{BF}}(\ell,k) &= \boldsymbol{w}^{\dagger}(\ell,k)\boldsymbol{z}(\ell,k) = \\
&\boldsymbol{g}^{\dagger}\left(\boldsymbol{C}^{\dagger}(\ell,k)\boldsymbol{C}(\ell,k)\right)^{-1}\boldsymbol{C}^{\dagger}(\ell,k)\left(\boldsymbol{H}(\ell,k)\boldsymbol{s}(\ell,k)+\boldsymbol{v}(\ell,k)\right) = \\
&\boldsymbol{g}^{\dagger}\boldsymbol{s}(\ell,k)+\boldsymbol{g}^{\dagger}\left(\boldsymbol{H}^{\dagger}(\ell,k)\boldsymbol{H}(\ell,k)\right)^{-1}\boldsymbol{H}^{\dagger}(\ell,k)\boldsymbol{v}(\ell,k) = \\
&\sum_{i=1}^{K} s_i^d(\ell,k)+\boldsymbol{g}^{\dagger}\left(\boldsymbol{H}^{\dagger}(\ell,k)\boldsymbol{H}(\ell,k)\right)^{-1}\boldsymbol{H}^{\dagger}(\ell,k)\boldsymbol{v}(\ell,k).
\end{aligned} \tag{3.10}$$

The beamformer output is therefore comprised of a sum two terms. One term is the sum of all desired sources and the second term is the response of the array to the sensor noise.

3.1.2 An Equivalent Constraints Set

The matrix $\boldsymbol{C}(\ell,k)$ in (3.8) is comprised of 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 (3.4) 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

$$N_i \triangleq N_s + N_{ns} \tag{3.11}$$

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[1] 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:

$$\boldsymbol{H}^i(\ell,k) = \boldsymbol{E}(\ell,k)\boldsymbol{\Theta}(\ell,k) \tag{3.12}$$

where $\boldsymbol{\Theta}_{N_i\times N_i}(\ell,k)$ is comprised of the projection coefficients of the original ATFs on the basis vectors. $\boldsymbol{\Theta}_{N_i\times N_i}(\ell,k)$ is usually an invertible matrix.

Define

$$\tilde{\boldsymbol{\Theta}}(\ell,k) \triangleq \begin{bmatrix} \boldsymbol{I}_{K\times K} & \boldsymbol{O}_{K\times N_i} \\ \boldsymbol{O}_{N_i\times K} & \boldsymbol{\Theta}(\ell,k) \end{bmatrix}_{N\times N}. \tag{3.13}$$

Multiplication by $(\tilde{\boldsymbol{\Theta}}^\dagger(\ell,k))^{-1}$ of both sides of the original constraints set in (3.3), with the definitions (3.8)-(3.9), yields:

$$(\tilde{\boldsymbol{\Theta}}^\dagger(\ell,k))^{-1}\boldsymbol{C}^\dagger(\ell,k)\boldsymbol{w}(\ell,k) = (\tilde{\boldsymbol{\Theta}}^\dagger(\ell,k))^{-1}\boldsymbol{g}. \tag{3.14}$$

Starting with the left-hand-side of (3.14) we have:

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

where the equivalent constraint matrix is defined as

$$\dot{\boldsymbol{C}}(\ell,k) \triangleq \begin{bmatrix} \boldsymbol{H}^d(\ell,k) & \boldsymbol{E}(\ell,k) \end{bmatrix}. \tag{3.15}$$

[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.

For the right-hand-side of (3.14)we have:

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

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

$$
\dot{\boldsymbol{C}}^{\dagger}(\ell,k)\boldsymbol{w}(\ell,k) = \boldsymbol{g}. \tag{3.16}
$$

3.1.3 A Modified Constraints Set

Both the original and equivalent constraints sets in (3.3) and (3.16) 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 (3.9)] 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:

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

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

$$
\tilde{\boldsymbol{w}}(\ell,k) \triangleq \dot{\boldsymbol{C}}(\ell,k)\big(\dot{\boldsymbol{C}}^{\dagger}(\ell,k)\dot{\boldsymbol{C}}(\ell,k)\big)^{-1}\tilde{\boldsymbol{g}}(\ell,k). \tag{3.17}
$$

Indeed, the beamformer output is now given by:

$$
\begin{aligned}
&y_{\mathrm{BF}}(\ell,k) = \tilde{\boldsymbol{w}}^{\dagger}(\ell,k)\boldsymbol{z}(\ell,k) = \\
&\quad \tilde{\boldsymbol{g}}^{\dagger}(\ell,k)\big(\dot{\boldsymbol{C}}^{\dagger}(\ell,k)\dot{\boldsymbol{C}}(\ell,k)\big)^{-1}\dot{\boldsymbol{C}}^{\dagger}(\ell,k)\left(\boldsymbol{H}(\ell,k)\boldsymbol{s}(\ell,k) + \boldsymbol{v}(\ell,k)\right) = \\
&\quad \tilde{\boldsymbol{g}}^{\dagger}(\ell,k)\boldsymbol{s}(\ell,k) + \tilde{\boldsymbol{g}}^{\dagger}(\ell,k)\big(\dot{\boldsymbol{C}}^{\dagger}(\ell,k)\dot{\boldsymbol{C}}(\ell,k)\big)^{-1}\dot{\boldsymbol{C}}^{\dagger}(\ell,k)\boldsymbol{v}(\ell,k) = \\
&\quad \sum_{i=1}^{K} h^{d}_{i1}(\ell,k)s^{d}_{i}(\ell,k) + \tilde{\boldsymbol{g}}^{\dagger}(\ell,k)\big(\dot{\boldsymbol{C}}^{\dagger}(\ell,k)\dot{\boldsymbol{C}}(\ell,k)\big)^{-1}\dot{\boldsymbol{C}}^{\dagger}(\ell,k)\boldsymbol{v}(\ell,k)
\end{aligned} \tag{3.18}
$$

as expected from the modified constraint response.

It is easily verified that the modified desired response is related to the original desired response (3.9) by:

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

where:

$$\boldsymbol{\Psi}(\ell,k) = \mathrm{diag}\left(\left[\begin{array}{ccc} h_{11}^{d}(\ell,k) & \dots & h_{K1}^{d}(\ell,k) \end{array}\right]\right)$$

and

$$\tilde{\boldsymbol{\Psi}}(\ell,k) = \left[\begin{array}{cc} \boldsymbol{\Psi}(\ell,k) & \mathbf{O}_{K\times N_i} \\ \mathbf{O}_{N_i\times K} & \boldsymbol{I}_{N_i\times N_i} \end{array}\right].$$

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,

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

Define

$$\tilde{\boldsymbol{C}}(\ell,k) \triangleq \dot{\boldsymbol{C}}(\ell,k)\tilde{\boldsymbol{\Psi}}^{-1}(\ell,k) = \left[\begin{array}{cc} \tilde{\boldsymbol{H}}^{d}(\ell,k) & \boldsymbol{E}(\ell,k) \end{array}\right] \tag{3.19}$$

where

$$\tilde{\boldsymbol{H}}^{d}(\ell,k) \triangleq \left[\begin{array}{ccc} \tilde{\boldsymbol{h}}_{1}^{d}(\ell,k) & \dots & \tilde{\boldsymbol{h}}_{K}^{d}(\ell,k) \end{array}\right] \tag{3.20}$$

with

$$\tilde{\boldsymbol{h}}_{i}^{d}(\ell,k) \triangleq \frac{\boldsymbol{h}_{i}^{d}(\ell,k)}{h_{i1}^{d}(\ell,k)} \tag{3.21}$$

defined as the RTF with respect to microphone #1.

Finally, the modified beamformer is given by:

$$\tilde{\boldsymbol{w}}(\ell,k) \triangleq \tilde{\boldsymbol{C}}(\ell,k)\big(\tilde{\boldsymbol{C}}(\ell,k)^{\dagger}\tilde{\boldsymbol{C}}(\ell,k)\big)^{-1}\boldsymbol{g} \tag{3.22}$$

and its corresponding output is therefore given by:

$$y_{\mathrm{BF}}(\ell,k) = \tilde{\boldsymbol{w}}^{\dagger}(\ell,k)\boldsymbol{z}(\ell,k) = \sum_{i=1}^{K} s_{i}^{d}(\ell,k)h_{i1}^{d}(\ell,k) + \boldsymbol{g}^{\dagger}\big(\tilde{\boldsymbol{C}}^{\dagger}(\ell,k)\tilde{\boldsymbol{C}}(\ell,k)\big)^{-1}\tilde{\boldsymbol{C}}^{\dagger}(\ell,k)\boldsymbol{v}(\ell,k). \tag{3.23}$$

The modified beamformer output is therefore comprised of the sum of the desired sources as measured at the reference microphone (arbitrarily chosen as microphone #1) and the sensor noise contribution.

