Design and Analysis of Quadratic and Multistage Beamformers

Gal Itzhak

Design and Analysis of Quadratic and Multistage Beamformers

Research Thesis

Submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

Gal Itzhak

Submitted to the Senate

of the Technion — Israel Institute of Technology

Tevet 5782 Haifa December 2021

This research was carried out under the supervision of Prof. Israel Cohen and Prof. Jacob Benesty in the Faculty of Electrical and Computer Engineering.

List of Publications

Journal Papers

- G. Itzhak, J. Benesty, and I. Cohen, "Nonlinear Kronecker product filtering for multichannel noise reduction," *Speech Communication*, vol. 114, pp. 49–59, 2019
- G. Itzhak, J. Benesty, and I. Cohen, "Quadratic approach for single-channel noise reduction," *EURASIP Journal of Audio, Speech and Music Processing*, 7 (2020), pp. 1–14, April 2020
- G. Itzhak, J. Benesty, and I. Cohen, "On the Design of Differential Kronecker Product Beamformers," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 29, pp.1397–1410, 2021
- G. Itzhak, J. Benesty, and I. Cohen, "Multistage Approach for Steerable Differential Beamforming with Rectangular Arrays," submitted to *Speech Communication*, December 2021

Conference Papers

- G. Itzhak, J. Benesty, and I. Cohen, "Quadratic Beamforming for Magnitude Estimation," in *Proc. 29th European Signal Processing Conference (EUSIPCO)*, August 2021
- G. Itzhak, I. Cohen, and J. Benesty, "Robust Differential Beamforming with Rectangular Arrays," in *Proc. 29th European Signal Processing Conference (EUSIPCO)*, August 2021

Acknowledgements

I would like to express my sincere gratitude to my principal supervisor, Prof. Israel Cohen, for his guidance. His great advice, consistent support and helpful ideas had a huge impact on my work.

I would also like to thank Prof. Jacob Benesty, for sharing his innovative ideas and expertise in the field.

Finally, I wish to thank my family for the understanding and help during the journey of my PhD.

Contents

List of Figures

Abstract

This dissertation introduces design approaches of quadratic and multistage beamformers whose properties may come in handy in a variety of practical applications, such as communication systems, speech recognition, human-to-machine interfaces and more.

The problem of reducing undesirable noise and interferences from observation signals is a fundamental problem in acoustic signal processing, for which numerous schemes and algorithms have been suggested over the years. Broadly, these algorithms can be divided into two classes: single-channel noise reduction (SCNR) and multichannel noise reduction (MCNR), whose primary difference is whether a single microphone or multiple microphones are utilized. In this thesis we address them both.

This research thesis focuses on four topics of interest. The first is the quadratic beamforming approach. Traditionally, beamforming in the frequency domain is performed by applying a complex-valued linear filter to a vector of noisy observations to yield an estimate of the desired signal (complex) value. Such approaches are almost exclusively based on the second-order statistics of the observations, even though significant information may underlie in higher-order statistics. In the presented study, we focus on the estimation of the desired signal power and propose a quadratic beamforming approach which makes a direct use of higher-order statistics. We show that the quadratic approach outperforms the traditional linear approach, in particular when the input signal-to-noise ratio (SNR) is low or the number of sensors is small.

The second topic this thesis addresses is a quadratic approach for SCNR, in which we consider the interframe correlation property in the short-time Fourier transform (STFT) domain. We utilize the quadratic formulation and propose a quadratic version for the maximum SNR filter. We demonstrate that the quadratic maximum SNR

filter is superior to its linear counterpart, and may in fact achieve a theoretical unbounded approximate SNR gain, assuming the noise statistics is accurately given. The performance gap is particularly significant in low input SNRs.

In the third study, we address the differential microphone array (DMA) beamforming concept. We propose a flexible approach for deriving multistage differential beamformers by employing the Kronecker product (KP) decomposition of the global beamformer into two independent sub-beamformers. We present the notion of multistage differential KP beamformers and analyze the influence of three inherent design parameters which allow a high design flexibility. We demonstrate that the multistage differential KP beamforming approach outperforms previous approaches, depending on the scenario and the selection of the design parameters.

In the fourth study, we focus on DMAs from the beam steering perspective. We present a multistage approach for steerable differential beamforming by exploiting uniform rectangular arrays (URAs). At first, we differentiate along the columns and rows of the observation signals, respectively; then, we design and apply a rectangular differential beamformer. We show that the proposed approach may significantly improve the directivity of the resulted beamformer at the expanse of white noise amplification, depending on the design parameters selection. We demonstrate that the proposed approach outperforms common linear approaches, particularly when the incident angle is far from the endfire direction.

Abbreviations and Notations

Abbreviations

DF	:	directivity factor
DMA	:	differential microphone array
DNR	:	diffuse noise reduction
ENR	:	error-to-noise ratio
FBR	:	front-to-back ratio
IR	:	interference reduction
KP	:	Kronecker-product
MCNR	:	Multichannel noise reduction
MDF	:	maximum directivity factor
MFBR	:	maximum front-to-back ratio
MWNG	:	maximum white noise gain
NCMDF	:	null-constrained maximum directivity factor
NCMWNG	:	null-constrained maximum white noise gain
NS	:	null steering
PESQ	:	perceptual evaluation of speech quality
RIR	:	room impulse response
RR	:	reverberations reduction
SCNR	:	Single-channel noise reduction
SNR	:	signal-to-noise ratio
STFT	:	short-time Fourier transform
STOI	:	short-time objective intelligibility
UCA	:	uniform circular array

ULA	:	uniform linear array
URA	:	uniform rectangular array
WNG	:	white noise gain
WNR	:	white noise reduction

Notations

$\mathbf{a}_\theta$	:	Steering vector corresponding to sub-beamformer a
$\mathbf{b}_\theta$	:	Steering vector corresponding to sub-beamformer b
$\mathcal{B}[\mathbf{h}]$	:	Beampattern of beamformer $\mathbf{h}$
c	:	Speed of sound
$\mathbf{d}_\theta$	:	Steering vector
$\mathcal{D}[\mathbf{h}]$	:	Directivity factor of beamformer $\mathbf{h}$
$E[\cdot]$	:	Mathematical expectation
f	:	Frequency
f_s	:	Sampling frequency
$\mathcal{G}[\mathbf{h}]$	:	SNR gain of beamformer $\mathbf{h}$
$\mathbf{h}$	:	Linear beamformer
$\widetilde{\mathbf{h}}$	:	Quadratic beamformer
$\mathbf{I}_M$	:	$M \times M$ identity matrix
M	:	Number of microphones in an array
$M_\mathbf{a}$	:	Number of microphones in sub-beamformer a
$M_\mathbf{b}$	:	Number of microphones in sub-beamformer b
M_x	:	Number of microphones in a ULA along the x-axis
M_y	:	Number of microphones in a ULA along the y-axis
N	:	Length of observations vector in the STFT domain
P	:	Number of multistage differentiation stages in a ULA
P_c	:	Number of multistage differentiation stages along the columns of a URA
P_r	:	Number of multistage differentiation stages along the rows of a URA
V_m	:	Additive noise component in the mth microphone
$\mathbf{v}$	:	Additive noise vector
$\mathcal{W}[\mathbf{h}]$	:	White noise gain of beamformer $\mathbf{h}$
X_m	:	Desired signal component in the mth microphone
$\mathbf{x}$	:	Desired signal observations vector
Y_m	:	Noisy observation of the mth microphone
$\mathbf{y}$	:	Noisy observations vector

$\mathbf{y}_{(P_c,P_r)}$	:	Vector of differentials of a URA
$\mathbf{y}_{(P)}$	:	Vector of differentials of a ULA
$\mathbf{\Gamma}_{\mathrm{d}}$	:	The pseudo-coherence matrix of the diffuse noise
δ	:	Interelement spacing
δ_{x}	:	Interelement spacing along the x-axis
δ_{y}	:	Interelement spacing along the y-axis
Δ^p	:	pth-order forward spatial difference operator
$\mathbf{\Delta}_{(P_c)}$	:	P_cth-order differentiation matrix along the columns of a URA
$\mathbf{\Delta}_{(P_r)}$	:	P_cth-order differentiation matrix along the rows of a URA
$\mathbf{\Delta}_{(P_c,P_r)}$	:	2-D differentiation matrix of a URA
θ	:	Azimuth angle
θ_{s}	:	Desired signal incident angle
$\mathbf{\Phi}_{\mathbf{v}}$	:	Correlation matrix of the additive noise vector
$\mathbf{\Phi}_{\mathbf{y}}$	:	Correlation matrix of the observations vector
$\mathbf{\Phi}_{\widetilde{\mathbf{y}}}$	:	Correlation matrix of the modified observations vector
ϕ	:	The variance of the diffuse noise
ϕ_V	:	Variance of the additive noise in the reference microphone
ϕ_X	:	Variance of the desired signal in the reference microphone
ω	:	Angular frequency
$\mathrm{tr}(\cdot)$	:	Linear trace operator
$\otimes$	:	Kronecker-product operator
$(\cdot)^*$	:	Complex conjugation
$(\cdot)^T$	:	Transpose operator
$(\cdot)^H$	:	Conjugate transpose operator
$\jmath$	:	Imaginary unit
$\mathrm{sinc}(x)$	:	$\sin x/x$

Chapter 1

Introduction

1.1 Background and Motivation

Many modern applications in a wide variety of areas, from speech recognition and communications to speaker identification and human-to-machine systems, are required to operate in noisy environments. Noise fields, in many cases, significantly deteriorate the speech signal quality, thus damaging the functionality of communication and speech recognition systems. The problem of enhancing speech, or reducing noise, has attracted many researchers over the years, who suggested numerous schemes and algorithms in multiple processing domains.

With the growing demand for robust noise reduction capabilities, MCNR methods are often employed in order to exploit spatial information. This additional information allows, in many cases, to attain a considerable amount of noise reduction while preserving the desired signal distortionless [BCHC09]. Often referred to as beamformers, MCNR methods may be designed and implemented in various domains.

Time-domain beamformers are the easiest to implement, as the filters are applied directly to the noisy observations, typically generating a single speech sample estimate at a time. It is also possible to estimate a vector of successive speech samples simultaneously. However, such beamformers tend to suffer from high computational complexity [BC11, BCC18, BCB19].

Most commonly with communication and speech signals, processing is done in the frequency domain. That is, a frame of consecutive time-domain samples is transformed

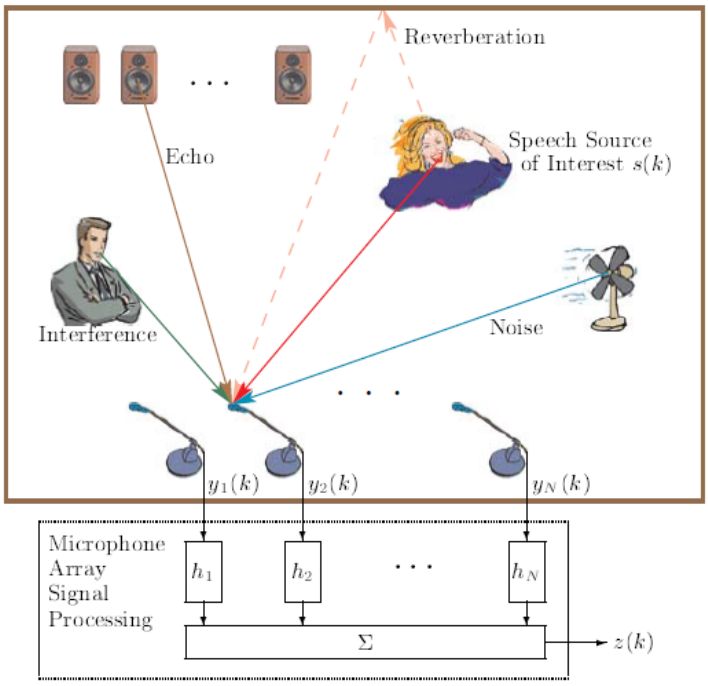


Figure 1.1: Illustration of microphone array signal processing with different sources of noise [BCH08].

into the frequency domain by applying the fast Fourier transform (FFT), yielding a set of analysis coefficients, which can be processed more efficiently than the time-domain samples. This is particularly significant with multichannel methods, in which time-domain noisy observations are sampled simultaneously in multiple sensors. These methods typically seek for a linear optimal solution with respect to some criterion, looking to estimate both the desired signal phase and magnitude [JD92, VT04, BCC18, BCC19].

Originally proposed in [Cap69], Capon's minimum variance distortionless response (MVDR) beamformer has been investigated in theoretical studies from a variety of aspects [SBA10]. The linear MVDR, which operates directly on a vector of transformed noisy observations, is shown to be optimal in terms of the residual noise energy, under the restriction of zero desired signal distortion. Moreover, it has inspired numerous variations, e.g., the minimum power distortionless response (MPDR) [VT04], which avoids the estimation of the noise-only correlation matrix, and the linearly constrained minimum variance (LCMV) beamformer [GJ82], which provides a convenient scheme to cope with spatial interferences by placing nulls in their respective directions.

In many cases, for example, with physically small or low cost systems, an array of sensors may not be available. Consequently, the noise reduction procedure is confined to observations samples acquired by merely a single microphone and the spatial information used in the multichannel case for the purpose of beamforming is lost.

Common approaches for SCNR in the frequncy domain include statistical estimation methods which typically consider the desired signal spectral power more paramount than its spectral phase. Indeed, this property is well-known for speech signals, whose spectral magnitude has received special attention in the context of statistical models and optimal estimators, e.g., a maximum-likelihood spectral magnitude estimator [MM80], short-time spectral [EM84], log-spectral [EM85] and optimally modified log-spectral [Coh02] magnitude estimators, and a maximum *a posteriori* spectral magnitude estimator [WG03]. These celebrated estimators assume that time trajectories in the STFT domain of clean speech and noise signals are independent complex Gaussian random processes. Other statistical models, e.g., super-Gaussian [Mar05, Coh05b, CHRJ09], Gamma [Mar02, EHHJ07] or Laplace [MB03, Coh06] distributions were also investigated, and were demonstrated to be potentially more effective, depending on the desired speech spectral magnitude estimator and the speech conditional variance evolution model. While all the foregoing estimators rely on the strong correlation between magnitudes of successive coefficients (in a fixed frequency) [CB01, Coh05a, BCH12], their derivation is typically cumbersome and requires one to numerically evaluate non-analytical functions following the assumed statistical speech and noise models. Moreover, with the aforementioned spectral magnitude correlation hidden behind first-order recursive temporal processes, additional parameters and lower boundaries must carefully be set to guarantee the model tracking over time.

In the recent years, there has been a growing attempt to exploit the framework of deep neural networks for speech enhancement tasks [ZWW19, PLV$^+$19]. These include, for example, convolutional neural networks [PW19], denoising autoencoders [YZW$^+$20], and feed-forward neural networks [XDDL15]. In contrast to the aforementioned statistical methods, no explicit assumption on the distribution of the speech signal is made. However, methods based on deep neural networks suffer from other drawbacks: their performance is heavily tied to the data used to train them (which, for

itself, is a computationally-expensive operation), their performance on newly-observed speech signals is hard to a priori predict or bound, and the interpretability of the layer blocks with respect to the loss function is lacking. Therefore, in our work, we focus on well-defined mathematical models, whose relation to noisy speech signals is clear.

Despite the lack of spatial information, a beamforming-like approach may be applied to single-channel observations by exploiting the self-correlation property of STFT domain coefficients in a linear manner. That is, instead of explicitly assuming statistical models which depend on unobserved measures, e.g., the *a priori* SNR, it was suggested to employ linear filters which require the second-order statistics of the desired signal and noise. These linear filters are derived within a multi-frame framework that takes into account the interframe correlation of the STFT coefficients from successive time frames and adjacent frequencies [BCH12, HB12, HBLC14]. The multi-frame formulation highly resembles a sensor array formulation, which implies that conventional array filters may be modified for the single-channel case, but with an interframe correlation interpretation rather than spatial sensing. Examples of such filters are the Wiener filter, the MVDR filter [BCH12, HB12], the LCMV filter [BCH12], and the maximum SNR filter [HBLC14].

Many systems are required to operate in broadband conditions. That is, the desired signal is not confined to narrow frequency bandwidth, but rather it consists of a wide range of frequencies. In such cases, it is highly likely for the system to operate similarly throughout the entire spectrum of interest, in order to avoid undesirable distortions. Assuming a multichannel structure, DMAs have been proposed and optimized for almost a century [WOM33, Ols46], with the underlying principle of exploiting acoustic pressure differences among adjacent microphones [BCB19]. This principle implies arrays of small sizes and frequency-invariant beampatterns [BCHD07, BC13, TJCB16].

Typically, in order to design high-order DMAs which were capable of obtaining a significant amount of noise reduction, a multistage approach was taken. That is, the operation of differentiating acoustic pressure observations was successively repeated, in analogy to high-order derivatives of analytic functions [SHC12]. This approach was implemented in the time domain. Unfortunately, it was highly susceptible to array mismatches and imperfections [Buc02, WC16], making it less appealing under prac-

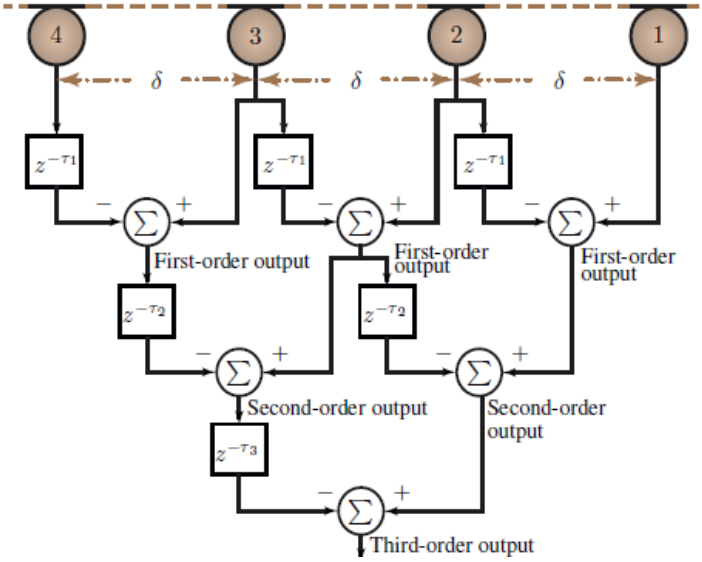


Figure 1.2: Traditional multistage structure of first-, second-, and third-order DMAs in the time domain [BCC15].

tical conditions. Consequently, DMA design in the STFT domain was introduced, providing a robust framework that is based on a single stage with linear matrix operations. Despite its simplicity, it is still capable of satisfying spatial constraints while simultaneously minimizing residual noise, in either fixed or adaptive settings [BCHC09, BCH12, CBP14]. With such an approach, the relation to the traditional time-domain differential beamformers remains vague. In addition, due to the simple single-stage structure and inherent linear nature, the noise reduction capabilities of DMAs are limited. Recently, an innovative DMA design approach has been proposed, generalizing the STFT-domain design approach from the common single-stage structure to a multistage one, in a way that highly resembles the traditional time-domain approach [HBCC20b]. This technique was shown to be effective to reduce diffuse noise and to handle reverberant environments, though a significant drawback was its white noise amplification.

Due to their usefulness, DMAs in the STFT domain were thoroughly analyzed and adapted into many variations. One paramount example is KP beamforming, in which a global beamformer is decomposed into a KP of independent sub-beamformers that may be individually designed and optimized [AFJ10, RdAM16, BCC19, CBC19]. The main advantage of KP beamformers is their great design flexibility. That is, each sub-

beamformer may be optimized by a different criterion, yielding a global beamformer that is "optimized" according to all criteria. The properties of the KP decomposition, that is, the relative sizes of each of the sub-beamformers, set the trade-off for the optimization of the global beamformer.

Differential ULAs have been most commonly addressed in the literature as a consequence of their simplicity and easy-to-analyze nature [BCHC09, BBAS19, BCC19, CBC19]. Unfortunately, they suffer from a few inherent drawbacks. For example, it is well known that in order to attain a high level of directivity, a desired signal source must be located in the endfire direction [CBP14]. In addition, ULAs suffer from a lower-upper plane ambiguity: the beampattern of any ULA is always symmetric with respect to the imaginary line connecting the sensors of the array. Therefore, more sophisticated geometric structures were explored, out of which differential uniform circular arrays (UCAs) have drawn most of the attention [HBC17, BCB18, HCB20a]. Other studies exploited the Jacobi-Anger expansion approximation to refer differential beamforming with arbitrarily-shaped planar arrays [HCB18, HCB$^+$20b]. These approaches did not assume any regular array shape but merely required the positions of the array sensors to be either known in advance or measurable. While they are indeed general, they are highly sensitive to the selection of the expansion's reference point and may result in frequency-variant beampatterns as the array size increases. Therefore, they might not embody a proper beamforming design approach with symmetric array geometries, for which it may be possible to take advantage of the symmetry to circumvent these drawbacks.

Rectangular-shaped arrays are symmetric and useful structures, which may be used to design differential beamformers with asymmetric beampatterns [VT04]. Such arrays, on top of being particularly suitable for rectangular-shaped appliances, may also be designed in a flexible manner. For example, a uniform rectangular beamformer may always be decomposed into two sub-beamformers by employing the Kronecker-product (KP) decomposition. This allows some level of flexibility: the KP decomposition is not unique and each of the sub-beamformers may be independently designed with respect to a different criterion [HBCC20a]. An alternative approach [ICB21] exploits the rectangular geometry to improve white noise robustness at the expanse of array direc-

tivity. However, with this approach, the beam steering property of its corresponding beamformers is not considered and the array directivity deteriorates.

1.2 Main Contributions

In this research thesis, the aforementioned topics and their corresponding drawbacks are addressed. In essence, the thesis provides four main contributions. Two of them relate the quadratic beamforming notion; the other two concern multistage beamforming:

- A quadratic (KP) approach for MCNR is presented and investigated which focuses on the estimate of a desired signal power in the frequency domain. Taking advantage of higher-order statistics of the noisy observations, it is demonstrated to outperform the traditional linear MCNR approach, in particular in low SNRs and with small microphone arrays.

- A quadratic generalization to the linear filtering approach in SCNR is proposed, which is based on the interframe correlation property. It is shown to outperform the traditional linear approach as well celebrated speech enhancement estimators, depending on the noise type.

- A multistage differential KP beamforming approach which provides a generalized and flexible framework to design DMAs whose array directivity and white noise robustness may carefully be tuned.

- A steerable approach for multistage differential beamforming, which exploits a differential URA geometry to allow high array directivity as well as beam steering flexibility.

1.3 Overview of the Thesis

In Chapter 3, a quadratic approach for MCNR in the frequency domain is presented. The underlying idea is to take advantage of higher-order statistics and apply a complex linear filter to a modified observation signal vector. The modified vector is constructed

from the original noisy observations and its elements may be interpreted as the instantaneous correlation coefficients. The filtering product of the KP approach is an estimate of the desired signal spectral power, which is considered more important than the spectral phase in many applications, such as speech enhancement. The spectral phase may be extracted from a conventional beamformer, e.g., the linear MVDR. A modified optimization criterion for deriving quadratic (or, as they are alternatively referred to, KP) filters and present the (quadratic) KP-MVDR and the (quadratic) KP-LCMV.

To evaluate the quadratic MCNR approach, a toy example is analyzed and a series of speech signals simulations in both anechoic and reverberant environments is carried out. Simulations indicate that the proposed KP-MVDR and KP-LCMV beamformers outperform their conventional counterparts when proper temporal smoothing is employed to estimate the correlation matrix of the modified observation signal vector. This is emphasized in particular when the number of sensors is small or when the input SNR is low.

In Chapter 4, a quadratic approach for SCNR is presented, which extends existing multi-frame approaches and takes advantage of the interframe correlation property of speech signals. This property is taken into account in the same manner as previously, but the noise reduction filters are not applied to the observations' vector directly, but rather to its modified version. The modified version is obtained from the Kronecker product of the observations' vector and its complex conjugate. In its mathematical formulation, this approach is rather similar to the quadratic beamforming approach which addresses the MCNR problem. On the contrary, while in the context of MCNR the essence of the innovation is the direct utilization of higher-order statistics, the key idea in this work is a generalization of the single-channel linear filtering approach.

The advantages of the quadratic approach are stressed through the maximum SNR beamformer in the STFT domain. The theoretical subband SNR gain is analyzed in a toy example. Unlike with the linear maximum SNR beamformer, whose SNR gain is strictly bounded, the approximate gain with the quadratic maximum SNR beamformer is potentially unbounded and heavily depends on the error-to-noise ratio (ENR). In practice, the quadratic beamformer is shown to outperform the linear beamformer in

terms of the quality and intelligibility of enhanced speech signals. Moreover, simulations in nonstationary noise environments demonstrate that the quadratic maximum SNR beamformer potentially outperforms well-known speech enhancement methods, depending on the nonstationary noise type.

In Chapter 5, a multistage differential KP beamforming approach is presented. This multistage approach considers a KP decomposition of the global beamformer into two independent sub-beamformers and provides a framework to derive differential KP beamformers according to different criteria. The beamformers are tuned by three design parameters. These parameters allow a high beamforming design flexibility; in particular, previous non-differential or non-KP beamformers may be obtained as a special case. Depending on the type of the beamformer, this flexibility enables one to either mitigate the white noise amplification or improve the array directivity.

Simulations indicate that desired signal reverberations may be attenuated to a greater extent with differential KP beamformers with respect to non-differential and non-KP beamformers. In addition, considering the quality and intelligibility of their respective time-domain enhanced signals, the differential KP approach is shown to outperform the previous approaches.

In Chapter 6, a multistage rectangular approach for steerable differential beamforming is proposed. As a first step, a 2-D differentiation scheme is employed, which operates independently on the columns and rows of the observation signals of a URA. This yields a differentials matrix controlled by two design parameters which indicate the number of differential stages with respect to the URA columns and rows, respectively. Then, as a second step, a rectangular differential beamformer is designed and applied to the vector form of the differentials matrix. It is shown that the first differentiation scheme may significantly improve the directivity of the resulted beamformer at the expanse of white noise amplification. The latter is heavily tied to the design parameters configuration, which is optimized with respect to the desired signal incident angle.

Simulations are performed with four types of multistage rectangular differential beamformers, and their performances are examined in terms of four reduction factors. These factors are calculated from the noisy and enhanced signals in the time domain.

In addition, the quality and intelligibility of the enhances signals are investigated. It is demonstrate that the proposed rectangular differential beamformers outperform common linear differential beamformers in terms of these measures- in particular when the incident angle is far from the endfire direction.

1.4 Organization

This research thesis is organized as follows. Chapter 2 provides a high-level scientific background for the performed research. The contribution of this thesis is elaborated in Chapters 3 to 6. Chapters 3 and 4 are dedicated to the presentation of quadratic beamforming in the for MCNR and SCNR settings, respectively. Chapter 5 introduces a multistage differential KP beamforming approach, and in Chapter 6 a steerable approach for multistage differential beamforming with URAs is proposed. Finally, Chapter 7 concludes this thesis and proposes directions for future research.

Chapter 2

Scientific Background

2.1 Linear filtering with microphone arrays

In this section, we provide a brief overview of the MCNR problem in the frequency domain. Consider an array consisting of M omnidirectional microphones. The received signals at the frequency index f are expressed as [BCH12, BCH08, BIB14]:

$$\begin{aligned} Y_m(f) &= G_m(f)S(f) + V_m(f) \\ &= X_m(f) + V_m(f),\ m = 1, 2, \ldots, M, \end{aligned} \tag{2.1}$$

where $Y_m(f)$ is the m-th microphone signal, $S(f)$ is the unknown speech source, $G_m(f)$ is the acoustic impulse response from the position of $S(f)$ to the mth microphone, $X_m(f) = G_m(f)S(f)$ is the zero-mean convolved speech signal, and $V_m(f)$ is the zero-mean additive noise. It is assumed that $X_i(f)$ and $V_j(f)$ are uncorrelated, i.e., $E\left[X_i(f)V_j^*(f)\right] = 0,\ \forall i, j = 1, 2, \ldots, M$. By definition, the terms $X_m(f),\ m = 1, 2, \ldots, M$ are correlated while the other terms $V_m(f),\ m = 1, 2, \ldots, M$ are only partially correlated. We consider the first microphone as the reference; then, the objective of multichannel noise reduction in the frequency domain is to estimate the desired signal, $X_1(f)$, from the M observations $Y_m(f),\ m = 1, 2, \ldots, M$.

It is more convenient to write the M frequency-domain microphone signals in a

vector notation:

$$\begin{aligned}\mathbf{y}(f) &= \mathbf{g}(f)S(f) + \mathbf{v}(f) \\ &= \mathbf{x}(f) + \mathbf{v}(f) \\ &= \mathbf{d}(f)X_1(f) + \mathbf{v}(f),\end{aligned} \tag{2.2}$$

where

$$\begin{aligned}\mathbf{y}(f) &= \begin{bmatrix} Y_1(f) & Y_2(f) & \cdots & Y_M(f) \end{bmatrix}^T, \\ \mathbf{x}(f) &= \begin{bmatrix} X_1(f) & X_2(f) & \cdots & X_M(f) \end{bmatrix}^T \\ &= S(f)\mathbf{g}(f), \\ \mathbf{g}(f) &= \begin{bmatrix} G_1(f) & G_2(f) & \cdots & G_M(f) \end{bmatrix}^T, \\ \mathbf{v}(f) &= \begin{bmatrix} V_1(f) & V_2(f) & \cdots & V_M(f) \end{bmatrix}^T,\end{aligned}$$

the superscript ^{T} is the transpose operator, and

$$\mathbf{d}(f) = \begin{bmatrix} 1 & \frac{G_2(f)}{G_1(f)} & \cdots & \frac{G_M(f)}{G_1(f)} \end{bmatrix}^T = \frac{\mathbf{g}(f)}{G_1(f)}. \tag{2.3}$$

Expression (2.2) depends explicitly on the desired signal, $X_1(f)$; therefore, (2.2) is the frequency-domain signal model for noise reduction. The vector $\mathbf{d}(f)$ represents the frequency-domain steering vector under the far-field assumption.

Since $\mathbf{y}(f)$ is the sum of two uncorrelated components, its correlation matrix is

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}}(f) &= E\left[\mathbf{y}(f)\mathbf{y}^H(f)\right] \\ &= \phi_{X_1}(f)\mathbf{d}(f)\mathbf{d}^H(f) + \mathbf{\Phi}_{\mathbf{v}}(f),\end{aligned} \tag{2.4}$$

where $\phi_{X_1}(f) = E\left[|X_1(f)|^2\right]$ is the variance of $X_1(f)$, and $\mathbf{\Phi}_{\mathbf{v}}(f) = E\left[\mathbf{v}(f)\mathbf{v}^H(f)\right]$ is the correlation matrix of $\mathbf{v}(f)$.

In the conventional (linear) filtering approach, multichannel noise reduction in the frequency (or time-frequency) domain is performed by applying a complex-valued linear

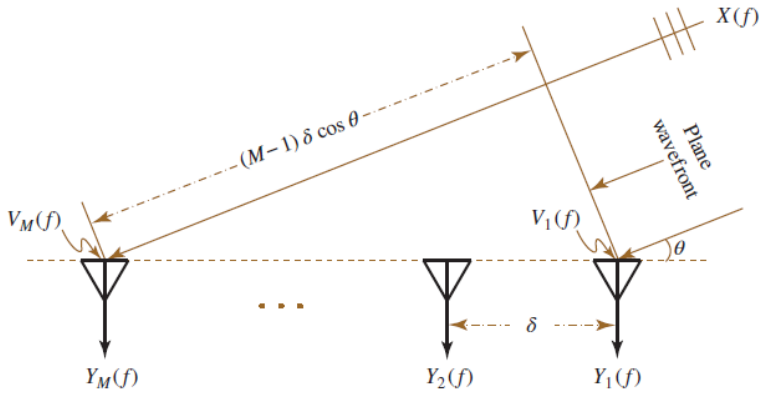


Figure 2.1: Physical illustration of the microphone array signal model [BCC18].

filter, $\mathbf{h}(f)$, of length M, to the observation signal vector, $\mathbf{y}(f)$ [BCH08], i.e.,

$$\begin{aligned}\widehat{X}(f) &= \mathbf{h}^H(f)\mathbf{y}(f) \\ &= X_{\text{fd}}(f) + V_{\text{rn}}(f),\end{aligned} \tag{2.5}$$

where the filter output, $\widehat{X}(f)$, is an estimate of $X_1(f)$, $X_{\text{fd}}(f) = X_1(f)\mathbf{h}^H(f)\mathbf{d}(f)$ is the filtered desired signal, and $V_{\text{rn}}(f) = \mathbf{h}^H(f)\mathbf{v}(f)$ is the residual noise.

The two terms on the right-hand side of (2.5) are uncorrelated. Hence, the variance of $\widehat{X}(f)$ is also the sum of two variances:

$$\begin{aligned}\phi_{\widehat{X}}(f) &= \mathbf{h}^H(f)\mathbf{\Phi}_{\mathbf{y}}(f)\mathbf{h}(f) \\ &= \phi_{X_{\text{fd}}}(f) + \phi_{V_{\text{rn}}}(f),\end{aligned} \tag{2.6}$$

where $\phi_{X_{\text{fd}}}(f) = \phi_{X_1}(f)\left|\mathbf{h}^H(f)\mathbf{d}(f)\right|^2$ is the variance of the filtered desired signal and $\phi_{V_{\text{rn}}}(f) = \mathbf{h}^H(f)\mathbf{\Phi}_{\mathbf{v}}(f)\mathbf{h}(f)$ is the variance of the residual noise. From (2.6), we deduce that the narrowband output SNR is

$$\text{oSNR}\left[\mathbf{h}(f)\right] = \frac{\phi_{X_1}(f)\left|\mathbf{h}^H(f)\mathbf{d}(f)\right|^2}{\mathbf{h}^H(f)\mathbf{\Phi}_{\mathbf{v}}(f)\mathbf{h}(f)}, \tag{2.7}$$

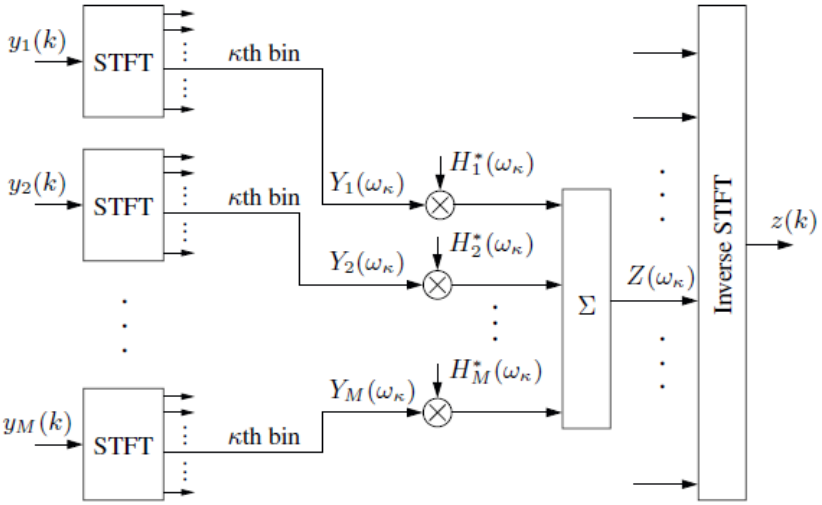


Figure 2.2: Illustration of linear beamforming in the time-frequency domain [BCC15].

which is upper bounded by [BCH08]

$$\begin{aligned} \text{oSNR}\left[\mathbf{h}(f)\right] &\leq \phi_{X_1}(f)\mathbf{d}^H(f)\mathbf{\Phi}_{\mathbf{v}}^{-1}(f)\mathbf{d}(f) \\ &= \text{oSNR}_{\max}. \end{aligned} \tag{2.8}$$

Additionally, the narrowband output SNR gain is defined as

$$\mathcal{G}\left[\mathbf{h}(f)\right] = \frac{\text{oSNR}\left[\mathbf{h}(f)\right]}{\text{iSNR}(f)}. \tag{2.9}$$

A well-known and widely used example of such a filter is obtained upon minimizing the variance of the filter output or the variance of the residual noise subject to the distortionless constraint, i.e., $\mathbf{h}^H(f)\mathbf{d}(f) = 1$. This optimization results in Capon's MVDR filter [Cap69], [Lac71]:

$$\begin{aligned} \mathbf{h}_{\text{MVDR}}(f) &= \frac{\mathbf{\Phi}_{\mathbf{y}}^{-1}(f)\mathbf{d}(f)}{\mathbf{d}^H(f)\mathbf{\Phi}_{\mathbf{y}}^{-1}(f)\mathbf{d}(f)} \\ &= \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}(f)\mathbf{d}(f)}{\mathbf{d}^H(f)\mathbf{\Phi}_{\mathbf{v}}^{-1}(f)\mathbf{d}(f)}, \end{aligned} \tag{2.10}$$

which can be rewritten as [BCH08]

$$\mathbf{h}_{\mathrm{MVDR}}(f) = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}(f)\mathbf{\Phi}_{\mathbf{y}}(f) - \mathbf{I}_M}{\mathrm{tr}\left[\mathbf{\Phi}_{\mathbf{v}}^{-1}(f)\mathbf{\Phi}_{\mathbf{y}}(f)\right] - M}\mathbf{i}, \tag{2.11}$$

where $\mathrm{tr}\left[\cdot\right]$ is the trace of a square matrix, $\mathbf{I}_M$ is the identity matrix of size $M \times M$, and $\mathbf{i}$ is the first column of $\mathbf{I}_M$. As a result, the estimate of $X_1(f)$ with the MVDR filter is

$$\begin{aligned} \widehat{X}_{\mathrm{MVDR}}(f) &= \mathbf{h}_{\mathrm{MVDR}}^H(f)\mathbf{y}(f) \\ &= X_1(f) + \mathbf{h}_{\mathrm{MVDR}}^H(f)\mathbf{v}(f). \end{aligned} \tag{2.12}$$

Another commonly used filter is the LCMV, which attempts, much like the MVDR, to minimize the variance of the residual noise. However, with the LCMV, this minimization is subject to a set of $2 \leq L \leq M$ linear constraints. The LCMV filter is usually effective in cases when some further information on the environment is known, thus allowing to a priori attenuate the output signal in noisy directions [GJ82]:

$$\mathbf{h}_{\mathrm{LCMV}}(f) = \mathbf{\Phi}_{\mathbf{y}}^{-1}(f)\mathbf{C}(f)[\mathbf{C}^H(f)\mathbf{\Phi}_{\mathbf{y}}^{-1}(f)\mathbf{C}(f)]^{-1}\boldsymbol{\beta}, \tag{2.13}$$

where $\mathbf{C}$ is an $M \times L$ matrix whose columns are the steering vectors in the directions of constraints, and $\boldsymbol{\beta}$ is an $L \times 1$ vector of the desired filter responses in these directions. Thus, the estimate of $X_1(f)$ with the LCMV filter is

$$\begin{aligned} \widehat{X}_{\mathrm{LCMV}}(f) &= \mathbf{h}_{\mathrm{LCMV}}^H(f)\mathbf{y}(f) \\ &= X_1(f) + \mathbf{h}_{\mathrm{LCMV}}^H(f)\mathbf{v}(f). \end{aligned} \tag{2.14}$$

2.2 Single-channel linear filtering

In this section, we address the SCNR problem in the STFT domain. We consider the classical single-channel noise reduction problem, where the noisy microphone signal at

the time index t is given by [Loi13, BCHC09]

$$y(t) = x(t) + v(t), \tag{2.15}$$

with $x(t)$ and $v(t)$ denoting the desired speech signal and additive noise, respectively. We assume that $x(t)$ and $v(t)$ are uncorrelated, and that all signals are real, zero mean, and broadband. Using the STFT, (2.15) can be rewritten in the time-frequency domain as [BCH12]

$$Y(k,n) = X(k,n) + V(k,n), \tag{2.16}$$

where the zero-mean complex random variables $Y(k,n)$, $X(k,n)$, and $V(k,n)$ are the STFTs of $y(t)$, $x(t)$, and $v(t)$, respectively, at the frequency bin $k \in \{0,1,\ldots,K-1\}$ and the time frame n. It is well known that the same signal at different time frames is correlated [Coh05a]. Therefore, the interframe correlation should be taken into account in order to improve the performance of noise reduction algorithms. In this case, we may consider forming an observation signal vector of length N, containing N most recent samples of $Y(k,n)$, i.e.,

$$\begin{aligned}\mathbf{y}(k,n) &= \begin{bmatrix} Y(k,n) & Y(k,n-1) & \cdots & Y(k,n-N+1) \end{bmatrix}^T \\ &= \mathbf{x}(k,n) + \mathbf{v}(k,n),\end{aligned} \tag{2.17}$$

where $\mathbf{x}(k,n)$ and $\mathbf{v}(k,n)$ are defined similarly to $\mathbf{y}(k,n)$. Since $x(t)$ and $v(t)$ are uncorrelated by assumption, the $N \times N$ correlation matrix of $\mathbf{y}(k,n)$ is

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}}(k,n) &= E\left[\mathbf{y}(k,n)\mathbf{y}^H(k,n)\right] \\ &= \mathbf{\Phi}_{\mathbf{x}}(k,n) + \mathbf{\Phi}_{\mathbf{v}}(k,n),\end{aligned} \tag{2.18}$$

where $\mathbf{\Phi}_{\mathbf{x}}(k,n)$ and $\mathbf{\Phi}_{\mathbf{v}}(k,n)$ are the correlation matrices of $\mathbf{x}(k,n)$ and $\mathbf{v}(k,n)$, respectively.

With the interframe correlation taken into account, single-channel noise reduction in the STFT domain is performed by applying a complex-valued filter, $\mathbf{h}(k,n)$, of length

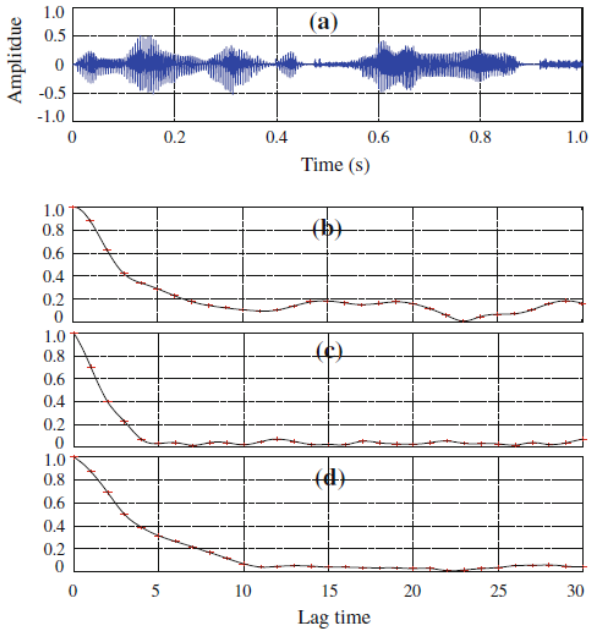


Figure 2.3: Illustration of the interframe correlation property. (a) A 1-second segment of speech signal, (b)-(d) magnitude of the interframe cross-correlation coefficients for different frequency bins [BCH12].

N, to the observation signal vector, $\mathbf{y}(k,n)$, i.e., [BCH12]

$$\begin{aligned}\widehat{X}(k,n) &= \mathbf{h}^H(k,n)\mathbf{y}(k,n) \\ &= X_{\text{fd}}(k,n) + V_{\text{rn}}(k,n),\end{aligned} \tag{2.19}$$

where the filter output, $\widehat{X}(k,n)$, is an estimate of $X(k,n)$, $X_{\text{fd}}(k,n) = \mathbf{h}^H(k,n)\mathbf{x}(k,n)$ a filtered version of the desired speech signal, and $V_{\text{rn}}(k,n) = \mathbf{h}^H(k,n)\mathbf{v}(k,n)$ is the residual noise. We note that the vector $\mathbf{x}(k,n)$ contains components that are uncorrelated with the desired signal $X(k,n)$. That is, every element $X(k,n-l), l \neq 0$ may be decomposed into two orthogonal components, one correlated with $X(k,n)$ and another orthogonal component that is considered as an interference

$$X(k,n-l) = \rho_X^*(k,n,l)X(k,n) + X_{\text{i}}(k,n-l), \tag{2.20}$$

where the interference component is defined by

$$X_{\text{i}}(k,n-l) = X(k,n-l) - \rho_X^*(k,n,l)X(k,n), \tag{2.21}$$

$$E\left[X(k,n)X_{\mathrm{i}}^{*}(k,n-l)\right]=0, \tag{2.22}$$

and

$$\rho_X(k,n,l)=\frac{E\left[X(k,n)X^{*}(k,n-l)\right]}{E\left[|X(k,n)|^2\right]}, \tag{2.23}$$

is the interframe correlation coefficient of $X(k,n)$ at time distance l. Stacking the elements of $\rho_X(k,n,l)$ in a vector form we have

$$\begin{aligned}\rho_X(k,n)&=\left[1,\rho_X(k,n,1),\ldots,\rho_X(k,n,N-1)\right]^T \\ &=\frac{E\left[X(k,n)\mathbf{x}^{*}(k,n)\right]}{E\left[|X(k,n)|^2\right]},\end{aligned} \tag{2.24}$$

which is referred to as the normalized interframe correlation vector between $X(k,n)$ and $\mathbf{x}(k,n)$. Now, it is possible to reformulate (2.19) as

$$\begin{aligned}\widehat{X}(k,n)&=\mathbf{h}^{H}(k,n)\left[\mathbf{x}_{\mathrm{d}}(k,n)+\mathbf{x}_{\mathrm{i}}(k,n)+\mathbf{v}(k,n)\right] \\ &=X_{\mathrm{f}}(k,n)+X_{\mathrm{ri}}(k,n)+V_{\mathrm{rn}}(k,n),\end{aligned} \tag{2.25}$$

where

$$\mathbf{x}_{\mathrm{d}}(k,n)=X(k,n)\rho_X^{*}(k,n), \tag{2.26}$$

and

$$\mathbf{x}_{\mathrm{i}}(k,n)=\left[0,X_{\mathrm{i}}(k,n-1),\ldots,X_{\mathrm{i}}(k,n-N+1)\right]^T, \tag{2.27}$$

are the desired and interference signal vectors, respectively. The three terms on the right-hand side of (2.25) are mutually uncorrelated. Hence, the variance of $\widehat{X}(k,n)$ is

$$\begin{aligned}\phi_{\widehat{X}}(k,n)&=\mathbf{h}^{H}(k,n)\mathbf{\Phi}_{\mathbf{y}}(k,n)\mathbf{h}(k,n) \\ &=\phi_{X_{\mathrm{f}}}(k,n)+\phi_{X_{\mathrm{ri}}}(k,n)+\phi_{V_{\mathrm{rn}}}(k,n),\end{aligned} \tag{2.28}$$

where $\phi_{X_{\rm f}}(k,n) = \mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{x}_{\rm d}}(k,n)\mathbf{h}(k,n)$, $\phi_{X_{\rm ri}}(k,n) = \mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{x}_{\rm i}}(k,n)\mathbf{h}(k,n)$ and $\phi_{V_{\rm rn}}(k,n) = \mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{v}}(k,n)\mathbf{h}(k,n)$ are the variances of the filtered speech signal, the residual interference and the residual noise, respectively.

In the conventional linear approach [BCH12], noise reduction is performed by applying a complex-valued filter, $\mathbf{h}(k,n)$ of length N, to the observation signal vector, $\mathbf{y}(k,n)$, i.e.,

$$\begin{aligned}\widehat{X}(k,n) &= \mathbf{h}^H(k,n)\mathbf{y}(k,n) \\ &= X_{\rm fd}(k,n) + V_{\rm rn}(k,n),\end{aligned} \tag{2.29}$$

where the filter output, $\widehat{X}(k,n)$, is an estimate of $X(k,n)$, $X_{\rm fd}(k,n) = \mathbf{h}^H(k,n)\mathbf{x}(k,n)$ is the filtered desired signal, and $V_{\rm rn}(k,n) = \mathbf{h}^H(k,n)\mathbf{v}(k,n)$ is the residual noise.

The two terms on the right-hand side of (2.29) are uncorrelated. Hence, the variance of $\widehat{X}(k,n)$ is

$$\begin{aligned}\phi_{\widehat{X}}(k,n) &= \mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{y}}(k,n)\mathbf{h}(k,n) \\ &= \phi_{X_{\rm fd}}(k,n) + \phi_{V_{\rm rn}}(k,n),\end{aligned} \tag{2.30}$$

where $\phi_{X_{\rm fd}}(k,n) = \mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{x}}(k,n)\mathbf{h}(k,n)$ is the variance of the filtered desired signal and $\phi_{V_{\rm rn}}(k,n) = \mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{v}}(k,n)\mathbf{h}(k,n)$ is the variance of the residual noise. Then, from (2.30), the subband output SNR is given by

$$\text{oSNR}\left[\mathbf{h}(k,n)\right] = \frac{\mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{x}}(k,n)\mathbf{h}(k,n)}{\mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{v}}(k,n)\mathbf{h}(k,n)}. \tag{2.31}$$

2.3 Design of multistage differential beamformers

In this section, we provide background to the multistage differential beamforming approach. We consider a ULA, which is composed of $M \geq 2$ omnidirectional microphones with an interelement spacing equal to δ. Let us assume that a farfield plane wave propagates from a potential azimuth angle, θ, in an anechoic acoustic environment at the speed of sound, i.e., $c = 340$ m/s, and impinges on this ULA. In this scenario, the

corresponding steering vector (of length M) is [JD92]

$$\mathbf{d}_{\theta,M}(f) = \Big[1 \quad e^{-\jmath 2\pi f \delta \cos\theta/c} \quad \cdots \quad e^{-\jmath(M-1)2\pi f \delta \cos\theta/c}\Big]^T, \tag{2.32}$$

where f is the temporal frequency, $\jmath = \sqrt{-1}$ is the imaginary unit, and the superscript ^{T} is the transpose operator.

In order to be in the optimal working conditions of differential beamforming, we assume that the desired source comes from the direction $\theta_{\mathrm{s}} = 0$ and δ is small [BC13]. In this case, we can express the frequency-domain observed signal vector of length M as [BCC18]

$$\begin{aligned}\mathbf{y}(f) &= \begin{bmatrix} Y_1(f) & Y_2(f) & \cdots & Y_M(f) \end{bmatrix}^T \\ &= \mathbf{x}(f) + \mathbf{v}(f) \\ &= \mathbf{d}_{0,M}(f) X(f) + \mathbf{v}(f), \end{aligned} \tag{2.33}$$

where $Y_m(f)$ is the mth microphone signal, $\mathbf{x}(f) = \mathbf{d}_{0,M}(f) X(f)$, $\mathbf{d}_{0,M}(f)$ is the steering vector at $\theta = 0$, $X(f)$ is the zero-mean desired source signal, $\mathbf{v}(f)$ is the zero-mean additive noise signal vector defined similarly to $\mathbf{y}(f)$, and $X(f)$ and $\mathbf{v}(f)$ are incoherent. In the rest, in order to simplify the notation, we drop the dependence on the temporal frequency, f. For a small and compact array, it is reasonable to assume that the variance of the noise is the same at all sensors, i.e., $\phi_V = \phi_{V_1} = \phi_{V_2} = \cdots = \phi_{V_M}$, with $\phi_{V_m} = E\left(|V_m|^2\right)$, $m = 1, 2, \ldots, M$ and $E(\cdot)$ denoting mathematical expectation. Therefore, the variance of Y_m is $\phi_{Y_m} = \phi_Y = \phi_X + \phi_V$, where ϕ_X is the variance of X.

Let P be a positive integer with $0 \leq P < M$. We can transform the observed signal vector $\mathbf{y}$ of length M to a Pth-order forward spatial difference of $\mathbf{y}$ of length $M(P) = M - P$, i.e., [HBCC20b]

$$\mathbf{y}_{(P)} = \mathbf{\Delta}_{(P)} \mathbf{y}, \tag{2.34}$$

with $\mathbf{y}_{(0)} = \mathbf{y}$, where

$$\boldsymbol{\Delta}_{(P)} = \begin{bmatrix} \mathbf{c}_{(P)}^T & 0 & \cdots & 0 \\ 0 & \mathbf{c}_{(P)}^T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{c}_{(P)}^T \end{bmatrix} \tag{2.35}$$

is a matrix of size $M(P) \times M$, with $\boldsymbol{\Delta}_{(0)} = \mathbf{I}_M$, which is the $M \times M$ identity matrix,

$$\mathbf{c}_{(P)} = \begin{bmatrix} (-1)^P \begin{pmatrix} P \\ 0 \end{pmatrix} & (-1)^{P-1} \begin{pmatrix} P \\ 1 \end{pmatrix} \\ & \cdots \quad -\begin{pmatrix} P \\ P-1 \end{pmatrix} \quad 1 \end{bmatrix}^T \tag{2.36}$$

is a vector of length $P + 1$, and

$$\begin{pmatrix} P \\ j \end{pmatrix} = \frac{P!}{j!\,(P-j)!}$$

is the binomial coefficient. The major benefit with the difference observed signal vector, $\mathbf{y}_{(P)}$ of length $M(P)$, for $P > 0$, is that it is much less sensitive to diffuse noise as compared to $\mathbf{y}$; in fact, the larger is the value of P, the higher is the signal-to-noise ratio (SNR) of $\mathbf{y}_{(P)}$. However, $\mathbf{y}_{(P)}$ is more sensitive to white noise.

It can be shown that (2.34) can be expressed as [HBCC20b]

$$\begin{aligned} \mathbf{y}_{(P)} &= \tau_0^P \mathbf{d}_{0,M(P)} X + \mathbf{v}_{(P)} \\ &= \mathbf{x}_{(P)} + \mathbf{v}_{(P)}, \end{aligned} \tag{2.37}$$

where

$$\tau_0 = e^{-\jmath 2\pi f \delta / c} - 1 \tag{2.38}$$

is a frequency-dependent variable, $\mathbf{d}_{0,M(P)}$ is the steering vector of length $M(P)$ at

$\theta = 0$, and $\mathbf{v}_{(P)} = \mathbf{\Delta}_{(P)}\mathbf{v}$. We deduce that the $M(P) \times M(P)$ covariance matrix of $\mathbf{y}_{(P)}$ is

$$\begin{aligned}
\mathbf{\Phi}_{\mathbf{y}_{(P)}} &= \phi_X \left|\tau_0\right|^{2P} \mathbf{d}_{0,M(P)}\mathbf{d}_{0,M(P)}^H + \mathbf{\Delta}_{(P)}\mathbf{\Phi}_{\mathbf{v}}\mathbf{\Delta}_{(P)}^T \\
&= \phi_X \left|\tau_0\right|^{2P} \mathbf{d}_{0,M(P)}\mathbf{d}_{0,M(P)}^H + \mathbf{\Phi}_{\mathbf{v}_{(P)}} \\
&= \phi_X \left|\tau_0\right|^{2P} \mathbf{d}_{0,M(P)}\mathbf{d}_{0,M(P)}^H + \phi_V \mathbf{\Delta}_{(P)}\mathbf{\Gamma}_{\mathbf{v}}\mathbf{\Delta}_{(P)}^T \\
&= \phi_X \left|\tau_0\right|^{2P} \mathbf{d}_{0,M(P)}\mathbf{d}_{0,M(P)}^H + \phi_V \mathbf{\Gamma}_{\mathbf{v}_{(P)}}, \qquad (2.39)
\end{aligned}$$

where $\mathbf{\Phi}_{\mathbf{v}}$ is the covariance matrix of $\mathbf{v}$, $\mathbf{\Gamma}_{\mathbf{v}} = \mathbf{\Phi}_{\mathbf{v}}/\phi_V$, and $\mathbf{\Gamma}_{\mathbf{v}_{(P)}} = \mathbf{\Delta}_{(P)}\mathbf{\Gamma}_{\mathbf{v}}\mathbf{\Delta}_{(P)}^T$.

