Dereverberation and Noise Suppression of Multichannel Speech Signals

Eliav Benisty

Dereverberation and Noise Suppression of Multichannel Speech Signals

Research Thesis

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

Eliav Benisty

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

Kislev 5768 Haifa August 2014

Acknowledgment

The Research Thesis Was Done Under The Supervision of Professor Israel Cohen from the Department of Electrical Engineering at the Technion and Professor Jacob Benesty of the University of Quebec.

I would like to thank Professor Cohen for his dedicated and professional support, and his encouragement to creativity and innovation. I would also like to express my sincere gratitude to professor Jacob Benesty for his valuable advices, which contributed much to this research.

Special thanks to my beloved family, Chaim, Varda, Liraz, Dorin and Gali Benisty. For their continues encouragement and support.

Contents

List of Figures

List of Tables

Abstract

Speech signals recorded in a real environment are often contaminated by non-stationary interfering signals (other speakers), as well as by stationary noise. Moreover, distortion imposed by the room impulse response (RIR) of the acoustic environment may severely degrade the fidelity and intelligibility of the speech signal. Extraction of desired speech signals recorded in such environments is required in many practical applications: hands-free communication systems, cellular phones, teleconferencing, hearing aids, and man-machine interfaces among others.

In this thesis, beamforming techniques are utilized to address the problem of joint dereverberation and noise reduction in the short-time Fourier transform (STFT) domain. The term beamforming refers to the design of a spatio-temporal filter used to process signals received by a microphone array in order to extract a desired sound. Obviously such a spatial filter can be designed to optimize different cost functions, depending on the application's demands and constraints. The main goal of this study is to develop and analyze beamformers specifically designed to extract clean, early reflections from noisy, reverberating observations.

The study is divided into two parts. In the first part we show that beamformers originally developed for the purpose of noise reduction can be adjusted to resolve the problem of joint dereverberation and noise reduction. The analysis is based on the estimation of the full to early relative transfer function (RTF), defined as the ratio between the acoustical transfer function (ATF) of the early speech reflections, and the ATFs of the array's microphones. Using this estimate we express the beamformer observations vector as a function of the desired signal, from this perspective performance measures are defined and existing beamformers are deduced. A comprehensive experimental study, consisting of simulated environments proves the applicability of the proposed algorithm to the task of joint dereverberation and noise reduction (JDNR). Furthermore, the proposed algorithm outperforms the transfer function generalized sidelobe canceler (TF-GSC) algorithm in extracting the clean, early reflections.

In the second part we introduce a different outlook on the design of beamformers in the STFT domain. Specifically we use the full to early RTF, to divide the beamformer

observation vector into three terms: the early reflection vector, the unwanted coherent interference vector (which includes the late reverberations), and the uncorrelated noise vector. This decomposition introduces new optimization problems, followed by beamformers with potentially high reverberation suppression capabilities. An experimental study consisting of simulated reverberant environments demonstrates the dereverberation and denoising capabilities of the suggested beamformers, compared to the minimum variance distortionless response (MVDR) beamformer.

Notation

A matrix

A^{-1} matrix inverse

A_{ij} the (i, j) element of the matrix A

$A(:, j)$ the jth column of the matrix A

$\text{diag}\{x\}$ diagonal matrix with the vector x on its diagonal

$E[\cdot]$ expectation operation

$I_{\text{N}\times\text{N}}$ N $\times$ N identity matrix

$Re\{x\}$ the real part of x

x scalar

$\bar{x}$ column vector

$\bar{x}_j$the jth element of the vector x

$X[l, m]$ time frequency coefficient

$\Phi_{xx}[l, m]$ correlation matrix of the time frequency coefficient vector $\bar{x}[l, m]$

$(\cdot)^*$conjugate operation

$\phi_x[l, m]$ power spectral density of the time frequency scalar $x[l, m]$

$\|\cdot\|$ Euclidian norm operation

$*$ linear convolution operation

$(\cdot)^T$ transpose operation

$(\cdot)^H$ transpose-conjugate operation

Abbreviations

ANC Adaptive Noise Canceler

BF Beamformer

BM Blocking Matrix

DS Delay and Sum

DRR Direct to Reverberant Ratio

DTF-GSC Dual Transfer Function Generalized Sidelobe Canceler

D-GSC Delay Generalized Sidelobe Canceler

FBF Fixed Beamformer

FIR Finite Impulse Response

GEVD Generalized Eigenvalue Decomposition

GSVD Generalized Singular Value Decomposition

GSC Generalized Sidelobe Canceler

JDNR Joint Dereverberation and Noise Reduction

LTI Linear Time Invariant

LCMV Linearly Constrained Minimum Variance

LMS Least Mean Squares

MMSE Minimum Mean Squared Error

MSNR Maximum Signal to Noise Ratio

MTF Multiplicative Transfer Function

MV Minimum Variance

MVDR Minimum Variance Distortionless Response

MWF Multichannel Wiener Filter

NR Noise Reduction

NLMS Normalized Least Mean Square

NLCMV Normalized Linearly Constrained Minimum Variance

OMLSA Optimally Modified Log Spectral Amplitude

oSNR Output Signal To Noise Ratio

iSNR Input Signal To Noise Ratio

iSCNR Input Signal To Coherent Noise Ratio

PSD Power Spectrum Density

RTF Relative Transfer Function

RIR Room Impulse Response

RNC Residual Noise Canceler

RBW Reverberation Block Wiener Beamformer

RBT Reverberation Block tradeoff Beamformer

SDW-MWF Speech Distortion Weighted Multichannel Wiener Filter

SIR Signal to Interference Ratio

SNR Signal to Noise Ratio

SRA Statistical Room Acoustics

STFT Short Time Fourier Transform

TF-GSC Transfer Function Generalized Sidelobe Canceler

VAD Voice Activity Detector

Chapter 1

Introduction

A sound wave recorded within an enclosed environment is a superposition of many delayed and attenuated copies of itself reflected from the enclosure's walls and the objects within it. These hundreds of reflections produce an effect known as reverberation. The reverberant signal consists of a direct path and multiple reflections that arrive at the microphone with different phases and gains. These delayed reflections are perceived as noise that degrades the fidelity and intelligibility of the speech signal. The degradation increases with the distance between the desired source and the microphones, and becomes even more severe as sound waves from other sources reverberate in the room.

The reverberant signal can be modeled as the output of an LTI system, in which the source signal is convolved with a RIR. The RIR describes the response of the room to a delta impulse located at the source position. It can be divided into three segments: the direct path, the early reflections, and the late reflections. The convolved product of these parts with the desired signal results in the direct sound, its early reflections, and its late reverberation, respectively. Looking at a typical RIR, the early reflections appear among its first taps as a small group of delayed impulses. The late reverberations assemble most of the RIR and can be characterized as a decaying exponential. The decay rate of the late reverberations has a crucial effect on the amount of distortion they impose. The time it takes the sound energy to reach a $60\,dB$ decay after switching off a sound source is denoted by T_{60}. This measure is used to quantify the severity of reverberation within an enclosure. It is affected by the volume of the enclosed space and the acoustic properties of the reflecting surfaces within it [20, chapter 2].

The main goal of dereverberation algorithms is to extract the clean, direct path and early reflections from the noisy reverberant signal. The early reflections arrive at the microphone a short time after the direct path signal, and tend to improve its intelligibility. Since these desired reflections usually occupy the same frequency bands as the late reverberations, the task of dereverberation becomes almost impossible when using only the temporal data.

Array processing and beamforming algorithms make use of both spectral and spatial information to extract sounds related to a desired source, making them optimal for addressing the JDNR problem. Unfortunately most of the existing beamforming algorithms focus on the problem of noise reduction and only partly suppress the late reverberations. A review on some of these algorithms is given in the next chapter.

1.1 Beamforming

Over the last four decades, numerous spatio-temporal filters have been developed; many of them designated for wireless communication systems. More recently beamforming algorithms were adapted by many speech enhancement applications. Most existing beamformers are designed to statistically optimize a certain quality measure, e.g., maximum output SNR, minimum mean-squared error, or minimum noise variance. Some do so while allowing the desired signal to be distorted; others constrain the optimization process to ensure the preservation of the desired signal. An unconstrained minimization of the MSE in the context of beamforming leads to the well-known multichannel Wiener filter (MWF) [1]. This beamformer tries to maximize the array output-SNR while allowing distortion of the desired signal to occur. Efficient implementation of the Weiner filter based on the generalized singular value decomposition (GSVD) was first introduced by Doclo and Moonen [2]. A narrowband implementation of the GSVD scheme was suggested by Spriet [3].

Beamformers in which additional linear constraints are imposed are known as linearly constrained minimum variance (LCMV) beamformers [4], of which the minimum variance distortionless response (MVDR) beamformer (also known as Capon's beamformer) [5] is a special case. In [6], Frost proposes an adaptive scheme for the MVDR beamformer capable of satisfying some desired frequency response in the look direction, while minimizing

output noise power. Frost suggests the use of a constrained LMS-type algorithm to minimize the total output power. Kaneda et al. [7] propose a noise reduction system for speech signals, termed AMNOR, which adopts a soft-constraint that controls tradeoff between speech distortion and noise reduction.

To avoid the constrained adaptation of the MVDR beamformer, Griffiths and Jim [8] propose a generalized sidelobe canceler (GSC) structure, which separates the output power minimization and the application of the constraint. The GSC structure is comprised of three blocks: the first being a fixed beamformer designed to satisfy the constraint; the second being a blocking matrix that blocks the desired signal and produces noise-only reference signals; and the third being an unconstrained adaptive algorithm (e.g., the LMS algorithm) that aims at canceling the residual noise at the fixed beamformer output, given the noise-only reference signals at the output of the blocking matrix. The main drawback of the GSC beamformer is that a single direction of arrival for the signal cannot be determined in reverberant environments, since reflections from different directions are also captured by the sensor array. If signals that are correlated with the desired signal leak into the noise reference signals the noise canceler filters will subtract speech components from the FSB output, causing self-cancellation of the desired speech, and hence a severe distortion. Even when the ANC filters are estimated or adapted during noise-only periods, the distortion is unavoidable.

Hoshuyama et al. [39] used a three-block structure similar to the GSC. However, the blocking matrix has been modified to operate adaptively. To limit the leakage of the desired signal, a quadratic constraint is imposed on the norm of the noise canceler coefficients. Alternatively, use of the leaky LMS algorithm has been suggested. Nordholm et al. [40] used a GSC solution in which the blocking matrix is realized by spatial highpass filtering, thus yielding improved noise-only reference signals. Meyer and Sydow [41] have suggested to construct the noise reference signals by steering the lobes of a multibeam beamformer toward the noise and desired signal directions separately. In [43] Doclo et al. suggested taking the speech distortion due to speech leakage directly into account in the design criterion of the multichannel ANC, introducing the speech distortion weighted multichannel Wiener filtering (SDW-MWF).

The GSC structure was re-derived in the frequency domain, and extended to deal

with general acoustic transfer functions (ATFs) as suggested by Affes and Grenier [9]. In practice, the desired source ATFs are unknown and difficult to acquire. Thus, sub-optimal solutions aimed at noise reduction rather than estimation of the desired source signal were developed. These solutions use estimates of the relative transfer functions (RTFs), which represent the coupling between pairs of sensors with respect to the desired source. A major contribution was made by Gannot et al. [10], they proposed a transfer function GSC (TF-GSC) which exploits the RTFs in the time–frequency domain. They showed that knowledge of the ratios of transfer functions from the source to the individual sensors is sufficient to block the desired signal. To estimate the RTFs they proposed using a least squares approach exploiting the non-stationarity of speech signals as opposed to the stationary character of the noise. Cohen extended the method using a weighted least squares approach which incorporates an indicator function for the presence of the desired signal [15]. Talmon et al. presented a GSC framework in the STFT domain using a complete representation of a linear convolution [44], and proposed a new practical algorithm relying on the convolutive transfer function approximation (CTF-GSC). In [45], the authors derive a procedure for obtaining the RTFs from a generalized eigenvector decomposition and develop an algorithm to compute them from multichannel input signals.

Some solutions suggest following a beamformer type algorithm by a post-processor. Zelinski [37] suggested a Wiener filter, followed by further noise reduction in a post processing configuration. Meyer and Simmer [38] combined Wiener filtering in the high-frequency band with spectral subtraction in the low-frequency band. Cohen et al. [34], proposed a multi-microphone post-processor for the TF-GSC . The proposed post-processor is designed to enhance the noise suppression, and does not reduce reverberation.

In [11], Bitzer et al. investigate the theoretical performance limits of the GSC in the case of diffuse noise. The tradeoff between noise reduction and speech distortion in the parametrized multichannel Wiener filtering was established in [12]. In [13] the authors analyzed the performance of the MVDR beamformer in different noise fields (viz., coherent, incoherent and diffused noise fields) and studied the effectiveness of the MVDR beamformer in reverberant environments. Their study showed that the amount of noise reduction sacrificed when complete dereverberation is required depends on the

DRR of the beamformers source to reference ATFs. Another notable effort to understand the functioning of the TF-GSC beamformer was published by Gannot and Cohen in [31]. They found that it is theoretically possible to achieve infinite noise reduction when only a coherent noise source is added to the desired source. In the presence of ambient noise, the latter can also be achieved by the LCMV beamformer.

In [14], Benesty and Chen, introduce a new formalism by which a beamformer is designed in the STFT domain. They propose writing the beamformer's observations vector as a function of the desired source, using the RTFs between each channel to the reference channel. Using these relations they introduce important quality measures, and re-derive known beamformers in their closed-form formulation. They also develop a tradeoff beamformer which allows a tradeoff between noise reduction and speech distortion. In [33], Habets and Benesty extend this framework by dividing the observed noise signals into spatially coherent and incoherent additive noise components. They introduce a general tradeoff filter that enables a compromise between noise reduction and speech distortion on the one hand, and coherent noise versus incoherent on the other hand.

The beamformers proposed in [10, 14, 45], aim at estimating the reverberant signal received by one of the microphones and thereby alleviate the GSC leakage problem. Because there is a tradeoff between noise reduction and (signal-channel) dereverberation as shown in [13], these beamformers also maximize the amount of noise reduction. An obvious disadvantage is that the reverberation of the received speech signal is not reduced.

In this contribution we propose an estimate for the full to early RTFs, defined as the ratio between the ATF of the early speech reflections at the reference channel, and the ATFs of the array's microphones. We then incorporate this estimate to the framework proposed in [14], introducing optimal beamformers for the JDNR problem. To further increase the beamformers reverberation suppression capabilities we suggest a new signal model for the multichannel observation vector. Using this model we derive new optimization problems, in which reverberation suppression, noise reduction, and speech distortion, are treated independently. Solutions of these optimization problem introduce beamformers with high dereverberation and noise reduction capabilities.

1.2 Thesis Structure

Chapter 2 introduces the problem of joint dereverberation and noise reduction. It explores two algorithms that deal with a subset of the general problem, discussing their drawbacks and the motivation for new research. Chapter 3 presents three RTF based optimal beamformers, designed for extracting the clean early signal, recorded in a noisy reverberant environment. The estimation processes of the full to early RTF and other components needed for the beamformers' construction are described in detail, followed by a vast experimental study demonstrating the beamformers' capabilities and comparing them to those of the TF-GSC beamformer. In Chapter 4 the microphone array signal is expressed as a sum of 3 terms: the early reflection vector, the unwanted coherent interference vector, and the uncorrelated noise vector. This decomposition introduces new constrained optimization problems, in which the suppression of reverberation is addressed directly. Close-form solutions of these optimization problems, introduce new beamformers with high dereverberation and noise reduction capabilities. We summarize this chapter with an experimental study demonstrating the proposed beamfomers capabilities. Chapter 5 concludes with proposals for future research.

Chapter 2

Background

In this chapter the mathematical investigation of the problem of joint dereverberation and noise reduction starts by examining two partial solutions to the dereverberation problem - one using single channel spectral subtraction algorithm OM-LSA [18], and the other using an array processing algorithm known as TF-GSC [10]. These solutions incorporate techniques used in the algorithm proposed in this contribution therefore constitute a good starting point for understanding the challenges in solving the JDNR problem.

Section 2.1 features a single channel dereverberation algorithm using spectral enhancement techniques. Section 2.2 details the TF-GSC beamforming algorithm for dereverbration in noisy environments, while section 2.3 discusses of the limitations of these methods, providing the rationale for the algorithms derived in the following chapters.

2.1 Dereverberation Using Spectral Enhancement

This section discusses a spectral enhancement method to suppress late reverberation and ambient noise. The method assumes a statistical model for the RIR, that relies on the fact that its late component is dependent mostly on the reverberation time of the room [47]. Using the statistical model, the algorithm provides a computation of the late reverberant spectral variance, and then suppresses it using the OM-LSA algorithm.

2.1.1 Problem Formulation

The reverberant signal results from the convolution of the anechoic speech signal $s(n)$ and a causal RIR h. The reverberant speech signal at discrete-time n is given by:

$$z(n) = h(n) * s(n). \tag{2.1}$$

In this section we assume that the RIR is time-invariant and that its length is infinite. We split the RIR into two components

$$h(n) = \begin{cases} h_e(n) & n \in [0, n_e] \\ h_l(n) & n \in [n_e + 1, \infty) \\ 0 & else \end{cases}. \tag{2.2}$$

n_e is chosen such that $h_e(n)$ consist of the direct path and a few early reflections and $h_l(n)$ contains of the late reflections. The fraction $\frac{n_e}{f_s}$ is used to define the time instance from which the late reverberations would be suppressed. In practice, it ranges from 20 to 40 ms. Using (2.2) the signal $y(n)$ received at the microphone becomes

$$y(n) = [h_e(n) + h_l(n)] * s(n) + v(n) = z_e(n) + z_l(n) + v(n), \tag{2.3}$$

where $z_e(n)$ is the early speech component, $z_l(n)$ denotes the late reverberant speech component, and $v(n)$ denotes the additive uncorrelated noise component. The goal of the enhancement algorithm is to suppress the late reverberation component $z_l(n)$ and the noise. The analysis begins by transforming the reverberant and noisy signal $y(n)$ into the STFT domain, using

$$Y[l, m] = \sum_{n=0}^{M-1} y(n + lR)\, w(n)\, e^{-j\frac{2\pi}{M}nm}, \tag{2.4}$$

where $w(n)$ is the analysis window of size M, and R is the number of samples separating two successive STFT frames. Using $Y[l, m]$, the spectral variance of the additive noise, $\lambda_v = E\left\{|V[l, m]|^2\right\}$ and the spectral variance of the late reverberant component, $\lambda_{Z_l} = E\left\{|Z_l[l, m]|^2\right\}$ can be estimated. The spectral variance of the noise can be estimated using the minima controlled recursive averaging (MCRA) approach [17]. By which past spectral cross-power values of the noisy observed signals are recursively averaged with a

time-varying frequency-dependent smoothing parameter

$$\lambda_v[l,m] = \tilde{\alpha}_v[l,m]\lambda_v[l-1,m] + \beta(1-\tilde{\alpha}_v[l,m])\max(|Y[l,m]|^2, \lambda_{v,\min}), \tag{2.5}$$

where $\tilde{\alpha}_v[l,m]$ is the smoothing parameter, $\beta \geq 1$ is a factor that compensates the bias when the desired signal is absent, and $\lambda_{v,\min}$ is a low level threshold parameter to reduce musical noise. The smoothing parameter is determined using the reference signal presence probability, $p[l,m]$, and α_v a constant that represents its minimal value

$$\tilde{\alpha}_v[l,m] = \alpha_v + (1-\alpha_v)\, p[l,m]\,. \tag{2.6}$$

The spectral variance of the late reverberation component is more complicated to estimate, as it changes rapidly and is highly correlated with the speech source. An estimate for λ_{Z_l} will be derived in section (2.1.3) using a statistical model for the RIR. For now let the availability of such an estimate suffice. The spectral enhancement problem can be formulated as deriving an estimator $Z_e[l,m]$ for the early reflections spectral coefficients such that a certain distortion measure is minimized.

2.1.2 Generalized Post-Flter

To obtain an estimate of the desired early reverberant signal, we apply the (OM-LSA) method for spectral enhancement [18]. Assuming two different hypotheses for the presence of the early speech signal $H_1(l,m)$ and its absence $H_0(l,m)$, we can write the recorded signal at the reference microphone as:

$$\begin{aligned} H_1[l,m]:\ Y[l,m] &= Z_e[l,m] + Z_l[l,m] + V[l,m] \\ H_0[l,m]:\ Y[l,m] &= Z_l[l,m] + V[l,m]. \end{aligned} \tag{2.7}$$

The *a posteriori* SIR is defined as

$$\gamma[l,m] = \frac{|Z[l,m]|^2}{\lambda_{Z_l}[l,m] + \lambda_v[l,m]}, \tag{2.8}$$

and the *a priori* SIR is defined as

$$\xi[l,m] = \frac{\lambda_{Z_e}[l,m]}{\lambda_{Z_l}[l,m] + \lambda_v[l,m]}. \tag{2.9}$$

The *a posteriori* SIR can be calculated directly, given the spectral variances of both noise and late reverberant component. On the other hand the *a priori* SIR cannot be calculated

directly since $Z_e[l,m]$ is not observable. Estimation of the *a priori* SIR is possible however using a decision directed estimator [18] which is to be described later on.

Under $H_1\,[l,m]$ the log spectral amplitude (LSA) gain function $F_{H_1}[l,m]$ is given by

$$F_{H_1}[l,m] = \frac{\xi[l,m]}{1+\xi[l,m]}\text{expint}\left\{\frac{\xi[l,m]\gamma[l,m]}{1+\xi[l,m]}\right\}, \tag{2.10}$$

where

$$\text{expint}\{x\} = \exp\left(\frac{1}{2}\int_x^{\infty}\frac{e^{-t}}{t}dt\right). \tag{2.11}$$

Under $H_0\,[l,m]$ a lower bound $F_{H_0}[l,m]$ given by

$$F_{H_0}[l,m] = F_{\min}\frac{\lambda_v[l,m]}{\lambda_v[l,m]+\lambda_{Z_l}[l,m]} \tag{2.12}$$

is applied, where $F_{\min}$ is a predefined lower bound. The OM-LSA spectral gain function is obtained as the weighted geometric mean of the gains associated with the speech presence probability denoted by $p\,[l,m]$

$$F_{OM-LSA}\,[l,m] = \{F_{H_1}\,[l,m]\}^{p[l,m]}\,\{F_{H_0}[l,m]\}^{(1-p[l,m])}\,. \tag{2.13}$$

The spectral early component can now be estimated by applying the OM-LSA spectral gain function to each component of the received signal, i.e.,

$$\hat{Z}_e\,[l,m] = F_{OM-LSA}\,[l,m]\,Y\,[l,m]\,. \tag{2.14}$$

Using the log spectral gain function the decision directed estimator for the *a priori* SIR is defined as

$$\hat{\xi}[l,m] = \max\left\{\alpha_e\frac{F_{OM-LSA}\,[l-1,m]\,|Y\,[l-1,m]|^2}{\lambda_v[l,m]+\lambda_{Z_l}[l,m]} + (1-\alpha_e)\,[\gamma[l,m]-1]\,,\xi_{\min}\right\}. \tag{2.15}$$

Where $\xi_{\min}$ is a predefined lower-bound of the *a priori* SIR designated to reduce musical noise effect. The smoothing parameter $\alpha_e \in [0,1)$ controls the tradeoff between the amount of noise reduction and distortion.

2.1.3 Statistical RIR Model

This section presents a statistical model for the RIR first introduced by Habets et al, in [19]. This model is a generalization of the statistical model suggested by Polack in [46].

To model the contribution of the direct-path, the RIR is divided into two segments:

$$h(n) = \begin{cases} h_d(n) & n \in [0, n_d] \\ h_r(n) & n \in [n_d + 1, \infty) \\ 0 & \text{else} \end{cases} \tag{2.16}$$

n_d is chosen such that $h_d(n)$ consist of the direct path and $h_r(n)$ contains all of the reflections. In practice, the direct-path is deterministic and could be modeled using a Dirac pulse. Unfortunately this precludes the creation of a statistical model. Therefore we choose to described the direct path as the following non stationary stochastic process:

$$h_d\left(n\right) = b_d\left(n\right) e^{-\zeta n}, \tag{2.17}$$

where $b_d\left(n\right)$ is a zero mean stationary Gaussian noise, and the decay factor ζ is given by

$$\zeta = \frac{3ln\left(10\right)}{T_{60} f_s}, \tag{2.18}$$

where T_{60} being the reverberation time and f_s the sampling frequency. In [46], Polack proved that the most interesting properties of room acoustics are statistical, when the echo density is high enough such that the space can be considered to be in a fully diffused or mixed state. The essential requirement for the validity of (2.17) is ergodicity, which requires that any given reflection trajectory in the space will eventually reach all points. The ergodicity assumption is determined by the shape of the enclosure and the surface reflection properties, which in turn determine the reverberation time and the model's exponential decay rate.