3.2 A Residual Noise Cancellation Postfilter

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 estimates thereof instead. The estimation procedure will be discussed in Sec. 3.3. 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. In this section we propose to use a linear postfilter to mitigate the residual interference signals.

Define the BM as a projection matrix to the null subspace of the column-space of $\tilde{\boldsymbol{C}}$:

$$\boldsymbol{B}(\ell,k) = \boldsymbol{I}_{M\times M} - \tilde{\boldsymbol{C}}(\ell,k)\big(\tilde{\boldsymbol{C}}^{\dagger}(\ell,k)\tilde{\boldsymbol{C}}(\ell,k)\big)^{-1}\tilde{\boldsymbol{C}}^{\dagger}(\ell,k). \tag{3.24}$$

Denote the noise reference signals:

$$\boldsymbol{u}(\ell,k) = \begin{bmatrix} u_1(\ell,k) & \dots & u_M(\ell,k) \end{bmatrix}^T \triangleq \boldsymbol{B}(\ell,k)\boldsymbol{z}(\ell,k). \tag{3.25}$$

We propose to use the signals $\boldsymbol{u}(\ell,k)$ to predict the residual noise component at the beamformer output $y_{\text{BF}}(\ell,k)$ by using a set of linear filters, updated according to the following Normalized Least Mean Squares (NLMS) equations [37]:

$$y(\ell,k) = y_{\text{BF}}(\ell,k) - \mathbf{q}^{\dagger}(\ell,k)\mathbf{u}(\ell,k) \tag{3.26}$$

$$\tilde{\mathbf{q}}(\ell+1,k) = \mathbf{q}(\ell,k) + \mu_q \frac{\mathbf{u}(\ell,k)y^*(\ell,k)}{p_{\text{est}}(\ell,k)} \tag{3.27}$$

$$\mathbf{q}(\ell+1,k) \xleftarrow{\text{FIR}} \tilde{\mathbf{q}}(\ell+1,k) \tag{3.28}$$

$$p_{\text{est}}(\ell,k) = \alpha_p p_{\text{est}}(\ell-1,k) + (1-\alpha_p)\|\mathbf{u}(\ell,k)\|^2 \tag{3.29}$$

The operator $\xleftarrow{\text{FIR}}$ consists of the following three stages, which are detailed in [38]. First, $\tilde{q}_m(\ell+1,k)$ is transformed to the time domain. Second, the resulting impulse response is truncated, namely an FIR structure constraint is imposed. Third, the result is transformed back to the frequency domain. Applying the $\xleftarrow{\text{FIR}}$ operator avoids cyclic convolution. When no estimation errors exist, $\boldsymbol{u}(\ell,k)$ are solely determined by the spatially white noise $\boldsymbol{v}(\ell,k)$. Reuven et al. [39] showed that spatially white noise cannot be canceled out by a linear enhancer, and that the resulting filters converge to $\boldsymbol{q}(\ell,k) = 0$. We note, however, that applying nonlinear postfilters such as spectral subtraction [40] or Optimally Modified Log Spectral Amplitude (OMLSA) [41] might be capable of further reducing the residual noise

level. The application of nonlinear noise reduction algorithms is beyond the scope of this paper. The proposed system comprising the LCMV beamformer and the RNC is depicted in Fig. 3.1. When the desired and interference signals leak through the BM, the filters $\mathbf{q}(\ell, k)$ go

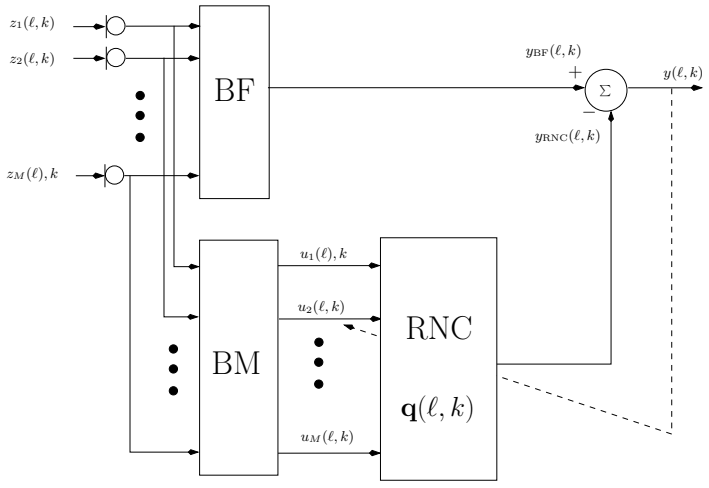


Figure 3.1: The proposed method comprising a LCMV beamformer and a RNC.

into action. Since the Signal to Noise Ratio (SNR) at the beamformer output is much higher than at the BM output, the RNC will tend to further increase the SNR without imposing severe distortion on the desired signals.

Note that although the proposed structure has tight resemblances to the well-known GSC structure, it exhibits a profound difference. While the purpose of the ANC in the GSC structure [22] is to eliminate the stationary noise passing through the BM, in the proposed structure the RNC is only responsible for the residual noise reduction as all signals, including the stationary directional noise signal, are treated by the LCMV beamformer.

Another role of the RNC block is to enhance the robustness of the algorithm to small changes of the ATFs relating the various sources and the microphones. We emphasize, however, that the structure is not capable of tracking major changes in the source position manifested as significant change in the ATFs.

We conclude this section by several remarks regarding the convergence of the NLMS algorithm in the proposed structure. Firstly, as both the desired sources and the interference sources are expected to leak through the BM, in order to avoid mis-convergence of the filters it is necessary to adapt $\mathbf{q}(\ell, k)$ only when the desired sources are inactive. Secondly, we argue that the residual signals leaking through the BM are more stationary than the interference signals themselves. This phenomenon can be verified experimentally. Since the convergence

ability of the NLMS algorithm is known to be affected by the stationarity level of its input signals ($\mathbf{u}(\ell, k)$ in our case), it is expected that the noise reduction of the postfilter will exhibit a more stable behavior.

3.3 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 the 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. 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 assumption, 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 we assume that the information is available.

In the rest of this section we discuss the subspace estimation procedure. This procedure can be regarded in this aspect as a multi-source extension of the single source subspace estimation method proposed by Affes and Grenier [21].

We further assume that the various filters are slowly time-varying filters, i.e $\boldsymbol{H}(\ell, k) \approx \boldsymbol{H}(k)$.

3.3.1 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 $\hat{\boldsymbol{\Phi}}_{zz}(\ell_i, k)$ be a PSD estimate at the noise-only frame ℓ_i. Using the EVD we have $\hat{\boldsymbol{\Phi}}_{zz}(\ell_i, k) = \boldsymbol{E}_i \boldsymbol{\Lambda}_i \boldsymbol{E}_i^{\dagger}$. Noise-only segments consist of both directional 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 thresholds $\mathrm{EV_{TH}}$ and $\mathrm{MEV_{TH}}$. All eigenvectors corresponding to eigenvalues that are more than $\mathrm{EV_{TH}}$ below the largest eigenvalue or not higher than $\mathrm{MEV_{TH}}$ 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. We denote the remaining eigenvectors as $\hat{\boldsymbol{E}}_i$, and their corresponding eigenvalues $\hat{\boldsymbol{\Lambda}}_i$. This procedure is repeated for each segment ℓ_i; $i = 1, 2, \ldots, N_{seg}$. The resulting eigenvectors are collected using the union operator, i.e. eigenvectors that are common to more than one segment are not counted more than once:

$$\hat{\boldsymbol{E}}(k) \triangleq \bigcup_{i=1}^{N_{seg}} \hat{\boldsymbol{E}}_i(k). \tag{3.30}$$

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. This erroneous rank estimation will result in the well-known desired signal cancelation phenomenon in beamformer structures. The union operator can be implemented in many ways. Here we chose to use the Orthogonal Triangular Decomposition (QR).

Consider the following QR:

$$\left[\begin{array}{ccc} \hat{\boldsymbol{E}}_1(k)\hat{\boldsymbol{\Lambda}}_1^{\frac{1}{2}}(k) & \ldots & \hat{\boldsymbol{E}}_{N_{seg}}(k)\hat{\boldsymbol{\Lambda}}_{N_{seg}}^{\frac{1}{2}}(k) \end{array} \right] \boldsymbol{P}(k) = \boldsymbol{Q}(k)\boldsymbol{R}(k) \tag{3.31}$$

where $\boldsymbol{Q}(k)$ is a unitary matrix, $\boldsymbol{R}(k)$ is an upper triangular matrix with decreasing diagonal absolute values and $\boldsymbol{P}(k)$ is a permutation matrix.