Then, in this context, linear beamforming is performed by applying a beamformer $\mathbf{h}$ of length $M(P)$ to $\mathbf{y}_{(P)}$, i.e.,

$$\begin{aligned}
Z &= \mathbf{h}^H \mathbf{y}_{(P)} \\
&= \mathbf{h}^H \mathbf{x}_{(P)} + \mathbf{h}^H \mathbf{v}_{(P)} \\
&= X_{\text{fd}} + V_{\text{rn}}, \qquad (2.40)
\end{aligned}$$

where Z is the estimate of the desired signal, X,

$$X_{\text{fd}} = \tau_0^P \mathbf{h}^H X \qquad (2.41)$$

is the filtered desired signal, and

$$V_{\text{rn}} = \mathbf{h}^H \mathbf{v}_{(P)} \qquad (2.42)$$

is the residual noise. We deduce that the variance of Z is

$$\begin{aligned}
\phi_Z = \left|\tau_0\right|^{2P} &\left|\mathbf{h}^H \mathbf{d}_{0,M(P)}\right|^2 \phi_X \\
&+ \phi_V \mathbf{h}^H \mathbf{\Gamma}_{\mathbf{v}_{(P)}} \mathbf{h}. \qquad (2.43)
\end{aligned}$$

We see from X_{fd} that the distortionless constraint is

$$\mathbf{h}^H \mathbf{d}_{0,M(P)} = \tau_0^{-P}. \qquad (2.44)$$

2.4 Design of Kronecker-product beamformers

In this section, we present in brief the KP beamforming approach. Let us now invoke the MCNR signal model of (2.2). Assume that $M = M_{\mathbf{a}} \times M_{\mathbf{b}}$, where $M_{\mathbf{a}}, M_{\mathbf{b}} \geq 1$. Then, one can verify that the steering vector, which is denoted here by $\mathbf{d}_\theta$:

$$\mathbf{d}_\theta = \Big[1 \quad e^{-\jmath 2\pi f \delta \cos\theta/c} \quad \dots \quad e^{-\jmath(M-1)2\pi f \delta \cos\theta/c} \Big]^T, \tag{2.45}$$

can be decomposed as [BCC19]:

$$\mathbf{d}_\theta = \mathbf{a}_\theta \otimes \mathbf{b}_\theta, \tag{2.46}$$

with θ being the desired signal incident angle, and

$$\mathbf{a}_\theta = \Big[1 \quad e^{-\jmath 2\pi f M_{\mathbf{b}} \delta \cos\theta/c} \quad \dots \quad e^{-\jmath(M_{\mathbf{a}}-1)2\pi f M_{\mathbf{b}} \delta \cos\theta/c} \Big]^T \tag{2.47}$$

is the steering vector (of length $M_{\mathbf{a}}$) corresponding to a ULA of $M_{\mathbf{a}}$ sensors with an interelement spacing equal to $M_{\mathbf{b}}\delta$, $\otimes$ is the Kronecker product, and

$$\mathbf{b}_\theta = \Big[1 \quad e^{-\jmath 2\pi f \delta \cos\theta/c} \quad \dots \quad e^{-\jmath(M_{\mathbf{b}}-1)2\pi f \delta \cos\theta/c} \Big]^T \tag{2.48}$$

is the steering vector (of length $M_{\mathbf{b}}$) corresponding to a ULA of $M_{\mathbf{b}}$ sensors with an interelement spacing equal to δ. As a consequence, the signal model in (2.2) becomes:

$$\mathbf{y} = (\mathbf{a}_\theta \otimes \mathbf{b}_\theta)\, X + \mathbf{v}, \tag{2.49}$$

and its covariance matrix is:

$$\mathbf{\Phi}_{\mathbf{y}} = \phi_X \left(\mathbf{a}_\theta \mathbf{a}_\theta^H\right) \otimes \left(\mathbf{b}_\theta \mathbf{b}_\theta^H\right) + \phi_V \mathbf{\Gamma}_{\mathbf{v}}. \tag{2.50}$$

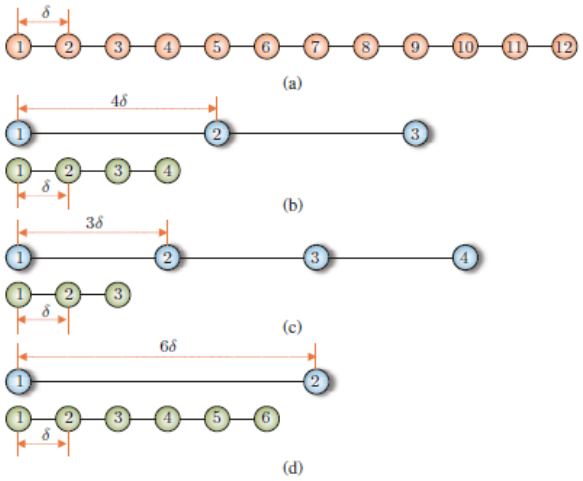


Figure 2.4: Examples of the KP decomposition of a global beamformer with $M = 12$ microphones into two sub-beamformers with varying number of microphones [BCC19].

Because of the particular structure of the steering vector in (2.46) and in order to fully exploit this structure, we propose (global) beamformers of the form:

$$\mathbf{h} = \mathbf{h}_{\mathbf{a}} \otimes \mathbf{h}_{\mathbf{b}}, \tag{2.51}$$

where $\mathbf{h}_{\mathbf{a}}$ and $\mathbf{h}_{\mathbf{b}}$ are two complex-valued linear filters of lengths $M_{\mathbf{a}}$ and $M_{\mathbf{b}}$, respectively. Then, in the proposed context, linear beamforming is performed by applying $\mathbf{h}$ to $\mathbf{y}$, i.e.,

$$\begin{aligned} Z &= \mathbf{h}^H \mathbf{y} \\ &= \mathbf{h}^H \mathbf{x} + \mathbf{h}^H \mathbf{v} \\ &= X_{\text{fd}} + V_{\text{rn}}, \end{aligned} \tag{2.52}$$

where Z is the estimate of the desired signal, X,

$$\begin{aligned} X_{\text{fd}} &= \left(\mathbf{h}_{\mathbf{a}} \otimes \mathbf{h}_{\mathbf{b}}\right)^H \left(\mathbf{a}_\theta \otimes \mathbf{b}_\theta\right) X \\ &= \left(\mathbf{h}_{\mathbf{a}}^H \mathbf{a}_\theta\right) \left(\mathbf{h}_{\mathbf{b}}^H \mathbf{b}_\theta\right) X \end{aligned} \tag{2.53}$$

is the filtered desired signal, and

$$V_{\mathrm{rn}} = \left(\mathbf{h_a} \otimes \mathbf{h_b}\right)^H \mathbf{v} \tag{2.54}$$

is the residual noise. We deduce that the variance of Z is

$$\begin{aligned}\phi_Z = &\left|\mathbf{h}_{\mathbf{a}}^H \mathbf{a}_\theta\right|^2 \left|\mathbf{h}_{\mathbf{b}}^H \mathbf{b}_\theta\right|^2 \phi_X \\ &+ \phi_V \left(\mathbf{h_a} \otimes \mathbf{h_b}\right)^H \mathbf{\Gamma_v} \left(\mathbf{h_a} \otimes \mathbf{h_b}\right).\end{aligned} \tag{2.55}$$

We see from X_{fd} that the distortionless constraint is

$$\left(\mathbf{h}_{\mathbf{a}}^H \mathbf{a}_\theta\right)\left(\mathbf{h}_{\mathbf{b}}^H \mathbf{b}_\theta\right) = 1. \tag{2.56}$$

Furthermore, we will often use the following relationships:

$$\mathbf{h_a} \otimes \mathbf{h_b} = \left(\mathbf{h_a} \otimes \mathbf{I}_{M_{\mathbf{b}}}\right) \mathbf{h_b} \tag{2.57}$$

$$= \left(\mathbf{I}_{M_{\mathbf{a}}} \otimes \mathbf{h_b}\right) \mathbf{h_a}, \tag{2.58}$$

where $\mathbf{I}_{M_{\mathbf{b}}}$ and $\mathbf{I}_{M_{\mathbf{a}}}$ are the identity matrices of sizes $M_{\mathbf{b}} \times M_{\mathbf{b}}$ and $M_{\mathbf{a}} \times M_{\mathbf{a}}$, respectively.

Chapter 3

Nonlinear Kronecker-product Filtering for Multichannel Noise Reduction

Speech Communication 114 (2019) 49–59

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Nonlinear Kronecker product filtering for multichannel noise reduction

Gal Itzhak[a,*], Jacob Benesty[b], Israel Cohen[a]

[a] Andrew and Erna Viterby Faculty of Electrical Engineering, Technion – Israel Institute of Technology, Technion City, Haifa 3200003, Israel
[b] INRS-EMT, University of Quebec, 800 de la Gauchetiere Ouest, Montreal QC H5A 1K6, Canada

ARTICLE INFO

Keywords:
Noise reduction
Speech enhancement
Microphone arrays
Multichannel
Frequency-domain filtering
Optimal filters
Nonlinear processing

ABSTRACT

Multichannel noise reduction in the frequency domain is a fundamental problem in the areas of speech processing and speech recognition. In this paper, we address this problem and propose an alternative approach to retrieve a speech signal out of microphone array noisy observations. We focus on the spectral amplitude of the speech signal and assume that the spectral phase is less significant. The estimate of the spectral amplitude squared, that is the spectral power, is obtained by applying a complex linear filter to a modified version of the observations vector. This modified version is obtained as a Kronecker product of the complex conjugate of the observations vector and the original observations vector. The complex speech signal estimate is obtained by multiplying the spectral amplitude estimate with a complex exponential whose phase may be extracted from the minimum variance distortionless response beamformer. We present a modified optimization criterion according to which the proposed filters may be derived, and compare their performances to conventional multichannel noise reduction filters. We show that the new approach is preferable, in particular when the input signal-to-noise ratio (SNR) is low or the number of sensors is small.

1. Introduction

Many modern applications in a wide variety of areas, from speech recognition and communications to speaker identification and human-to-machine systems, are required to operate in noisy environments. Noise fields, in many cases, significantly deteriorate the speech signal quality, thus damaging the functionality of communication and speech recognition systems. The problem of enhancing speech, or reducing noise, has attracted many researchers over the years, who suggested numerous schemes and algorithms in multiple processing domains.

With the growing demand for robust noise reduction capabilities, multichannel noise reduction (MCNR) methods are often employed in order to exploit spatial information. This additional information allows, in many cases, to attain a considerable amount of noise reduction while preserving the desired signal distortionless (Benesty et al., 2009a). Often referred to as beamformers, MCNR methods may be designed and implemented in various domains.

Time-domain beamformers are the easiest to implement, as the filters are applied directly to the noisy observations, typically generating a single speech sample estimate at a time. It is also possible to estimate a vector of successive speech samples simultaneously. However, such beamformers tend to suffer from high computational complexity (Benesty and Chen, 2011; Benesty et al., 2017; Buchris et al., 2019).

Transform-domain beamformers, as in Chen et al. (2003) and Benesty et al. (2008, 2007, 2009b), are typically formulated on a frame basis. That is, the noisy signal is transformed into another domain, the optimal filter is derived and applied in the transformed domain, and subsequently the filtered observations are transformed back to the time domain using an appropriate inverse transform. Time-domain beamformers may be derived in a transform domain by appropriately adjusting the criterion for optimization (Benesty et al., 2012). Choosing an appropriate transform may be beneficial in terms of the quality of the enhanced speech and the computational complexity of the noise reduction filter application. The generalized Karhunen-Loève (KL) domain and the frequency domain, for instance, constitute common choices for these particular reasons.

The generalized KL domain is obtained by projecting noisy observations onto an orthonormal basis of eigenvectors of a speech signal correlation matrix. This projection results in uncorrelated analysis coefficients which are independently processed (Benesty et al., 2009b; Ephraim and Trees, 1995; Lacouture-Parodi et al., 2014). Moreover, it was shown (Ephraim and Trees, 1995) that by taking a signal subspace approach, KL-domain processing may separate a speech signal-plus-noise subspace from noise-only subspace. With this approach, the latter is employed to estimate the noise-only statistics and then applied on the former to form an estimate of the clean speech. Assuming the

* Corresponding author.
E-mail address: galitz@campus.technion.ac.il (G. Itzhak).

https://doi.org/10.1016/j.specom.2019.10.001
Received 21 February 2019; Received in revised form 21 August 2019; Accepted 2 October 2019
Available online 3 October 2019

estimate of the clean speech correlation matrix is accurate, it is guaranteed that no aliasing problems emerge (Benesty et al., 2009b). Nonetheless, frequency-domain beamformers (Dmochowski and Benesty, 2010; Gannot and Cohen, 2008; Tavakoli et al., 2016), which are typically implemented in the short-time Fourier transform (STFT) domain, are considered easier to employ. That is, unlike the KL domain speech signal-dependent eigenvectors, the Fourier basis functions are global. Consequently, frequency-domain beamforming does not require the overhead of estimating and diagonalizing the speech correlation matrix, yet the frequency coefficients remain uncorrelated provided that the analysis window is long enough.

Originally proposed in Capon (1969), Capon's minimum variance distortionless response (MVDR) beamformer has been investigated in theoretical studies from a variety of aspects (Souden et al., 2010). The linear MVDR, which operates directly on a vector of transformed noisy observations, is shown to be optimal in terms of the residual noise energy, under the restriction of zero desired signal distortion. Moreover, it has inspired numerous variations, e.g., the minimum power distortionless response (MPDR) (Van Trees, 2004), which avoids the estimation of the noise-only correlation matrix. This sort of flexibility, combined with its proved noise-reduction capabilities and easy-to-analyze linear nature, has made the MVDR beamformer very common in real-world applications.

Another important variation, which may also be seen as a generalization of the MVDR, is the linearly constrained minimum variance (LCMV) beamformer (Griffiths and Jim, 1982). While maintaining the desired signal distortionless in the same spatial manner as the MVDR does, the LCMV provides a convenient scheme to cope with spatial interferences by placing nulls in their respective directions. Additionally, it is optimal in the sense of the residual noise energy minimization. However, its noise reduction performance is known to be inferior in comparison to the MVDR, unless the number of sensors is significantly greater than the number of interfered directions. This limitation results from the linear nature of the two beamformers.

For classical speech analysis purposes, higher-order statistics were shown to be informative, though typically in the context of single-channel noise reduction (SCNR). In Moreno and Fonollosa (1992), the third-order statistics of noisy speech was used to determine its pitch, suggesting that unlike speech, most common noises exhibit a nearly-zero skewness. In Nemer et al. (2001), a voice activity detector (VAD) was presented assuming an underlying zero-phase harmonic representation of speech. Closed-form expressions of the third and fourth-order cumulants were derived and combined with second-order measures to yield what was demonstrated to be a more robust VAD. In Nemer et al. (2002), it was suggested to take advantage of fourth-order cumulants to estimate some widely-used parameters, such as the SNR, the speech autocorrelation and the probability of speech presence. These estimates may then feed any speech enhancement method in which they are required as input parameters.

In this paper, we present a Kronecker product (KP) approach for MCNR in the frequency domain. We propose to take advantage of higher-order statistics and apply a complex linear filter to a modified observation signal vector. The modified vector is constructed from the original noisy observations and its elements may be interpreted as the instantaneous correlation coefficients. The filtering product of the KP approach is an estimate of the desired signal spectral power, which is considered more important than the spectral phase in many applications, such as speech enhancement. The spectral phase may be extracted from a conventional beamformer, e.g., the linear MVDR. We propose a modified optimization criterion for deriving KP filters and present the KP-MVDR and the KP-LCMV. We demonstrate that when the array size is small or the SNR is low the KP filters outperform the conventional ones, provided that temporal smoothing is employed to properly estimate the correlation matrix of the modified observation signal vector.

The rest of the paper is organized as follows. In Section 2, we formulate the MCNR problem in the frequency domain. In Section 3, we review the conventional filtering approach and derive the MVDR and LCMV beamformers, which are used as benchmarks for comparison. In Section 4, we formulate the KP filtering approach and derive the KP filters based on a new optimization criterion. In Section 5, we compare the conventional and KP filters analytically through a stationary toy example. Finally, in Section 6, we evaluate the performances of the two approaches by a set of nonstationary speech signals simulations in anechoic and reverberant environments.

2. Signal model and problem formulation

Consider an array consisting of M omnidirectional microphones. The received signals at the frequency index f are expressed as (Benesty et al., 2008; 2012; Bai et al., 2014)

$$\begin{aligned} Y_m(f) &= G_m(f)S(f) + V_m(f) \\ &= X_m(f) + V_m(f), \ m = 1, 2, \dots, M, \end{aligned} \tag{1}$$

where $Y_m(f)$ is the mth microphone signal, $S(f)$ is the unknown speech source, $G_m(f)$ is the acoustic room transfer function from the position of $S(f)$ to the mth microphone, $X_m(f) = G_m(f)S(f)$ is the zero-mean speech signal which takes into account the acoustic room transfer function, and $V_m(f)$ is the zero-mean additive noise. It is assumed that $X_i(f)$ and $V_j(f)$ are uncorrelated, i.e., $E\left[X_i(f)V_j^*(f)\right] = 0, \ \forall i, j = 1, 2, \dots, M$, where $E[\cdot]$ denotes mathematical expectation and the superscript * is the complex-conjugate operator. By definition, the terms $X_m(f), \ m = 1, 2, \dots, M$ are correlated while the other terms $V_m(f), \ m = 1, 2, \dots, M$, depending on the nature of the noise, may only be partially correlated. We consider the first microphone as the reference; then, the objective of multichannel noise reduction in the frequency domain is to estimate the desired signal, $X_1(f)$, from the M observations $Y_m(f), \ m = 1, 2, \dots, M$, in the best possible way.

It is more convenient to write the M frequency-domain microphone signals in a vector notation:

$$\begin{aligned} \mathbf{y}(f) &= \mathbf{g}(f)S(f) + \mathbf{v}(f) \\ &= \mathbf{x}(f) + \mathbf{v}(f) \\ &= \mathbf{d}(f)X_1(f) + \mathbf{v}(f), \end{aligned} \tag{2}$$

where

$$\begin{aligned} \mathbf{y}(f) &= \left[Y_1(f) \quad Y_2(f) \quad \cdots \quad Y_M(f)\right]^T, \\ \mathbf{x}(f) &= \left[X_1(f) \quad X_2(f) \quad \cdots \quad X_M(f)\right]^T \\ &= S(f)\mathbf{g}(f), \\ \mathbf{g}(f) &= \left[G_1(f) \quad G_2(f) \quad \cdots \quad G_M(f)\right]^T, \\ \mathbf{v}(f) &= \left[V_1(f) \quad V_2(f) \quad \cdots \quad V_M(f)\right]^T, \end{aligned}$$

the superscript ^{T} is the transpose operator, and

$$\mathbf{d}(f) = \left[1 \quad \frac{G_2(f)}{G_1(f)} \quad \cdots \quad \frac{G_M(f)}{G_1(f)}\right]^T = \frac{\mathbf{g}(f)}{G_1(f)}. \tag{3}$$

Expression (2) depends explicitly on the desired signal, $X_1(f)$; therefore, (2) is the frequency-domain signal model for noise reduction. The vector $\mathbf{d}(f)$ can be seen as the frequency-domain steering vector (Dmochowski and Benesty, 2010). This general formulation implies that we are interested in recovering the noise-free signal and not necessarily the clean speech signal.

Since $\mathbf{y}(f)$ is the sum of two uncorrelated components, its correlation matrix is

$$\begin{aligned} \mathbf{\Phi}_{\mathbf{y}}(f) &= E\left[\mathbf{y}(f)\mathbf{y}^H(f)\right] \\ &= \phi_{X_1}(f)\mathbf{d}(f)\mathbf{d}^H(f) + \mathbf{\Phi}_{\mathbf{v}}(f), \end{aligned} \tag{4}$$

where the superscript ^{H} is the conjugate-transpose operator, $\phi_{X_1}(f) = E\left[|X_1(f)|^2\right]$ is the variance of $X_1(f)$ which may also be interpreted as

the power spectral density (PSD) of the time-domain representation of $X_1(f)$ (Welch, 1967), and $\mathbf{\Phi}_{\mathbf{v}}(f) = E\left[\mathbf{v}(f)\mathbf{v}^H(f)\right]$ is the correlation matrix of $\mathbf{v}(f)$. The narrowband input SNR is given by

$$\text{iSNR}(f) = \frac{\phi_{X_1}(f)}{\phi_{V_1}(f)}, \tag{5}$$

where $\phi_{V_1}(f) = E\left[|V_1(f)|^2\right]$ is the variance of $V_1(f)$.

3. Conventional filtering approach

In the conventional filtering approach, multichannel noise reduction in the frequency domain is performed by applying a complex-valued linear filter, $\mathbf{h}(f)$, of length M, to the observation signal vector, $\mathbf{y}(f)$ (Dmochowski and Benesty, 2010; Benesty et al., 2008), i.e.,

$$\begin{aligned}\widehat{X}(f) &= \mathbf{h}^H(f)\mathbf{y}(f)\\ &= X_{\text{fd}}(f) + V_{\text{rn}}(f),\end{aligned} \tag{6}$$

where the filter output, $\widehat{X}(f)$, is an estimate of $X_1(f)$, $X_{\text{fd}}(f) = X_1(f)\mathbf{h}^H(f)\mathbf{d}(f)$ is the filtered desired signal, and $V_{\text{rn}}(f) = \mathbf{h}^H(f)\mathbf{v}(f)$ is the residual noise.

The two terms on the right-hand side of (6) are uncorrelated. Hence, the variance of $\widehat{X}(f)$ is also the sum of two variances:

$$\begin{aligned}\phi_{\widehat{X}}(f) &= \mathbf{h}^H(f)\mathbf{\Phi}_{\mathbf{y}}(f)\mathbf{h}(f)\\ &= \phi_{X_{\text{fd}}}(f) + \phi_{V_{\text{rn}}}(f),\end{aligned} \tag{7}$$

where $\phi_{X_{\text{fd}}}(f) = \phi_{X_1}(f)\left|\mathbf{h}^H(f)\mathbf{d}(f)\right|^2$ is the variance of the filtered desired signal and $\phi_{V_{\text{rn}}}(f) = \mathbf{h}^H(f)\mathbf{\Phi}_{\mathbf{v}}(f)\mathbf{h}(f)$ is the variance of the residual noise. From (7), we deduce that the narrowband output SNR is

$$\text{oSNR}[\mathbf{h}(f)] = \frac{\phi_{X_1}(f)\left|\mathbf{h}^H(f)\mathbf{d}(f)\right|^2}{\mathbf{h}^H(f)\mathbf{\Phi}_{\mathbf{v}}(f)\mathbf{h}(f)}, \tag{8}$$

which is upper bounded by Benesty et al. (2008)

$$\begin{aligned}\text{oSNR}[\mathbf{h}(f)] &\le \phi_{X_1}(f)\mathbf{d}^H(f)\mathbf{\Phi}_{\mathbf{v}}^{-1}(f)\mathbf{d}(f)\\ &= \text{oSNR}_{\text{max}}.\end{aligned} \tag{9}$$

Additionally, the narrowband output SNR gain is defined as

$$\mathcal{G}[\mathbf{h}(f)] = \frac{\text{oSNR}[\mathbf{h}(f)]}{\text{iSNR}(f)}. \tag{10}$$

A well-known and widely used example of such a filter is obtained upon minimizing the variance of the filter output or the variance of the residual noise subject to the distortionless constraint, i.e., $\mathbf{h}^H(f)\mathbf{d}(f) = 1$. This optimization results in Capon's MVDR filter (Capon, 1969), (Lacoss, 1971):

$$\begin{aligned}\mathbf{h}_{\text{MVDR}}(f) &= \frac{\mathbf{\Phi}_{\mathbf{y}}^{-1}(f)\mathbf{d}(f)}{\mathbf{d}^H(f)\mathbf{\Phi}_{\mathbf{y}}^{-1}(f)\mathbf{d}(f)}\\ &= \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}(f)\mathbf{d}(f)}{\mathbf{d}^H(f)\mathbf{\Phi}_{\mathbf{v}}^{-1}(f)\mathbf{d}(f)},\end{aligned} \tag{11}$$

which can be rewritten as (Benesty et al., 2008)

$$\mathbf{h}_{\text{MVDR}}(f) = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}(f)\mathbf{\Phi}_{\mathbf{y}}(f) - \mathbf{I}_M}{\text{tr}\left[\mathbf{\Phi}_{\mathbf{v}}^{-1}(f)\mathbf{\Phi}_{\mathbf{y}}(f)\right] - M}\mathbf{i}, \tag{12}$$

where $\text{tr}[\cdot]$ is the trace of a square matrix, $\mathbf{I}_M$ is the identity matrix of size $M\times M$, and $\mathbf{i}$ is the first column of $\mathbf{I}_M$. As a result, the estimate of $X_1(f)$ with the MVDR filter is

$$\begin{aligned}\widehat{X}_{\text{MVDR}}(f) &= \mathbf{h}_{\text{MVDR}}^H(f)\mathbf{y}(f)\\ &= X_1(f) + \mathbf{h}_{\text{MVDR}}^H(f)\mathbf{v}(f).\end{aligned} \tag{13}$$

Another commonly used filter is the LCMV, which attempts, much like the MVDR, to minimize the variance of the residual noise. However, with the LCMV, this minimization is subject to a set of $2 \le L \le M$ linear constraints. The LCMV filter is usually effective in cases when some further information on the environment is known, thus allowing to a priori attenuate the output signal in noisy directions (Griffiths and Jim, 1982):

$$\mathbf{h}_{\text{LCMV}}(f) = \mathbf{\Phi}_{\mathbf{y}}^{-1}(f)\mathbf{C}(f)[\mathbf{C}^H(f)\mathbf{\Phi}_{\mathbf{y}}^{-1}(f)\mathbf{C}(f)]^{-1}\boldsymbol{\beta}, \tag{14}$$

where $\mathbf{C}$ is an $M\times L$ matrix whose columns are the steering vectors in the directions of constraints, and $\boldsymbol{\beta}$ is an $L\times 1$ vector of the desired filter responses in these directions. Thus, the estimate of $X_1(f)$ with the LCMV filter is

$$\begin{aligned}\widehat{X}_{\text{LCMV}}(f) &= \mathbf{h}_{\text{LCMV}}^H(f)\mathbf{y}(f)\\ &= X_1(f) + \mathbf{h}_{\text{LCMV}}^H(f)\mathbf{v}(f).\end{aligned} \tag{15}$$

4. Kronecker product filtering approach

The idea behind the new approach is to estimate the spectral power of the desired signal, i.e., $|X_1(f)|^2$, rather than the complex signal, $X_1(f)$, as it was suggested in Ephraim and Malah (1984). We can express the spectral power of $\widehat{X}(f)$ defined in (6) as

$$\begin{aligned}\left|\widehat{X}(f)\right|^2 &= \mathbf{h}^H(f)\mathbf{y}(f)\mathbf{y}^H(f)\mathbf{h}(f)\\ &= \text{tr}\left[\mathbf{h}(f)\mathbf{h}^H(f)\mathbf{y}(f)\mathbf{y}^H(f)\right]\\ &= \text{vec}^H\left[\mathbf{h}(f)\mathbf{h}^H(f)\right]\text{vec}\left[\mathbf{y}(f)\mathbf{y}^H(f)\right]\\ &= \left[\mathbf{h}^*(f)\otimes\mathbf{h}(f)\right]^H\left[\mathbf{y}^*(f)\otimes\mathbf{y}(f)\right]\\ &= \left[\mathbf{h}^*(f)\otimes\mathbf{h}(f)\right]^H\widetilde{\mathbf{y}}(f),\end{aligned} \tag{16}$$

where vec[.] is the vectorization operation, $\otimes$ is the Kronecker product, and $\widetilde{\mathbf{y}}(f) = \mathbf{y}^*(f)\otimes\mathbf{y}(f)$.

Now, let $\widetilde{\mathbf{h}}(f)$ be a complex-valued filter of length M^2 which is not necessarily of the form $\widetilde{\mathbf{h}}(f) = \mathbf{h}^*(f)\otimes\mathbf{h}(f)$. Eq. (16) suggests that we can estimate $|X_1(f)|^2$ by applying $\widetilde{\mathbf{h}}(f)$ to $\widetilde{\mathbf{y}}(f) = \mathbf{y}^*(f)\otimes\mathbf{y}(f)$, i.e.,

$$Z(f) = \widetilde{\mathbf{h}}^H(f)\widetilde{\mathbf{y}}(f). \tag{17}$$

We note that by not restricting $\widetilde{\mathbf{h}}(f)$ to have the Kronecker product structure of the last line of (16), we generate extra degrees of freedom which may potentially yield improved noise reduction capabilities with respect to $\mathbf{h}(f)$. When $\widetilde{\mathbf{h}}(f)$ is derived, we can estimate the desired signal, $X_1(f)$, with

$$\widehat{X}(f) = e^{\eta\psi(f)}\sqrt{|Z(f)|}, \tag{18}$$

where $\psi(f)$ is the desired signal estimated phase that can be obtained in any given way. Practically, $\psi(f)$ may be the phase of $\widehat{X}_{\text{MVDR}}(f)$ or $\widehat{X}_{\text{LCMV}}(f)$, for example. Clearly, this approach is highly nonlinear.

It should be pointed out that the concept of extending the dimension of filtering beyond the observations signal dimension was, for example, suggested in Benesty et al. (2010) in the context of single-channel noise reduction with a gain. However, the differences between the widely linear filter of Benesty et al. (2010) and the work we present here are significant. The widely linear approach is essentially linear, while the approach taken here is clearly nonlinear. Furthermore, as it can be observed from the definition of $\widetilde{\mathbf{y}}(f)$, in our approach a squared-dimensional filter is applied, not to the observations vector directly, but rather to their instantaneous correlation terms. As we will show, this implies that we exploit higher-order statistics.

The expression for $\widetilde{\mathbf{y}}(f)$ can be further developed

$$\begin{aligned}\widetilde{\mathbf{y}}(f) &= \mathbf{y}^*(f)\otimes\mathbf{y}(f)\\ &= \left[\mathbf{x}^*(f) + \mathbf{v}^*(f)\right]\otimes\left[\mathbf{x}(f) + \mathbf{v}(f)\right]\\ &= \left|X_1(f)\right|^2\widetilde{\mathbf{d}}(f) + \mathbf{x}^*(f)\otimes\mathbf{v}(f)\\ &\quad + \mathbf{v}^*(f)\otimes\mathbf{x}(f) + \widetilde{\mathbf{v}}(f),\end{aligned} \tag{19}$$

where $\widetilde{\mathbf{d}}(f) = \mathbf{d}^*(f) \otimes \mathbf{d}(f)$ and $\widetilde{\mathbf{v}}(f) = \mathbf{v}^*(f) \otimes \mathbf{v}(f)$. Exploiting (19) to analyze the variance of the estimated desired signal $\widehat{X}(f)$, we have

$$\begin{aligned}\phi_{\widehat{X}}(f) &= E[|Z(f)|] \\ &\approx |E[Z(f)]| \\ &= \left|\widetilde{\mathbf{h}}^H(f)E\left[\widetilde{\mathbf{y}}(f)\right]\right| \\ &= \left|\phi_{X_1}(f)\widetilde{\mathbf{h}}^H(f)\widetilde{\mathbf{d}}(f) + \widetilde{\mathbf{h}}^H(f)E\left[\widetilde{\mathbf{v}}(f)\right]\right| \\ &= \left|\widetilde{\mathbf{h}}^H(f)\mathrm{vec}\left[\mathbf{\Phi}_{\mathbf{x}}(f)\right] + \widetilde{\mathbf{h}}^H(f)\mathrm{vec}\left[\mathbf{\Phi}_{\mathbf{v}}(f)\right]\right|.\end{aligned} \tag{20}$$

Note that according to the approximation in the second row of (20), $Z(f)$ is assumed to be real and positive.

We may define the output SNR and the output SNR gain for the KP filtering by analogy to the expressions in (8) and (10), respectively, by

$$\mathrm{oSNR}\left[\widetilde{\mathbf{h}}(f)\right] = \frac{\left|\widetilde{\mathbf{h}}^H(f)\mathrm{vec}\left[\mathbf{\Phi}_{\mathbf{x}}(f)\right]\right|}{\left|\widetilde{\mathbf{h}}^H(f)\mathrm{vec}\left[\mathbf{\Phi}_{\mathbf{v}}(f)\right]\right|}, \tag{21}$$

$$\mathcal{G}\left[\widetilde{\mathbf{h}}(f)\right] = \frac{\mathrm{oSNR}\left[\widetilde{\mathbf{h}}(f)\right]}{\mathrm{iSNR}(f)}. \tag{22}$$

To begin with, we note that when $\widetilde{\mathbf{h}}(f) = \mathbf{h}^*(f) \otimes \mathbf{h}(f)$ Eq. (21) immediately reduces to (8) and the conventional filtering approach is obtained as a special case of the KP filtering approach. In this case the approximation in (20) is always correct, and therefore the output SNR and the output SNR gain expressions are mathematically accurate. Alternatively, when $\widetilde{\mathbf{h}}(f)$ does not follow this structure, there is no guarantee that $Z(f)$ is real and positive, and hence in such cases Eq. (21) is merely an approximation. Nonetheless, and as it will be further explained in Section 6, we make sure that $Z(f)$ is always real and positive, implying that (21) is, in fact, a reasonable output SNR approximation.

Consider the following criterion:

$$\begin{aligned}\mathcal{J}\left[\widetilde{\mathbf{h}}(f)\right] &= E\left[|Z(f)|^2\right] \\ &= \widetilde{\mathbf{h}}^H(f)\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(f)\widetilde{\mathbf{h}}(f),\end{aligned} \tag{23}$$

where $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(f) = E\left[\widetilde{\mathbf{y}}(f)\widetilde{\mathbf{y}}^H(f)\right]$. We would like to minimize $\mathcal{J}\left[\widetilde{\mathbf{h}}(f)\right]$ subject to the distortionless constraint, i.e., $\widetilde{\mathbf{h}}^H(f)\widetilde{\mathbf{d}}(f) = 1$. The optimal filter is given by

$$\widetilde{\mathbf{h}}_{\mathrm{MVDR}}(f) = \frac{\mathbf{\Phi}_{\widetilde{\mathbf{y}}}^{-1}(f)\widetilde{\mathbf{d}}(f)}{\widetilde{\mathbf{d}}^H(f)\mathbf{\Phi}_{\widetilde{\mathbf{y}}}^{-1}(f)\widetilde{\mathbf{d}}(f)}, \tag{24}$$

which we refer to as the KP-MVDR filter. Note that $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(f)$ is the fourth moment matrix of the noisy observations vector $\mathbf{y}(f)$ and is of size $M^2 \times M^2$. Unlike former studies, in which higher-order statistics were primarily used to obtain estimates of input parameters of speech enhancement method (Nemer et al., 2001; 2002), in this study the fourth-order statistics is directly utilized to derive the speech enhancement filter. As for the computational complexity, calculating the inverse of $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(f)$, for example, by using the classic Gauss-Jordan method, would require a time complexity of $O(M^6)$ operations in comparison to the $O(M^3)$ required by the conventional $\mathbf{\Phi}_{\mathbf{y}}(f)$. Nonetheless, when M is not very big, the additional complexity is insignificant.

We can generalize this approach and minimize $\mathcal{J}\left[\widetilde{\mathbf{h}}(f)\right]$ subject to a set of linear constraints as is done with the conventional LCMV:

$$\widetilde{\mathbf{h}}^H(f)\widetilde{\mathbf{C}}(f) = \boldsymbol{\beta}^H, \tag{25}$$

where $\widetilde{\mathbf{C}}(f)$ is an $M^2 \times L$ matrix whose columns are the KP filtering steering vectors $\widetilde{\mathbf{d}}(f)$ in the directions of constraints, and $\boldsymbol{\beta}$ is the same as in (14). Then, the derivation of the KP-LCMV is straightforward

$$\widetilde{\mathbf{h}}_{\mathrm{LCMV}}(f) = \mathbf{\Phi}_{\widetilde{\mathbf{y}}}^{-1}(f)\widetilde{\mathbf{C}}(f)[\widetilde{\mathbf{C}}^H(f)\mathbf{\Phi}_{\widetilde{\mathbf{y}}}^{-1}(f)\widetilde{\mathbf{C}}(f)]^{-1}\boldsymbol{\beta}. \tag{26}$$

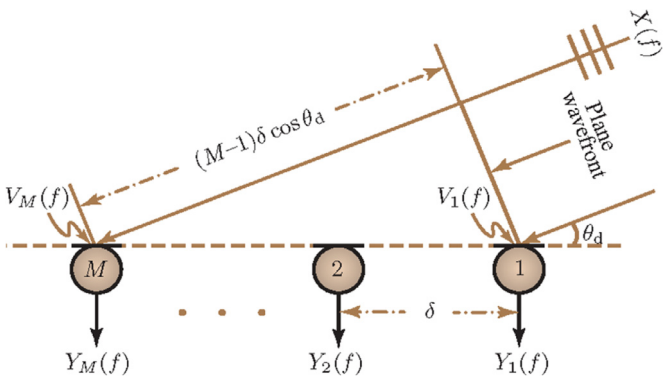


Fig. 1. A typical uniform linear array with M sensors.

5. Analysis of a toy example

In a non reverberant environment, consider a uniform linear array (ULA) of M sensors with an interelement spacing δ (see Fig. 1) satisfying $T_s = \delta/c$, where T_s is the sampling interval and c the speed of sound in the air. Assume that a white Gaussian signal of interest, $x(t) \sim \mathcal{N}(0, f_x)$, is impinging on the array from the broadside direction, i.e., $\theta_d = 90°$, and corrupted by thermal white Gaussian noise, $v_m(t) \sim \mathcal{N}(0, f_v)$, $m = \{1, \ldots, M\}$. For simplicity, we consider the case of $M = 2$ sensors, and assume that the signals in the array are sampled $N = Tf_s$ times within the signal duration T at $f_s = 1/T_s = 16$ kHz. The correlation matrices of $\mathbf{x}(f)$ and $\mathbf{v}(f)$ are given, respectively, by

$$\mathbf{\Phi}_{\mathbf{x}}(f) = K\epsilon_x\mathbf{d}(f, \cos\theta_d)\mathbf{d}^H(f, \cos\theta_d), \tag{27}$$

$$\mathbf{\Phi}_{\mathbf{v}}(f) = K\epsilon_v\mathbf{I}_M, \tag{28}$$

where $\mathbf{d}(f, \cos\theta_d) = [1\ 1]^T$ is the desired signal steering vector and K is an appropriate scaling constant resulting from the frequency domain transform. The narrowband input SNR is given by

$$\begin{aligned}\mathrm{iSNR}(f) &= \frac{\phi_{X_1}(f)}{\phi_{V_1}(f)} \\ &= \frac{\epsilon_x}{\epsilon_v},\end{aligned} \tag{29}$$

where $\phi_{X_1}(f)$ and $\phi_{V_1}(f)$ are the variances of the desired signal and noise at the first microphone.

It is clear that the optimal conventional MVDR filter is given by $\mathbf{h}_{\mathrm{MVDR}}(f) = [0.5\ 0.5]^T$, which results in an output SNR gain of approximately 3dB. This gain is independent of the input SNR. For the purpose of deriving the KP-MVDR filter, we first have to evaluate its corresponding correlation matrix $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(f)$. The latter is a sum of 16 $M^2 \times M^2$ matrices and therefore might be difficult to calculate in general. However, in our toy example, by recalling complex-normal distribution properties, it may be shown (as derived in the Appendix) that

$$\begin{aligned}\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(f) \propto\ & 2\,\mathrm{iSNR}\begin{bmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{bmatrix} \\ &+ \frac{1}{\mathrm{iSNR}}\begin{bmatrix}2&0&0&1\\0&1&0&0\\0&0&1&0\\1&0&0&2\end{bmatrix} + 2\begin{bmatrix}2&1&1&1\\1&1&0&1\\1&1&0&1\\1&1&1&2\end{bmatrix},\end{aligned} \tag{30}$$

which is indeed input SNR-dependent. We note that when the input SNR is high, $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(f)$ is nearly singular. This implies that regularization should be used, and that the optimal KP filter is approximately $\widetilde{\mathbf{h}}_{\mathrm{MVDR}}(f) =$

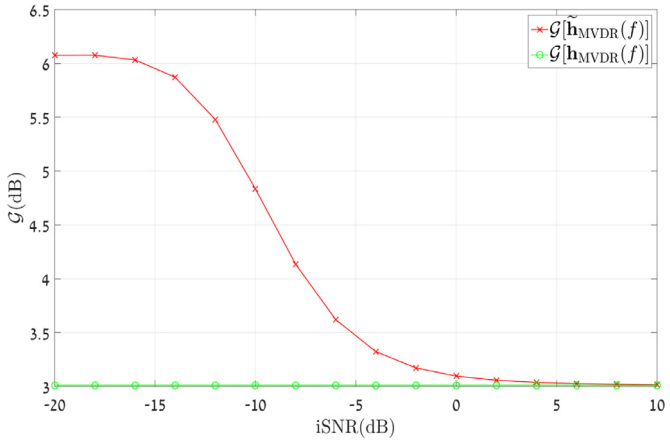


Fig. 2. Output SNR gain curve of a desired white Gaussian signal corrupted by thermal white Gaussian noise as a function of the input SNR. The array consists of $M = 2$ sensors.

$0.25\widetilde{\mathbf{d}}(f, \cos\theta_{\mathrm{d}}) = 0.25[1\ 1\ 1\ 1]^T$. Recalling that in this case $\mathrm{vec}\left[\mathbf{\Phi}_{\mathbf{v}}(f)\right] = K\epsilon_{\mathrm{v}}[1\ 0\ 0\ 1]^T$, we have

$$\begin{aligned}\mathcal{G}\left[\widetilde{\mathbf{h}}(f)\right] &= \frac{\mathrm{oSNR}\left[\widetilde{\mathbf{h}}(f)\right]}{\mathrm{iSNR}(f)}\\ &= \frac{\left|\widetilde{\mathbf{h}}^H(f)\mathrm{vec}\left[\mathbf{\Phi}_{\mathbf{x}}(f)\right]\right|}{\left|\widetilde{\mathbf{h}}^H(f)\mathrm{vec}\left[\mathbf{\Phi}_{\mathbf{v}}(f)\right]\right|}\frac{\phi_{V_1}(f)}{\phi_{X_1}(f)}\\ &= \frac{1}{0.25[1\ 1\ 1\ 1][1\ 0\ 0\ 1]^T}\\ &\approx 3\mathrm{dB}, \end{aligned} \tag{31}$$

which is identical to $\mathcal{G}\left[\mathbf{h}_{\mathrm{MVDR}}(f)\right]$. However, when the input SNR is very low, the second component on the right hand side of (30) is dominant, and we have

$$\mathbf{\Phi}_{\tilde{\mathbf{y}}}^{-1}(f) \propto \begin{bmatrix} 2/3 & 0 & 0 & -1/3\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ -1/3 & 0 & 0 & 2/3 \end{bmatrix},$$

which implies that $\widetilde{\mathbf{h}}_{\mathrm{MVDR}}(f) = 1/8[1\ 3\ 3\ 1]^T$. Thus

$$\mathcal{G}\left[\widetilde{\mathbf{h}}_{\mathrm{MVDR}}(f)\right] = \frac{1}{1/8[1\ 3\ 3\ 1][1\ 0\ 0\ 1]^T} \approx 6\mathrm{dB}. \tag{32}$$

Fig. 2 depicts the output SNR gain curves of the conventional and KP MVDRs in the foregoing scenario as a function of the input SNR. Indeed, in high input SNRs both approaches perform equally well. However, as the input SNR decreases, the KP-MVDR yields higher SNR gain than the conventional MVDR.

It may also be informative to discuss the time complexity differences between the two approaches in the context of this example. Let us begin with the conventional MVDR. The most expensive calculation it requires in terms of computational costs is indeed the correlation matrix inversion, which requires roughly $O(M^3)$ multiplications. We assume for simplicity that the cost of M^3 multiplications is exact. Then, the calculation of the numerator, i.e., the term $\mathbf{\Phi}_{\mathbf{y}}^{-1}(f)\mathbf{d}(f)$, requires M^2 multiplications, whereas in order to compute the denominator another M multiplications must be performed. Calculating the filter takes another (real) division operation, and applying the filter to the noisy observations vector costs additional M multiplications. Thus, the total cost of generating a desired signal estimate out of the observations is roughly $M^3 + M^2 + 2M = 16$ complex multiplications, with $M = 2$, plus another real division operation. We move on to the KP-MVDR. It requires the same operations as the conventional MVDR, but rather with squared-size correlation matrix and steering vector. In addition, generating the modified observations vector $\widetilde{\mathbf{y}}(f)$ requires another M^2 multiplications, whereas the desired signal power estimate requires an additional (real) square-root operation. Thus, the total computational cost of the KP-MVDR is about $M^6 + M^4 + 3M^2 = 92$ complex multiplications, a single real division and a single real square-root operation. In practice, running the toy example code with MATLAB software on an ordinary CPU takes 250 μs to complete with the conventional MVDR and 520 μs with the KP-MVDR. Increasing the array size to $M = 7$ yields a total runtime of 350 μs with the conventional MVDR and 1000 μs with the KP-MVDR.

6. Simulations

6.1. Speech signals simulations in an anechoic environment

We are interested in examining the KP approach in more practical scenarios, i.e., with nonstationary desired signals and in the presence of spatial interferences. Therefore, the noise reduction procedure is modified as follows. The observed signals are transformed into the STFT domain using 50% overlapping time frames and a Hamming analysis window of length 512 (32 ms). We derive each of the aforementioned filters in the STFT domain. That is, the two conventional filters: $\mathbf{h}_{\mathrm{MVDR}}(f, t)$ and $\mathbf{h}_{\mathrm{LCMV}}(f, t)$; and the two KP filters: $\widetilde{\mathbf{h}}_{\mathrm{MVDR}}(f,t)$ and $\widetilde{\mathbf{h}}_{\mathrm{LCMV}}(f,t)$. The filters are applied in the STFT domain to generate estimates of the desired signal. Finally, the inverse STFT is applied to yield time-domain speech signals.

We evaluate the performances of the different filters by comparing the output SNR gains. In the STFT domain, the input and output SNR expressions in (5), (8), and (21) are modified to

$$\overline{\mathrm{iSNR}}(f) = \frac{\sum_t \phi_{X_1}(f,t)}{\sum_t \phi_{V_1}(f,t)}, \tag{33}$$

$$\overline{\mathrm{oSNR}}[\mathbf{h}(f)] = \frac{\sum_t \phi_{X_1}(f,t)\left|\mathbf{h}^H(f,t)\mathbf{d}(f)\right|^2}{\sum_t \mathbf{h}^H(f,t)\mathbf{\Phi}_{\mathbf{v}}(f,t)\mathbf{h}(f,t)}, \tag{34}$$

and

$$\overline{\mathrm{oSNR}}\left[\widetilde{\mathbf{h}}(f)\right] = \frac{\sum_t \left|\widetilde{\mathbf{h}}^H(f,t)\ \mathrm{vec}\left[\mathbf{\Phi}_{\mathbf{x}}(f,t)\right]\right|}{\sum_t \left|\widetilde{\mathbf{h}}^H(f,t)\ \mathrm{vec}\left[\mathbf{\Phi}_{\mathbf{v}}(f,t)\right]\right|}, \tag{35}$$

where $\phi_{X_1}(f,t)$ and $\phi_{V_1}(f,t)$ are the STFT-domain variances of the desired signal and noise at the first microphone, and $\mathbf{\Phi}_{\mathbf{x}}(f, t)$ and $\mathbf{\Phi}_{\mathbf{v}}(f, t)$ are the STFT-domain correlation matrices of the desired signal and noise. The average output SNR gains are given by

$$\overline{\mathcal{G}}[\mathbf{h}(f)] = \frac{\overline{\mathrm{oSNR}}[\mathbf{h}(f)]}{\overline{\mathrm{iSNR}}(f)} \tag{36}$$

and

$$\overline{\mathcal{G}}\left[\widetilde{\mathbf{h}}(f)\right] = \frac{\overline{\mathrm{oSNR}}\left[\widetilde{\mathbf{h}}(f)\right]}{\overline{\mathrm{iSNR}}(f)}, \tag{37}$$

respectively. We employ the average output SNR gain as our main performance measure.

There is another modification that should be made with the KP approach in order to obtain a reliable desired signal estimation and keep the expressions in (35) and (37) valid. We recall that for each time-frequency bin the STFT modification of (17) provides a local estimate of the desired signal spectral power:

$$Z(f,t) = \widetilde{\mathbf{h}}^H(f,t)\widetilde{\mathbf{y}}(f,t). \tag{38}$$

While it is easy to show that with both $\widetilde{\mathbf{h}}_{\mathrm{MVDR}}(f)$ and $\widetilde{\mathbf{h}}_{\mathrm{LCMV}}(f)$ this expression is real, there is no guarantee that it is strictly positive. In practice, when a desired speech signal is present, it is very likely that the inner product in (38) is indeed positive, hence yielding a valid estimate of the desired signal spectral power. This may be seen by applying

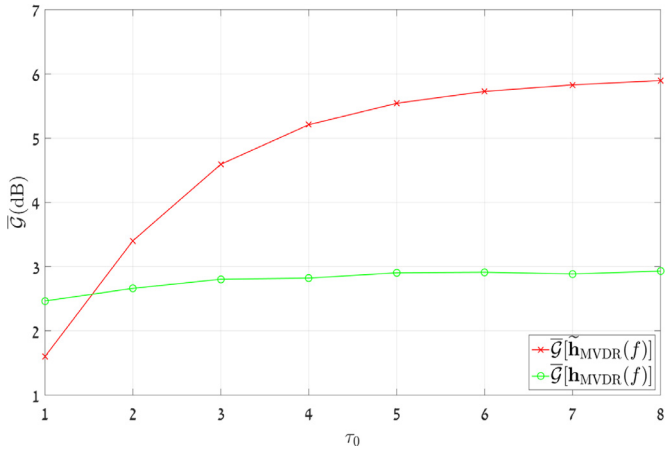


Fig. 3. Output SNR gain curve of a desired white Gaussian signal corrupted by thermal white Gaussian noise as a function of the temporal smoothing parameter τ_0. The array consists of $M = 2$ sensors and the input SNR is -20dB.

one of the KP filters to the last equality of (19) in which the first term, that is associated with the true desired signal power, is guaranteed to be positive. Nevertheless, when a desired signal is absent, this positive term is approximately zero and the power estimate might turn out to be negative. Clearly, such an estimate is non-physical and should be clipped to zero. Consequently, (38) is modified to

$$Z(f,t) = \max\left\{\widetilde{\mathbf{h}}^H(f,t)\widetilde{\mathbf{y}}(f,t), 0\right\}, \tag{39}$$

whereas when a zero estimate is obtained both the filtered desired signal and the residual noise are zeroed out, and are referred to accordingly in (35) and (37).

The correlation matrices in the STFT domain are obtained as a straightforward temporal smoothing of the appropriate instantaneous signals, i.e.,

$$\mathbf{\Phi}_{\mathbf{y}}(f,t) = \frac{1}{2\tau_0+1}\sum_{\tau=t-\tau_0}^{t+\tau_0}\mathbf{y}(f,\tau)\mathbf{y}^H(f,\tau) \tag{40}$$

and

$$\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(f,t) = \frac{1}{2\tau_0+1}\sum_{\tau=t-\tau_0}^{t+\tau_0}\widetilde{\mathbf{y}}(f,\tau)\widetilde{\mathbf{y}}^H(f,\tau), \tag{41}$$

where τ_0 is a smoothing parameter indicating the temporal duration of the smoothing process. As $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(f,t)$ is considerably larger than $\mathbf{\Phi}_{\mathbf{y}}(f,t)$, the KP filter is more vulnerable to correlation estimation errors. That is, when τ_0 is not large enough the off-diagonal elements of the correlation matrices are inaccurately estimated, thus adding a significant amount of noise which corrupts the optimal filters. This behaviour is demonstrated, for example, for the stationary toy example described above in Fig. 3, in which the average output SNR gain of the KP and conventional MVDRs are plotted as a function of the smoothing parameter τ_0. Indeed, we observe that the KP-MVDR requires a considerably longer temporal smoothing in order to achieve its theoretical output SNR gain. We set $\tau_0 = 5$ for all further simulations we discuss next.

We examine the average output SNR gains of the four filters in similar settings to the previous scenario, but we employ speech signals which are taken from the TIMIT database (Darpa timit acoustic phonetic continuous speech corpus cdrom, 1993) as the desired signal, and add an interference impinging on the sensor array from the endfire direction (0°) that is three times more powerful than the background thermal white Gaussian noise. This implies that the overall noise consists of both directional and thermal noise terms. Naturally, all four filters satisfy the distortionless constraint, while the conventional and KP LCMVs also place a null at 0° ($L = 2$).

The average output SNR gains for the six combinations of $M \in \{3, 5, 7\}$ with input SNR $\in \{0, 7\}$dB for the two KP filters and their two conventional counterparts are depicted in Fig. 4. We note that the input SNR takes into account both the background noise and the interference. To begin with, we observe that the KP-LCMV achieves a significantly preferable average output SNR gain in comparison to the conventional LCMV, in particular in low frequencies. Similarly, the KP-MVDR maintains a substantial output SNR gain gap over the conventional MVDR throughout the entire frequency spectrum. However, as M and the input SNR increase the output SNR gain gap decreases- specifically in frequencies lower than 3 kHz, which are typically of high interest in speech signals. The significant output SNR gain gap in frequencies higher than 3 kHz may be explained by the lower bound in (39), which eliminates much of the residual noise with the KP filters when the desired speech signal is absent.

Next, we employ the perceptual evaluation of speech quality (PESQ) score (Rix et al., 2001) as an additional objective performance measure, which is applied on the enhanced speech in the time domain. Table 1 summarises the average PESQ score of the four filters in the aforementioned six settings combinations in addition to the PESQ score of the noisy signal in the reference microphone $Y_1(f)$. We point out that the PESQ score and the average output SNR gain in low frequencies exhibit some sort of correlation. That is, according to both measures the KP approach is shown to produce cleaner enhanced signals, particularly for low input SNRs or small arrays. For high input SNRs and large arrays, the KP approach may still be preferable, but the PESQ score and average output SNR gain differences are significantly reduced.

We end this section by comparing the spectrograms of the desired, noisy and four enhanced signals with $M = 5$ and iSNR = 0dB, which are depicted in Fig. 5. We observe that the enhanced signals with the KP approach exhibit a higher resemblance to the desired signal, with the background noise strongly attenuated. This is stressed out in particular in very low frequencies, in which the conventional LCMV, for example, amplifies the background noise while the KP-LCMV attenuates it.

6.2. Speech signals simulations in reverberant environments

In this section, we address the noise reduction performance in reverberant environments. We use a room impulse response (RIR) generator (Habets, 2014) to simulate the reverberant noise-free signal received in each of the microphones. The RIR generator is based on the image method of Allen and Berkley (1979). We point out that for the sake of verification, some of the following scenarios were repeated using the randomized image method presented in Sena et al. (2015). To begin with, we are interested in examining the desired speech signal reverberation influence on the noise reduction performance for different RIRs. Hence, the simulation settings is as follows. A $6\times6\times3$ m room contains a desired speech signal source located at $(x, y, z) = (3, 1, 1.5)$, and $M = 3$ microphones located, respectively, at (2.95, 5, 1.5), (3, 5, 1.5) and (3.05, 5, 1.5). The microphone signals contain thermal white Gaussian noise. We simulate 3 scenarios with a varying value of $T_{60} \in \{0, 250, 400\}$ msec (as defined by Sabin-Franklin's formula Pierce (1991)), and use the conventional and KP-MVDR filters to perform noise reduction. We note that the filters are derived according to the non-reverberant model with $\theta_{\mathrm{d}} = 90°$.

The simulations are carried out for both iSNR = 0 and iSNR = 7 and the same set of TIMIT speech signals used in the previous part. We compare the average PESQ scores of the time-domain enhanced speech signals to the clean and noisy reverberant signals. The results are shown in Table 2. We observe that the KP-MVDR obtains higher PESQ scores in all the foregoing scenarios, however, the performance gap is more significant for the lower values of the input SNR and T_{60}. That is, when T_{60} is high, the speech quality is mainly deteriorated by the reverberation and not by the white background noise. Hence, in such cases the PESQ score improvement due to the noise reduction is limited, and an additional dereverberation stage should be incorporated.