The reverberant component $h_r(n)$ is modeled in a similar way:

$$h_r\left(n\right) = b_r\left(n\right) e^{-\zeta n}, \tag{2.19}$$

where $b_r\left(n\right)$ is a zero mean stationary Gaussian noise. Assuming the SRA conditions stated in [30] are fulfilled, the direct and reverberant components of the RIR are considered uncorrelated. The energy envelope of the RIR can be expressed as

$$E\left\{h^2(n)\right\} = \begin{cases} \sigma_d^2 e^{-2\zeta n} & n \in [0, n_d] \\ \sigma_r^2 e^{-2\zeta n} & n \in [n_d + 1, \infty), \\ 0 & else \end{cases} \tag{2.20}$$

where σ_d^2 and σ_r^2 are the variances of $b_d(n)$ and $b_r(n)$, respectively. We assume that $\sigma_d^2 \geq \sigma_r^2$ otherwise the direct path can be disregarded.

2.1.4 Late Reverberant Spectral Variance Estimator

In this section a spectral variance estimator for the late reverberant spectral component, $Z_l(k)$, is derived using the statistical reverberation model described in the last section. The analysis starts with the calculation of the reverberant signal $z(n)$, defined at a discrete time n and with lag τ:

$$r_{zz}(n, n+\tau; h) = E\{z(n)z(n+\tau)\}. \tag{2.21}$$

Using (2.2) and (2.1) we obtain

$$\begin{aligned} r_{zz}(n, n+\tau; h) = & \sum_{l=n-n_d+1}^{n} \sum_{l'=n-n_d+\tau}^{n+\tau} E\{s(l)\, s(l')\}\, h_d(n-l)\, h_d(n+\tau-l') \\ & + \sum_{l=-\infty+1}^{n-n_d} \sum_{l'=-\infty}^{n-n_d+\tau} E\{s(l)\, s(l')\}\, h_r(n-l)\, h_r(n+\tau-l'). \end{aligned} \tag{2.22}$$

Using (2.20) and the assumption that $b_d(n)$ and $b_r(n)$ are uncorrelated zero mean white Gaussian noise sequences it follows that

$$E_\theta\{h_v(n-l)\, h_v(n+\tau-l')\} = \sigma_v^2 e^{-2\zeta n} e^{\zeta(l+l'-\tau)} \delta(l-l'+\tau),\ v \in [d, r] \tag{2.23}$$

and

$$E_\theta\{h_r(n)\, h_d(n+\tau)\} = 0 \tag{2.24}$$

where E_θ is the spatial expectation operator (averaging all possible source microphone positions), $\delta()$ denotes the Kronecker delta function. Assuming the source signal and the RIR are mutually independent, the spatially averaged auto-correlation results in

$$r_{zz}(n, n+\tau) = r_{z_d z_d}(n, n+\tau) + r_{z_r z_r}(n, n+\tau) \tag{2.25}$$

with

$$r_{z_d z_d}(n, n+\tau) = e^{-2\zeta n} \sum_{l=n-n_d+\tau}^{n} E\{s(l)\, s(l')\}\, \sigma_d^2 e^{2\zeta l} \tag{2.26}$$

and

$$\begin{aligned} r_{z_r z_r}(n, n+\tau) &= e^{-2\zeta n} \sum_{l=-\infty}^{n-n_d} E\{s(l)\, s(l+\tau)\}\, \sigma_r^2 e^{2\zeta l} \\ &= e^{-2\zeta n} \sum_{l=n-2n_d+1}^{n-n_d} E\{s(l)\, s(l+\tau)\}\, \sigma_r^2 e^{2\zeta l} \\ &+ e^{-2\zeta n} \sum_{l=-\infty}^{n-2n_d} E\{s(l)\, s(l+\tau)\}\, \sigma_r^2 e^{2\zeta l}. \end{aligned} \tag{2.27}$$

The first term in (2.27) depends on the direct signal between time $n - n_d + 1$ and n , and the second depends on the reverberant signal. Using the last two expressions the auto-correlations function r_{zz} at time $n - n_d$ is given by

$$\begin{aligned} r_{zz}(n-n_d, n-n_d+\tau) = e^{-2\zeta(n-n_d)} \Bigg[&\sum_{l=n-2n_d+1}^{n-n_d} E\{s(l)\, s(l+\tau)\}\, \sigma_d^2 e^{2\zeta l} \\ &+ \sum_{l=-\infty}^{n-2n_d} E\{s(l)\, s(l+\tau)\}\, \sigma_r^2 e^{2\zeta l} \Bigg]. \end{aligned} \tag{2.28}$$

Using (2.27) and (2.28) the auto correlation of the reverberant part can be written as

$$\begin{aligned} r_{z_r z_r}(n, n+\tau) &= \kappa e^{-2\zeta n_d} r_{z_d z_d}(n-n_d, n-n_d+\tau) \\ &+ e^{-2\zeta n_d} r_{z_r z_r}(n-n_d, n-n_d+\tau) \end{aligned} \tag{2.29}$$

with $\kappa = \frac{\sigma_r^2}{\sigma_d^2}$. Using (2.25), (2.29) can be rewritten as:

$$\begin{aligned} r_{z_r z_r}(n, n+\tau) &= (1-\kappa)\, e^{-2\zeta n_d} r_{z_r z_r}(n-n_d, n-n_d+\tau) \\ &+ \kappa e^{-2\zeta n_d} r_{zz}(n-n_d, n-n_d+\tau). \end{aligned} \tag{2.30}$$

The late reverberant component can now be obtained using

$$r_{z_l z_l}(n, n+\tau) = (1-\kappa)\, e^{-2\zeta(n_e-n_d)} r_{z_r z_r}(n-n_e+n_d, n-n_e+n_d+\tau). \tag{2.31}$$

The parameter κ is related to the Direct to Reverberant Ratio (DRR), which is defined as

$$DRR = \frac{E_d}{E_r} = \frac{\sum_{l=0}^{n_d} h^2(n)}{\sum_{l=n_d+1}^{\infty} h^2(n)} = \frac{\sum_{l=0}^{n_d} \sigma_d^2 e^{-2\zeta l}}{\sum_{l=n_d+1}^{\infty} \sigma_r^2 e^{-2\zeta l}} = \frac{\sigma_d^2 \left(1 - e^{-2\zeta n_d}\right)}{\sigma_r^2 e^{-2\zeta n_d}}. \tag{2.32}$$

From (2.32) κ can be expressed in terms of the DRR

$$\kappa = \frac{\sigma_r^2}{\sigma_d^2} = \frac{\left(1 - e^{-2\zeta n_d}\right)}{e^{-2\zeta n_d}} DRR^{-1}. \tag{2.33}$$

In general the DRR is frequency dependent, as shown in [20, chapter 3]. Hence, in order to improve the accuracy of the model, κ is made frequency dependent. Furthermore, since the DRR, and thus κ, depend on the distance between source and microphone, spatial averaging can only be performed over microphone signals that have the same source-microphone distance. κ can also be estimated adaptively as described in [19].

Assuming that $n_e \ll T_{60} f_s$, and that n_d is equal to R, the number of samples separating two successive STFT frames, the spectral variance of the expression in (2.30) is obtained:

$$\lambda_{z_r}[l,m] = e^{-2\zeta[m]R}\left(1-\kappa[m]\right)\lambda_{z_r}[l-1,m] + \kappa[m]\, e^{-2\zeta[m]R}\lambda_z[l-1,m]\,. \tag{2.34}$$

The late reverberant spectral variance (LRSV) is then given by:

$$\lambda_{z_l}[l,m] = e^{-2\zeta[m](n_e-R)}\lambda_{z_r}\left[l-\frac{n_e}{R}+1,m\right]. \tag{2.35}$$

2.2 TF-GSC

2.2.1 Problem formulation

Consider the problem of extracting the early reflection component, from signals received by K sensors arranged in an arbitrary array, where each sensor captures a convolved source signal in some noise field. The received signals at the discrete-time index n are expressed as:

$$\begin{aligned} y_k(n) &= h_k(n) * s(n) + v_k(n) \\ &= x_k(n) + v_k(n),\ k = 1,2,...,K \end{aligned} \tag{2.36}$$

where $h_k(n)$ is the Linear Time Invariant (LTI) RIRs relating the desired sources, $s(n)$ and each sensor k, $v_k(n)$ is the additive noise at the k-th microphone. We assume that the signals $x_k(n) = h_k(n) * s(n)$ and $v_k(n)$ are uncorrelated, zero mean, real, and broadband. Expression (2.36) can be rewritten in the STFT domain, at time index l and the frequency bin m, as:

$$\begin{aligned} Y_k[l,m] &= H_k[m]S[l,m] + V_k[l,m] \\ &= X_k[l,m] + V_k[l,m]\,, k = 1,2,...,K, \end{aligned} \tag{2.37}$$

where $Y_k[l,m]$, $V_k[l,m]$, $S[l,m]$, are the STFT representations of $y_k(n)$, $v_k(n)$, and $s(n)$ respectively. $H_k[m]$ is the frequency representation of $h_k(n)$. For the sake of the formulation simplicity we assume Multiplicative Transfer Function (MTF) approximation when replacing convolution in multiplication in the STFT domain. The assumption that the STFT window length is much larger than the RIR length ensures the MTF approximation validness. Expression (2.36) can be described in vector notation as:

$$\begin{aligned}\overline{y}[l,m] &= \bar{x}[l,m] + \overline{v}[l,m] \\ &= \overline{H}[m]S[l,m] + \overline{v}[l,m],\end{aligned} \tag{2.38}$$

where

$$\begin{aligned}\overline{y}[l,m] &= [Y_1[l,m], Y_2[l,m], ..., Y_K[l,m]]^T \\ \overline{x}[l,m] &= [X_1[l,m], X_2[l,m], ..., X_K[l,m]]^T \\ \overline{v}[l,m] &= [V_1[l,m], V_2[l,m], ..., V_K[l,m]]^T\end{aligned} \tag{2.39}$$

and

$$\overline{H}[m] = [H_1[m], H_2[m], ..., H_K[m]]^T . \tag{2.40}$$

Assuming the desired speech signals, the interference and the noise signals to be uncorrelated, we have from (2.38)

$$\begin{aligned}\Phi_{yy}[l,m] &= E\left[\overline{y}[l,m]\overline{y}^H[l,m]\right] = \Phi_{xx}[l,m] + \Phi_{vv}[l,m] \\ &= S\left[l,m\right]\overline{H}[m]\overline{H}^H[m] + \Phi_{vv}[l,m].\end{aligned} \tag{2.41}$$

A beamformer is constructed by applying a set of filters $\overline{W}\left[l,m\right]$ to each microphone signal and summing up all the signals:

$$Z[l,m] = \sum_{k=0}^{K} w_k^*[l,m]Y_k[l,m] = \overline{W}^H[l,m]\overline{y}[l,m] \tag{2.42}$$

where $Z\left[l,m\right]$ is the beamformer's output STFT representation, and $\overline{W}\left[l,m\right]$ is the beamformer's weights vector at time frame l and frequency bin m.

2.2.2 The TF-GSC Beamformer

A multi-channel beamforming approach for signal enhancement based on the non-stationarity of the desired signal was proposed by Gannot et al., [10]. This approach,

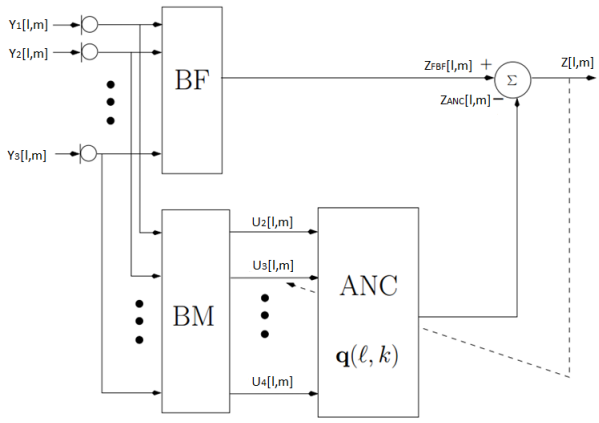


Figure 2.1: Generalized sidelobe canceler.

aiming at enhancing a single desired source, can be exploited for extracting the clean source signal from a set of noisy reverberant observations. The beamformer, $\overline{W}[l,m]$, is determined by solving the following minimization problem

$$\min_{\overline{h}[l,m]} \left\{ \overline{W}^H[l,m]\Phi_{yy}[l,m]\overline{W}[l,m] \right\} \text{ subject to } \overline{W}^H[l,m]\bar{H}[m] = 1. \tag{2.43}$$

The vector $\overline{W}[l,m]$ which solves this optimization problem is called the MVDR beamformer. The output of the MVDR beamformer is equal to the desired signal $S[l,m]$. It is shown, that if the constraint is relaxed, such that the desired signal component at the output is given by $X_1[l,m] = S[l,m]\,H_1[m]$, i.e. the reverberant signal as received by the first (reference) microphone, the Relative Transfer Function (RTF)

$$\widetilde{\overline{H}}[m] = \frac{\overline{H}[m]}{H_1[m]}, \tag{2.44}$$

suffices for implementing the MVDR beamformer. The MVDR minimization problem can be efficiently implemented by constructing a GSC structure as described in Fig. 2.1. which contains an FBF responsible for aligning the desired signal component, a BM which blocks the desired signal and constructs the noise reference signals $\overline{U}[l,m]$, and a multichannel ANC which cancels out all interference components from the FBF output by using the reference signals. It is shown in [10], that the FBF can be implemented by:

$$\overline{W}_0[l,m] = \frac{\widetilde{\overline{H}}[m]}{\left\|\widetilde{\overline{H}}[m]\right\|^2} \tag{2.45}$$

and that the blocking matrix can be chosen as,

$$B[m] = \begin{pmatrix} -\widetilde{H}_0^*[m] & -\widetilde{H}_2^*[m] & \dots & -\widetilde{H}_N^*[m] \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & \dots & \ddots & \\ 0 & 0 & \dots & 1 \end{pmatrix} \tag{2.46}$$

which satisfies the constraint $\overline{\widetilde{H}}^H[m]\,B[m] = 0$ and constructs the reference signals $\overline{U}[l,m] = B^H[m]\,\overline{Y}[l,m]$. The output of the GSC beamformer is a difference of two terms, both of which operate on input signal $Y[l,m]$

$$Z[l,m] = Z_{FBF}[l,m] - Z_{ANC}[l,m] \tag{2.47}$$

where

$$\begin{aligned} Z_{FBF}[l,m] &= \overline{W}_0^H[l,m]\,\overline{Y}[l,m] \\ Z_{ANC}[l,m] &= \overline{G}^H[l,m]\,\overline{U}[l,m]\,. \end{aligned} \tag{2.48}$$

The ANC filters, $\overline{G}[l,m]$, are adjusted to reduce the residual noise at the beamformer output $Z[l,m]$, using the minimum output power criteria:

$$\min_{\mathrm{G}} \left\{ \left[\overline{W}_0[l,m] - B[m]\,\overline{G}[l,m]\right]^H \Phi_{\mathrm{yy}}[l,m] \left[\overline{W}_0[l,m] - B[m]\,\overline{G}[l,m]\right] \right\}. \tag{2.49}$$

The minimization problem is in fact the classical multichannel noise cancellation problem, which can be solved using an adaptive LMS algorithm as proposed by Widrow [29]:

$$\overline{G}[l+1,m] = \begin{cases} \overline{G}[l,m] + \frac{\mu_g \overline{U}[l,m] Z^*[l,m]}{P_{est}[l,m]} & \text{Noise only frames} \\ \overline{G}[l,m] & \text{otherwise} \end{cases} \tag{2.50}$$

where

$$P_{est}[l,m] = \alpha_p P_{est}[l-1,m] + (1-\alpha_p)\,\|U[l,m]\|^2 \tag{2.51}$$

represents the power of the noise reference signals, μ_g is a step-size that regulates the convergence rate and α_p is a smoothing parameter in the PSD estimation process. The filters are usually constrained to an FIR structure for stabilizing the update algorithm. This can be achieved by keeping only the non-aliased samples in accordance with the

overlap & save method. In practice the RTF's needed for the construction of the blocking matrix are unknown and should be estimated, a method for estimating them is presented in [10].

Define the cross PSD between the ith and jth observation as $\Phi_{Y_i,Y_j}[l,m]$ and $\Phi_{U_K,Y_j}[l,m]$ the cross PSD between $U_k[l,m]$ and the j-th observation $Y_j[l,m]$. Let $\hat{\Phi}_{Y_i,Y_j}[l,m]$ and $\hat{\Phi}_{U_K,Y_j}[l,m]$ represent the corresponding estimates. Further define the error term $\epsilon_j[l,m] = \hat{\Phi}_{Y_i,Y_j}[l,m] - \hat{\Phi}_{U_K,Y_j}[l,m]$. An unbiased estimate for the RTF's vector $\widetilde{\overline{H}}[m]$ is obtained by applying a least squares criteria to the following set L of over-determined equations, with L being the number of STFT frames.

$$\begin{bmatrix} \Phi_{Y_1,Y_j}[1,m] \\ \Phi_{Y_1,Y_j}[2,m] \\ . \\ . \\ . \\ \Phi_{Y_1,Y_j}[L,m] \end{bmatrix} = \begin{bmatrix} \Phi_{Y_1,Y_1}[1,m] & 1 \\ \Phi_{Y_1,Y_1}[2,m] & 1 \\ . & . \\ . & . \\ . & . \\ \Phi_{Y_1,Y_1}[L,m] & 1 \end{bmatrix} \times \begin{bmatrix} \widetilde{\overline{H}}_j[m] \\ \hat{\Phi}_{U_K,Y_j}[l,m] \end{bmatrix} + \begin{bmatrix} \epsilon_j[1,m] \\ \epsilon_j[2,m] \\ . \\ . \\ . \\ \epsilon_j[L,m] \end{bmatrix} \tag{2.52}$$

where a separate set of equations is used for each microphone signal ($j = 2, ..., K$) and frequency bin m. A major limitation of this estimation process, is the absence of reference to the presence of speech when applying the least squares criteria. Since noise only frames hold no valid information regarding the source microphone position, accounting for them in the estimation process will severely degrade it. Furthermore, the RTF is assumed to be constant during the observation interval. Hence, very long observation intervals also restrict the capability of the system identification technique to track varying environmental changes. In Chapter 3 we describe an improved estimation process, based on the one suggested by Cohen in [15], where these limitations are mitigated.

2.2.3 Discussion

The drawbacks of the algorithms [10, 18] presented above limit the ability to deal with the JDNR problem. While the generalized OM-LSA algorithm presented in the first part of this chapter is considered to be a robust and efficient algorithm for noise reduction, its strong dependence upon the estimation process of the LRSVE makes it vulnerable to errors of overestimation which are particularly common when the DRR is low. In addition since the OM-LSA is a single-channel algorithm, it is bound to be outperformed by multi-channel algorithms using spatial information that block noise and reverberation coming from different directions. The OM-LSA is mostly used as a post-filter at the output of a beamformer in order to reduce remaining residual noise.

A major drawback of the TF-GSC lies in its requirement of a highly accurate BM estimation. This is known as the leakage problem. If the desired speech leaks into the noise reference signals, the noise canceling filters will subtract speech components from the FBF output. The BM is constructed out of the RTFs from the speech source to the microphone array. Estimation of the RTFs is a challenging task, especially in reverberant environments where the RTF's bandwidth is long. A TF-GSC scheme based on inaccurate RTF representation, does not meet the MVDR criterion, since both the power minimization and the constraint of the TF-GSC are inaccurately formed, causing self-cancellation and inflicting high speech distortion.

Another drawback of the TF-GSC lies in its inability to suppress reverberation arriving at its look direction. This occurs since the beamformer is designed to extract the clean, but reverberant part of the signal in the array reference channel. To tackle this problem a post-filter (OM-LSA) [34] is usually applied. Such an operation, aside from being costly is vulnerable to error caused by the beamforming operation, and may further increase speech distortion.

The aim of this thesis therefore is to suggest algorithms that might circumvent these limitations. Where the emphasis is placed on developing beamforerms with high dereverberation capabilities, capable of maintaining low noise and speech distortion levels at their output. Our proposed methods are presented in Chapters 3 and 4.

Chapter 3

Optimal STFT-Domain Beamformers for Joint Dereverberation and Noise Reduction

3.1 Introduction

This chapter introduces a new framework in which known optimal beamformers are adjusted to deal with the JDNR problem. The outline of this chapter is as follows. In Section 3.2 we formulate the JDNR problem, and outline fundamental assumptions and definitions. In Section 3.3 we introduce the general linear array model. In section 3.4 we describe important performance measures using the linear array model. In Section 3.5 we describe different optimal beamformers that were deduced using the signal model. In Section 3.6 we describe an algorithm for the estimation of the full to early RTFs needed for the construction of the suggested beamformers. Implementation of the beamformers is described in Section 3.7. In Section 3.8 we introduce results, and analyze the performance of the beamformers. Finally, Section 3.9 concludes this chapter.

3.2 Signal Model and Problem Formulation

Consider the conventional signal model in which a microphone array with N sensors captures a convolved source signal in some noise field. The received signals at the discrete-time index k are expressed as [13, 14],

$$\begin{aligned} y_n(k) &= g_n(k) * s(k) + v_n(k) \\ &= x_n(k) + v_n(k),\ n = 1, 2, ..., N \end{aligned} \tag{3.1}$$

where $g_n(k)$ is the acoustic impulse response from the unknown source's $s(k)$ position to the nth microphone, and $v_n(k)$ is the additive noise at the nth microphone. We assume that signals $x_n(k) = g_n(k) * s(k)$ and $v_n(k)$ are uncorrelated, zero-mean, real, and broadband. The signal recorded at the first microphone is defined as the reference signal. This signal will be used later on for calculation of several important performance measures. It should be noted that any other microphone could be considered as the reference, and that the choice could influence the performance of the beamformer. The convolved speech signal at microphone 1 $x_1(k)$ can be further decomposed into two parts:

$$\begin{aligned} x_1(k) = g_1(k) * s(k) &= [g_{1,c}(k) + g_{1,l}(k)] * s(k) \\ &= x_{1,c}(k) + x_{1,\mathrm{l}}(k). \end{aligned} \tag{3.2}$$

$x_{1,c}(k)$ represents the clean early speech signal at the reference microphone, which corresponds to the early reflections of the RIR $g_1(k)$ [i.e., its first part, $g_{1,\mathrm{c}}(k)$], while $x_{1,l}(k)$ represents the reverberant speech signal due to the late reflections of $g_1(k)$.

Joint dereverberation and noise reduction consist of removing both the noise component $v_1(k)$ and the reverberant component $x_{1,l}(k)$ from the reference observation $y_1(k)$. In array processing methodology, this would mean extracting the clean early signal $x_{1,c}(k)$ from the N observations $y_n(k)$ $n = 1, 2, ..., N$. The extraction process should impose as little distortion as possible upon the desired speech signal, while minimizing the contribution of both the noise component and late reverberations. Though $x_{1,c}(k)$ varies from the actual source signal, we choose to define it as the desired signal. This choice is backed up by hearing tests showing that early reflections do not affect the intelligibility of speech signals, making them both clearer and more pleasant to the listener [20].