All vectors in $\boldsymbol{Q}(k)$ that corresponds to values on the diagonal of $\boldsymbol{R}(k)$ that are lower than $\mathrm{U_{TH}}$ below their largest value, or less then $\mathrm{MU_{TH}}$ above their lowest value are not counted as a basis vector of the directional interference subspace. The collection of all vectors passing the designated thresholds, denoted $\hat{\boldsymbol{E}}(k)$, are used as an estimate of the interference subspace necessary for the implementation of the algorithm.

The novel procedure is advantageous over the more commonly used procedures, since the equivalent interference sources subspace is estimated directly, relaxing the requirement for non-overlapping activity periods of the distinct 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.

Finally, we draw again the reader's attention to a major difference between the proposed method and the GSC structure, i.e. that while the stationary sources are passing through the BM in the GSC structure, in the proposed method they are treated by the LCMV beamformer as any other directional source. One of the reasons for equally treating the stationary and the non-stationary sources is the ability to circumvent the need for the application of the less robust GEVD procedure while estimating interferences subspace $\boldsymbol{E}(k)$.

3.3.2 Desired Sources RTF Estimation

Consider time frames for which only the stationary sources are active and estimate the corresponding PSD matrix, $\hat{\boldsymbol{\Phi}}^{s}_{zz}(\ell,k)$. Assume that there exist 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:

$$\hat{\boldsymbol{\Phi}}^{d,i}_{zz}(\ell_i,k) \approx (\sigma^d_i(\ell_i,k))^2 \boldsymbol{h}^d_i(\ell_i,k)\big(\boldsymbol{h}^d_i(\ell_i,k)\big)^{\dagger} + \hat{\boldsymbol{\Phi}}^{s}_{zz}(\ell,k). \tag{3.32}$$

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:

$$\hat{\boldsymbol{\Phi}}^{d,i}_{zz}(\ell_i,k)\boldsymbol{f}_i(k) = \lambda_i(k)\hat{\boldsymbol{\Phi}}^{s}_{zz}(\ell,k)\boldsymbol{f}_i(k) \tag{3.33}$$

The desired source subspace is spanned by the generalized eigenvectors, corresponding to the generalized eigenvalues with values other than 1, after a rotation by $\hat{\boldsymbol{\Phi}}^{s}_{zz}(\ell,k)$. 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 $\boldsymbol{f}_i(k), \lambda_i(k)$ be the largest eigenvalue and the corresponding eigenvector at segment ℓ_i. Substituting in (3.33) $\hat{\boldsymbol{\Phi}}^{d,i}_{zz}(\ell_i,k)$ with (3.32) yields:

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

Hence we have

$$(\sigma^d_i(\ell_i,k))^2 \boldsymbol{h}^d_i(\ell_i,k)\big(\boldsymbol{h}^d_i(\ell_i,k)\big)^{\dagger}\boldsymbol{f}_i(k) = \big(\lambda_i(k)-1\big)\hat{\boldsymbol{\Phi}}^{s}_{zz}(\ell,k)\boldsymbol{f}_i(k).$$

Now, since by assumption $\lambda_i(k) \neq 1$ we finally have

$$\underbrace{\frac{(\sigma^d_i(\ell_i,k))^2\big(\boldsymbol{h}^d_i(\ell_i,k)\big)^{\dagger}\boldsymbol{f}_i(k)}{\lambda_i(k)-1}}_{scalar}\boldsymbol{h}^d_i(\ell_i,k) = \hat{\boldsymbol{\Phi}}^{s}_{zz}(\ell,k)\boldsymbol{f}_i(k) \quad \blacksquare$$

As we are interested in the RTFs in respect to the first microphone rather than the entire ATFs, the scaling ambiguity can be resolved by the normalization:

$$\hat{\tilde{\boldsymbol{h}}}_i^d(\ell,k) \triangleq \frac{\boldsymbol{\Phi}_{zz}^s(\ell,k)\boldsymbol{f}_i(k)}{\left(\boldsymbol{\Phi}_{zz}^s(\ell,k)\boldsymbol{f}_i(k)\right)_1} \tag{3.34}$$

where $(\cdot)_1$ is the first component of the vector.

3.4 Algorithm Summary

The entire algorithm is summarized in Alg. 2.

Algorithm 2 Summary of the multiple constraint beamforming algorithm.

1) Output signal:
$y(\ell,k) \triangleq y_{\mathrm{BF}}(\ell,k) - \mathbf{q}^{\dagger}(\ell,k)\mathbf{u}(\ell,k)$
2) Beamformer with modified constraints set :
$y_{\mathrm{BF}}(\ell,k) \triangleq \tilde{\boldsymbol{w}}^{\dagger}(\ell,k)\boldsymbol{z}(\ell,k)$
where
$$\tilde{\boldsymbol{w}}(\ell,k) \triangleq \tilde{\boldsymbol{C}}(\ell,k)\big(\tilde{\boldsymbol{C}}(\ell,k)^{\dagger}\tilde{\boldsymbol{C}}(\ell,k)\big)^{-1}\boldsymbol{g}$$
$$\tilde{\boldsymbol{C}}(\ell,k) \triangleq \begin{bmatrix} \tilde{\boldsymbol{H}}^{d}(\ell,k) & \boldsymbol{E}(\ell,k) \end{bmatrix}$$
$$\boldsymbol{g} \triangleq \begin{bmatrix} \underbrace{1 \ \dots \ 1}_{K} & \underbrace{0 \ \dots \ 0}_{N-K} \end{bmatrix}^{T}.$$
$\tilde{\boldsymbol{H}}^{d}(\ell,k)$ 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:
$\tilde{\mathbf{q}}(\ell+1,k) = \mathbf{q}(\ell,k) + \mu_q \frac{\mathbf{u}(\ell,k)y^{*}(\ell,k)}{p_{\mathrm{est}}(\ell,k)}$
$\mathbf{q}(\ell+1,k) \xleftarrow{\mathrm{FIR}} \tilde{\mathbf{q}}(\ell+1,k)$
$p_{\mathrm{est}}(\ell,k) = \alpha_p p_{\mathrm{est}}(\ell-1,k) + (1-\alpha_p)\|\mathbf{u}(\ell,k)\|^2$
5) Estimation:
a) Estimate the stationary noise PSD using Welch method: $\boldsymbol{\Phi}^{s}_{zz}(\ell,k)$
b) Estimate time-invariant desired sources RTFs $\tilde{\boldsymbol{H}}^{d}(k) \triangleq \begin{bmatrix} \tilde{\boldsymbol{h}}^{d}_{1}(k) \dots \tilde{\boldsymbol{h}}^{d}_{K}(k) \end{bmatrix}$
Using GEVD and normalization:
i) $\boldsymbol{\Phi}^{d,i}_{zz}(\ell_i,k)\boldsymbol{f}_i(k) = \lambda_i \boldsymbol{\Phi}^{s}_{zz}(\ell,k)\boldsymbol{f}_i(k)$
ii) $\tilde{\boldsymbol{h}}^{d}_{i}(k) = \frac{1}{f_{i,1}(k)}\boldsymbol{f}_i(k)$.
c) Interferences subspace:
QR factorization of eigen-spaces $\begin{bmatrix} \boldsymbol{E}_1(k)\boldsymbol{\Lambda}_1^{\frac{1}{2}}(k) & \dots & \boldsymbol{E}_{N_{seg}}(k)\boldsymbol{\Lambda}_{N_{seg}}^{\frac{1}{2}}(k) \end{bmatrix}$
Where $\hat{\boldsymbol{\Phi}}_{zz}(\ell_i,k) = \boldsymbol{E}_i(k)\boldsymbol{\Lambda}_i(k)\boldsymbol{E}_i^{\dagger}(k)$ for time segment i