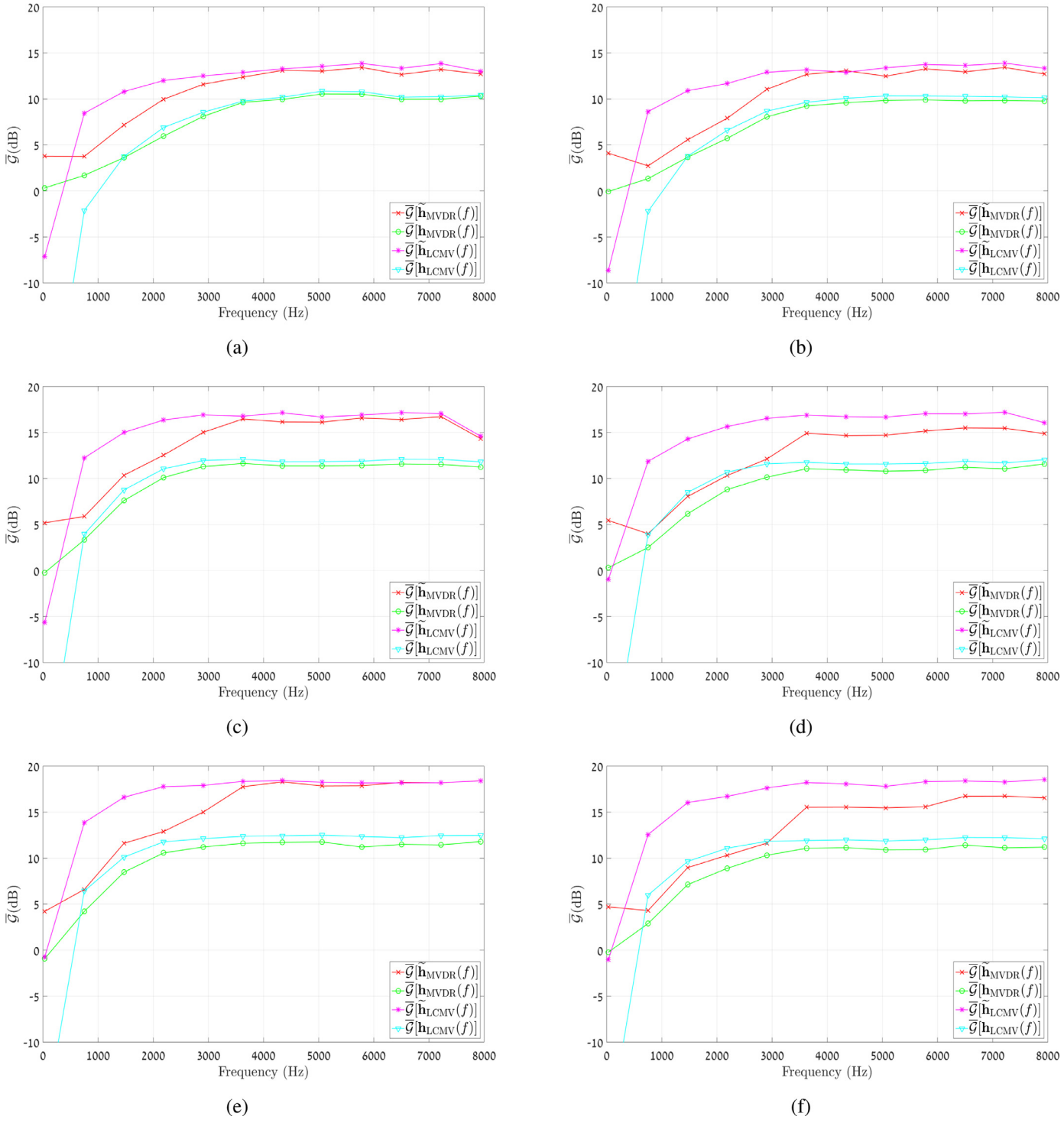


Fig. 4. Average output SNR gains as a function of the frequency for various array sizes and input SNRs. (a) iSNR = 0 dB and $M = 3$, (b) iSNR = 7 dB and $M = 3$, (c) iSNR = 0 dB and $M = 5$, (d) iSNR = 7 dB and $M = 5$, (e) iSNR = 0 dB and $M = 7$, and (f) iSNR = 7 dB and $M = 7$.

Table 1
Average PESQ scores for iSNR = 0dB and iSNR = 7dB with varying array sizes.

	iSNR = 0 dB, $M = 3$	iSNR = 0 dB, $M = 5$	iSNR = 0 dB, $M = 7$	iSNR = 7 dB, $M = 3$	iSNR = 7 dB, $M = 5$	iSNR = 7 dB, $M = 7$
$Y_1(f)$	1.502	1.502	1.502	1.982	1.982	1.982
$\widetilde{\mathbf{h}}_{\mathrm{MVDR}}(f)$	2.235	2.741	2.762	2.644	2.987	3.027
$\mathbf{h}_{\mathrm{MVDR}}(f)$	1.939	2.313	2.537	2.494	2.769	2.977
$\widetilde{\mathbf{h}}_{\mathrm{LCMV}}(f)$	2.123	2.748	2.817	2.627	3.033	3.082
$\mathbf{h}_{\mathrm{LCMV}}(f)$	1.628	2.155	2.478	2.192	2.738	3.022

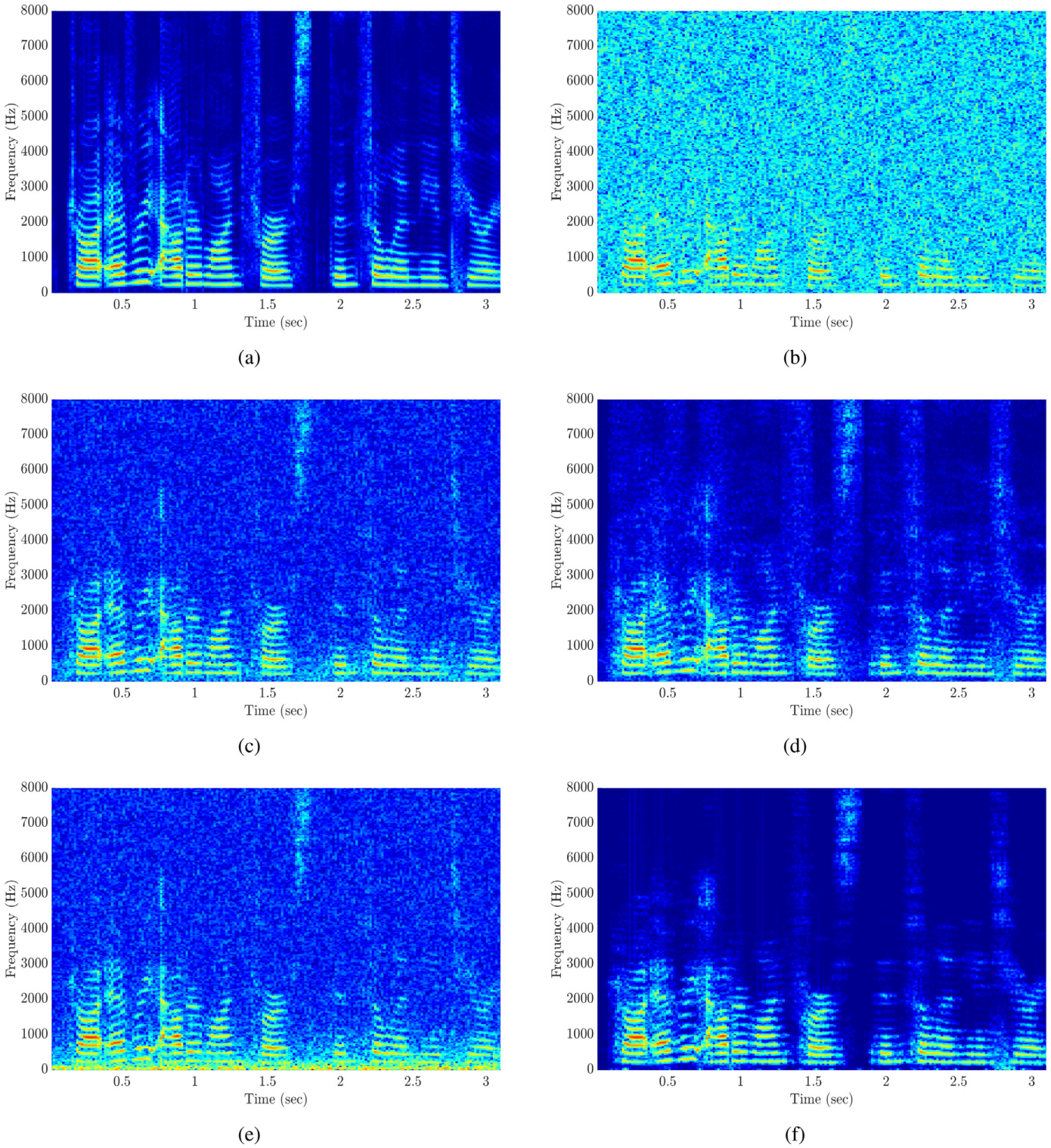


Fig. 5. Spectrograms of the desired signal, noisy input signal, and the enhanced signals in the presence of a directional interference and white thermal noise. The array size is $M = 5$ microphones and the input SNR is iSNR = 0 dB. (a) Clean desired signal, (b) noisy input signal at the first (reference) microphone, (c) enhanced signal with $\mathbf{h}_{\mathrm{MVDR}}(f)$, (d) enhanced signal with $\widetilde{\mathbf{h}}_{\mathrm{MVDR}}(f)$, (e) enhanced signal with $\mathbf{h}_{\mathrm{LCMV}}(f)$, and (f) enhanced signal with $\widetilde{\mathbf{h}}_{\mathrm{LCMV}}(f)$.

Next, we examine a more complicated set of scenarios in which in addition to the thermal white Gaussian noise, there are up to three reverberant spatial interferences. The problem settings are modified as follows. A uniform linear array of $M = 5$ microphones is located around (3, 5, 1.5) in the same room described above. The microphone array spacing is 0.05m, which implies that the total array length is 0.2m. Additionally, 3 white interference sources are placed on the $z = 1.5$ plane: U_1@(1, 5, 1.5), U_2@(1, 1, 1.5), and U_3@(5, 1, 1.5). An illustration of the $z = 1.5$ plane of the reverberant room is depicted in Fig. 6. The activity of the interference sources is set according to the 3

Table 2
Average PESQ scores for iSNR = 0 dB and iSNR = 7 dB with varying values of T_{60}. The background noise is thermal white Gaussian noise and $M = 3$.

	iSNR = 0 dB, T_{60} = 0ms	iSNR = 0 dB, T_{60} = 250ms	iSNR = 0 dB, T_{60} = 400ms	iSNR = 7 dB, T_{60} = 0ms	iSNR = 7 dB, T_{60} = 250ms	iSNR = 7 dB, T_{60} = 400ms
Noisy reverberant signal	1.87	2.04	1.98	2.34	2.33	2.2
Reverberant enhanced signal with $\mathbf{h}_{\mathrm{MVDR}}(f)$	2.19	2.18	2.13	2.64	2.41	2.3
Reverberant enhanced signal with $\tilde{\mathbf{h}}_{\mathrm{MVDR}}(f)$	2.38	2.27	2.2	2.75	2.46	2.33
Clean reverberant signal	4.5	2.72	2.43	4.5	2.72	2.43

Table 3
Average PESQ scores for the 3 aforementioned scenarios with iSNR = 0 dB and iSNR = 7 dB, T_{60} = 250msec and $M = 5$.

	iSNR = 0 dB, Scen. (a)	iSNR = 0 dB, Scen. (b)	iSNR = 0 dB, Scen. (c)	iSNR = 7 dB, Scen. (a)	iSNR = 7 dB, Scen. (b)	iSNR = 7 dB, Scen. (c)
Noisy reverberant signal	2.11	2.19	2.24	2.39	2.45	2.49
Reverberant enhanced signal with $\mathbf{h}_{\mathrm{LCMV}}(f)$	2.27	2.22	1.34	2.43	2.45	1.58
Reverberant enhanced signal with $\tilde{\mathbf{h}}_{\mathrm{LCMV}}(f)$	2.46	2.46	2.43	2.49	2.47	2.46
Reverberant enhanced signal with $\mathbf{h}_{\mathrm{MVDR}}(f)$	2.3	2.34	2.36	2.44	2.49	2.5
Reverberant enhanced signal with $\tilde{\mathbf{h}}_{\mathrm{MVDR}}(f)$	2.48	2.5	2.48	2.49	2.51	2.5
Clean reverberant signal	2.72	2.72	2.72	2.72	2.72	2.72

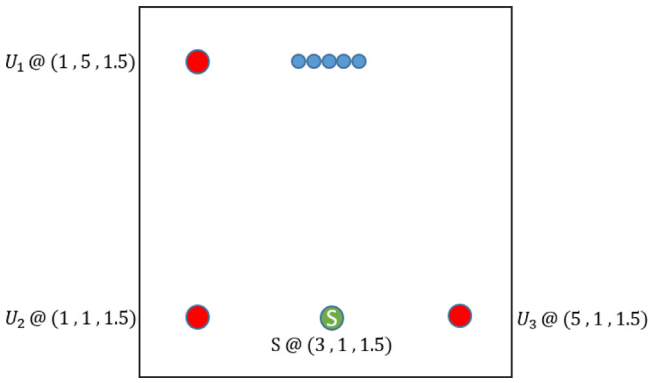


Fig. 6. An illustration of the $z = 1.5$ plane of the reverberant room.

following scenarios:

Scen. (a): only U_1 is active,
Scen. (b): only U_1 and U_2 are active,
Scen. (c): all interference sources are active.

In addition, a thermal white Gaussian noise is present. We note that in all scenarios, each interference and background noise are equally powerful. The received signal is individually simulated for each of the microphones using the RIR generator with $T_{60} = 250$ms. The set of simulations is performed for both iSNR = 0 and iSNR = 7, where the input noise, according to which the input SNR is calculated, takes into account the free field interferences and the thermal white Gaussian noise. Since in these scenarios interferences are present, it is interesting to examine the LCMV filters as well, for which the appropriate $L - 1$ null constraints are set in the direction of the direct path of the interferences. This implies that we have $L = 2$ in Scen. (a), $L = 3$ in Scen. (b), and $L = 4$ in Scen. (c). The PESQ scores of all 4 filters in the 3 scenarios and both input SNRs are shown in Table 3. We observe the following. Indeed, as the input SNR increases, the speech quality with the conventional MVDR and LCMV significantly improves. In contrast, this is not the case with the two KP filters, whose average PESQ scores remain roughly unchanged upon increasing the input SNR from iSNR = 0 to iSNR = 7. This is somewhat similar to the anechoic environment simulations of Section 6.1: The KP approach is of a better potential particularly in low input SNRs. Additionally, we note that while the conventional LCMV significantly enhances the noise in both input SNRs of Scen. (c), the KP-LCMV attains a considerable noise reduction in the low input SNR and only a slight quality deterioration in the high input SNR. Since both LCMV filters are designed to zero the direct paths of the directional interferences, we deduce that the KP-LCMV is potentially preferable in terms of white noise gain.

7. Conclusions

Conventional multichannel filters such as the MVDR and the LCMV are often used to estimate desired signals in noisy environments. Although they perform well when the input SNR is relatively high and with a large number of channels (i.e., sensors), their performances may not always be sufficient when the arrays contain only a few sensors. We have introduced a KP filtering approach to estimate the spectral power of the desired signal by exploiting higher-order statistics. We analyzed a toy example and performed a series of speech signals simulations in both anechoic and reverberant environments. We demonstrated that the proposed KP-MVDR and KP-LCMV outperform their conventional counterparts when proper temporal smoothing is employed to estimate the correlation matrix of the modified observation signal vector. This is emphasized in particular when the number of sensors is small or when the input SNR is low.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the Israel Science Foundation (grant no. 576/16) and the ISF-NSFC joint research program (grant no. 2514/17).

Appendix

Derivation of (30)

Let us evaluate the KP correlation matrix $\mathbf{\Phi}_{\tilde{\mathbf{y}}}(f)$. We have

$$\begin{aligned}\mathbf{\Phi}_{\tilde{\mathbf{y}}}(f) = & E\Big\{\left[(\mathbf{x}(f)+\mathbf{v}(f))^{*}\otimes(\mathbf{x}(f)+\mathbf{v}(f))\right]\\ & \left[(\mathbf{x}(f)+\mathbf{v}(f))^{T}\otimes(\mathbf{x}(f)+\mathbf{v}(f))^{H}\right]\Big\}\\ = & E\Big\{\left[\mathbf{x}^{*}(f)\otimes\mathbf{x}(f)+\mathbf{x}^{*}(f)\otimes\mathbf{v}(f)+\mathbf{v}^{*}(f)\otimes\mathbf{x}(f)+\mathbf{v}^{*}(f)\otimes\mathbf{v}(f)\right]\\ & \left[\mathbf{x}^{T}(f)\otimes\mathbf{x}^{H}(f)+\mathbf{x}^{T}(f)\otimes\mathbf{v}^{H}(f)+\mathbf{v}^{T}(f)\otimes\mathbf{x}^{H}(f)\right.\\ & \left.+\mathbf{v}^{T}(f)\otimes\mathbf{v}^{H}(f)\right]\Big\},\end{aligned} \tag{42}$$

which is a sum of 16 $M^2\times M^2$ matrices. However, since $E[\mathbf{x}(f)]=E[\mathbf{v}(f)]=\mathbf{0}$, and $E\left[\mathbf{x}(f)\mathbf{v}^{H}(f)\right]=E\left[\mathbf{v}(f)\mathbf{x}^{H}(f)\right]=\mathbf{0}$, precisely 8 out of these 16 matrices are strictly zero. The other 8 matrices are the desired signal-only matrix, the noise-only matrix and 6 non-zero mixed matrices. We begin with the desired signal-only matrix. We recall that $M=2$, and the desired signal is normally distributed, i.e., $x_1(nT_s)=x_2(nT_s)\sim\mathcal{N}(0,f_x)$. Hence, its frequency domain representation is complex-normally distributed, that is

$$X_1(f)\sim\mathcal{CN}\left[\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}K\epsilon_x & 0\\ 0 & K\epsilon_x\end{bmatrix}\right], \tag{43}$$

where K is half of the number of bins in a single STFT frame. Then

$$\begin{aligned}&E\left\{\left[\mathbf{x}^{*}(f)\otimes\mathbf{x}(f)\right]\left[\mathbf{x}^{T}(f)\otimes\mathbf{x}^{H}(f)\right]\right\}\\ &=E\left[|X_1(f)|^4\right]\begin{bmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{bmatrix}\\ &=8\epsilon_x^2\begin{bmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{bmatrix}.\end{aligned}$$

We move on to the noise-only matrix. Since the noise distribution is normal as well, the frequency domain representations of the noise terms are complex-normally distributed, and we easily obtain

$$E\left\{\left[\mathbf{v}^{*}(f)\otimes\mathbf{v}(f)\right]\left[\mathbf{v}^{T}(f)\otimes\mathbf{v}^{H}(f)\right]\right\}=4\epsilon_v^2\begin{bmatrix}2&0&0&1\\0&1&0&0\\0&0&1&0\\1&0&0&2\end{bmatrix}.$$

Similarly, the following 6 mixed matrices are given by

$$\begin{aligned}&E\left\{\left[\mathbf{v}^{*}(f)\otimes\mathbf{x}(f)\right]\left[\mathbf{x}^{T}(f)\otimes\mathbf{v}^{H}(f)\right]\right\}\\ &=E\left\{\left[\mathbf{x}^{*}(f)\otimes\mathbf{v}(f)\right]\left[\mathbf{v}^{T}(f)\otimes\mathbf{x}^{H}(f)\right]\right\}\\ &=\mathbf{0},\end{aligned}$$

$$\begin{aligned}&E\left\{\left[\mathbf{x}^{*}(f)\otimes\mathbf{v}(f)\right]\left[\mathbf{x}^{T}(f)\otimes\mathbf{v}^{H}(f)\right]\right\}\\ &=4\epsilon_x\epsilon_v\begin{bmatrix}1&0&1&0\\0&1&0&1\\1&0&1&0\\0&1&0&1\end{bmatrix},\\ &E\left\{\left[\mathbf{v}^{*}(f)\otimes\mathbf{x}(f)\right]\left[\mathbf{v}^{T}(f)\otimes\mathbf{x}^{H}(f)\right]\right\}\\ &=4\epsilon_x\epsilon_v\begin{bmatrix}1&1&0&0\\1&1&0&0\\0&0&1&1\\0&0&1&1\end{bmatrix},\end{aligned}$$

$$\begin{aligned}&E\left\{\left[\mathbf{x}^{*}(f)\otimes\mathbf{x}(f)\right]\left[\mathbf{v}^{T}(f)\otimes\mathbf{v}^{H}(f)\right]\right\}\\ &=4\epsilon_x\epsilon_v\begin{bmatrix}1&0&0&1\\1&0&0&1\\1&0&0&1\\1&0&0&1\end{bmatrix},\end{aligned}$$

and

$$\begin{aligned}&E\left\{\left[\mathbf{v}^{*}(f)\otimes\mathbf{v}(f)\right]\left[\mathbf{x}^{T}(f)\otimes\mathbf{x}^{H}(f)\right]\right\}\\ &=4\epsilon_x\epsilon_v\begin{bmatrix}1&1&1&1\\0&0&0&0\\0&0&0&0\\1&1&1&1\end{bmatrix}.\end{aligned}$$

Combining all the matrices together and dividing by $4\epsilon_x\epsilon_v$, we obtain the expression in (30).

References

Allen, J.B., Berkley, D.A., 1979. Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. 65 (4), 943–950.

Bai, M., IH, J., Benesty, J., 2014. Acoustic Array Systems: Theory, Implementation, and Application. Wiley-IEEE Press.

Benesty, J., Chen, J., 2011. Optimal Time-Domain Noise Reduction Filters; A Theoretical Study. Springer-Verlag, Berlin Heidelberg doi:10.1007/978-3-642-19601-0.

Benesty, J., Chen, J., Habets, E.A.P., 2012. Speech Enhancement in the STFT Domain. Springer-Verlag, Berlin Heidelberg doi:10.1007/978-3-642-23250-3.

Benesty, J., Chen, J., Huang, Y., 2008. Microphone Array Signal Processing. Springer-Verlag, Berlin Heidelberg doi:10.1007/978-3-540-78612-2.

Benesty, J., Chen, J., Huang, Y., 2010. A widely linear distortionless filter for single-channel noise reduction. IEEE Signal Process. Lett. 17 (5), 469–472. doi:10.1109/LSP.2010.2043152.

Benesty, J., Chen, J., Huang, Y., Cohen, I., 2009. Noise Reduction in Speech Processing, first ed. Springer-Verlag, Berlin Heidelberg.

Benesty, J., Chen, J., Huang, Y., Dmochowski, J., 2007. On microphone-array beamforming from a mimo acoustic signal processing perspective. IEEE Trans. Audio Speech Lang. Process. 15 (3), 1053–1065. doi:10.1109/TASL.2006.885251.

Benesty, J., Chen, J., Huang, Y.A., 2009. Noise reduction algorithms in a generalized transform domain. IEEE Trans. Audio Speech Lang. Process. 17 (6), 1109–1123. doi:10.1109/TASL.2009.2020415.

Benesty, J., Cohen, I., Chen, J., 2017. Fundamentals of Signal Enhancement and Array Signal Processing. Wiley-IEEE Press.

Buchris, Y., Cohen, I., Benesty, J., 2019. On the design of time-domain differential microphone arrays. Appl. Acoust. 148, 212–222.

Capon, J., 1969. High-resolution frequency-wavenumber spectrum analysis. Proc. IEEE 57 (8), 1408–1418. doi:10.1109/PROC.1969.7278.

Chen, J., Huang, Y., Benesty, J., 2003. Filtering Techniques for Noise Reduction and Speech Enhancement. Springer, Berlin Heidelberg, pp. 129–154. doi:10.1007/978-3-662-11028-7_5.

Darpa timit acoustic phonetic continuous speech corpus cdrom, 1993.

Dmochowski, J., Benesty, J., 2010. Microphone Arrays: Fundamental Concepts. Springer, Berlin Heidelberg, pp. 199–223. doi:10.1007/978-3-642-11130-3_8.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. 32 (6), 1109–1121. doi:10.1109/TASSP.1984.1164453.

Ephraim, Y., Trees, H.L.V., 1995. A signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Process. 3 (4), 251–266. doi:10.1109/89.397090.

Gannot, S., Cohen, I., 2008. Adaptive Beamforming and Postfiltering. Springer, Berlin Heidelberg, pp. 945–978. doi:10.1007/978-3-540-49127-9_47.

Griffiths, L., Jim, C., 1982. An alternative approach to linearly constrained adaptive beamforming. IEEE Trans. Antennas Propag. 30 (1), 27–34. doi:10.1109/TAP.1982.1142739.

Habets, E. A. P.,. Rir-generator 2014. https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator.

Lacoss, R.T., 1971. Data adaptive spectral analysis methods. Geophysics 36 (4), 661–675. doi:10.1190/1.1440203.

Lacouture-Parodi, Y., Habets, E.A.P., Chen, J., Benesty, J., 2014. Multichannel noise reduction in the Karhunen-Loeve expansion domain. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (5), 923–936. doi:10.1109/TASLP.2014.2311299.

Moreno, A., Fonollosa, J., 1992. Pitch determination of noisy speech using higher order statistics. Vol. 1, pp. 133–136 vol.1. doi:10.1109/ICASSP.1992.225954.

Nemer, E., Goubran, R., Mahmoud, S., 2001. Robust voice activity detection using higher-order statistics in the lpc residual domain. IEEE Trans. Speech Audio Process. 9 (3), 217–231. doi:10.1109/89.905996.

Nemer, E., Goubran, R., Mahmoud, S., 2002. Speech enhancement using fourth-order cumulants and optimum filters in the subband domain. Speech Commun. 36 (3), 219–246.

Pierce, A., 1991. Acoustics: An Introduction to Its Physical Principles and Applications. Acoustical Society of America.

Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P., 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings, 2 doi:10.1109/ICASSP.2001.941023. 749–752 vol.2.

Sena, E.D., Antonello, N., Moonen, M., van Waterschoot, T., 2015. On the modeling of rectangular geometries in room acoustic simulations. IEEE/ACM Trans. Audio Speech Lang. Process. 23 (4), 774–786. doi:10.1109/TASLP.2015.2405476.

Souden, M., Benesty, J., Affes, S., 2010. A study of the LCMV and MVDR noise reduction filters. IEEE Trans. Signal Process. 58 (9), 4925–4935. doi:10.1109/TSP.2010.2051803.

Tavakoli, V., Jensen, J., Christensen, M., Benesty, J., 2016. A framework for speech enhancement with ad hoc microphone arrays. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (6), 1038–1051.

Van Trees, H., 2004. Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory. Wiley.

Welch, P., 1967. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans. Audio Electroacoust. 15 (2), 70–73. doi:10.1109/TAU.1967.1161901.

Chapter 4

Quadratic Approach for Single-channel Noise Reduction

Itzhak et al. EURASIP Journal on Audio, Speech, and Music Processing (2020) 2020:7
https://doi.org/10.1186/s13636-020-00174-4

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Quadratic approach for single-channel noise reduction

Gal Itzhak[1]*, Jacob Benesty[2] and Israel Cohen[1]

Abstract

In this paper, we introduce a quadratic approach for single-channel noise reduction. The desired signal magnitude is estimated by applying a linear filter to a modified version of the observations' vector. The modified version is constructed from a Kronecker product of the observations' vector with its complex conjugate. The estimated signal magnitude is multiplied by a complex exponential whose phase is obtained using a conventional linear filtering approach. We focus on the linear and quadratic maximum signal-to-noise ratio (SNR) filters and demonstrate that the quadratic filter is superior in terms of subband SNR gains. In addition, in the context of speech enhancement, we show that the quadratic filter is ideally preferable in terms of perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) scores. The advantages, compared to the conventional linear filtering approach, are particularly significant for low input SNRs, at the expanse of a higher computational complexity. The results are verified in practical scenarios with nonstationary noise and in comparison to well-known speech enhancement methods. We demonstrate that the quadratic maximum SNR filter may be superior, depending on the nonstationary noise type.

Keywords: Quadratic filtering, Maximum SNR filter, Frequency-domain filtering, Optimal filters, Nonlinear processing, Kronecker product

1 Introduction

Communications and signal processing systems are very likely to operate in adverse environments, which are characterized by the presence of background noise that might severely degrade the quality of desired signals. Noise reduction methods are designed and applied to noisy signals with the objective of improving their quality and attenuating the background noise. Single-channel noise reduction (SCNR) methods are often implemented in physically small or low cost systems. SCNR filters are usually derived by minimizing a given distortion function between the clean signal and its estimate, or by minimizing the energy of the residual noise under some constraints.

Frequency-domain methods, e.g., [1–6], are typically formulated on a frame basis, that is, a frame of noisy observations is transformed into the frequency (or time-frequency) domain using the short-time Fourier transform (STFT). Then, the optimal filter is derived in the chosen domain and applied to the transformed observations. Finally, the filtered observations are transformed back to the time domain using the inverse STFT.

It is clear by construction that signals in the frequency domain are complex. Nonetheless, in many cases, most of the information in a desired signal is stored in its spectral magnitude. Indeed, this property is well known for speech signals, whose spectral magnitude has received special attention in the context of statistical models and optimal estimators, e.g., a maximum-likelihood spectral magnitude estimator [1], short-time spectral [2], log-spectral [3] and optimally modified log-spectral [7] magnitude estimators, and a maximum a posteriori spectral magnitude estimator [8]. These celebrated estimators assume that time trajectories in the STFT domain of clean speech and noise signals are independent complex Gaussian random processes. Other statistical models, e.g., super-Gaussian [9–11], Gamma [12, 13], or Laplace

*Correspondence: sgalitz007@gmail.com
[1]Technion, Israel Institute of Technology, 32000, Haifa, Israel
Full list of author information is available at the end of the article

[14, 15] distributions, were also investigated and were demonstrated to be potentially more effective, depending on the desired speech spectral magnitude estimator and the speech conditional variance evolution model. While all the foregoing estimators rely on the strong correlation between magnitudes of successive coefficients (in a fixed frequency) [5, 16, 17], their derivation is typically cumbersome and requires one to numerically evaluate non-analytical functions following the assumed statistical speech and noise models. Moreover, with the aforementioned spectral magnitude correlation hidden behind first-order recursive temporal processes, additional parameters and lower boundaries must carefully be set to guarantee the model tracking over time.

Recently, it has been proposed to exploit the self-correlation property of STFT domain coefficients in a linear manner. That is, instead of explicitly assuming statistical models which depend on unobserved measures, e.g., the a priori SNR, it was suggested to employ linear filters which require the second-order statistics of the desired signal and noise. These linear filters are derived within a multi-frame framework that takes into account the interframe correlation of the STFT coefficients from successive time frames and adjacent frequencies [5, 18, 19]. The multi-frame formulation highly resembles a sensor array formulation, which implies that conventional array filters may be modified for the single-channel case, but with an interframe correlation interpretation rather than spatial sensing. Examples of such filters are the Wiener filter, the minimum variance distortionless response (MVDR) filter [5, 18], the linearly constrained minimum variance (LCMV) filter [5], and the maximum SNR filter [19].

In this paper, we present a quadratic approach for SCNR which extends the multi-frame approach suggested in [18]. The interframe correlation property is taken into account in the same manner as in [18], but the noise reduction filters are not applied to the observations' vector directly, but rather to its modified version. The modified version is obtained from the Kronecker product of the observations' vector and its complex conjugate. In its mathematical formulation, this approach is similar to the approach presented in [20] in the context of multi-channel noise reduction. On the contrary, while in [20] the essence of the innovation is the direct utilization of higher-order statistics, the key idea in this work is a generalization of the single-channel linear filtering approach. We demonstrate that by focusing on the estimation of the desired signal magnitude in the transform domain, we are able to achieve further reduction of the background noise. More specifically, we propose the quadratic maximum SNR filter, which may potentially achieve a theoretically unbounded subband output SNR. We compare the quadratic and the linear maximum SNR filters and demonstrate that the quadratic filter is superior, in particular in low input SNR environments.

The rest of the paper is organized as follows. In Section 2, we present the signal model and formulate the SCNR problem. In Section 3, we introduce the quadratic filtering approach, from which quadratic filters may be derived. In Section 4, we propose a quadratic maximum SNR filter and derive it from two different perspectives. In Section 5, we focus on a toy example and theoretically evaluate the performances of the linear and quadratic maximum SNR filters. Finally, in Section 6, we demonstrate the noise reduction capabilities of the quadratic maximum SNR filter. We compare its performance to existing speech enhancement methods in ideal and practical conditions and in the presence of nonstationary noise.

2 Signal model and problem formulation

We consider the classical single-channel noise reduction problem, where the noisy signal at time index t is given by [21, 22]:

$$y(t) = x(t) + v(t), \tag{1}$$

with $x(t)$ and $v(t)$ denoting the desired signal and additive noise, respectively. We assume that $x(t)$ and $v(t)$ are uncorrelated and that all signals are real, zero mean, and broadband.

By employing the STFT or any other appropriate transform as suggested in [23], (1) can be rewritten in terms of the transform domain coefficients as:

$$Y(k,n) = X(k,n) + V(k,n), \tag{2}$$

where the zero-mean complex random variables $Y(k,n)$, $X(k,n)$, and $V(k,n)$ are the analysis coefficients of $y(t)$, $x(t)$, and $v(t)$, respectively, at the frequency index $k \in \{0,1,\ldots,K-1\}$ and time-frame index n. It is well known that the same signal at different time frames is correlated [17]. Therefore, the interframe correlation should be taken into account in order to improve the performance of noise reduction algorithms. In this case, we may consider forming an observation signal vector of length N, containing the N most recent samples of $Y(k,n)$, i.e.,

$$\begin{aligned}\mathbf{y}(k,n) &= \left[\, Y(k,n) \;\cdots\; Y(k,n-N+1) \,\right]^T \\ &= \mathbf{x}(k,n) + \mathbf{v}(k,n),\end{aligned} \tag{3}$$

where the superscript ^{T} is the transpose operator, and $\mathbf{x}(k,n)$ and $\mathbf{v}(k,n)$ are defined similarly to $\mathbf{y}(k,n)$. Then, the objective of noise reduction is to estimate the desired signal $X(k,n)$ from the noisy observation signal vector $\mathbf{y}(k,n)$.

Since $x(t)$ and $v(t)$ are uncorrelated by assumption, the $N \times N$ correlation matrix of $\mathbf{y}(k,n)$ is

$$\begin{aligned}\Phi_{\mathbf{y}}(k,n) &= E\left[\mathbf{y}(k,n)\mathbf{y}^H(k,n)\right] \\ &= \Phi_{\mathbf{x}}(k,n) + \Phi_{\mathbf{v}}(k,n),\end{aligned} \tag{4}$$

where the superscript ^{H} is the conjugate-transpose operator, and $\Phi_{\mathbf{x}}(k,n)$ and $\Phi_{\mathbf{v}}(k,n)$ are the correlation matrices of $\mathbf{x}(k,n)$ and $\mathbf{v}(k,n)$, respectively.

We end this part by defining the subband input SNR as:

$$\text{iSNR}(k,n) = \frac{\phi_X(k,n)}{\phi_V(k,n)}, \tag{5}$$

where $\phi_X(k,n) = E\left[|X(k,n)|^2\right]$ and $\phi_V(k,n) = E\left[|V(k,n)|^2\right]$ are the variances of $X(k,n)$ and $V(k,n)$, respectively.

3 Quadratic filtering approach

In the conventional linear approach [5], noise reduction is performed by applying a complex-valued filter, $\mathbf{h}(k,n)$ of length N, to the observation signal vector, $\mathbf{y}(k,n)$, i.e.,

$$\begin{aligned}\widehat{X}(k,n) &= \mathbf{h}^H(k,n)\mathbf{y}(k,n) \\ &= X_{\text{fd}}(k,n) + V_{\text{rn}}(k,n),\end{aligned} \tag{6}$$

where the filter output, $\widehat{X}(k,n)$, is an estimate of $X(k,n)$; $X_{\text{fd}}(k,n) = \mathbf{h}^H(k,n)\mathbf{x}(k,n)$ is the filtered desired signal; and $V_{\text{rn}}(k,n) = \mathbf{h}^H(k,n)\mathbf{v}(k,n)$ is the residual noise.

The two terms on the right-hand side of (6) are uncorrelated. Hence, the variance of $\widehat{X}(k,n)$ is:

$$\begin{aligned}\phi_{\widehat{X}}(k,n) &= \mathbf{h}^H(k,n)\Phi_{\mathbf{y}}(k,n)\mathbf{h}(k,n) \\ &= \phi_{X_{\text{fd}}}(k,n) + \phi_{V_{\text{rn}}}(k,n),\end{aligned} \tag{7}$$

where $\phi_{X_{\text{fd}}}(k,n) = \mathbf{h}^H(k,n)\Phi_{\mathbf{x}}(k,n)\mathbf{h}(k,n)$ is the variance of the filtered desired signal and $\phi_{V_{\text{rn}}}(k,n) = \mathbf{h}^H(k,n)\Phi_{\mathbf{v}}(k,n)\mathbf{h}(k,n)$ is the variance of the residual noise. Then, from (7), the subband output SNR is given by:

$$\text{oSNR}\left[\mathbf{h}(k,n)\right] = \frac{\mathbf{h}^H(k,n)\Phi_{\mathbf{x}}(k,n)\mathbf{h}(k,n)}{\mathbf{h}^H(k,n)\Phi_{\mathbf{v}}(k,n)\mathbf{h}(k,n)}. \tag{8}$$

The quadratic filtering approach emerges from a different perspective. First, assuming that the desired signal is estimated with the linear approach, we find an expression for the energy of the estimated desired signal $\left|\widehat{X}(k,n)\right|^2$. We have:

$$\begin{aligned}\left|\widehat{X}(k,n)\right|^2 &= \mathbf{h}^H(k,n)\mathbf{y}(k,n)\mathbf{y}^H(k,n)\mathbf{h}(k,n) \\ &= \text{tr}\left[\mathbf{y}(k,n)\mathbf{y}^H(k,n)\mathbf{h}(k,n)\mathbf{h}^H(k,n)\right] \\ &= \text{vec}^H\left[\mathbf{h}(k,n)\mathbf{h}^H(k,n)\right]\text{vec}\left[\mathbf{y}(k,n)\mathbf{y}^H(k,n)\right] \\ &= \left[\mathbf{h}^*(k,n)\otimes\mathbf{h}(k,n)\right]^H\left[\mathbf{y}^*(k,n)\otimes\mathbf{y}(k,n)\right] \\ &= \left[\mathbf{h}^*(k,n)\otimes\mathbf{h}(k,n)\right]^H\widetilde{\mathbf{y}}(k,n),\end{aligned} \tag{9}$$

where $\text{tr}[\cdot]$ is the trace of a square matrix; $\text{vec}[\cdot]$ is the vectorization operator, which consists of converting a matrix into a vector; $\otimes$ denotes the Kronecker product [24]; and $\widetilde{\mathbf{y}}(k,n) = \mathbf{y}^*(k,n)\otimes\mathbf{y}(k,n)$ is a vector of length N^2.

Let $\widetilde{\mathbf{h}}(k,n)$ be a general complex-valued filter of length N^2, which is not necessarily of the form $\mathbf{h}^*(k,n)\otimes\mathbf{h}(k,n)$. From (9), we can generate an estimate of $\left|\widehat{X}(k,n)\right|^2$ by applying the filter $\widetilde{\mathbf{h}}(k,n)$ to $\widetilde{\mathbf{y}}(k,n)$, i.e.,

$$Z(k,n) = \widetilde{\mathbf{h}}^H(k,n)\widetilde{\mathbf{y}}(k,n), \tag{10}$$

where $Z(k,n)$ is the estimate of the desired signal energy. Indeed, this approach generalizes the conventional linear approach, since (10) reduces to (9) with quadratic filters of the form $\widetilde{\mathbf{h}}(k,n) = \mathbf{h}^*(k,n)\otimes\mathbf{h}(k,n)$.

With $Z(k,n)$, we can obtain an estimate of the desired signal:

$$\widehat{X}(k,n) = e^{J\psi(k,n)}\sqrt{|Z(k,n)|}, \tag{11}$$

where the phase $\psi(k,n)$ can be taken from the linear approach (6). We note that in practice, this implies an additional computational complexity, as a linear filter might have to be implemented for the purpose of obtaining a desired signal phase estimate. Clearly, this approach is highly nonlinear.

Next, we would like to derive a theoretical expression for the subband output SNR with the quadratic approach. We have:

$$\begin{aligned}\widetilde{\mathbf{y}}(k,n) &= \mathbf{y}^*(k,n)\otimes\mathbf{y}(k,n) \\ &= \left[\mathbf{x}^*(k,n)+\mathbf{v}^*(k,n)\right]\otimes\left[\mathbf{x}(k,n)+\mathbf{v}(k,n)\right] \\ &= \widetilde{\mathbf{x}}(k,n) + \mathbf{x}^*(k,n)\otimes\mathbf{v}(k,n) \\ &\quad + \mathbf{v}^*(k,n)\otimes\mathbf{x}(k,n) + \widetilde{\mathbf{v}}(k,n),\end{aligned} \tag{12}$$

where $\widetilde{\mathbf{x}}(k,n) = \mathbf{x}^*(k,n)\otimes\mathbf{x}(k,n)$ and $\widetilde{\mathbf{v}}(k,n) = \mathbf{v}^*(k,n)\otimes\mathbf{v}(k,n)$. Taking mathematical expectation on both sides of (12), we have:

$$\begin{aligned}E\left[\widetilde{\mathbf{y}}(k,n)\right] &= E\left[\widetilde{\mathbf{x}}(k,n)\right] + E\left[\widetilde{\mathbf{v}}(k,n)\right] \\ &= \text{vec}\left[\Phi_{\mathbf{x}}(k,n)\right] + \text{vec}\left[\Phi_{\mathbf{v}}(k,n)\right] \\ &= \text{vec}\left[\Phi_{\mathbf{y}}(k,n)\right].\end{aligned} \tag{13}$$

We deduce that:

$$\begin{aligned}E\left[Z(k,n)\right] &= \widetilde{\mathbf{h}}^H(k,n)E\left[\widetilde{\mathbf{y}}(k,n)\right] \\ &= \widetilde{\mathbf{h}}^H(k,n)\text{vec}\left[\Phi_{\mathbf{x}}(k,n)\right] \\ &\quad + \widetilde{\mathbf{h}}^H(k,n)\text{vec}\left[\Phi_{\mathbf{v}}(k,n)\right].\end{aligned} \tag{14}$$

Consequently, the variance of $\widehat{X}(k,n)$ is:

$$\begin{aligned}\phi_{\widehat{X}}(k,n) &= E\left[|Z(k,n)|\right] \\ &\approx \left|E\left[Z(k,n)\right]\right| \\ &= \left|\widetilde{\mathbf{h}}^H(k,n)E\left[\widetilde{\mathbf{y}}(k,n)\right]\right| \\ &= \Big|\widetilde{\mathbf{h}}^H(k,n)\text{vec}\left[\Phi_{\mathbf{x}}(k,n)\right] \\ &\quad + \widetilde{\mathbf{h}}^H(k,n)\text{vec}\left[\Phi_{\mathbf{v}}(k,n)\right]\Big|,\end{aligned} \tag{15}$$

where the approximation in the second row of (15) assumes $Z(k,n)$ to be real and positive. Thus, we can

define the subband output SNR corresponding to a general quadratic filter $\widetilde{\mathbf{h}}(k,n)$ of length N^2 as:

$$\text{oSNR}\left[\widetilde{\mathbf{h}}(k,n)\right] = \frac{\left|\widetilde{\mathbf{h}}^H(k,n)\text{vec}\left[\Phi_{\mathbf{x}}(k,n)\right]\right|}{\left|\widetilde{\mathbf{h}}^H(k,n)\text{vec}\left[\Phi_{\mathbf{v}}(k,n)\right]\right|} \quad (16)$$

$$= \sqrt{\frac{\widetilde{\mathbf{h}}^H(k,n)\text{vec}\left[\Phi_{\mathbf{x}}(k,n)\right]\text{vec}^H\left[\Phi_{\mathbf{x}}(k,n)\right]\widetilde{\mathbf{h}}(k,n)}{\widetilde{\mathbf{h}}^H(k,n)\text{vec}\left[\Phi_{\mathbf{v}}(k,n)\right]\text{vec}^H\left[\Phi_{\mathbf{v}}(k,n)\right]\widetilde{\mathbf{h}}(k,n)}}.$$

In Sections 4 and 5, in order to simplify the notation, we drop the dependence on the time and frequency indices. For example, (10) would be written as $Z = \widetilde{\mathbf{h}}^H\widetilde{\mathbf{y}}$.

4 Quadratic maximum SNR filter

In this section, we derive a filter $\widetilde{\mathbf{h}}$ that maximizes the output SNR given in (16). For theoretical completeness, the filter is derived from two different perspectives: by performing an eigenvalue decomposition to a rank deficient matrix defined by the noise statistics or by using an appropriate matrix projection operator.

The matrix $\text{vec}\left(\Phi_{\mathbf{v}}\right)\text{vec}^H\left[\Phi_{\mathbf{v}}\right]$ may be diagonalized using the eigenvalue decomposition [25] as:

$$\mathbf{U}^H\text{vec}\left(\Phi_{\mathbf{v}}\right)\text{vec}^H\left(\Phi_{\mathbf{v}}\right)\mathbf{U} = \Lambda, \quad (17)$$

where

$$\mathbf{U} = \left[\,\mathbf{u}_1\;\;\mathbf{U}'\,\right] \quad (18)$$

is a unitary matrix and

$$\Lambda = \text{diag}\left(\lambda_{\max}, 0, \ldots, 0\right) \quad (19)$$

is a diagonal matrix. The vector:

$$\mathbf{u}_1 = \frac{\text{vec}\left(\Phi_{\mathbf{v}}\right)}{\sqrt{\text{vec}^H\left(\Phi_{\mathbf{v}}\right)\text{vec}\left(\Phi_{\mathbf{v}}\right)}} \quad (20)$$

is the eigenvector corresponding to the only nonzero eigenvalue $\lambda_{\max} = \text{vec}^H\left(\Phi_{\mathbf{v}}\right)\text{vec}\left(\Phi_{\mathbf{v}}\right)$ of the matrix $\text{vec}\left(\Phi_{\mathbf{v}}\right)\text{vec}^H\left(\Phi_{\mathbf{v}}\right)$, while $\mathbf{U}'$ contains the other N^2-1 eigenvectors of the zero eigenvalues. It is clear from (17) that:

$$\mathbf{U}'^H\text{vec}\left(\Phi_{\mathbf{v}}\right) = \mathbf{0}. \quad (21)$$

Now, let us consider filters of the form:

$$\widetilde{\mathbf{h}}_{\max} = \mathbf{U}'\widetilde{\mathbf{h}}'_{\max}, \quad (22)$$

where $\widetilde{\mathbf{h}}'_{\max} \neq \mathbf{0}$ is a filter of length N^2-1. Substituting (22) into (16), we infer that the subband output SNR with $\widetilde{\mathbf{h}}_{\max}$ may be unbounded, as opposed to the strictly bounded subband output SNR with the linear maximum SNR filter [19].

We point out the following observation. Despite achieving a potentially unbounded subband output SNR, the filter $\widetilde{\mathbf{h}}_{\max}$ is not expected to result in zero residual noise, as in practice it is applied to a vector of instantaneous analysis coefficients, while it is designed to eliminate the statistical noise PSD. Nonetheless, we recall that any linear filter may be extended to an appropriate quadratic filter but not vice versa. That is, the linear filtering approach may be regarded as a constrained version of the quadratic filtering approach. Hence, we deduce that the subband output SNR with the quadratic maximum SNR filter should be equal or larger than the subband output SNR with the linear maximum SNR filter.

With the subband output SNR maximized, it is possible to find $\widetilde{\mathbf{h}}'_{\max}$ in such a way that the desired signal distortion is minimized. Since the first term on the right-hand side of (14) corresponds to the filtered desired signal, we take this term equal to the variance of the desired signal, i.e.,

$$\widetilde{\mathbf{h}}^H_{\max}\text{vec}\left(\Phi_{\mathbf{x}}\right) = \phi_X. \quad (23)$$

Substituting (22) into (23) and noting that $\widetilde{\mathbf{h}}'_{\max}$ should equal the vector $\mathbf{U}'^H\text{vec}\left(\Phi_{\mathbf{x}}\right)$ up to appropriate scaling factors, we obtain:

$$\widetilde{\mathbf{h}}'_{\max} = \frac{\mathbf{U}'^H\text{vec}\left(\Phi_{\mathbf{x}}\right)\phi_X}{\text{vec}^H\left(\Phi_{\mathbf{x}}\right)\mathbf{U}'\mathbf{U}'^H\text{vec}\left(\Phi_{\mathbf{x}}\right)}. \quad (24)$$

Therefore,

$$\widetilde{\mathbf{h}}_{\max} = \frac{\mathbf{U}'\mathbf{U}'^H\text{vec}\left(\Phi_{\mathbf{x}}\right)\phi_X}{\text{vec}^H\left(\Phi_{\mathbf{x}}\right)\mathbf{U}'\mathbf{U}'^H\text{vec}\left(\Phi_{\mathbf{x}}\right)}. \quad (25)$$

There is an alternative way to derive $\widetilde{\mathbf{h}}_{\max}$ from the first row of (16). That is, we may derive a filter $\widetilde{\mathbf{h}}_{\max,2}$ that is orthogonal to $\text{vec}\left[\Phi_{\mathbf{v}}\right]$, i.e., $\widetilde{\mathbf{h}}^H_{\max,2}\text{vec}\left(\Phi_{\mathbf{v}}\right) = 0$. While the previous derivation of $\widetilde{\mathbf{h}}_{\max}$ may be considered more comparable to $\mathbf{h}_{\max}$ as both filters employ an eigenvalue decomposition, the alternative derivation of $\widetilde{\mathbf{h}}_{\max,2}$ may be more convenient to implement and analyze, and is indeed utilized for the thoretical performance analysis in Section 5. Any filter whose form is:

$$\widetilde{\mathbf{h}}_{\max,2} = \widetilde{\mathbf{h}}'_{\max,2} - \frac{\text{vec}\left(\Phi_{\mathbf{v}}\right)\text{vec}^H\left(\Phi_{\mathbf{v}}\right)}{\text{vec}^H\left(\Phi_{\mathbf{v}}\right)\text{vec}\left(\Phi_{\mathbf{v}}\right)}\widetilde{\mathbf{h}}'_{\max,2}$$

$$= \mathbf{P}\widetilde{\mathbf{h}}'_{\max,2} \quad (26)$$

satisfies the condition, where $\widetilde{\mathbf{h}}'_{\max,2} \neq \mathbf{0}$ is an arbitrary complex-valued filter,

$$\mathbf{P} = \mathbf{I}_{N^2} - \frac{\text{vec}\left(\Phi_{\mathbf{v}}\right)\text{vec}^H\left(\Phi_{\mathbf{v}}\right)}{\text{vec}^H\left(\Phi_{\mathbf{v}}\right)\text{vec}\left(\Phi_{\mathbf{v}}\right)}, \quad (27)$$

and $\mathbf{I}_{N^2}$ is the identity matrix of size $N^2 \times N^2$.

Next, we wish to minimize the distortion, i.e., find $\widetilde{\mathbf{h}}_{\max,2}$ such that:

$$\widetilde{\mathbf{h}}^H_{\max,2}\text{vec}\left(\Phi_{\mathbf{x}}\right) = \phi_X. \quad (28)$$

Substituting (26) into (28), we have:

$$\widetilde{\mathbf{h}}'_{\max,2} = \frac{\mathbf{P}\text{vec}\left(\Phi_{\mathbf{x}}\right)\phi_X}{\text{vec}^H\left(\Phi_{\mathbf{x}}\right)\mathbf{P}\text{vec}\left(\Phi_{\mathbf{x}}\right)}. \quad (29)$$

Since $\mathbf{P}^2 = \mathbf{P}$, we have:

$$\widetilde{\mathbf{h}}_{\max,2} = \widetilde{\mathbf{h}}'_{\max,2}. \quad (30)$$

Finally, by observing that $\mathbf{P} = \mathbf{U}'\mathbf{U}'^H$, we deduce that:

$$\widetilde{\mathbf{h}}_{\max} = \widetilde{\mathbf{h}}_{\max,2}. \tag{31}$$

It should be noted that the formulation of (9) was already suggested in [20] in the context of multichannel noise reduction in the frequency domain. However, in this work, the quadratic approach is applied to a single-channel observation vector in an arbitrary linear filtering domain, in which the interframe correlation is considered. Additionally, while the optimal filters suggested in [20] are designed to minimize the squared output energy and may be seen as the quadratic approach counterparts of the conventional MVDR and LCMV, this work provides a more general perspective to derive quadratic filters and proposes the quadratic maximum SNR filter $\widetilde{\mathbf{h}}_{\max}$ as a special case.

5 Performance analysis

In this section, we analyze a toy example for which we derive the linear and quadratic maximum SNR filters. We theoretically evaluate and compare their corresponding subband SNR gains.

From Section 4, the theoretical subband SNR gain with the quadratic maximum SNR filter may be potentially unbounded. However, this would only be possible when the noise PSD matrix is precisely known. Since this assumption is never true in practice, it is important to analyze robustness to estimation errors in order to determine how practical the quadratic approach may be. Thus, our objective in this section is to evaluate the performance of the quadratic maximum SNR filter in the presence of estimation errors and compare it to the linear maximum SNR filter. This is done through a theoretical analysis of the following toy example in the STFT domain. Let us begin by assuming that the background noise is white and Gaussian, i.e., $v(t) \sim \mathcal{N}\left(0, \sigma_v^2\right)$. It can be shown that in the STFT domain with 50% overlapping rectangular analysis windows, the correlation matrix of the $N = 2$ element noise vector:

$$\mathbf{v}(k,n) = \left[\, V(k,n) \;\; V(k,n-1) \,\right]^T, \tag{32}$$

is given by:

$$\Phi_{\mathbf{v}} = \frac{N_{\text{FFT}}\,\sigma_v^2}{2}\begin{bmatrix} 2 & (-1)^k \\ (-1)^k & 2 \end{bmatrix}, \tag{33}$$

where N_{FFT} is the number of FFT bins in a single frame. Next, we model the noise PSD matrix estimation errors as independent centralized complex Gaussian variables $\epsilon_{ij}, 1 \le i,j \le 2$, whose variance is denoted by σ_ϵ^2. Additionally, we use the notation $\sigma_V = N_{\text{FFT}}\,\sigma_v^2/2$. Thus, the noise PSD matrix estimate with errors is given by:

$$\Phi_{\mathbf{v},\epsilon} = \sigma_V \begin{bmatrix} 2 & (-1)^k \\ (-1)^k & 2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} & \epsilon_{12} \\ \epsilon_{21} & \epsilon_{22} \end{bmatrix}. \tag{34}$$

In order to derive the optimal filters, we also require the PSD matrix of the desired signal. Since our goal is to analyze the effect of the noise PSD matrix estimation errors, we assume for simplicity a fully coherent desired signal, that is:

$$\Phi_{\mathbf{x}} = \phi_X \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}. \tag{35}$$

The first step in deriving the quadratic maximum SNR filter $\widetilde{\mathbf{h}}_{\max}$ involves calculation of the projection operator. Following the simplified notation, we have:

$$\mathbf{P} = \mathbf{I}_4 - \frac{\text{vec}\left(\Phi_{\mathbf{v},\epsilon}\right)\text{vec}^H\left(\Phi_{\mathbf{v},\epsilon}\right)}{\text{vec}^H\left(\Phi_{\mathbf{v},\epsilon}\right)\text{vec}\left(\Phi_{\mathbf{v},\epsilon}\right)}, \tag{36}$$

in which the matrix $\text{vec}\left(\Phi_{\mathbf{v},\epsilon}\right)\text{vec}^H\left(\Phi_{\mathbf{v},\epsilon}\right)$ and the scalar $\text{vec}^H\left(\Phi_{\mathbf{v},\epsilon}\right)\text{vec}\left(\Phi_{\mathbf{v},\epsilon}\right)$ should be computed. We have:

$$\begin{aligned} &\text{vec}^H\left(\Phi_{\mathbf{v},\epsilon}\right)\text{vec}\left(\Phi_{\mathbf{v},\epsilon}\right) = 10\sigma_V^2 \\ &+ 4\sigma_V \Re\{\epsilon_{11}+\epsilon_{22}\} + 2(-1)^k\sigma_V\Re\{\epsilon_{12}+\epsilon_{21}\} \\ &+ |\epsilon_{11}|^2 + |\epsilon_{12}|^2 + |\epsilon_{21}|^2 + |\epsilon_{22}|^2, \end{aligned} \tag{37}$$

where $|\epsilon_{ij}|^2, 1 \le i,j \le 2$ are independent exponentially distributed random variables, that is, $|\epsilon_{ij}|^2 \sim \exp\left(1/2\sigma_\epsilon^2\right)$.