Expression (3.1) can be rewritten in the STFT domain, at time index l and frequency bin m, as:

$$\begin{aligned} Y_n[l,m] &= G_n[m]S[l,m] + V_n[l,m] \\ &= X_n[l,m] + V_n[l,m]\,, n = 1,2,...,N, \end{aligned} \tag{3.3}$$

where $Y_n[l,m]$, $S[l,m]$, $X_n[l,m] = G_n[m]S[l,m]$, and $V_n[l,m]$ are the STFT-domain representations of $y_n(k)$, $s(k)$, $x_n(k) = g_n(k) * s(k)$ and $v_n[l,m]$ respectively. $G_n[m]$ is the frequency representation of $g_n(k)$. For the sake of simplicity we assume the Multiplicative Transfer Function (MTF) approximation when replacing convolution in multiplication in the STFT domain [24]. This approximation tends to become less accurate as the T_{60} increases i.e., the environment becomes more reverberant. However it is within reasonable bounds. In accordance with (3.2), the decomposition of the reference signal $X_1[l,m]$ in the STFT domain is given by

$$X_1[l,m] = X_{1,c}[l,m] + X_{1,l}[l,m] \tag{3.4}$$

where $X_{1,c}[l,m] = G_{1,c}[m]S[l,m]$ and $X_{1,l}[l,m] = G_{1,l}[m]S[l,m]$ are STFT representation of the early clean speech signal and the late reverberation signal respectively. Using the N, STFT-domain microphone signals, the problem can be formulated in vector notation as follows:

$$\begin{aligned} \overline{y}[l,m] &= \bar{x}[l,m] + \overline{v}[l,m] \\ &= \bar{d}_{X_{1,c},\overline{x}}[m]X_{1,c}[l,m] + \overline{v}[l,m], \end{aligned} \tag{3.5}$$

where

$$\begin{aligned} \overline{y}[l,m] &= [Y_1[l,m], Y_2[l,m], ..., Y_N[l,m]]^T \\ \overline{x}[l,m] &= [X_1[l,m], X_2[l,m], ..., X_N[l,m]]^T \\ &= S\,[l,m]\,[G_1\,[m]\,, G_2\,[m]\,, ..G_N\,[m]] = S\,[l,m]\,\bar{g}[m] \\ \overline{v}[l,m] &= [V_1[l,m], V_2[l,m], ..., V_N[l,m]]^T \\ \overline{d}_{X_{1,c},\overline{x}}[m] &= \left[\frac{G_1[m]}{G_{1,c}[m]}, \frac{G_2[m]}{G_{1,c}[m]}, ..., \frac{G_N[m]}{G_{1,c}[m]}\right]^T = \frac{\bar{g}[m]}{G_{1,c}[m]}. \end{aligned} \tag{3.6}$$

Expression (3.5) depends explicitly on the desired signal $X_{1,c}[l,m]$, making it the STFT-domain signal model for joint reverberation and noise reduction. The vector $\overline{d}_{X_{1,c},\overline{x}}[m]$

is defined as the STFT-domain steering vector for the case of joint dereverberation and noise reduction, since it determines the direction of the desired signal $X_{1,c}[l,m]$. Here we assume that $G_{1,c}[m] \neq 0\,\forall m$, so that the full to early RTFs that construct $\overline{d}_{X_{1,c},\overline{x}}[m]$ are always defined. This definition is a generalization of the classical steering vector [25, 26] for a reverberant environment, and in a way an adaptation of the work in [13] to RTF based formulation. Using (3.5) we can deduce the $N \times N$ covariance matrix of $\overline{y}[l,m]$

$$\begin{aligned}\Phi_{yy}[l,m] &= E\left[\overline{y}[l,m]\overline{y}^H[l,m]\right] \\ &= \phi_{X_{1,c}}[l,m]\bar{d}_{X_{1,c},\overline{x}}[m]\bar{d}^H_{X_{1,c},\overline{x}}[m] + \Phi_{vv}[l,m] \\ &= \Phi_{xx}[l,m] + \Phi_{vv}[l,m], \end{aligned} \tag{3.7}$$

where $\phi_{X_{1,c}}[l,m] = E\left[|X_{1,c}[l,m]|^2\right]$ is the variance of $X_{1,c}[l,m]$, $\Phi_{xx}[l,m] = E\left[\overline{x}[l,m]\overline{x}^H[l,m]\right]$ and $\Phi_{vv}[l,m] = E\left[\overline{v}[l,m]\overline{v}^H[l,m]\right]$ are covariance matrices of $\overline{v}[l,m]$ and $\overline{x}[l,m]$, respectively.

3.3 Linear Array Model

In the STFT domain, the conventional beamforming for multi-channel joint dereverberation and noise reduction is performed by applying a complex weight to the output of each sensor at frequency-bin m and time frame l, then summing across the aperture [13, 14]:

$$\begin{aligned} Z[l,m] &= \sum_{n=1}^{N} H_n^*[l,m]Y_n[l,m] = \overline{h}^H[l,m]\overline{y}[l,m] \\ &= X_{1,c}[l,m]\overline{h}^H[l,m]\overline{d}_{X_{1,c},\overline{x}}[m] + \overline{h}^H[l,m]\overline{v}[l,m] \\ &= X_{fd}[l,m] + V_{rn}[l,m], \end{aligned} \tag{3.8}$$

where

$$\overline{h}[l,m] = [H_1[l,m], H_2[l,m], ..., H_N[l,m]]^T \tag{3.9}$$

is a filter of length N containing all the complex gains applied to the microphone outputs. $X_{fd}[l,m] = X_{1,c}[l,m]\overline{h}^H[l,m]\overline{d}_{X_{1,c},\overline{x}}[m]$ is the filtered desired signal, and $V_{rn}(f) = \overline{h}^H[l,m]\overline{v}[l,m]$ is the filtered residual noise. The two terms on the right-hand side of (3.8) are uncorrelated, hence the variance of $Z[l,m]$ is also the sum of two variances:

$$\phi_Z[l,m] = \overline{h}^H[l,m]\Phi_{yy}[l,m]\overline{h}[l,m] = \phi_{X_{fd}}[l,m] + \phi_{V_{rn}}[l,m], \tag{3.10}$$

where $\phi_{X_{fd}}[l,m] = \phi_{X_{1,c}}[l,m] \left|h^H[l,m]\overline{d}_{X_{1,c},\overline{x}}[m]\right|^2$and $\phi_{V_{rn}}[l,m] = \overline{h}^H[l,m]\Phi_{vv}[l,m]\overline{h}[l,m]$. These variances will be used in the next chapter for the definition of important performance measures.

3.4 Performance Measures

This section discusses useful performance measures necessary for the proper design of beamformers in the STFT domain. One of the most fundamental measures in all aspects of speech enhancement is the signal-to-noise ratio (SNR). The input SNR (before any processing) is derived from the signal model (3.5), and the output SNR (after beamforming) is derived from the array model (3.8). Since microphone 1 is the reference microphone, all measures will be defined with respect to the signal it records. In the STFT domain we differentiate between narrowband (i.e., single frequency) SNR and broadband (i.e., across the entire frequency range) SNR. The narrowband input SNRs at time-frame l is given by:

$$\begin{aligned} iSNR[l,m] &= \frac{\phi_{Y_1}[l,m]}{\phi_{V_1}[l,m]} \\ &= \left|\overline{d}_{X_{1,c},\overline{x}}[l,m]\right|^2 \frac{\phi_{X_{1,c}}[l,m]}{\phi_{V_1}[l,m]}, m = 0,1,....,M-1 \end{aligned} \tag{3.11}$$

where $\phi_{X_{1,c}}[l,m]$, $\phi_{V_1}[l,m]$, and $\phi_{Y_1}[l,m]$ are the variances of $X_{1,c}[l,m]$, $V_1[l,m]$, and $Y_1[l,m]$, respectively. The time-dependent broadband input SNR is defined as

$$iSNR[l] = \frac{\sum_{m=0}^{M-1} \phi_{X_{1,c}}[l,m]}{\sum_{m=0}^{M-1} \phi_{V_1}[l,m] \left|\overline{d}_{X_{1,c},X_1}[m]\right|^{-2}}. \tag{3.12}$$

Property 3.1

The broadband iSNR is smaller or equal to the summation of the narrowband iSNR across all frequency bands

$$iSNR[l] \leq \sum_{m=0}^{M-1} iSNR[l,m]. \tag{3.13}$$

Proof. For any two positive real scalars a_m and b_m , the following inequality stands

$$\frac{\sum_{m=0}^{M-1} a_m}{\sum_{m=0}^{M-1} b_m} = \sum_{m=0}^{M-1} \left(\frac{a_m}{b_m} \frac{b_m}{\sum_{m=0}^{M-1} b_m} \right) \leq \sum_{m=0}^{M-1} \frac{a_m}{b_m} \tag{3.14}$$

which completes the proof.

□

The narrowband and broadband output SNRs at time frame l and frequency bin m are defined in the following way:

$$\begin{aligned} oSNR\left\{\overline{h}[l,m]\right\} &= \frac{\phi_{X_{fd}}[l,m]}{\phi_{V_{rn}}[l,m]} \\ &= \frac{\phi_{X_{1,c}}[l,m]\left|\overline{h}^H[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]\right|^2}{\overline{h}^H[l,m]\Phi_{vv}[l,m]\overline{h}[l,m]}, \end{aligned} \tag{3.15}$$

$$oSNR\left\{\overline{h}[l,:]\right\} = \frac{\sum_{m=0}^{M-1}\phi_{X_{1,c}}[l,m]\left|\overline{h}^H[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]\right|^2}{\sum_{m=0}^{M-1}\overline{h}^H[l,m]\Phi_{vv}[l,m]\overline{h}[l,m]}. \tag{3.16}$$

For the particular filter

$$i_{X_{1,c}X_1}[l,m] = \frac{1}{\overline{d}_{X_{1,c},X_1}[l,m]}\left[1,0,....,0\right]^T = \frac{1}{\overline{d}_{X_{1,c},X_1}[l,m]}i_1 \tag{3.17}$$

of length N, where i_1 is the first column of the N $\times$ N identity matrix I_N, the oSNR equals the iSNR.

For any two vectors $\overline{h}[l,m]$ and $\overline{d}_{X_{1,c},\overline{x}}[l,m]$ and a positive definite matrix $\Phi_{vv}[l,m]$, we have

$$\left|\overline{h}^H[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]\right|^2 \leq \left[\overline{h}^H[l,m]\Phi_{vv}[l,m]\overline{h}[l,m]\right]\left[\overline{d}_{X_{1,c},\overline{x}}[l,m]\Phi_{vv}^{-1}[l,m]\overline{d}^H_{X_{1,c},\overline{x}}[l,m]\right]. \tag{3.18}$$

Using the previous inequality in (3.18), we deduce an upper bound for the narrowband oSNR:

$$oSNR\left[\overline{h}[l,m]\right] \leq \phi_{X_{1,c}}[l,m]\overline{d}^H_{X_{1,c},\overline{x}}[l,m]\Phi_{vv}^{-1}[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]\ \forall\overline{h}[l,m] \tag{3.19}$$

and in particular for the identity filter,

$$oSNR\left[i_{X_{1,c}X_1}[l,m]\right] \leq \phi_{X_{1,c}}[l,m]\overline{d}^H_{X_{1,c},\overline{x}}[l,m]\Phi_{vv}^{-1}[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m] \tag{3.20}$$

this implies:

$$\overline{d}^H_{X_{1,c},\overline{x}}[l,m]\Phi_{vv}^{-1}[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m] \geq \frac{1}{\phi_{V_1}[l,m]}. \tag{3.21}$$

To measure the distortion of the desired speech signal due to the filtering operation, we introduce the speech distortion index, which is defined as the mean-square error between the desired signal and its estimate, normalized by the variance of the desired signal. The

narrowband and broadband speech distortion indices are defined as

$$\begin{aligned}\nu_{sd}[\overline{h}[l,m]] &= \frac{E\left[|X_{fd}[l,m]-X_{1,c}[l,m]|^2\right]}{E\left[|X_{1,c}[l,m]|^2\right]} \\ &= \left|\overline{h}^H[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]-1\right|^2,\end{aligned} \tag{3.22}$$

$$\nu_{sd}[\overline{h}[l,:]] = \frac{\sum_{m=0}^{M-1}\phi_{X_{1,c}}[l,m]\left|\overline{h}^H[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]-1\right|^2}{\sum_{m=0}^{M-1}\phi_{X_{1,c}}[l,m]}. \tag{3.23}$$

To produce an undistorted output signal the beamformers should fulfill the constraint $\nu_{sd}=0$. Therefore, if there is no distortion the speech distortion index is equal to 0 and is expected to be greater than 0 whenever distortion occurs.

3.4.1 Mean-Square Error Criterion

Error criteria play a critical role in deriving optimal beamformers. Although many different criteria can be defined, the mean-square error (MSE) is by far the most common one due to its simplicity in terms of deriving useful and practical beamforming algorithms. This section presents relevant error signals from which meaningful MSE-based criteria are built in the STFT domain.

We define the error signal between the estimated and desired signals at frequency bin m and time frame l as:

$$\begin{aligned}\varepsilon[l,m] &= Z[l,m]-X_{1,c}[l,m] \\ &= \bar{h}^H[l,m]\overline{y}[l,m]-X_{1,c}[l,m] \\ &= X_{fd}[l,m]+V_{rn}[l,m]-X_{1,c}[l,m].\end{aligned} \tag{3.24}$$

This error can also be expressed as:

$$\varepsilon[l,m] = \varepsilon_d[l,m]+\varepsilon_r[l,m] \tag{3.25}$$

where

$$\varepsilon_d[l,m] = \left|\overline{h}^H[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]-1\right| X_{1,c}[l,m] \tag{3.26}$$

is the speech distortion due to the complex filter, and $\varepsilon_r[l,m]=\bar{h}^H[l,m]\overline{v}[l,m]$ represents residual noise. Since we assume no correlation between the desired speech and the additive

noise, the error signals $\varepsilon_d[l,m]$ and $\varepsilon_r[l,m]$ are considered incoherent. From (3.24) the STFT domain narrowband MSE is given by:

$$\begin{aligned} J[\bar{h}[l,m]] &= E\left[|\varepsilon[l,m]|^2\right] \\ &= \phi_{X_{1,c}}[l,m] + \overline{h}^H[l,m]\Phi_{yy}[l,m]\overline{h}[l,m] - 2Re\left\{\overline{h}^H[l,m]E\left[X_{1,c}[l,m]\overline{x}^*[l,m]\right]\right\}. \end{aligned} \tag{3.27}$$

Using (3.25) the narrowband MSE can also be written as:

$$\begin{aligned} J[\bar{h}[l,m]] &= E\left[|\varepsilon_d[l,m]|^2\right] + E\left[|\varepsilon_r[l,m]|^2\right] \\ &= J_d[\bar{h}[l,m]] + J_r[\bar{h}[l,m]] \end{aligned} \tag{3.28}$$

where

$$J_d[\bar{h}[l,m]] = \phi_{X_{1,c}}[l,m]\left|\overline{h}^H[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m] - 1\right|^2, \tag{3.29}$$

and

$$J_r[\bar{h}[l,m]] = \overline{h}^H[l,m]\Phi_{vv}[l,m]\overline{h}[l,m]. \tag{3.30}$$

In a similar way we define the broadband MSE:

$$\begin{aligned} J[\bar{h}[l,:]] &= \frac{1}{M}\sum_{m=0}^{M-1} J[\bar{h}[l,m]] \\ &= \frac{1}{M}\sum_{m=0}^{M-1} J_d[\bar{h}[l,m]] + \frac{1}{M}\sum_{m=0}^{M-1} J_r[\bar{h}[l,m]] \\ &= J_d[\bar{h}[l,:]] + J_r[\bar{h}[l,:]]. \end{aligned} \tag{3.31}$$

It is straightforward to see that minimizing the narrowband MSE for each frequency bin m is equivalent to minimizing the broadband MSE. For the particular filters $i_{X_{1,c}X_1}[l,m]$ (3.17) and $0_{\mathrm{N}\times 1}[l,m]$ (N zeros), the minimum square errors are:

$$J\left[i_{X_{1,c}X_1}[l,m]\right] = \frac{\phi_{V_1}[l,m]}{\left|\overline{d}_{X_{1,c},\overline{x}}[l,m]\right|^2}, \tag{3.32}$$

$$J\left[0_{Nx1}[l,m]\right] = \phi_{X_{1,c}}[l,m]. \tag{3.33}$$

Using (3.32) and (3.33) we can write the input SNR as:

$$iSNR[l,m] = \frac{J\left[0_{Nx1}[l,m]\right]}{J\left[i_{X_{1,c}X_1}[l,m]\right]}. \tag{3.34}$$

We can also rewrite the narrow band speech distortion index

$$\nu_{sd}\left[\overline{h}[l,m]\right] = \frac{J_d[\bar{h}[l,m]]}{J\left[0_{Nx1}[l,m]\right]}. \tag{3.35}$$

In the STFT domain, we are interested in filters for which

$$J_d\left[i_{X_{1,c}X_1}[l,m]\right] \leq J_d[\bar{h}[l,m]] < J_d\left[0_{Nx1}[l,m]\right] \tag{3.36}$$

and

$$J_r\left[0_{Nx1}[l,m]\right] < J_r[\bar{h}[l,m]] < J_r\left[i_{X_{1,c}X_1}[l,m]\right]. \tag{3.37}$$

The expressions on (3.36) implies that

$$0 \leq \nu_{sd}\left[\overline{h}[l,m]\right] < 1. \tag{3.38}$$

The objective of STFT domain beamforming with the linear array model in (3.8), is to find optimal filters that either minimize $J[\bar{h}[l,m]]$ or minimize $J_d[\bar{h}[l,m]]$ or $J_r[\bar{h}[l,m]]$ subject to some constraint.

3.5 Optimal Filters

In this section, we derive important STFT domain beamformers and their quality measures, using the signal model described in section 3.2.

3.5.1 Maximum SNR

From (3.16) the narrowband oSNR is given by:

$$oSNR\left[\overline{h}[l,m]\right] = \frac{\phi_{X_{1,c}}[l,m]\overline{h}^H[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]\overline{d}^H_{X_{1,c},\overline{x}}[l,m]\overline{h}[l,m]}{\overline{h}^H[l,m]\Phi_{vv}[l,m]\overline{h}[l,m]}. \tag{3.39}$$

Looking at the expression in (3.39) we recognize the generalized Rayleigh quotient, which is known to be maximized by the eigenvector of the matrix

$$M_{\max}[l,m] = \phi_{X_{1,c}}[l,m]\Phi^{-1}_{vv}[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]\overline{d}^H_{X_{1,c},\overline{x}}[l,m] \tag{3.40}$$

denoted by $\overline{h}_{\max}[l,m]$. Since this is a matrix of rank 1, the maximum eigenvalue $\lambda_{\max}[l,m]$ corresponding to the maximum egienvector is given by:

$$\begin{aligned}\lambda_{\max}[l,m] &= \text{trace}\left\{\phi_{X_{1,c}}[l,m]\Phi^{-1}_{vv}[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]\overline{d}^H_{X_{1,c},\overline{x}}[l,m]\right\} \\ &= \phi_{X_{1,c}}[l,m]\overline{d}^H_{X_{1,c},\overline{x}}[l,m]\Phi^{-1}_{vv}[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m].\end{aligned} \tag{3.41}$$

As a result,

$$oSNR\left[\overline{h}_{\max}[l,m]\right] = \lambda_{\max}[l,m] \tag{3.42}$$

and

$$\overline{h}_{\max}[l,m] = \varsigma\left[l,m\right]\Phi_{vv}^{-1}[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m] \tag{3.43}$$

where $\varsigma\left[l,m\right]$ is an arbitrary time frequency-dependent scaling factor other than zero. While this factor has no effect on the narrowband oSNR, it does affect the broadband oSNR and the broadband speech distortion index. In fact, all beamformers derived in the rest of this section are equivalent up to a scaling factor. The different scaling factors depend upon the optimization problem the beamformer solves.

With the inclusion principle [23] for the matrix in (3.40), it can be shown that:

$$\lambda_{\max}^{(N)}\left[l,m\right] \geq \lambda_{\max}^{(N-1)}\left[l,m\right] \geq \cdots \geq \lambda_{\max}^{(2)}\left[l,m\right] \geq \lambda_{\max}^{(1)}\left[l,m\right] = \mathrm{iSNR}\left[l,m\right] \tag{3.44}$$

where $\lambda_{\max}^{(n)}\left[l,m\right]$ is the maximum narrowband oSNR of a microphone array with N sensors. Therefore, the maximum narrowband oSNR can be improved by increasing the number of microphones. If there is only one microphone, the narrowband oSNR cannot be improved as would be expected [14].

3.5.2 Wiener

The Wiener beamformer is found by minimizing the narrowband MSE, $J[\bar{h}[l,m]]$ (3.28):

$$\begin{aligned}\overline{h}_W[l,m] &= \Phi_{yy}^{-1}[l,m]E\left[X_{1,c}[l,m]\overline{x}^*[l,m]\right] \\ &= \phi_{X_{1,c}}[l,m]\Phi_{yy}^{-1}[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m],\end{aligned} \tag{3.45}$$

that we can rewrite as:

$$\begin{aligned}\overline{h}_W[l,m] &= \left[I_N - \Phi_{yy}^{-1}[l,m]\Phi_{vv}[l,m] - \Phi_{yy}^{-1}[l,m]\Phi_{yx_r}[l,m]\right] i_1 \\ &= \overline{h}_{W,n}[l,m] - \Phi_{yy}^{-1}[l,m]\Phi_{yx_r}[l,m]i_1\end{aligned} \tag{3.46}$$

where

$$h_{W,n}[l,m] = \left[I_N - \Phi_{yy}^{-1}[l,m]\Phi_{vv}[l,m]\right] i_1 \tag{3.47}$$

is the multichannel Wiener beamformer for noise reduction only. The challenge in (3.46) is in the estimation of $\Phi_{yx_r}[l,m] = E\left[\overline{y}[l,m]\overline{x}_r^H[l,m]\right]$, since $\overline{x}_r[l,m]$ is not observable. We

can rewrite the general form of the Wiener beamformer in another way that will make comparison to other beamformers easier, and in addition may also make implementation easier. Assuming that the noise is not correlated with the speech signal the observations correlation matrix is given by:

$$\Phi_{yy}[l,m] = \phi_{X_{1,c}}[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]\overline{d}^{H}_{X_{1,c},\overline{x}}[l,m] + \Phi_{vv}[l,m]. \tag{3.48}$$

To calculate the inverse of $\Phi_{yy}[l,m]$ we use Woodbury's identity on the previous expression, giving:

$$\Phi^{-1}_{yy}[l,m] = \Phi^{-1}_{vv}[l,m] - \frac{\phi_{X_{1,c}}[l,m]\Phi^{-1}_{vv}[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]\overline{d}^{H}_{X_{1,c},\overline{x}}[l,m]\Phi^{-1}_{vv}[l,m]}{1+\phi_{X_{1,c}}[l,m]\overline{d}^{H}_{X_{1,c},\overline{x}}[l,m]\Phi^{-1}_{vv}[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]}. \tag{3.49}$$

Substituting (3.49) into (3.45) gives

$$h_W[l,m] = \frac{\phi_{X_{1,c}}[l,m]\Phi^{-1}_{vv}[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]}{1+\phi_{X_{1,c}}[l,m]\overline{d}^{H}_{X_{1,c},\overline{x}}[l,m]\Phi^{-1}_{vv}[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]}. \tag{3.50}$$

From (3.43) and (3.50) it follows that the Wiener beamformer is a scaled version of $\overline{h}_{max}[l,m]$, with

$$\varsigma\,[l,m] = \left[\phi^{-1}_{X_{1,c}}[l,m] + \overline{d}^{H}_{X_{1,c},\overline{x}}[l,m]\Phi^{-1}_{vv}[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]\right]^{-1}. \tag{3.51}$$

Since a scaling factor does not affect the narrowband oSNR, the oSNR of the Wiener beamformer equals the one of $\overline{h}_{max}[l,m]$,

$$\begin{aligned} oSNR[h_W[l,m]] &= \lambda_{\max}[l,m] \\ &= \text{trace}\left[\Phi^{-1}_{vv}[l,m]\Phi_{yy}[l,m]\right] - N. \end{aligned} \tag{3.52}$$

From (3.52) it is clear that $oSNR[h_W[l,m]] \geq iSNR[h_W[l,m]]$. The broadband oSNR of the Wiener beamfoemer is found by applying (3.50) into (3.16)

$$oSNR[h_W[l,:]] = \frac{\sum_{m=0}^{M-1}\phi_{X_{1,c}}[l,m]\frac{\lambda^2_{\max}[l,m]}{[1+\lambda_{\max}[l,m]]^2}}{\sum_{m=0}^{M-1}\phi_{X_{1,c}}[l,m]\frac{\lambda_{\max}[l,m]}{[1+\lambda_{\max}[l,m]]^2}}. \tag{3.53}$$

The narrowband and broadband speech distortion indices are found by using (3.50) in (3.22), and (3.23)

$$\nu_{sd}[l,m] = \frac{1}{[1+\lambda_{\max}[l,m]]^2}, \tag{3.54}$$

$$\nu_{sd}[l,:] = \frac{\sum_{m=0}^{M-1}\phi_{X_{1,c}}[l,m]\,[1+\lambda_{\max}[l,m]]^{-2}}{\sum_{m=0}^{M-1}\phi_{X_{1,c}}[l,m]}. \tag{3.55}$$

3.5.3 Minimum Variance Distortion-less Response (MVDR)

The well-known minimum variance distortion-less response (MVDR) beamformer proposed by Capon [5] is derived by minimizing the narrowband MSE of the residual noise, $J_r[\overline{h}[l,m]]$, with the constraint that the desired signal is not distorted. Mathematically, this is equivalent to:

$$\min_{\overline{h}[l,m]} \left\{ \overline{h}^H[l,m]\Phi_{vv}[l,m]\overline{h}[l,m] \right\} \quad \text{S.T.} \quad \overline{h}^H[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m] = 1. \tag{3.56}$$

The solution for this minimization problem is known to be

$$\overline{h}_{MVDR}[l,m] = \frac{\Phi_{vv}^{-1}[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]}{\overline{d}^H_{X_{1,c},\overline{x}}[l,m]\Phi_{vv}^{-1}[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]}, \tag{3.57}$$

and can rewritten as

$$\begin{aligned}\overline{h}_{MVDR}[l,m] &= \frac{1}{d_{X_{1,c}X_1}[l,m]} \cdot \frac{\Phi_{vv}^{-1}[l,m]\Phi_{\mathbf{yy}}[l,m] - I_N}{\text{trace}\left[\Phi_{vv}^{-1}[l,m]\Phi_{\mathbf{yy}}[l,m]\right] - N} i_1 \\ &= \frac{1}{d_{X_{1,c}X_1}[l,m]} \cdot \overline{h}_{MVDR,n}[l,m]\end{aligned} \tag{3.58}$$

where

$$\overline{h}_{MVDR,n}[l,m] = \frac{\Phi_{vv}^{-1}[l,m]\Phi_{yy}[l,m] - I_N}{\text{trace}\left[\Phi_{vv}^{-1}[l,m]\Phi_{yy}[l,m]\right] - N} i_1 \tag{3.59}$$

is the MVDR filter for noise reduction only.