Chapter 4

Experimental Study

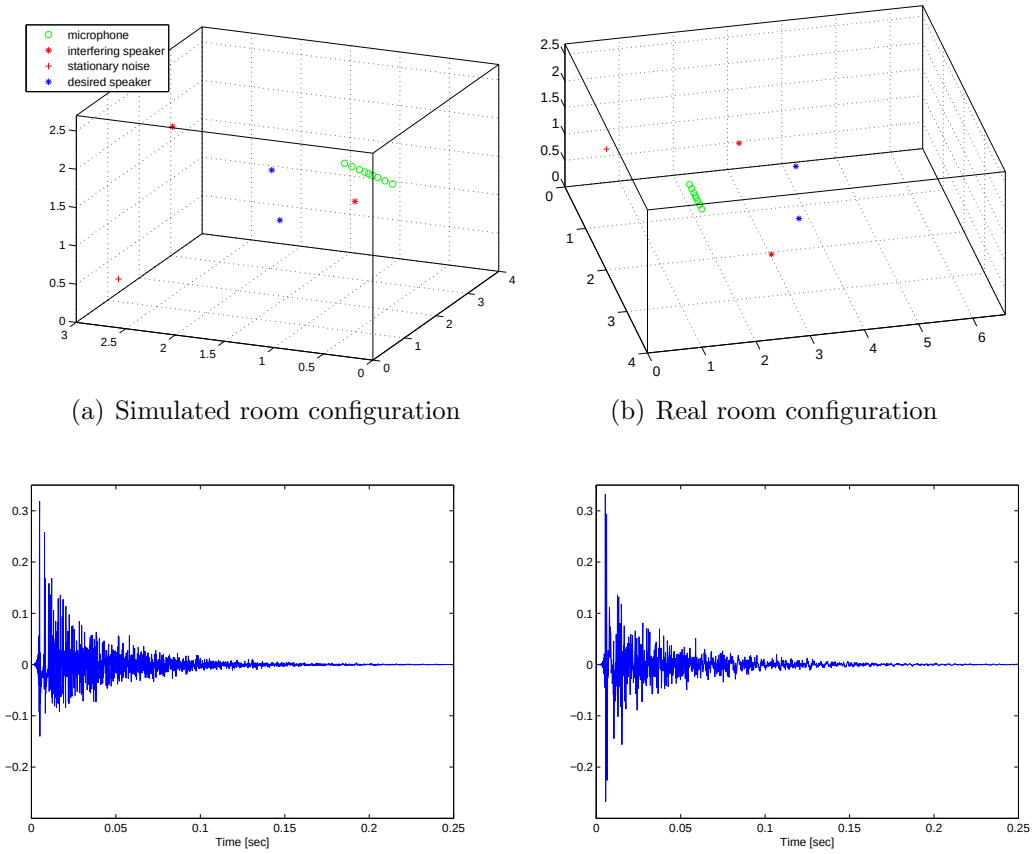
In this chapter the performance of the proposed algorithm is evaluated in different scenarios. In Section 4.1 the test environment and performance criteria are described. In Section 4.2 some implementation considerations are discussed. Two scenarios are then evaluated: the two desired speakers scenario, in both simulated and real rooms is presented in Section 4.3. The performance of the proposed algorithm in the single desired speaker scenario is presented in Section 4.4 and compared with the performance of the TF-GSC algorithm.

4.1 The Test Scenario

The proposed algorithm was tested both in simulated and real room environments. Four sources, two males and two females were drawn from the TIMIT [42] database. A fifth directional signal was the stationary speech-like noise drawn from NOISEX-92 [43] database.

In the simulated room scenario the image method [44] was used to generate the RIR using the simulator in [45]. All the signals were then convolved with the corresponding time-invariant RIRs $\boldsymbol{h}_{i\,m}(k)$ relating each source $i = 1, 2, \ldots, N$ with each microphone $m = 1, 2, \ldots, M$. The microphone signals $z_m(\ell, k)$; $m = 1, 2, \ldots, M$ were finally obtained by summing up the contribution of all directional sources with an additional uncorrelated sensor noise. The reverberation time was set to $T_{60} = 300$mSec. The simulated environment was a $4m \times 3m \times 2.7m$ room. A linear array consisting of 11 microphones was used to perform the beamforming task. The microphone and the various sources positions are depicted in Fig. 4.1(a). A typical RIR relating a source and one of the microphones is depicted in 4.1(c).

The algorithm's performance was also verified using real medium-size conference room equipped with furniture, book shelves, a large meeting table, chairs and other standard items.



(a) Simulated room configuration

(b) Real room configuration

(c) A typical simulated RIR

(d) A typical measured RIR

Figure 4.1: Room configuration and the corresponding typical RIR for simulated and real scenarios.

The room dimensions are $6.6m \times 4m \times 2.7m$. A linear array consists of 8 omni-directional microphones (AKG CK32) was used to pick up the sound signals. The various sources were played separately from point loudspeakers (FOSTEX 6301BX). The signals $\mathbf{z}(\ell, k)$ were finally constructed by summing up all recorded microphone signals with a gain related to the desired input SIR. The source-microphone constellation is depicted in Fig. 4.1(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. 4.1(d).

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 RNC) 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 and the RNC were *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 RNC), 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 $y^d_{\mathrm{BF},i}(\ell, k)$, $y^d_i(\ell, k)$; $i = 1, \ldots, K$, the desired signals components at the beamformer output and the total output (including the RNC), respectively, $y^{ns}_{\mathrm{BF},i}(\ell, k)$, $y^{ns}_i(\ell, k)$; $i = 1, \ldots, N_{ns}$ the corresponding non-stationary interference components, $y^s_{\mathrm{BF},i}(\ell, k)$, $y^s_i(\ell, k)$; $i = 1, \ldots, N_s$ the stationary interference components, and $y^v_{\mathrm{BF}}(\ell, k)$, $y^v(\ell, k)$ the sensor noise component at the beamformer and total output respectively. The entire test procedure is depicted in Fig. 4.2.

The quality measure used for evaluating the performance of the proposed algorithm is the improvement in the SIR. Since, generally, there are several desired sources and interference sources we will use the worst-case SIR for quantifying the performance. The worst-case input SIR relative to the non-stationary signals as measured on microphone m_0 is defined as follows:

$$\mathrm{SIR}^{ns}_{\mathrm{in}}[\mathrm{dB}] = \min_{1 \le i \le K, 1 \le j \le N_{ns}} 10 \log_{10} \frac{\mathrm{var}\left(s^d_i(\ell, k) h^d_{im_0}(\ell, k)\right)}{\mathrm{var}\left(s^{ns}_j(\ell, k) h^{ns}_{jm_0}(\ell, k)\right)}.$$

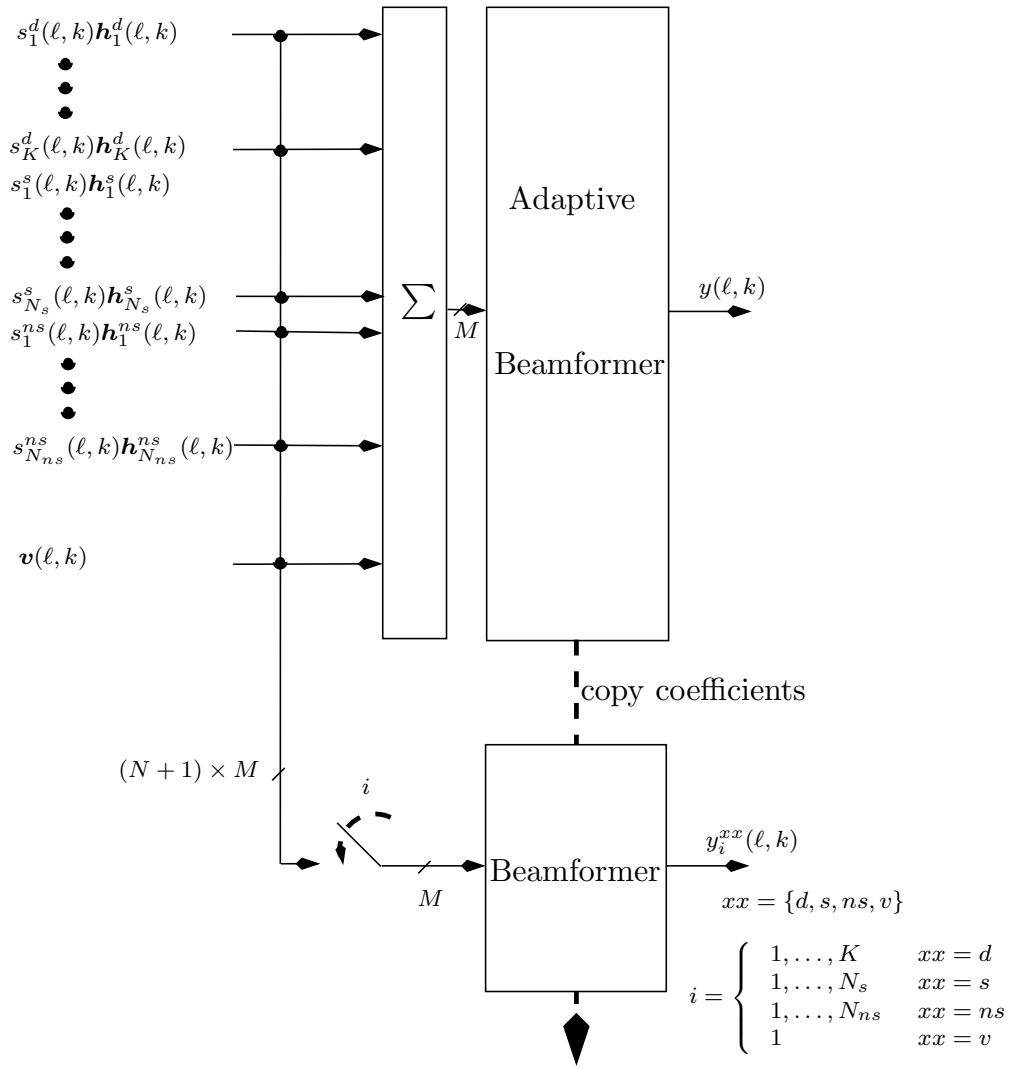


Figure 4.2: Test procedure for evaluating the performance of the algorithm.