Next, we compute the elements of the 4×4 matrix $\text{vec}\left(\Phi_{\mathbf{v},\epsilon}\right)\text{vec}^H\left(\Phi_{\mathbf{v},\epsilon}\right)$, by which we may approximate the expected value of $\mathbf{P}$, a key value required to approximate the theoretical subband SNR gain. We have:

$$\begin{aligned} E(\mathbf{P}) &= \mathbf{I}_4 - E\left[\frac{\text{vec}\left(\Phi_{\mathbf{v},\epsilon}\right)\text{vec}^H\left(\Phi_{\mathbf{v},\epsilon}\right)}{\text{vec}\left(\Phi_{\mathbf{v},\epsilon}\right)^H\text{vec}\left(\Phi_{\mathbf{v},\epsilon}\right)}\right] \\ &\approx \mathbf{I}_4 - \frac{E\left[\text{vec}\left(\Phi_{\mathbf{v},\epsilon}\right)\text{vec}^H\left(\Phi_{\mathbf{v},\epsilon}\right)\right]}{E\left[\text{vec}^H\left(\Phi_{\mathbf{v},\epsilon}\right)\text{vec}\left(\Phi_{\mathbf{v},\epsilon}\right)\right]}, \end{aligned} \tag{38}$$

where we used a first-order approximation [26]. Defining the error-to-noise ratio (ENR):

$$R_\epsilon = \frac{\sigma_\epsilon^2}{\sigma_V^2}, \tag{39}$$

we obtain:

$$\begin{aligned} E(\mathbf{P}) &\approx \frac{1}{2\left(4R_\epsilon+5\right)} \\ &\times \begin{bmatrix} 6\left(R_\epsilon+1\right) & 2(-1)^{k+1} & 2(-1)^{k+1} & -4 \\ 2(-1)^{k+1} & 6R_\epsilon+9 & -1 & 2(-1)^{k+1} \\ 2(-1)^{k+1} & -1 & 6R_\epsilon+9 & 2(-1)^{k+1} \\ -4 & 2(-1)^{k+1} & 2(-1)^{k+1} & 6\left(R_\epsilon+1\right) \end{bmatrix}. \end{aligned} \tag{40}$$

Rewriting (15) to calculate the PSD of the estimated desired signal with the random filter $\widetilde{\mathbf{h}}$, we have

$$\begin{aligned}\phi_{\widehat{X}} &= E\left(|Z|\right) \\ &\approx |E\left(Z\right)| \\ &= \left|E\left[E\left[\widetilde{\mathbf{h}}^H\widetilde{\mathbf{y}}|\{\epsilon_{ij}\}\right]\right]\right| \\ &= \left|E\left(\widetilde{\mathbf{h}}^H\right)E\left(\widetilde{\mathbf{y}}\right)\right| \\ &= \left|E\left(\widetilde{\mathbf{h}}^H\right)\mathrm{vec}\left(\Phi_{\mathbf{x}}\right)+E\left(\widetilde{\mathbf{h}}^H\right)\mathrm{vec}\left(\Phi_{\mathbf{v}}\right)\right|,\end{aligned} \tag{41}$$

which implies that the subband output SNR is:

$$\mathrm{oSNR}\left(\widetilde{\mathbf{h}}\right)=\frac{\left|E\left(\widetilde{\mathbf{h}}^H\right)\mathrm{vec}\left(\Phi_{\mathbf{x}}\right)\right|}{\left|E\left(\widetilde{\mathbf{h}}^H\right)\mathrm{vec}\left(\Phi_{\mathbf{v}}\right)\right|}, \tag{42}$$

and its corresponding subband SNR gain is:

$$\mathcal{G}\left(\widetilde{\mathbf{h}}\right)=\frac{\left|E\left(\widetilde{\mathbf{h}}^H\right)\mathrm{vec}\left(\Phi_{\mathbf{x}}\right)\right|}{\left|E\left(\widetilde{\mathbf{h}}^H\right)\mathrm{vec}\left(\Phi_{\mathbf{v}}\right)\right|}\times\frac{\phi_V}{\phi_X}. \tag{43}$$

Thus, in order to evaluate the subband SNR gain, we must first compute the expected value of the random filter $\widetilde{\mathbf{h}}_{\max}$. We have:

$$\begin{aligned}E\left(\widetilde{\mathbf{h}}_{\max}\right) &= E\left[\frac{\mathbf{P}\mathrm{vec}\left(\Phi_{\mathbf{x}}\right)\phi_X}{\mathrm{vec}^H\left(\Phi_{\mathbf{x}}\right)\mathbf{P}\mathrm{vec}\left(\Phi_{\mathbf{x}}\right)}\right] \\ &\approx \frac{E\left[\mathbf{P}\mathrm{vec}\left(\Phi_{\mathbf{x}}\right)\phi_X\right]}{E\left[\mathrm{vec}^H\left(\Phi_{\mathbf{x}}\right)\mathbf{P}\mathrm{vec}\left(\Phi_{\mathbf{x}}\right)\right]} \\ &= \frac{E\left(\mathbf{P}\right)\mathrm{vec}\left(\Phi_{\mathbf{x}}\right)\phi_X}{\mathrm{vec}^H\left(\Phi_{\mathbf{x}}\right)E\left(\mathbf{P}\right)\mathrm{vec}\left(\Phi_{\mathbf{x}}\right)},\end{aligned} \tag{44}$$

where we used a first-order approximation in the second row of (44). Substituting (35) and (44) into (43), the subband SNR gain reduces to:

$$\begin{aligned}&\mathcal{G}\left(\widetilde{\mathbf{h}}_{\max}\right) \\ &= \frac{\mathrm{vec}^H\left(\Phi_{\mathbf{x}}\right)E\left(\mathbf{P}\right)\mathrm{vec}\left(\Phi_{\mathbf{x}}\right)}{\mathrm{vec}^H\left(\Phi_{\mathbf{x}}\right)E\left(\mathbf{P}\right)\mathrm{vec}\left(\Phi_{\mathbf{v}}\right)} \\ &= \frac{1}{R_\epsilon}\frac{2\left[4(-1)^{k+1}+5\right]}{3\left[2+(-1)^k\right]}+\frac{4}{2+(-1)^k}+O\left(\frac{1}{R_\epsilon^2}\right).\end{aligned} \tag{45}$$

We deduce that when the ENR approaches zero, the theoretical subband SNR gain goes to infinity, and when the ENR is large, the subband SNR gain is finite and frequency dependent.

The derivation of the linear filter $\mathbf{h}_{\max}$ of [19], which is used as a baseline for performance evaluation, begins by assessing the eigenvector corresponding to the maximum eigenvalue of the matrix $\Phi_{\mathbf{v},\epsilon}^{-1}\Phi_{\mathbf{x}}$. We have:

$$\begin{aligned}\Phi_{\mathbf{v},\epsilon}^{-1}\Phi_{\mathbf{x}} &= \frac{\phi_X}{\left|\Phi_{\mathbf{v},\epsilon}\right|} \\ &\times\begin{bmatrix}\left[2+(-1)^{k+1}\right]\sigma_V+\epsilon_{22}-\epsilon_{12}, & \left[2+(-1)^{k+1}\right]\sigma_V+\epsilon_{22}-\epsilon_{12} \\ \left[2+(-1)^{k+1}\right]\sigma_V+\epsilon_{11}-\epsilon_{21}, & \left[2+(-1)^{k+1}\right]\sigma_V+\epsilon_{11}-\epsilon_{21}\end{bmatrix},\end{aligned} \tag{46}$$

whose eigenvalues are:

$$\lambda_{\min}=0, \tag{47}$$

$$\begin{aligned}\lambda_{\max} &= \frac{\phi_X}{\left|\Phi_{\mathbf{v},\epsilon}\right|} \\ &\times\left[2\sigma_V\left[2+(-1)^{k+1}\right]+\epsilon_{11}+\epsilon_{22}-\epsilon_{12}-\epsilon_{21}\right].\end{aligned} \tag{48}$$

It is easily verified that the (unnormalized) eigenvector $\mathbf{b}_{\max}$ that corresponds to $\lambda_{\max}$ is given by:

$$\mathbf{b}_{\max}=\begin{bmatrix}\left[2+(-1)^{k+1}\right]\sigma_V+\epsilon_{22}-\epsilon_{12} \\ \left[2+(-1)^{k+1}\right]\sigma_V+\epsilon_{11}-\epsilon_{21}\end{bmatrix}, \tag{49}$$

which implies that:

$$\begin{aligned}&E\left(\mathbf{b}_{\max}\mathbf{b}_{\max}^H\right) \\ &=\begin{bmatrix}\left[2+(-1)^{k+1}\right]^2\sigma_V^2+4\sigma_\epsilon^2 & \left[2+(-1)^{k+1}\right]^2\sigma_V^2 \\ \left[2+(-1)^{k+1}\right]^2\sigma_V^2 & \left[2+(-1)^{k+1}\right]^2\sigma_V^2+4\sigma_\epsilon^2\end{bmatrix}.\end{aligned} \tag{50}$$

Formulating the PSD expression of the estimated desired signal with the random linear filter $\mathbf{h}_{\max}$ in a similar manner to (41), its subband SNR gain is:

$$\mathcal{G}\left(\mathbf{h}_{\max}\right)=\frac{\phi_V}{\phi_X}\times\frac{E\left(\mathbf{h}_{\max}^H\right)\Phi_{\mathbf{x}}E\left(\mathbf{h}_{\max}\right)}{E\left(\mathbf{h}_{\max}^H\right)\Phi_{\mathbf{v}}E\left(\mathbf{h}_{\max}\right)}, \tag{51}$$

where the expected value of $\mathbf{h}_{\max}$ is given by:

$$\begin{aligned}E\left(\mathbf{h}_{\max}\right) &= E\left(\frac{\mathbf{b}_{\max}\mathbf{b}_{\max}^H\Phi_{\mathbf{x}}\mathbf{i}_1}{\mathbf{b}_{\max}^H\Phi_{\mathbf{x}}\mathbf{b}_{\max}}\right) \\ &\approx \frac{E\left(\mathbf{b}_{\max}\mathbf{b}_{\max}^H\Phi_{\mathbf{x}}\mathbf{i}_1\right)}{E\left(\mathbf{b}_{\max}^H\Phi_{\mathbf{x}}\mathbf{b}_{\max}\right)} \\ &= \frac{E\left(\mathbf{b}_{\max}\mathbf{b}_{\max}^H\right)\Phi_{\mathbf{x}}\mathbf{i}_1}{E\left(\mathbf{b}_{\max}^H\right)\Phi_{\mathbf{x}}E\left(\mathbf{b}_{\max}\right)} \\ &= [0.5\ 0.5]^T,\end{aligned} \tag{52}$$

where we used a first-order approximation in the second row of (52). Substituting (35) and (52) into (51), the subband SNR gain is finally:

$$\mathcal{G}\left(\mathbf{h}_{\max}\right)=\frac{4}{2+(-1)^k}, \tag{53}$$

which is ENR independent, but frequency dependent.

We infer that when the ENR is low, i.e., when the relative noise PSD estimation error is negligible, the quadratic approach achieves a highly preferable subband SNR gain. However, when the estimation error is in the same order of the noise energy, the two approaches exhibit a similar subband SNR gain. To illustrate the latter result, we return to

(40) in the limit of ENR that approaches infinity. We have:

$$\lim_{R_\epsilon \to \infty} E\left(\mathbf{P}\right) \propto \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{54}$$

and hence:

$$\lim_{R_\epsilon \to \infty} E\left(\widetilde{\mathbf{h}}_{\max}\right) = \left[0.25\ 0.25\ 0.25\ 0.25\right]^T. \tag{55}$$

This implies that in the high ENR limit, the quadratic max SNR filter converges to a version of the linear max SNR filter of (52), in which case both filters are simple averaging filters. While this result is explicitly derived for the toy example, we would expect such a behavior in any high ENR scenario in which the errors are modeled as normal identically distributed independent random variables. Additionally, we have:

$$\lim_{R_\epsilon \to \infty} E\left(\widetilde{\mathbf{h}}_{\max}\right) = E\left(\mathbf{h}_{\max}\right) \otimes E\left(\mathbf{h}_{\max}\right), \tag{56}$$

which, by recalling (10) and the elaboration underneath, explains why in this limit the subband SNR gains are identical. The theoretical gain plots for odd and even values of k as a function of the ENR are illustrated in Fig. 1.

We end this part by addressing the computational complexity issue. On top of the additional complexity required with the quadratic maximum SNR filter in order to generate a desired signal phase estimate, the computational costs of the two filters are not straightforward to theoretically compare. That is, while deriving the quadratic maximum SNR filter typically requires matrix multiplications of a squared dimension, with the linear maximum SNR filter derivation, a matrix inversion and an eigenvalue decomposition are computed. In practice, running the toy example with MATLAB software on an ordinary CPU takes 13 msec with the linear maximum SNR filter and 22 msec with the quadratic maximum SNR filter. Increasing the observation signal vector length to $N = 7$ yields a total runtime of 15 msec with the linear maximum SNR filter and 27 msec with the quadratic maximum SNR filter. Combining the runtime of both filters, we deduce that with a serial processor, the quadratic maximum SNR filter requires about a three-time longer runtime than the linear maximum SNR filter in order to yield a desired signal amplitude and phase estimates.

6 Experimental results

In this section, we demonstrate the noise reduction capabilities of the quadratic maximum SNR filter in the context of speech enhancement. We perform extensive experiments in ideal and practical conditions, and compare its performance to well-known speech enhancement methods in stationary and nonstationary noise environments.

In the rest of the paper, for the sake of clarity, we return to explicit time and frequency indices notation.

6.1 Simulations in ideal conditions

We have shown that in the lack of estimation errors, the quadratic filter $\widetilde{\mathbf{h}}_{\max}(k,n)$ is designed to eliminate the residual noise, provided it is applied to the vector form of the additive noise correlation matrix. However, in practice, noise reduction filters are usually applied to instantaneous observation signal vectors, in which the noise term is of the form $\mathbf{v}^*(k,n) \otimes \mathbf{v}(k,n)$. Indeed, the latter may significantly differ from the statistical noise correlation matrix, which implies that the noise reduction performance might be far from optimal. It is therefore beneficial to employ a preliminary temporal smoothing step to the observation signal vector and then apply the quadratic filtering approach to a time-smoothed vector. Define:

$$\begin{aligned} \widetilde{\mathbf{y}}_a(k,n;\tau_{\mathrm{y}}) &= \frac{1}{2\tau_{\mathrm{y}}+1} \sum_{n'=-\tau_{\mathrm{y}}}^{\tau_{\mathrm{y}}} \widetilde{\mathbf{y}}(k,n+n') \\ &= \frac{1}{2\tau_{\mathrm{y}}+1} \sum_{n'=-\tau_{\mathrm{y}}}^{\tau_{\mathrm{y}}} \mathbf{y}^*(k,n+n') \otimes \mathbf{y}(k,n+n') \\ &= \widetilde{\mathbf{x}}_a(k,n;\tau_{\mathrm{y}}) + \widetilde{\mathbf{v}}_a(k,n;\tau_{\mathrm{y}}) \\ &\quad + \frac{1}{2\tau_{\mathrm{y}}+1} \sum_{n'=-\tau_{\mathrm{y}}}^{\tau_{\mathrm{y}}} \{\mathbf{x}^*(k,n+n') \otimes \mathbf{v}(k,n+n') \\ &\qquad + \mathbf{v}^*(k,n+n') \otimes \mathbf{x}(k,n+n')\}, \end{aligned} \tag{57}$$

where:

$$\begin{aligned} &\widetilde{\mathbf{x}}_a(k,n;\tau_{\mathrm{y}}) \\ &= \frac{1}{2\tau_{\mathrm{y}}+1} \sum_{n'=-\tau_{\mathrm{y}}}^{\tau_{\mathrm{y}}} \mathbf{x}^*(k,n+n') \otimes \mathbf{x}(k,n+n'), \end{aligned} \tag{58}$$

$$\begin{aligned} &\widetilde{\mathbf{v}}_a(k,n;\tau_{\mathrm{y}}) \\ &= \frac{1}{2\tau_{\mathrm{y}}+1} \sum_{n'=-\tau_{\mathrm{y}}}^{\tau_{\mathrm{y}}} \mathbf{v}^*(k,n+n') \otimes \mathbf{v}(k,n+n'), \end{aligned} \tag{59}$$

and τ_{y} is the temporal smoothing preprocessing step parameter. We note that this implies a minor algorithmic delay of τ_{y} frames. Clearly, when the desired signal and noise are stationary and ergodic, we should choose a high value for τ_{y}, as:

$$E\left[\widetilde{\mathbf{y}}_a(k,n;\tau_{\mathrm{y}})\right] = E\left[\widetilde{\mathbf{y}}(k,n)\right], \tag{60}$$

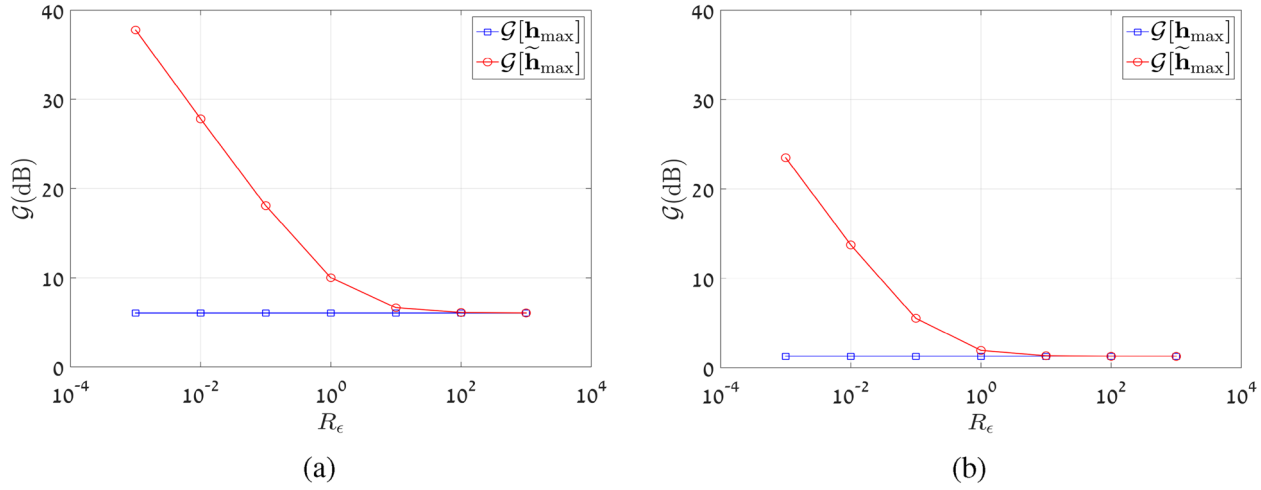


Fig. 1 Theoretical subband SNR gain with the linear and quadratic maximum SNR filters as a function of the ENR for **a** odd k and **b** even k

meaning that the temporal smoothing step does not distort the desired signal in terms of its second-order statistics. On the contrary, we have:

$$E\left[\widetilde{\mathbf{v}}_a^i(k,n;\tau_{\mathrm{y}}) - \mathrm{vec}^i\left[\Phi_{\mathbf{v}}(k,n)\right]\right]^2 < E\left[\widetilde{\mathbf{v}}^i(k,n) - \mathrm{vec}^i\left[\Phi_{\mathbf{v}}(k,n)\right]\right]^2, \tag{61}$$

for every vector element $1 \leq i \leq N^2$, meaning the time-smoothed version of the noise observations' vector better resembles the theoretical noise PSD statistics than its instantaneous version. In addition, with the left-hand side of (61) being a monotonically decreasing function of τ_{y}, we have:

$$\begin{aligned}\lim_{\tau_{\mathrm{y}}\to\infty}\left\{\frac{\left|\widetilde{\mathbf{h}}_{\max}^H(k,n)\widetilde{\mathbf{x}}_a(k,n;\tau_{\mathrm{y}})\right|}{\left|\widetilde{\mathbf{h}}_{\max}^H(k,n)\widetilde{\mathbf{v}}_a(k,n;\tau_{\mathrm{y}})\right|}\right\} &= \frac{\left|\widetilde{\mathbf{h}}_{\max}^H(k,n)\mathrm{vec}[\Phi_{\mathbf{x}}(k,n)]\right|}{\left|\widetilde{\mathbf{h}}_{\max}^H(k,n)\mathrm{vec}[\Phi_{\mathbf{v}}(k,n)]\right|} \\ &= \mathrm{oSNR}\left[\widetilde{\mathbf{h}}_{\max}(k,n)\right],\end{aligned} \tag{62}$$

which was previously shown to be potentially unbounded. On the contrary, for nonstationary desired signals, there is an inherent trade-off in setting τ_{y}: as τ_{y} increases the mean-squared estimation error of the left-hand side of (61) decreases, resulting in a lower residual noise. However, by further increasing τ_{y}, the equality in (60) does not hold as the non stationary desired signal is smeared over time and hence distorted.

In order to demonstrate this trade-off, we consider a clean speech signal $x(t)$ that is sampled at a sampling rate of $f_s = 1/T_s = 16$ kHz within the signal duration T. The desired speech signal is formed by concatenating 24 speech signals (12 speech signals per gender) with varying dialects that are taken from the TIMIT database [27]. The clean speech signal is corrupted by an uncorrelated white Gaussian additive noise $v(t)$. The noisy signal is transformed into the STFT domain using 50% overlapping time frames and a Hamming analysis window of length 256 (16 msec). Next, it undergoes the foregoing temporal smoothing step, and then filtered by the two maximum SNR filters, i.e., the quadratic $\widetilde{\mathbf{h}}_{\max}(k,n)$ and the linear $\mathbf{h}_{\max}(k,n)$ of [19] to generate estimates of the desired speech signal. It is important to mention that both filters use the exact same desired speech and noise signal statistics estimates. As in this part we assume ideal conditions in which the desired speech and noise signals are known, their statistics are calculated by smoothing the corresponding signals over time. We want to compare the two approaches fairly. Hence, we allow a temporal smoothing preprocessing step for the conventional filter as well. However, we note that while with the quadratic filter $\widetilde{\mathbf{h}}_{\max}(k,n)$ the temporal smoothing step is employed over $\widetilde{\mathbf{y}}(k,n)$, with the linear $\mathbf{h}_{\max}(k,n)$ the smoothing is employed over $\mathbf{y}(k,n)$.

There is another modification that should be made with the quadratic approach in order to obtain a reliable desired signal estimation and keep the desired signal variance expression in (15) valid. While it is easy to show that with $\widetilde{\mathbf{h}}_{\max}(k,n)$ the expression in (10) is real, there is no guarantee that it is strictly positive. In practice, when a desired speech signal is present, it is very likely that the inner product is indeed positive, hence yielding a valid estimate of the desired signal spectral energy. This may be seen by applying the quadratic filter to the last equality of (12) in which the first term, that is associated with the true desired signal energy and the positive inter-frame correlation of adjacent time-frequency speech bins, is likely to be positive. Nevertheless, when a desired signal is absent, this positive term is approximately zero and

the energy estimate may turn out negative. Clearly, such an estimate is non-physical and should be clipped to zero. Consequently, (10) is modified to:

$$Z(k,n) = \max\left\{\widetilde{\mathbf{h}}^{H}_{\max}(k,n)\widetilde{\mathbf{y}}(k,n), 0\right\}. \tag{63}$$

Once the noise reduction procedure is completed, an inverse STFT transform is applied to yield the enhanced signals in the time domain. Then, it is possible to compute the PESQ [28] and STOI [29] scores, which function as a complementary performance measure to the subband SNR gain. We employ these scores to demonstrate the aforementioned trade-off in setting τ_{y} by computing them from the time-domain enhanced signals with the two maximum SNR filters. This simulation is carried out multiple times with varying values of τ_{y} with $N = 3$ and for time-domain input SNRs of -5 dB and 15 dB, where the time-domain input SNR is defined by:

$$\mathrm{iSNR} = \frac{E\left[x^2(t)\right]}{E\left[v^2(t)\right]}. \tag{64}$$

The PESQ and STOI scores of the enhanced signals are shown in Fig. 2. We note that in this part, the desired signal and noise are assumed to be known and are used to respectively generate their estimated statistics by performing a straightforward temporal smoothing. To begin with, it is clear that with the linear $\mathbf{h}_{\max}(k,n)$ for both time-domain input SNRs, the optimal τ_{y} is zero. This is not surprising, of course, as the time-smoothed version of $\mathbf{y}(k,n)$ converges to zero according to the signal model assumption. On the contrary, while for the high input SNR a small value of τ_{y} should be used with $\widetilde{\mathbf{h}}_{\max}(k,n)$ (as the noise is very weak and the optimal filter should resemble

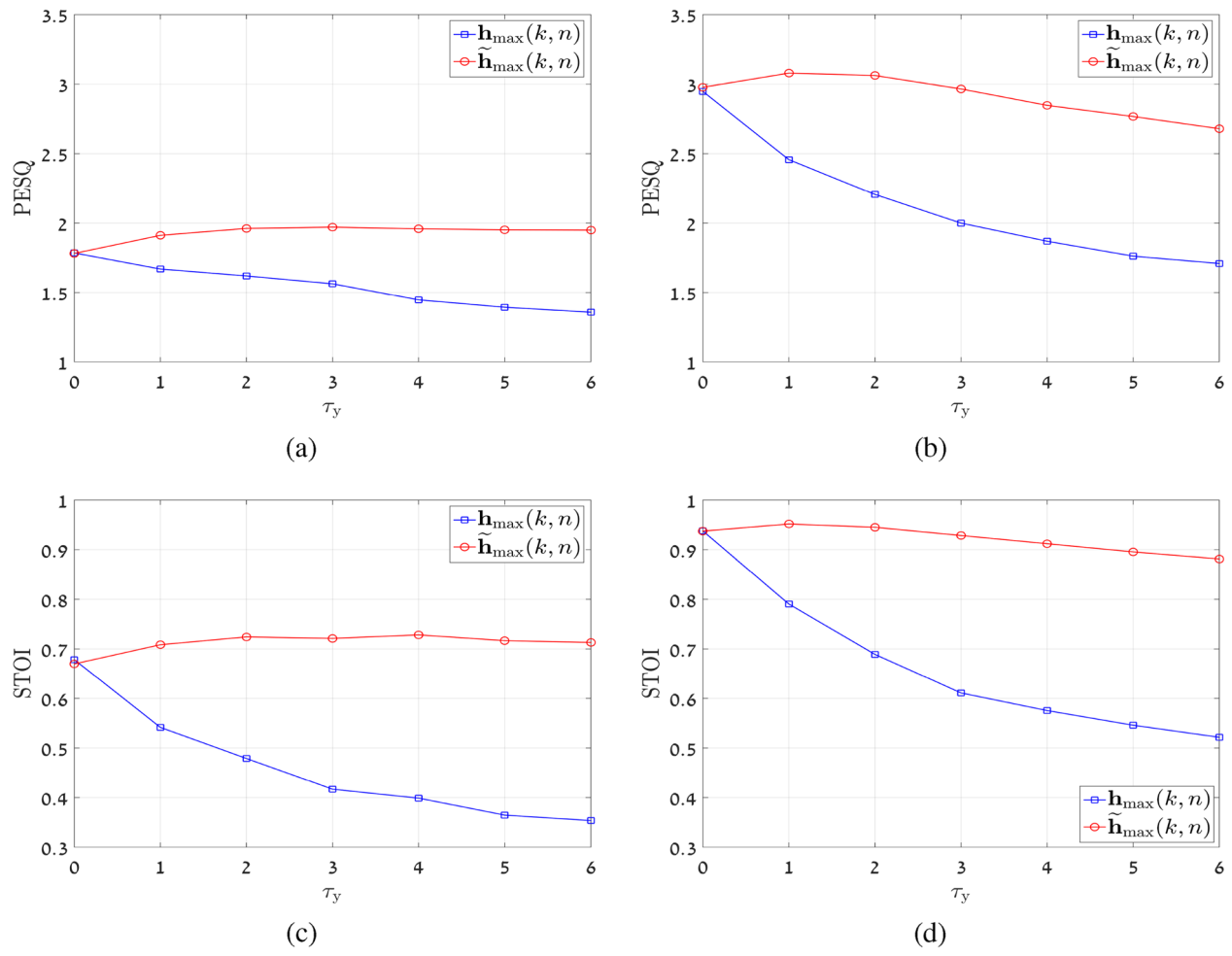


Fig. 2 PESQ and STOI scores of TIMIT speech signals as a function of the temporal smoothing preprocessing parameter τ_{y} for $N = 3$ in the presence of white Gaussian noise: **a** PESQ scores with iSNR $= -5$ dB, **b** PESQ scores with iSNR $= 15$ dB, **c** STOI scores with iSNR $= -5$ dB, and **d** STOI scores with iSNR $= 15$ dB. The PESQ scores of the input noisy observation signal are 1.47 and 2.78 with iSNR $= -5$ dB and iSNR $= 15$ dB, respectively; their corresponding STOI scores are 0.61 and 0.97

the identity filter), for a low input SNR, the convergence of the noise term $\widetilde{\mathbf{v}}_a(k,n;\tau_{\mathrm{y}})$ in $\widetilde{\mathbf{y}}_a(k,n;\tau_{\mathrm{y}})$ to the true noise correlation matrix is essential, and the optimal value of τ_{y} is found to be approximately 4. Clearly, when $\tau_{\mathrm{y}} \leq 4$, the approximation in (60) holds and the desired speech signal remains roughly distortionless. Thus, the mean-squared estimation error of the left-hand side of (61) decreases as τ_{y} increases. However, we observe that while further increasing τ_{y}, i.e., when $\tau_{\mathrm{y}} > 4$, reduces the mean-squared estimation error of the noise, it also distorts the desired speech signal. Consequently, we infer that τ_{y} should be set to a value ranging $1-4$, with 1 being optimal for very high input SNRs and 3 or 4 being optimal for low input SNRs.

Next, in Fig. 3, we investigate the PESQ and STOI scores as a function of the input SNR for $N = 3$ and $N = 7$. We note that as a compromise between high and low input SNRs, we fix $\tau_{\mathrm{y}} = 2$. We observe that in both cases, the quadratic maximum SNR filter is preferable, in particular in low input SNRs where the noise reduction capabilities are stressed. As the input SNR increases, the linear and quadratic filter performances converge. This is intuitively explained as in the limit of zero additive noise, the PESQ and STOI score improvements should converge to zero and both the linear and quadratic filters should converge to a version of the identity filter. Nevertheless, we exhibit a minor STOI score degradation in higher input SNRs. In essence, this is an artifact of the desired signal statistics estimation errors used to derive both the linear and the quadratic filters. That is, even with a stationary background noise, we expect estimation errors to emerge due to the highly nonstationary nature of the speech signals. The estimation errors inevitably result in some minor enhanced signal distortion which is more dominant in such scenarios. Finally, we note that the performance gap between the $N = 3$ and $N = 7$ cases, as exhibited in both filters, is a consequence of the stationary background noise. That is, we would not expect such a gap with an abruptly varying noise.

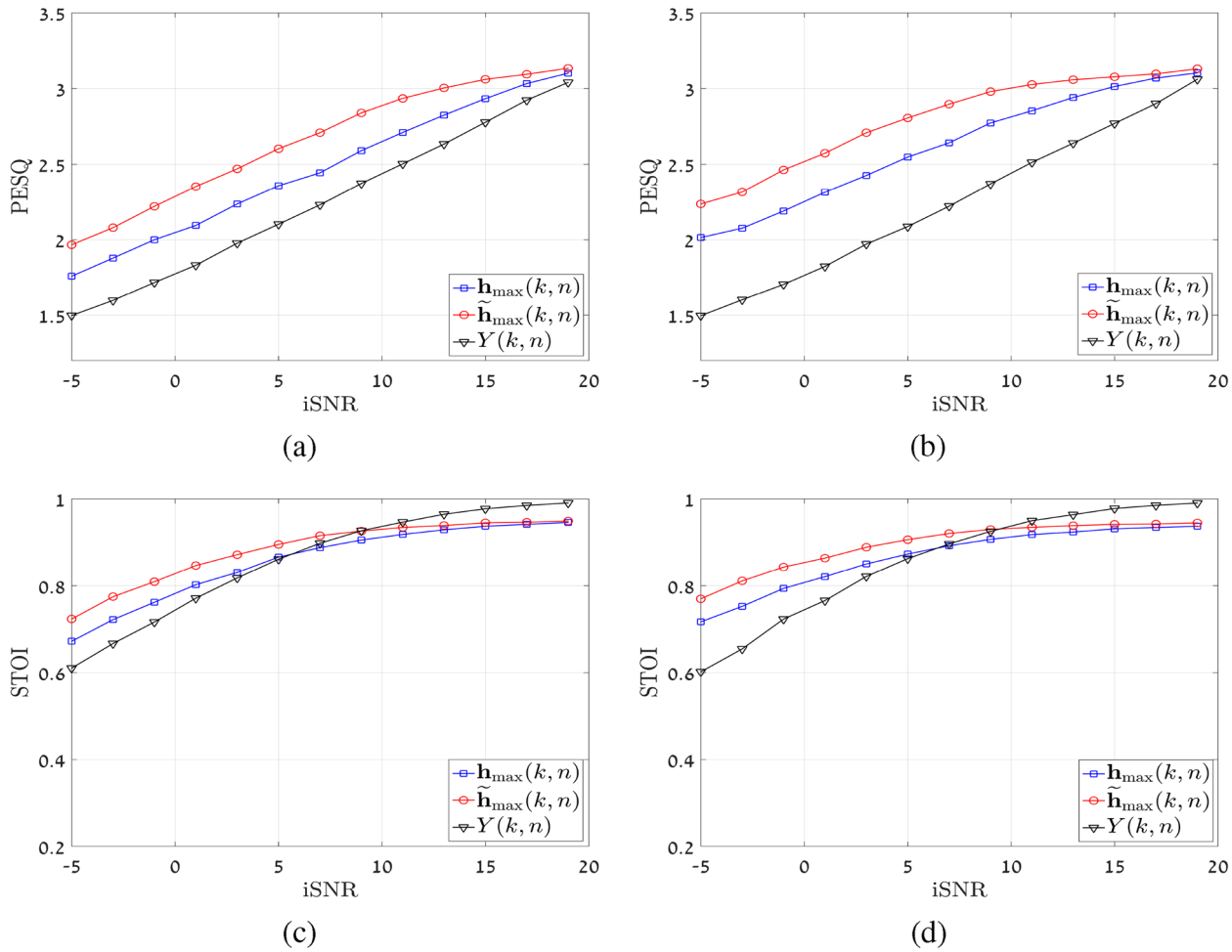


Fig. 3 PESQ and STOI scores of TIMIT speech signals as a function of the iSNR for $N = 3$ and $N = 7$ in the presence of white Gaussian noise. **a** PESQ scores with $N = 3$. **b** PESQ scores with $N = 7$. **c** STOI scores with $N = 3$. **d** STOI scores with $N = 7$. We set $\tau_{\mathrm{y}} = 2$ for the quadratic filter $\widetilde{\mathbf{h}}_{\max}(k,n)$

We return to the aforementioned subband SNR gain. In the STFT domain, it is convenient to average the subband input and output SNR expressions of (5), (8), and (16) over time, i.e.,

$$\overline{\mathrm{iSNR}}(k,:) = \frac{\sum_n \phi_X(k,n)}{\sum_n \phi_V(k,n)}, \tag{65}$$

$$\overline{\mathrm{oSNR}}\left[\mathbf{h}(k,:)\right] = \frac{\sum_n \mathbf{h}^H(k,n)\Phi_{\mathbf{x}}(k,n)\mathbf{h}(k,n)}{\sum_n \mathbf{h}^H(k,n)\Phi_{\mathbf{v}}(k,n)\mathbf{h}(k,n)}, \tag{66}$$

and

$$\overline{\mathrm{oSNR}}\left[\widetilde{\mathbf{h}}(k,:)\right] = \frac{\sum_n \left|\widetilde{\mathbf{h}}^H(k,n)\mathrm{vec}\left[\Phi_{\mathbf{x}}(k,n)\right]\right|}{\sum_n \left|\widetilde{\mathbf{h}}^H(k,n)\mathrm{vec}\left[\Phi_{\mathbf{v}}(k,n)\right]\right|}. \tag{67}$$

Consequently, the average subband SNR gains are given by:

$$\overline{\mathcal{G}}\left[\mathbf{h}(k,:)\right] = \frac{\overline{\mathrm{oSNR}}\left[\mathbf{h}(k,:)\right]}{\overline{\mathrm{iSNR}}(k,:)} \tag{68}$$

and

$$\overline{\mathcal{G}}\left[\widetilde{\mathbf{h}}(k,:)\right] = \frac{\overline{\mathrm{oSNR}}\left[\widetilde{\mathbf{h}}(k,:)\right]}{\overline{\mathrm{iSNR}}(k,:)}, \tag{69}$$

respectively.

We use expressions (68) and (69), respectively, to compare $\widetilde{\mathbf{h}}_{\max}(k,n)$ and $\mathbf{h}_{\max}(k,n)$ in terms of the average subband SNR gain. The results for iSNR $= 0$ dB and for $N = 3$ and 7 are depicted in Fig. 4. According to the analysis above, we set $\tau_{\mathrm{y}} = 2$ with the quadratic maximum SNR filter, which is shown to result in a significantly preferable gain. This is true for both values of N. Moreover, as it is observed in Fig. 4 and in a similar fashion to the previously discussed average PESQ and STOI scores, the performance of the linear maximum SNR filter with $N = 7$ is somewhat close to the performance of the quadratic maximum SNR filter with $N = 3$. That is, the quadratic filter is demonstrated to better utilize a given noisy observation signals vector from the subband SNR gain perspective.

6.2 Experiments in practical scenarios

Next, we are interested in comparing the two approaches in practical scenarios and with nonstationary noise. Four scenarios are simulated with the additive noise signal being either a stationary white Gaussian noise or one of the following three nonstationary noise types: a motor crank noise, a wind noise, or a traffic noise. The TIMIT set of clean desired speech signals is maintained. We set iSNR $= 0$ dB and analyze the PESQ and STOI scores with the following six methods: two practical versions of the linear and quadratic maximum SNR filters, their two ideal versions (as presented in the previous part), the celebrated log-spectral amplitude estimator (LSA) [3], and the spectral subtraction in the short-time modulation domain (STSS) of [30]. We set $N = 3$ for all four maximum SNR filters and perform the STFT transform with the same analysis window and overlap factor in all methods except the STSS. The STSS is employed in its default parameters as defined by the authors of [30], with acoustic and modulation frame lengths and overlap factors of 32 msec and 75%, and 256 msec and 87.5%, respectively. According to the previous part, we fix $\tau_{\mathrm{y}} = 2$ with $\widetilde{\mathbf{h}}_{\max}(k,n)$, whereas no smoothing is performed with $\mathbf{h}_{\max}(k,n)$.

The practical versions of the linear and quadratic maximum SNR filters, denoted, respectively, by $\mathbf{h}_{\max,\mathrm{prac}}(k,n)$ and $\widetilde{\mathbf{h}}_{\max,\mathrm{prac}}(k,n)$, require estimates of the desired speech and noise correlation matrices to be computed out of the noisy observations. In this experiment, we employ a somewhat naive estimation approach that is inspired by [31] and leave more sophisticated schemes for future research. The noisy observation correlation matrix is updated over time by a first-order recursive temporal smoothing:

$$\begin{aligned}\Phi_{\mathbf{y}}(k,n) &= \lambda\Phi_{\mathbf{y}}(k,n-1)\\ &\quad + (1-\lambda)\,\mathbf{y}(k,n)\mathbf{y}^H(k,n),\end{aligned} \tag{70}$$

with $0 < \lambda < 1$ being the smoothing parameter. We found $\lambda = 0.5$ to be an optimal choice to cope with both stationary and quickly-varying nonstationary noise. Then, the noise correlation matrix is given by:

$$\Phi_{\mathbf{v}}(k,n) = \min\{\Phi_{\mathbf{v}}(k,n-1), \Phi_{\mathbf{y}}(k,n)\}\,(1+\epsilon)\,, \tag{71}$$

with ϵ set to yield a power increase of 5 dB/s. Finally, the desired signal correlation matrix is estimated by

$$\Phi_{\mathbf{x}}(k,n) = \max\{\Phi_{\mathbf{y}}(k,n) - \Phi_{\mathbf{v}}(k,n), 0\}. \tag{72}$$

We note the following. To begin with, the minimum and maximum operations above are considered element-wise, whereas the first 100 frames are used to generate an initial noise correlation matrix estimate, i.e., the first 808 msec is assumed to be silent. In addition, we verify that $\Phi_{\mathbf{x}}(k,n)$ is obtained as a positive-definite matrix, which is the case in practically all the simulations we have performed. Finally, the presented correlation matrices' estimation approach requires setting the optimal values of additional parameters in a similar manner to traditional approaches as described in Section 1.

The experimental results in terms of the average PESQ and STOI scores with their respective confidence (standard deviation) intervals computed over 24 speech utterances are described in Fig. 5. To begin with, we observe that in terms of PESQ scores, the ideal quadratic maximum SNR filter performs significantly better than the other methods in the three nonstationary noise scenarios, whereas it is slightly inferior to the STSS in the white noise scenario. In addition, the ideal quadratic maximum SNR filter is highly superior in terms of STOI scores in all the examined scenarios. In particular, the ideal quadratic

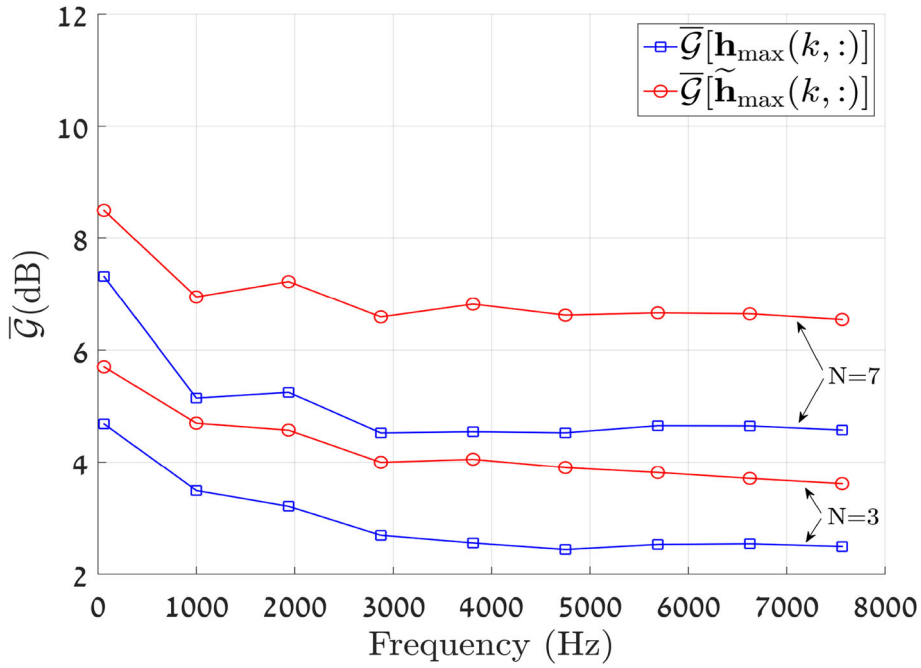


Fig. 4 Average subband SNR gains of TIMIT speech signals for iSNR = 0 dB and for $N = 3$ and $N = 7$, with white Gaussian noise. We set $\tau_y = 2$ for the quadratic filter $\widetilde{\mathbf{h}}_{\max}(k,n)$

maximum SNR filter outperforms its linear counterpart, which implies that the former's potential is preferable. Analyzing the practical versions of the maximum SNR filters, we note that in general, the quadratic filter is superior to the linear filter in terms of PESQ scors, whereas in terms of STOI scores, the performances are overall roughly equal. A comparison to the LSA and the STSS indicates that both are significantly inferior to the practical quadratic maximum SNR filter in the motor crank noise and wind noise scenarios. On the contrary, in the white noise and traffic noise scenarios, the performance gap is opposite, with the LSA and the STSS performing better than the practical quadratic maximum SNR filter, which is however preferable to the practical linear maximum SNR filter. The performance difference between noise types for the different methods is resulted in by the nature of the noise signals and the method we used to estimate and track their statistics. For example, this could be due to their level of nonstationarity, i.e., the coherence time during which the statistics of the noise remain roughly unchanged. We deduce that the quadratic maximum SNR filter is ideally of a high potential and may also be successfully applied in practice, even with naive desired signal and noise statistics estimation techniques.

We end this part by relating an informal listening experiment we conducted to verify the foregoing results. This included extensive comparisons between enhanced signals with all the presented methods in the different noise scenarios. While no musical noise nor reverberation effects were detected with any of the methods, their distinctive natures were observable. That is, while it was apparent that the four maximum SNR filters preserved the desired signals distortionless, the noise reduction capabilities of their two practical versions were relatively limited with respect to the LSA and STSS, which featured less residual noise in the white noise and traffic noise scenarios. On the contrary, the LSA and STSS did exhibit some desired signal distortion in most cases, particularly in frequencies higher than 3 kHz. This was more stressed in the motor crank noise and the wind noise scenarios, in which their respective residual noise was significant. Considering the ideal versions of the linear and quadratic maximum SNR filters, the enhanced signals they yielded sounded considerably clearer than all other methods, with the ideal quadratic maximum SNR filter being superior to its linear counterpart particularly in the white noise and the traffic noise scenarios.

7 Conclusions

We have presented a quadratic filtering approach for single-channel noise reduction, which generalizes the conventional linear filtering approach. The advantage of the quadratic approach was demonstrated by focusing on the maximum SNR filter in the STFT domain. We have analyzed the theoretical subband SNR gain in a toy example and showed that while with the linear maximum SNR

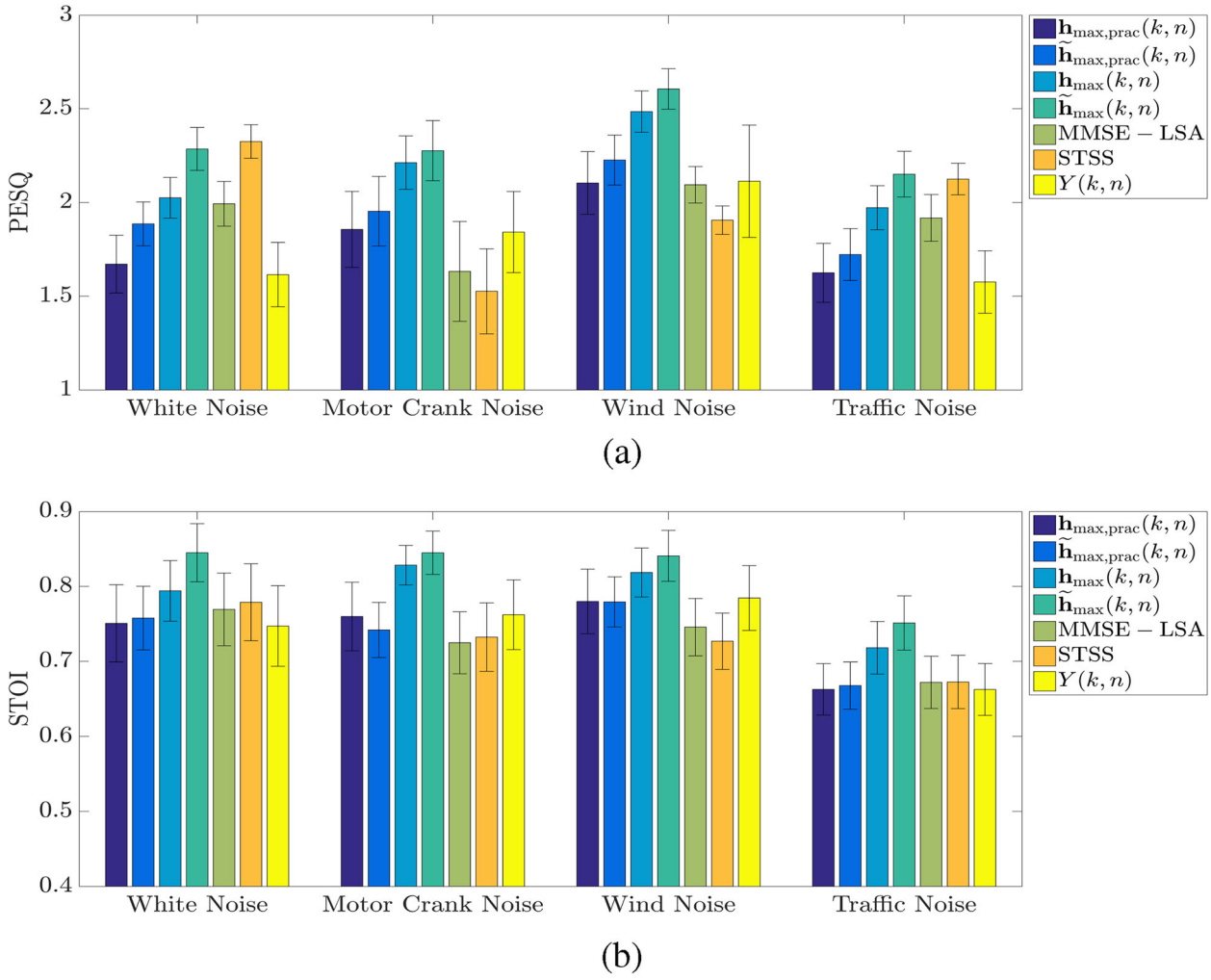


Fig. 5 A comparison of average PESQ and STOI scores of TIMIT speech signals in practical scenarios with iSNR = 0. The vertical margins exhibit the confidence (standard deviation) intervals computed over 24 speech utterances. $\mathbf{h}_{\max,\mathrm{prac}}(k,n)$ and $\widetilde{\mathbf{h}}_{\max,\mathrm{prac}}(k,n)$ denote, respectively, the linear and quadratic maximum SNR filters in practical conditions; $\mathbf{h}_{\max}(k,n)$ and $\widetilde{\mathbf{h}}_{\max}(k,n)$ denote, respectively, the linear and quadratic maximum SNR filters in ideal conditions; MMSE-LSA denotes the log-spectral amplitude estimator of [3]; STSS denotes the spectral subtraction in the short-time modulation domain method of [30]; $Y(k,n)$ denotes the input noisy observation: **a** PESQ scores and **b** STOI scores. We set $N = 3$ for all four maximum SNR filters and $\tau_y = 2$ for the two quadratic maximum SNR filters

filter, the subband SNR gain is strictly bounded, with the quadratic maximum SNR filter, the gain is potentially unbounded and heavily depends on the ENR. We have proposed the temporal smoothing preprocessing step and verified the performance on speech signals. In ideal and practical conditions, the quadratic maximum SNR filter was compared to the linear maximum SNR filter and to two well-known speech enhancement methods in both stationary and nonstationary noise environments. We have demonstrated that the quadratic maximum SNR filter outperforms the linear maximum SNR filter, in particular in low input SNRs, at the expanse of a higher computational complexity. In addition, the former was shown to perform better than commonly used methods in practice in some of the scenarios we examined, even with naive desired signal and noise statistics estimation techniques, whereas in other scenarios, the performance gap was the opposite. In future work, we may improve these estimation techniques to reach closer to the performance of the ideal quadratic maximum SNR filter, and possibly estimate the desired signal phase directly, i.e., not through a separate linear filter.

Abbreviations

SNR: Signal-to-noise ratio; PESQ: Perceptual evaluation of speech quality; SCNR: Single-channel noise reduction; STFT: Short-time Fourier transform; HMM: Hidden Markov model; MVDR: Minimum variance distortionless response; LCMV: Linearly constrained minimum variance

Acknowledgements
The authors thank the associate editor and the anonymous reviewers for their constructive comments and useful suggestions.

Authors' contributions
The authors' contributions are equal. The authors read and approved the final manuscript.

Funding
This research was supported by the Israel Science Foundation (grant no. 576/16) and the ISF-NSFC joint research program (grant no. 2514/17).

Availability of data and materials
Please contact the author for data requests.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Technion, Israel Institute of Technology, 32000, Haifa, Israel. [2]INRS-EMT, University of Quebec, 800 de la Gauchetiere Ouest, Suite 6900, Montreal, QC, H5A 1K6, Canada.

Received: 27 September 2019 Accepted: 20 March 2020
Published online: 15 April 2020

References

1. R. McAulay, M. Malpass, Speech enhancement using a soft-decision noise suppression filter. IEEE Trans. Acoust. Speech Sig. Proc. **28**(2), 137–145 (1980)
2. Y. Ephraim, D. Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Sig. Proc. **32**(6), 1109–1121 (1984)
3. Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Sig. Proc. **33**(2), 443–445 (1985)
4. I. Cohen, S. Gannot, *Spectral Enhancement Methods*. (J. Benesty, M. Sondhi, Y. Huang, eds.) (Springer, Berlin, Heidelberg, 2008), pp. 873–902
5. J. Benesty, J. Chen, E. Habets, *Speech Enhancement in the STFT Domain*. (Springer-Verlag Berlin Heidelberg, Berlin, 2012)
6. J. Benesty, I. Cohen, J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*. (Wiley-IEEE Press, Singapore, 2018)
7. I. Cohen, Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. IEEE Sig. Proc. Lett. **9**(4), 113–116 (2002)
8. P. J. Wolfe, J. S. Godsill, Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement. EURASIP J. Adv. Sig. Proc. **2003**(10) (2003)
9. R. Martin, Speech enhancement based on minimum mean-square error estimation and supergaussian priors. IEEE Trans. Speech Audio Proc. **13**, 845–856 (2005). https://doi.org/10.1109/TSA.2005.851927
10. I. Cohen, Speech enhancement using super-gaussian speech models and noncausal a priori snr estimation. Speech Commun. **47**(3), 336–350 (2005)
11. R. C. Hendriks, H. Richard, J. Jensen, *Log-Spectral Magnitude MMSE Estimators Under Super-Gaussian Densities*, (2009), pp. 1319–1322
12. R. Martin, in *Proceedings of the 27th IEEE International Conference Acoustics Speech Signal Processing, ICASSP-02*. Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors, vol. 1, (2002), pp. 253–256
13. J. S. Erkelens, R. C. Hendriks, R. Heusdens, J. Jensen, Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors. IEEE Trans. Audio Speech Lang. Proc. **15**(6), 1741–1752 (2007). https://doi.org/10.1109/TASL.2007.899233
14. R. R. Martin, C. Breithaupt, in *Proceedings of the 8th International Workshop on Acoustic Echo and Noise Control (IWAENC)*. Speech enhancement in the DFT domain using Laplacian speech priors, (2003), pp. 87–90
15. I. Cohen, Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models. Sig. Proc. **86**(4), 698–709 (2006)
16. I. Cohen, B. Berdugo, Speech enhancement for non-stationary noise environments. Sig. Proc. **81**(11), 2403–2418 (2001)
17. I. Cohen, Relaxed statistical model for speech enhancement and a priori SNR estimation. IEEE Trans. Speech Audio Proc. **13**(5), 870–881 (2005)
18. Y. A. Huang, J. Benesty, A multi-frame approach to the frequency-domain single-channel noise reduction problem. IEEE Trans Audio Speech Lang. Proc. **20**(4), 1256–1269 (2012)
19. G. Huang, J. Benesty, T. Long, J. Chen, A family of maximum SNR filters for noise reduction. IEEE/ACM Trans. Audio Speech Lang. Proc. **22**(12), 2034–2047 (2014)
20. G. Itzhak, J. Benesty, I. Cohen, Nonlinear kronecker product filtering for multichannel noise reduction. Speech Commun. **114**, 49–59 (2019)
21. P. C. Loizou, *Speech Enhancement: Theory and Practice, 2nd edn.* (CRC Press, Inc., Boca Raton, 2013)
22. J. Benesty, J. Chen, Y. Huang, I. Cohen, *Noise Reduction in Speech Processing, 1st edn.* (Springer-Verlag Berlin Heidelberg, Berlin, 2009)
23. J. Benesty, J. Chen, Y. A. Huang, Noise reduction algorithms in a generalized transform domain. IEEE Trans. Audio Speech Lang. Proc. **17**(6), 1109–1123 (2009)
24. D. A. Harville, *Matrix Algebra from a Statistician's Perspective, 1st edn.* (Springer-Verlag New York, New York, 1997)
25. G. H. Golub, C. F. V. Loan, *Matrix Computations, 3rd edn.* (Baltimore, Maryland: The Johns Hopkins University Press, Baltimore, 1996)
26. A. Stuart, K. Ord, *Kendall's Advanced Theory of Statistics, Volume1: Distribution Theory, 6th edn.* (Wiley, New York, 2010)
27. DARPA TIMIT acoustic phonetic continuous speech corpus CDROM. NIST (1993)
28. A. W. Rix, J. G. Beerends, M. P. Hollier, A. P. Hekstra, in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs, vol. 2, (2001), pp. 749–7522
29. C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Trans. Audio Speech. Lang. Proc. **19**(7), 2125–2136 (2011). https://doi.org/10.1109/TASL.2011.2114881
30. K. Paliwal, K. Wójcicki, B. Schwerin, Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. Speech Commun., 450–475 (2010)
31. A. Schasse, R. Martin, Estimation of subband speech correlations for noise reduction via MVDR processing. IEEE/ACM Trans. Speech Lang. Proc. **22**(9), 1355–1365 (2014)

Publisher's Note

Chapter 5

On the Design of Differential Kronecker-product Beamformers

On the Design of Differential Kronecker Product Beamformers

Gal Itzhak, Jacob Benesty, and Israel Cohen

***Abstract*—In this paper, we present a generalized approach for differential microphone array (DMA) beamforming in the short-time Fourier transform (STFT) domain. We propose a multistage beamforming approach, which considers a Kronecker product (KP) decomposition of the global beamformer into two independent sub-beamformers. We derive differential KP beamformers according to different criteria and analyze their performances, which are tuned by three design parameters. These parameters allow a high beamforming design flexibility; in particular, non-differential or non-KP beamformers may be obtained as special cases. Depending on the selection of parameters, we demonstrate a preferable performance with the new approach with respect to the white noise gain and directivity factor measures. In addition, we consider the task of speech enhancement. We show that differential KP beamformers perform better than non-differential and non-KP beamformers in terms of the quality and intelligibility of their respective time-domain enhanced signals, particularly in moderately reverberant environments.**

***Index Terms*—Microphone arrays, uniform linear arrays, differential beamforming, optimal beamformer, Kronecker product decomposition.**

I. Introduction

Communication systems in the real world involving audio and speech signals are typically required to operate in the presence of undesirable background noise that might heavily corrupt their quality. Taking advantage of a multichannel structure, the notion of beamforming has been in the center of attention to cope with this problem, being addressed in numerous researches [1]–[4]. Traced back to almost a century ago [5], [6], differential microphone arrays (DMAs) have been proposed and optimized with the underlying principle of exploiting acoustic pressure differences among adjacent microphones [7]–[11]. This principle implies arrays of small sizes and frequency-invariant beampatterns [12]–[16].