The MVDR beamformer can also be seen as a scaled version of $\overline{h}_{\max}[l,m]$, with $\varsigma[l,m] = \frac{\phi_{X_{1,c}}[l,m]}{\lambda_{\max}[l,m]}$. Since both the MVDR and Wiener beamformers are a scaled version of $\overline{h}_{max}[l,m]$, it is clear they are related by a scaling factor as well i.e.

$$\overline{h}_W[l,m] = C_W[l,m]\overline{h}_{MVDR}[l,m] \tag{3.60}$$

where

$$C_W[l,m] = \overline{h}_W[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]. \tag{3.61}$$

While the narrowband oSNRs of the Wiener and MVDR beamformers are strictly equal, their broadband oSNRs are not due to this scaling factor. This difference is significant, as speech signals are broadband in nature, making the broadband oSNR a meaningful measure when analyzing the beamformers noise reduction capabilities.

The broadband oSNR of the MVDR is given by:

$$oSNR\left[\overline{h}_{MVDR}[l,:]\right] = \frac{\sum_{m=0}^{M-1}\phi_{X_{1,c}}[l,m]}{\sum_{m=0}^{M-1}\phi_{X_{1,c}}[l,m]\lambda_{\max}^{-1}[l,m]}, \tag{3.62}$$

and (from property 3.2)

$$oSNR\left[\overline{h}_{MVDR}[l,:]\right] \leq oSNR\left[\overline{h}_{W}[l,:]\right]. \tag{3.63}$$

The MVDR beamformer rejects the maximum level of noise allowable without distorting the desired signal. From the constraint on (3.56) it is clear that the MVDR speech distortion index is always 0, clearly this means the broadband speech distortion index of the MVDR beamformer also equals to 0.

3.5.4 Tradeoff Beamformer

As we have learned from the two previous sub-sections, the Wiener and MVDR beamformers do not have much flexibility since we never know in advance by how much the narrowband oSNR will be improved. However, in many practical applications, we wish to control the compromise between noise reduction, reverberation suppression and speech distortion. The best known way to do this is via the so-called tradeoff beamformer [14]. In the tradeoff approach, we minimize the narrowband speech distortion with the constraint that the residual noise level is equal to a value smaller than the level of the original noise. Mathematically, this is equivalent to:

$$\min_{\overline{h}[l,m]} \left\{J_d[\overline{h}[l,m]]\right\} \quad \text{S.T.} \quad J_r[\overline{h}[l,m]] = \beta J_r\left[i_{X_{1,c}X_1}[l,m]\right] \tag{3.64}$$

where $\beta \in (0,1)$ to insure that we get some noise reduction. By using a Lagrange multiplier, $\mu > 0$, to adjoin the constraint to the cost function, we formulate the tradeoff filter:

$$\begin{aligned}
\overline{h}_{T,\mu}[l,m] &= \phi_{X_{1,c}}[l,m]\left[\Phi_{xx}[l,m] + \mu\Phi_{vv}[l,m]\right]^{-1}\overline{d}_{X_{1,c},\overline{x}}[l,m] \\
&= \frac{\phi_{X_{1,c}}[l,m]\Phi_{vv}^{-1}[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]}{\mu + \phi_{X_{1,c}}[l,m]\overline{d}^{H}_{X_{1,c},\overline{x}}[l,m]\Phi_{vv}^{-1}[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]} \\
&= \frac{1}{d_{X_{1,c}X_1}[l,m]} \cdot \frac{\Phi_{vv}^{-1}[l,m]\Phi_{\mathbf{yy}}[l,m] - I_N}{\text{trace}\left[\Phi_{vv}^{-1}[l,m]\Phi_{\mathbf{yy}}[l,m]\right] + \mu - N} i_1 \\
&= \frac{1}{d_{X_{1,c}X_1}[l,m]}\overline{h}_{n,T,\mu}[l,m]
\end{aligned} \tag{3.65}$$

where

$$\overline{h}_{n,T,\mu}[l,m] = \frac{\Phi_{vv}^{-1}[l,m]\Phi_{\mathbf{yy}}[l,m] - I_N}{\text{trace}\left[\Phi_{vv}^{-1}[l,m]\Phi_{\mathbf{yy}}[l,m]\right] + \mu - N} i_1 \tag{3.66}$$

is the multichannel tradeoff filter for noise reduction only. The Lagrange multiplier μ, satisfies

$$J_r\left[\overline{h}_{T,\mu}[l,m]\right] = \beta J_r\left[i_{X_{1,c}X_1}[l,m]\right]. \tag{3.67}$$

In practice it is hard to determine the optimal μ, however when chosen in an ad-hoc way the following statements hold:

- For $\mu = 1$ the tradeoff beamformer equals the Wiener beamformer $\overline{h}_{T,1}[l,m] = \overline{h}_W[l,m]$.
- For $\mu = 0$ the tradeoff beamformer equals the MVDR beamformer $\overline{h}_{T,0}[l,m] = \overline{h}_{MVDR}[l,m]$.
- For $\mu > 1$ the tradeoff beamformer will reduce residual noise better, though at the expense of higher speech distortion.
- For $\mu < 1$ the tradeoff beamformer will inflict less speech distortion, though at the expense of higher residual noise.

The tradeoff beamformer is a scaled version of $\overline{h}_{max}[l,m]$, with

$$\varsigma\left[l,m\right] = \left[\mu\phi^{-1}_{X_{1,c}}[l,m] + \overline{d}^H_{X_{1,c},\overline{x}}[l,m]\Phi^{-1}_{vv}[l,m]\overline{d}_{X_{1,c},\overline{x}}[l,m]\right]^{-1}. \tag{3.68}$$

As a result, the narrowband oSNR of the tradeoff beamformer is independent of μ and is identical to the narrowband output SNR of the $\overline{h}_{max}[l,m]$, i.e.,

$$oSNR\left[\overline{h}_{T,\mu}[l,m]\right] = oSNR\left[\overline{h}_{\max}[l,m]\right],\ \forall\mu \geq 0. \tag{3.69}$$

By applying (3.41) to (3.65) we deduce the broadband oSNR

$$oSNR[\overline{h}_{T,\mu}[l,:]] = \frac{\sum_{m=0}^{M-1}\phi_{X_{1,c}}[l,m]\frac{\lambda^2_{\max}[l,m]}{[\mu+\lambda_{\max}[l,m]]^2}}{\sum_{m=0}^{M-1}\phi_{X_{1,c}}[l,m]\frac{\lambda_{\max}[l,m]}{[\mu+\lambda_{\max}[l,m]]^2}}. \tag{3.70}$$

The narrowband speech distortion index in this case is given by:

$$\nu_{sd}\left[\overline{h}_{T,\mu}[l,m]\right] = \left[\frac{\mu}{\mu+\lambda_{\max}[l,m]}\right]^2, \tag{3.71}$$

and the broadband speech distortion index is given by:

$$\nu_{sd}[\overline{h}_{T,\mu}[l,:]] = \frac{\sum_{m=0}^{M-1}\phi_{X_{1,c}}[l,m]\frac{\mu^2}{[\mu+\lambda_{\max}[l,m]]^2}}{\sum_{m=0}^{M-1}\phi_{X_{1,c}}[l,m]}. \tag{3.72}$$

Property 3.2

The broadband oSNR of the tradeoff beamformer is an increasing function of the parameter μ.

Proof. By replacing the integrals by sums, in the proof given in [28] , it is easy to show that:

$$\frac{d}{d\mu} oSNR[\overline{h}_{T,\mu}[l,:]] \geq 0 \tag{3.73}$$

which completes the proof. □

Property 3.3 The broadband speech distortion index of the tradeoff beamformer is an increasing function of the parameter μ.

Proof. The derivative of (3.72) is given by:

$$\frac{d}{d\mu}\nu_{sd}[\overline{h}_{T,\mu}[l,:] = \frac{\sum_{m=0}^{M-1} \phi_{X_{1,c}}[l,m] \frac{2\lambda_{\max}[l,m]\mu}{[\mu+\lambda_{\max}[l,m]]^3}}{\sum_{m=0}^{M-1} \phi_{X_{1,c}}[l,m]}. \tag{3.74}$$

Since both $\lambda_{max}[l,m]$ and μ, are non negative scalars we have

$$\frac{d}{d\mu}\nu_{sd}[\overline{h}_{T,\mu}[l,:] \geq 0 \tag{3.75}$$

proving that the broadband speech distortion index increases with μ. □

From (3.71) it is clear that

$$0 \leq \nu_{sd}\left[\overline{h}_{T,\mu}[l,m]\right] \leq 1, \ \forall \mu \geq 0. \tag{3.76}$$

Therefore, as μ increases, the broadband oSNR increases at the price of more distortion to the desired signal.

We note yet another interesting property.

Property 3.4 With the tradeoff beamformer, the broadband oSNR is always greater than or equal to the iSNR, i.e., oSNR $\left[\overline{h}_{T,\mu}[l,:]\right] \geq$ iSNR $[l,:]$, $\forall \mu \geq 0$.

Proof. From (3.44) we know that the

$$\lambda_{\max}[l,m] \geq iSNR\,[l,m] \tag{3.77}$$

implying that

$$\sum_{m=0}^{M-1} \frac{\phi_{V_1}[l,m] iSNR\,[l,m]}{\left|\overline{d}_{X_{1,c},\overline{x}}[l,m]\right|^2 \lambda_{\max}[l,m]} \leq \sum_{m=0}^{M-1} \frac{\phi_{V_1}[l,m]}{\left|\overline{d}_{X_{1,c},\overline{x}}[l,m]\right|^2}. \tag{3.78}$$

Hence,

$$\begin{aligned} oSNR[\overline{h}_{T,0}[l,:]] &= \frac{\sum_{m=0}^{M-1} \phi_{X_{1,c}}[l,m]}{\sum_{m=0}^{M-1} \frac{\phi_{V_1}[l,m] iSNR[l,m]}{\left|\overline{d}_{X_{1,c},\overline{x}}[l,m]\right|^2 \lambda_{\max}[l,m]}} \\ &\geq \frac{\sum_{m=0}^{M-1} \phi_{X_{1,c}}[l,m]}{\sum_{m=0}^{M-1} \frac{\phi_{V_1}[l,m]}{\left|\overline{d}_{X_{1,c},\overline{x}}[l,m]\right|^2}} = iSNR\,[l,:]\,. \end{aligned} \tag{3.79}$$

From property 3.2 it is clear that

$$oSNR[\overline{h}_{T,\mu}[l,:]] \geq oSNR[\overline{h}_{T,0}[l,:]]\,\forall \mu \geq 0 \tag{3.80}$$

therefore

$$oSNR[\overline{h}_{T,\mu}[l,:]] \geq iSNR\,[l]\ \forall \mu \geq 0 \tag{3.81}$$

which is what we wanted to prove . □

Using property 3.2 and 3.3, we deduce that for $\mu \geq 1$,

$$\begin{aligned} &\text{oSNR}\left[\overline{h}_{MVDR}[l,m]\right] \leq \text{oSNR}\left[\overline{h}_{\text{W}}[l,m]\right] \leq \text{oSNR}\left[\overline{h}_{T,\mu}[l,m]\right] \\ &\nu_{sd}\left[\overline{h}_{MVDR}[l,m]\right] \leq \nu_{sd}\left[\overline{h}_{W}[l,m]\right] \leq \nu_{sd}\left[\overline{h}_{T,\mu}[l,m]\right] \end{aligned} \tag{3.82}$$

and for $0 \leq \mu \leq 1$,

$$\begin{aligned} &\text{oSNR}\left[\overline{h}_{MVDR}[l,m]\right] \leq \text{oSNR}\left[\overline{h}_{T,\mu}[l,m]\right] \leq \text{oSNR}\left[\overline{h}_{\text{W}}[l,m]\right] \\ &\nu_{sd}\left[\overline{h}_{MVDR}[l,m]\right] \leq \nu_{sd}\left[\overline{h}_{T,\mu}[l,m]\right] \leq \nu_{sd}\left[\overline{h}_{W}[l,m]\right]. \end{aligned} \tag{3.83}$$

The fact that this beamformer is somewhat "adjustable" can be used to maximize its performance by adaptively changing μ as function of the speech presence probability $p\,[l,m]$. In doing so, we can use the tradeoff beamformer when speech is absence, thus increasing

the overall noise reduction, and the MVDR beamformer when speech is present which will ensure less distortion in the speech parts of the signal. To achieve this we propose the following function, $\mu[l,m] = [1 - p[l,m]]\,\mu_0$ where $\mu_0 > 0$. When $p[l,m]$ is close to 1 (i.e. speech is present) $\mu[l,m]$ will decrease to zero, making the tradeoff beamformer converge toward the MVDR beamformer. On the other hand when $p[l,m]$ decreases to 0, $\mu[l,m]$ will grow, increasing the oSNR at the expense of high signal distortion.

3.6 Relative Transfer Function Estimation

In this section, we propose an estimator for the full to early RTF vector $\overline{d}_{X_{1,c},\overline{x}}[l,m]$. The estimator we proposed is based on the work done by Cohen [15], in which a robust estimator for the classical RTF is introduced.

Assuming that the presence of the desired speech signal in the STFT domain is uncertain, we employ the speech presence probability to derive an estimator based only on the sub intervals that contains speech. Using equation (3.5) we find the cross correlation between the reference microphone and the j-th microphone

$$\Phi_{Y_1,Y_j}[l,m] = \phi_{X_{1,c}}[l,m] d^*_{X_{1,c},X_1}[m] d_{X_{1,c},X_j}[m] + \Phi_{V_1,V_j}[l,m]. \tag{3.84}$$

It is easy to see that

$$\begin{aligned} \phi_{X_{1,c}}[l,m] d^*_{X_{1,c},X_1}[m] &= E\left\{X_{1,c}[l,m] X^*_{1,c}[l,m] d^*_{X_{1,c},X_1}[m]\right\} \\ &= E\left\{X_{1,c}[l,m] X^*_1[l,m]\right\} = \Phi_{X_{1,c},X_1}[l,m]. \end{aligned} \tag{3.85}$$

Writing (3.84) in terms of PSD estimates and using (3.85), we obtain

$$\hat{\Phi}_{Y_1,Y_j}[l,m] - \hat{\Phi}_{V_1,V_j}[l,m] = \hat{\Phi}_{X_1,X_{1,c}}[l,m] \hat{d}_{X_{1,c},X_j}[m] + \bar{\epsilon}_j[l,m] \tag{3.86}$$

where $\hat{d}_{X_{1,c},X_j}[l,m]$ is an estimator for $d_{X_{1,c},X_j}[l,m]$, the j-th element in $\overline{d}_{X_{1,c},\overline{x}}[l,m]$ and $\bar{\epsilon}_j$ is the estimation error. Given that in every frequency bin m there are L time frames, the expression in (3.86) can be stacked into a vector, defining the following set of L

over-determined equations:

$$\Psi_{1,j}[m] = \begin{bmatrix} \hat{\Phi}_{Y_1,Y_j}[1,m] - \hat{\Phi}_{V_1,V_j}[1,m] \\ \hat{\Phi}_{Y_1,Y_j}[2,m] - \hat{\Phi}_{V_1,V_j}[2,m] \\ . \\ . \\ . \\ \hat{\Phi}_{Y_1,Y_j}[L,m] - \hat{\Phi}_{V_1,V_j}[L,m] \end{bmatrix}$$

$$= \begin{bmatrix} \hat{\Phi}_{X_1,X_{1,c}}[1,m] \\ \hat{\Phi}_{X_1,X_{1,c}}[2,m] \\ . \\ . \\ . \\ \hat{\Phi}_{X_1,X_{1,c}}[L,m] \end{bmatrix} \hat{d}_{X_{1,c},X_j}[m] + \begin{bmatrix} \bar{\epsilon}_j[1,m] \\ \bar{\epsilon}_j[2,m] \\ . \\ . \\ . \\ \bar{\epsilon}_j[L,m] \end{bmatrix}. \tag{3.87}$$

Since the RTF $\hat{d}_{X_{1,c},X_j}$ describes the correlation between the desired early signal and the signal at jth microphone, the optimization criterion for the estimation of $\hat{d}_{X_{1,c},X_j}$ has to take under account only time frames in which the desired signal is present. Time frames containing only noise, convey no information regarding the direction of the desired signal, and will severely degrade the estimation process. To avoid this degradation we formulate the following estimation process:

Let $I[l,m]$ denote an indicator function for signal presence i.e $I[l,m] = 1$ if $X_{1,c}[l,m] \neq 0$, and $I[l,m] = 0$ otherwise, and let $\underline{I}[m]$ be the following matrix $\underline{I}[m] = \text{diag}\left(\begin{bmatrix} I[1,m] & I[2,m] & \dots & , & I[L,m] \end{bmatrix}\right)$. Then the Weighted Least Square estimate of $\hat{d}_{X_{1,c},X_j}$ is obtained by [16]:

$$\begin{aligned} \hat{d}_{X_{1,c},X_j}[m] &= \min_d \left\{ [\underline{I}\overline{\epsilon_j}]^H W [\underline{I}\overline{\epsilon_j}] \right\} \\ &= \left[\hat{\Phi}^T_{X_1,X_{1,c}} \underline{IWI} \hat{\Phi}_{X_1,X_{1,c}} \right]^{-1} \\ &\times \hat{\Phi}^T_{X_1,X_{1,c}} \underline{IWI} \Psi_{1,j} \end{aligned} \tag{3.88}$$

where the frequency argument m was omitted for simplicity.

An exact indicator function, $I[l,m]$, for the time frames containing speech, cannot be obtained. However it can estimated, applying a static threshold on the speech presence

probability $p\,[l,m]$ [18],

$$\hat{I}\,[l,m] = \begin{cases} 1 & \text{if } p\,[l,m] \geq p_0 \\ 0 & \text{otherwise} \end{cases} \tag{3.89}$$

where $p_0 \in \big[0,1\big)$ is a predefined threshold parameter, controlling the tradeoff between false alarms and true detections. A small value of p_0 will allow more STFT frames to be included in the estimation process, however when p_0 is too small, the probability of a false alarm increases, which in turn reduces the robustness of the estimation process.

By applying an online normalized LMS solution on (3.88) as described in [15, 29], we derive the following recursive estimator :

$$\begin{aligned} \hat{d}_{X_{1,c},X_j}[l,m] &= \hat{d}_{X_{1,c},X_j}[l-1,m] - \frac{\mu_{RTF}}{\hat{\Phi}^2{}_{X_1,X_{1,c}}[l,m]} \frac{\partial}{\partial \hat{d}^*_{X_{1,c},X_j}} \\ &\times \left(\hat{I}\,[l,m] \left| \hat{\Phi}_{Y_1,Y_j}[l,m] - \hat{\Phi}_{V_1,V_j}[l,m] - \hat{d}_{X_{1,c},X_j} \hat{\Phi}_{X_1,X_{1,c}}[l,m] \right|^2 \right) \Big|_{d_{X_{1,c},X_j} = \hat{d}_{X_{1,c},X_j}[l-1,k]} \\ &= \hat{d}_{X_{1,c},X_j}[l-1,m] - \mu_{RTF} \frac{\hat{I}\,[l,m]\,\hat{\epsilon}_j[l,m]}{\hat{\Phi}_{X_1,X_{1,c}}[l,m]} \end{aligned} \tag{3.90}$$

where

$$\begin{aligned} \hat{\epsilon}_j[l,m] &= \hat{\Phi}_{Y_1,Y_j}[l,m] - \hat{\Phi}_{V_1,V_j}[l,m] \\ &\quad - \hat{\Phi}_{X_1,X_{1,c}}[l,m] \hat{d}_{X_{1,c},X_j}\,[l-1,m] \end{aligned} \tag{3.91}$$

is the estimation error, and μ_{RTF} is a parameter describing the LMS algorithm update step size. Estimation of the cross PSD $\Phi_{X_1,X_{1,c}}[l,m]$ will be discussed in the next sub section.

3.7 Implementation

The construction of the optimal beamformers described in Section 3.5 and the RTF estimation described in Section 3.6, requires an estimation of $\Phi_{vv}[l,m]$, $\Phi_{yy}[l,m]$, $\hat{\phi}_{X_{1,c}}[l,m]$ and $\Phi_{X_1,X_{1,c}}[l,m]$. Methods and algorithms used to acquire these elements will be outlined in the following section.

The correlation matrix $\Phi_{yy}[l,m]$, is estimated by applying a first order recursive smoothing approach (Periodegram estimation) to the cross correlation of the observed

signals

$$\hat{\Phi}_{yy}[l,m] = \alpha_y \hat{\Phi}_{yy}[l-1,m] + (1-\alpha_y)\max(\overline{y}^H[l,m]\overline{y}[l,m], \Phi_{y,\min}), \tag{3.92}$$

where the smoothing parameter $\alpha_y \in [0,1)$ determines the number of cross-correlations that are averaged. Typically, speech periodograms are recursively smoothed with an equivalent rectangular window of 0.2 sec length, which represents a good compromise between tracking the speech spectral variations and smoothing the noise. Therefore, for a sampling rate of $16\,KHz$, a STFT window length of 512 samples and a frame update step of 128 samples, we use $\alpha_y \approx \frac{(0.2\cdot 16K/128-1)}{0.2\cdot 16K/128+1} = 0.85$.

To estimate the noise correlation matrix $\Phi_{vv}[l,m]$, we use the MRCA algorithm [17], and recursively average the cross power values of the interfering signal using a time frequency dependent smoothing factor $\tilde{\alpha}_v[l,m]$

$$\hat{\Phi}_{vv}[l,m] = \tilde{\alpha}_v[l,m]\hat{\Phi}_{vv}[l-1,m] + \beta(1-\tilde{\alpha}_v[l,m])\max(\overline{y}[l,m]\overline{y}^H[l,m], \Phi_{vv,\min}) \tag{3.93}$$

$\beta \geq 1$ is a factor that compensates the bias when the desired signal is absent, and $\Phi_{v,min}$ is a low level threshold parameter to reduce musical noise. The smoothing parameter is determined using the reference signal presence probability $p\,[l,m]$, and α_v is a constant that represents its minimal value of

$$\tilde{\alpha}_v[l,m] = \alpha_v + (1-\alpha_v)\,p\,[l,m]\,. \tag{3.94}$$

When the desired signal is present, the value of $\tilde{\alpha}_v$ is close to 1. This helps prevent the noise correlation matrix estimate from increasing as a result of speech signal components. It decreases linearly with the probability of speech presence to allow a faster update of the noise estimate. The value α_v is set to compromise between the response rate to abrupt changes in noise statistics and the variance of the noise estimate. In case of non-stationary noise, α_v is typically set to 0.85.