Similarly, The worst-case input SIR relative to the stationary signals is:

$$\mathrm{SIR}^{s}_{\mathrm{in}}[\mathrm{dB}] = \min_{1\le i\le K, 1\le j\le N_s} 10\log_{10}\frac{\mathrm{var}\left(s^{d}_{i}(\ell,k)h^{d}_{im_0}(\ell,k)\right)}{\mathrm{var}\left(s^{s}_{j}(\ell,k)h^{s}_{jm_0}(\ell,k)\right)}.$$

These quantities are compared with the corresponding beamformer and total outputs (including the RNC) SIR:

$$\begin{aligned}
\mathrm{SIR}^{ns}_{\mathrm{BF}}[\mathrm{dB}] &= \min_{1\le i\le K, 1\le j\le N_{ns}} 10\log_{10}\frac{\mathrm{var}\left(y^{d}_{\mathrm{BF},i}(\ell,k)\right)}{\mathrm{var}\left(y^{ns}_{\mathrm{BF},j}(\ell,k)\right)}\\
\mathrm{SIR}^{s}_{\mathrm{BF}}[\mathrm{dB}] &= \min_{1\le i\le K, 1\le j\le N_{s}} 10\log_{10}\frac{\mathrm{var}\left(y^{d}_{\mathrm{BF},i}(\ell,k)\right)}{\mathrm{var}\left(y^{s}_{\mathrm{BF},j}(\ell,k)\right)}\\
\mathrm{SIR}^{ns}_{\mathrm{out}}[\mathrm{dB}] &= \min_{1\le i\le K, 1\le j\le N_{ns}} 10\log_{10}\frac{\mathrm{var}\left(y^{d}_{i}(\ell,k)\right)}{\mathrm{var}\left(y^{ns}_{j}(\ell,k)\right)}\\
\mathrm{SIR}^{s}_{\mathrm{out}}[\mathrm{dB}] &= \min_{1\le i\le K, 1\le j\le N_{s}} 10\log_{10}\frac{\mathrm{var}\left(y^{d}_{i}(\ell,k)\right)}{\mathrm{var}\left(y^{s}_{j}(\ell,k)\right)}.
\end{aligned}$$

4.2 Implementation Considerations

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 [38]. The PSD of the stationary interferences and the desired sources is estimated using the Welch method, with an 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 (DFT) 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 an 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 4.1.

4.3 Two Desired Sources Scenario

This scenario included 5 sources: two desired speakers (a male and a female), two competing speakers (a male and a female) and a stationary directional noise. The performance of the proposed algorithm was evaluated in a simulated and in a real room with different SIR values.

Table 4.1: The parameters used by the subspace beamformer algorithm.

Parameter	Description	Value
General Parameters		
f_s	Sampling frequency	8KHz
σ_v^2	Sensor noise variance	1
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%
EV_{TH}	Eigenvectors corresponding to eigenvalues that are more than EV_{TH} lower below the largest eigenvalue are discarded from the signal subspace	40dB
MEV_{TH}	Eigenvectors corresponding to eigenvalues not higher than MEV_{TH} above the sensor noise are discarded from the signal subspace	5dB
U_{TH}	Vectors of $\boldsymbol{Q}(k)$ corresponding to values of $\boldsymbol{R}(k)$ that are more than U_{TH} below the largest value on the diagonal of $\boldsymbol{R}(k)$	40dB
MU_{TH}	Vectors of $\boldsymbol{Q}(k)$ corresponding to values of $\boldsymbol{R}(k)$ not higher than MU_{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 RNC filters	500 taps
RL_l	Noncausal part of the RNC filters	500 taps
RNC Parameters		
μ_0	NLMS adaptation factor	0.18
ρ	Forgetting factor for the estimation of the normalization power $p_{est}(\ell, k)$	0.9

4.3.1 Simulated Environment

The SIR improvement obtained by the beamformer and by the additional RNC in the two desired sources simulated scenario is depicted in Table 4.2. The results in the table were obtained using the second phase of the test procedure described in Sec. 4.1. It is shown that the beamformer can gain approximately 18dB SIR improvement relative to the stationary interference signal and approximately 14dB relative to the non-stationary signals. The RNC further improves the SIR performance by more than 4dB for the stationary sources and only around 2dB for the non-stationary signals. Across all input SIRs the total SIR improvement is impressively high. Assessment of the sonograms in Fig. 4.3 visually verifies the objective

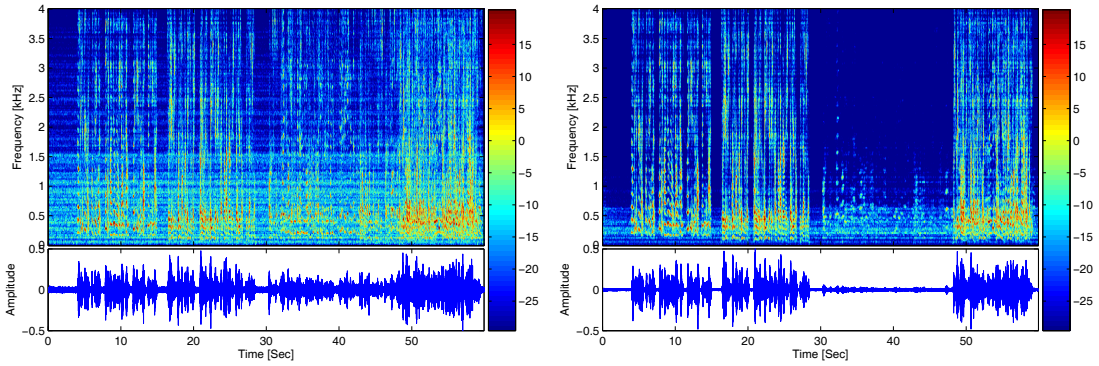
Table 4.2: SIR improvement in dB for the beamformer and the RNC outputs for various input SIR levels in the simulated room two desired sources scenario.

Input	RNC		Beamformer	
$\mathrm{SIR}_{\mathrm{in}}^{ns} = \mathrm{SIR}_{\mathrm{in}}^{s}$	$\mathrm{SIR}_{\mathrm{out}}^{s} - \mathrm{SIR}_{\mathrm{BF}}^{s}$	$\mathrm{SIR}_{\mathrm{out}}^{ns} - \mathrm{SIR}_{\mathrm{BF}}^{ns}$	$\mathrm{SIR}_{\mathrm{BF}}^{s} - \mathrm{SIR}_{\mathrm{in}}^{s}$	$\mathrm{SIR}_{\mathrm{BF}}^{ns} - \mathrm{SIR}_{\mathrm{in}}^{ns}$
−10	3.69	2.24	17.68	12.85
−8	4.73	2.00	17.83	13.70
−6	5.5	2.06	18.18	14.14
−4	5.24	1.89	18.56	14.05
−2	6.02	2.08	18.03	14.05
0	4.49	1.77	18.31	14.49
2	4.61	2.03	18.52	14.64
4	5.06	1.74	18.61	14.54
6	4.31	1.73	18.42	14.27
8	4.67	1.66	17.92	13.68
10	4.39	1.73	18.31	14.29

results presented in Table 4.2 for $\mathrm{SIR}_{\mathrm{in}}^{ns} = 6\mathrm{dB}$. It can be shown that the interference signals are significantly attenuated while the desired sources remain almost undistorted. Finally, we show some of the waveforms comparing the various components at microphone #1 and at the algorithm's output. The results are shown in Fig. 4.4.

4.3.2 Real Environment

In Fig. 4.5 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 RNC's contribution was $1.32dB$ for the competing speakers, and $3.15dB$ for the stationary noise.