Typically, in order to design high-order DMAs which were capable of obtaining a significant amount of noise reduction, a multistage approach was taken. That is, the operation of differentiating acoustic pressure observations was successively repeated, in analogy to high-order derivatives of analytic functions [17], [18]. This approach was implemented in the time domain. Unfortunately, it was highly susceptible to array mismatches and imperfections [19]–[21], making it less appealing under practical conditions. Consequently, DMA design in the short-time Fourier transform (STFT) domain was introduced, providing a robust framework that is based on a single stage with linear matrix operations. Despite its simplicity, it is still capable of satisfying spatial constraints while simultaneously minimizing residual noise, in either fixed or adaptive settings [22], [23]. Nevertheless, due to the simple single-stage structure and inherent linear nature, the noise reduction capabilities of DMAs are limited. Recently, the design of DMAs in the STFT-domain was generalized to a multistage structure [24]. This technique was shown to be effective to reduce diffuse noise and to handle reverberant environments, though a significant drawback was its white noise amplification.

DMAs in the STFT domain were thoroughly analyzed and adapted into many variations. One recent example is Kronecker product (KP) beamforming, in which a global beamformer is decomposed into a KP of independent sub-beamformers that may be individually designed and optimized [25]–[30]. The main advantage of KP beamformers is their great design flexibility. That is, each sub-beamformer may be optimized by a different criterion, yielding a global beamformer that is "optimized" according to all criteria. The relative sizes of the sub-beamformers set the trade-off for the optimization of the global beamformer.

In this paper, we present a differential KP beamforming approach in the STFT domain, which generalizes the approach in [24]. We propose a multistage approach, which considers a KP decomposition of the global beamformer into two independent sub-beamformers. We derive differential KP beamformers according to different criteria and analyze their performances, tuned by three design parameters. These parameters facilitate high design flexibility; in particular, non-differential or non-KP beamformers may be obtained as special cases. Depending on the selection of parameters, we demonstrate a preferable performance using the new approach with respect to the white noise gain (WNG) and directivity factor (DF) measures. This may turn important in practice when considering microphones whose self-noise is significant or scenarios of considerable reverberation. In addition, we consider the task of speech enhancement. We show that differential KP beamformers perform better than non-differential and non-KP beamformers in terms of the quality and intelligibility of their respective time-domain enhanced signals, particularly in moderately reverberant environments.

The rest of the paper is organized as follows. In Section II, we briefly review the multistage differential beamforming approach in the STFT domain. In Section III, we present

This research was supported by the Pazy Research Foundation and the ISF-NSFC joint research program (grant No. 2514/17).

G. Itzhak and I. Cohen are with Andrew and Erna Viterby Faculty of Electrical Engineering, Technion–Israel Institute of Technology, Technion City, Haifa 3200003, Israel (e-mail: galitz@campus.technion.ac.il, icohen@ee.technion.ac.il).

J. Benesty is with INRS-EMT, University of Quebec, 800 de la Gauchetiere Ouest, Montreal, QC H5A 1K6, Canada (e-mail: Jacob.Benesty@inrs.ca).

the differential KP beamforming approach and reformulate the problem in terms of two independent sub-beamformers. Then, in Section IV, performance measures are derived accordingly. We present the familiar beamforming design measures as they are expressed in the context of differential KP beamforming. Section V is dedicated to the derivation of five differential KP beamformers, each of which is designed with respect to a different optimization criterion. Then, in Section VI, we perform a simulative study in an anechoic environment, followed by simulations of speech signals in a reverberant room as well as with array imperfections. Finally, we summarize this work in Section VII.

II. Signal Model and problem Formulation

We consider a uniform linear array (ULA), consisting of $M \geq 2$ omnidirectional microphones with an interelement spacing equal to δ. Let us assume that a farfield plane wave propagates from an azimuth angle θ in an anechoic acoustic environment at the speed of sound, i.e., $c = 340$ m/s, and impinges on this ULA. In this scenario, the corresponding steering vector (of length M) is [31]

$$\mathbf{d}_{\theta,M}(f) = \begin{bmatrix} 1 & e^{-\jmath 2\pi f \delta \cos\theta/c} & \cdots & e^{-\jmath(M-1)2\pi f\delta\cos\theta/c} \end{bmatrix}^T, \quad (1)$$

where f is the temporal frequency, $\jmath = \sqrt{-1}$ is the imaginary unit, and the superscript ^{T} is the transpose operator.

In order to be in the optimal working conditions of differential beamforming, we assume that the desired source comes from the direction $\theta_{\mathrm{s}} = 0$ and δ is small [13]. Note that in Section VI-C we demonstrate that our proposed approach is robust to small deviations in the values of θ_{s} and δ. In this case, we can express the frequency-domain observed signal vector of length M as [32]

$$\begin{aligned} \mathbf{y}(f) &= \begin{bmatrix} Y_1(f) & Y_2(f) & \cdots & Y_M(f) \end{bmatrix}^T \\ &= \mathbf{x}(f) + \mathbf{v}(f) \\ &= \mathbf{d}_{0,M}(f) X(f) + \mathbf{v}(f), \end{aligned} \quad (2)$$

where $Y_m(f)$ is the mth microphone signal, $\mathbf{x}(f) = \mathbf{d}_{0,M}(f) X(f)$, $\mathbf{d}_{0,M}(f)$ is the steering vector at $\theta = 0$, $X(f)$ is the zero-mean desired source signal, $\mathbf{v}(f)$ is the zero-mean additive noise signal vector defined similarly to $\mathbf{y}(f)$, and $X(f)$ and $\mathbf{v}(f)$ are incoherent. In the rest, in order to simplify the notation, we drop the dependence on the temporal frequency, f. For a small and compact array, it is reasonable to assume that the variance of the noise is the same at all sensors, i.e., $\phi_V = \phi_{V_1} = \phi_{V_2} = \cdots = \phi_{V_M}$, with $\phi_{V_m} = E\left(|V_m|^2\right)$, $m = 1, 2, \ldots, M$ and $E(\cdot)$ denoting mathematical expectation. The meaning of this assumption is as follows. Considering the self-noise of the microphones, it implies that all of them are, roughly, of the same kind and have the same level of imperfections. Considering directional interferences and spherically isotropic (diffuse) noise, it implies that the power of the received signals is similar in all microphones, that is, the distance between the signal source to the array reference microphone is much bigger than the interelement spacing. Clearly, this is a restatement of the farfield signal model. Consequently, the variance of Y_m is $\phi_{Y_m} = \phi_Y = \phi_X + \phi_V$, where ϕ_X is the variance of X.

Let P be a positive integer with $0 \leq P < M$. We can transform the observed signal vector $\mathbf{y}$ of length M to a Pth-order forward spatial difference of $\mathbf{y}$ of length $M(P) = M - P$, i.e., [24]

$$\mathbf{y}_{(P)} = \mathbf{\Delta}_{(P)}\mathbf{y}, \quad (3)$$

with $\mathbf{y}_{(0)} = \mathbf{y}$, where

$$\mathbf{\Delta}_{(P)} = \begin{bmatrix} \mathbf{c}_{(P)}^T & 0 & \cdots & 0 \\ 0 & \mathbf{c}_{(P)}^T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{c}_{(P)}^T \end{bmatrix} \quad (4)$$

is a matrix of size $M(P) \times M$, with $\mathbf{\Delta}_{(0)} = \mathbf{I}_M$, which is the $M \times M$ identity matrix,

$$\mathbf{c}_{(P)} = \begin{bmatrix} (-1)^P \binom{P}{0} & (-1)^{P-1}\binom{P}{1} & \cdots & -\binom{P}{P-1} & 1 \end{bmatrix}^T \quad (5)$$

is a vector of length $P+1$, and

$$\binom{P}{j} = \frac{P!}{j!\,(P-j)!}$$

is the binomial coefficient. The major benefit with the difference observed signal vector, $\mathbf{y}_{(P)}$ of length $M(P)$, for $P > 0$, is that it is less sensitive to diffuse noise as compared to $\mathbf{y}$; in fact, the larger is the value of P, the higher is the signal-to-noise ratio (SNR) of $\mathbf{y}_{(P)}$. However, $\mathbf{y}_{(P)}$ is more sensitive to white noise.

It can be shown that (3) can be expressed as [24]

$$\begin{aligned} \mathbf{y}_{(P)} &= \tau_0^P \mathbf{d}_{0,M(P)} X + \mathbf{v}_{(P)} \\ &= \mathbf{x}_{(P)} + \mathbf{v}_{(P)}, \end{aligned} \quad (6)$$

where

$$\tau_0 = e^{-\jmath 2\pi f\delta/c} - 1 \quad (7)$$

is a frequency-dependent variable, $\mathbf{d}_{0,M(P)}$ is the steering vector of length $M(P)$ at $\theta = 0$, and $\mathbf{v}_{(P)} = \mathbf{\Delta}_{(P)}\mathbf{v}$. We deduce that the $M(P) \times M(P)$ covariance matrix of $\mathbf{y}_{(P)}$ is

$$\begin{aligned} \mathbf{\Phi}_{\mathbf{y}_{(P)}} &= \phi_X |\tau_0|^{2P} \mathbf{d}_{0,M(P)}\mathbf{d}_{0,M(P)}^H + \mathbf{\Delta}_{(P)}\mathbf{\Phi}_{\mathbf{v}}\mathbf{\Delta}_{(P)}^T \\ &= \phi_X |\tau_0|^{2P} \mathbf{d}_{0,M(P)}\mathbf{d}_{0,M(P)}^H + \mathbf{\Phi}_{\mathbf{v}_{(P)}} \\ &= \phi_X |\tau_0|^{2P} \mathbf{d}_{0,M(P)}\mathbf{d}_{0,M(P)}^H + \phi_V \mathbf{\Delta}_{(P)}\mathbf{\Gamma}_{\mathbf{v}}\mathbf{\Delta}_{(P)}^T \\ &= \phi_X |\tau_0|^{2P} \mathbf{d}_{0,M(P)}\mathbf{d}_{0,M(P)}^H + \phi_V \mathbf{\Gamma}_{\mathbf{v}_{(P)}}, \end{aligned} \quad (8)$$

where $\mathbf{\Phi}_{\mathbf{v}}$ is the covariance matrix of $\mathbf{v}$, $\mathbf{\Gamma}_{\mathbf{v}} = \mathbf{\Phi}_{\mathbf{v}}/\phi_V$, and $\mathbf{\Gamma}_{\mathbf{v}_{(P)}} = \mathbf{\Delta}_{(P)}\mathbf{\Gamma}_{\mathbf{v}}\mathbf{\Delta}_{(P)}^T$.

III. Differential Kronecker Product Beamforming

Assume that $M(P) = M - P = M_{\mathbf{a}} \times M_{\mathbf{b}}$, where $M_{\mathbf{a}}, M_{\mathbf{b}} \geq 1$. Then, one can verify that the steering vector $\mathbf{d}_{\theta,M(P)}$ can be decomposed as [27]

$$\mathbf{d}_{\theta,M(P)} = \mathbf{a}_\theta \otimes \mathbf{b}_\theta, \tag{9}$$

where

$$\mathbf{a}_\theta = \Big[1 \quad e^{-\jmath 2\pi f M_{\mathbf{b}} \delta \cos\theta/c} \quad \cdots \quad e^{-\jmath(M_{\mathbf{a}}-1)2\pi f M_{\mathbf{b}} \delta \cos\theta/c} \Big]^T \tag{10}$$

is the steering vector (of length $M_{\mathbf{a}}$) corresponding to a ULA of $M_{\mathbf{a}}$ sensors with an interelement spacing equal to $M_{\mathbf{b}}\delta$, $\otimes$ is the Kronecker product, and

$$\mathbf{b}_\theta = \Big[1 \quad e^{-\jmath 2\pi f \delta \cos\theta/c} \quad \cdots \quad e^{-\jmath(M_{\mathbf{b}}-1)2\pi f \delta \cos\theta/c} \Big]^T \tag{11}$$

is the steering vector (of length $M_{\mathbf{b}}$) corresponding to a ULA of $M_{\mathbf{b}}$ sensors with an interelement spacing equal to δ. As a consequence, the signal model in (6) becomes

$$\mathbf{y}_{(P)} = \tau_0^P \left(\mathbf{a}_0 \otimes \mathbf{b}_0\right) X + \mathbf{v}_{(P)} \tag{12}$$

and its covariance matrix is

$$\mathbf{\Phi}_{\mathbf{y}_{(P)}} = \phi_X \left|\tau_0\right|^{2P} \left(\mathbf{a}_0\mathbf{a}_0^H\right) \otimes \left(\mathbf{b}_0\mathbf{b}_0^H\right) + \phi_V \mathbf{\Gamma}_{\mathbf{v}_{(P)}}. \tag{13}$$

Because of the particular structure of the steering vector in (9) and in order to fully exploit this structure, we propose (global) beamformers of the form:

$$\mathbf{h} = \mathbf{h}_{\mathbf{a}} \otimes \mathbf{h}_{\mathbf{b}}, \tag{14}$$

where $\mathbf{h}_{\mathbf{a}}$ and $\mathbf{h}_{\mathbf{b}}$ are two complex-valued linear filters of lengths $M_{\mathbf{a}}$ and $M_{\mathbf{b}}$, respectively. Then, in the proposed context, linear beamforming is performed by applying $\mathbf{h}$ [from (14)] to $\mathbf{y}_{(P)}$ [from (12)], i.e.,

$$\begin{aligned} Z &= \mathbf{h}^H \mathbf{y}_{(P)} \\ &= \mathbf{h}^H \mathbf{x}_{(P)} + \mathbf{h}^H \mathbf{v}_{(P)} \\ &= X_{\mathrm{fd}} + V_{\mathrm{rn}}, \end{aligned} \tag{15}$$

where Z is the estimate of the desired signal, X,

$$\begin{aligned} X_{\mathrm{fd}} &= \tau_0^P \left(\mathbf{h}_{\mathbf{a}} \otimes \mathbf{h}_{\mathbf{b}}\right)^H \left(\mathbf{a}_0 \otimes \mathbf{b}_0\right) X \\ &= \tau_0^P \left(\mathbf{h}_{\mathbf{a}}^H \mathbf{a}_0\right) \left(\mathbf{h}_{\mathbf{b}}^H \mathbf{b}_0\right) X \end{aligned} \tag{16}$$

is the filtered desired signal, and

$$V_{\mathrm{rn}} = \left(\mathbf{h}_{\mathbf{a}} \otimes \mathbf{h}_{\mathbf{b}}\right)^H \mathbf{v}_{(P)} \tag{17}$$

is the residual noise. We deduce that the variance of Z is

$$\begin{aligned} \phi_Z = {} & \left|\tau_0\right|^{2P} \left|\mathbf{h}_{\mathbf{a}}^H \mathbf{a}_0\right|^2 \left|\mathbf{h}_{\mathbf{b}}^H \mathbf{b}_0\right|^2 \phi_X \\ & + \phi_V \left(\mathbf{h}_{\mathbf{a}} \otimes \mathbf{h}_{\mathbf{b}}\right)^H \mathbf{\Gamma}_{\mathbf{v}_{(P)}} \left(\mathbf{h}_{\mathbf{a}} \otimes \mathbf{h}_{\mathbf{b}}\right). \end{aligned} \tag{18}$$

We see from X_{fd} that the distortionless constraint is

$$\left(\mathbf{h}_{\mathbf{a}}^H \mathbf{a}_0\right) \left(\mathbf{h}_{\mathbf{b}}^H \mathbf{b}_0\right) = \tau_0^{-P}. \tag{19}$$

In the rest, we choose $\mathbf{h}_{\mathbf{a}}^H \mathbf{a}_0 = \tau_0^{-P}$ and $\mathbf{h}_{\mathbf{b}}^H \mathbf{b}_0 = 1$, so that (19) is satisfied.

Furthermore, we will often use the following relationships:

$$\begin{aligned} \mathbf{h}_{\mathbf{a}} \otimes \mathbf{h}_{\mathbf{b}} &= \left(\mathbf{h}_{\mathbf{a}} \otimes \mathbf{I}_{M_{\mathbf{b}}}\right) \mathbf{h}_{\mathbf{b}} && (20) \\ &= \left(\mathbf{I}_{M_{\mathbf{a}}} \otimes \mathbf{h}_{\mathbf{b}}\right) \mathbf{h}_{\mathbf{a}}, && (21) \end{aligned}$$

where $\mathbf{I}_{M_{\mathbf{b}}}$ and $\mathbf{I}_{M_{\mathbf{a}}}$ are the identity matrices of sizes $M_{\mathbf{b}} \times M_{\mathbf{b}}$ and $M_{\mathbf{a}} \times M_{\mathbf{a}}$, respectively.

IV. Performance Measures

In this section, we express common performance measures according to the differential KP beamforming approach.

The first useful measure discussed in this section is the beampattern, which describes the sensitivity of the beamformer to a plane wave (source signal) impinging on the ULA from the direction θ. Mathematically, it is defined as

$$\begin{aligned} \mathcal{B}_\theta\left(\mathbf{h}\right) &= \tau_\theta^P \mathbf{h}^H \mathbf{d}_{\theta,M(P)} \\ &= \tau_\theta^P \times \mathbf{h}_{\mathbf{a}}^H \mathbf{a}_\theta \times \mathbf{h}_{\mathbf{b}}^H \mathbf{b}_\theta \\ &= \mathcal{B}_{\theta,\mathbf{a}}\left(\mathbf{h}_{\mathbf{a}}\right) \times \mathcal{B}_{\theta,\mathbf{b}}\left(\mathbf{h}_{\mathbf{b}}\right), \end{aligned} \tag{22}$$

where

$$\tau_\theta = e^{-\jmath 2\pi f \delta \cos\theta/c} - 1, \tag{23}$$

$\mathcal{B}_{\theta,\mathbf{a}}\left(\mathbf{h}_{\mathbf{a}}\right) = \tau_\theta^P \mathbf{h}_{\mathbf{a}}^H \mathbf{a}_\theta$, and $\mathcal{B}_{\theta,\mathbf{b}}\left(\mathbf{h}_{\mathbf{b}}\right) = \mathbf{h}_{\mathbf{b}}^H \mathbf{b}_\theta$. The global beampattern is composed of three terms: the first one, τ_θ^P, emphasizes the directivity of the pattern; the second term, $\mathbf{h}_{\mathbf{a}}^H \mathbf{a}_\theta$, is the beampattern of the first ULA with an interelement spacing equal to $M_{\mathbf{b}}\delta$, and the last term, $\mathbf{h}_{\mathbf{b}}^H \mathbf{b}_\theta$, corresponds to the beampattern of the second ULA with an interelement spacing equal to δ.

From (2), we easily find that the input SNR is

$$\mathrm{iSNR} = \frac{\phi_X}{\phi_V}. \tag{24}$$

The output SNR is defined [from the variance of Z, see (18)] as

$$\mathrm{oSNR}\left(\mathbf{h}\right) = \frac{\phi_X}{\phi_V} \times \frac{\left|\tau_0\right|^{2P} \left|\mathbf{h}^H \mathbf{d}_{\theta,M(P)}\right|^2}{\mathbf{h}^H \mathbf{\Gamma}_{\mathbf{v}_{(P)}} \mathbf{h}}. \tag{25}$$

The definition of the gain in SNR is obtained from the previous definitions, i.e.,

$$\begin{aligned} \mathcal{G}\left(\mathbf{h}\right) &= \frac{\mathrm{oSNR}\left(\mathbf{h}\right)}{\mathrm{iSNR}} \\ &= \frac{\left|\tau_0\right|^{2P} \left|\mathbf{h}^H \mathbf{d}_{\theta,M(P)}\right|^2}{\mathbf{h}^H \mathbf{\Gamma}_{\mathbf{v}_{(P)}} \mathbf{h}}. \end{aligned} \tag{26}$$

We can rewrite this gain as

$$\mathcal{G}\left(\mathbf{h}_{\mathbf{a}}, \mathbf{h}_{\mathbf{b}}\right) = \frac{\left|\tau_0\right|^{2P} \left|\mathbf{h}_{\mathbf{a}}^H \mathbf{a}_0\right|^2 \left|\mathbf{h}_{\mathbf{b}}^H \mathbf{b}_0\right|^2}{\left(\mathbf{h}_{\mathbf{a}} \otimes \mathbf{h}_{\mathbf{b}}\right)^H \mathbf{\Gamma}_{\mathbf{v}_{(P)}} \left(\mathbf{h}_{\mathbf{a}} \otimes \mathbf{h}_{\mathbf{b}}\right)}. \tag{27}$$

When $\mathbf{h}_{\mathbf{b}}$ is fixed, given, and satisfies the distortionless constraint, i.e., $\mathbf{h}_{\mathbf{b}}^H \mathbf{b}_0 = 1$; then, we can express the gain as

$$\mathcal{G}\left(\mathbf{h}_{\mathbf{a}} \left| \mathbf{h}_{\mathbf{b}} \right.\right) = \frac{\left|\tau_0\right|^{2P} \left|\mathbf{h}_{\mathbf{a}}^H \mathbf{a}_0\right|^2}{\mathbf{h}_{\mathbf{a}}^H \mathbf{\Gamma}_{\mathbf{v}_{(P)},\mathbf{b}} \mathbf{h}_{\mathbf{a}}}, \tag{28}$$

where

$$\mathbf{\Gamma}_{\mathbf{v}_{(P)},\mathbf{b}} = \left(\mathbf{I}_{M_{\mathbf{a}}} \otimes \mathbf{h}_{\mathbf{b}}\right)^H \mathbf{\Gamma}_{\mathbf{v}_{(P)}} \left(\mathbf{I}_{M_{\mathbf{a}}} \otimes \mathbf{h}_{\mathbf{b}}\right). \tag{29}$$

In the same way, when $\mathbf{h_a}$ is fixed, given, and satisfies the distortionless constraint, i.e., $\mathbf{h_a}^H\mathbf{a}_0 = \tau_0^{-P}$; then, we can express the gain as

$$\mathcal{G}\left(\mathbf{h_b}\left|\mathbf{h_a}\right.\right) = \frac{\left|\mathbf{h_b}^H\mathbf{b}_0\right|^2}{\mathbf{h_b}^H\mathbf{\Gamma}_{\mathbf{v}_{(P)},\mathbf{a}}\mathbf{h_b}}, \tag{30}$$

where

$$\mathbf{\Gamma}_{\mathbf{v}_{(P)},\mathbf{a}} = \left(\mathbf{h_a}\otimes\mathbf{I}_{M_\mathbf{b}}\right)^H\mathbf{\Gamma}_{\mathbf{v}_{(P)}}\left(\mathbf{h_a}\otimes\mathbf{I}_{M_\mathbf{b}}\right). \tag{31}$$

One important measure which expresses the sensitivity of the ULA to spatially white noise is the WNG, which is defined by taking $\mathbf{\Gamma_v} = \mathbf{I}_M$, i.e.,

$$\begin{aligned}\mathcal{W}\left(\mathbf{h}\right) &= \frac{\left|\tau_0\right|^{2P}\left|\mathbf{h}^H\mathbf{d}_{\theta,M(P)}\right|^2}{\mathbf{h}^H\mathbf{\Delta}_{(P)}\mathbf{\Delta}_{(P)}^T\mathbf{h}}\\ &= \frac{\left|\tau_0\right|^{2P}\left|\mathbf{h_a}^H\mathbf{a}_0\right|^2\left|\mathbf{h_b}^H\mathbf{b}_0\right|^2}{\left(\mathbf{h_a}\otimes\mathbf{h_b}\right)^H\mathbf{\Delta}_{(P)}\mathbf{\Delta}_{(P)}^T\left(\mathbf{h_a}\otimes\mathbf{h_b}\right)}\\ &= \mathcal{W}\left(\mathbf{h_a},\mathbf{h_b}\right),\end{aligned} \tag{32}$$

which can only be expressed as a product of the WNGs of the two sub-beamformers for $P = 0$.

When $\mathbf{h_b}$ is fixed, given, and satisfies the distortionless constraint, i.e., $\mathbf{h_b}^H\mathbf{b}_0 = 1$; then, we can express the WNG as

$$\mathcal{W}\left(\mathbf{h_a}\left|\mathbf{h_b}\right.\right) = \frac{\left|\tau_0\right|^{2P}\left|\mathbf{h_a}^H\mathbf{a}_0\right|^2}{\mathbf{h_a}^H\left[\left(\mathbf{I}_{M_\mathbf{a}}\otimes\mathbf{h_b}\right)^H\mathbf{\Delta}_{(P)}\mathbf{\Delta}_{(P)}^T\left(\mathbf{I}_{M_\mathbf{a}}\otimes\mathbf{h_b}\right)\right]\mathbf{h_a}}. \tag{33}$$

Similarly, when $\mathbf{h_a}$ is fixed, given, and satisfies the distortionless constraint, i.e., $\mathbf{h_a}^H\mathbf{a}_0 = \tau_0^{-P}$ we have

$$\mathcal{W}\left(\mathbf{h_b}\left|\mathbf{h_a}\right.\right) = \frac{\left|\mathbf{h_b}^H\mathbf{b}_0\right|^2}{\mathbf{h_b}^H\left[\left(\mathbf{h_a}\otimes\mathbf{I}_{M_\mathbf{b}}\right)^H\mathbf{\Delta}_{(P)}\mathbf{\Delta}_{(P)}^T\left(\mathbf{h_a}\otimes\mathbf{I}_{M_\mathbf{b}}\right)\right]\mathbf{h_b}}. \tag{34}$$

Another important measure, which quantifies how the microphone array performs in the presence of reverberation is the DF. Considering the spherically isotropic (diffuse) noise field, the DF is defined as

$$\begin{aligned}\mathcal{D}\left(\mathbf{h}\right) &= \frac{\left|\mathcal{B}_0\left(\mathbf{h}\right)\right|^2}{\frac{1}{2}\int_0^\pi\left|\mathcal{B}_\theta\left(\mathbf{h}\right)\right|^2\sin\theta d\theta}\\ &= \frac{\left|\mathcal{B}_{0,\mathbf{a}}\left(\mathbf{h_a}\right)\right|^2\left|\mathcal{B}_{0,\mathbf{b}}\left(\mathbf{h_b}\right)\right|^2}{\frac{1}{2}\int_0^\pi\left|\mathcal{B}_{\theta,\mathbf{a}}\left(\mathbf{h_a}\right)\right|^2\left|\mathcal{B}_{\theta,\mathbf{b}}\left(\mathbf{h_b}\right)\right|^2\sin\theta d\theta}\\ &= \frac{\left|\tau_0\right|^{2P}\left|\mathbf{h_a}^H\mathbf{a}_0\right|^2\left|\mathbf{h_b}^H\mathbf{b}_0\right|^2}{\left(\mathbf{h_a}\otimes\mathbf{h_b}\right)^H\mathbf{\Delta}_{(P)}\mathbf{\Gamma}_\mathrm{d}\mathbf{\Delta}_{(P)}^T\left(\mathbf{h_a}\otimes\mathbf{h_b}\right)}\\ &= \mathcal{D}\left(\mathbf{h_a},\mathbf{h_b}\right),\end{aligned} \tag{35}$$

where the elements of the diffuse noise coherence matrix $\mathbf{\Gamma}_\mathrm{d}$ are given by

$$\left(\mathbf{\Gamma}_\mathrm{d}\right)_{ij} = \frac{\sin\left[2\pi f(j-i)\delta/c\right]}{2\pi f(j-i)\delta/c}. \tag{36}$$

When $\mathbf{h_b}$ is fixed, given, and satisfies the distortionless constraint, i.e., $\mathbf{h_b}^H\mathbf{b}_0 = 1$; then, we can express the DF as

$$\mathcal{D}\left(\mathbf{h_a}\left|\mathbf{h_b}\right.\right) = \frac{\left|\tau_0\right|^{2P}\left|\mathbf{h_a}^H\mathbf{a}_0\right|^2}{\mathbf{h_a}^H\mathbf{\Gamma}_{\mathrm{d},\mathbf{b}}\mathbf{h_a}}, \tag{37}$$

where

$$\mathbf{\Gamma}_{\mathrm{d},\mathbf{b}} = \frac{1}{2}\int_0^\pi\left|\tau_\theta\right|^{2P}\mathbf{a}_\theta\mathbf{a}_\theta^H\left|\mathcal{B}_{\theta,\mathbf{b}}\left(\mathbf{h_b}\right)\right|^2\sin\theta d\theta. \tag{38}$$

In an analogous way, when $\mathbf{h_a}$ is fixed, given, and satisfies the distortionless constraint, i.e., $\mathbf{h_a}^H\mathbf{a}_0 = \tau_0^{-P}$; then, we can write the DF as

$$\mathcal{D}\left(\mathbf{h_b}\left|\mathbf{h_a}\right.\right) = \frac{\left|\mathbf{h_b}^H\mathbf{b}_0\right|^2}{\mathbf{h_b}^H\mathbf{\Gamma}_{\mathrm{d},\mathbf{a}}\mathbf{h_b}}, \tag{39}$$

where

$$\mathbf{\Gamma}_{\mathrm{d},\mathbf{a}} = \frac{1}{2}\int_0^\pi\mathbf{b}_\theta\mathbf{b}_\theta^H\left|\mathcal{B}_{\theta,\mathbf{a}}\left(\mathbf{h_a}\right)\right|^2\sin\theta d\theta. \tag{40}$$

Finally, the last measure of interest in this section is the front-to-back ratio (FBR), which is defined as the ratio of the power of the output of the array to signals propagating from the front-half plane to the output power for signals arriving from the rear-half plane. This ratio, for the spherically isotropic (diffuse) noise field, is mathematically defined as

$$\begin{aligned}\mathcal{F}\left(\mathbf{h}\right) &= \frac{\int_0^{\pi/2}\left|\mathcal{B}_\theta\left(\mathbf{h}\right)\right|^2\sin\theta d\theta}{\int_{\pi/2}^{\pi}\left|\mathcal{B}_\theta\left(\mathbf{h}\right)\right|^2\sin\theta d\theta}\\ &= \frac{\int_0^{\pi/2}\left|\mathcal{B}_{\theta,\mathbf{a}}\left(\mathbf{h_a}\right)\right|^2\left|\mathcal{B}_{\theta,\mathbf{b}}\left(\mathbf{h_b}\right)\right|^2\sin\theta d\theta}{\int_{\pi/2}^{\pi}\left|\mathcal{B}_{\theta,\mathbf{a}}\left(\mathbf{h_a}\right)\right|^2\left|\mathcal{B}_{\theta,\mathbf{b}}\left(\mathbf{h_b}\right)\right|^2\sin\theta d\theta}\\ &= \mathcal{F}\left(\mathbf{h_a},\mathbf{h_b}\right).\end{aligned} \tag{41}$$

When $\mathbf{h_b}$ is fixed and given; then, we can express the FBR as

$$\mathcal{F}\left(\mathbf{h_a}\left|\mathbf{h_b}\right.\right) = \frac{\mathbf{h_a}^H\mathbf{\Gamma}_{\mathrm{f},\mathbf{b}}\mathbf{h_a}}{\mathbf{h_a}^H\mathbf{\Gamma}_{\mathrm{b},\mathbf{b}}\mathbf{h_a}}, \tag{42}$$

where

$$\mathbf{\Gamma}_{\mathrm{f},\mathbf{b}} = \int_0^{\pi/2}\left|\tau_\theta\right|^{2P}\mathbf{a}_\theta\mathbf{a}_\theta^H\left|\mathcal{B}_{\theta,\mathbf{b}}\left(\mathbf{h_b}\right)\right|^2\sin\theta d\theta, \tag{43}$$

$$\mathbf{\Gamma}_{\mathrm{b},\mathbf{b}} = \int_{\pi/2}^{\pi}\left|\tau_\theta\right|^{2P}\mathbf{a}_\theta\mathbf{a}_\theta^H\left|\mathcal{B}_{\theta,\mathbf{b}}\left(\mathbf{h_b}\right)\right|^2\sin\theta d\theta. \tag{44}$$

Similarly, when $\mathbf{h_a}$ is fixed and given; then, we can express the FBR as

$$\mathcal{F}\left(\mathbf{h_b}\left|\mathbf{h_a}\right.\right) = \frac{\mathbf{h_b}^H\mathbf{\Gamma}_{\mathrm{f},\mathbf{a}}\mathbf{h_b}}{\mathbf{h_b}^H\mathbf{\Gamma}_{\mathrm{b},\mathbf{a}}\mathbf{h_b}}, \tag{45}$$

where

$$\mathbf{\Gamma}_{\mathrm{f},\mathbf{a}} = \int_0^{\pi/2}\mathbf{b}_\theta\mathbf{b}_\theta^H\left|\mathcal{B}_{\theta,\mathbf{a}}\left(\mathbf{h_a}\right)\right|^2\sin\theta d\theta, \tag{46}$$

$$\mathbf{\Gamma}_{\mathrm{b},\mathbf{a}} = \int_{\pi/2}^{\pi}\mathbf{b}_\theta\mathbf{b}_\theta^H\left|\mathcal{B}_{\theta,\mathbf{a}}\left(\mathbf{h_a}\right)\right|^2\sin\theta d\theta. \tag{47}$$

V. Examples of Optimal Differential KP Beamformers

In this section, we propose five types of differential KP beamformers, each designed with respect to a different optimization criterion.

A. Maximum White Noise Gain

Let us start with the WNG measure. It is not possible to get a closed-form expression beamformer from the maximization of the WNG. However, using the alternating least-squares (ALS) strategy [33], [34], we can derive the maximum WNG (MWNG) beamformer iteratively from

$$\min_{\mathbf{h_a}} \mathbf{h_a}^H \mathbf{\Gamma}_{\Delta,\mathbf{b}} \mathbf{h_a} \quad \text{s. t.} \quad \mathbf{h_a}^H \mathbf{a}_0 = \tau_0^{-P}, \tag{48}$$

$$\min_{\mathbf{h_b}} \mathbf{h_b}^H \mathbf{\Gamma}_{\Delta,\mathbf{a}} \mathbf{h_b} \quad \text{s. t.} \quad \mathbf{h_b}^H \mathbf{b}_0 = 1, \tag{49}$$

whose solution is easily obtained by

$$\mathbf{h}_{\mathbf{a},\mathrm{MWNG}} = \frac{{\mathbf{\Gamma}_{\Delta,\mathbf{b}}}^{-1} \mathbf{a}_0}{\left(\tau_0^P\right)^* \mathbf{a}_0^H {\mathbf{\Gamma}_{\Delta,\mathbf{b}}}^{-1} \mathbf{a}_0}, \tag{50}$$

$$\mathbf{h}_{\mathbf{b},\mathrm{MWNG}} = \frac{{\mathbf{\Gamma}_{\Delta,\mathbf{a}}}^{-1} \mathbf{b}_0}{\mathbf{b}_0^H {\mathbf{\Gamma}_{\Delta,\mathbf{a}}}^{-1} \mathbf{b}_0}, \tag{51}$$

where

$$\mathbf{\Gamma}_{\Delta,\mathbf{a}} = (\mathbf{h}_{\mathbf{a},\mathrm{MWNG}} \otimes \mathbf{I}_{M_\mathbf{b}})^H \mathbf{\Delta}_{(P)} \mathbf{\Delta}_{(P)}^T (\mathbf{h}_{\mathbf{a},\mathrm{MWNG}} \otimes \mathbf{I}_{M_\mathbf{b}}), \tag{52}$$

$$\mathbf{\Gamma}_{\Delta,\mathbf{b}} = (\mathbf{I}_{M_\mathbf{a}} \otimes \mathbf{h}_{\mathbf{b},\mathrm{MWNG}})^H \mathbf{\Delta}_{(P)} \mathbf{\Delta}_{(P)}^T (\mathbf{I}_{M_\mathbf{a}} \otimes \mathbf{h}_{\mathbf{b},\mathrm{MWNG}}). \tag{53}$$

As a result, at iteration n, the global MWNG beamformer is

$$\mathbf{h}_{\mathrm{MWNG}}^{(n)} = \mathbf{h}_{\mathbf{a},\mathrm{MWNG}}^{(n)} \otimes \mathbf{h}_{\mathbf{b},\mathrm{MWNG}}^{(n)}, \tag{54}$$

where

$$\mathbf{h}_{\mathbf{a},\mathrm{MWNG}}^{(n)} = \frac{{\mathbf{\Gamma}_{\Delta,\mathbf{b}}^{(n)}}^{-1} \mathbf{a}_0}{\left(\tau_0^P\right)^* \mathbf{a}_0^H {\mathbf{\Gamma}_{\Delta,\mathbf{b}}^{(n)}}^{-1} \mathbf{a}_0}, \tag{55}$$

$$\mathbf{h}_{\mathbf{b},\mathrm{MWNG}}^{(n)} = \frac{{\mathbf{\Gamma}_{\Delta,\mathbf{a}}^{(n)}}^{-1} \mathbf{b}_0}{\mathbf{b}_0^H {\mathbf{\Gamma}_{\Delta,\mathbf{a}}^{(n)}}^{-1} \mathbf{b}_0}, \tag{56}$$

and the iteratively updated coherence matrices are given by

$$\mathbf{\Gamma}_{\Delta,\mathbf{a}}^{(n)} = (\mathbf{h}_{\mathbf{a},\mathrm{MWNG}}^{(n)} \otimes \mathbf{I}_{M_\mathbf{b}})^H \mathbf{\Delta}_{(P)} \mathbf{\Delta}_{(P)}^T (\mathbf{h}_{\mathbf{a},\mathrm{MWNG}}^{(n)} \otimes \mathbf{I}_{M_\mathbf{b}}), \tag{57}$$

$$\mathbf{\Gamma}_{\Delta,\mathbf{b}}^{(n)} = (\mathbf{I}_{M_\mathbf{a}} \otimes \mathbf{h}_{\mathbf{b},\mathrm{MWNG}}^{(n-1)})^H \mathbf{\Delta}_{(P)} \mathbf{\Delta}_{(P)}^T (\mathbf{I}_{M_\mathbf{a}} \otimes \mathbf{h}_{\mathbf{b},\mathrm{MWNG}}^{(n-1)}). \tag{58}$$

B. Null Steering

Now, assume that we have one interference source impinging on the array from the direction $\theta_\mathrm{i} \neq \theta_\mathrm{s} = 0$ that we would like to completely cancel, i.e., to steer a null in that direction and, meanwhile, recover the desired source coming from the direction $\theta_\mathrm{s} = 0$. One possible approach to achieve such a behaviour is to have a null in the beampattern $\mathcal{B}_{\theta,\mathbf{b}}(\mathbf{h_b})$, which implies a null in the global beampattern $\mathcal{B}_\theta(\mathbf{h})$, and a null constraint on the filter $\mathbf{h_b}$. Then, by including the distortionless constraint, we can write the constraint equation as

$$\mathbf{C}^H \mathbf{h_b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \tag{59}$$

where

$$\mathbf{C} = \begin{bmatrix} \mathbf{b}_0 & \mathbf{b}_{\theta_\mathrm{i}} \end{bmatrix} \tag{60}$$

is the constraint matrix of size $M_\mathbf{b} \times 2$ whose two columns are linearly independent. To find this filter, we maximize the WNG by taking (59) into account, i.e.,

$$\min_{\mathbf{h_b}} \mathbf{h_b}^H \mathbf{\Gamma}_{\Delta,\mathbf{a}} \mathbf{h_b} \quad \text{s. t.} \quad \mathbf{C}^H \mathbf{h_b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{61}$$

From this criterion, we get the following null-steering (NS) beamformer:

$$\mathbf{h}_{\mathbf{b},\mathrm{NS}} = \mathbf{\Gamma}_{\Delta,\mathbf{a}}^{-1} \mathbf{C} \left[\mathbf{C}^H \mathbf{\Gamma}_{\Delta,\mathbf{a}}^{-1} \mathbf{C}\right]^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{62}$$

For the other beamformer, we may choose $\mathbf{h}_{\mathbf{a},\mathrm{MWNG}}$ in (50). Therefore, the (global) proposed NS beamformer is obtained, iteratively, by

$$\mathbf{h}_{\mathrm{NS}}^{(n)} = \mathbf{h}_{\mathbf{a},\mathrm{MWNG}}^{(n)} \otimes \mathbf{h}_{\mathbf{b},\mathrm{NS}}^{(n)}, \tag{63}$$

where $\mathbf{h}_{\mathbf{a},\mathrm{MWNG}}^{(n)}$ is identical to the expression in (55) and

$$\mathbf{h}_{\mathbf{b},\mathrm{NS}}^{(n)} = {\mathbf{\Gamma}_{\Delta,\mathbf{a}}^{(n)}}^{-1} \mathbf{C} \left[\mathbf{C}^H {\mathbf{\Gamma}_{\Delta,\mathbf{a}}^{(n)}}^{-1} \mathbf{C}\right]^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{64}$$

C. Maximum Directivity Factor

Let us focus on the DF measure. In a similar manner to the MWNG beamformer, we can derive the maximum DF (MDF) beamformer iteratively from

$$\min_{\mathbf{h_a}} \mathbf{h_a}^H \mathbf{\Gamma}_{\mathrm{d},\mathbf{b}} \mathbf{h_a} \quad \text{s. t.} \quad \mathbf{h_a}^H \mathbf{a}_0 = \tau_0^{-P}, \tag{65}$$

$$\min_{\mathbf{h_b}} \mathbf{h_b}^H \mathbf{\Gamma}_{\mathrm{d},\mathbf{a}} \mathbf{h_b} \quad \text{s. t.} \quad \mathbf{h_b}^H \mathbf{b}_0 = 1. \tag{66}$$

We get

$$\mathbf{h}_{\mathbf{a},\mathrm{MDF}} = \frac{\mathbf{\Gamma}_{\mathrm{d},\mathbf{b}}^{-1} \mathbf{a}_0}{\left(\tau_0^P\right)^* \mathbf{a}_0^H \mathbf{\Gamma}_{\mathrm{d},\mathbf{b}}^{-1} \mathbf{a}_0}, \tag{67}$$

$$\mathbf{h}_{\mathbf{b},\mathrm{MDF}} = \frac{\mathbf{\Gamma}_{\mathrm{d},\mathbf{a}}^{-1} \mathbf{b}_0}{\mathbf{b}_0^H \mathbf{\Gamma}_{\mathrm{d},\mathbf{a}}^{-1} \mathbf{b}_0}. \tag{68}$$

As a result, at iteration n, the MDF beamformer is

$$\mathbf{h}_{\mathrm{MDF}}^{(n)} = \mathbf{h}_{\mathbf{a},\mathrm{MDF}}^{(n)} \otimes \mathbf{h}_{\mathbf{b},\mathrm{MDF}}^{(n)}, \tag{69}$$

where

$$\mathbf{h}_{\mathbf{a},\mathrm{MDF}}^{(n)} = \frac{{\mathbf{\Gamma}_{\mathrm{d},\mathbf{b}}^{(n)}}^{-1} \mathbf{a}_0}{\left(\tau_0^P\right)^* \mathbf{a}_0^H {\mathbf{\Gamma}_{\mathrm{d},\mathbf{b}}^{(n)}}^{-1} \mathbf{a}_0}, \tag{70}$$

$$\mathbf{h}_{\mathbf{b},\mathrm{MDF}}^{(n)} = \frac{{\mathbf{\Gamma}_{\mathrm{d},\mathbf{a}}^{(n)}}^{-1} \mathbf{b}_0}{\mathbf{b}_0^H {\mathbf{\Gamma}_{\mathrm{d},\mathbf{a}}^{(n)}}^{-1} \mathbf{b}_0}, \tag{71}$$

and the iteratively updated coherence matrices are

$$\mathbf{\Gamma}_{\mathrm{d},\mathbf{a}}^{(n)} = \left(\mathbf{h}_{\mathbf{a},\mathrm{MDF}}^{(n)} \otimes \mathbf{I}_{M_\mathbf{b}}\right)^H \mathbf{\Gamma}_\mathrm{d} \left(\mathbf{h}_{\mathbf{a},\mathrm{MDF}}^{(n)} \otimes \mathbf{I}_{M_\mathbf{b}}\right), \quad (72)$$

$$\mathbf{\Gamma}_{\mathrm{d},\mathbf{b}}^{(n)} = \left(\mathbf{I}_{M_\mathbf{a}} \otimes \mathbf{h}_{\mathbf{b},\mathrm{MDF}}^{(n-1)}\right)^H \mathbf{\Gamma}_\mathrm{d} \left(\mathbf{I}_{M_\mathbf{a}} \otimes \mathbf{h}_{\mathbf{b},\mathrm{MDF}}^{(n-1)}\right). \quad (73)$$

We end this part by addressing a possible compromise between the WNG and DF measures. That is, taking advantage of the KP decomposition, we can combine the MWNG and MDF beamformers. For example, $\mathbf{h}_\mathbf{a}$ may be designed as a MWNG beamformer (50) whereas $\mathbf{h}_\mathbf{b}$ can be chosen as the MDF beamformer in (68). This would yield the global combined MWNG/MDF beamformer, which, at iteration n, is given by

$$\mathbf{h}_{\mathrm{MWNG/MDF}}^{(n)} = \mathbf{h}_{\mathbf{a},\mathrm{MWNG}}^{(n)} \otimes \mathbf{h}_{\mathbf{b},\mathrm{MDF}}^{(n)}. \quad (74)$$

D. Maximum Front-to-Back Ratio

We turn our attention to the FBR measure. Employing the ALS strategy, the maximum FBR (MFBR) beamformer is derived from

$$\max_{\mathbf{h}_\mathbf{a}} \mathcal{F}\left(\mathbf{h}_\mathbf{a} \,|\mathbf{h}_\mathbf{b}\right), \quad (75)$$

$$\max_{\mathbf{h}_\mathbf{b}} \mathcal{F}\left(\mathbf{h}_\mathbf{b} \,|\mathbf{h}_\mathbf{a}\right). \quad (76)$$

Let $\mathbf{t}_\mathbf{a}$ (resp. $\mathbf{t}_\mathbf{b}$) be the eigenvector corresponding to the maximum eigenvalue of $\mathbf{\Gamma}_{\mathrm{b},\mathbf{b}}^{-1}\mathbf{\Gamma}_{\mathrm{f},\mathbf{b}}$ (resp. $\mathbf{\Gamma}_{\mathrm{b},\mathbf{a}}^{-1}\mathbf{\Gamma}_{\mathrm{f},\mathbf{a}}$). Then, the solutions are

$$\mathbf{h}_{\mathbf{a},\mathrm{MFBR}} = \frac{\mathbf{t}_\mathbf{a}}{\left(\tau_0^P\right)^* \mathbf{a}_0^H \mathbf{t}_\mathbf{a}}, \quad (77)$$

$$\mathbf{h}_{\mathbf{b},\mathrm{MFBR}} = \frac{\mathbf{t}_\mathbf{b}}{\mathbf{b}_0^H \mathbf{t}_\mathbf{b}}, \quad (78)$$

where we took into account the distortionless constraints. We deduce that, at iteration n, the MFBR beamformer is

$$\mathbf{h}_{\mathrm{MFBR}}^{(n)} = \mathbf{h}_{\mathbf{a},\mathrm{MFBR}}^{(n)} \otimes \mathbf{h}_{\mathbf{b},\mathrm{MFBR}}^{(n)}, \quad (79)$$

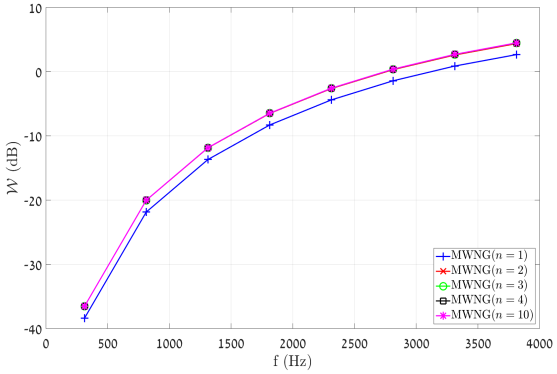
in which the appropriate coherence matrices are iteratively updated in an identical manner to the MWNG and MDF beamformers.

VI. Simulations

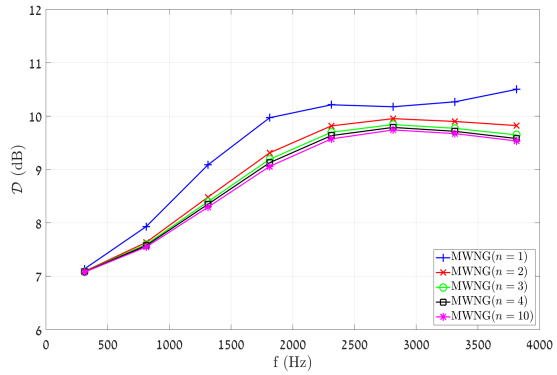
A. Performance Study

In this part, we investigate the performance of each of the differential KP beamformers presented in the former section in anechoic environments and with different settings.

Let us begin with the MWNG differential KP beamformer, $\mathbf{h}_{\mathrm{MWNG}}$. We note that as all other differential KP beamformers presented in this paper, it is derived iteratively. Therefore, it is important to get a notion of the convergence rate, i.e., to find a satisfactory value of the number of iterations n. Fig. 1 shows the WNG and DF measures with $(P = 2, M_\mathbf{a} = 4, M_\mathbf{b} = 2)$ for varying values of n and with initial sub-beamformers which implement the identity function: a one in the first element and zeros in the others. We observe that even for $n = 2$ convergence is achieved, whereas the performances with $n = 4$ and $n = 10$ nearly overlap. This result was verified with all other beamformers and with varying values of



(a)



(b)

Fig. 1: WNG and DF measures with the MWNG differential KP beamformer, $\mathbf{h}_{\mathrm{MWNG}}$, with a varying number of iterations n. Simulation parameters: $M = 10$, $(P = 2, M_\mathbf{a} = 4, M_\mathbf{b} = 2)$ and $\delta = 1$ cm. (a) WNG and (b) DF.

$(P, M_\mathbf{a}, M_\mathbf{b})$, to which we refer to from now on as the "array settings." Consequently, we will use the value of $n = 4$ in the rest.

Next, we explore the WNG and DF performance of $\mathbf{h}_{\mathrm{MWNG}}$ as a function of the array settings. These are demonstrated in Fig. 2. To begin with, we note that in the non-differential case ($P = 0$) the KP ($M_\mathbf{a} \neq 1$ and $M_\mathbf{b} \neq 1$) and non-KP ($M_\mathbf{a} = 1$ or $M_\mathbf{b} = 1$) settings result in similar beamformers whose performances perfectly overlap. This is with accordance to Section V-A as in this case $\mathbf{\Delta}_{(P)}\mathbf{\Delta}_{(P)}^T = \mathbf{I}_M$, $\mathbf{h}_{\mathrm{MWNG}}$ converges to the MWNG beamformer of [28] and $\mathcal{W}\left(\mathbf{h}_{\mathrm{MWNG}}\right) = M$ regardless of $M_\mathbf{a}$ and $M_\mathbf{b}$. As P increases, the DF performance improves but the WNG deteriorates. In addition, for a fixed value of P, the influence of $M_\mathbf{a}$ and $M_\mathbf{b}$ is clearly observed. That is, with the non-KP settings, a preferable WNG performance is attained, whereas with the KP settings, the DF performance is preferable. Nonetheless, we observe, for example, that with $(2, 4, 2)$ both measures are higher than with $(4, 1, 6)$ for frequencies above 2500 Hz. We infer that the parameter P has a more dominant influence on the WNG-DF performance trade-off, whereas $M_\mathbf{a}$ and $M_\mathbf{b}$ enable a more flexible, finer, tuning.

Next, we analyze the NS differential KP beamformer, $\mathbf{h}_{\mathrm{NS}}$.

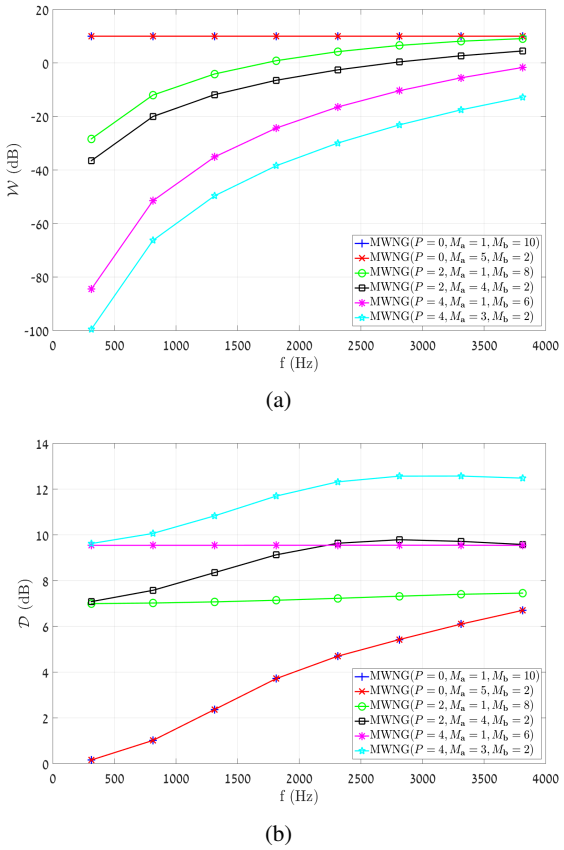


Fig. 2: WNG and DF measures with the MWNG differential KP beamformer, $\mathbf{h}_{\mathrm{MWNG}}$, with different array settings $(P, M_{\mathbf{a}}, M_{\mathbf{b}})$. Simulation parameters: $M = 10$ and $\delta = 1$ cm. (a) WNG and (b) DF.

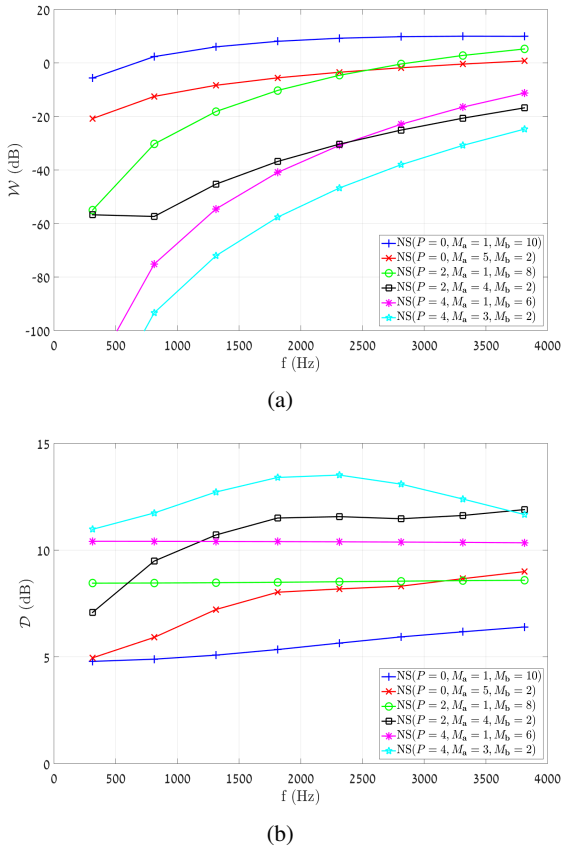
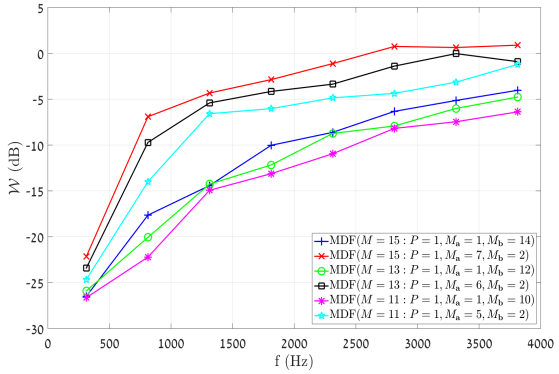


Fig. 3: WNG and DF measures with the NS differential KP beamformer, $\mathbf{h}_{\mathrm{NS}}$, with different array settings $(P, M_{\mathbf{a}}, M_{\mathbf{b}})$. Simulation parameters: $M = 10$ and $\delta = 1$ cm. (a) WNG and (b) DF.