To obtain an estimate for $\Phi_{Y_1,X_{1,c}}[l,m]$, the cross PSD between the reverberant signal at the reference channel to the desired signal, we use the OM-LSA algorithm described in Sub-section 2.1.2 and calculate the log-spectral amplitude gain function:

$$F_{1,c}[l,m] = \frac{\xi[l,m]}{1+\xi[l,m]}\text{expint}\left\{\frac{\xi[l,m]\gamma[l,m]}{1+\xi[l,m]}\right\}, \tag{3.95}$$

where

$$\xi_{1,c}[l,m] = \frac{\hat{\phi}_{X_{1,c}[l,m]}}{\hat{\phi}_{X_{1,l}}[l,m] + \hat{\phi}_{V_1,V_1}[l,m]}$$
$$\gamma_{1,c}[l,m] = \frac{|Y_1[l,m]|^2}{\hat{\phi}_{X_{1,l}}[l,m] + \hat{\phi}_{V_1,V_1}[l,m]} \tag{3.96}$$

are the *a priori* and *a posteriori* SIR, and $\hat{\phi}_{X_{1,l}}[l,m]$ is an estimate for the PSD of the late reverberant part of the signal calculated using (2.35). In a similar way we obtain $F_1[l,m]$ the LSA function that accounts only for noise reduction i.e. it's *a priori* and *a posteriori* SIRs are given by

$$\xi_1[l,m] = \frac{\hat{\phi}_{X_{1,c}[l,m]}}{\hat{\phi}_{V_1,V_1}[l,m]},$$
$$\gamma_1[l,m] = \frac{|Y_1[l,m]|^2}{\hat{\phi}_{V_1,V_1}[l,m]}.$$

Once the LSA functions are obtained we use them to recursively smooth the estimated STFT coefficients of the desired cross PSD

$$\hat{\Phi}_{X_1,X_{1,c}}[l,m] = \alpha_y \hat{\Phi}_{X_1,X_{1,c}}[l-1,m] + (1-\alpha_y)\left[|Y_1[l,m]|\right]^2 \left[F_1[l,m]F^*_{1,c}[l,m]\right]. \tag{3.97}$$

From the signal model in (3.5) the PSD of the desired signal can be written as:

$$\phi_{X_{1,c}}[l,m] = \phi_{X_1}[l,m]\left|d_{X_{1,c},X_1}[l,m]\right|^2 \tag{3.98}$$

where $\hat{\phi}_{X_1}[l,m]$ is the PSD of the clean but reverberant signal at the reference channel. An estimate for this PSD is given by the first element clean but reverberant cross correlation matrix $\hat{\Phi}_{xx}[l,m]$ defined as:

$$\hat{\Phi}_{xx}[l,m] = \hat{\Phi}_{yy}[l,m] - \hat{\Phi}_{vv}[l,m]. \tag{3.99}$$

Using (3.98) and (3.99) we deduce the following estimator for the clean early signal PSD

$$\hat{\phi}_{X_{1,c}}[l,m] = \alpha_y \hat{\phi}_{X_{1,c}}[l-1,m] + (1-\alpha_y)\left(\left\{\hat{\Phi}_{xx}[l,m]\right\}_{1,1}\left|\hat{d}_{X_{1,c},X_1}[l,m]\right|^2\right) \tag{3.100}$$

where $\left\{\hat{\Phi}_{xx}[l,m]\right\}_{1,1}$ is the element indexed by $(1,1)$ in the matrix $\hat{\Phi}_{xx}[l,m]$.

Another element needed for the construction of the suggested beamformers is the inverse of the noise correlation matrix $\hat{\Phi}_{vv}[l,m]$. A straight forward inversion will not

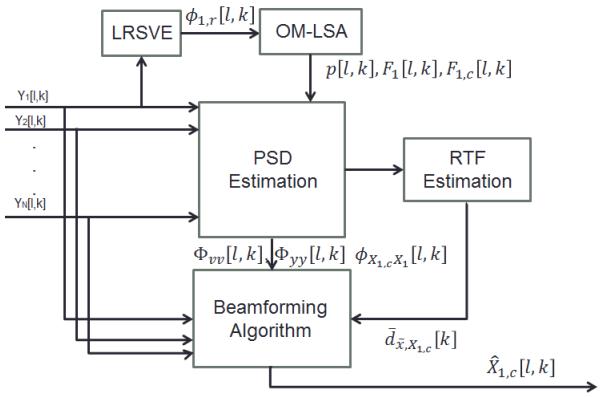


Figure 3.1: System flow.

suffices, as the correlation matrix tends to be ill-posed. Therefore regularization is needed prior to the inversion, we use the regularization method introduced in [27]

$$\tilde{\Phi}_{vv}[l,m] = \hat{\Phi}_{vv}[l,m] + \frac{\delta}{N} trace\left\{\tilde{\Phi}_{vv}[l,m]\right\} I_{\mathrm{NxN}} \tag{3.101}$$

where δ is a small regularization constant. High values of δ will ensure robust inversion with low accuracy, low values of δ will increase the inversion accuracy while risking in inverting ill posed matrices. In our experiments we set $\delta = 0.02$ which ensures robust inversion with reasonable numeric accuracy. The predefined parameters used in our algorithm are summed in Table. 3.1, and the algorithm flow is described in Fig. 3.1.

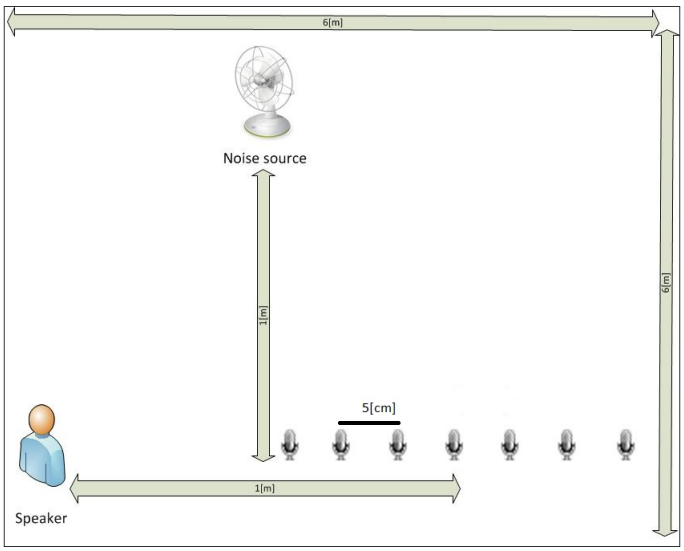


Figure 3.2: Directional noise field experimental setup.

Algorithm 3.1 Pseudo code

Initialize variables on the first frame for all frequency bins m:

$\hat{\Phi}_{yy}[0,m] = \hat{\Phi}_{vv}[0,m] = \bar{y}[0,m]\,\bar{y}^T[0,m]$;

$\hat{\phi}_{X_{1,c}}[0,m] = \hat{\Phi}_{X_1,X_{1,c}}[0,m] = 0$;

$\hat{d}_{X_{1,c},\bar{x}}[0,m] = \frac{\hat{\Phi}_{y,1:N}[0,m]}{\hat{\Phi}_{y,1:1}[0,m]}$;

For all time frames l

For all frequency bins m

- Compute the recursively averaged Correlation matrix $\Phi_{yy}[l,m]$ by using (3.92).
- Compute the reference signal presence probability $p[l,m]$ using [18], and the adaptive smoothing parameter $\tilde{\alpha}_v[l,m]$.
- Compute the noise correlation matrix $\hat{\Phi}_{vv}[l,m]$ using (3.93).
- Compute the log spectral function $F_1[l,m]$ using [18].
- Compute the LRSVE $\hat{\phi}_{X_{1,l}}[l,m]$ using (2.35).
- Compute the reverberation adapted log-spectral function $F_{1,c}[l,m]$ using [19, chapter 3].
- Compute the cross PSD $\hat{\Phi}_{X_1,X_{1,c}}[l,m]$ using (3.97).
- Compute the estimation error $\hat{\epsilon}_j[l,m]$ using (3.91), and the estimated indication function $\hat{I}[l,m]$ (3.89).
- Compute the estimated RTF vector $\hat{d}_{X_{1,c},\overline{x}}[l,m]$ using (3.90).
- Compute the PSD of the desired signal $\hat{\phi}_{X_{1,c}}[l,m]$ using (3.100).
- Regulate the noise correlation matrix using (3.101), and inverse it.
- Compute one of the suggested beamformers using (3.57), (3.50) , or (3.65).
- Apply the beamformer on the observation vector $\overline{y}[l,m]$, to achieve the estimated early clean STFT coefficient $Z[l,m]$.

3.8 Experimental Study

In this section, we present an experimental analysis of the proposed STFT domain beamformers. Using the performance measures introduced in subsection 3.4 we evaluate the suggested beamformers performance and compare it with that of the well-celebrated TF-GSC beamformer (Sub-section 2.2). Our evaluation is composed of two parts: in the first part, we measure the noise reduction capabilities of each beamformer and the amount of distortion it imposes; in the second part, we test the beamformers' reverberation suppression capabilities.

The results of our simulations are presented in terms of broadband performance measures. The beamformers are computed and applied on a frame-by-frame basis, and the performance measures are evaluated per frame and subsequently averaged over all frames. The SNRs are averaged in the logarithm domain while the rest of the measures are averaged in the linear domain.

3.8.1 Experimental Setup

In the following experiments we use Habet's simulator [21] to simulate the RIRs of a 7 microphone narrow band array. The RIRs are simulated in a rectangular room, 6 m wide by 6 m long, and 4 m high. The microphones are placed along a straight line ranging from [3.8 m, 3 m, 2 m] to [4.15 m, 3 m, 2 m], with inner microphone distance of 5 cm. A speech source is placed at [3 m ,3 m, 2 m], 1 m away from the reference microphone (forth in the array). The speech source is composed out of 20 different recorded speech signals (10 men and 10 women) taken from the TIMIT database [22] sampled at 16 KHz. We simulate two different noise fields, the first one is a diffused noise field simulated using the method proposed in [35]. The second one is a directional noise field, it is created using RIRs that describe the path from a noise source placed 1 m from the array at [4 m, 4 m, 2 m], to each microphones Fig. 3.2. To simulate the microphones' thermal noise, a computer-generated white Gaussian noise with SNR of 20 dB is added to each microphone.

Table 3.1: Predefined Parameters.

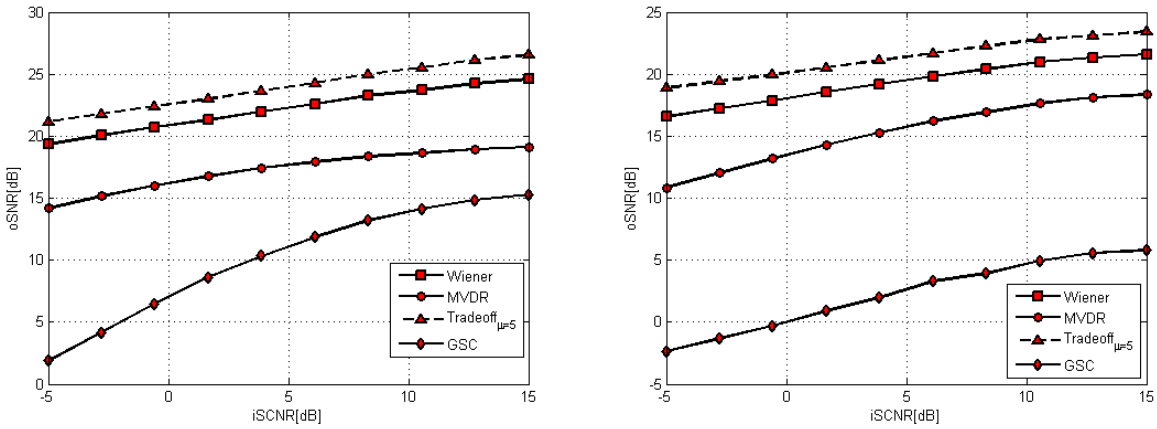
Algorithm section	Needed Parameters	
RTF Estimation:	μ_{RTF}=0.65	p_0=0.75
Signal Correlation Matrix:	α_y=0.85	$\Phi_{y,min}$ =1.4e-6
Noise Correlation Matrix:	α_v=0.85	$\Phi_{v,min}$1.4e-7
	β=1.05	
OM-LSA:	α=0.92	F_{min}=-20 dB
LRSVE:	R=128	L_e=356 taps
Matrix Regulator:	δ=0.02	

In our analysis we compare the performances of the proposed beamformers to the one of the TF-GSC beamformer [10]. The TF-GSC beamformer is implemented using the same RTF estimator as the rest of the discussed beamformers. The blocking matrix $\widetilde{H}_{BM}$ we used in our implementation of the TF-GSC , is aimed at eliminating the desired clean early signal, and constructing the reference noise signals. It is given by $\widetilde{H}_{BM} = H_{BM}\hat{d}^*_{X_{1,c},X_1}$, where H_{BM} is the blocking matrix introduced in [10].

3.8.2 Influence of Noise Field iSNR

In this subsection, we investigate the beamformers' performance with regard to change in the noise field input SNR. We consider two different noise fields one diffused and the other directional, and change their input signal to coherent noise ratio (iSCNR). The iSCNR is measured with respect to the signal in the reference channel and accounts only for the changes in the coherent noise field (i.e. not including thermal noise). The reverberation time is set to $T_{60} = 0.8\,\text{sec}$.

The broadband output SNR and speech distortion curves for the directional and diffuse noise fields, are shown in Fig. 3.3 (a)-(b), and Fig. 3.4 (a)-(b). Each curve was obtained by averaging over 20 different speaker signals that were processed by the suggested beamformers. The curves verify our claim that regarding oSNR, the Weiner and Tradeoff beamformers perform better, however the speech distortion they produce is larger than that produced by MVDR beamformer. As expected we achieve higher oSNR when simulating a directional noise field. In that case the desired signal and the noise are com-



(a) Directional noise field. (b) Diffused noise field.

Figure 3.3: Broadband output SNR as a function of the input signal to coherent noise ratio.

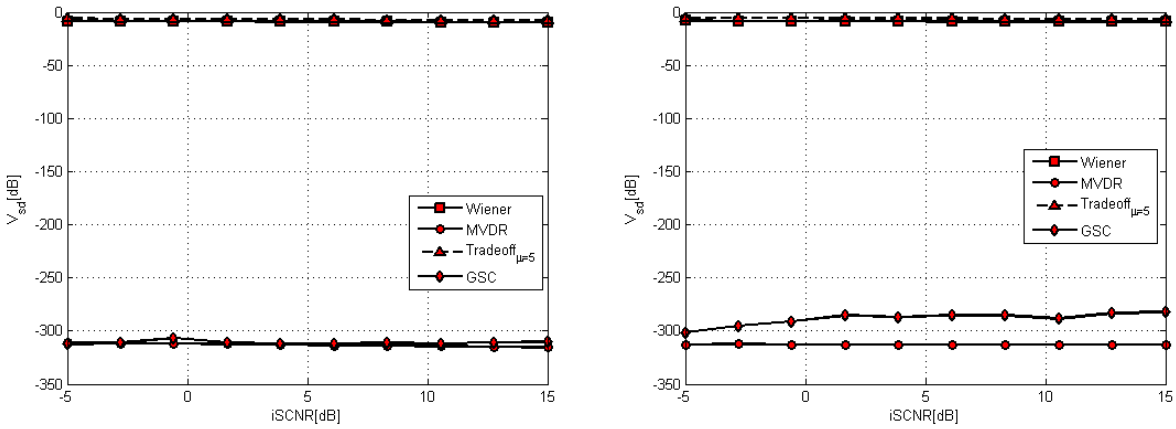
ing from a different direction, unlike the diffused noise field case, in which noise arrives to the array from all directions (including the beamformers' look direction). We note that the TF-GSC speech distortion curve is close to that of the MVDR beamformer, but its oSNR curve is lower. This behavior is to be expected, as the FBF used in the TF-GSC implementation manages to sufficiently enforce the no distortion constraint. However the ANC block doesn't manage to converge fast enough to an optimal solution that maximizes the narrowband oSNR. Hence only a semi-optimal solution is applied to the BM output signals, resulting in lower oSNR.

3.8.3 Influence of Reverberation Time

In this subsection, we investigate the beamformers' reverberation suppression capabilities. We examine the effect of the room reverberation time T_{60}, on the oSNR and Log Spectral Distortion (LSD) measures. In our experiment we simulate both diffuse and directional noise fields, and change T_{60} while keeping a constant iSCNR of $5\,dB$.

The LSD is a subjective measure, it is defined as the Lp norm of the difference between the STFT transforms of the early clean signal $X_{1,c}[l,m]$, and $Z[l,m]$ the beamformer's output. For a time frame l the narrowband LSD is given by:

$$LSD[l] = \frac{1}{M}\sum_{m=0}^{M-1}\left|\hat{I}[l,m]\,10\log_{10}\left(\frac{\chi\left\{\left|X_{1,c}[l,m]\right|^2\right\}}{\chi\left\{\left|Z[l,m]\right|^2\right\}}\right)\right| dB \tag{3.102}$$



(a) Coherent noise field.

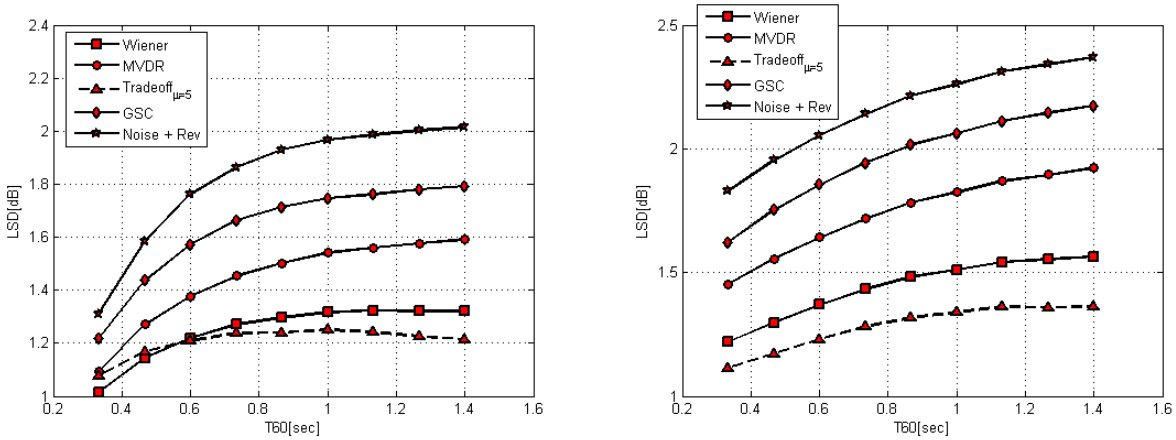
(b) Diffused noise field.

Figure 3.4: Broadband speech distortion index as function of the input signal to coherent noise ratio.

where M is the number of frequency bins, and $\chi\left\{|Z[l,m]|^2\right\} = \max\left\{|Z[l,m]|^2, \epsilon\right\}$ denotes a clipping operator which confines the log-spectrum dynamic range to about 50 dB, i.e.,$\epsilon = 10^{-50/10}\max_{l,m}\left\{|Z[l,m]|^2\right\}$. By using the indicator function $\hat{I}[l,m]$ (3.89), we make sure that the LSD is averaged only on relevant frequency bins that contain speech. $X_{1,c}[l,m]$ is calculated by convolving the speech signal with the first n_d taps (mentioned in Sub-section 2.1.4) of the reference channel RIR. The broadband LSD curve is defined as the average of the narrowband LSD over all time frames.

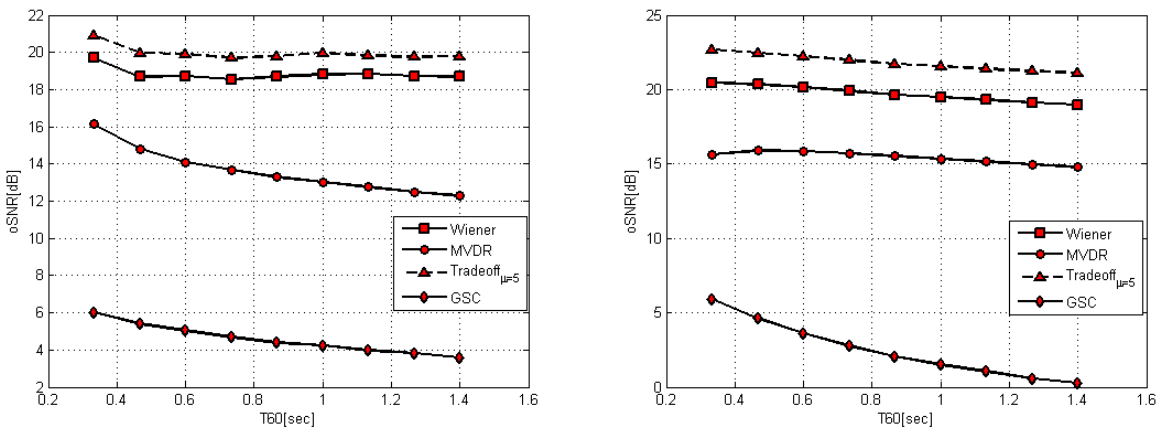
Similar to the previous experiment, we compare the beamformers performance to that of the TF-GSC beamformer. In Fig. 3.5 (a)-(b), the broadband LSD plots for the directional and diffused noise field are shown. Looking at the LSD curves we notice an improvement in dereverberation relative to the TF-GSC. The high spectral distortion of the TF-GSC can be traced back to the leakage problem. This problem becomes severe in reverberant environments, where the RTF estimation becomes less accurate and non optimal BM is applied. The beamformers we propose show less sensitive to these estimation errors, thus maintaining a lower level of distortion. We observe that the LSD is higher in the case of diffused noise field. In that scenario more noise enters the beamformers look direction, adding a bias to all the LSD curves.

Plots of the broadband oSNR as a function of the reverberation time are shown in



(a) Coherent noise field. (b) Diffuse noise field.

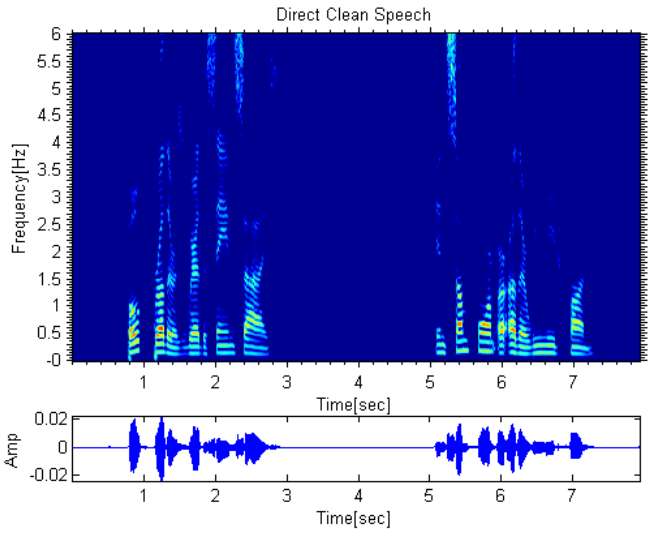
Figure 3.5: Broadband log spectral distortion as function of T_{60}.



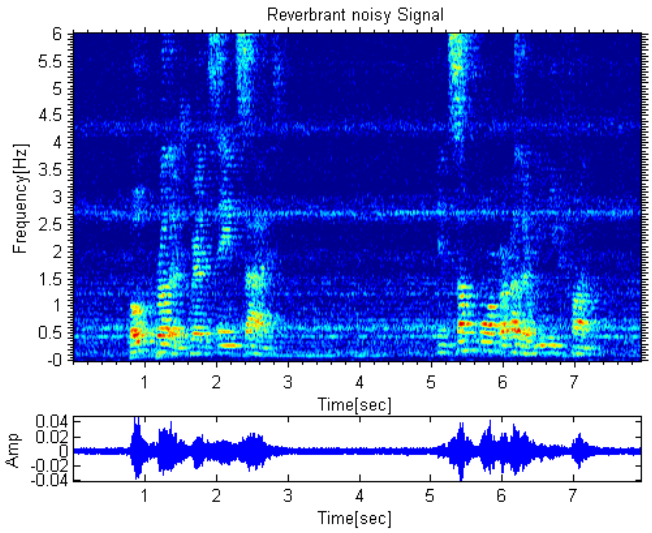
(a) Coherent noise field. (b) Diffuse noise field.

Figure 3.6: Broadband output SNR as a function of T_{60}.

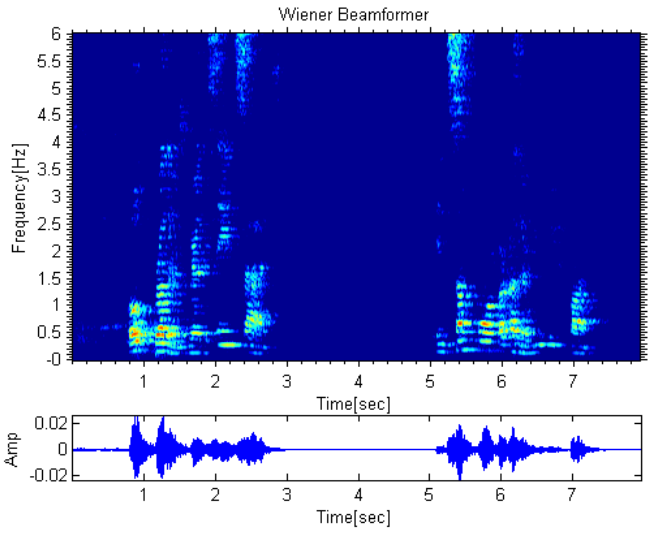
Fig. 3.6 (a)-(b), we note that the reverberation time has lesser effect on the oSNR than that of the iSCNR. The small decrease in oSNR demonstrates the proposed beamformers capability to maintain high levels of oSNR in reverberant environments, contrary to the TF-GSC for which the oSNR decrease below the level of the iSCNR. The results also show that the beamformers that achieve higher noise reduction also achieve greater reverberation suppression. This effect is expected as the late reverberations are modeled as a directive interference.