(a) Microphone #1 signal (b) Algorithm's output

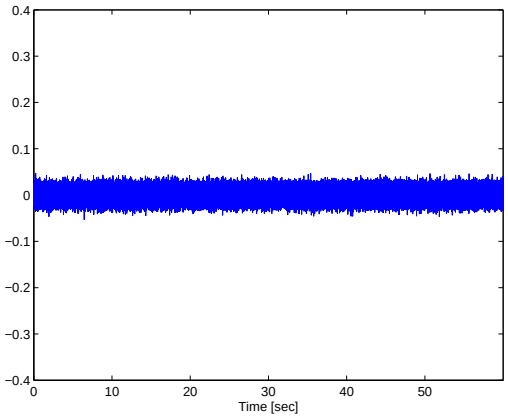
Figure 4.3: Sonograms and waveforms for the simulated room environment for two desired sources and two interfering sources scenario depicting the algorithm's SIR improvement

4.4 One Desired Source Scenario

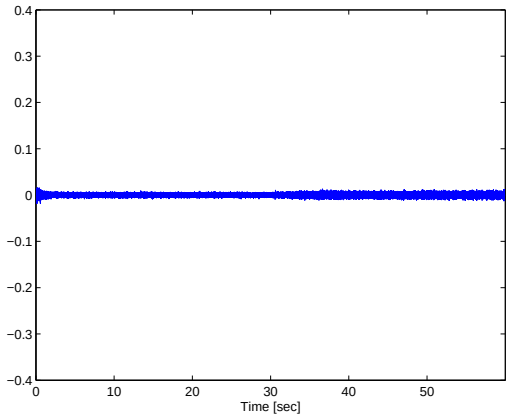
In this section we test the performance of the proposed algorithm in the single desired source scenario.

Two paradigms can be adopted for designing a beamformer for enhancing a desired signal contaminated by both noise and interferences. These paradigms differ in their treatment of the interferences (competing speech signals and/or directional noise). The straight-forward alternative is to apply the single constraint beamformer (The GSC [18], or in reverberant environment the TF-GSC [22]), in which a beam is steered towards the desired signal, while all other interference signals are treated by the ANC. Another alternative suggests steering nulls towards the interference signals. Our contribution adopts the latter approach, in which the FBF steers a beam towards the desired signal while simultaneously directs nulls towards all interferers. The BM blocks both the desired and all interfering sources. Hence, the ANC is only responsible for the residual noise signal mitigation.

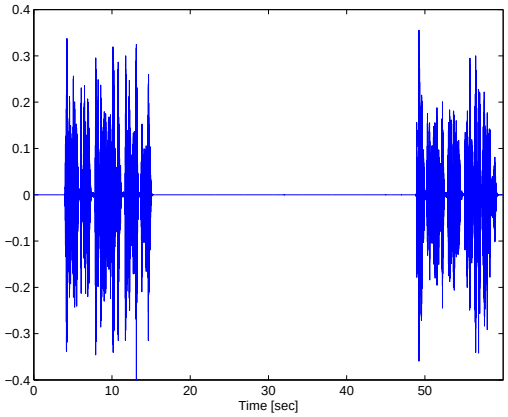
A major difference between the TF-GSC structure and the proposed structure lies in the adaptation mechanism of the ANC. While for the TF-GSC the signals fed to the ANC consist of both non-stationary competing speech signals and stationary noise signal, the input to the corresponding block in the proposed structure is comprised of residual noise, leaking through the BM. The need of the ANC in the TF-GSC to adapt during both types of signals impose contradicting requirements on the adaptation rate. On one hand, the rate should be high enough to allow for tracking of a fast varying signal, and on the other-hand it should be low



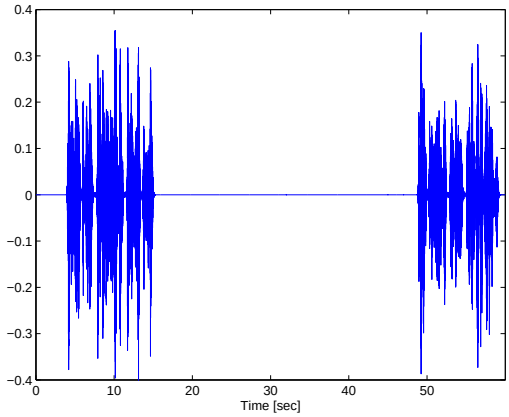
(a) Directional stationary noise at microphone #1



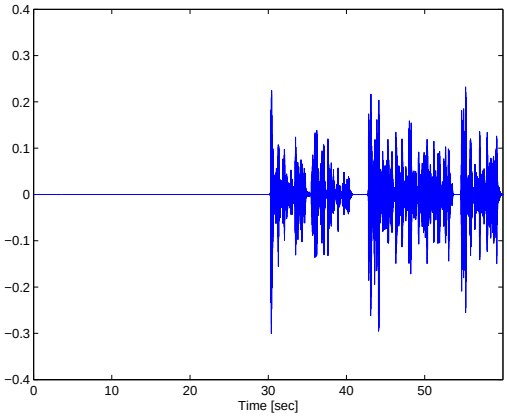
(b) Directional stationary noise at the output



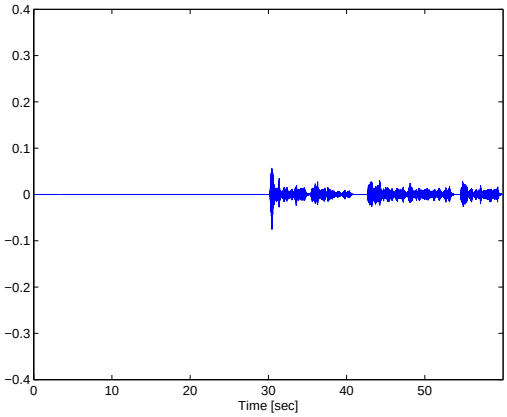
(c) Desired speaker #1 at microphone #1



(d) Desired speaker #1 at the output

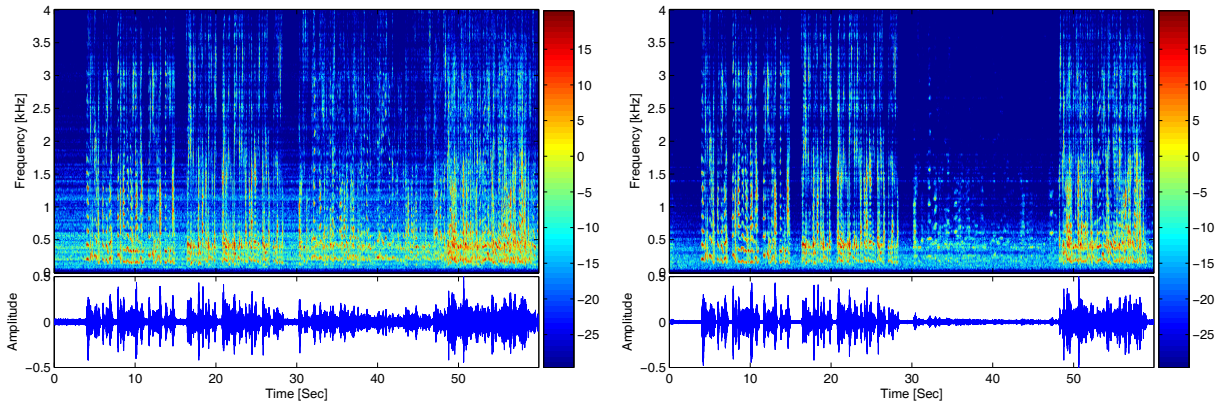


(e) Interfering speaker #2 at microphone #1



(f) Interfering speaker #2 at the output

Figure 4.4: Algorithm performance per component in the two desired sources simulated scenario



(a) Microphone #1 signal

(b) Algorithm's output

Figure 4.5: Sonograms and waveforms for the real room two desired sources scenario depicting the algorithm's SIR improvement.

enough to enable sufficient reduction of the stationary noise level.

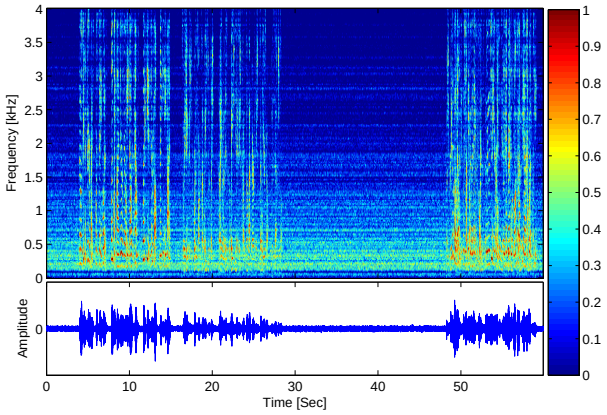
We compared the performance of the two paradigms by using simulated room recordings. We used an array of 11 microphones and a female speaker as the single desired source while various number of competing speakers were used as the interference sources. In all cases single stationary directional noise was used. The performance of the proposed algorithm, in terms of worst case SIR, is presented for the simulated room environment with various SIR values and compared with the performance of the TF-GSC algorithm. The results are depicted in Table 4.3. Examination of the results reveals that the proposed algorithm achieves an excessive 10dB for the stationary noise suppression, and 7dB for the competing speakers suppression with respect to the TF-GSC figures. The SIR improvement was consistent with the number of competing speakers.