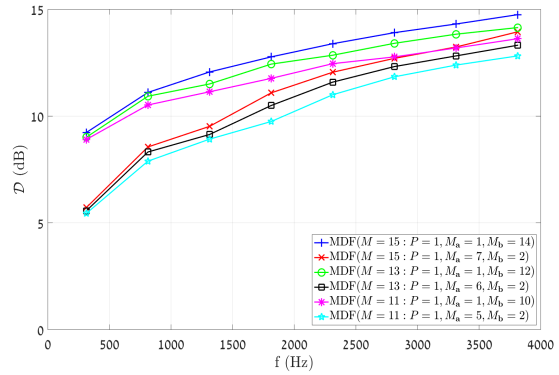
In our scenarios, the NS beamformer $\mathbf{h}_{\mathbf{b},\mathrm{NS}}$ is designed with a unique null constraint in the direction $\theta_{\mathrm{i}} = 90^o$. The corresponding WNG and DF measures are depicted in Fig. 3. To begin with, we observe that in contrast to the MWNG differential KP beamformer, with the NS differential KP beamformer the values of $M_{\mathbf{a}}$ and $M_{\mathbf{b}}$ set the performance trade-off for a fixed value of P even with $P = 0$. As with $\mathbf{h}_{\mathrm{MWNG}}$, the larger P the better the DF at the expense of the WNG, whereas the values of $M_{\mathbf{a}}$ and $M_{\mathbf{b}}$ tune the trade-off even further. That is, with the KP settings the DF performance is preferable with respect to the non-KP settings but the WNG is worse. Additionally, comparing the WNG performance of $\mathbf{h}_{\mathrm{MWNG}}$ and $\mathbf{h}_{\mathrm{NS}}$ for fixed settings, we note that the latter is lower as a consequence of the additional null constraint.

We now focus on the MDF differential KP beamformer, $\mathbf{h}_{\mathrm{MDF}}$, and the MWNG/MDF differential KP beamformer, $\mathbf{h}_{\mathrm{MWNG/MDF}}$, whose performances with arrays of $M = 11$, 13, and 15 microphones are shown in Fig. 4 and Fig. 5, respectively. As for high values of P white noise amplification was shown to be significant, we set $P = 1$ and examine the performance differences between the KP and non-KP versions of the differential beamformers. We point out that we use a small regularization factor of $\lambda = 10^{-2}$ when inverting the diffuse noise coherence matrix. Firstly, we note that with the non-KP settings $\mathbf{h}_{\mathrm{MDF}}$ and $\mathbf{h}_{\mathrm{MWNG/MDF}}$ are identical. Moreover, we observe that the two types of settings result in two distinct classes considering both performance measures, with the KP versions exhibiting a superior WNG performance to the non-KP versions but inferior DF performance. This separation is more dominant with $\mathbf{h}_{\mathrm{MWNG/MDF}}$ than with $\mathbf{h}_{\mathrm{MDF}}$, in particular in higher frequencies. In all cases, increasing M improves both performance measures.

Let us turn to the MFBR differential KP beamformer, $\mathbf{h}_{\mathrm{MFBR}}$. We maintain the same array settings used with $\mathbf{h}_{\mathrm{MDF}}$ and $\mathbf{h}_{\mathrm{MWNG/MDF}}$, and set a regularization factor of 10^{-2} with $\mathbf{\Gamma}_{\mathrm{b},\mathbf{a}}$ and $\mathbf{\Gamma}_{\mathrm{b},\mathbf{b}}$. As demonstrated in Fig. 6, the non-KP versions of $\mathbf{h}_{\mathrm{MFBR}}$ are of a better DF performance than its KP versions but of a worse WNG, particularly in high frequencies. The larger M, the lower the frequency above which the former is true. We deduce that with $\mathbf{h}_{\mathrm{MFBR}}$, in a similar manner to $\mathbf{h}_{\mathrm{MDF}}$ and $\mathbf{h}_{\mathrm{MWNG/MDF}}$, the flexibility in setting $M_{\mathbf{a}}$ and $M_{\mathbf{b}}$ allows one performance measure to improve at the expense of the other.

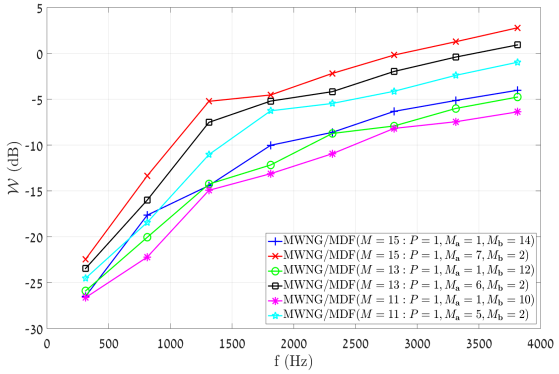


(a)

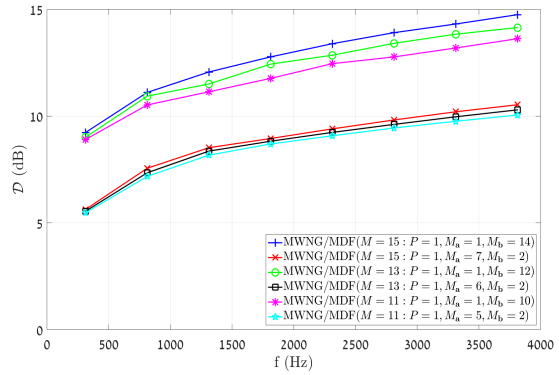


(b)

Fig. 4: WNG and DF measures with the MDF differential KP beamformer, $\mathbf{h}_{\mathrm{MDF}}$, with different array settings $(P, M_{\mathbf{a}}, M_{\mathbf{b}})$. Simulation parameters: $n = 4$, $\delta = 1$ cm and $\lambda = 10^{-2}$. (a) WNG and (b) DF.



(a)



(b)

Fig. 5: WNG and DF measures with the MWNG/MDF differential KP beamformer, $\mathbf{h}_{\mathrm{MWNG/MDF}}$, with different array settings $(P, M_{\mathbf{a}}, M_{\mathbf{b}})$ and a varying number of microphones M. Simulation parameters: $\delta = 1$ cm and $\lambda = 10^{-3}$. (a) WNG and (b) DF.

B. Speech Signals Simulations in Reverberant Environments

In this part, we demonstrate the performances of differential KP beamformers on speech signals in practical simulated scenarios and reverberant environments. We investigate and compare the performances of all the beamformers presented in the paper with four distinct array settings of a $M = 9$ microphone array.

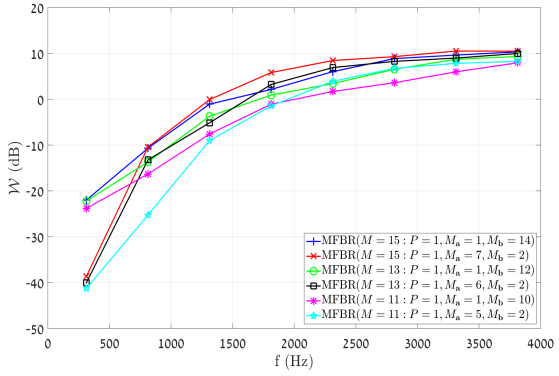
The reverberant simulations are performed as follows. We use a room impulse response (RIR) generator [35] to simulate the reverberant noise-free signal received in each of the microphones. The RIR generator is based on the image method of Allen and Berkley [36]. We simulate a $6 \times 6 \times 3$ m room in which a desired speech signal source is located at $(x, y, z) = (1, 1, 1.5)$ and an uncorrelated directional interference is located at $(x, y, z) = (3, 3, 1.5)$. The desired speech signal, $x(t)$, is a concatenation of 24 speech signals (12 speech signals per gender) with varying dialects that are taken from the TIMIT database [37]. It is sampled at a sampling rate of $f_{\mathrm{s}} = 1/T_{\mathrm{s}} = 16$ kHz within the signal duration T. A ULA consisting of $M = 9$ microphones is located on the $(1, y, 1.5)$-axis, with $y = 2.96 : 3.04$. In addition to the directional interference, two uncorrelated noise fields are present: a white thermal Gaussian noise and a spherically isotropic diffuse noise. The latter and the directional interference are equally powerful, whereas the white thermal noise is 20 dB weaker than each of them. Denoting the combined noise signal at the reference (first) microphone in the time domain by $v(t)$, we may define the time-domain SNR (which is identical to the broadband SNR) by

$$\mathrm{iSNR}_{\mathrm{t}} = \frac{\int_t x^2(t)dt}{\int_t v^2(t)dt}, \tag{80}$$

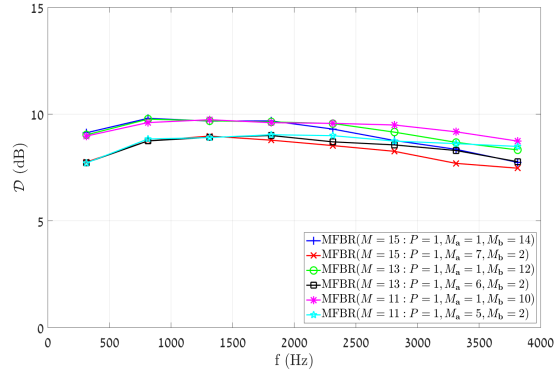
which is set to $\mathrm{iSNR}_{\mathrm{t}} = 0$ dB.

The noisy observations signal is transformed into the STFT domain using 75% overlapping time frames and a Hamming analysis window of length 256 (16 msec). Next, differential KP beamformers with different array settings are independently applied to the noisy signal to yield clean signal estimates in the STFT domain, followed by an inverse STFT procedure to obtain time-domain enhanced signals. The latter is carried out by using the overlap-and-add method.

We simulate three reverberant scenarios with $T_{60} \in \{130, 250, 400\}$ msec, where T_{60} is defined by Sabine-Franklin's formula [38]. In each scenario, we design each of the five beamformers presented in the paper with five different



(a)



(b)

Fig. 6: WNG and DF measures with the MFBR differential KP beamformer, $\mathbf{h}_{\mathrm{MFBR}}$, with different array settings $(P, M_{\mathbf{a}}, M_{\mathbf{b}})$. Simulation parameters: $n = 4$ and $\delta = 1$ cm. (a) WNG and (b) DF.

array settings: $(0, 1, 9)$- a "classical" non-differential non-KP beamformer, $(0, 3, 3)$- a non-differential KP beamformer [27], $(1, 1, 8)$- a differential non-KP beamformer [24], and $(1, 2, 4)$ and $(1, 4, 2)$- two distinct differential KP beamformers. The NS differential KP beamformer $\mathbf{h}_{\mathrm{NS}}$ is designed with a single unique null in the direction of the directional interference.

Next, we are interested in objectively quantifying the performance of each of the five beamformers with the five aforementioned array settings. We shall do that by individually examining the power ratio of the following components of the time-domain enhanced signals and the noisy observation signals of the first microphone: the white thermal Gaussian noise, the diffuse noise, the reverberant directional interference and the desired speech signal reverberations. Formulating the noisy observation signal in the time-domain in microphone m, we have

$$\begin{aligned} y_m &= x_{\mathrm{d}} * g_{\mathrm{d},m} + v_{\mathrm{i}} * g_{\mathrm{i},m} + v_{\mathrm{d},m} + v_{\mathrm{w},m} \\ &= x_m + x_{\mathrm{r},m} + v_{\mathrm{r},m} + v_{\mathrm{d},m} + v_{\mathrm{w},m}, \end{aligned} \tag{81}$$

where $*$ is the linear convolution operator, x_{d} is the desired speech signal, v_{i} is the directional interference, $v_{\mathrm{d},m}$ and $v_{\mathrm{w},m}$ are, respectively, the additive diffuse noise and white noise in microphone m, $g_{\mathrm{d},m}$ is the RIR from the desired signal to microphone m, $g_{\mathrm{i},m}$ is the RIR from the directional interference to microphone m, and x_m, $x_{\mathrm{r},m}$, and $v_{\mathrm{r},m}$ are the direct path desired signal, its reverberations and the reverberant interference as received in microphone m, respectively. Additionally, we define the same components of the second row of (81) with respect to the time-domain enhanced speech signals by using the subscript f. For example, x_{f} is the enhanced direct path desired signal and $v_{\mathrm{w,f}}$ is the white noise component in the enhanced speech signal. We note that the energies of the direct path desired signal component in the first microphone and the enhanced direct path desired signal were verified to equal up to a scale of 0.1 dB in all scenarios and with all the beamformers.

We now address the white thermal noise and the diffuse noise. Using the notations of (81), we define the diffuse noise reduction (DNR) factor by

$$\mathrm{DNR} = \frac{E\left[v_{\mathrm{d},1}^2\right]}{E\left[v_{\mathrm{d,f}}^2\right]} \tag{82}$$

and the white noise reduction (WNR) factor by

$$\mathrm{WNR} = \frac{E\left[v_{\mathrm{w},1}^2\right]}{E\left[v_{\mathrm{w,f}}^2\right]}. \tag{83}$$

The DNR and WNR with each of the presented beamformers and the discussed array settings appear in Table I. To begin with, we note that, broadly, the $(0, 1, 9)$- and $(0, 3, 3)$-beamformers attain a better (higher) WNR but a worse (lower) DNR (except with $\mathbf{h}_{\mathrm{MDF}}$). Clearly, this is due to their non-differential nature. Focusing on the three other array settings, we observe that the KP beamformers exhibit a preferable WNR with $\mathbf{h}_{\mathrm{MDF}}$, $\mathbf{h}_{\mathrm{MWNG/MDF}}$, and $\mathbf{h}_{\mathrm{MFBR}}$, hence mitigating the white noise amplification issue of differential beamformers. This is with accordance to the WNG performance addressed in the previous sub-section and is considerably more significant with $(1, 4, 2)$. On the other hand, the DNR with $(1, 1, 8)$ and these three beamformers is equal to or better than with the KP settings. Turning to the $\mathbf{h}_{\mathrm{MWNG}}$ and $\mathbf{h}_{\mathrm{NS}}$, we observe that the KP beamformers exhibit a preferable DNR but worse WNR, with the performance gaps being more significant with the latter beamformer. We infer that the differential KP approach constitute a mean to flexibly tune the desirable performance trade-off of mitigating the white noise amplification and improving the diffuse noise attenuation.

Next, we define the desired signal reverberations reduction (RR) factor by

$$\mathrm{RR} = \frac{E\left[x_{\mathrm{r},1}^2\right]}{E\left[x_{\mathrm{r,f}}^2\right]} \tag{84}$$

and the interference reduction (IR) factor by

$$\mathrm{IR} = \frac{E\left[v_{\mathrm{r},1}^2\right]}{E\left[v_{\mathrm{r,f}}^2\right]}. \tag{85}$$

We note that both factors are a function of the RIRs, as opposed to the DNR and WNR. The RR and IR with the discussed beamformers and array settings for $T_{60} = 130$ msec

are shown in Table II. It is stressed that except for with $\mathbf{h}_{\mathrm{NS}}$, the differential beamformers attain a significantly higher IR than their non-differential counterparts. In addition, with $\mathbf{h}_{\mathrm{NS}}$, the latter measure is considerably higher with $(1,2,4)$ and $(1,4,2)$ than with $(1,1,8)$, whereas with the other beamformers the IRs are roughly equal. Focusing on the RR, the differential beamformers exhibit a preferable performance, with the $(1,4,2)$ settings attaining the highest value with $\mathbf{h}_{\mathrm{NS}}$ and the $(1,1,8)$ settings attaining the highest values with $\mathbf{h}_{\mathrm{MDF}}$, $\mathbf{h}_{\mathrm{MWNG/MDF}}$, and $\mathbf{h}_{\mathrm{MFBR}}$. The RR performance with $\mathbf{h}_{\mathrm{MWNG}}$ is equal with the three differential settings.

We turn to the $T_{60} = 250$ msec scenario for which the RR and IR are shown in Table III. Firstly, we observe that, in general, the RR increases and the IR decreases in comparison to the former scenario. This may be explained as follows. In a non-reverberant scenario, the RR is zero (which is low) and it is straightforward to attain a very high value of the IR (for example, by placing a null in the appropriate direction). As T_{60} increases, the desired signal reverberations are more paramount and, therefore, there is more room for a beamformer to reject these reverberations, implying a higher possible RR value. On the other hand, as T_{60} increases, the reverberations of the directional interference are greater scattered across the azimuth angle. Thus, the task of rejecting the spatial interference turns more complex. In spite of this observation, we note that the performance analysis of the previous scenario remains valid, with the values of the RR and IR change as explained but their corresponding differences with varying settings and fixed beamformers are maintained.

Lastly, we relate the $T_{60} = 400$ msec scenario of Table IV. We observe that indeed, according to the formerly discussed trend, the RR values are higher but the IR values are lower. In addition, the performance analysis of the two previous scenarios applies for this scenario as well.

We end this part by examining the same three reverberant scenarios discussed above from a different perspective. That is, we analyze the average PESQ [39] and STOI [40] scores of the time-domain enhanced speech signals. The results are depicted in Fig 7. It is clear that in terms of the PESQ score the $(1,4,2)$-beamformers are superior to their counterparts, excluding the $\mathbf{h}_{\mathrm{MWNG}}$. Particularly, the performance gap is the most significant in the $T_{60} = 130$ msec scenario. We observe that in the three scenarios $\mathbf{h}_{\mathrm{NS}}$ exhibits the highest overall PESQ score with the aforementioned settings, which is of a great contrast to the low PESQ scores of the same beamformer with $(1,1,8)$. Considering the reduction factors from the previous part, these results may be explained as follows. When T_{60} is low, the reverberations of the desired signal and interference are less significant. Hence, the white noise amplification becomes more considerable. As it is mitigated with the KP settings with $\mathbf{h}_{\mathrm{MDF}}$, $\mathbf{h}_{\mathrm{MWNG/MDF}}$, and $\mathbf{h}_{\mathrm{MFBR}}$ with respect to the non-KP settings, their corresponding time-domain enhanced signals are of a higher quality. On the contrary, while with the $(1,4,2)$ version of $\mathbf{h}_{\mathrm{NS}}$ the white noise is even greater amplified, its preferable DNR, RR, and IR values with respect to most or all of its counterparts result in higher quality enhanced signals. Focusing on the STOI scores in the $T_{60} = 130$ msec scenario, we observe that indeed, the beamformers with the differential KP settings, that is, $(1,2,4)$ and $(1,4,2)$, are of a better intelligibility, with the latter outperforming the former. Increasing T_{60} to 250 msec reduces the performance gap, whereas with $T_{60} = 400$ msec the STOI scores are roughly equal with most combinations of beamformers and array settings.

Considering all the presented beamformers and array settings, it is beneficial to conclude by proposing some design rules of thumbs. For starters, one must choose an appropriate beamformer type. For example, in case powerful directional interferences are present $\mathbf{h}_{\mathrm{NS}}$ is likely to be preferred, and in case it is desirable to substantially attenuate the array response in the rear-half plane $\mathbf{h}_{\mathrm{MFBR}}$ should be chosen. Alternatively, if the top-priority attribute of the array is white-noise attenuation $\mathbf{h}_{\mathrm{MWNG}}$ is an appropriate choice, whereas $\mathbf{h}_{\mathrm{MWNG/MDF}}$ and $\mathbf{h}_{\mathrm{MDF}}$ should be considered if the array directivity is of the highest significance. Then, one should determine the array settings, with P being the more dominant parameter, whereas $M_{\mathbf{a}}$ and $M_{\mathbf{b}}$ enable a finer tuning. That is, the value of P sets a level of array directivity at the expanse of white-noise sensitivity. Depending on the selection of beamformer type, the values of $M_{\mathbf{a}}$ and $M_{\mathbf{b}}$ may either improve the array directivity or white-noise attenuation. Nevertheless, the high flexibility of this approach requires a careful design, as it is not guaranteed that every differential KP combination of the array settings $(P, M_{\mathbf{a}}, M_{\mathbf{b}})$, that is, $P > 0$ and $M_{\mathbf{a}}, M_{\mathbf{b}} > 1$, would yield a better choice than non-differential ($P = 0$) or non-KP ($M_{\mathbf{a}} = 1$ or $M_{\mathbf{b}} = 1$) settings.

C. Reverberant Simulations with Array Imperfections

In this part, we focus on the effects of two common types of array imperfections: deviation of the speech source incident angle and misplacements of the array microphones. We maintain the same simulation settings of Section VI-B, set $T_{60} = 130$ msec, and analyze the performance of a subset of the previously discussed beamformers with two array sizes.

Let us describe the two types of array imperfections. To begin with, we move the desired speech signal source along the x-axis to generate an incident angle of $\theta_{\mathrm{s}} = 10^{o}$. In addition, we simulate microphones misplacements by adding independent normally-distributed values of zero mean and standard deviation of 1 mm to each of the microphones positions along the y-axis. Then, we compare the three following scenarios.

- Scen. (a): no array imperfections (reference scenario).
- Scen. (b): microphones misplacements but true speech source incident angle.
- Scen. (c): microphones misplacements and deviation of the speech source incident angle.

We examine the PESQ and STOI scores in each of these scenarios with $\mathbf{h}_{\mathrm{MWNG}}$, $\mathbf{h}_{\mathrm{MDF}}$ and $\mathbf{h}_{\mathrm{MWNG/MDF}}$ with two array sizes: $M = 5$ and $M = 9$. In each, we show examples of non-differential non-KP beamformers, differential non-KP beamformers and differential KP beamformers. We note that with the $M = 9$ array we maintain the same array settings as above and show a subset of three out of five of them: $(0,1,9)$, $(1,1,8)$, and $(1,4,2)$. Additionally, we employ a

TABLE I: The DNR and WNR for Differential KP Beamformers with Different Array Settings $(P, M_{\mathbf{a}}, M_{\mathbf{b}})$. Simulation Parameters: $M = 9$, $\delta = 1$ cm, $n = 4$, and $\mathrm{iSNR}_{\mathrm{t}} = 0$ dB.

	DNR (dB)					WNR (dB)				
Settings	$(\mathbf{0,1,9})$	$(\mathbf{0,3,3})$	$(\mathbf{1,1,8})$	$(\mathbf{1,2,4})$	$(\mathbf{1,4,2})$	$(\mathbf{0,1,9})$	$(\mathbf{0,3,3})$	$(\mathbf{1,1,8})$	$(\mathbf{1,2,4})$	$(\mathbf{1,4,2})$
$\mathbf{h}_{\mathrm{MWNG}}$	4.6	4.6	6.3	7.0	6.8	**9.5**	**9.5**	$\mathbf{-1.4}$	$\mathbf{-3.5}$	$\mathbf{-3.0}$
$\mathbf{h}_{\mathrm{NS}}$	6.3	7.3	7.5	8.5	10.1	-1.4	-11.0	-22.9	-26.0	-28.6
$\mathbf{h}_{\mathrm{MDF}}$	**10.7**	**10.4**	**11.8**	**11.1**	**10.4**	-2.4	-4.6	-23.4	-21.7	-15.4
$\mathbf{h}_{\mathrm{MWNG/MDF}}$	**10.7**	7.9	**11.8**	10.4	8.9	-2.4	-0.9	-23.4	-21.4	-15.6
$\mathbf{h}_{\mathrm{MFBR}}$	5.5	4.7	8.0	8.0	7.4	2.8	2.3	-24.7	-23.8	-17.9

TABLE II: The RR and IR for Differential KP Beamformers with Different Array Settings $(P, M_{\mathbf{a}}, M_{\mathbf{b}})$. Simulation Parameters: $T_{60} = 130$ msec, $M = 9$, $\delta = 1$ cm, $n = 4$, and $\mathrm{iSNR}_{\mathrm{t}} = 0$ dB.

	RR (dB)					IR (dB)				
Settings	$(\mathbf{0,1,9})$	$(\mathbf{0,3,3})$	$(\mathbf{1,1,8})$	$(\mathbf{1,2,4})$	$(\mathbf{1,4,2})$	$(\mathbf{0,1,9})$	$(\mathbf{0,3,3})$	$(\mathbf{1,1,8})$	$(\mathbf{1,2,4})$	$(\mathbf{1,4,2})$
$\mathbf{h}_{\mathrm{MWNG}}$	0.2	0.2	3.2	3.3	3.3	6.2	6.2	**20.7**	19.7	20.4
$\mathbf{h}_{\mathrm{NS}}$	3.2	3.6	5.5	3.9	**6.4**	**20.7**	**21.1**	17.5	20.5	20.0
$\mathbf{h}_{\mathrm{MDF}}$	**4.0**	**4.4**	**5.8**	**5.3**	4.4	16.1	15.1	**20.7**	**20.9**	**21.0**
$\mathbf{h}_{\mathrm{MWNG/MDF}}$	**4.0**	2.4	**5.8**	**5.3**	4.2	16.1	12.6	**20.7**	20.6	20.5
$\mathbf{h}_{\mathrm{MFBR}}$	2.6	2.3	5.0	4.8	4.2	14.3	13.0	19.6	19.8	19.9

TABLE III: The RR and IR for Differential KP Beamformers with Different Array Settings $(P, M_{\mathbf{a}}, M_{\mathbf{b}})$. Simulation Parameters: $T_{60} = 250$ msec, $M = 9$, $\delta = 1$ cm, $n = 4$, and $\mathrm{iSNR}_{\mathrm{t}} = 0$ dB.

	RR (dB)					IR (dB)				
Settings	$(\mathbf{0,1,9})$	$(\mathbf{0,3,3})$	$(\mathbf{1,1,8})$	$(\mathbf{1,2,4})$	$(\mathbf{1,4,2})$	$(\mathbf{0,1,9})$	$(\mathbf{0,3,3})$	$(\mathbf{1,1,8})$	$(\mathbf{1,2,4})$	$(\mathbf{1,4,2})$
$\mathbf{h}_{\mathrm{MWNG}}$	0.5	0.5	4.2	4.4	4.3	3.5	3.5	9.7	8.6	9.6
$\mathbf{h}_{\mathrm{NS}}$	4.2	4.7	6.1	**6.8**	**7.9**	**9.7**	**10.2**	6.8	10.6	9.9
$\mathbf{h}_{\mathrm{MDF}}$	**5.4**	**5.7**	**7.1**	6.7	5.7	9.3	9.0	**10.9**	**11.0**	**11.2**
$\mathbf{h}_{\mathrm{MWNG/MDF}}$	**5.4**	3.5	**7.1**	6.6	5.4	9.3	7.0	**10.9**	10.7	10.2
$\mathbf{h}_{\mathrm{MFBR}}$	3.8	3.5	6.3	6.2	5.4	6.3	5.6	8.7	9.0	9.0

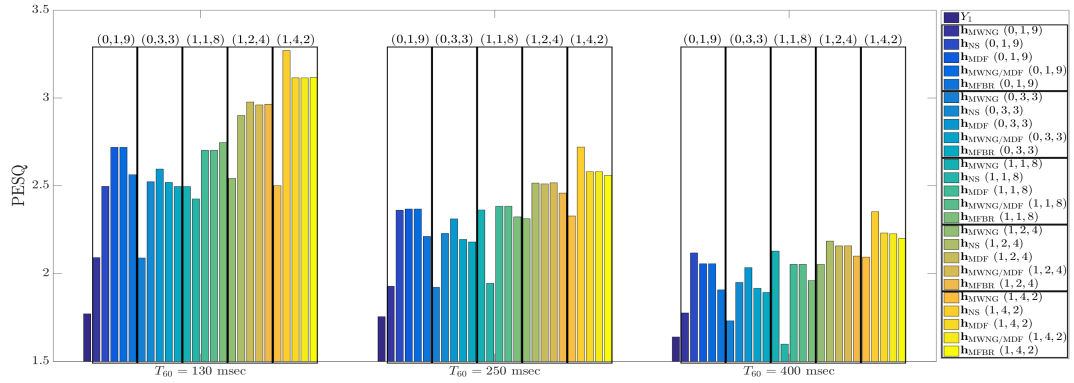
TABLE IV: The RR and IR for Differential KP Beamformers with Different Array Settings $(P, M_{\mathbf{a}}, M_{\mathbf{b}})$. Simulation Parameters: $T_{60} = 400$ msec, $M = 9$, $\delta = 1$ cm, $n = 4$, and $\mathrm{iSNR}_{\mathrm{t}} = 0$ dB.

	RR (dB)					IR (dB)				
Settings	$(\mathbf{0,1,9})$	$(\mathbf{0,3,3})$	$(\mathbf{1,1,8})$	$(\mathbf{1,2,4})$	$(\mathbf{1,4,2})$	$(\mathbf{0,1,9})$	$(\mathbf{0,3,3})$	$(\mathbf{1,1,8})$	$(\mathbf{1,2,4})$	$(\mathbf{1,4,2})$
$\mathbf{h}_{\mathrm{MWNG}}$	0.6	0.6	5.1	5.4	5.3	1.8	1.8	6.8	5.8	6.4
$\mathbf{h}_{\mathrm{NS}}$	5.1	5.7	6.6	7.7	**8.8**	**6.6**	**7.3**	4.4	7.5	6.8
$\mathbf{h}_{\mathrm{MDF}}$	**6.6**	**6.9**	**8.3**	**7.8**	6.8	6.2	6.0	**8.0**	**7.9**	**7.7**
$\mathbf{h}_{\mathrm{MWNG/MDF}}$	**6.6**	4.4	**8.3**	**7.8**	6.5	6.2	4.5	**8.0**	7.7	7.0
$\mathbf{h}_{\mathrm{MFBR}}$	4.9	4.6	7.6	7.4	6.5	3.5	2.9	5.9	6.2	5.8

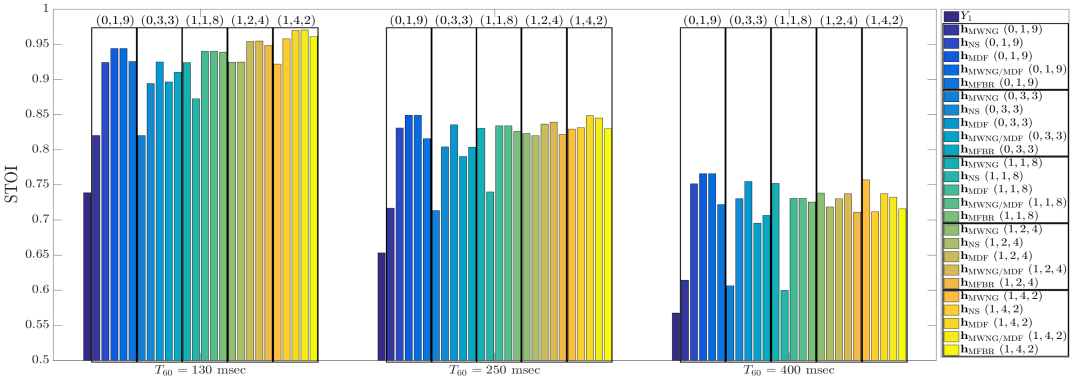
$M = 5$ array (formed by dropping the the first-two and last-two microphones of the $M = 9$ array) as a practical example for small-size arrays. With the latter, we use the following array settings: $(0, 1, 5)$, $(1, 1, 4)$, and $(1, 2, 2)$.

The PESQ and STOI scores with the $M = 5$ and $M = 9$ arrays are shown, respectively, in Fig. 8 and Fig. 9. To begin with, we observe that in the reference scenario, i.e., Scen. (a), and with both arrays, the beamformers with the differential KP settings outperform their counterparts in terms of both the PESQ and STOI scores. In particular, the performance gap is accentuated with $\mathbf{h}_{\mathrm{MDF}}$ and $\mathbf{h}_{\mathrm{MWNG/MDF}}$. As we consider the array imperfections of Scen. (b) and Scen. (c), we note that while the performances of all the presented beamformers slightly degrade- the performance gap remains. We infer that even with small arrays and practical imperfections the differential KP approach may outperform the rest of the discussed approaches.

Finally, for the sake of completeness, we address the beampattern measure. As an example, we focus on the MWNG differential KP beamformer, $\mathbf{h}_{\mathrm{MWNG}}$, with the six combinations of array settings and sizes discussed in this part. The beampatterns are plotted in Fig. 10. We observe that with the $M = 5$ array the $(0, 1, 5)$ version of $\mathbf{h}_{\mathrm{MWNG}}$ exhibits a supercardioid-like shape, whereas its $(1, 1, 4)$ and $(1, 2, 2)$ versions exhibit a dipole-like shape. Nevertheless, we observe that the back lobe with the latter is roughly 5 dB lower than with the former, indicating a preferable directivity. Examining the $M = 9$ array, we note that the beampatterns of the $(0, 1, 9)$ and $(1, 1, 8)$ versions are, roughly, similarly-shaped, but the latter offers a preferable directivity. On the contrary, the



(a)



(b)

Fig. 7: Average PESQ and STOI scores of enhanced speech signals with the five differential KP beamformers presented in the paper and with five different array settings $(P, M_{\mathbf{a}}, M_{\mathbf{b}})$. Simulation parameters: $M = 9$, $\delta = 1$ cm, and $\mathrm{iSNR}_{\mathrm{t}} = 0$ dB. (a) PESQ scores and (b) STOI scores.

$(1, 4, 2)$ version exhibits a different beampattern shape whose side lobe is significantly lower than the side lobes of the over versions- a paramount attribute in scenarios of substantial early reverberations.

VII. Conclusions

We have generalized the multistage differential beamforming approach by applying a KP decomposition to a global differential beamformer, and independently optimizing the two sub-beamformers. Previous non-differential or non-KP beamformers may be obtained by an appropriate selection of the array settings parameters. We proposed five types of differential KP beamformers and demonstrated that each may perform better than previous approaches in terms of the WNG or DF measure at the expense of the complementary measure, depending on the array settings. This flexibility enables one to mitigate the white noise amplification with the differential MDF, MWNG/MDF, and MFBR beamformers or improve the directivity with the differential MWNG and NS beamformers. In addition, we showed that signal reverberations are attenuated to the greatest extent using the NS differential KP beamformer, whereas reverberations of a directional interference are equally attenuated using the other beamformers with differential KP and differential non-KP settings. Finally, we examined the average PESQ and STOI scores of the respective time-domain enhanced signals and demonstrated that both are higher with the new approach, even under array imperfections. This is in particular true for moderately reverberant environments.

Acknowledgement

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

References

[1] G. W. Elko, "Microphone array systems for hands-free telecommunication," *Speech Communication*, vol. 20, no. 3, pp. 229 – 240, 1996, Acoustic Echo Control and Speech Enhancement Techniques.

[2] H.L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*, Detection, Estimation, and Modulation Theory. Wiley, 2004.

[3] Y. Buchris, I. Cohen, and J. Benesty, "Frequency-Domain Design of Asymmetric Circular Differential Microphone Arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 760–773, 2018.

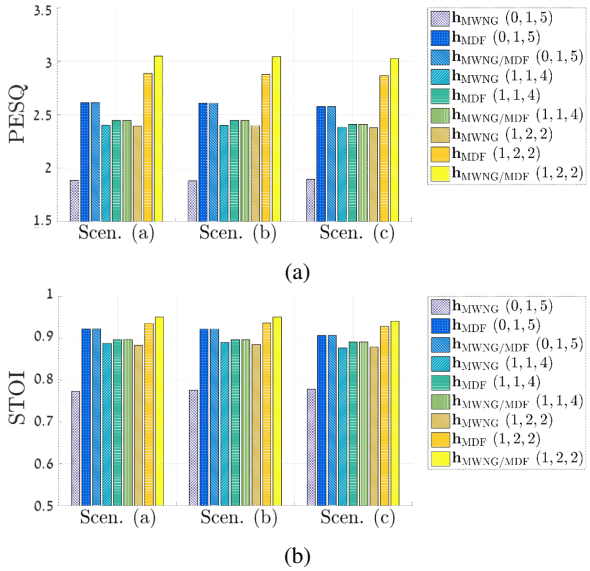


Fig. 8: Average PESQ and STOI scores of enhanced speech signals in the three array imperfection scenarios described in the paper. We simulate three differential KP beamformers with three different array settings $(P, M_{\mathbf{a}}, M_{\mathbf{b}})$. Simulation parameters: $M = 5$, $T_{60} = 130$ msec, $\delta = 1$ cm, and $\text{iSNR}_{\text{t}} = 0$ dB. (a) PESQ scores and (b) STOI scores.

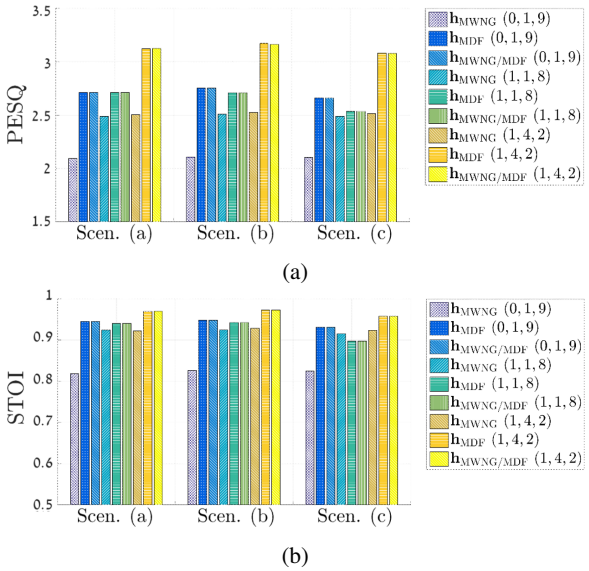


Fig. 9: Average PESQ and STOI scores of enhanced speech signals in the three array imperfection scenarios described in the paper. We simulate three differential KP beamformers with three different array settings $(P, M_{\mathbf{a}}, M_{\mathbf{b}})$. Simulation parameters: $M = 9$, $T_{60} = 130$ msec, $\delta = 1$ cm, and $\text{iSNR}_{\text{t}} = 0$ dB. (a) PESQ scores and (b) STOI scores.

[4] G. Huang, J. Chen, and J. Benesty, "On the design of differential beamformers with arbitrary planar microphone array geometry," *The Journal of the Acoustical Society of America*, vol. 144, no. 1, pp. EL66–EL70, 2018.

[5] J. Weinberger, H.F. Olson, and F. Massa, "A uni-directional ribbon microphone," *The Journal of the Acoustical Society of America*, vol. 5, no. 2, pp. 139–147, 1933.

[6] H. F. Olson, "Gradient microphones," *The Journal of the Acoustical Society of America*, vol. 17, no. 3, pp. 192–198, 1946.

[7] G. W. Elko, *Differential Microphone Arrays*, pp. 11–65, Springer US, Boston, MA, 2004.

[8] M. Kolundzija, C. Faller, and M. Vetterli, "Spatiotemporal gradient analysis of differential microphone arrays," *journal of the audio engineering society*, vol. 59, no. 1, pp. 20–28, Jan. 2011.

[9] Y. Buchris, I. Cohen, and J. Benesty, "On the design of time-domain differential microphone arrays," *Applied Acoustics*, vol. 148, pp. 212 – 222, 2019.

[10] J. Benesty, I. Cohen, and J. Chen, *Array Beamforming with Linear Difference Equations*, Springer, 2021.

[11] J. Jin, G. Huang, X. Wang, J. Chen, J. Benesty, and I. Cohen, "Steering study of linear differential microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 158–170, 2021.

[12] S. Gannot and I. Cohen, *Adaptive Beamforming and Postfiltering*, pp. 945–978, Springer Berlin Heidelberg, 2008.

[13] J. Benesty and J. Chen, *Study and Design of Differential Microphone Arrays*, Springer-Verlag Berlin Heidelberg, 2013.

[14] T. Long, J. Benesty, J. Chen, and I. Cohen, "Differential beamformers derived from approximate performance measures," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 66–70.

[15] A. Bernardini, F. Antonacci, and A. Sarti, "Wave digital implementation of robust first-order differential microphone arrays," *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 253–257, 2018.

[16] F. Borra, A. Bernardini, F. Antonacci, and A. Sarti, "Uniform linear arrays of first-order steerable differential microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 1906–1918, 2019.

[17] G. W. Elko and J. Meyer, *Microphone arrays*, pp. 1021–1041, Springer Berlin Heidelberg, 2008.

[18] E. De Sena, H. Hacihabiboglu, and Z. Cvetkovic, "On the Design and Implementation of Higher Order Differential Microphones," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 162–174, 2012.

[19] M. Buck, "Aspects of first-order differential microphone arrays in the presence of sensor imperfections," *European Transactions on Telecommunications*, vol. 13, pp. 115–122, 2002.

[20] X. Wu, H. Chen, J. Zhou, and T. Guo, "Study of the mainlobe misorientation of the first-order steerable differential array in the presence of microphone gain and phase errors," *IEEE Signal Processing Letters*, vol. 21, no. 6, pp. 667–671, 2014.

[21] X. Wu and H. Chen, "Directivity Factors of the First-Order Steerable Differential Array With Microphone Mismatches: Deterministic and Worst-Case Analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 300–315, 2016.

[22] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*, Springer-Verlag Berlin Heidelberg, Berlin, 1st edition, 2009.

[23] S. He and H. Chen, "Closed-form DOA estimation using first-order differential microphone arrays via joint temporal-spectral-spatial processing," *IEEE Sensors Journal*, vol. 17, no. 4, pp. 1046–1060, 2017.

[24] G. Huang, J. Benesty, I. Cohen, and J. Chen, "A simple theory and new method of differential beamforming with uniform linear microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1079–1093, 2020.

[25] Y. I. Abramovich, G. J. Frazer, and B. A. Johnson, "Iterative Adaptive Kronecker MIMO Radar Beamformer: Description and Convergence Analysis," *IEEE Transactions on Signal Processing*, vol. 58, no. 7, pp. 3681–3691, 2010.

[26] L. N. Ribeiro, A. L. F. de Almeida, and J. C. M. Mota, "Tensor beamforming for multilinear translation invariant arrays," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2966–2970.

[27] J. Benesty, I. Cohen, and J. Chen, *Array Processing - Kronecker Product Beamforming*, Springer-Verlag, Switzerland, 2019.

[28] I. Cohen, J. Benesty, and J. Chen, "Differential kronecker product beamforming," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 5, pp. 892–902, 2019.

[29] G. Huang, J. Benesty, J. Chen, and I. Cohen, "Robust and steerable kronecker product differential beamforming with rectangular microphone arrays," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 211–215.

[30] G. Huang, I. Cohen, J. Benesty, and J. Chen, "Kronecker product beamforming with multiple differential microphone arrays," in *2020*

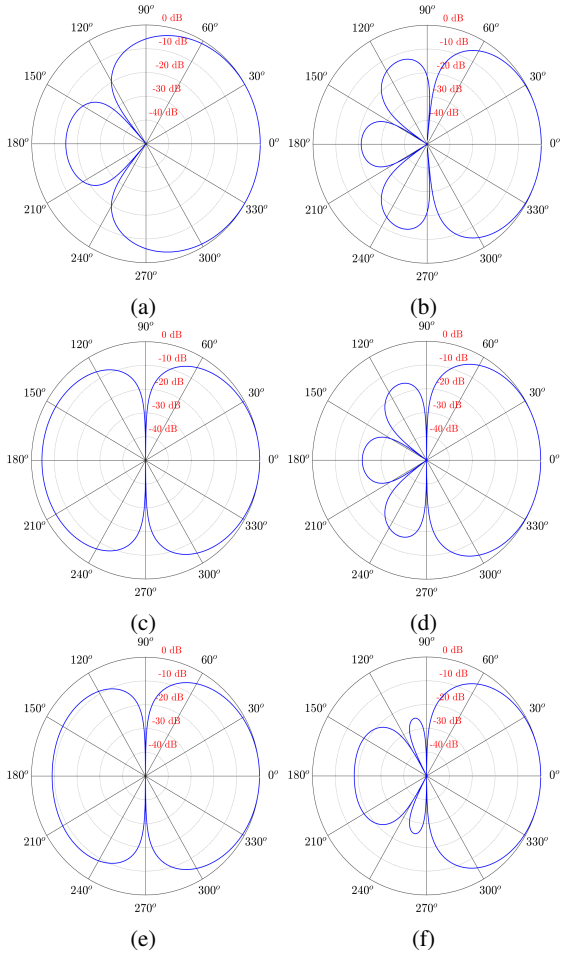


Fig. 10: Beampatterns of the MWNG differential KP beamformer, $\mathbf{h}_{\text{MWNG}}$, with six different array settings $(P, M_{\mathbf{a}}, M_{\mathbf{b}})$, and two values of M. Simulation parameters: $f = 2$ kHz and $\delta = 1$ cm. (a) $M = 5$; $(P = 0, M_{\mathbf{a}} = 1, M_{\mathbf{b}} = 5)$, (b) $M = 9$; $(P = 0, M_{\mathbf{a}} = 1, M_{\mathbf{b}} = 9)$, (c) $M = 5$; $(P = 1, M_{\mathbf{a}} = 1, M_{\mathbf{b}} = 4)$, (d) $M = 9$; $(P = 1, M_{\mathbf{a}} = 1, M_{\mathbf{b}} = 8)$; (e) $M = 5$; $(P = 1, M_{\mathbf{a}} = 2, M_{\mathbf{b}} = 2)$, and (f) $M = 9$; $(P = 1, M_{\mathbf{a}} = 4, M_{\mathbf{b}} = 2)$.

IEEE 11th Sensor Array and Multichannel Signal Processing Workshop (SAM), 2020, pp. 1–5.

[31] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Simon and Schuster, Inc., USA, 1992.

[32] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*, Wiley-IEEE Press, Singapore, 2018.

[33] F. Yates, "The Analysis of Replicated Experimetns when the Field Results are Incomplete," *Empire Journal of Experimental Agriculture*, vol. 1, no. 2, pp. 129–142, 1933.

[34] H. Wold and F. Lyttkens, "Nonlinear iterative partial least squares (NIPALS) estimation procedures," *Bulletin of the International Statistical Institute*, vol. 43, no. 1, pp. 29–47, 1969.

[35] E. A. P. Habets, "Rir-generator," https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator.

[36] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am*, vol. 65, no. 4, pp. 943–950, 1979.

[37] "Darpa timit acoustic phonetic continuous speech corpus cdrom," 1993.

[38] A. Pierce, *Acoustics: An Introduction to Its Physical Principles and Applications*, Acoustical Society of America, 1991.

[39] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, 2001, vol. 2, pp. 749–752 vol.2.

[40] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep 2011.

Chapter 6

Multistage Approach for Steerable Differential Beamforming with Rectangular Arrays (Unpublished)

Multistage Approach for Steerable Differential Beamforming with Rectangular Arrays

Gal Itzhak[a,*], Jacob Benesty[b], Israel Cohen[a]

[a]*Andrew and Erna Viterby Faculty of Electrical and Computer Engineering, Technion – Israel Institute of Technology, Technion City, Haifa 3200003, Israel*
[b]*INRS-EMT, University of Quebec, 800 de la Gauchetiere Ouest, Montreal, QC H5A 1K6, Canada*

Abstract

This paper presents a multistage rectangular approach for steerable differential beamforming. As a first step, we propose employing a two-dimensional (2-D) differentiation scheme that operates independently on the columns and rows of a uniform rectangular array (URA). This yields a differentials matrix controlled by two parameters, P_c and P_r, which indicate the number of differential stages for the URA columns and rows. Then, as a second step, we design a rectangular differential beamformer and apply it to the vector form of the differentials matrix. We show that the first differentiation scheme may significantly improve the directivity of the resulted beamformer at the expense of white noise amplification. The latter is heavily tied to selecting the (P_c, P_r) configuration optimized for the desired signal incident angle. Next, we propose four rectangular differential beamformers and analyze their performances in terms of the white noise gain (WNG) and directivity factor (DF) measures. Finally, we address reverberant scenarios with three distinct incident angles of the desired signal. We examine the performances of each beamformer in terms of four reduction factors that are calculated from the noisy and enhanced signals and investigate their quality and intelligibility. We demonstrate that the proposed rectangular differential beamformers outperform common linear differential beamformers in terms of these measures, mainly when the incident angle is far from the endfire direction.

Keywords: Microphone arrays, uniform rectangular arrays (URAs), differential beamforming, two-dimensional (2-D) arrays.

1. Introduction

Observation signals in communication systems are very likely to be degraded by undesired noise and reverberations, which might heavily damage the performance of such systems. These include, for example, speech and audio appliances, internet-of-things devices, and larger-scale systems as radars and sonars. Sensor arrays are employed to simultaneously capture samples in different locations in space while aiming to attenuate undesirable artifacts. These samples feed spatial filters, typically referred to as "beamformers". Beamformers, and their properties in time, frequency and space, have been widely studied and optimized for different criteria and geometric configurations [1, 2, 3].

Within the sensor array processing framework, differential microphone arrays (DMAs) have received particular attention due to their small physical size and frequency-invariant beampattern, two appealing characteristics for practical purposes DMA[4, 5, 6, 7, 8, 9]. DMAs were initially inspired by differentiating the acoustic pressure of successive microphones in the time domain [10, 11], potentially in a multistage manner, these type of beamformers were later modified to overcome their inherent intolerance to microphone imperfections [12, 13]. These modifications typically considered the input observation signals in the short-time Fourier transform (STFT) domain, in which a DMA design was formulated as a single-stage linear equations system [14, 15].

Due to their simplicity and easy-to-analyze nature, differential uniform linear arrays (ULAs) have been most commonly addressed in the literature [16, 17, 18, 19]. Unfortunately, they suffer from a few inherent drawbacks. For example, it is well known that to attain a high level of directivity, the desired signal is preferably in the endfire direction [8]. In addition, ULAs suffer from a lower-upper plane ambiguity: the beampattern of any ULA is always symmetric concerning the imaginary line connecting the sensors of the array. Therefore, more sophisticated geometric structures were explored, out of which differential uniform circular arrays (UCAs) have drawn the most attention [20, 21, 22]. Other studies exploited the Jacobi-Anger expansion approximation to refer to differential beamforming with arbitrarily-shaped planar arrays [23, 24]. These approaches did not assume any regular array shape but merely required the positions of the array sensors to be either known in advance or measurable. While they are general, they are susceptible to selecting the expansion's reference point and may result in frequency-variant beampatterns as the array size increases. Therefore, they might not embody a proper beamforming design approach with symmetric array geometries, for which it may be possible to take advantage of the symmetry to circumvent these drawbacks.

Rectangular-shaped arrays are symmetric and valuable structures, which may be used to design differential beamformers with asymmetric beampatterns [2]. On top of being particularly suitable for rectangular-shaped appliances, such arrays may also be designed flexibly. For example, a uniform rectangular beamformer may always be decomposed into two

*Corresponding author
Email address: galitz@campus.technion.ac.il (Gal Itzhak)

sub-beamformers by employing the Kronecker-product (KP) decomposition. This allows some flexibility: the KP decomposition is not unique, and each sub-beamformer may be independently designed for a different criterion [25]. An alternative approach [26] exploits the rectangular geometry to improve white noise robustness at the expense of array directivity. However, with this approach, the beam steering property of its corresponding beamformers is not considered.

This paper presents a multistage rectangular approach for steerable differential beamforming. As a first step, we propose to employ a 2-D differentiation scheme that operates independently on the columns and rows of the observation signals of a URA. This yields a differentials matrix controlled by two parameters, P_c and P_r, which indicate the number of differential stages for the URA columns and rows, respectively. Then, as a second step, we design a rectangular differential beamformer and apply it to the vector form of the differentials matrix. At some level, this approach may be seen as the URA generalization of the work in [27, 28], which addresses ULAs. We show that the first differentiation scheme may significantly improve the directivity of the resulted beamformer at the expense of white noise amplification. The latter is heavily tied to selecting the (P_c, P_r) configuration optimized for the desired signal incident angle. Next, we propose four rectangular differential beamformers and analyze their performance in the WNG and DF measures. Finally, we address reverberant scenarios with three distinct incident angles of the desired signal. We examine the versions of each presented beamformer in terms of four reduction factors calculated from the noisy and enhanced signals in the time domain and investigate their quality and intelligibility. We demonstrate that the proposed rectangular differential beamformers outperform common linear differential beamformers in terms of these measures, mainly when the incident angle is far from the endfire direction.

The rest of the paper is organized as follows. In Section 2, we discuss the rectangular array signal model. We define the signals of interest and formulate the observations in vector and matrix forms. In Section 3, we present a multistage differential scheme that independently operates on both axes of the rectangular array. We formulate the resulting 2-D differentials matrix controlled by the (P_c, P_r) configuration and express the corresponding SNR gains between the observation differentials and the raw observations. Then, in Section 4, we discuss the application of a rectangular differential beamformer onto the vector form of the differentials matrix. We derive the standard performance measures, including the WNG, DF, and the power beampattern. Section 5 is dedicated to deriving four types of multistage rectangular differential beamformers. Section 6 consists of two parts. The first part analyzes the WNG and DF measures. The second part shows simulations with speech signals in reverberant environments and varying incident angles of the desired speech. We compare the multistage rectangular differential beamformers to their existing linear counterparts. Finally, we summarize this study in Section 7.

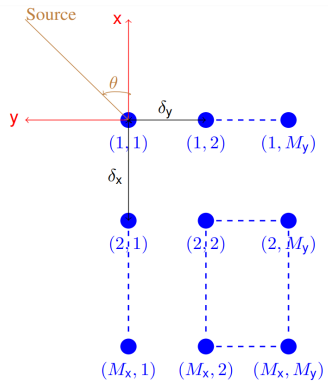


Figure 1: Illustration of the studied rectangular microphone array.

2. Signal Model

Consider a 2-D microphone URA. Given the Cartesian coordinate system with microphone $(1,1)$ as its origin, the URA is composed of M_{x} omnidirectional sensors along the x (negative) axis with a uniform interelement spacing equal to δ_{x} and M_{y} omnidirectional sensors along the y (negative) axis with a uniform interelement spacing equal to δ_{y}. Thus, the total number of microphones is equal to $M_{\mathsf{x}}M_{\mathsf{y}}$, whose positions are denoted $(m_{\mathsf{x}}, m_{\mathsf{y}})$, with $m_{\mathsf{x}} = 1, 2, \ldots, M_{\mathsf{x}}$ and $m_{\mathsf{y}} = 1, 2, \ldots, M_{\mathsf{y}}$. Notice that in the direction of the x axis, we have M_{y} parallel ULAs composed of M_{x} microphones each with a spacing δ_{x}, while in the direction of the y axis, we have M_{x} parallel ULAs composed of M_{y} microphones each with a spacing δ_{y}. An illustration of the 2-D URA studied in this paper is depicted in Fig. 1.

We assume that a farfield desired source signal (plane wave), on the same plane of the 2-D array, propagates from the azimuth angle, θ, in an anechoic acoustic environment at the speed of sound, i.e., $c = 340$ m/s, and impinges on the above described array. Then, the corresponding steering matrix (of size $M_{\mathsf{x}} \times M_{\mathsf{y}}$) is [2]:

$$\begin{aligned}\mathbf{D}_\theta(\omega) &= \left[\begin{array}{ccc} B_{\theta,1}(\omega)\,\mathbf{a}_\theta(\omega) & \cdots & B_{\theta,M_{\mathsf{y}}}(\omega)\,\mathbf{a}_\theta(\omega)\end{array}\right] \\ &= \mathbf{b}_\theta^T(\omega) \otimes \mathbf{a}_\theta(\omega), \end{aligned} \tag{1}$$

where

$$\begin{aligned}\mathbf{a}_\theta(\omega) &= \left[\begin{array}{cccc} A_{\theta,1}(\omega) & A_{\theta,2}(\omega) & \cdots & A_{\theta,M_{\mathsf{x}}}(\omega)\end{array}\right]^T \\ &= \left[\begin{array}{cccc} 1 & e^{-\jmath\varpi_{\theta,\mathsf{x}}(\omega)} & \cdots & e^{-\jmath(M_{\mathsf{x}}-1)\varpi_{\theta,\mathsf{x}}(\omega)}\end{array}\right]^T \end{aligned} \tag{2}$$

is the steering vector associated with the x axis,

$$\begin{aligned}\mathbf{b}_\theta(\omega) &= \left[\begin{array}{cccc} B_{\theta,1}(\omega) & B_{\theta,2}(\omega) & \cdots & B_{\theta,M_{\mathsf{y}}}(\omega)\end{array}\right]^T \\ &= \left[\begin{array}{cccc} 1 & e^{-\jmath\varpi_{\theta,\mathsf{y}}(\omega)} & \cdots & e^{-\jmath(M_{\mathsf{y}}-1)\varpi_{\theta,\mathsf{y}}(\omega)}\end{array}\right]^T \end{aligned} \tag{3}$$

is the steering vector associated with the y axis,

$$\varpi_{\theta,\mathsf{x}}(\omega) = \frac{\omega\delta_{\mathsf{x}}\cos\theta}{c},$$
$$\varpi_{\theta,\mathsf{y}}(\omega) = \frac{\omega\delta_{\mathsf{y}}\sin\theta}{c},$$

the superscript ^{T} denotes the transpose operator, $\otimes$ is the KP operator, $\jmath = \sqrt{-1}$ is the imaginary unit, $\omega = 2\pi f$ is the angular frequency, and $f > 0$ is the temporal frequency.