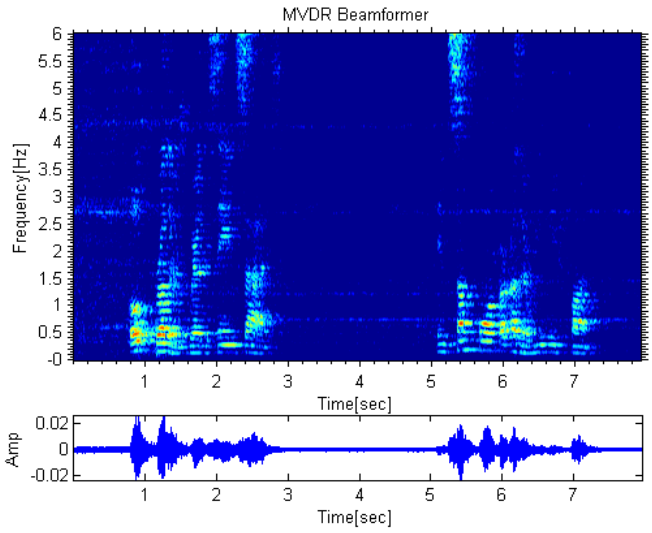
(a) Clean speech signal.



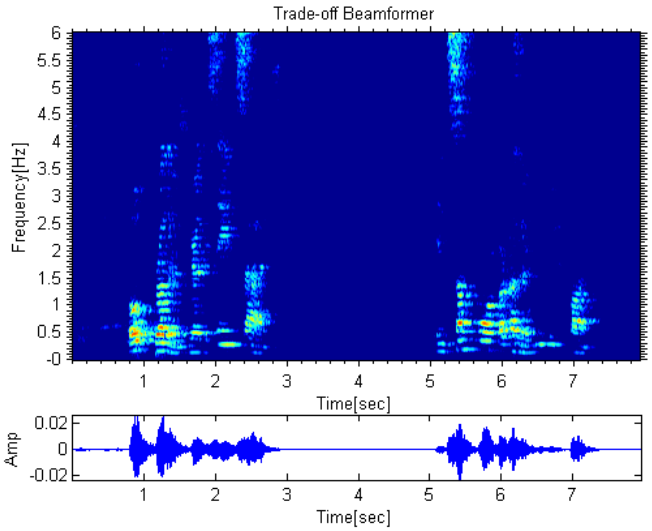
(b) Reverberant and noisy signal.



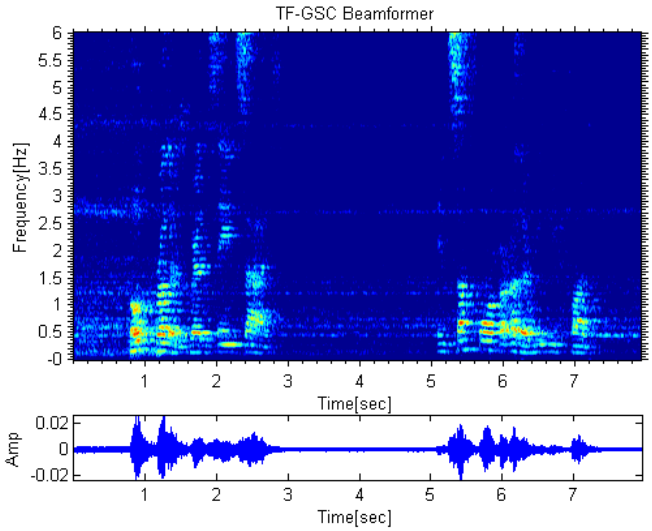
(c) Output of the Weiner beamformer.



(d) Output of the minimum variance distortionless response beamformer.



(e) Output of the tradeoff beamformer.



(f) Output of the transfer function generalized sidelobe canceler beamformer.

Figure 3.7: Spectrograms of the beamformers' outputs.

Examples of speech spectrograms for the case of coherent noise field with iSCNR of 5 dB and $T_{60} = 0.8\,\text{sec}$, can be seen in Fig. 3.7 (a)-(f). The spectrograms visually demonstrate the smoothing effect of the reverberations on the speech signal. We clearly observe that the enhanced signals obtained by the proposed beamformers are less noisy and less distorted than the enhanced signal obtained by the TF-GSC.

3.9 Conclusions

In this chapter we proposed a multichannel speech enhancement algorithm designed for the task of joint noise reduction and reverberation suppression. We've suggested algorithms for the calculation of three different beamformers adapted to perform in reverberant environments.

The proposed algorithms were based on a simple signal model, in which the array observation vector is written as a function of the clean early reflections signal. This was achieved by incorporating the full to early RTF steering vector into the signal model. Using this model, we developed important quality measures and a closed-form formulation for the MVDR, Wiener, and tradeoff beamformers.

We proposed an iterative estimator for the full to early RTFs which is adapted specifically to speech signals, and incorporate it into an on-line algorithm for the calculation of the beamformers. We provided a thorough description of the estimation processes of the different PSDs, needed for the construction of the suggested beamformers.

Performance analysis and simulations results show that our method outperforms the TF-GSC beamformer in aspects of both noise reduction and signal distortion. It also show impressive reverberation suppression capabilities, in both diffused and directional noise fields.

Chapter 4

STFT-Domain Late Reverberation Blocking Beamformers

4.1 Introduction

In Chapter 3 we suggested three different beamformers for the task of joint dereverberation and noise reduction. These beamformers use the full to early RTF steering vector to extract the early clean speech signal. Though this approach shows promising capabilities in aspects of denoising and dereverberation, it still hasn't incorporated the problem of reverberation suppression into the minimization process. In other words, the dereverberation mechanism is based on successful extraction of the early clean reflections, rather than on the direct suppression of the late reverberations.

In this chapter we express the microphone array signal as a sum of three terms: the early reflection vector, the unwanted coherent interference vector (which includes the late reverberations), and the uncorrelated noise vector. Using the MSE measure we show that there's a direct correlation between the speech distortion error signal and the reverberation suppression error signal. To remove it we impose a decorrelating constraint, allowing us to treat them as statistically independent. From this perspective new optimization problems emerge, in which the reverberation suppression process undertakes an active part. This in turn will allow us to design low distortion beamformers with high dereverberation and noise reduction capabilities.

The chapter is organized as follows. In Section 4.2, a new signal model is introduced.

In Section 4.3 we define new quality measure, and use them to derive and evaluate the relation between noise reduction, reverberation suppression and speech distortion. In Section 4.4 beamformers with high dereverberation capabilities are introduced. Section 4.5 contains some simulation results that corroborate our study. Finally, Section 4.6 concludes this chapter.

4.2 Signal Model

In Chapter 3, Section 3.2 we presented the classical model for the STFT domain representation of a microphone array observations vector, recorded in a reverberant room. In this section we use a different decomposition of the observations vector, extending it to a sum of three terms. The first term describes the desired clean early signal. The second term describes the coherent interference signals that arrive at each of the array microphones and includes the late reverberation. The third part describes the uncorrelated residual noise.

Recalling the notation introduced in (3.5), the observation vector of a microphone array with N microphones is given by:

$$\begin{aligned}\overline{y}[l,m] &= \bar{x}[l,m] + \overline{v}[l,m] \\ &= \overline{d}[m] X_{1,c}[l,m] + \overline{v}[l,m]\end{aligned} \tag{4.1}$$

where

$$\begin{aligned}\overline{y}[l,m] &= [Y_1[l,m], Y_2[l,m], ..., Y_N[l,m]]^T = S\,[l,m]\,\bar{g}[m] \\ \overline{x}[l,m] &= [X_1[l,m], X_2[l,m], ..., X_N[l,m]]^T \\ &= S\,[l,m]\,[G_1\,[m]\,, G_2\,[m]\,, ..G_N\,[m]] = S\,[l,m]\,\bar{g}[m] \\ \overline{v}[l,m] &= [V_1[l,m], V_2[l,m], ..., V_N[l,m]]^T \\ \overline{d}_{X_{1,c},\overline{x}}[m] &= \left[\frac{G_1[m]}{G_{1,c}[m]}, \frac{G_2[m]}{G_{1,c}[m]}, ..., \frac{G_N[m]}{G_{1,c}[m]}\right]^T = \frac{\bar{g}[m]}{G_{1,c}[m]}.\end{aligned} \tag{4.2}$$

The signal representation in (4.2) allows us to account for reverberation coming at the beamformer look direction (without post filtering operation), however beamformers designed based on this model overlook the direct suppression of the late reverberation and

focuse on an undistorted extraction of the early clean signal. In real life scenarios, the late reverberations are correlated to the early reflections, therefore suppressing the late reverberation without inflicting distortion to the desired signal when using the model in (4.1), is somewhat limited.

We suggest decomposing the clean but reverberant observation vector $\bar{x}[l,m]$ into the part that includes the early reflection in each channel, and the one that includes the coherent interference. The coherent interference vector is derived by subtracting the desired speech signal from each element of the reverberant observations vector, clearly this vector includes the late reverberations.

This decomposition will allow us to address dereverberation, noise reduction, and speech distortion separately in the optimization scheme. Which will lead to better reverberation suppression, and increase the flexibility of the design process, introducing new measures and tradeoffs.

We start by a simple manipulation of the signal model in (4.2):

$$\begin{aligned}\overline{y}[l,m] &= \overline{d}_{X_{1,c},\overline{x}}[m]X_{1,c}[l,m] + \overline{v}[l,m] \\ &= i_N X_{1,c}[l,m] + \left(\overline{d}_{X_{1,c},\overline{x}}[m] - i_N\right) X_{1,c}[l,m] + \overline{v}[l,m] \\ &= i_N X_{1,c}[l,m] + \overline{r}\,[m]\, X_{1,c}[l,m] + \overline{v}[l,m]\end{aligned} \tag{4.3}$$

where $i_N = [1,1,...,1]^T$ is a $1 \times$ N vector of ones, and

$$\begin{aligned}\overline{r}\,[m] &= \overline{d}_{X_{1,c},\overline{x}}[m] - i_N \\ &= \left[\frac{G_1[m]}{G_{1,c}[m]} - 1, \frac{G_2[m]}{G_{1,c}[m]} - 1, ..., \frac{G_N[m]}{G_{1,c}[m]} - 1\right]^T \\ &= \left[\frac{G_{1,l}[m]}{G_{1,c}[m]}, \frac{G_2[m] - G_{1,c}[m]}{G_{1,c}[m]}, ..., \frac{G_N[m] - G_{1,c}[m]}{G_{1,c}[m]}\right]^T\end{aligned} \tag{4.4}$$

is the vector holding the ratio between the unwanted parts of the RIR, and the desired part. We note that its first element contains the ratio between the RIR of the late reverberations and the RIR of the desired signal.

In the following sections we show that this manipulation is a solid basis for the design of new and interesting beamformers.

4.3 Performance Measures

In this section, we define different quality measures based on the model presented in (4.3). These measures allow us to analyze the suggested beamformers' noise reduction and reverberation suppression capabilities independently of each other, and will be used to define new minimization problems. As in Section 3.4, we define the microphone indexed by 1 as the reference microphone; all measures will be defined with respect to the signal it records.

4.3.1 MSE Criterion

We define the error signal between the estimated and the desired signals at frequency bin m and time frame l as:

$$\begin{aligned}\varepsilon[l,m] &= Z[l,m] - X_{1,c}[l,m] = \bar{h}^H[l,m]\overline{y}[l,m] - X_{1,c}[l,m] \\ &= \bar{h}^H[l,m]\left[(i_N - 1)\, X_{1,c}[l,m] + \overline{r}\,[m]\, X_{1,c}[l,m] + \overline{v}[l,m]\right]. \end{aligned} \tag{4.5}$$

This error can also be expressed as:

$$\varepsilon[l,m] = \varepsilon_d[l,m] + \varepsilon_c[l,m] + \varepsilon_r[l,m] \tag{4.6}$$

where

$$\varepsilon_d[l,m] = \left|\overline{h}^H[l,m]i_N - 1\right| X_{1,c}[l,m] \tag{4.7}$$

is the speech distortion due to the complex filter,

$$\varepsilon_c[l,m] = \bar{h}^H[l,m]\overline{r}[l,m]X_{1,c}[l,m] \tag{4.8}$$

is the coherent interference due to the late reverberations, and

$$\varepsilon_r[l,m] = \bar{h}^H[l,m]\overline{v}[l,m] \tag{4.9}$$

is the residual noise. The error signals $\varepsilon_d[l,m]$ and $\varepsilon_c[l,m]$ are statistically independent with $\varepsilon_r[l,m]$, but are obviously correlated, as they both include $X_{1,c}[l,m]$.

The STFT domain narrowband MSE is given by:

$$J\left[\bar{h}[l,m]\right] = E\left[|\varepsilon[l,m]|^2\right] = \phi_{X_{1,c}}[l,m]\left|\left(\overline{h}^H[l,m]i_N - 1\right)^2\right| \tag{4.10}$$
$$+ 2\phi_{X_{1,c}}[l,m]Re\left\{\left(\overline{h}^H[l,m]i_N - 1\right)\overline{r}^H[m]\overline{h}[l,m]\right\}$$
$$+ \phi_{X_{1,c}}[l,m]\left|\overline{h}^H[l,m]\overline{r}[m]\overline{r}^H[m]\overline{h}[l,m]\right|^2 + \overline{h}^H[l,m]\left[\Phi_{vv}[l,m]\right]\overline{h}[l,m].$$

The expression above demonstrates the main challenge inherent in the JDNR problem, the correlation between the speech distortion signal and the coherent interference signal does not allow treating the dereverberation challenge independently from the distortionless extraction of the desired speech signal. However if we impose a constraint that decorrelates these error signals, i.e. $2Re\left\{\left(\overline{h}^H[l,m]i_N - 1\right)\overline{r}^H[m]\overline{h}[l,m]\right\} = 0$, the error expression in (4.10) becomes simpler and can be used as a basis for different optimization problems.

Under such constraint we get the following MSE,

$$J\left[\bar{h}[l,m]\right] = \phi_{X_{1,c}}[l,m]\left|\left(\overline{h}^H[l,m]i_N - 1\right)\right|^2 + \overline{h}^H[l,m]\left[\Phi_{vv}[l,m]\right]\overline{h}[l,m] \tag{4.11}$$
$$+ \overline{h}^H[l,m]\left[\Phi_{rr}[l,m]\right]\overline{h}[l,m]$$

where $\Phi_{rr}[l,m] = \phi_{X_{1,c}}[l,m]\overline{r}[m]\overline{r}^H[m]$ is the correlation matrix of the coherent interference. The expression in (4.11) is composed out of three terms which will later on be used in the design of the suggested beamformers

$$J_d[\bar{h}[l,m]] = E\left[|\varepsilon_d[l,m]|^2\right] = \phi_{X_{1,c}}[l,m]\left|\overline{h}^H[l,m]i_N - 1\right|^2 \tag{4.12}$$

$$J_{rc}[\bar{h}[l,m]] = E\left[|\varepsilon_c[l,m]|^2\right] = \overline{h}^H[l,m]\left[\Phi_{rr}[l,m]]\right]\overline{h}[l,m] \tag{4.13}$$

$$J_{rn}[\bar{h}[l,m]] = E\left[|\varepsilon_r[l,m]|^2\right] = \overline{h}^H[l,m]\left[\Phi_{vv}[l,m]]\right]\overline{h}[l,m]. \tag{4.14}$$

The broadband MSE is defined as:

$$\begin{aligned} J[\bar{h}[l,:]] &= \frac{1}{M}\sum_{m=0}^{M-1} J[\bar{h}[l,m]] \\ &= \frac{1}{M}\sum_{m=0}^{M-1} J_d[\bar{h}[l,m]] + \frac{1}{M}\sum_{m=0}^{M-1} J_{rl}[\bar{h}[l,m]] \\ &= J_d[\bar{h}[l,:]] + J_{rc}[\bar{h}[l,:]] + J_{rn}[\bar{h}[l,:]]. \end{aligned} \tag{4.15}$$

Obviously straightforward minimization of (4.15) will lead directly to the Wiener beamformer (3.50). However, new optimization problems can be derived, each addressing the

minimization of $J_{rc}[\bar{h}[l,:]]$, $J_{rn}[\bar{h}[l,m]]$ or $J_d[\bar{h}[l,:]]$ under some constraint, while enforcing the constraint that

$$2Re\left\{\left(\overline{h}^H[l,m]i_N - 1\right)\overline{r}^H[m]\overline{h}[l,m]\right\} = 0. \tag{4.16}$$

In practice enforcing (4.16) can be achieved by enforcing no distortion

$$\overline{h}^H[l,m]i_N = 1 \tag{4.17}$$

or by enforcing reverberation blockage

$$\overline{r}^H[m]\overline{h}[l,m] = 0. \tag{4.18}$$

4.3.2 Noise Reduction and Reverberation Suppression

One of the most fundamental measures in all aspects of speech enhancement is the signal-to-noise ratio (SNR). The iSNR is a second-order measure which quantifies the level of noise and reverberations present, relative to the level of the desired signal.

The narrowband iSNR is defined as the following ratio

$$\begin{aligned} iSNR[l,m] &= \frac{\phi_{X_{1,c}}[l,m]}{\phi_{Y_1}[l,m] - \phi_{X_{1,c}}[l,m]} \\ &= \frac{\phi_{X_{1,c}}[l,m]}{\phi_{X_{1,c}}[l,m]\left[|\bar{r}[1]|^2 + 2Re\left\{\bar{r}[1]\right\}\right] + \phi_{V_1}[l,m]}. \end{aligned} \tag{4.19}$$

The broadband iSNR is given by

$$iSNR[m] = \frac{\sum_{m=0}^{M-1}\phi_{X_{1,c}}[l,m]}{\sum_{m=0}^{M-1}\phi_{X_{1,c}}[l,m]\left[|\bar{r}[1]|^2 + 2Re\left\{\bar{r}[1]\right\}\right] + \phi_{V_1}[l,m]}. \tag{4.20}$$

The narrowband and broadband oSNRs are defined as follows:

$$oSNR\left\{\overline{h}[l,m]\right\} = \frac{\phi_{X_{1,c}}[l,m]\left|\overline{h}^H[l,m]i_N\right|^2}{\overline{h}^H[l,m]\left[\Phi_{tt}[l,m]\right]\overline{h}[l,m]} \tag{4.21}$$

$$oSNR\left\{\overline{h}[l,:]\right\} = \frac{\sum_{m=0}^{M-1}\phi_{X_{1,c}}[l,m]\left|\overline{h}^H[l,m]i_N\right|^2}{\sum_{m=0}^{M-1}\left[\overline{h}^H[l,m]\left[\Phi_{tt}[l,m]\right]\overline{h}[l,m]\right]}, \tag{4.22}$$

where $\Phi_{tt}[l,m] = \Phi_{rr}[l,m] + \Phi_{vv}[l,m]$ is the sum of the noise and coherent interference correlation matrices. Looking at expression (4.21) we recognize the generalized Rayleigh

quotient, which is maximized by the eigenvector of the matrix $\phi_{X_{1,c}}[l,m]\Phi_{tt}^{-1}[l,m]i_N i_N^T$ denoted by $\overline{h}_{Nmax}[l,m]$. Since this matrix is of rank 1, the maximum eigenvalue $\lambda_{\mathrm{Nmax}}[l,m]$, which corresponds to the maximum egienvector is given by:

$$\begin{aligned}\lambda_{N\max}[l,m] &= \operatorname{trace}\left\{\phi_{X_{1,c}}[l,m]\Phi_{tt}^{-1}[l,m]i_N i_N^T\right\} \\ &= \phi_{X_{1,c}}[l,m]\sum\sum\Phi_{tt}^{-1}[l,m]. \end{aligned} \tag{4.23}$$

As a result

$$oSNR\left[\overline{h}_{Nmax}[l,m]\right] = \lambda_{\mathrm{Nmax}}[l,m] \tag{4.24}$$

and

$$\overline{h}_{Nmax}[l,m] = \varsigma\left[l,m\right]\Phi_{tt}^{-1}[l,m]i_N. \tag{4.25}$$

The inverse of $\Phi_{tt}[l,m]$ is obtained using Woodbury's identity

$$\Phi_{tt}^{-1}[l,m] = \Phi_{vv}^{-1}[l,m] - \phi_{X_{1,c}}[l,m]\frac{\Phi_{vv}^{-1}[l,m]\bar{r}\bar{r}^H\Phi_{vv}^{-1}[l,m]}{1+\phi_{X_{1,c}}[l,m]\bar{r}^H\Phi_{vv}^{-1}[l,m]\bar{r}}.$$

We note that for beamformers that enforce (4.18), the oSNR degenerates to it's known formulation (3.39)

$$\overline{h}^H[l,m]\left[\Phi_{tt}[l,m]\right]\overline{h}[l,m] = \overline{h}^H[l,m]\left[\Phi_{vv}[l,m]\right]\overline{h}[l,m]. \tag{4.26}$$

The role of the beamformer is to produce a signal whose SNR is higher than that which was received. To that end, the array gain is defined as the ratio of the oSNR (after beamforming) over the iSNR (at the reference microphone) [36]. This leads to the definition of the narrowband array gain:

$$A\left\{\overline{h}[l,m]\right\} = \frac{oSNR\left\{\overline{h}[l,m]\right\}}{iSNR\left[l,m\right]}, \tag{4.27}$$

and the broadband array gain

$$A\left\{\overline{h}[:,m]\right\} = \frac{oSNR\left\{\overline{h}[:,m]\right\}}{iSNR\left[m\right]}. \tag{4.28}$$

From (4.24) and (4.27) it is clear that

$$A\left\{\overline{h}[l,m]\right\} \leq \frac{\lambda_{\mathrm{Nmax}}[l,m]}{iSNR\left[m\right]}. \tag{4.29}$$

To quantify the amount of late reverberation suppressed by the beamformer, we define the narrowband reverberation suppression index

$$\begin{aligned}\nu_{sr}[\overline{h}[l,m]] &= \frac{\phi_{X_{1,c}}[l,m]\left|\overline{h}^H[l,m]\overline{r}[l,m]\right|^2}{\phi_{X_{1,c}}[l,m]} \\ &= \left|\overline{h}^H[l,m]\overline{r}[l,m]\right|^2. \end{aligned} \tag{4.30}$$

The reverberation suppression index is expected to be close to zero, the higher it is the more late reflections enter the beamformer's look direction. The broadband reverberation suppression index is defined as the summation of the narrowband reverberation suppression index across all frequency bands

$$\nu_{sr}\left\{\overline{h}[l,:]\right\} = \frac{\sum_{m=0}^{M-1}\phi_{X_{1,c}}[l,m]\left|\overline{h}^H[l,m]\overline{r}[l,m]\right|^2}{\sum_{m=0}^{M-1}\phi_{X_{1,c}}[l,m]}. \tag{4.31}$$

4.3.3 Speech Distortion

To measure the distortion imposed by the beamformers we use the speech distortion index, defined as the mean-square error between the desired signal and the filtered desired signal, normalized by the variance of the desired signal, i.e.,

$$\nu_{sd}[\overline{h}[l,m]] = \frac{\phi_{X_{1,c}}[l,m]\left|\overline{h}^H[l,m]i_N - 1\right|^2}{\phi_{X_{1,c}}[l,m]} \tag{4.32}$$

in the narrowband case, and

$$\nu_{sd}[\overline{h}[l,:]] = \frac{\sum_{m=0}^{M-1}\phi_{X_{1,c}}[l,m]\left|\overline{h}^H[l,m]i_N - 1\right|^2}{\sum_{m=0}^{M-1}\phi_{X_{1,c}}[l,m]} \tag{4.33}$$

in the broadband case.

4.4 Reverberation Block Beamformers

In this sub-section, we derive three fundamental beamformers using the adjusted linear model in (4.3). We use the mean-square error (MSE) criteria to define new optimization problems, addressing a variety of application and performance needs.

4.4.1 Reverberation-Block Beamformer

The Reverberation Block Wiener beamformer (RBW) is designed to minimize the speech distortion index $J_d[\bar{h}[l,m]]$ and the residual noise $J_{rn}[\bar{h}[l,m]]$ under the constraint that no late reverberations be allowed to enter the beamformer look direction (4.18). Mathematically, this is equivalent to:

$$\begin{aligned} \min_{\bar{h}[l,m]} \phi_{X_{1,c}}[l,m] \left|\overline{h}^H[l,m] i_N - 1\right|^2 \\ + \overline{h}^H[l,m] \Phi_{vv}[l,m] \overline{h}[l,m] \\ \text{S.T.} \quad \overline{h}^H[l,m] \overline{r}[m] = 0. \end{aligned} \tag{4.34}$$

The beamformer that solves this minimization problem is given by:

$$\overline{h}_{RBW}[l,m] = \phi_{X_{1,c}}[l,m] \Phi_p^{-1}[l,m] \left[I_{NxN} - \frac{\overline{r}[m]\, \overline{r}^H[m]\, \Phi_p^{-1}[l,m]}{\overline{r}^H[m]\, \Phi_p^{-1}[l,m] \overline{r}[m]} \right] i_N \tag{4.35}$$

where the inverse of $\Phi_p[l,m] = \phi_{X_{1,c}}[l,m] i_N^H i_N + \Phi_{vv}[l,m]$ is derived using Woodbury's identity

$$\Phi_p^{-1}[l,m] = \Phi_{vv}^{-1}[l,m] - \phi_{X_{1,c}}[l,m] \frac{\Phi_{vv}^{-1}[l,m] i_N i_N^H \Phi_{vv}^{-1}[l,m]}{1 + \phi_{X_{1,c}}[l,m] i_N^H \Phi_{vv}^{-1}[l,m] i_N}.$$

The narrowband and broadband speech reverberation indices of the RBW are obviously equal to 0 as the constraint in (4.34) requires.