It is clear that in static scenarios, well-designed nulls towards all interfering signals yields much better undesired signal reduction than adaptive cancelers.

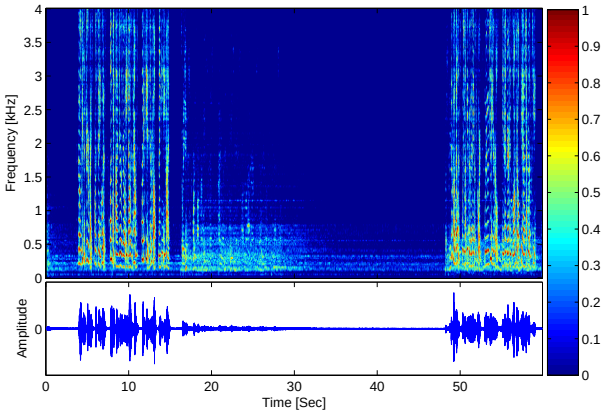
In Fig. 4.7 the stationary noise component at the beamformer output, $y_1^s(n)$, is compared. It is evident that the noise level at the proposed algorithm is much lower than the corresponding noise level at the TF-GSC output. Moreover, the noise signal at the TF-GSC output exhibits severe level fluctuations, while the noise level at the proposed method output is much more stable. This phenomenon can be attributed to contradicting adaptation demands during competing speakers activity.

On the other hand, the proposed algorithm, which directs a constant null towards all interference sources, does not have to adapt, and hence can maintain a stationary noise component at the output. This property ensures proper use of any further post-filter.

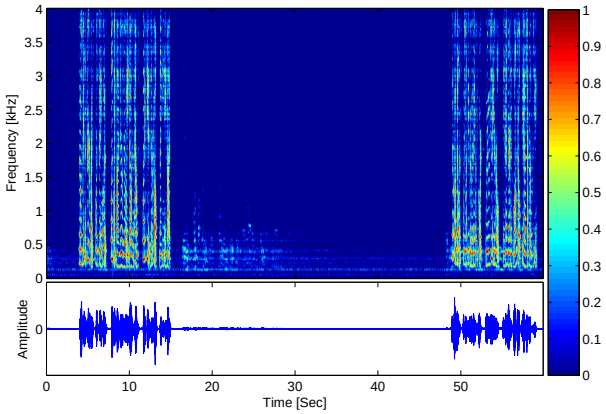
The TF-GSC output is depicted in Fig. 4.6(b) while the output of the proposed algorithm is shown in Fig. 4.6(c). It is evident that the latter outperforms the TF-GSC especially in terms of the competing speaker cancellation, and that it is closer to the microphone signal as shown in Fig. 4.6(a).



(a) Microphone #1 signal



(b) TF-GSC output

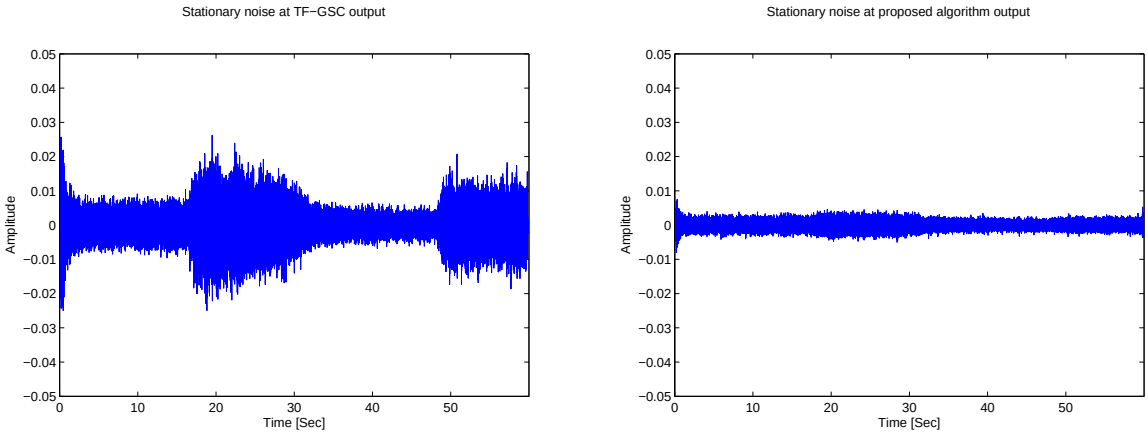


(c) The proposed algorithm output

Figure 4.6: TF-GSC and proposed algorithm comparison - Sonograms and waveforms

Table 4.3: Single desired source SIR improvement (in dB) for the TF-GSC and the proposed algorithm for various numbers of competing speakers in the simulated room scenario.

	TF-GSC		Proposed algorithm	
Competing speakers	$\mathrm{SIR}^{s}_{\mathrm{out}} - \mathrm{SIR}^{s}_{\mathrm{BF}}$	$\mathrm{SIR}^{ns}_{\mathrm{out}} - \mathrm{SIR}^{ns}_{\mathrm{BF}}$	$\mathrm{SIR}^{s}_{\mathrm{BF}} - \mathrm{SIR}^{s}_{\mathrm{in}}$	$\mathrm{SIR}^{ns}_{\mathrm{BF}} - \mathrm{SIR}^{ns}_{\mathrm{in}}$
2	$10.88dB$	–	$19.39dB$	–
3	$8.79dB$	$6.61dB$	$20.46dB$	$16.47dB$
4	$9.96dB$	$7.69dB$	$20.29dB$	$16.32dB$
5	$9.54dB$	$7.83dB$	$14.25dB$	$13.01dB$



(a) TF-GSC (b) Proposed algorithm

Figure 4.7: Noise signals compared at the output of the algorithms

4.5 Conclusion

In this chapter we presented an experimental study of the proposed algorithm. It was shown that the proposed algorithm obtained satisfying results in both simulated and real room environments. Comparison of the SIR improvement of the TF-GSC and the proposed algorithm, emphasize the advantages of the proposed when the acoustical environment is assumed to be time-invariant.

Chapter 5

Summary and Future Directions

In this chapter we summarize our work and propose some future research directions.

5.1 Research Summary

In this thesis we addressed the problem of extracting several desired sources in a reverberant environment contaminated by both nonstationary (competing speakers) and stationary interferences utilizing a microphone array.

A literature survey depict the limitations of exiting methods in solving the general problem of extracting multiple desired sources contaminated by several competing speakers and directional noise sources. These limitations are emphasized in highly reverberant environment.

We propose a novel method for array beam-pattern design based on the LCMV criterion. The array is designed to satisfy a set of linear constraints, that maintain multiple desired sources while cancelling out the interference sources. A practical procedure for estimating the constraints set is then derived.

Due to erroneous estimate of the constraint matrix residual interference signal leaks to the beamformer's output. We propose to construct a BM which outputs refer to the residual interference signal, enabling further enhancement of the desired sources by using an adaptive signal canceler, denoted RNC.

Unlike common GSC structures, we chose to block all directional signals, including the stationary noise signals, in the beamformer. By treating the stationary source as a directional signal we obtain more stable nulls, which do not suffer from fluctuations caused by the adaptive process. However, in time-varying environment different, more adaptive forms, might be adopted.

Experimental results for both simulated and real environments demonstrate that the proposed method can be applied to extract several desired sources from a combination of multiple sources in a complicated acoustic environment, and outperform existing algorithms.

5.2 Future Directions

The estimation method in the proposed algorithm is not applicable in case of time-varying acoustic environment. In the future we propose to add tracking ability to the RTF and the interferences subspace estimation procedures.

Another limitation of the proposed algorithm is the requirement for a single talk segment for each of the desired sources. In the future it is important to add a method that can also be applied in double-talk scenarios. It is also interesting to investigate the ability of updating the interferences subspace while the desired sources are active. Relaxing these limitations will make the algorithm more robust, easier to manage and applicable to more situations.

Bibliography

[1] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," vol. 8, no. 5, pp. 497–507, Sep. 2000.

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

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

[4] 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.

[5] 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, June 1994.

[6] 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.

[7] S. Amari, A. Chichocki, and H. H. Yang, "Blind signal separation and extraction: Neural and information-theoretic approaches," *Unsupervised Adaptive Filtering*, vol. 1, 2000.

[8] H. Wu and J.C. 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.

[9] 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, June 2000.

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

[11] 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.

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

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

[14] 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, Sept. 2002.

[15] 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, Dec. 2004.