Exploiting (1), the observed signal matrix of size $M_{\mathsf{x}} \times M_{\mathsf{y}}$ of the URA can be expressed in the frequency domain as [15]:

$$\begin{aligned}\mathbf{Y}(\omega) &= \mathbf{X}(\omega) + \mathbf{V}(\omega) \\ &= \mathbf{D}_\theta(\omega) X(\omega) + \mathbf{V}(\omega),\end{aligned} \tag{4}$$

where $X(\omega)$ is the zero-mean desired source signal and $\mathbf{V}(\omega)$ is the zero-mean additive noise signal matrix.

It is also convenient to express (4) in a vector form. Defining the steering vector $\mathbf{d}_\theta(\omega)$ of length $M_{\mathsf{x}} M_{\mathsf{y}}$, which is formed by concatenating the columns of $\mathbf{D}_\theta(\omega)$, by:

$$\mathbf{d}_\theta = \mathbf{b}_\theta \otimes \mathbf{a}_\theta, \tag{5}$$

we have

$$\begin{aligned}\mathbf{y}(\omega) &= \left[\begin{array}{cccc} \mathbf{y}_1^T(\omega) & \mathbf{y}_2^T(\omega) & \cdots & \mathbf{y}_{M_{\mathsf{y}}}^T(\omega) \end{array}\right]^T \\ &= \mathbf{x}(\omega) + \mathbf{v}(\omega) \\ &= \mathbf{d}_\theta(\omega) X(\omega) + \mathbf{v}(\omega),\end{aligned} \tag{6}$$

where

$$\begin{aligned}\mathbf{y}_{m_{\mathsf{y}}}(\omega) &= \left[\begin{array}{ccc} Y_{m_{\mathsf{y}},1}(\omega) & \cdots & Y_{m_{\mathsf{y}},M_{\mathsf{x}}}(\omega) \end{array}\right]^T \\ &= \mathbf{x}_{m_{\mathsf{y}}}(\omega) + \mathbf{v}_{m_{\mathsf{y}}}(\omega) \\ &= B_{\theta,m_{\mathsf{y}}}(\omega)\, \mathbf{a}_\theta(\omega) X(\omega) + \mathbf{v}_{m_{\mathsf{y}}}(\omega),\end{aligned} \tag{7}$$

for $m_{\mathsf{y}} = 1, 2, \ldots, M_{\mathsf{y}}$, is the observed signal vector of length M_{x} of the m_{y}th ULA parallel to the x axis, and $\mathbf{v}(\omega)$ and $\mathbf{v}_{m_{\mathsf{y}}}(\omega)$ are defined in a similar manner. Dropping the dependence on ω to simplify the notation and assuming a distinct incident angle θ_{s}, equation (6) becomes:

$$\mathbf{y} = (\mathbf{b}_{\theta_{\mathrm{s}}} \otimes \mathbf{a}_{\theta_{\mathrm{s}}}) X + \mathbf{v}, \tag{8}$$

where $\mathbf{b}_{\theta_{\mathrm{s}}} \otimes \mathbf{a}_{\theta_{\mathrm{s}}} = \mathbf{d}_{\theta_{\mathrm{s}}}$ is the steering vector at the incident angle θ_{s}, and the covariance matrix of $\mathbf{y}$ is:

$$\mathbf{\Phi}_{\mathbf{y}} = E\left(\mathbf{y}\mathbf{y}^H\right) = \phi_X \mathbf{d}_{\theta_{\mathrm{s}}} \mathbf{d}_{\theta_{\mathrm{s}}}^H + \mathbf{\Phi}_{\mathbf{v}}, \tag{9}$$

where $E(\cdot)$ denotes mathematical expectation, the superscript ^{H} is the conjugate-transpose operator, $\phi_X = E\left(|X|^2\right)$ is the variance of X, and $\mathbf{\Phi}_{\mathbf{v}} = E\left(\mathbf{v}\mathbf{v}^H\right)$ is the covariance matrix of $\mathbf{v}$. Assuming that the variance of the noise is approximately the same at all sensors, we can express (9) as:

$$\mathbf{\Phi}_{\mathbf{y}} = \phi_X \mathbf{d}_{\theta_{\mathrm{s}}} \mathbf{d}_{\theta_{\mathrm{s}}}^H + \phi_V \mathbf{\Gamma}_{\mathbf{v}}, \tag{10}$$

where ϕ_V is the variance of the noise at the reference microphone (i.e., the origin of the Cartesian coordinate system) and $\mathbf{\Gamma}_{\mathbf{v}} = \mathbf{\Phi}_{\mathbf{v}}/\phi_V$ is the pseudo-coherence matrix of the noise. From (10), we deduce that the input signal-to-noise ratio (SNR) is:

$$\mathrm{iSNR} = \frac{\mathrm{tr}\left(\phi_X \mathbf{d}_{\theta_{\mathrm{s}}} \mathbf{d}_{\theta_{\mathrm{s}}}^H\right)}{\mathrm{tr}\left(\phi_V \mathbf{\Gamma}_{\mathbf{v}}\right)} = \frac{\phi_X}{\phi_V}, \tag{11}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a square matrix. In the case of the spherically isotropic (diffuse) noise field, (10) becomes:

$$\mathbf{\Phi}_{\mathbf{y}} = \phi_X \mathbf{d}_{\theta_{\mathrm{s}}} \mathbf{d}_{\theta_{\mathrm{s}}}^H + \phi \mathbf{\Gamma}_{\mathrm{d}}, \tag{12}$$

where ϕ is the variance of the diffuse noise and $\mathbf{\Gamma}_{\mathrm{d}}$ is the pseudo-coherence matrix of the diffuse noise. We have

$$\mathbf{\Gamma}_{\mathrm{d}} = \left[\begin{array}{cccc} \mathbf{\Gamma}_{\mathrm{d},1} & \mathbf{\Gamma}_{\mathrm{d},2} & \cdots & \mathbf{\Gamma}_{\mathrm{d},M_{\mathsf{y}}} \\ \mathbf{\Gamma}_{\mathrm{d},2} & \mathbf{\Gamma}_{\mathrm{d},1} & \cdots & \mathbf{\Gamma}_{\mathrm{d},M_{\mathsf{y}}-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Gamma}_{\mathrm{d},M_{\mathsf{y}}} & \mathbf{\Gamma}_{\mathrm{d},M_{\mathsf{y}}-1} & \cdots & \mathbf{\Gamma}_{\mathrm{d},1} \end{array}\right], \tag{13}$$

which is a symmetric block Toeplitz matrix, and the elements of the M_{y} symmetric Toeplitz matrices $\mathbf{\Gamma}_{\mathrm{d},m_{\mathsf{y}}}$, $m_{\mathsf{y}} = 1, 2, \ldots, M_{\mathsf{y}}$ (of size $M_{\mathsf{x}} \times M_{\mathsf{x}}$) are given by:

$$\left(\mathbf{\Gamma}_{\mathrm{d},m_{\mathsf{y}}}\right)_{ij} = \mathrm{sinc}\left[\frac{\omega\sqrt{(i-j)^2\delta_{\mathsf{x}}^2 + (m_{\mathsf{y}}-1)^2\delta_{\mathsf{y}}^2}}{c}\right], \tag{14}$$

with $i, j = 1, 2, \ldots, M_{\mathsf{x}}$ and $\mathrm{sinc}(x) = \sin x / x$.

3. Multistage Rectangular Differentials

This section proposes a multistage differential scheme that operates on both axes of the rectangular array.

Let us consider the signal model given in (7). We define the first-order forward spatial difference of $\mathbf{y}_{m_{\mathsf{y}}}$ ($m_{\mathsf{y}} = 1, 2, \ldots, M_{\mathsf{y}}$) as:

$$\begin{aligned}\Delta Y_{m_{\mathsf{y}},i} &= Y_{m_{\mathsf{y}},i+1} - Y_{m_{\mathsf{y}},i} \\ &= Y_{m_{\mathsf{y}},(1),i},\ i = 1, 2, \ldots, M_{\mathsf{x}} - 1,\end{aligned} \tag{15}$$

where Δ is the forward spatial difference operator. In a vector/matrix form, (15) is:

$$\mathbf{\Delta}_{(1)} \mathbf{y}_{m_{\mathsf{y}}} = \mathbf{y}_{m_{\mathsf{y}},(1)}, \tag{16}$$

where

$$\mathbf{\Delta}_{(1)} = \left[\begin{array}{ccccc} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{array}\right] \tag{17}$$

is a matrix of size $(M_{\mathsf{x}} - 1) \times M_{\mathsf{x}}$. In the same way, the second-order forward spatial difference of $\mathbf{y}_{m_{\mathsf{y}}}$ is:

$$\begin{aligned}\Delta^2 Y_{m_{\mathsf{y}},i} &= \Delta\left(\Delta Y_{m_{\mathsf{y}},i}\right) = \Delta Y_{m_{\mathsf{y}},i+1} - \Delta Y_{m_{\mathsf{y}},i} \\ &= Y_{m_{\mathsf{y}},i+2} - 2Y_{m_{\mathsf{y}},i+1} + Y_{m_{\mathsf{y}},i} = Y_{m_{\mathsf{y}},(2),i}\end{aligned} \tag{18}$$

with $i = 1, 2, \ldots, M_{\mathsf{x}} - 2$, which can be rewritten as:

$$\mathbf{\Delta}_{(2)} \mathbf{y}_{m_{\mathsf{y}}} = \mathbf{y}_{m_{\mathsf{y}},(2)}, \tag{19}$$

where

$$\mathbf{\Delta}_{(2)} = \left[\begin{array}{cccccc} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & -2 & 1 \end{array}\right] \tag{20}$$

is a matrix of size $(M_{\mathsf{x}}-2)\times M_{\mathsf{x}}$. More generally, let $p=0,1,\ldots,P_c$, with $1\leq P_c<M_{\mathsf{x}}$. By definition, we write $\mathbf{\Delta}_{(0)}=\mathbf{I}_{M_{\mathsf{x}}}$,where $\mathbf{I}_{M_{\mathsf{x}}}$ is the $M_{\mathsf{x}}\times M_{\mathsf{x}}$ identity matrix. Therefore,

$$\mathbf{\Delta}_{(0)}\mathbf{y}_{m_{\mathsf{y}}}=\mathbf{I}_{M_{\mathsf{x}}}\mathbf{y}_{m_{\mathsf{y}}}=\mathbf{y}_{m_{\mathsf{y}}}. \tag{21}$$

We define the pth-order forward spatial difference of $\mathbf{y}_{m_{\mathsf{y}}}$ as:

$$\begin{aligned}\Delta^pY_{m_{\mathsf{y}},i}&=\Delta^{p-1}\left(\Delta Y_{m_{\mathsf{y}},i}\right)\\&=\Delta^{p-1}Y_{m_{\mathsf{y}},i+1}-\Delta^{p-1}Y_{m_{\mathsf{y}},i}\\&=\sum_{j=0}^{p}(-1)^{p-j}\left(\begin{array}{c}p\\j\end{array}\right)Y_{m_{\mathsf{y}},i+j},\end{aligned} \tag{22}$$

where $i=1,2,\ldots,M_{\mathsf{x}}-p$ and

$$\left(\begin{array}{c}p\\j\end{array}\right)=\frac{p!}{j!\,(p-j)!}$$

is the binomial coefficient. In a vector/matrix form, (22) is:

$$\mathbf{\Delta}_{(P_c)}\mathbf{y}_{m_{\mathsf{y}}}=\mathbf{y}_{m_{\mathsf{y}},(P_c)}, \tag{23}$$

where

$$\mathbf{\Delta}_{(P_c)}=\left[\begin{array}{cccc}\mathbf{c}_{(P_c)}^T&0&\cdots&0\\0&\mathbf{c}_{(P_c)}^T&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&\mathbf{c}_{(P_c)}^T\end{array}\right] \tag{24}$$

is a matrix of size $(M_{\mathsf{x}}-P_c)\times M_{\mathsf{x}}$, with

$$\begin{aligned}\mathbf{c}_{(P_c)}=\Bigg[(-1)^{P_c}\left(\begin{array}{c}P_c\\0\end{array}\right)\quad(-1)^{P_c-1}\left(\begin{array}{c}P_c\\1\end{array}\right)\\\cdots\quad-\left(\begin{array}{c}P_c\\P_c-1\end{array}\right)\quad1\Bigg]^T\end{aligned} \tag{25}$$

being a vector of length P_c+1.

Now, substituting (7) (with $\theta=\theta_{\mathsf{s}}$) into (22), it can be shown that:

$$\begin{aligned}\mathbf{\Delta}_{(P_c)}\mathbf{y}_{m_{\mathsf{y}}}&=B_{\theta_{\mathsf{s}},m_{\mathsf{y}}}\mathbf{\Delta}_{(P_c)}\mathbf{a}_{\theta_{\mathsf{s}}}X+\mathbf{\Delta}_{(P_c)}\mathbf{v}_{m_{\mathsf{y}}}\\&=B_{\theta_{\mathsf{s}},m_{\mathsf{y}}}\tau_{\theta_{\mathsf{s}},\mathsf{x}}^{P_c}\mathbf{a}_{\theta_{\mathsf{s}},(P_c)}X+\mathbf{v}_{m_{\mathsf{y}},(P_c)}\\&=\mathbf{y}_{m_{\mathsf{y}},(P_c)},\end{aligned} \tag{26}$$

where

$$\tau_{\theta_{\mathsf{s}},\mathsf{x}}=e^{-\jmath\varpi_{\theta_{\mathsf{s}},\mathsf{x}}}-1, \tag{27}$$

$$\begin{aligned}\mathbf{a}_{\theta_{\mathsf{s}},(P_c)}=\big[1\;\;e^{-\jmath\varpi_{\theta_{\mathsf{s}},\mathsf{x}}}\\\cdots\quad e^{-\jmath(M_{\mathsf{x}}-P_c-1)\varpi_{\theta_{\mathsf{s}},\mathsf{x}}}\;\big]^T\end{aligned} \tag{28}$$

is the steering vector of length $M_{\mathsf{x}}-P_c$ of the ULA at $\theta=\theta_{\mathsf{s}}$, and $\mathbf{v}_{m_{\mathsf{y}},(P_c)}=\mathbf{\Delta}_{(P_c)}\mathbf{v}_{m_{\mathsf{y}}}$. In an analogous manner, equations (15)-(28) can be rewritten with the roles of x and y axes interchanged. That is, we may differentiate over the rows of $\mathbf{Y}$ instead of over its columns. Define:

$$\tau_{\theta_{\mathsf{s}},\mathsf{y}}=e^{-\jmath\varpi_{\theta_{\mathsf{s}},\mathsf{y}}}-1, \tag{29}$$

$$\begin{aligned}\mathbf{b}_{\theta_{\mathsf{s}},(P_r)}=\big[1\;\;e^{-\jmath\varpi_{\theta_{\mathsf{s}},\mathsf{y}}}\\\cdots\quad e^{-\jmath(M_{\mathsf{y}}-P_r-1)\varpi_{\theta_{\mathsf{s}},\mathsf{y}}}\;\big]^T,\end{aligned} \tag{30}$$

with $1\leq P_r<M_{\mathsf{y}}$. Recalling the matrix form in (4), we may define the 2-D differentials matrix of $\mathbf{Y}$ with P_c column differentials and P_r row differentials, $\mathbf{Y}_{(P_c,P_r)}$, by:

$$\begin{aligned}\mathbf{Y}_{(P_c,P_r)}&=\mathbf{\Delta}_{(P_c)}\mathbf{Y}\mathbf{\Delta}_{(P_r)}^T\\&=\mathbf{\Delta}_{(P_c)}\left[\mathbf{X}+\mathbf{V}\right]\mathbf{\Delta}_{(P_r)}^T\\&=\tau_{\theta_{\mathsf{s}},\mathsf{x}}^{P_c}\tau_{\theta_{\mathsf{s}},\mathsf{y}}^{P_r}\left(\mathbf{b}_{\theta_{\mathsf{s}},(P_r)}^T\otimes\mathbf{a}_{\theta_{\mathsf{s}},(P_c)}\right)X\\&\quad+\mathbf{\Delta}_{(P_c)}\mathbf{V}\mathbf{\Delta}_{(P_r)}^T.\end{aligned} \tag{31}$$

Applying the (column-wise) vectorization operator, $\mathrm{vec}[\cdot]$, to $\mathbf{Y}_{(P_c,P_r)}$, we obtain:

$$\begin{aligned}\mathbf{y}_{(P_c,P_r)}&=\mathrm{vec}\left[\mathbf{Y}_{(P_c,P_r)}\right]\\&=\tau_{\theta_{\mathsf{s}},\mathsf{x}}^{P_c}\tau_{\theta_{\mathsf{s}},\mathsf{y}}^{P_r}\left(\mathbf{b}_{\theta_{\mathsf{s}},(P_r)}\otimes\mathbf{a}_{\theta_{\mathsf{s}},(P_c)}\right)X\\&\quad+\mathrm{vec}\left[\mathbf{\Delta}_{(P_c)}\mathbf{V}\mathbf{\Delta}_{(P_r)}^T\right]\\&=\tau_{\theta_{\mathsf{s}},\mathsf{x}}^{P_c}\tau_{\theta_{\mathsf{s}},\mathsf{y}}^{P_r}\mathbf{d}_{\theta_{\mathsf{s}},(P_c,P_r)}X\\&\quad+\left(\Delta_{(P_r)}\otimes\Delta_{(P_c)}\right)\mathbf{v},\end{aligned} \tag{32}$$

where $\mathbf{d}_{\theta_{\mathsf{s}},(P_c,P_r)}=\mathbf{b}_{\theta_{\mathsf{s}},(P_r)}\otimes\mathbf{a}_{\theta_{\mathsf{s}},(P_c)}$ is the 2-D differential steering vector of length $(M_{\mathsf{x}}-P_c)(M_{\mathsf{y}}-P_r)$. We deduce that the $(M_{\mathsf{x}}-P_c)(M_{\mathsf{y}}-P_r)\times(M_{\mathsf{x}}-P_c)(M_{\mathsf{y}}-P_r)$ covariance matrix of $\mathbf{y}_{(P_c,P_r)}$ is:

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}_{(P_c,P_r)}}&=\phi_X\left|\tau_{\theta_{\mathsf{s}},\mathsf{x}}\right|^{2P_c}\left|\tau_{\theta_{\mathsf{s}},\mathsf{y}}\right|^{2P_r}\\&\quad\times\mathbf{d}_{\theta_{\mathsf{s}},(P_c,P_r)}\mathbf{d}_{\theta_{\mathsf{s}},(P_c,P_r)}^H\\&\quad+\phi_V\left(\Delta_{(P_r)}\otimes\Delta_{(P_c)}\right)\mathbf{\Gamma}_{\mathbf{v}}\\&\quad\times\left(\Delta_{(P_r)}\otimes\Delta_{(P_c)}\right)^T.\end{aligned} \tag{33}$$

We immediately obtain the SNR corresponding to $\mathbf{y}_{(P_c,P_r)}$ by:

$$\begin{aligned}\mathrm{SNR}_{\mathbf{y}_{(P_c,P_r)}}&=\mathrm{tr}\Big(\phi_X\left|\tau_{\theta_{\mathsf{s}},\mathsf{x}}\right|^{2P_c}\left|\tau_{\theta_{\mathsf{s}},\mathsf{y}}\right|^{2P_r}\\&\quad\times\mathbf{d}_{\theta_{\mathsf{s}},(P_c,P_r)}\mathbf{d}_{\theta_{\mathsf{s}},(P_c,P_r)}^H\Big)\\&\quad\times\Big[\mathrm{tr}\Big(\phi_V\left(\Delta_{(P_r)}\otimes\Delta_{(P_c)}\right)\mathbf{\Gamma}_{\mathbf{v}}\\&\quad\times\left(\Delta_{(P_r)}\otimes\Delta_{(P_c)}\right)^T\Big)\Big]^{-1}\\&=\phi_X\left|\tau_{\theta_{\mathsf{s}},\mathsf{x}}\right|^{2P_c}\left|\tau_{\theta_{\mathsf{s}},\mathsf{y}}\right|^{2P_r}\\&\quad\times(M_{\mathsf{x}}-P_c)(M_{\mathsf{y}}-P_r)\\&\quad\times\Big[\mathrm{tr}\Big(\phi_V\left(\Delta_{(P_r)}\otimes\Delta_{(P_c)}\right)\mathbf{\Gamma}_{\mathbf{v}}\\&\quad\times\left(\Delta_{(P_r)}\otimes\Delta_{(P_c)}\right)^T\Big)\Big]^{-1},\end{aligned} \tag{34}$$

and the SNR gain between $\mathbf{y}_{(P_c,P_r)}$ and $\mathbf{y}$:

$$\mathcal{G}_{(P_c,P_r)} = \frac{|\tau_{\theta_\mathrm{s},\mathsf{x}}|^{2P_c}\,|\tau_{\theta_\mathrm{s},\mathsf{y}}|^{2P_r}\,(M_\mathsf{x}-P_c)(M_\mathsf{y}-P_r)}{\mathrm{tr}\left(\left(\Delta_{(P_r)}\otimes\Delta_{(P_c)}\right)\mathbf{\Gamma}_\mathbf{v}\left(\Delta_{(P_r)}\otimes\Delta_{(P_c)}\right)^T\right)}. \tag{35}$$

Addressing the white noise case, we obtain:

$$\begin{aligned}\mathcal{W}_{(P_c,P_r)} &= \frac{|\tau_{\theta_\mathrm{s},\mathsf{x}}|^{2P_c}\,|\tau_{\theta_\mathrm{s},\mathsf{y}}|^{2P_r}\,(M_\mathsf{x}-P_c)(M_\mathsf{y}-P_r)}{\mathrm{tr}\left(\left(\Delta_{(P_r)}\otimes\Delta_{(P_c)}\right)\left(\Delta_{(P_r)}\otimes\Delta_{(P_c)}\right)^T\right)}\\ &= \frac{|\tau_{\theta_\mathrm{s},\mathsf{x}}|^{2P_c}\,|\tau_{\theta_\mathrm{s},\mathsf{y}}|^{2P_r}\,(M_\mathsf{x}-P_c)}{\mathrm{tr}\left((\mathbf{c}^T_{(P_r)}\otimes\Delta_{(P_c)})(\mathbf{c}_{(P_r)}\otimes\Delta^T_{(P_c)})\right)}\\ &= \frac{|\tau_{\theta_\mathrm{s},\mathsf{x}}|^{2P_c}\,|\tau_{\theta_\mathrm{s},\mathsf{y}}|^{2P_r}\,(M_\mathsf{x}-P_c)}{\begin{pmatrix}2P_r\\P_r\end{pmatrix}\Delta_{(P_c)}\Delta^T_{(P_c)}}\\ &= \frac{|\tau_{\theta_\mathrm{s},\mathsf{x}}|^{2P_c}}{\begin{pmatrix}2P_c\\P_c\end{pmatrix}}\times\frac{|\tau_{\theta_\mathrm{s},\mathsf{y}}|^{2P_r}}{\begin{pmatrix}2P_r\\P_r\end{pmatrix}}\\ &= \mathcal{W}_{(P_c)}\times\mathcal{W}_{(P_r)},\end{aligned} \tag{36}$$

where $\mathcal{W}_{(P_c)}$ and $\mathcal{W}_{(P_r)}$ are the linear white noise gains with respect to the columns and the rows of the URA, respectively. Substituting (13) into (35), the diffuse noise gain is given by:

$$\mathcal{D}_{(P_c,P_r)} = \frac{|\tau_{\theta_\mathrm{s},\mathsf{x}}|^{2P_c}\,|\tau_{\theta_\mathrm{s},\mathsf{y}}|^{2P_r}\,(M_\mathsf{x}-P_c)(M_\mathsf{y}-P_r)}{\mathrm{tr}\left(\left(\Delta_{(P_r)}\otimes\Delta_{(P_c)}\right)\mathbf{\Gamma}_\mathrm{d}\left(\Delta_{(P_r)}\otimes\Delta_{(P_c)}\right)^T\right)}. \tag{37}$$

4. Multistage Rectangular Differential Beamforming

Next, we would like to apply a differential beamformer $\mathbf{w}_{(P_c,P_r)}$ of length $(M_\mathsf{x}-P_c)(M_\mathsf{y}-P_r)$ to the vector $\mathbf{y}_{(P_c,P_r)}$. Then, the beamformer output signal is:

$$\begin{aligned}Z_{(P_c,P_r)} &= \mathbf{w}^H_{(P_c,P_r)}\mathbf{y}_{(P_c,P_r)}\\ &= X_{\mathrm{fd},(P_c,P_r)} + V_{\mathrm{rn},(P_c,P_r)},\end{aligned} \tag{38}$$

where $Z_{(P_c,P_r)}$ is the estimate of X,

$$X_{\mathrm{fd},(P_c,P_r)} = \tau^{P_c}_{\theta_\mathrm{s},\mathsf{x}}\tau^{P_r}_{\theta_\mathrm{s},\mathsf{y}}\left(\mathbf{w}^H\mathbf{d}_{\theta_\mathrm{s},(P_c,P_r)}\right)X \tag{39}$$

is the filtered desired signal, and:

$$V_{\mathrm{rn},(P_c,P_r)} = \mathbf{w}^H_{(P_c,P_r)}\mathbf{v}_{(P_c,P_r)} \tag{40}$$

is the residual noise, where $\mathbf{v}_{(P_c,P_r)} = \left(\Delta_{(P_r)}\otimes\Delta_{(P_c)}\right)\mathbf{v}$. Consequently, the variance of $Z_{(P_c,P_r)}$ is:

$$\begin{aligned}\phi_{Z_{(P_c,P_r)}} &= \mathbf{w}^H_{(P_c,P_r)}\mathbf{\Phi}_{\mathbf{y}_{(P_c,P_r)}}\mathbf{w}_{(P_c,P_r)}\\ &= \phi_{X_{\mathrm{fd},(P_c,P_r)}} + \phi_{V_{\mathrm{rn},(P_c,P_r)}},\end{aligned} \tag{41}$$

where

$$\begin{aligned}\phi_{X_{\mathrm{fd},(P_c,P_r)}} &= \phi_X\,|\tau_{\theta_\mathrm{s},\mathsf{x}}|^{2P_c}\,|\tau_{\theta_\mathrm{s},\mathsf{y}}|^{2P_r}\\ &\quad\times\left|\mathbf{w}^H_{(P_c,P_r)}\mathbf{d}_{\theta_\mathrm{s},(P_c,P_r)}\right|^2,\end{aligned} \tag{42}$$

$$\phi_{V_{\mathrm{rn},(P_c,P_r)}} = \mathbf{w}^H_{(P_c,P_r)}\mathbf{\Phi}_{\mathbf{v}_{(P_c,P_r)}}\mathbf{w}_{(P_c,P_r)}, \tag{43}$$

and $\mathbf{\Phi}_{\mathbf{v}_{(P_c,P_r)}}$ is the correlation matrix of $\mathbf{v}_{(P_c,P_r)}$ which is given by:

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{v}_{(P_c,P_r)}} &= \phi_V\left(\Delta_{(P_r)}\otimes\Delta_{(P_c)}\right)\mathbf{\Gamma}_\mathbf{v}\\ &\quad\times\left(\Delta_{(P_r)}\otimes\Delta_{(P_c)}\right)^T\\ &= \phi_V\mathbf{\Gamma}_{\mathbf{v}_{(P_c,P_r)}}.\end{aligned} \tag{44}$$

Ultimately, it is clear that the distortionless constraint is given by:

$$\mathbf{w}^H_{(P_c,P_r)}\mathbf{d}_{\theta_\mathrm{s},(P_c,P_r)} = \tau^{-P_c}_{\theta_\mathrm{s},\mathsf{x}}\tau^{-P_r}_{\theta_\mathrm{s},\mathsf{y}}. \tag{45}$$

Now, let us relate the most prominent performance measures corresponding to $\mathbf{w}_{(P_c,P_r)}$. The output SNR and SNR gain are, respectively,

$$\begin{aligned}\mathrm{oSNR}\left(\mathbf{w}_{(P_c,P_r)}\right) &= |\tau_{\theta_\mathrm{s},\mathsf{x}}|^{2P_c}\,|\tau_{\theta_\mathrm{s},\mathsf{y}}|^{2P_r}\times\frac{\phi_X}{\phi_V}\\ &\quad\times\frac{\left|\mathbf{w}^H_{(P_c,P_r)}\mathbf{d}_{\theta_\mathrm{s},(P_c,P_r)}\right|^2}{\mathbf{w}^H_{(P_c,P_r)}\mathbf{\Gamma}_{\mathbf{v}_{(P_c,P_r)}}\mathbf{w}_{(P_c,P_r)}},\end{aligned} \tag{46}$$

and

$$\begin{aligned}\mathcal{G}\left(\mathbf{w}_{(P_c,P_r)}\right) &= |\tau_{\theta_\mathrm{s},\mathsf{x}}|^{2P_c}\,|\tau_{\theta_\mathrm{s},\mathsf{y}}|^{2P_r}\\ &\quad\times\frac{\left|\mathbf{w}^H_{(P_c,P_r)}\mathbf{d}_{\theta_\mathrm{s},(P_c,P_r)}\right|^2}{\mathbf{w}^H_{(P_c,P_r)}\mathbf{\Gamma}_{\mathbf{v}_{(P_c,P_r)}}\mathbf{w}_{(P_c,P_r)}}.\end{aligned} \tag{47}$$

Consequently, we deduce that the WNG is given by:

$$\begin{aligned}\mathcal{W}\left(\mathbf{w}_{(P_c,P_r)}\right) &= |\tau_{\theta_\mathrm{s},\mathsf{x}}|^{2P_c}\,|\tau_{\theta_\mathrm{s},\mathsf{y}}|^{2P_r}\\ &\quad\times\frac{\left|\mathbf{w}^H_{(P_c,P_r)}\mathbf{d}_{\theta_\mathrm{s},(P_c,P_r)}\right|^2}{\mathbf{w}^H_{(P_c,P_r)}\mathbf{\Xi}_{(P_c,P_r)}\mathbf{w}_{(P_c,P_r)}},\end{aligned} \tag{48}$$

and the DF:

$$\begin{aligned}\mathcal{D}\left(\mathbf{w}_{(P_c,P_r)}\right) &= |\tau_{\theta_\mathrm{s},\mathsf{x}}|^{2P_c}\,|\tau_{\theta_\mathrm{s},\mathsf{y}}|^{2P_r}\\ &\quad\times\frac{\left|\mathbf{w}^H_{(P_c,P_r)}\mathbf{d}_{\theta_\mathrm{s},(P_c,P_r)}\right|^2}{\mathbf{w}^H_{(P_c,P_r)}\mathbf{\Gamma}_{\mathrm{d},(P_c,P_r)}\mathbf{w}_{(P_c,P_r)}},\end{aligned} \tag{49}$$

where

$$\mathbf{\Xi}_{(P_c,P_r)} = \left(\Delta_{(P_r)}\otimes\Delta_{(P_c)}\right)\left(\Delta_{(P_r)}\otimes\Delta_{(P_c)}\right)^T, \tag{50}$$

$$\mathbf{\Gamma}_{\mathrm{d},(P_c,P_r)} = \left(\Delta_{(P_r)}\otimes\Delta_{(P_c)}\right)\mathbf{\Gamma}_\mathrm{d}\left(\Delta_{(P_r)}\otimes\Delta_{(P_c)}\right)^T. \tag{51}$$

Finally, we may define the power beampattern by:

$$\begin{aligned}\left|\mathcal{B}_\theta\left(\mathbf{w}_{(P_c,P_r)}\right)\right|^2 &= |\tau_{\theta,\mathsf{x}}|^{2P_c}\,|\tau_{\theta,\mathsf{y}}|^{2P_r}\\ &\quad\times\left|\mathbf{w}^H_{(P_c,P_r)}\mathbf{d}_{\theta,(P_c,P_r)}\right|^2,\end{aligned} \tag{52}$$

where $\tau_{\theta,\mathsf{x}} = e^{-\jmath\varpi_{\theta,\mathsf{x}}} - 1$ and $\tau_{\theta,\mathsf{y}} = e^{-\jmath\varpi_{\theta,\mathsf{y}}} - 1$.

We end this part by referring to the work in [26], which highly differs from this work. The main idea iteitzhak2021a is to employ a 2-D mean operator to improve white noise robustness at the expense of the array directivity. In addition, with this approach, the beam steering property of its corresponding beamformers is not considered. In contrast, in our proposed method, we exploit the rectangular array geometry to improve the array directivity while allowing beam steering, that is, while roughly maintaining the same level of performance regardless of the desired signal incident angle.

5. Optimal Rectangular Differential Beamformers

5.1. Maximum SNR Gain Rectangular Differential Beamformers

Let us start with the maximization of the WNG measure. Recalling (48), we deduce that the maximum WNG (MWNG) beamformer may be derived from:

$$\begin{aligned}&\min_{\mathbf{w}_{(P_c,P_r)}} \mathbf{w}_{(P_c,P_r)}^H \boldsymbol{\Xi}_{(P_c,P_r)} \mathbf{w}_{(P_c,P_r)} \\ &\text{s. t.} \quad \mathbf{w}_{(P_c,P_r)}^H \mathbf{d}_{\theta_\text{s},(P_c,P_r)} = \tau_{\theta_\text{s},\mathsf{x}}^{-P_c} \tau_{\theta_\text{s},\mathsf{y}}^{-P_r},\end{aligned} \tag{53}$$

with the distortionless constraint of (45) taken into account. Then, it is straightforward to derive the solution:

$$\begin{aligned}\mathbf{w}_{\text{MWNG}(P_c,P_r)} = &\frac{1}{\left(\tau_{\theta_\text{s},\mathsf{x}}^{P_c} \tau_{\theta_\text{s},\mathsf{y}}^{P_r}\right)^*} \\ &\times \frac{\boldsymbol{\Xi}_{(P_c,P_r)}^{-1} \mathbf{d}_{\theta_\text{s},(P_c,P_r)}}{\mathbf{d}_{\theta_\text{s},(P_c,P_r)}^H \boldsymbol{\Xi}_{(P_c,P_r)}^{-1} \mathbf{d}_{\theta_\text{s},(P_c,P_r)}}.\end{aligned} \tag{54}$$

We note that in case the desired signal incident angle θ_s equals either 0^o, 180^o, 90^o, or 270^o the solution is undefined. In the first two cases, to achieve a valid solution, we will never differentiate with respect to the y-axis. Therefore, in these cases, we always have $P_r = 0$. The same issue and solution apply for the two other cases concerning x-axis differentiation, in which we always have $P_c = 0$. In addition, for the sake of mathematical completeness, we define:

$$\tau_{0,\mathsf{y}}^0 = \tau_{\pi,\mathsf{y}}^0 = \tau_{\pi/2,\mathsf{x}}^0 = \tau_{3\pi/2,\mathsf{x}}^0 = 1. \tag{55}$$

We proceed by considering equation (49). The maximum DF (MDF) beamformer is derived from:

$$\begin{aligned}&\min_{\mathbf{w}_{(P_c,P_r)}} \mathbf{w}_{(P_c,P_r)}^H \boldsymbol{\Gamma}_{\text{d},(P_c,P_r)} \mathbf{w}_{(P_c,P_r)} \\ &\text{s. t.} \quad \mathbf{w}_{(P_c,P_r)}^H \mathbf{d}_{\theta_\text{s},(P_c,P_r)} = \tau_{\theta_\text{s},\mathsf{x}}^{-P_c} \tau_{\theta_\text{s},\mathsf{y}}^{-P_r}.\end{aligned} \tag{56}$$

The solution is therefore given by:

$$\begin{aligned}\mathbf{w}_{\text{MDF}(P_c,P_r)} = &\frac{1}{\left(\tau_{\theta_\text{s},\mathsf{x}}^{P_c} \tau_{\theta_\text{s},\mathsf{y}}^{P_r}\right)^*} \\ &\times \frac{\boldsymbol{\Gamma}_{\text{d},(P_c,P_r)}^{-1} \mathbf{d}_{\theta_\text{s},(P_c,P_r)}}{\mathbf{d}_{\theta_\text{s},(P_c,P_r)}^H \boldsymbol{\Gamma}_{\text{d},(P_c,P_r)}^{-1} \mathbf{d}_{\theta_\text{s},(P_c,P_r)}}.\end{aligned} \tag{57}$$

5.2. Null-constrained Rectangular Differential Beamformers

We now turn to null-constrained beamformers. In practice, in order to give a desired shape to a beampattern or attenuate directional interferences, spatial null constraints may be required. Therefore, with N distinct null constraints (53) is transformed into:

$$\begin{aligned}&\min_{\mathbf{w}} \mathbf{w}^H \boldsymbol{\Xi}_{(P_c,P_r)} \mathbf{w} \\ &\text{s. t.} \quad \mathbf{C}^H \left(\Delta_{(P_r)} \otimes \Delta_{(P_c)}\right)^T \mathbf{w}_{(P_c,P_r)} = \beta,\end{aligned} \tag{58}$$

where β is the first column of the identity matrix of size $(N+1) \times (N+1)$, and $\mathbf{C}$ is a constraint matrix of size $M_\mathsf{x} M_\mathsf{y} \times (N+1)$:

$$\mathbf{C} = \left[\begin{array}{cccc} \mathbf{d}_{\theta_\text{s}} & \mathbf{d}_{\theta_1} & \cdots & \mathbf{d}_{\theta_N} \end{array}\right], \tag{59}$$

whose first column is the steering vector in the direction of the desired signal, and the remaining independent columns are the steering vectors in the directions of the desired nulls. The resulting null-constrained maximum WNG (NCMWNG) beamformer is given by:

$$\begin{aligned}\mathbf{w}_{\text{NCMWNG}(P_c,P_r)} = &\boldsymbol{\Xi}_{(P_c,P_r)}^{-1} \left(\Delta_{(P_r)} \otimes \Delta_{(P_c)}\right) \mathbf{C} \\ &\times \Big[\mathbf{C}^H \left(\Delta_{(P_r)} \otimes \Delta_{(P_c)}\right)^T \boldsymbol{\Xi}_{(P_c,P_r)}^{-1} \\ &\times \left(\Delta_{(P_r)} \otimes \Delta_{(P_c)}\right) \mathbf{C}\Big]^{-1} \beta.\end{aligned} \tag{60}$$

In a similar manner, we may optimize with respect to the DF criterion:

$$\begin{aligned}&\min_{\mathbf{w}} \mathbf{w}^H \boldsymbol{\Gamma}_{\text{d},(P_c,P_r)} \mathbf{w} \\ &\text{s. t.} \quad \mathbf{C}^H \left(\Delta_{(P_r)} \otimes \Delta_{(P_c)}\right)^T \mathbf{w}_{(P_c,P_r)} = \beta.\end{aligned} \tag{61}$$

Then, the null-constrained maximum DF (NCMDF) beamformer is obtained as:

$$\begin{aligned}\mathbf{w}_{\text{NCMDF}(P_c,P_r)} = &\boldsymbol{\Gamma}_{\text{d},(P_c,P_r)}^{-1} \left(\Delta_{(P_r)} \otimes \Delta_{(P_c)}\right) \mathbf{C} \\ &\times \Big[\mathbf{C}^H \left(\Delta_{(P_r)} \otimes \Delta_{(P_c)}\right)^T \boldsymbol{\Gamma}_{\text{d},(P_c,P_r)}^{-1} \\ &\times \left(\Delta_{(P_r)} \otimes \Delta_{(P_c)}\right) \mathbf{C}\Big]^{-1} \beta.\end{aligned} \tag{62}$$

6. Simulations

6.1. Performance Study

In this part, we investigate the performance of each of the four rectangular differential beamformers presented in the former section in terms of the WNG and DF measures.

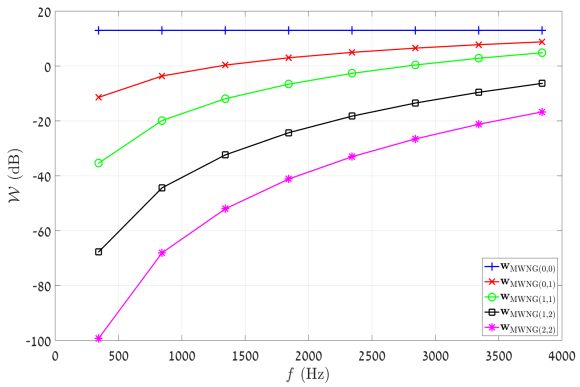
Let us assume the existence of the following URA: $M_\mathsf{x} = 5$, $M_\mathsf{y} = 4$, $\delta_\mathsf{x} = 1$ cm and $\delta_\mathsf{y} = 1.2$ cm. We will exploit this URA example through the entire section. To begin with, it is valuable to evaluate the WNG and DF between the observation differentials vector and the observations vector, that is, $\mathcal{W}_{(P_c,P_r)}$ and $\mathcal{D}_{(P_c,P_r)}$. Recalling equations (36) and (37), we realize that the

optimal configurations are scenario-dependent and are a function of the rectangular array structure and the desired signal incident angle θ_{s}. Table 1 and Table 2, respectively, show all possible values $\mathcal{D}_{(P_c,P_r)}$ and $\mathcal{W}_{(P_c,P_r)}$ for three distinct values of θ_{s}: 15^o, 45^o and 75^o with $f = 4$ kHz. We note that in terms of both measures, optimal values of the (P_c, P_r) configuration are obtained with accordance to θ_{s}: when $\theta_{\mathrm{s}} = 15^o$, P_r should be set to zero whereas P_c sets the WNG-DF trade-off; when $\theta_{\mathrm{s}} = 75^o$, the roles of P_r and P_c interchange; and when $\theta_{\mathrm{s}} = 45^o$ optimal values are obtained for $|P_r - P_c| \leq 1$. We deduce that the multistage rectangular differential approach enables controlled directivity gains and beam steering flexibility.

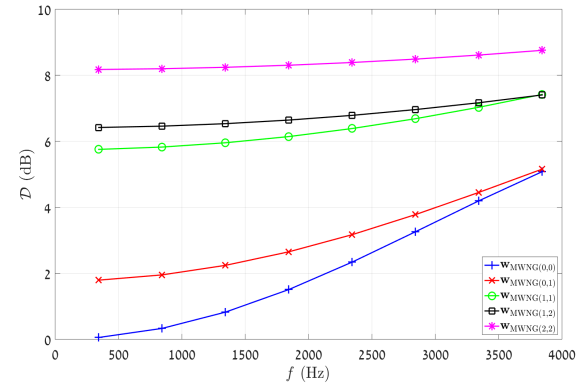
We move on to analyzing the WNG and DF performance measures of each of the beamformers presented in the former section with $\theta_{\mathrm{s}} = 45^o$. We begin with the MWNG rectangular beamformer, $\mathbf{w}_{\mathrm{MWNG}(P_c,P_r)}$, whose WNG and DF measures are depicted in Fig. 2. We observe the following. For $(P_c, P_r) = (0, 0)$, $\mathbf{w}_{\mathrm{MWNG}(0,0)}$ is the well-known Delay-and-Sum (DS) rectangular beamformer, whose frequency-independent WNG equals to the number of array sensors. As the configurations of (P_c, P_r) change to increase $\mathcal{D}_{(P_c,P_r)}$, according to the selected configurations of Tables 1 and 2, the DF performance improves, but the WNG performance degrades. We note that the DF improvement of $\mathbf{w}_{\mathrm{MWNG}(P_c,P_r)}$ complies with $\mathcal{D}_{(P_c,P_r)}$: the DFs of $\mathbf{w}_{\mathrm{MWNG}(0,0)}$ and $\mathbf{w}_{\mathrm{MWNG}(0,1)}$ are rather close and low; $\mathbf{w}_{\mathrm{MWNG}(1,1)}$ and $\mathbf{w}_{\mathrm{MWNG}(1,2)}$ exhibit a significant DF improvement with respect to the former configurations; with $\mathbf{w}_{\mathrm{MWNG}(2,2)}$ the DF is further improved at the expense of a significantly worse WNG performance. We deduce that, indeed, the configuration of the parameters (P_c, P_r) highly affects the WNG-DF performance of the rectangular differential beamformer, with a strong correlation to the gain between the observation differentials vector and the raw observations vector.

Next, we focus on $\mathbf{w}_{\mathrm{MDF}(P_c,P_r)}$, whose WNG and DF performances are depicted in Fig. 3. We observe a similar WNG-DF trade-off as the previous beamformer upon changing the configuration of (P_c, P_r). In contrast to the former case, and as one would expect, the DF performance of $\mathbf{w}_{\mathrm{MDF}(P_c,P_r)}$ is significantly better than $\mathbf{w}_{\mathrm{MWNG}(P_c,P_r)}$'s at the expense of the WNG performance. This is true regardless of the selection of (P_c, P_r). In addition, we note that with the higher values of (P_c, P_r), the WNG performance is poor, particularly in low frequencies. This unappealing behaviour implies that, in practice, $\mathbf{w}_{\mathrm{MDF}(P_c,P_r)}$ should only be designed with a modest $\mathcal{D}_{(P_c,P_r)}$ performance, that is, with small values of P_c and P_r.

We turn to the NCMWNG and NCMDF rectangular differential beamformers, whose performances are shown in Figs. 4 and 5. We note that both beamformers are designed with $N = 2$ distinct nulls located, respectively, at 160^o and -90^o. Considering $\mathbf{w}_{\mathrm{NCMWNG}(P_c,P_r)}$, we clearly observe a DF performance improvement with respect to $\mathbf{w}_{\mathrm{MWNG}(P_c,P_r)}$. This improvement is at the expense of worse WNG performance. Turning to $\mathbf{w}_{\mathrm{NCMDF}(P_c,P_r)}$, we note that its directivity is better than $\mathbf{w}_{\mathrm{NCMWNG}(P_c,P_r)}$'s but its WNG performance is indeed worse. Nevertheless, the DF performance of $\mathbf{w}_{\mathrm{NCMDF}(P_c,P_r)}$ is worse than $\mathbf{w}_{\mathrm{MDF}(P_c,P_r)}$'s, but its WNG performance is preferable.



(a)



(b)

Figure 2: WNG and DF measures with the MWNG rectangular differential beamformer, $\mathbf{w}_{\mathrm{MWNG}(P_c,P_r)}$, with varying values of (P_c, P_r). Simulation parameters: $\theta_{\mathrm{s}} = 45^o$, $M_{\mathsf{x}} = 5$, $M_{\mathsf{y}} = 4$, $\delta_{\mathsf{x}} = 1$ cm and $\delta_{\mathsf{y}} = 1.2$ cm. (a) WNG and (b) DF.

We deduce that the null-constrained versions of the MWNG and MDF rectangular differential beamformers enable additional WNG-DF performance tuning flexibility.

6.2. *Speech Signals Simulations in Reverberant Environments*

In this part, we demonstrate the performance of differential rectangular beamformers on speech signals in practical simulated scenarios in reverberant environments.

The reverberant simulations are performed as follows. We use a room impulse response (RIR) generator [29] to simulate the reverberant noise-free signal received in a URA consisting of $M_{\mathsf{x}} \times M_{\mathsf{y}} = 5 \times 4$ microphones. The URA is located on the $z = 1.5$ m plane, where the first microphone is located at the $(x, y) = (2\,\mathrm{m}, 2\,\mathrm{m})$ coordinate, with $\delta_{\mathsf{x}} = 1$ cm and $\delta_{\mathsf{y}} = 1.2$ cm. The RIR generator is based on the image method of Allen and Berkley [30]. We simulate three distinct scenarios in a $6 \times 6 \times 3$ m room, which differ by the value of the desired speech signal incident angle θ_{s}: 15^o, 45^o or 75^o. In all scenarios, we set $T_{60} = 600$ msec, where T_{60} is defined by Sabin-Franklin's formula [31]. In addition, two uncorrelated directional interferences are located on the $z = 1.5$ m plane in the same null directions of $\mathbf{w}_{\mathrm{NCMDF}(P_c,P_r)}$ and $\mathbf{w}_{\mathrm{NCMDF}(P_c,P_r)}$ from the former part, that is, 160^0 and -90^o. The former in-

Table 1: The DF in dB units between $\mathbf{y}_{(P_c,P_r)}$ and $\mathbf{y}$ for varying values of (P_c, P_r) and three distinct values of the incident angle θ_{s}. Gray background color indicates optimal configurations. Simulation parameters: $f = 4$ kHz, $M_{\mathsf{x}} = 5$, $M_{\mathsf{y}} = 4$, $\delta_{\mathsf{x}} = 1$ cm and $\delta_{\mathsf{y}} = 1.2$ cm.

$\theta_{\mathrm{s}} = 15^o$						$\theta_{\mathrm{s}} = 45^o$						$\theta_{\mathrm{s}} = 75^o$					
		P_r						P_r						P_r			
		0	1	2	3			0	1	2	3			0	1	2	3
P_c	0	0	-6.8	-16.1	-26.2	P_c	0	0	1.8	1.1	-0.3	P_c	0	0	4.4	6.3	7.4
	1	4.4	-0.3	-8.2	-17.2		1	1.8	5.7	6.4	6.0		1	-6.9	-0.4	2.9	5.1
	2	6.3	3.0	-3.8	-12.0		2	1.1	6.4	8.1	8.6		2	-16.2	-8.3	-4.0	-1.0
	3	7.5	5.2	-0.8	-8.3		3	-0.4	6.0	8.5	9.7		3	-26.3	-17.4	-12.2	-8.5
	4	8.3	6.9	1.5	-5.4		4	-2.2	5.0	8.2	10.0		4	-36.8	-27.0	-21.6	-16.9

Table 2: The WNG in dB units between $\mathbf{y}_{(P_c,P_r)}$ and $\mathbf{y}$ for varying values of (P_c, P_r) and three distinct values of the incident angle θ_{s}. Gray background color indicates optimal configurations. Simulation parameters: $f = 4$ kHz, $M_{\mathsf{x}} = 5$, $M_{\mathsf{y}} = 4$, $\delta_{\mathsf{x}} = 1$ cm and $\delta_{\mathsf{y}} = 1.2$ cm.

$\theta_{\mathrm{s}} = 15^o$						$\theta_{\mathrm{s}} = 45^o$						$\theta_{\mathrm{s}} = 75^o$					
		P_r						P_r						P_r			
		0	1	2	3			0	1	2	3			0	1	2	3
P_c	0	0	-15.8	-33.4	-51.4	P_c	0	0	-7.2	-16.2	-25.6	P_c	0	0	-4.6	-11.0	-17.8
	1	-6.1	-21.9	-39.5	-57.5		1	-8.7	-15.9	-24.9	-34.3		1	-17.4	-22.0	-28.4	-35.2
	2	-14.0	-29.8	-47.4	-65.4		2	-19.2	-26.4	-35.4	-44.8		2	-36.5	-41.2	-47.5	-54.4
	3	-22.3	-38.2	-55.7	-73.8		3	-30.2	-37.4	-46.4	-55.8		3	-56.1	-60.8	-67.1	-74.0
	4	-30.9	-46.7	-64.3	-82.3		4	-41.4	-48.6	-57.6	-67.0		4	-76.0	-80.6	-87.0	-93.8

terference is 10 dB weaker than the latter. On top of the directional interferences, two uncorrelated noise fields are present: a white thermal Gaussian noise, which is 30 dB weaker than the more powerful interference, and a spherically-isotropic diffuse noise, which is 3 dB stronger than the more powerful interference. The desired speech signal, $x\left(t\right)$, is a concatenation of 24 speech signals (12 speech signals per gender) with varying dialects that are taken from the TIMIT database [32]. It is sampled at a sampling rate of $f_{\mathrm{s}} = 1/T_{\mathrm{s}} = 16$ kHz within the signal duration T.

The noisy observations signal is transformed into the STFT domain using 75% overlapping time frames and a Hamming analysis window of length 256 (16 msec). Next, differential KP beamformers with different array settings are independently applied to the noisy signal to yield clean signal estimates in the STFT domain, followed by an inverse STFT procedure to obtain time-domain enhanced signals.

Next, we are interested in objectively quantifying the performance of each of the four beamformers. We shall do that by individually examining the power ratio between the noise and reverberation components of the first microphone and their respective components in the time-domain enhanced signals. This includes the white thermal Gaussian noise, the diffuse noise, the reverberant directional interferences, and the desired speech signal reverberations. Formulating the noisy observation signal in the time domain in microphone m we have

$$\begin{aligned} y_m &= x_{\mathrm{d}} * g_{\mathrm{d},m} + v_{\mathrm{i},1} * g_{\mathrm{i},1,m} + v_{\mathrm{i},2} * g_{\mathrm{i},2,m} \\ &\quad + v_{\mathrm{d},m} + v_{\mathrm{w},m} \\ &= x_m + x_{\mathrm{r},m} + v_{\mathrm{r},1,m} + v_{\mathrm{r},2,m} + v_{\mathrm{d},m} + v_{\mathrm{w},m}, \end{aligned} \quad (63)$$

where $*$ is the linear convolution operator, x_{d} is the desired speech signal, $v_{\mathrm{i},1}$ and $v_{\mathrm{i},2}$ are, respectively, the two directional interferences, $v_{\mathrm{d},m}$ is the additive diffuse noise and $v_{\mathrm{w},m}$ is white noise in microphone m. Additionally, $g_{\mathrm{d},m}$ is the RIR from the desired signal source to microphone m, $g_{\mathrm{i},1,m}$ and $g_{\mathrm{i},2,m}$ are, respectively, the RIR from the directional interference sources to microphone m, whereas x_m, $x_{\mathrm{r},m}$, $v_{\mathrm{r},1,m}$ and $v_{\mathrm{r},2,m}$ are the direct path desired signal, its reverberations and the two reverberant interferences as received in microphone m, respectively. Lastly, we define the same components of the second row of (63) with respect to the time-domain enhanced speech signals by using the subscript f. For example, x_{f} is the enhanced direct path desired signal and $v_{\mathrm{w,f}}$ is the white noise component in the enhanced speech signal.

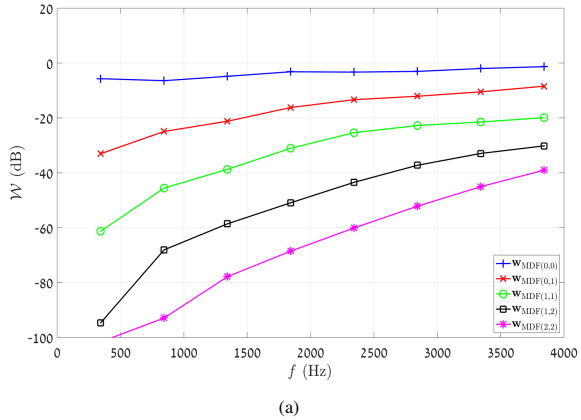
We now address the white thermal noise and the diffuse noise. Using the notations of (63), we define the Diffuse Noise Reduction (DNR) factor by:

$$\mathrm{DNR} = \frac{E\left[v_{\mathrm{d},1}^2\right]}{E\left[v_{\mathrm{d},f}^2\right]}, \quad (64)$$

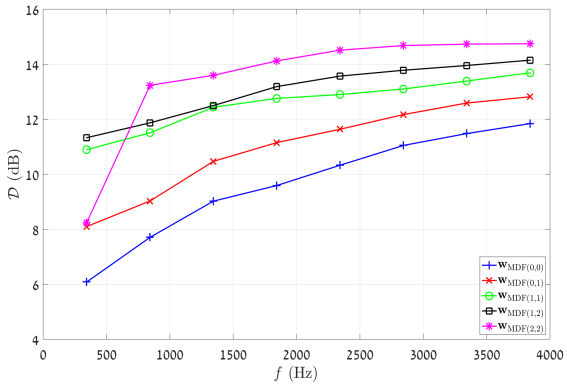
and the White Noise Reduction (WNR) factor by:

$$\mathrm{WNR} = \frac{E\left[v_{\mathrm{w},1}^2\right]}{E\left[v_{\mathrm{w},f}^2\right]}. \quad (65)$$

To demonstrate the advantages of the proposed multistage rectangular differential approach compared to the linear approach of [27], in the following, we compare the performance the aforementioned URA to the performance a ULA consisting of $M_{\mathsf{x}} \times M_{\mathsf{y}} = 20 \times 1$ microphones. The URA, ULA and each of the presented beamformers appear in Table 3. We note that while the optimal (P_c, P_r) configurations with the URA are taken from Tables 1 and 2, P_r is strictly zero

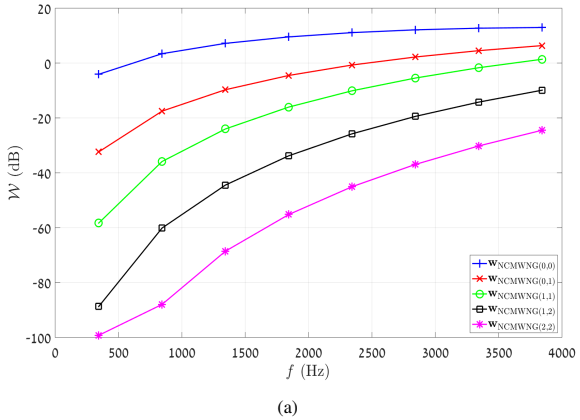


(a)

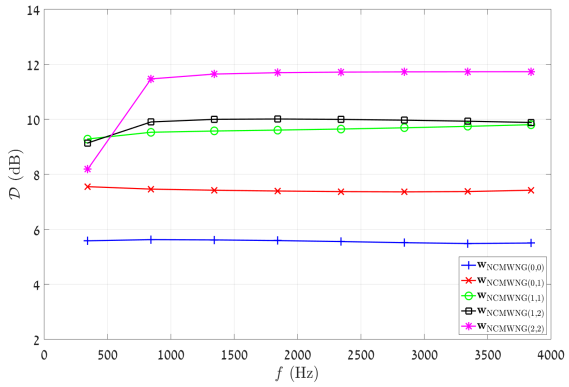


(b)

Figure 3: WNG and DF measures with the MDF rectangular differential beamformer, $\mathbf{w}_{\mathrm{MDF}(P_c,P_r)}$, with varying values of (P_c, P_r). Simulation parameters: $\theta_{\mathrm{s}} = 45^o$, $M_{\mathsf{x}} = 5$, $M_{\mathsf{y}} = 4$, $\delta_{\mathsf{x}} = 1$ cm and $\delta_{\mathsf{y}} = 1.2$ cm. (a) WNG and (b) DF.