4.4.2 Normalized MVDR Beamformer

The Normalized Minimum Variance Distortionless Response (NMVDR) beamformer minimizes the sum of $J_{rc}[\bar{h}[l,m]]$ and $J_{rn}[\bar{h}[l,m]]$, under the constraint that the desired signal is not distorted $J_d[\bar{h}[l,m]] = 0$. Mathematically, this is equivalent to

$$\min_{\bar{h}[l,m]} \overline{h}^H[l,m] \left[\Phi_{tt}[l,m]\right] \overline{h}[l,m] \text{ S.T. } \overline{h}^H[l,m] i_N = 1. \tag{4.36}$$

The solution for this minimization problem is given by

$$\overline{h}_{NMVDR}[l,m] = \frac{\Phi_{tt}^{-1}[l,m] i_N}{i_N^H \Phi_{tt}^{-1}[l,m] i_N}. \tag{4.37}$$

We can derive a LCMV beamformer [14], by adding the reverberation blockage constraint (4.18) to the no distortion constraint (4.17). The two constraints are put together

in matrix form as:

$$\overline{h}^H[l,m]C[m] = [1,0]^T \tag{4.38}$$

where $C[m] = [i_N, \overline{r}[m]]$ is a constraint matrix of size $L \times 2$.

The optimal filter is obtained by minimizing the residual uncorrelated noise, with the constraints that the correlated late reverberations be blocked and the desired speech be preserved, i.e.,

$$h_{NLCMV}[l,m] = \Phi_{tt}^{-1}[l,m]C[m] \times \left[C^H[m]\Phi_{tt}^{-1}[l,m]C[m]\right]^{-1}[1,0]^T. \tag{4.39}$$

By developing (4.39), it can be shown that the NLCMV can be written as a combination of two terms, i.e.,

$$h_{NLCMV}[l,m] = \frac{1}{1-|\theta(l,m)|^2}h_{NMVDR}[l,m] - \frac{|\theta(l,m)|^2}{1-|\theta(l,m)|^2}\Omega[l,m] \tag{4.40}$$

where

$$|\theta(l,m)|^2 = \frac{\left|\overline{r}^H[m]\Phi_{tt}^{-1}[l,m]i_N\right|^2}{\left|i_N^T\Phi_{tt}^{-1}[l,m]i_N\right|\left|\overline{r}^H[m]\Phi_{tt}^{-1}[l,m]\overline{r}[m]\right|} \tag{4.41}$$

with $|\theta(l,m)|^2 \in [0,1]$ and

$$\Omega[l,m] = \frac{\Phi_{tt}^{-1}[l,m]\overline{r}^H[m]}{i_N^T\Phi_{tt}^{-1}[l,m]\overline{r}[m]}. \tag{4.42}$$

We observe from (4.40) that when $|\theta(l,m)|^2$ approaches 0, the NLCMV filter tends to the NMVDR filter; however, when $|\theta(l,m)|^2$ approaches 1, there is no solution since we have conflicting requirements. Obviously, we always have

$$\begin{aligned}
\text{oSNR}\left[\overline{h}_{NLCMV}[l,m]\right] &\leq \text{oSNR}\left[\overline{h}_{NMVDR}[l,m]\right] \\
&= \text{oSNR}\left[\overline{h}_{Nmax}[l,m]\right] \\
\nu_{sd}\left[\overline{h}_{NLCMV}[l,m]\right] &= \nu_{sd}\left[\overline{h}_{NMVDR}[l,m]\right] = 0 \\
\nu_{sr}\left[\overline{h}_{NCLMV}[l,m]\right] &= 0.
\end{aligned} \tag{4.43}$$

4.4.3 Reverberation Block Tradeoff Beamformer

The RBW and NMVDR beamformers, do not allow much flexibility in that we cannot quantify in advance the improvement in the narrowband oSNR and the reverberation

suppression index. However, in many practical situations, we do wish to control the compromise between noise reduction, reverberation suppression, and speech distortion. In this subsection we will introduce a beamformer which allows some flexibility with regard to these measures while still enforcing the constraint in (4.16).

The Reverberation Block tradeoff (RBT) beamformer minimizes the speech distortion index $J_d[\bar{h}[l,m]]$, with the constraints that a certain amount of noise be allowed to enter the beamformer's look direction, and that the coherent interference is blocked. Mathematically, this is equivalent to

$$\begin{aligned} \min_{\bar{h}[l,m]} \phi_{X_{1,c}}[l,m] \left| \overline{h}^H[l,m] i_N - 1 \right|^2 \\ \text{S.T.} \quad \overline{h}^H[l,m] \overline{r}[m] = 0 \\ \overline{h}^H[l,m] \Phi_{vv}[l,m] \overline{h}[l,m] = \gamma \left[\phi_{V_1}[l,m] \right] \end{aligned} \tag{4.44}$$

where $\gamma \in (0,1)$ to insure that we get some noise reduction. By using a Lagrange multiplier, $\eta > 0$ to adjoin the constraint to the cost function, we formulate the RBT beamformer:

$$\overline{h}_{RBT,\eta}[l,m] = \phi_{X_{1,c}}[l,m] \Phi_{p,\eta}^{-1}[l,m] \times \left[I_{\text{NxN}} - \frac{\overline{r}[m] \overline{r}^H[m] \Phi_{p,\eta}^{-1}[l,m]}{\overline{r}^H[m] \Phi_{p,\eta}^{-1}[l,m] \overline{r}[m]} \right] i_N \tag{4.45}$$

where

$$[\Phi_{p,\eta}[l,m]]^{-1} = [\eta \Phi_{vv}[l,m]]^{-1} - \phi_{X_{1,c}}[l,m] \frac{\Phi_{vv}^{-1}[l,m] i_N i_N^H \Phi_{vv}^{-1}[l,m]}{\eta^2 + \eta \phi_{X_{1,c}}[l,m] i_N^H \Phi_{vv}^{-1}[l,m] i_N}. \tag{4.46}$$

The derivation of this beamformer is described in details in Appendix 1. We note that for this beamformer η must be positive, when η is chosen in an ad-hoc way the following statements hold:

- For $\eta \to 0$ the RBT beamformer converges to the NMVDR beamformer $\lim_{\eta \to 0^+} \overline{h}_{RBT,\eta}[l,m] \to \overline{h}_{NLCMV}[l,m]$.
- For $\eta = 1$ the RBT beamformer converges to the RBW beamformer $\lim_{\eta \to 1^+} \overline{h}_{RBT,\eta}[l,m] \to \overline{h}_{RBW}[l,m]$.
- For $\eta > 1$ the RBT beamformer reduces residual noise better, although at the expense of higher speech distortion.

- For $\eta \in (0, 1)$ the RBT beamformer reduces speech distortion better, although at the expense of some residual noise.

The narrowband and broadband speech reverberation indices of the RBT are obviously equal to 0 as the constraint in (4.44) requires.

4.5 Experimental Study

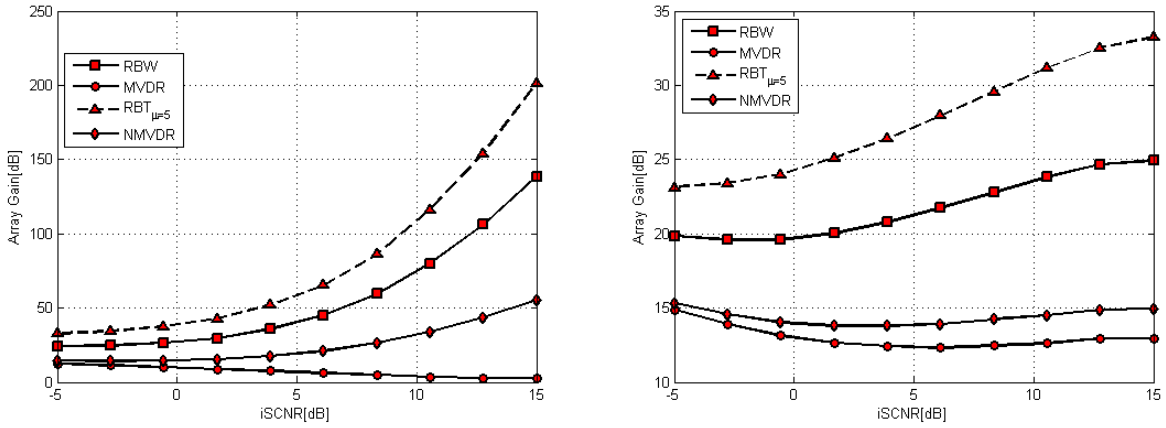
In this section, we present an experimental analysis of the proposed beamformers. Using the measures introduced in Sub-section 4.3, we evaluate the suggested beamformers' performances, in the context of speech distortion, noise reduction and reverberation suppression. We test the suggested beamformers capabilities in different environmental setups, changing either the noise field input SNR or the room reverberation time. The results of our simulations are presented in terms of broadband performance measures. The beamformers are computed and applied on a frame-by-frame basis, and the performance measures are evaluated per frame and subsequently averaged over all frames. The SNRs are averaged in the logarithm domain while the rest of the measures are averaged in the linear domain.

4.5.1 Experimental Setup

In the following experiments we use Habet's simulator [21] to simulate the RIRs of a 7 microphone narrow band array. The RIRs are simulated in a rectangular room, 6 m wide by 6 m long, and 4 m high. The microphones are placed along a straight line ranging from [3.8 m, 3 m, 2 m] to [4.15 m, 3 m, 2 m], with inner microphone distance of 5 cm. A speech source is placed at [3 m ,3 m, 2 m], 1 m away from the reference microphone (forth in the array). The speech source is composed out of 20 different recorded speech signals (10 men and 10 women) taken from the TIMIT database [22] sampled at 16 KHz. We simulate two different noise fields, the first one is a diffused noise field simulated using the method proposed in [35]. The second one is a directional noise field, it is created using RIRs that describe the path from a noise source placed 1 m from the array at [4 m, 4 m, 2 m], to each microphones, as seen in Fig 3.1. To simulate the microphones' thermal

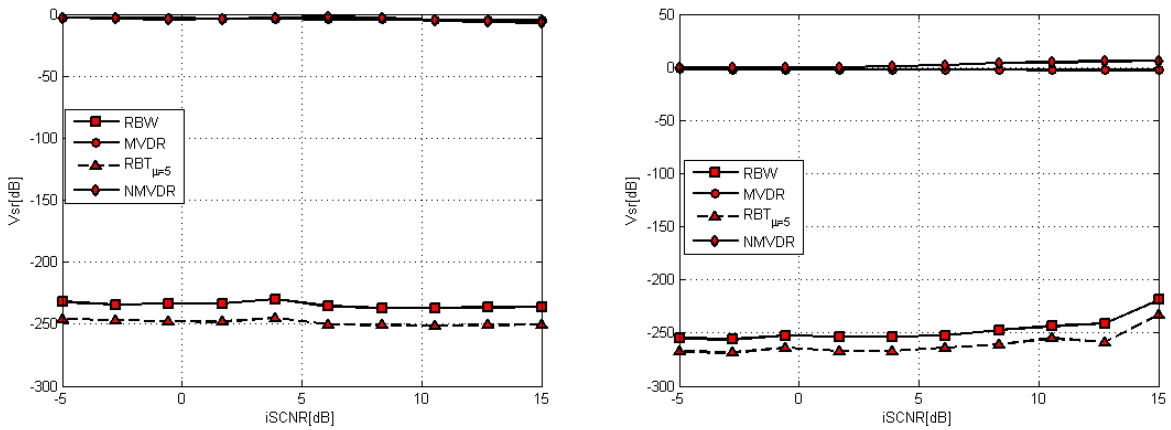
noise, a computer-generated white Gaussian noise with SNR of 20 dB is added to each microphone. Predefined parameters used in our algorithm are summed in Table 4.1.

The PSDs and RTFs needed for the beamformers' consturction are calculated using the implementation described in Section 3.7. To estimate $\bar{r}\,[l,m]$, we first the estimate the steering vector $\hat{d}_{X_{1,c},\overline{x}}[m]$ using (3.90), and then decrease 1 from each one of it's elements (4.4).



(a) Coherent noise field. (b) Diffused noise field.

Figure 4.1: Broadband array gain as function of the input signal to coherent noise ratio.



(a) Coherent noise field. (b) Diffused noise field.

Figure 4.2: Broadband reverberation suppression index as function of the input signal to coherent noise ratio.

Table 4.1: Predefined Parameters Reverberation Block Beamformers

Algorithm section	Needed Parameters	
RTF Estimation:	μ_{RTF}=0.025	p_0=0.65
Signal Correlation Matrix:	α_y=0.85	$\Phi_{y,min}$ =1.4e-6
Noise Correlation Matrix:	α_v=0.85	$\Phi_{v,min}$1.4e-7
	β=1.25	
OM-LSA:	α=0.92	F_{min}=-20 dB
LRSVE:	R=128	L_e=356 taps
Matrix Regulator:	δ=0.01	

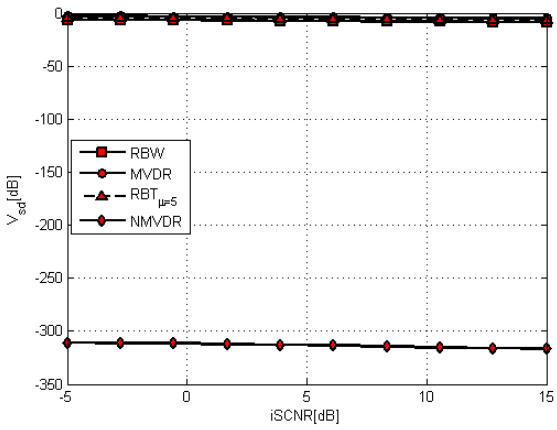
4.5.2 Influence of Noise Field iSNR

In this subsection, we investigate the beamformers' performance with regard to change in the noise field input SNR. We consider two different noise fields one diffused and the other directional, and change their iSCNR. The iSCNR is measured with respect to the signal at the reference channel and accounts only for changes in the coherent noise field (i.e. not including reverberation or thermal noise). The reverberation time is set to a constant value of 0.8 sec.

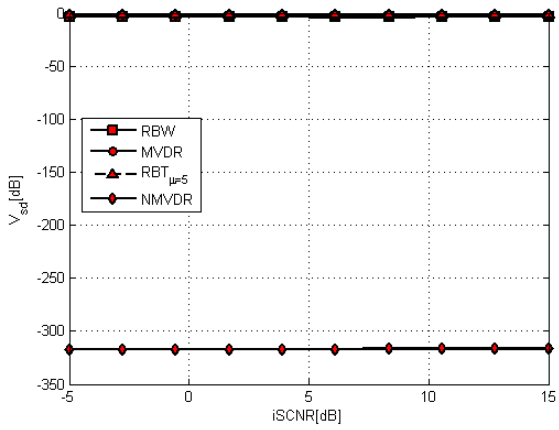
To demonstrate the effect of the signal decomposition described in (4.3), we compare the suggested beamformers performance to that of the MVDR beamformer (3.57), which is based on the classical signal model (4.1).

The broadband array gain curves for the diffused and directional noise fields are shown in Fig. 4.1 (a)-(b). From the plots we observe that all beamformers achieve some level of noise reduction and reverberation suppression. We note that for the diffuse noise case the array gain does not monotonically increase or decrease with iSCNR. In that case the statistics of the noise and the late reverberation are somewhat similar, as they both disperse in the room and arrive at the array from all directions. Therefore a complete separation between them is hard to achieve, and the expected monotonic behavior of the array gain is broken.

The tradeoff between reverberation suppression and speech distortion is demonstrated in Figs. 4.2-4.3 (a)-(b), where the speech distortion and reverberation suppression indices are plotted as a function of the iSCNR. We clearly see that beamformers that enforce the

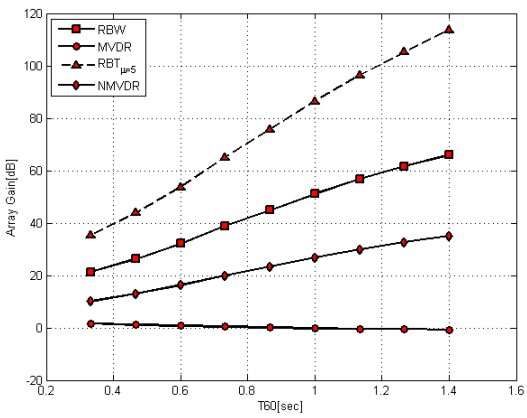


(a) Coherent noise field.

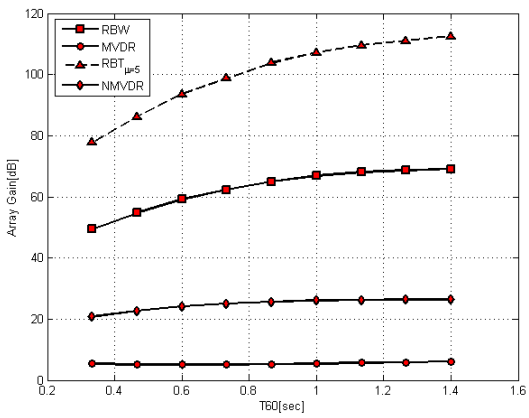


(b) Diffused noise field.

Figure 4.3: Broadband speech distortion index as function of the input signal to coherent noise ratio.

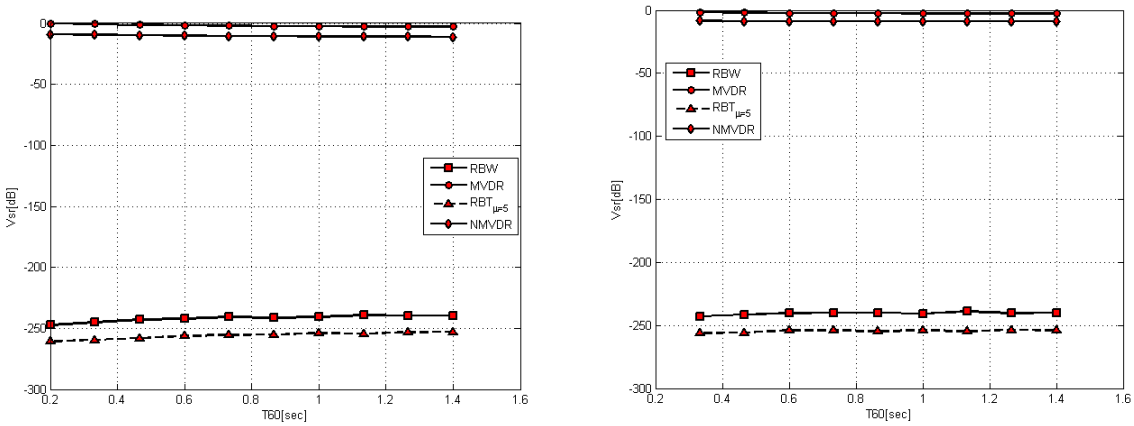


(a) Coherent noise field.



(b) Diffused noise field.

Figure 4.4: Broadband array gain as function of T_{60}.



(a) Coherent noise field. (b) Diffuse noise field.

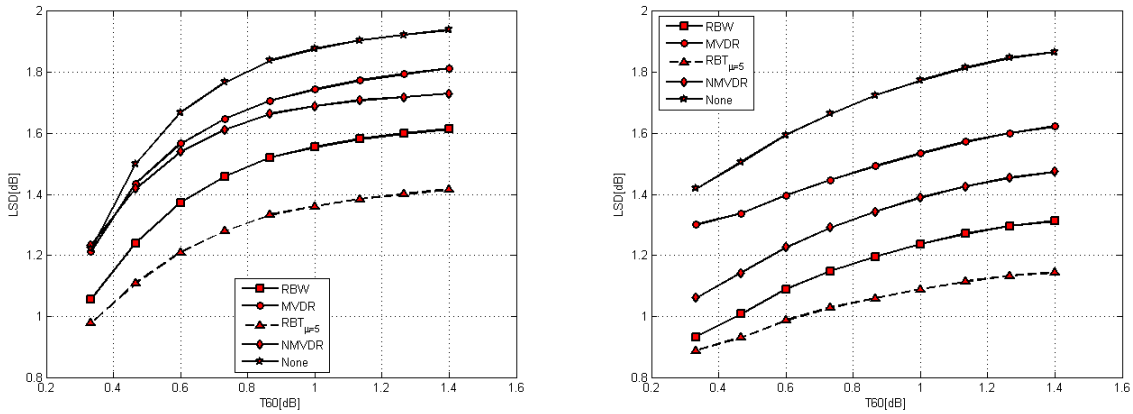
Figure 4.5: Broadband reverberation suppression index as function of T_{60}.

reverberation blockage constraint (RBW, RBT) achieve higher reverberation suppression on account of higher speech distortion, while the NMVDR beamformer which inflicts no speech distortion, suppresses less reverberations. We note that enforcing both constraints (NLCMV), degrades the beamformer noise reduction capabilities to a point where it is no longer interesting. From the plots we clearly see that the suggested beamformers outperform the MVDR beamformer in both noise reduction, and reverberation suppression aspects.

4.5.3 Influence of Reverberation Time

In this experiment we change the reverberation time T_{60}, while keeping a constant iSCNR of $5\,dB$. To test the beamformers dereverberation capabilities we use the array gain, the reverberation suppression index, and the broadband LSD measure (3.102).

Similar to the previous experiment we analyze the beamformers' performance in both diffused and directional noise fields, and compare them to that of the MVDR beamformer (3.57). Plots of the broadband array gain as a function of the reverberation time are shown in Fig. 4.4 (a)-(b). The array gain increases with T_{60}, since the decrease in oSNR due to an increase in reverberation time, is much smaller than the decrease in iSNR. In Fig. 4.5(a)-(b), the broadband speech reverberation index is plotted as a function of T_{60}. As expected the beamformers that enforce the reverberation blockage constraint (4.18)



(a) Coherent noise field. (b) Diffuse noise field.

Figure 4.6: Broadband log spectral distortion as function of T_{60}.

attain the highest reverberation suppression. We observe that the level of reverberation blockage only slightly changes as the reverberation time increases, demonstrating the proposed beamformers ability to preform in highly reverberant environments.

In Fig. 4.6 (a)-(b), the broadband LSD curves for the diffused and directional noise fields are shown. We observe that beamformers that enforce the reverberation blockage achieve lower spectral distortions, which implies higher reverberation suppression. Clearly from the plots, the suggested beamformers manage to suppress more of the late reflections than the standard MVDR beamformer.

4.6 Conclusions

In this chapter, we proposed new time frequency domain beamformers in room acoustic. We showed that by using the full to early RTFs we can rewrite the classical signal model for the observed signals as a sum of three terms: early reflection, unwanted coherent interference and noise. Using this model, we showed that the speech distortion error signal and the reverberation suppression error signal are correlated. To allow independent treatment of reverberation suppression in the design process, we've suggested constraining the beamformers' optimization problems, so that only solutions that decorrelate these error signals are allowed. From this perspective we've proposed interesting quality measures and developed different low distortion beamformers, with high noise reduction and late rever-

berations suppression capabilities. Through numerical experiments we've demonstrated the capabilities of the proposed beamformers in both diffused and directional noise fields, showing their superiority over the MVDR beamformer, and justifying the advantages of our approach in the design of beamformers for the JDNR problem.

Chapter 5

Conclusion

5.1 Research Summary

In this thesis we addresed the problem of joint dereverberation and noise reduction of speech signals. The review of the literature clearly depicts the limitations of exiting methods in solving the general problem of dereverberating signals contaminated by some noise field. To try and mitigate these limitations we proposed two approaches for the design of beamformers in the STFT domain.

In the first approach, we defined the full to early RTF as the ratio between the ATF of the early, clean speech reflections at the reference channel, and the ATFs of the array's microphones. We then used the full to early RTF to redefine the classical array signal model, writing the array's observations vector as a function of the desired clean early reflections. Using this model we adjusted the optimization problems of the Wiener, MVDR and tradeoff beamformers, to account for both noise reduction and reverberation suppression. Solutions of these optimization problem introduced beamformers, capable of extracting the clean early speech signal. We proposed a method for the estimation of full to early RTFs from speech signals recorded in reverberant environments, and incorporated it to an online algorithm for the construction of the suggested beamformers. We've concluded with an experimental study demonstrating the suggested beamformers capabilities to dereverberate noisy signals, recorded in directional or diffused noise fields. The experiments we've preformed, show the advantages of the proposed beamformers over the TF-GSC algorithm.