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

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

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

[19] 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.

[20] 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, June 2002.

[21] 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.

[22] 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.

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

[24] Y. Hu and P.C. 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, July 2003.

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

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

[27] 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.

[28] 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.

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

[30] 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, June 1996.

[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 canceller," 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] K. U. Simmer, J. Bitzer, and C. Marro, *Microphone Arrays: Signal Processing Techniques and Applications*, chapter Post-Filtering Techniques, pp. 39–60, Springer-Verlag, berlin, germany edition, 2001.

[36] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 6, pp. 561–571, Nov. 2004.

[37] B. Widrow, J.R. Glover, J.M. McCool, J. Kaunitz, C.S. Williams, R.H. Hearn, J.R. Zeidler, E. Dong, and R.C. Goodlin, "Adaptive noise cancelling: Principals and applications," *Proc. IEEE*, vol. 63, no. 12, pp. 1692–1716, Dec. 1975.

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

[39] G. Reuven, S. Gannot, and I. Cohen, "Performance analysis of the dual source transfer-function generalized sidelobe canceller," *Speech Communication*, vol. 49, no. 7-8, pp. 602–622, July 2007.

[40] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[41] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Lett.*, vol. 9, no. 4, pp. 113–116, Apr. 2002.

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

[43] 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, July 1993.

[44] J.B. Allen and D.A. 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.

[45] E.A.P Habets, "Room impulse response (RIR) generator," `http://home.tiscali.nl/ehabets/rir_generator.html`, July 2006.

תקציר

במקרים רבים אנו מעוניינים למצות מספר דוברים רצויים, אשר מופרעים ע"י מספר מקורות לא סטציונרים (דוברים נוספים, או מקורות רעש שונים), כמו גם רעשים סטציונרים. בנוסף, האותות הנקלטים לרוב מעוותים ע"י תגובה האקוסטית (המהדהדת) של הסביבה. אנו מניחים שהתגובה להלם בסביבה משתנה לאיטה ביחס לשינויים המהירים באותות.
דוגמאות טיפוסיות לבעיה הנ"ל הן שיחות וועידה מרובות משתתפים הנערכות בחדרי דיונים; שיחה באמצעות דיבורית ברכב כאשר הנוסעים האחרים מדברים ומפריעים; ותרחיש *מסיבת הקוקטייל*, אשר בו מעוניינים להאזין לשיחה אחת מני רבות המתנהלות במהלך מסיבה.
פיתרונות פרקטיים לתרחישים הללו דורשים שימוש במערך של מיקרופונים. את רב האלגוריתמים שמנסים לטפל בבעיה ניתן לסווג לאחת מבין שתי המשפחות הבאות (או לשילוב שלהן), הנקראות: הפרדת מקורות עיוורת (Blind Source Separation, BSS), משפחה זו של אלגוריתמים מפרידה את המקורות הקיימים בבעיה ע"י אילוץ חוסר תלות סטטיסטית (Independent Component Analysis, ICA) בין האותות המופרדים, והמשפחה השנייה, משפחת מעצבי האלומות, אשר מנצלים את ההתפשטות המרחבית של האותות כדי לקלוט טוב יותר את האותות הרצויים ולהקטין את השפעת האותות המפריעים
האלגוריתם שאנו מציגים בעבודה הזו משתייך למשפחה השנייה, משפחת מעצבי האלומות.
בדומה, לשיטות אחרות במשפחה אנו מציעים לבנות את מעצב האלומה בכפוף למערכת אילוצים ליניארית. לעומת רב השיטות, אשר נוהגות להגדיר אילוצים רק לגבי האותות הרצויים, ומטפלות בשאר המפריעים באמצעות מסננים מסתגלים אשר שואפים למזער את התרומה של האותות המפריעים במוצא מעצב האלומה, אנו מציעים להגדיר אילוץ לכל מקור מפריע. המסננים המסתגלים מטפלים באופן טוב מאוד ברעשים הסטציונריים, אולם לא מספיק טוב במפריעים הלא סטציונריים, אשר משתנים בקצב מהיר מקצב ההתכנסות של המסננים המסתגלים (אם ננסה להגדיל את קצב ההתכנסות של המסנן, נקבל שונות שגיאה גבוהה יותר).
בדומה, לשיטות אחרות, כדי להגדיר את האילוצים עבור הדוברים הרצויים, אנו משתמשים בתגובה להלם היחסית של הסביבה (ביחס למיקרופון ייחוס כלשהו), ולא בתגובה להלם עצמה אשר קשה לשערוך. כדי לשערך את התגובה להלם היחסית של דובר רצוי מסויים, אנו זקוקים לקטע זמן אשר בו רק הוא פעיל, וכמובן הרעשים הסטציונריים אשר פעילים בקביעות. אנו משתמשים בפירוק לוקטורים עצמיים, וערכי עצמיים מוכללים, של מטריצת הקוואריאנס של המיקרופונים המולבנת ע"י מטריצת הקוואריינס של הרעשים הסטציונריים במוצא המיקרופונים. הווקטור העצמי, בעל הערך העצמי החזק ביותר בתהליך פירוק זה משמש לשיערוך התגובה להלם היחסית של הדובר הרצוי. אנו חוזרים על תהליך זה לגבי כל אחד מהדוברים הרצויים.
את האילוצים לגבי האותות המפריעים אנו משיגים בצורה שונה וחדשנית, בכל קטע זמן שבו לא פעיל אף דובר רצוי אנו משערכים את מרחב האותות הפעילים (המפריעים). מרחב האותות משוערך ע"י בחירת הווקטורים העצמיים המתאימים לערכים העצמיים הדומיננטיים בפירוק של מטריצת הקוואריינס של המיקרופונים באותו קטע. מכיוון שבכל קטע וקטע איננו מאלצים שכל המפריעים יופיעו (אילוץ קשה ולא סביר), אנו מאחדים את כל מרחבי המפריעים למרחב הפרעות כולל ומגדירים אילוץ לכל וקטור בבסיס של מרחב זה.
על מנת לייצב את הפיתרון, ולהפוך אותו לפחות רגיש לשינויים קלים בסביבה, ולשגיאות שערוך אנו מוסיפים את מבטל הרעשים השיוריים (Residual Noise Canceller, RNC). מנגנון זה הפועל על הרעש השיורי, הרכיב של אותות המיקרופונים השייך למרחב המשלים למרחב האותות (המרחב הנפרש ע"י האותות הרצויים והמפריעים), שואף לחסר את החלק הקורלטיבי לרעש השיורי ממוצא מעצב האלומה.
לסיכום, אנו מעריכים את ביצועי האלגוריתם החדש הן בסביבת סימולציה, בה אנו משתמשים במודל של תגובה להלם אקראית בחדר, והן בסביבת אמיתית של חדר דיונים. את ביצועי האלגוריתם אנו בודקים תחת שינוי מספר פרמטרים: מספר הדוברים הרצויים, מספר הדוברים המפריעים, יחס האות לרעש בין האותות הרצויים למפריעים ומספר המיקרופונים. עבור המקרה של דובר רצוי אחד אנו משווים את ביצועי האלגוריתם המוצע לאלו של ה-TFGSC.

עבודת הגמר נעשתה בהנחיית ד"ר שרון גנות מבית הספר להנדסה באוניברסיטת בר אילן ופרופ' ישראל כהן מהפקולטה להנדסת חשמל בטכניון.

תודות

אני מעוניין להביע את תודתי העמוקה והערכתי למנחים ד"ר שרון גנות ופרופ' ישראל כהן על הנחייתם המסורה. תודה על עזרתכם המקצועית, על עידודכם למצוינות ועל הרבה עצות מועילות לאורך כל שלבי המחקר. בנוסף מבקש להודות לד"ר עמנואל הבטס על הרבה דיונים פוריים.

תודה מיוחדת לאהובתי לירן אשר עודדה אותי ותמכה בי לאורך כל הדרך, לאימי אורית, לאבי המנוח יעקב ז"ל ולאלי ואורלי גולן.

אני מודה לטכניון על התמיכה הכספית הנדיבה בהשתלמותי.

מעצב אלומות רב ערוצי מבוסס מרחבים עצמיים בסביבה מהדהדת עם מספר מקורות דיבור מפריעים

חיבור על עבודת גמר

לשם מילוי חלקי של הדרישות לקבלת התואר
מגיסטר למדעים בהנדסת חשמל

שמוליק מרקוביץ'

הוגש לסנט הטכניון - מכון טכנולוגי לישראל

אב תשס"ח חיפה אוגוסט 2008

מעצב אלומות רב ערוצי
מבוסס מרחבים עצמיים בסביבה מהדהדת
עם מספר מקורות דיבור מפריעים

שמוליק מרקוביץ'