In the second approach we suggested a new decomposition for the array's observation vector, writing it as the sum of three terms: the early reflection, the unwanted coherent interference and the uncorrelated noise. Using this decomposition, we formulated the MSE error signal, and showed the correlation between the speech distortion and the coherent interference error signals. To simplify the formulation and allow derivation of new optimization problems, we've suggested a framework in which constrains are imposed to decorrelate these error signals. This framework allowed us to address reverberation suppression directly in the optimization problem, and develope new beamformers, capable of high reverberation suppression. We finished our study with an experimental analysis demonstrating the capabilities of the proposed beamformers to suppress both noise and reverberations in various simulated environments. The experimental results we've presented prove the superiority of the proposed beamformers over the MVDR beamformer, and justify the advantages of the suggested signal decomposition.

5.2 Future Research

Our methods though promising are far from being efficiently implemented. More work can be done on adaptive LMS frameworks similar to those of the TF-GSC but adjusted to other beamformers beside the MVDR.

Different approaches to the estimation of the RTF can be examined and integrated into the algorithm, the GEVD method suggested in [10] would be a good starting point.

A statistical model which account for the correlation between the early reflection and late reverberation should be considered, allowing a better estimation of the full to early RTF.

Integrating a multi-channel hypothesis model, by which the probability of speech presence is estimated can reduce estimation errors and improve performance.

Inter-frame correlation can be utilized in the design of the beamformer, replacing gains with temporal filters as suggested in [14].

The proposed algorithms can also be used as a framework for blind source separation, switching between competing speakers.

Bibliography

[1] Ann Spriet, Marc Moonen , Jan Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Process.*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004.

[2] Simon. Doclo and Marc. Moonen, "Robustness of SVD-based optimal filtering for noise reduction in multi-microphone speech signals," in *Proc. Int. Workshop Acoust. Echo Noise Contr.*, Pocono Manor, PA, pp. 80–83, Sept. 1999.

[3] A. Spriet, M. Moonen, and J. Wouters, "A multi-channel subband generalized singular value decomposition approach to speech enhancement," *European Transactions on Telecommunications*, vol. 13, no. 2, pp. 149–158, Mar.-Apr. 2002.

[4] Meng Er, A Cantoni, "Derivative constraints for broadband element space antenna array processors," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-31, no. 6, pp. 1378–1393, Dec. 1983.

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

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

[7] Yutaka Kaneda, Juro Ohga, "Adaptive microphone array system for noise reduction," *J. Acoust. Soc. Am.* vol. 76, Issue S1, pp. S84-S84, 1984.

[8] L.J. Griffiths, C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag,* AP-30, pp. 27–34, 1982.

[9] 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, Sept. 1997.

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

[11] J. Bitzer, K. Simmer, and K.-D. Kammeyer, "Theoretical noise reduction limits of the generalized sidelobe canceller for speech enhancement," in *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Process*. (ICASSP), vol. 5, pp. 2965–2968, Mar. 1999.

[12] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Language Processing*, vol. 14, pp. 1218-1234, July. 2006.

[13] E.A.P. Habets, J. Benesty, I. Cohen, S. Gannot and J. Dmochowski, "New insights into the MVDR beamformer in room acoustics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 158-170, Jan. 2010.

[14] J. Benesty, J. Chen and E.A.P. Habets, *"Speech Enhancement in the STFT Domain"*, *Springer Briefs in Electrical and Computer Engineering*, Springer-Verlag, ISBN 978-3-642-23249-7, Sept. 2011.

[15] I. Cohen, "Relative transfer function identification using speech signals," *Special Issue of the IEEE Trans. Speech and Audio Processing on Multi-channel Signal Processing for Audio and Acoustics Applications*, vol. 12, pp. 451-459, Sep. 2004.

[16] D. G. Manolakis, V. K. Ingle, and S. M. Kogan, *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering and Array Processing*. New York: McGraw-Hill, 2000.

[17] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech and Audio Processing*, vol. 11, pp. ,466-475, Sep. 2003.

[18] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, pp. 113-116, Apr. 2002.

[19] E.A.P. Habets, S. Gannot and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Processing Letters*, vol. 16, pp. 770-773, Sept. 2009.

[20] E.A.P. Habets "Speech dereverberation using statistical reverberation models," in *Speech Dereverberation* Chapter 3, P.A. Naylor and N.D. Gaubitch (Eds.), Springer, 2010.

[21] http://home.tiscali.nl/ehabets/rir_generator.html

[22] The DARPA TIMIT acoustic-phonetic continuous speech corpus, Nat. Inst. Standards Technology. NIST Speech Disc 1-1.1 [CDROM], Oct. 1991.

[23] Franklin, J. N. *Matrix Algebra.* New Jersey: Prentice-Hall, 1968.

[24] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short time Fourier transform domain," *IEEE Signal Processing Letters*, vol. 14, pp. 337-340, Sept. 2007.

[25] E. Habets and J. Benesty, "A perspective on frequency-domain beamformers in room acoustics," *IEEE Trans. Audio, Speech, Language Processing*, vol. 20, pp. 947-960, Mar. 2012.

[26] W. Herbordt and W. Kellermann, "Adaptive beamforming for audio signal acquisition," in *Adaptive Signal Processing: Applications to Real-World Problems,* Chapter 6, J. Benesty and Y. Huang, (Eds.) Berlin, Germany: Springer-Verlag,, ch. 6, pp. 155-194, 2003.

[27] Y. Huang and J. Benesty, "A multi-frame approach to the frequency-domain single-channel noise reduction problem," *IEEE Trans. Audio, Speech, Language Processing*, vol. 20, pp. 1256-1269, May 2012.

[28] M. Souden, J. Benesty, and S. Affes, "On the global output SNR of the parametrized frequency-domain multichannel noise reduction Wiener filter," *IEEE Signal Processing Letters*, vol. 17, pp. 425-428, Sep. 2010.

[29] B. Widrow and S. D. Stearns, Eds., *Adaptive Signal Processing*. New York: Prentice-Hall, 1985.

[30] H. Kuttruff, *Room Acoustics,* 4th edn. Taylor and Frances, 2000.

[31] S. Gannot and I. Cohen, *"Adaptive Beamforming and Postfiltering*," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, (Eds.) Berlin, Germany: Springer-Verlag, ch. 47, 2008.

[33] E. Habets and J. Benesty, "A perspective on frequency-domain beamformers in room acoustics," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 3, pp. 947–960, Mar. 2012.

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

[35] E. A. P. Habets and I. Cohen and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *Journal of the Acoustical Society of America*, vol. 124, Issue 5, pp. 2911-2917, Nov. 2008.

[36] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

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

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

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

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

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

[42] G. Reuven, S. Gannot, and I. Cohen, "Dual-source transfer-function generalized sidelobe canceller," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 711–727, May 2008.

[43] S. Doclo, A. Spriet, and M. Moonen, "Efficient frequency-domain implementation of speech distortion weighted multi-channel Wiener filtering for noise reduction," in Proc. EUSIPCO, pp. 2007–2010, 2004.

[44] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1420–1434, Sept. 2009.

[45] A. Krueger, E. Warsitz, and R. Haeb-Umbach, "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 206–219, Jan. 2011.

[46] Polack, J.D, "Playing billiards in the concert hall: the mathematical foundations of geometrical room acoustics," *Appl. Acoust*, 235–244, 1993.

[47] M. R. Schroeder, "Frequency-correlation functions of frequency responses in rooms," *The Journal of the Acoustical Society of America*, vol. 34, no. 12, pp. 1819–1823, 1962.

Appendix A

.1 Derivation of the RBT beamformer

We start by writing the optimization problem the RBT beamformer solves

$$\min_{\overline{h}[l,m]} \phi_{X_{1,c}}[l,m] \left| \overline{h}^H[l,m] i_N - 1 \right|^2 S.T. \quad \overline{h}^H[l,m] \overline{r}\,[l,m] = 0$$
$$\overline{h}^H[l,m] \Phi_{vv}[l,m] \overline{h}[l,m] = \gamma \left[\phi_{V_1}[l,m]\right]. \tag{1}$$

The Lagrangian of the minimization problem in (1) is given by

$$L\,[l,m] = \phi_{X_{1,c}}[l,m] \left(\overline{h}^H[l,m] i_N i_N^H \overline{h}[l,m] - \overline{h}^H[l,m] i_N - i_N^H \overline{h}[l,m] + 1 \right)$$
$$+ \lambda \overline{h}^H[l,m] \overline{r}\,[l,m] + \eta \overline{h}^H[l,m] \Phi_{vv}[l,m] \overline{h}[l,m].$$

To find the optimal filter $\overline{h}[l,m]$ we compare the derivative of $L\,[l,m]$ to zero

$$\frac{d}{d\overline{h}^H} L\,[l,m] = \phi_{X_{1,c}}[l,m] i_N i_N^H \overline{h}[l,m] - \phi_{X_{1,c}}[l,m] i_N + \lambda \overline{r}\,[l,m] + \eta \Phi_{vv}[l,m] \overline{h}[l,m]$$
$$= \left(\phi_{X_{1,c}}[l,m] i_N i_N^H + \eta \Phi_{vv}[l,m] \right) \overline{h}[l,m] - \phi_{X_{1,c}}[l,m] i_N + \lambda \overline{r}\,[l,m]$$
$$= \Phi_{p,\eta}[l,m] \overline{h}[l,m] - \phi_{X_{1,c}}[l,m] i_N + \lambda \overline{r}\,[l,m]$$

where $\Phi_{p,\eta}[l,m] = \left(\phi_{X_{1,c}}[l,m] i_N i_N^H + \eta \Phi_{vv}[l,m] \right)$,

$$\frac{d}{d\overline{h}^H} L\,[l,m] = 0 \iff \Phi_{p,\eta}[l,m] \overline{h}[l,m] - \phi_{X_{1,c}}[l,m] i_N + \lambda \overline{r}\,[l,m] = 0$$
$$\iff \overline{h}[l,m] = \Phi_{p,\eta}^{-1}[l,m] \left(\phi_{X_{1,c}}[l,m] i_N - \lambda \overline{r}\,[l,m] \right). \tag{2}$$

To find to λ we apply (2) to the first constraint in (1)

$$
\begin{aligned}
\overline{h}^H[l,m]\overline{r}\,[l,m] &= 0 \iff \\
\left(\phi_{X_{1,c}}[l,m]i_N^H - \lambda\bar{r}^H\,[l,m]\right)\Phi_{p,\eta}^{-1}[l,m]\overline{r}\,[l,m] &= 0 \iff \\
\lambda\bar{r}^H\,[l,m]\,\Phi_{p,\eta}^{-1}[l,m]\overline{r}\,[l,m] &= \phi_{X_{1,c}}[l,m]i_N^H\Phi_{p,\eta}^{-1}[l,m]\overline{r}\,[l,m] \\
\lambda &= \frac{\phi_{X_{1,c}}[l,m]i_N^H\Phi_{p,\eta}^{-1}[l,m]\overline{r}\,[l,m]}{\bar{r}^H\,[l,m]\,\Phi_{p,\eta}^{-1}[l,m]\overline{r}\,[l,m]} \qquad (3) \\
&= \frac{\phi_{X_{1,c}}[l,m]\overline{r}^H\,[l,m]\,\Phi_{p,\eta}^{-1}[l,m]i_N}{\bar{r}^H\,[l,m]\,\Phi_{p,\eta}^{-1}[l,m]\overline{r}\,[l,m]}.
\end{aligned}
$$

Applying (3) to (2) we get the final form of the RBT beamformer

$$
\begin{aligned}
\overline{h}[l,m] &= \Phi_{p,\eta}^{-1}[l,m]\left(\phi_{X_{1,c}}[l,m]i_N - \bar{r}\,[l,m]\,\frac{\phi_{X_{1,c}}[l,m]\overline{r}^H\,[l,m]\,\Phi_{p,\eta}^{-1}[l,m]i_N}{\bar{r}^H\,[l,m]\,\Phi_{p,\eta}^{-1}[l,m]\overline{r}\,[l,m]}\right) \\
&= \phi_{X_{1,c}}[l,m]\Phi_{p,\eta}^{-1}[l,m] \times \left[I_{NxN} - \frac{\overline{r}\,[l,m]\,\overline{r}^H\,[l,m]\,\Phi_{p,\eta}^{-1}[l,m]}{\overline{r}^H\,[l,m]\,\Phi_{p,\eta}^{-1}[l,m]\overline{r}\,[l,m]}\right] i_N.
\end{aligned}
$$

The inverse of $\Phi_{p,\eta}[l,m]$ is derived using Woodbary's identity

$$
\begin{aligned}
\Phi_{p,\eta}^{-1}[l,m] &= \left(\phi_{X_{1,c}}[l,m]i_N i_N^H + \eta\Phi_{vv}[l,m]\right)^{-1} \\
&= \frac{1}{\eta}\Phi_{vv}[l,m]^{-1} - \phi_{X_{1,c}}[l,m]\frac{\frac{1}{\eta}\Phi_{vv}[l,m]^{-1}i_N i_N^H\frac{1}{\eta}\Phi_{vv}[l,m]^{-1}}{1+\phi_{X_{1,c}}[l,m]i_N^H\frac{1}{\eta}\Phi_{vv}[l,m]^{-1}i_N} \\
&= \frac{1}{\eta}\Phi_{vv}[l,m]^{-1} - \phi_{X_{1,c}}[l,m]\frac{\Phi_{vv}[l,m]^{-1}i_N i_N^H\Phi_{vv}[l,m]^{-1}}{\eta^2+\eta\phi_{X_{1,c}}[l,m]i_N^H\Phi_{vv}[l,m]^{-1}i_N}.
\end{aligned}
$$

.

הפחתת הדהודים ורעשי רקע מאותות דיבור באמצעות מערכי מיקרופונים

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר
מגיסטר למדעים בהנדסת חשמל

אליאב בניסטי

הוגש לסנט הטכניון – מכון טכנולוגי לישראל

אוגוסט 2014 חיפה אב תשע"ד

הפחתת הדהודים ורעשי רקע מאותות דיבור באמצעות מערכי מיקרופונים

אליאב בניסטי

עבודת הגמר נעשתה בהנחיית פרופ' ישראל כהן מהפקולטה להנדסת חשמל בטכניון ופרופ' ג'ייקוב בניסטי מאוניברסיטאת קוויבק.

תודות

ברצוני להביע את תודתי והערכתי, לפרופסור ישראל כהן על הנחייתו המסורה המדרבנת למצויינות, עזרתו המקצועית, ויחסו המבין והאנושי. בנוסף ארצה להודות לפרופסור ג'ייקוב בניסטי על עצותיו המועילות.

תודה מיוחדת מגיעה למשפחתי אמי ורדה ואבי חיים ואחיותיי לירז ודורין שעמדו לצדי עודדו ודירבנו להצלחה. ולאחיינתי הראשונה גלי, מי יתן ותגדלי לעולם יוצר ומתחדש.

תקציר

אותות דיבור המוקלטים בסביבה סגורה, לרוב מלווים בהדהודים (reverberations) המוחזרים מקירות החדר והעצמים הנמצאים בו. את האות המהדהד ניתן למדל כתוצר של קונבולוציה בן סיגנל הדיבור הישיר לתגובת התדר של החדר (RoomImpulseResponse). את תגובת התדר ניתן לחלק לשלושה חלקים עיקריים: הסיגנל הישיר המגיע ראשון אל המיקרופון, הדהודים מוקדמים המשפרים מעט את איכות האות המוקלט, והדהודים מאוחרים הפוגעים במובנות הסגינל והופכים אותו לפחות ברור. ההדהודים המאוחרים מרכיבים את רובה של תגובת התדר של החדר, ומאופיינים על ידי מעטפת הדועכת אקספוננציאלית. הזמן שבו לוקח לאנרגיית המעטפת של תגובת התדר של החדר לרדת לכ - $60\,dB$ מהמקסימום נקרא זמן הרווברציה של החדר ומסומן ב T_{60}. ככל שפרק זמן זה ארוך יותר, כמות ההדהודים המאוחרים בחדר גדולה יותר. המטרה העיקרת של אלגוריתמים להסרת הדהודים היא חילוץ הסיגנל הישיר וההדהודים המוקדמים הנלווים אליו. מכיוון שהסיגנל הרצוי וההדהודים המאוחרים חולקים מספר תדרים משותפים, אלגוריתמים המסתמכים על אינפורמציה זמנית בלבד עשויים להנחית חלקים רצויים בסיגנל ולגרום לעיוותים חמורים. מכאן שאלגוריתם אלו אינם מתאימים לטיפול בבעית הסרת הידהודים מאוחרים בסביבה רועשת ויש צרוך בפתרונות אחרים.

בשנים האחרונות ישנה התעניינות רבה באלגוריתמים המבוססים על מערכי מיקרופונים. אלגוריתמים אלו שמקורם בתקשורת אלחוטית, עושים שימוש במידע המרחבי והזמני בכדי לחלץ אות רצוי ממספר אותות המוקלטים על ידי מערך המכיל מספר מיקרופונים. היכולת של אלגוריתמים אלו להעדיף כיוון מסויים ממנו מגיע אות אל מערך המיקרופונים תוך כדי הנחתה של כיוונים אחרים, הופכת אותם אופטימלים להפחתת הדהודים מאוחרים ורעשי רקע. רוב האלגורתמים המוזכרים בספרות ומבוססים על מערכי מקרופונים, תוככנו להתמודד עם רעש המגיע למערך מכיוון מסויים. שלא כמו רעש כיווני, ההדהודים המאוחרים מגיעים מהרבה כיוונים בו זמנית, ולכן מרבית האלגוריתם הקיימים אינם מתאימים לטיפול בבעית הפחתת הדהודים מאוחרים בסביבה רועשת (dereverberation), ויש צרוך בהתמאה שלהם לסביבות מהדהדות.

המטרה העיקרית של עבודה זו היא לפתח אלגורתמים לחילוץ הסיגנל הנקי, המגיע ישירות מהדובר למיקרופון, מתוך סיגנלים שהוקלטו בסביבה רועשות ומהדהדת. בכדי להשיג מטרה זו, נעשה שימוש במעצבי אלומה (beamformers) אדפטיבים, המוגדרים במרחב הזמן תדר (ShortTimeFourierTransform). המונח מעצב אלומה (beamformer) מתייחס לתכנון מסנן זמני ומרחבי, המאפשר חילוץ אות רצוי ממספר אותות המוקלטים במערך מיקרופונים, זאת על ידי משקול כל אות המגיע מאחד המקרופונים וסכימה לאורך מפתח המערך. המשקלים נבחרים מתוך מזעור פונקציות עלות שונות כתלות בדרישות היישום והאילוצים הנלווים אליו. מעצבי אלומה מתחלקים לשני קבוצות עיקריות, כאלה המאפשרים עיוות של הסיגנל הרצוי על חשבון הגדלת יחס הסיגנל לרעש ביציאת מעצב האלומה, וכאלה

המאלצים פיתרון שאינו מאפשר עיוות. לאחרונה הוצג מצעב אלומה המאפשר שקלול בין כמות הרעש במוצא לעיוות הנגרם לסיגנל הרצוי (Tradeoffbeamformers). מעצב אלומה יעיל מוגדר ככזה המצליח להגדיל את יחס האות לרעש (SignaltoNoiseRatio), תוך כדי שהוא מצליח להפחית כמה שיותר הידהודים, ומעוות כמה שפחות את הסיגנל הרצוי. פתרון אמין להפחת הדהודים מאוחרים ורעשי רקע, דרוש בישומים רבים בהם קיים מרחק בן המיקרופון לדובר. יישומים אלו כוללים בין היתר: טלפונים סוללרים, שיחות ועידה, עזרי שמע, וממשקי אדם מכונה למינהם.

את האנליזה בעבודה זו, ניתן לחלק לשני חלקים עיקריים. בחלק הראשון נציע דרך להמיר מעצבי אלומות שיועדו להסרת רעשים בלבד, לכאלו המסוגלים להפחית הידהודים בסביבה רועשת. הניתוח בגישה זו מתבסס על שיערוך וקטור פונקציות התמסורת היחסיות (RelativeTransferFunction), המוגדר כווקטור המכיל את היחסים בין תגובת התדר המתארת את ההדהודים המוקדמים המגיעים למיקרופון הרפרנס (ראשון במערך), לתגובות התדר של שאר המיקרופונים במערך. בעזרת וקטור התמסורות היחסיות, נתאר את וקטור המדידות של מעצבי האלומה כפונקציה של אות הדיבור הרצוי, המורכב מהסיגנל הישיר והנקי, ומההדהודים המוקדמים המגיעים למיקרופון. באמצעות תיאור זה נגדיר מדדי ביצועים מעניינים, שיאפשרו לנו למדוד את ביצועי מעצבי האלומה המוצעים. המדדים המהותיים בהקשר לבעיית הסרת ההידודים המאורחים בסביבה רועשת הם: כמות הרעש במוצע מעצב האלומה, העיוות הנגרם לסיגנל הרצוי, וכמות ההדהודים המאוחרים שנחסמו על ידי מעצב האלומה. בעזרת מדדים אלו נציג שלוש בעיות אופטימיזציה שפתרונן יניב שלושה מעצבי אלומה ידועים (Wiener,MVDR,Tradeoff), המותאמים להתמודד עם בעיית הסרת ההידהודים המאוחרים בסביבה רועשת. בכדי לשערך את וקטור התמסורות היחסיות נשתשמש בגישה המבוססת על מודל סטיסטי לתגובת התדר של החדר, בעזרת מודל זה נשערך את הספקטרום של ההידהודים המאוחרים ונשתמש בו בכדי לייצר מערכת משוואות שפתרונה יניב את המשערך הרצוי. בפתרון מערכת המשוואות הנ"ל נתשמש אך ורק בחלונות זמן המכילים בתוכם דיבור, שכן חלונות זמן המכילים רק רעש אינם נושאים אינפורמציה על הקשר בן המקור לסיגנל הנקלט, ועלולים להפחית את דיוק המשערך. פתרון מערכת המשוואת מתקבל על ידי הפעלת אלגוריתמים רקורסבי (NormalizeLeastMeans Squares). על ידי ניתוח נסיוני מעמיק, נראה כי מעצבי האלומה המוצעים מצליחים לחלץ בצורה טובה את הסיגנל הרצוי בסביבות מהדהדות שונות, ותחת שתי שדות רעש שונים, רעש כיווני שעיקרו מגיע אל המערך מכיוון מסויים, ורעש דיפוזיבי המתפזר בחדר באופן אחיד. בנוסף נראה כי ביצועיי מעצבי האלומה המוצעים טובים מאלו של אלגוריתם ה־Transfer) FunctionGeneralizedSidelobeCanceler), באספקטים של הפחתת רעשים, עיוות סיגנל נמוך והפחתת הידהודים מאוחרים.

בחלק השני נציג גישה חדשה לפיתוח מעצבי אלומה המסוגלים להפחית הדהודים. הגישה מתבססת על שינוי המודל הקלאסי לאות הנקלט במערך המיקורפונים, ופירוקו לשלוש גורמים מרכזיים: הראשון מייצג את האות הישיר המגיע אל המיקרופון, השני מייצג הפרעה

קורלטיבית לא רצויה הכוללת גם את ההדהודים המאוחרים, והשלישי מייצג את רעש הרקע הלא קורלטיבי. בעזרת פירוק זה נוכיח כי ישנה קורלציה בין סיגנל עיוות אות הדיבור, לסיגנל ההפרעה הקורלטיבית. סיגנלים אלו מהווים חלק מהשגיאה הריבועית הממוצעת (Mean SquareError) בן האות הישיר לאות במוצא מעצב האלומה. לשגיאה הריבועית חשיבות רבה בתהליך תכנון מעצבי האלומה, והיא עומדת בבסיס בעיות האופטימציה בעבודה זו. בכדי לפשט את תהליך התכנון ולאפשר טיפול נפרד בהפחתת הדהודים מאוחרים והפחתת עייוותים, נציע לאפס את הקורלציה בן סיגנלי השגיאה על ידי אכיפת אילוץ. מנקודת מבט זו עולות בעיות אופטימזציה חדשות, המערבות בתוכן את הליך הסרת ההדהודים המאוחרים באופן ישיר ובלתי תלוי בתהליך הסרת רעשי הרקע. הפרדה זו מגדילה את פוטנציאל הפחתת ההדהודים המאוחרים, ומאפשרת יותר גמישות בתהליך התכנון של מעצב האלומה. מפתרון בעיות אופטימצזיה אלו מתקבלים מעצבי אלומה חדשים המסוגלים להפחית בצורה ניכרת רעש והדהודים, ואינם מעוותים משמעותית את אות הדיבור. בחינת מעצבי האלומה המוצעים בתנאי סביבה ורעש שונים, מדגימה את יכולתם להפחית הדהודים מאוחרים ורעשי רקע, ומראה שיפור ניכר בביצועים ביחס לביצועיי מעצב האלומה המוכר - Minimum (VarianceDistortionlessResponse). שיפור זה מוכיח את כדאיות הגישה, ונותן משנה תוקף לפירוק המוצע.