(a)



(b)

Figure 4: WNG and DF measures with the null-constrained MWNG rectangular differential beamformer, $\mathbf{w}_{\mathrm{NCMWNG}(P_c,P_r)}$, with varying values of (P_c, P_r). Simulation parameters: $\theta_{\mathrm{s}} = 45^o$, $M_{\mathsf{x}} = 5$, $M_{\mathsf{y}} = 4$, $\delta_{\mathsf{x}} = 1$ cm and $\delta_{\mathsf{y}} = 1.2$ cm. (a) WNG and (b) DF.

for the ULA, implying that P_c exclusively determines the directivity of the array. To begin with, it is shown that the DNR is typically maximized with $\mathbf{w}_{\mathrm{MDF}(P_c,P_r)}$, whereas the DNR with $\mathbf{w}_{\mathrm{NCMDF}(P_c,P_r)}$ is superior to the DNR with both $\mathbf{w}_{\mathrm{MWNG}(P_c,P_r)}$ and $\mathbf{w}_{\mathrm{NCMWNG}(P_c,P_r)}$. Addressing the URA, we observe that the higher the (P_c, P_r) configuration, the preferable the DNR. In addition, the values of the DNR remain roughly similar regardless of the incident angle θ_{s}. On the contrary, with the ULA, the DNR performance deteriorates as θ_{s} deviates from the endfire direction: when $\theta_{\mathrm{s}} = 15^o$, the DNR is better than the DNR with the URA, and it improves upon increasing the (P_c, P_r) configuration; when $\theta_{\mathrm{s}} = 45^o$ the DNR is significantly worse than in the former case and compared to the URA, and when $\theta_{\mathrm{s}} = 75^o$ the DNR worsens even further, and might also turn negative (which implies diffuse noise amplification) with $P_c = 2$.

Let us focus on the WNR whose values with the discussed beamformers appear in Table 4. It is worth noting that when $(P_c, P_r) = (0, 0)$ white noise is typically attenuated. However, with other configurations, white noise might be significantly amplified. We note that this is in accordance with Section 6.1: the higher the configuration, the worse the WNR. Therefore, we infer that, in practice, whenever the microphones suffer from considerable inherent imperfections, the (P_c, P_r) configuration should be kept low.

Next, we define the desired signal Reverberations Reduction (RR) factor by:

$$\mathrm{RR} = \frac{E\left[x_{\mathrm{r},1}^2\right]}{E\left[x_{\mathrm{r},f}^2\right]}, \tag{66}$$

and the Interference Reduction (IR) factor by:

$$\mathrm{IR} = \frac{E\left[v_{\mathrm{r},1,1}^2\right] + E\left[v_{\mathrm{r},2,1}^2\right]}{E\left[v_{\mathrm{r},1,f}^2\right] + E\left[v_{\mathrm{r},2,f}^2\right]}. \tag{67}$$

We note that both factors are a function of the RIRs instead of the DNR and WNR. The RR and IR with the discussed URA, ULA, and their corresponding beamformers are shown in Table 5 and Table 6, respectively. We begin by addressing the RRs. It is clear that, roughly, the RR with the ULA is slightly better than with the URA for $\theta_{\mathrm{s}} = 15^o$. However, for $\theta_{\mathrm{s}} = 45^o$, and in particular for $\theta_{\mathrm{s}} = 75^o$- the URA is superior. In addition, unlike with the WNR and DNR, the RR performance does not monotonically improve or worsen upon increasing the (P_c, P_r) configuration as it highly depends on the beampatterns and the

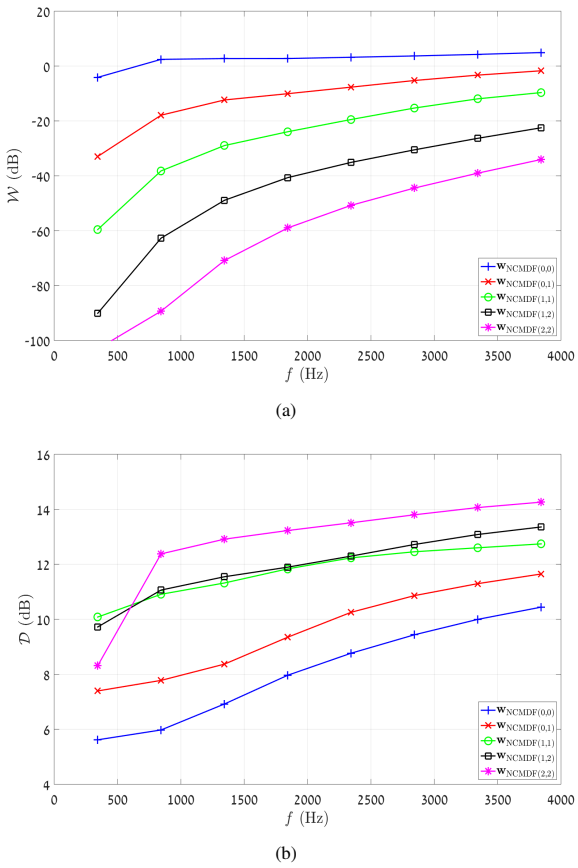


Figure 5: WNG and DF measures with the null-constrained MDF rectangular differential beamformer, $\mathbf{w}_{\mathrm{NCMDF}(P_c,P_r)}$, with varying values of (P_c, P_r). Simulation parameters: $\theta_{\mathrm{s}} = 45^o$, $M_{\mathsf{x}} = 5$, $M_{\mathsf{y}} = 4$, $\delta_{\mathsf{x}} = 1$ cm and $\delta_{\mathsf{y}} = 1.2$ cm. (a) WNG and (b) DF.

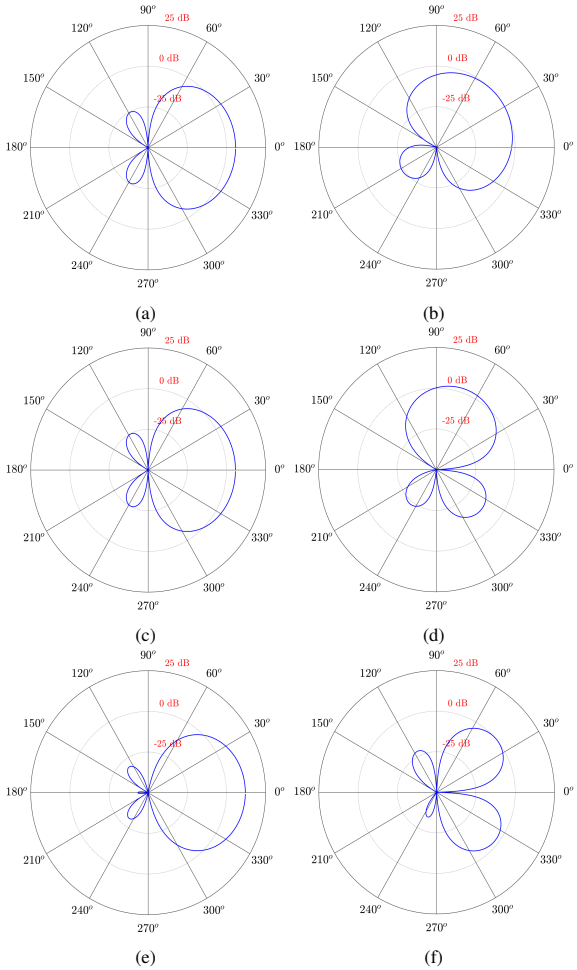


Figure 6: Power beampatterns of the NCMWNG rectangular differential beamformer, $\mathbf{w}_{\mathrm{NCMWNG}(P_c,P_r)}$, for $M_{\mathsf{x}}M_{\mathsf{y}} = 20$ microphones, two array geometries and three varying values of (P_c, P_r) with each geometry. (a) $M_{\mathsf{x}} = 20$, $M_{\mathsf{y}} = 1$ and $(P_c, P_r) = (0, 0)$, (b) $M_{\mathsf{x}} = 5$, $M_{\mathsf{y}} = 4$ and $(P_c, P_r) = (0, 0)$, (c) $M_{\mathsf{x}} = 20$, $M_{\mathsf{y}} = 1$ and $(P_c, P_r) = (1, 0)$, (d) $M_{\mathsf{x}} = 5$, $M_{\mathsf{y}} = 4$ and $(P_c, P_r) = (0, 1)$, (e) $M_{\mathsf{x}} = 20$, $M_{\mathsf{y}} = 1$ and $(P_c, P_r) = (2, 0)$, and (f) $M_{\mathsf{x}} = 5$, $M_{\mathsf{y}} = 4$ and $(P_c, P_r) = (1, 1)$. Simulation parameters: $\theta_{\mathrm{s}} = 45^0$, $\delta_{\mathsf{x}} = 1$ cm, $\delta_{\mathsf{y}} = 1.2$ cm and $f = 1$ kHz.

specific reverberations regime. However, we note that with the URA, it is highly likely for the RR to improve by modifying from $(P_c, P_r) = (0, 0)$ to the successive configuration. This is true for all examined values of θ_{s}. On the contrary, with the ULA, when $\theta_{\mathrm{s}} = 15^o$, the RR does improve by such a modification; when $\theta_{\mathrm{s}} = 45^o$ the RR improvement depends on the type of the beamformer, and when $\theta_{\mathrm{s}} = 75^o$, the RR typically worsens. We infer that to obtain a reduction in the reverberations of the desired signal with considerable beam steering requirements, a multistage differential URA should be preferred over a multistage differential ULA and its configuration should not be set to $(P_c, P_r) = (0, 0)$. Addressing the IR, we note that a similar analysis applies for this measure as well. However, it is shown that except for with $\mathbf{w}_{\mathrm{MWNG}(P_c,P_r)}$, the IR with the URA is typically superior to the IR with the ULA- even for $\theta_{\mathrm{s}} = 15^o$.

Next, we turn to analyze the power beampatterns of the NCMWNG beamformer with the URA, ULA, and their three corresponding configurations for $\theta_{\mathrm{s}} = 45^o$, which are depicted in Fig. 6 for $f = 1$ kHz. For starters, we observe that with the ULA, the maximum of all three beampatterns is not at 45^o but rather at the endfire direction. This implies a potential amplification of diffuse noise and reverberations impinging on the array from all directions between -45^o and 45^o. We also note that the higher the (P_c, P_r) configuration, the more significant the undesirable amplification at the endfire direction. This phenomenon translates into negative RR and IR values with $(P_c, P_r) = (2, 0)$, as previously elaborated. On the contrary, with the URA, the main lobe is directly steered to 45^o with $(P_c, P_r) = (0, 0)$ and $(P_c, P_r) = (1, 1)$, implying no amplification in undesirable directions exists, whereas with $(P_c, P_r) = (0, 1)$ some minor such amplification exists between 45^o and 90^o.

We end this section by analyzing the average PESQ [33] and STOI [34] scores of the time-domain enhanced signals with $\mathbf{w}_{\mathrm{NCMWNG}(P_c,P_r)}$ and $\mathbf{w}_{\mathrm{NCMDF}(P_c,P_r)}$ with both the URA and ULA, and with all three analyzed incident angles. The corresponding scores are depicted in Fig. 7. To begin with, we observe that in terms of both scores and with all three incident angles- superior performance is obtained with the URA,

Table 3: The DNR in dB units for the three investigated values of θ_{s}, three configurations of the ULA with $M_{\mathsf{x}} = 20$ and $M_{\mathsf{y}} = 1$, and three configurations of the URA with $M_{\mathsf{x}} = 5$ and $M_{\mathsf{y}} = 4$. Simulation parameters: $\delta_{\mathsf{x}} = 1$ cm and $\delta_{\mathsf{y}} = 1.2$ cm.

	$\theta_{\mathrm{s}} = 15^o$						$\theta_{\mathrm{s}} = 45^o$						$\theta_{\mathrm{s}} = 75^o$					
	$M_{\mathsf{x}} = 20, M_{\mathsf{y}} = 1$			$M_{\mathsf{x}} = 5, M_{\mathsf{y}} = 4$			$M_{\mathsf{x}} = 20, M_{\mathsf{y}} = 1$			$M_{\mathsf{x}} = 5, M_{\mathsf{y}} = 4$			$M_{\mathsf{x}} = 20, M_{\mathsf{y}} = 1$			$M_{\mathsf{x}} = 5, M_{\mathsf{y}} = 4$		
$(P_c, P_r) =$	$(0,0)$	$(1,0)$	$(2,0)$	$(0,0)$	$(1,0)$	$(2,0)$	$(0,0)$	$(1,0)$	$(2,0)$	$(0,0)$	$(0,1)$	$(1,1)$	$(0,0)$	$(1,0)$	$(2,0)$	$(0,0)$	$(0,1)$	$(0,2)$
$\mathbf{w}_{\mathrm{MWNG}(P_c,P_r)}$	6.8	8.5	9.8	3.8	6.3	7.6	5.5	6.1	5.4	3.8	4.7	7.5	5.0	0.0	-8.9	3.7	6.2	7.4
$\mathbf{w}_{\mathrm{MDF}(P_c,P_r)}$	**11.7**	**12.6**	**12.9**	**10.2**	**12.0**	**12.6**	7.5	**8.0**	**8.1**	**10.3**	**11.5**	**13.0**	**6.6**	4.1	-3.1	**10.1**	**11.6**	**11.9**
$\mathbf{w}_{\mathrm{NCMWNG}(P_c,P_r)}$	8.7	9.2	10.2	6.3	8.5	10.0	6.7	6.6	5.8	6.4	8.0	9.9	1.8	0.3	-7.5	5.9	8.9	10.9
$\mathbf{w}_{\mathrm{NCMDF}(P_c,P_r)}$	10.5	11.3	12.0	8.5	10.2	11.6	**7.8**	7.7	7.7	8.6	9.8	11.6	6.4	**4.8**	**6.5**	8.3	10.3	11.8

Table 4: The WNR in dB units for the three investigated values of θ_{s}, three configurations of the ULA with $M_{\mathsf{x}} = 20$ and $M_{\mathsf{y}} = 1$, and three configurations of the URA with $M_{\mathsf{x}} = 5$ and $M_{\mathsf{y}} = 4$. Simulation parameters: $\delta_{\mathsf{x}} = 1$ cm and $\delta_{\mathsf{y}} = 1.2$ cm.

	$\theta_{\mathrm{s}} = 15^o$						$\theta_{\mathrm{s}} = 45^o$						$\theta_{\mathrm{s}} = 75^o$					
	$M_{\mathsf{x}} = 20, M_{\mathsf{y}} = 1$			$M_{\mathsf{x}} = 5, M_{\mathsf{y}} = 4$			$M_{\mathsf{x}} = 20, M_{\mathsf{y}} = 1$			$M_{\mathsf{x}} = 5, M_{\mathsf{y}} = 4$			$M_{\mathsf{x}} = 20, M_{\mathsf{y}} = 1$			$M_{\mathsf{x}} = 5, M_{\mathsf{y}} = 4$		
$(P_c, P_r) =$	$(0,0)$	$(1,0)$	$(2,0)$	$(0,0)$	$(1,0)$	$(2,0)$	$(0,0)$	$(1,0)$	$(2,0)$	$(0,0)$	$(0,1)$	$(1,1)$	$(0,0)$	$(1,0)$	$(2,0)$	$(0,0)$	$(0,1)$	$(0,2)$
$\mathbf{w}_{\mathrm{MWNG}(P_c,P_r)}$	**12.9**	**5.0**	**-34.9**	**12.9**	**-6.3**	**-53.1**	**12.9**	**2.6**	**-40.3**	**12.9**	**-9.3**	**-51.6**	**12.9**	**-5.8**	-57.7	**12.9**	**-6.6**	**-54.5**
$\mathbf{w}_{\mathrm{MDF}(P_c,P_r)}$	1.1	-25.9	-67.6	-2.5	-33.1	-71.8	0.4	-28.0	-71.7	-2.5	-33.7	-69.6	-0.5	-35.0	-83.6	-2.5	-32.4	-70.7
$\mathbf{w}_{\mathrm{NCMWNG}(P_c,P_r)}$	7.1	-19.7	-63.0	4.2	-27.2	-65.7	5.9	-22.7	-68.1	4.2	-29.1	-66.1	2.5	-26.4	-60.1	3.7	-27.0	-65.2
$\mathbf{w}_{\mathrm{NCMDF}(P_c,P_r)}$	10.2	-17.9	-63.9	7.3	-20.5	-60.7	10.1	-20.0	-67.0	7.3	-20.0	-58.1	9.7	-19.9	**-54.2**	7.0	-19.2	-57.7

that is, by applying a multistage rectangular differential beamformer. Moreover, considering the multistage linear differential beamformers of the ULA, it is evident that high values of P_c degrade the PESQ and STOI scores. In contrast, with the URA higher (P_c, P_r) configurations improve these scores. For example, for $\theta_{\mathrm{s}} = 15^o$, with $(P_c, P_r) = (2, 0)$ the PESQ score is significantly higher with both rectangular beamformers with respect to with $(P_c, P_r) = (0, 0)$, and for $\theta_{\mathrm{s}} = 45^o$ both scores are maximized with $\mathbf{w}_{\mathrm{NCMWNG}(1,1)}$. On the contrary, for $\theta_{\mathrm{s}} = 75^o$, both rectangular beamformers considerably degrade speech signal with $(P_c, P_r) = (0, 2)$, with great accordance, for instance, to the IR measure discussed previously. We infer that considering beam steering, the multistage rectangular differential approach outperforms the multistage linear differential approach in terms of enhanced speech quality and intelligibility.

7. Conclusions

In this paper, we have presented a multistage rectangular approach for steerable differential beamforming. As a first step, we have proposed to employ a 2-D differentiation scheme that operates independently on the columns and rows of the observation signals of the rectangular array. This yields a differentials matrix of the noisy observations. Then, as a second step, a rectangular differential beamformer is designed and applied to the vector form of the differentials matrix. We have shown that the first differentiation scheme may significantly improve the directivity of the resulting beamformer at the expense of white noise amplification. The latter depends on an appropriate selection of the (P_c, P_r) configuration, which is optimized compared to the desired signal incident angle. For example, when the incident angle is close to the endfire direction, P_c should be larger than P_r; when the incident angle is close to the broadside direction, P_r should be larger than P_c; and when the incident angle is 45^0, P_c and P_r should be roughly equal. We have proposed four rectangular differential beamforming: two maximum SNR gain beamformers, $\mathbf{w}_{\mathrm{MWNG}(P_c,P_r)}$ and $\mathbf{w}_{\mathrm{MDF}(P_c,P_r)}$, as well as their two null-constrained versions, $\mathbf{w}_{\mathrm{NCMWNG}(P_c,P_r)}$ and $\mathbf{w}_{\mathrm{NCMDF}(P_c,P_r)}$. We have analyzed their performance in terms of both the WNG and DF measures for different (P_c, P_r) configurations and demonstrated that, indeed, a configuration change significantly influences the WNG-DF performance trade-off. Finally, addressing reverberant scenarios, we have analyzed and compared the WNR, DNR, RR, and IR of all the presented beamformers. We have shown that when the desired signal incident angle is not close to the endfire direction, the DNR, RR, and IR are highly preferable with the proposed rectangular differential approach compared to the former linear differential approach. Furthermore, considering average PESQ and STOI scores of time-domain enhanced signals, we have demonstrated the null-constrained rectangular differential beamformers to outperform their linear counterparts- for all three analyzed incident angles.

References

[1] D. H. Johnson, D. E. Dudgeon, Array Signal Processing: Concepts and Techniques, Simon and Schuster, Inc., USA, 1992 (1992).

[2] H. Van Trees, Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory, Detection, Estimation, and Modulation Theory, Wiley, New York, 2004 (2004).

[3] M. Bai, J. IH, J. Benesty, Acoustic Array Systems: Theory, Implementation, and Application, Wiley-IEEE Press, New York, 2014 (2014).

[4] G. W. Elko, Differential Microphone Arrays, Springer US, Boston, MA, 2004, pp. 11–65 (2004).

[5] M. Kolundzija, C. Faller, M. Vetterli, Spatiotemporal gradient analysis of differential microphone arrays, journal of the audio engineering society 59 (1) (2011) 20–28 (Jan 2011).

[6] E. D. Sena, H. Hacihabiboglu, Z. Cvetkovic, On the design and implementation of higher order differential microphones, IEEE Transactions on Audio, Speech, and Language Processing 20 (1) (2012) 162–174 (2012).

[7] J. Benesty, J. Chen, Study and Design of Differential Microphone Arrays, Springer-Verlag Berlin Heidelberg, Berlin, 2013 (2013).

[8] J. Chen, J. Benesty, C. Pan, On the design and implementation of linear differential microphone arrays, The Journal of the Acoustical Society of America 136 (6) (2014) 3097–3113 (2014).

Table 5: The RR in dB units for the three investigated values of $\theta_{\rm s}$, three configurations of the ULA with $M_{\rm x}=20$ and $M_{\rm y}=1$, and three configurations of the URA with $M_{\rm x}=5$ and $M_{\rm y}=4$. Simulation parameters: $\delta_{\rm x}=1$ cm, $\delta_{\rm y}=1.2$ cm, and $T_{60}=600$msec.

	$\theta_{\rm s}=15^o$						$\theta_{\rm s}=45^o$						$\theta_{\rm s}=75^o$					
	$M_{\rm x}=20, M_{\rm y}=1$			$M_{\rm x}=5, M_{\rm y}=4$			$M_{\rm x}=20, M_{\rm y}=1$			$M_{\rm x}=5, M_{\rm y}=4$			$M_{\rm x}=20, M_{\rm y}=1$			$M_{\rm x}=5, M_{\rm y}=4$		
$(P_c,P_r)=$	$(0,0)$	$(1,0)$	$(2,0)$	$(0,0)$	$(1,0)$	$(2,0)$	$(0,0)$	$(1,0)$	$(2,0)$	$(0,0)$	$(0,1)$	$(1,1)$	$(0,0)$	$(1,0)$	$(2,0)$	$(0,0)$	$(0,1)$	$(0,2)$
$\mathbf{w}_{\rm MWNG}(P_c,P_r)$	1.4	5.5	**7.3**	0.1	4.6	**4.6**	-0.3	3.0	**1.8**	0.0	3.6	6.2	-0.3	-3.7	-14.2	0.1	4.6	4.5
$\mathbf{w}_{\rm MDF}(P_c,P_r)$	**8.2**	**8.8**	-0.2	**3.8**	6.6	-10.7	2.7	3.5	-3.0	3.7	6.5	-1.4	1.4	**4.0**	-12.9	**3.7**	6.8	4.4
$\mathbf{w}_{\rm NCMWNG}(P_c,P_r)$	6.3	6.4	-4.8	2.4	6.5	-2.4	3.3	3.3	-9.4	2.9	5.9	2.1	-3.5	-5.6	-14.4	1.7	6.5	4.4
$\mathbf{w}_{\rm NCMDF}(P_c,P_r)$	6.7	7.4	2.9	3.4	**7.2**	0.9	**4.1**	**4.1**	0.0	**3.9**	**7.3**	**7.1**	**2.6**	-0.1	**4.1**	3.0	**7.5**	**8.9**

Table 6: The IR in dB units for the three investigated values of $\theta_{\rm s}$, three configurations of the ULA with $M_{\rm x}=20$ and $M_{\rm y}=1$, and three configurations of the URA with $M_{\rm x}=5$ and $M_{\rm y}=4$. Simulation parameters: $\delta_{\rm x}=1$ cm, $\delta_{\rm y}=1.2$ cm, and $T_{60}=600$msec.

	$\theta_{\rm s}=15^o$						$\theta_{\rm s}=45^o$						$\theta_{\rm s}=75^o$					
	$M_{\rm x}=20, M_{\rm y}=1$			$M_{\rm x}=5, M_{\rm y}=4$			$M_{\rm x}=20, M_{\rm y}=1$			$M_{\rm x}=5, M_{\rm y}=4$			$M_{\rm x}=20, M_{\rm y}=1$			$M_{\rm x}=5, M_{\rm y}=4$		
$(P_c,P_r)=$	$(0,0)$	$(1,0)$	$(2,0)$	$(0,0)$	$(1,0)$	$(2,0)$	$(0,0)$	$(1,0)$	$(2,0)$	$(0,0)$	$(0,1)$	$(1,1)$	$(0,0)$	$(1,0)$	$(2,0)$	$(0,0)$	$(0,1)$	$(0,2)$
$\mathbf{w}_{\rm MWNG}(P_c,P_r)$	7.8	11.6	**10.0**	4.6	10.8	1.9	7.0	10.2	**5.7**	5.1	2.5	7.7	4.5	5.6	-10.3	5.0	4.3	-7.1
$\mathbf{w}_{\rm MDF}(P_c,P_r)$	**12.7**	12.7	-0.9	**13.2**	**15.7**	0.2	10.4	10.8	-4.3	**13.6**	13.3	3.4	6.4	9.4	-14.5	11.5	12.1	-10.7
$\mathbf{w}_{\rm NCMWNG}(P_c,P_r)$	11.6	11.6	-5.6	10.1	11.9	4.8	10.6	9.9	-10.7	10.7	12.0	6.4	7.0	4.6	-5.1	9.8	11.8	-4.1
$\mathbf{w}_{\rm NCMDF}(P_c,P_r)$	**12.7**	**13.2**	2.2	12.2	13.8	**5.5**	**11.2**	**11.2**	-0.3	13.2	**13.7**	**9.1**	**11.7**	**10.7**	**10.1**	**11.7**	**12.9**	**-1.8**

[9] A. Bernardini, F. Antonacci, A. Sarti, Wave digital implementation of robust first-order differential microphone arrays, IEEE Signal Processing Letters 25 (2) (2018) 253–257 (2018).

[10] J. Weinberger, H. Olson, F. Massa, A uni-directional ribbon microphone, The Journal of the Acoustical Society of America 5 (2) (1933) 139–147 (1933).

[11] H. F. Olson, Gradient microphones, The Journal of the Acoustical Society of America 17 (3) (1946) 192–198 (1946).

[12] M. Buck, Aspects of first-order differential microphone arrays in the presence of sensor imperfections, European Transactions on Telecommunications 13 (2002) 115–122 (2002).

[13] X. Wu, H. Chen, Directivity factors of the first-order steerable differential array with microphone mismatches: Deterministic and worst-case analysis, IEEE/ACM Transactions on Audio, Speech, and Language Processing 24 (2) (2016) 300–315 (2016).

[14] J. Benesty, J. Chen, E. Habets, Speech Enhancement in the STFT Domain, Springer-Verlag Berlin Heidelberg, Berlin, 2012 (2012).

[15] J. Benesty, I. Cohen, J. Chen, Fundamentals of Signal Enhancement and Array Signal Processing, 1st Edition, Wiley-IEEE Press, New York, 2017 (2017).

[16] J. Benesty, J. Chen, Y. Huang, I. Cohen, Noise Reduction in Speech Processing, 1st Edition, Springer-Verlag Berlin Heidelberg, Berlin, 2009 (2009).

[17] F. Borra, A. Bernardini, F. Antonacci, A. Sarti, Uniform linear arrays of first-order steerable differential microphones, IEEE/ACM Transactions on Audio, Speech, and Language Processing 27 (12) (2019) 1906–1918 (2019).

[18] J. Benesty, I. Cohen, J. Chen, Array Processing - Kronecker Product Beamforming, Springer-Verlag, Switzerland, 2019 (2019).

[19] I. Cohen, J. Benesty, J. Chen, Differential kronecker product beamforming, IEEE/ACM Transactions on Audio, Speech, and Language Processing 27 (5) (2019) 892–902 (2019).

[20] G. Huang, J. Benesty, J. Chen, Design of robust concentric circular differential microphone arrays, The Journal of the Acoustical Society of America 141 (5) (2017) 3236–3249 (2017).

[21] Y. Buchris, I. Cohen, J. Benesty, Frequency-domain design of asymmetric circular differential microphone arrays, IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (4) (2018) 760–773 (2018).

[22] G. Huang, J. Chen, J. Benesty, Design of planar differential microphone arrays with fractional orders, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020) 116–130 (2020).

[23] G. Huang, J. Chen, J. Benesty, On the design of differential beamformers with arbitrary planar microphone array geometry, The Journal of the Acoustical Society of America 144 (1) (2018) EL66–EL70 (2018).

[24] G. Huang, J. Chen, J. Benesty, I. Cohen, X. Zhao, Steerable differential beamformers with planar microphone arrays, EURASIP Journal on Audio, Speech, and Music Processing 2020 (2020).

[25] G. Huang, J. Benesty, J. Chen, I. Cohen, Robust and steerable kronecker product differential beamforming with rectangular microphone arrays, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 211–215 (2020).

[26] G. Itzhak, I. Cohen, J. Benesty, Robust differential beamforming with rectangular arrays, in: Proc. 29th European Signal Processing Conference, EUSIPCO-2021, 2021 (Aug 2021).

[27] G. Huang, J. Benesty, I. Cohen, J. Chen, A simple theory and new method of differential beamforming with uniform linear microphone arrays, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020) 1079–1093 (2020).

[28] G. Itzhak, J. Benesty, I. Cohen, On the design of differential kronecker product beamformers, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 1397–1410 (2021).

[29] E. A. P. Habets, Room impulse response (rir) generator (2008).

[30] J. B. Allen, D. A. Berkley, Image method for efficiently simulating small-room acoustics, J. Acoust. Soc. Am 65 (4) (1979) 943–950 (1979).

[31] A. Pierce, Acoustics: An Introduction to Its Physical Principles and Applications, Springer International Publishing, Switzerland, 2019 (2019).

[32] Darpa timit acoustic phonetic continuous speech corpus cdrom (1993).

[33] A. W. Rix, J. G. Beerends, M. P. Hollier, A. P. Hekstra, Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs, in: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings, Vol. 2, 2001, pp. 749–752 vol.2 (2001).

[34] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time–frequency weighted noisy speech, IEEE Transactions on Audio, Speech, and Language Processing 19 (7) (2011) 2125–2136 (Sep 2011).

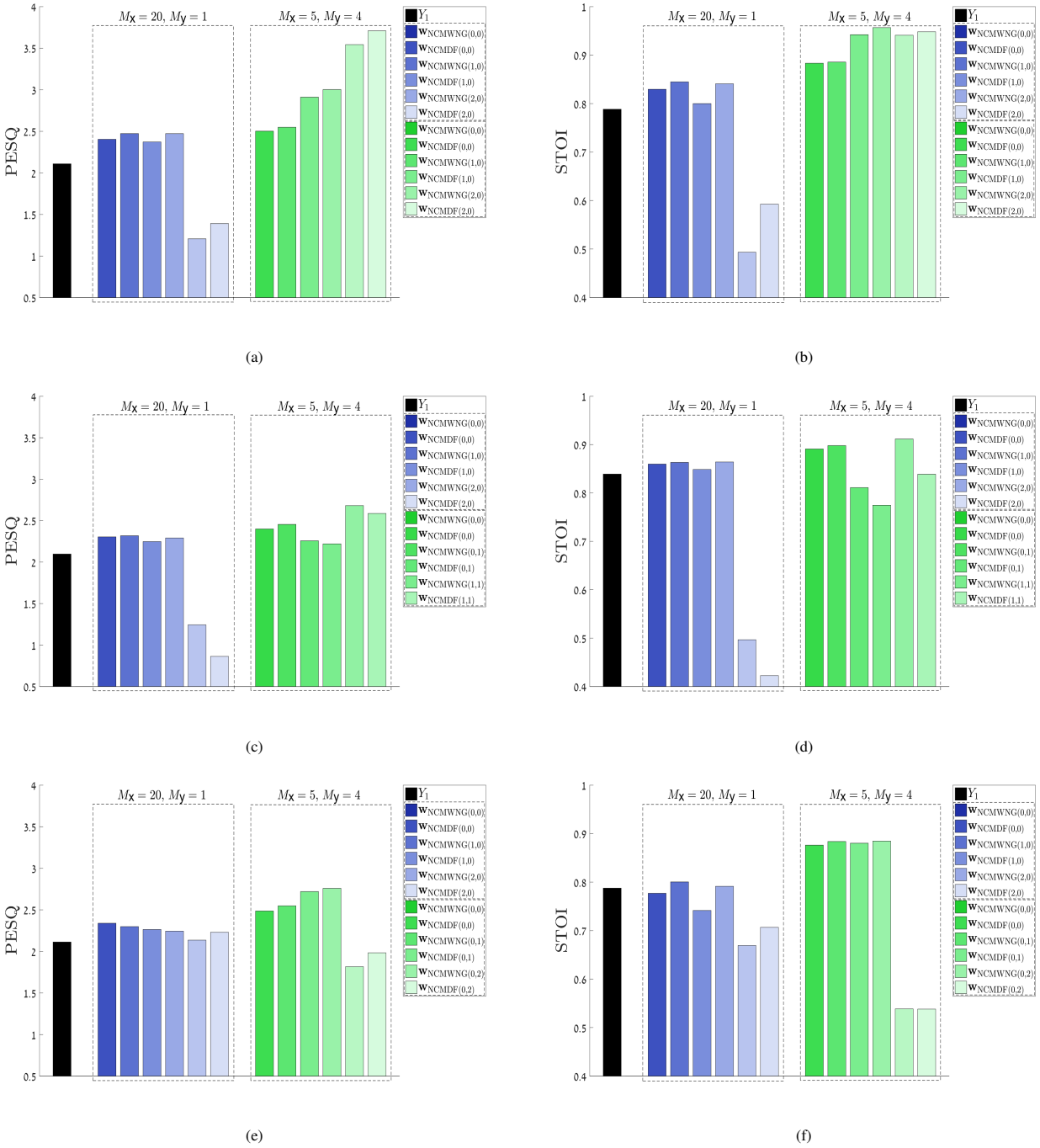


Figure 7: Average PESQ and STOI scores with the NCMWNG and NCMDF rectangular differential beamformers, $\mathbf{w}_{\mathrm{NCMWNG}(P_c,P_r)}$ and $\mathbf{w}_{\mathrm{NCMDF}(P_c,P_r)}$, respectively, for three values of the incident angle θ_{s}, two array geometries and three varying values of (P_c, P_r) with each geometry. (a) Average PESQ scores and $\theta_{\mathrm{s}} = 15^0$, (b) average STOI scores and $\theta_{\mathrm{s}} = 15^0$, (c) average PESQ scores and $\theta_{\mathrm{s}} = 45^0$, (d) average STOI scores and $\theta_{\mathrm{s}} = 45^0$, (e) average PESQ scores and $\theta_{\mathrm{s}} = 75^0$, and (f) average STOI scores and $\theta_{\mathrm{s}} = 75^0$. Simulation parameters: $\delta_{\mathrm{x}} = 1$ cm and $\delta_{\mathrm{y}} = 1.2$ cm.

Chapter 7

Discussion and Conclusions

7.1 Discussion and Conclusions

This research thesis introduces quadratic and multistage approaches for beamforming. Their advantages are embodied in their preferable noise reduction performances and design flexibility with respect to common existing beamforming approaches.

Traditionally, reducing background noise and interferences from noisy samples has been a fundamental task in the fields of signal and speech processing. Nowadays, with the growing demand for communication automation and control, it has become even more essential. Depending on the application, this task may be carried out in different settings (single microphone/microphone array), various types and intensity of noise, and additional desirable properties, such as beam steering.

To accomplish the noise reduction task, numerous approaches and schemes have been proposed in the literature over the years. For starters, with microphone arrays, it is very common to employ linear filtering to a vector of noisy samples as the derivation and evaluation of the filter (beamformer) is the easiest to design and implement. Unfortunately, with linear approaches, the noise reduction potential is strictly limited as merely the second-order statistics of the desired signal and noise is exploited. This implies that more sophisticated approaches, which take advantage of higher-order statistics, may potentially achieve a significantly preferable noise reduction. On the contrary, with single microphone settings, desired speech estimators are typically highly nonlinear, explicitly assume an underlying statistical model of the desired speech and

rely on the strong correlation between magnitudes of successive analysis coefficients in a fixed frequency. Unfortunately, their derivation is typically cumbersome and requires numerical non-analytical function evaluations, which are resulted by the assumed statistical model. Moreover, with the aforementioned spectral magnitude correlation hidden behind first-order recursive temporal processes, additional parameters and lower boundaries must carefully be set to guarantee the model tracking over time.

In other cases, in particular when the array directivity is of a high importance, DMAs are used. Typically, in order to design high-order DMAs which were capable of obtaining a significant amount of noise reduction, a multistage approach was taken. That is, the operation of differentiating acoustic pressure observations was successively repeated, in analogy to high-order derivatives of analytic functions. This approach was originally implemented in the time domain, however, was highly susceptible to array mismatches and imperfections. As a consequence, a framework for design of DMA design in the STFT was introduced which is based on a single stage with linear matrix operations. Despite its simplicity, it is still capable of satisfying spatial constraints while simultaneously minimizing the residual noise. Nevertheless, its inherent single-stage and linear nature restricted the array directivity. A more recent approach, which employs a multistage structure for processing in the STFT domain was shown to be effective to reduce diffuse noise and to handle reverberant environments, albeit significantly more sensitive to white noise.

In the foregoing, the beam steering property has not been addressed, although its importance may be critical in many scenarios, i.e., when the desired signal incident angle is not a priori known. Invoking the well-known preference of linear DMAs to the endfire direction, it is clear that obtaining high array directivity in a wide range of incident angles is an actual challenge, in particular when the flexibility to control the trade-off between array directivity and white noise amplification is required.

To address these problems, this thesis proposes four beamforming approaches:

1) A KP filtering approach for MCNR is introduced. As opposed to linear approaches, it focuses on the estimate of the spectral power of the desired signal by exploiting higher-order statistics. This is carried out by linearly applying a quadratic beamformer to a modified vector of the noisy observations. This study proposes two

KP beamformers, the KP-MVDR and KP-LCMV, which may be seen as the quadratic counterparts of the celebrated linear MVDR and LCMV beamformers. It is shown that the quadratic beamformers are able to outperform the linear beamformers, achieve a greater noise attenuation, and as a consequence yield enhanced speech signals of a higher quality and intelligibility. This is emphasized in particular when the number of sensors is small or when the input SNR is low.

2) This study introduces a quadratic filtering approach for single-channel noise reduction, which takes advantage of the interframe correlation property and generalizes the conventional linear filtering approach. This approach exploits the quadratic formulation to focus on the desired signal power estimate and achieves a theoretically unbounded approximate SNR gain, depending on the estimation error of the second order statistics of the noise. The advantage of the quadratic approach are demonstrated by focusing on the maximum SNR beamformer in the STFT domain. It is shown that the quadratic maximum SNR beamformer outperforms the linear maximum SNR beamformer, in particular in low input SNRs, at the expanse of a higher computational complexity. In addition, a comparison to commonly used methods is conducted. The quadratic maximum SNR beamformer former is shown to perform better than these methods in some of the examined scenarios, even with naive desired signal and noise statistics estimation techniques.

3) This study presents a generalized framework for design of multistage differential beamforming. This is achieved by applying a KP decomposition to a global differential beamformer, and independently optimizing the two sub-beamformers. Previous non-differential or non-KP beamformers may be obtained by an appropriate selection of the array settings parameters. The study proposes five types of differential KP beamformers which are flexibly tuned by three design parameters. This flexibility enables one to mitigate the white noise amplification or improve the directivity, depending on the beamformer type. It is shown that with differential KP beamformers reverberations of the desired signal may be attenuated to a greater extent than with non-differential and non-KP beamformers, whereas the quality and intelligibility of the enhanced signals with the former are superior. This is validated even under array imperfections and is in particular true for moderately reverberant environments.

4) The study described in the former part is extended to allow steerable multistage differential beamforming by exploiting URAs. As a first step, the study proposes to employ a 2-D differentiation scheme which operates independently on the columns and rows of the observation signals of the rectangular array. This yields a differentials matrix of the noisy observations. Then, as a second step, a rectangular differential beamformer is designed and applied to the vector form of the differentials matrix. It is shown that the first differentiation scheme may significantly improve the directivity of the resulted beamformer at the expanse of white noise amplification. The latter depends on an appropriate selection of the design parameters configuration, which is optimized with respect to the desired signal incident angle. Simulations in reverberant environments and distinct incident angles indicate that undesirable reverberations and diffuse noise are better attenuated with the multistage rectangular differential approach, with respect to the multistage linear differential approach. As a consequence, beamformers of the former approach yield more intelligible enhanced signals whose quality is preferable.

7.2 Future Research Directions

Within this thesis, the notion of quadratic beamforming has been introduced and developed, and the concept of multistage differential beamforming has been generalized and extended. Following these advancements, further research can be performed:

1) **Higher-order beamforming**. The study described in this thesis indicates that under certain conditions quadratic beamforming outperforms the traditional linear beamforming. This is of a great importance, for example, in low SNR scenarios. We suggest to extend this concept one step further and perform higher-order beamforming ,i.e., cubic beamforming, fourth-order beamforming and so on. We believe that this would enable further reduction of background noise depending on its properties, by taking advantage of its higher-order statistics.

2) **Quadratic multistage differential beamforming**. This thesis has demonstrated that while multistage differential beamforming improves the array directivity, it evidently increases its sensitivity to white noise. On the contrary, we have shown

that quadratic beamforming allows further reduction of the background noise with respect to linear beamforming. Consequently, taking advantage of the two concepts by first applying the spatial difference operator and then applying a quadratic beamformer, quadratic multistage differential beamforming is likely to benefit from both: achieve high array directivity and mitigate white noise amplification, in particular in low frequencies.

3) **Directive constant-beamwidth beamforming**. In many cases of broadband signals of interest, the constant beamwidth property is essential. Unfortunately, constant-beamwidth (with respect to the azimuth angle) beamforming is typically characterized by poor array directivity. Therefore, a promising direction for research may emerge from the combination of a differential (or multistage differential) beamformer with a constant-beamwidth beamformer. This may be achieved, for example, by applying the KP to two linear beamformers, thus forming a combined rectangular beamformer: a differential beamformer located in the endfire direction with respect to the desired signal incident angle and a constant-beamwidth beamformer located in the broadside direction with respect to it.

4) **Directive and steerable constant-beamwidth beamforming**. While the former research lead is indeed promising, it lacks the property of beam steering, implying that the characteristics of the beamformer might heavily change as the desired signal incident angle deviates from the described direction (endfire with respect to the differential beamformer; broadside with respect to the constant-beamwidth beamformer). Consequently, we suggest to replace the linear constant-beamwidth beamformer with a uniform circular beamformer which is well-known to allow steerable beamforming but suffers from poor array directivity. We may also employ a concentric circular beamformer instead of the uniform circular beamformer to further improve the array directivity and potentially allow constant beamwidth with respect to the elevation angle as well. The array directivity may improve to an even greater extent by applying a multistage differentiation scheme across adjacent concentric rings.

Appendix A

Quadratic Beamforming for Magnitude Estimation

Quadratic Beamforming for Magnitude Estimation

Gal Itzhak[1], Jacob Benesty[2] and Israel Cohen[1]
[1]*Faculty of Electrical and Computer Engineering, Technion–Israel Institute of Technology*, Haifa 3200003, Israel
[2]*INRS-EMT, University of Quebec*, Montreal, QC H5A 1K6, Canada
galitz@campus.technion.ac.il; benesty@emt.inrs.ca; icohen@ee.technion.ac.il

***Abstract*—In this paper, we introduce an optimal quadratic Wiener beamformer for magnitude estimation of a desired signal. For simplicity, we focus on a two-microphone array and develop an iterative algorithm for magnitude estimation based on a quadratic multichannel noise reduction approach. We analyze two test cases, with uncorrelated and correlated noises. In each, we derive the appropriate versions of the Wiener beamformer, as well as their corresponding unbiased magnitude estimators. We compare the root-mean-squared errors (RMSEs) for the linear and quadratic Wiener beamformers and show that for low input signal-to-noise ratios (SNRs), the RMSE obtained with the proposed approach is either lower than or equal to the RMSE obtained with the linear Wiener beamformer, depending on the type of noise and its distribution.**

I. Introduction

Magnitude estimation of a desired signal of interest is a common task in a wide area of fields, including communications, target detection, and speech enhancement. The desired signal is typically only observable through noisy samples, that is, it is corrupted by noise, which may critically damage the performance of the application at hand. Consequently, a large number of studies have addressed this issue by exploiting data either from a single microphone or a sensor array.

Most commonly with communications and speech signals, processing is done in the frequency domain. That is, a frame of consecutive time-domain samples is transformed into the frequency domain by applying the fast Fourier transform (FFT), yielding a set of analysis coefficients, which can be processed more efficiently than the time-domain samples. This is particularly significant with multichannel methods, in which time-domain noisy observations are sampled simultaneously in multiple sensors. These methods typically seek for a linear optimal solution with respect to some criterion, looking to estimate both the desired signal phase and magnitude [1]–[4]. On the contrary, single-channel approaches may either attempt to estimate the complex desired signal [5]–[8] or may directly attempt to estimate its magnitude [9]–[17], which is known to be more prominent than its phase for some applications.

Recently, a quadratic noise reduction approach was suggested [17], [18]. The idea behind the quadratic approach is to estimate the spectral power of the desired signal by applying a complex-valued beamformer, which takes into consideration data from higher-order moments. The quadratic beamformer is applied to a modified version of the noisy observations vector and may be seen as a generalization of the traditional linear beamformer.

In this paper, we introduce a quadratic beamformer for a desired signal magnitude estimation, which is optimal in terms of the RMSE. For simplicity, we focus on a two-microphone array and develop an iterative algorithm for magnitude estimation based on the quadratic multichannel noise reduction approach. We analyze two test cases, with uncorrelated and correlated noises. In each case, we derive the appropriate version of the Wiener beamformer, as well as the corresponding unbiased magnitude estimator. We compare the RMSEs with the linear and quadratic Wiener beamformers and show that for low input SNRs, the RMSE obtained with the proposed approach is either lower than or equal to the RMSE obtained with the linear Wiener beamformer, depending on the type of noise and its distribution.

The rest of the paper is organized as follows. In Section II, we present the signal model. In Section III, we introduce the quadratic beamforming approach. In Section IV, we analyze two test cases, with both uncorrelated and correlated noises. We derive the quadratic optimal beamformers in terms of minimum RMSE and use them to derive unbiased magnitude estimators. Then, in Section V, we demonstrate the advantage of the quadratic approach over the linear one through simulations. Finally, we summarize this work in Section VI.

II. Signal Model

Consider an array consisting of M omnidirectional microphones. The received signals at the frequency index f are expressed as [3], [19]

$$Y_m(f) = X_m(f) + V_m(f),\ m = 1, 2, \ldots, M, \qquad (1)$$

where $Y_m(f)$ is the mth microphone signal, $X_m(f)$ is the zero-mean desired speech signal, and $V_m(f)$ is the zero-mean additive noise. It is assumed that the desired signal and noise are uncorrelated.

Considering the first microphone as the reference, we may express (1) in a vector notation:

$$\mathbf{y}(f) = \mathbf{d}_{\theta_{\rm d}}(f)X_1(f) + \mathbf{v}(f), \qquad (2)$$

where

$$\mathbf{y}(f) = [Y_1(f)\ \ Y_2(f)\ \ \cdots\ \ Y_M(f)]^T,$$

This research was supported by the ISF-NSFC joint research program (grant No. 2514/17), and the Pazy Research Foundation.

$\mathbf{v}(f)$ is defined in a similar manner to $\mathbf{y}(f)$, the superscript ^{T} is the transpose operator, and

$$\mathbf{d}_{\theta_\mathrm{d}}(f) = \left[1 \quad e^{-\jmath 2\pi f \delta \cos\theta_\mathrm{d}/c} \quad \cdots \quad e^{-\jmath 2\pi f \delta (M-1)\cos\theta_\mathrm{d}/c}\right]^T \tag{3}$$

is the frequency-domain steering vector, considering the farfield planar wave model [1], [2]. In addition, θ_d is the desired speech signal incident angle, δ is the inter-element spacing, $c = 340$ m/s is the speed of sound, and $\jmath = \sqrt{-1}$ is the imaginary unit.

Since $\mathbf{y}(f)$ is the sum of two uncorrelated components, its correlation matrix is

$$\begin{aligned}\mathbf{\Phi}_\mathbf{y}(f) &= E\left[\mathbf{y}(f)\mathbf{y}^H(f)\right] \\ &= \phi_{X_1}(f)\mathbf{d}_{\theta_\mathrm{d}}(f)\mathbf{d}_{\theta_\mathrm{d}}^H(f) + \mathbf{\Phi}_\mathbf{v}(f),\end{aligned} \tag{4}$$

where $E[\cdot]$ denotes mathematical expectation, the superscript ^{H} is the conjugate-transpose operator, $\phi_{X_1}(f) = E\left[|X_1(f)|^2\right]$ is the variance of $X_1(f)$, and $\mathbf{\Phi}_\mathbf{v}(f) = E\left[\mathbf{v}(f)\mathbf{v}^H(f)\right]$ is the 2nd-order correlation matrix of $\mathbf{v}(f)$ whose top-left element is $\phi_{V_1}(f) = E\left[|V_1(f)|^2\right]$.

III. Quadratic Beamforming

Conventionally, with an array of M sensors, beamforming is performed by applying a complex-valued linear filter, $\mathbf{h}(f)$ of length M, to the observation signal vector, $\mathbf{y}(f)$, i.e., [3], [19]

$$\begin{aligned}\widehat{X}(f) &= \mathbf{h}^H(f)\mathbf{y}(f) \\ &= X_1(f)\mathbf{h}^H(f)\mathbf{d}_{\theta_\mathrm{d}}(f) + \mathbf{h}^H(f)\mathbf{v}(f),\end{aligned} \tag{5}$$

where the filter output, $\widehat{X}(f)$, is an estimate of $X_1(f)$. We note that $\widehat{X}(f)$ is complex, that is, it carries information on both the magnitude and phase of the desired signal.

Recently, a quadratic noise reduction approach was suggested in [18]. According to this technique, we can estimate the spectral power of $\widehat{X}(f)$ defined in (5) for a given complex-valued beamformer $\widetilde{\mathbf{h}}(f)$ of length M^2 by

$$\left|\widehat{X}(f)\right|^2 = \widetilde{\mathbf{h}}^H(f)\widetilde{\mathbf{y}}(f), \tag{6}$$

where $\widetilde{\mathbf{y}}(f) = \mathbf{y}^*(f) \otimes \mathbf{y}(f)$, with the superscript * being the complex-conjugate operator and $\otimes$ the Kronecker product. Additionally, it was shown that

$$\begin{aligned}\phi_{\widehat{X}}(f) &= E\left[\left|\widetilde{\mathbf{h}}^H(f)\widetilde{\mathbf{y}}(f)\right|\right] \\ &\approx \left|\phi_{X_1}(f)\widetilde{\mathbf{h}}^H(f)\widetilde{\mathbf{d}}_{\theta_\mathrm{d}}(f) + \widetilde{\mathbf{h}}^H(f)\mathrm{vec}\left[\mathbf{\Phi}_\mathbf{v}(f)\right]\right|,\end{aligned} \tag{7}$$

where $\widetilde{\mathbf{d}}_{\theta_\mathrm{d}}(f) = \mathbf{d}_{\theta_\mathrm{d}}^*(f) \otimes \mathbf{d}_{\theta_\mathrm{d}}(f)$ is the quadratic steering vector and $\mathrm{vec}[\cdot]$ is the vectorization operator.

IV. Analysis of Two Test Cases

As of this point, for the sake of simplicity, let us assume that $M = 2$ and $\theta_\mathrm{d} = 90^o$ (it should be noted, though, that this approach is indeed general and not limited to certain array sizes or incident angles). Hence, (2) reduces to

$$\begin{aligned}\mathbf{y} &= \begin{bmatrix} X \\ X \end{bmatrix} + \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \\ &= \begin{bmatrix} ae^{\jmath\phi_a} \\ ae^{\jmath\phi_a} \end{bmatrix} + \begin{bmatrix} V_1 \\ V_2 \end{bmatrix},\end{aligned} \tag{8}$$

where a and ϕ_a are the magnitude and phase of the desired signal, respectively, and the explicit dependence on frequency is dropped to lighten the notation. In this model, we assume the desired signal is a deterministic and unknown variable. With a known to be more prominent than ϕ_a for speech enhancement purposes [10], we further assume that $\phi_a = 0$. Consequently, our objective is to derive optimal estimators for the real and positive variable a in two key cases: uncorrelated and correlated noises.

A. Uncorrelated Noise

Let us assume that V_1 and V_2 are independent, real, zero-mean, identically-distributed random variables whose variances is σ^2. Adopting the RMSE as the performance criterion, the optimal linear beamformer is given by the linear Wiener beamformer [20], [21]:

$$\begin{aligned}\mathbf{h}_\mathrm{W} &= \phi_X \mathbf{\Phi}_\mathbf{y}^{-1}\mathbf{d}_{\theta_\mathrm{d}} \\ &= a^2\mathbf{\Phi}_\mathbf{y}^{-1}\mathbf{1}_2 \\ &\doteq \left[H_\mathrm{W}(1) \quad H_\mathrm{W}(2)\right]^T,\end{aligned} \tag{9}$$

where $\mathbf{1}_2$ is an "all-ones" vector of length 2, and we assume that the correlation matrix $\mathbf{\Phi}_\mathbf{y}$ is either given or can be estimated from the noisy observations. Then, considering (8), an unbiased estimator for a is given by

$$\hat{a}\left[\mathbf{h}_\mathrm{W}\right] = \sqrt{\max\left\{\frac{|\mathbf{h}_\mathrm{W}^T\mathbf{y}|^2 - \sigma^2\|\mathbf{h}_\mathrm{W}\|^2}{|\mathbf{h}_\mathrm{W}^T\mathbf{1}_2|^2}, 0\right\}}, \tag{10}$$

where $\|\cdot\|$ is the Euclidean norm.

We follow a similar protocol with $\widetilde{\mathbf{h}}$ where (6) is employed. That is, we are interested in solving the following optimization criterion:

$$\min_{\widetilde{\mathbf{h}}} E\left|\widetilde{\mathbf{h}}^T\widetilde{\mathbf{y}} - a^2\right|^2, \tag{11}$$

whose solution is given by

$$\begin{aligned}\widetilde{\mathbf{h}}_\mathrm{W} &= a^2\mathbf{\Phi}_{\widetilde{\mathbf{y}}}^{-1}E\left(\widetilde{\mathbf{y}}\right) \\ &= a^2\mathbf{\Phi}_{\widetilde{\mathbf{y}}}^{-1}\begin{bmatrix} a^2+\sigma^2 \\ a^2 \\ a^2 \\ a^2+\sigma^2 \end{bmatrix} \\ &\doteq \left[\widetilde{H}_\mathrm{W}(1) \quad \widetilde{H}_\mathrm{W}(2) \quad \widetilde{H}_\mathrm{W}(3) \quad \widetilde{H}_\mathrm{W}(4)\right]^T,\end{aligned} \tag{12}$$

where the 4th-order correlation matrix $\mathbf{\Phi}_{\widetilde{\mathbf{y}}} = E\left[\widetilde{\mathbf{y}}\widetilde{\mathbf{y}}^H\right]$ is assumed to be known or can be estimated from the noisy observations. We refer to $\widetilde{\mathbf{h}}_{\mathrm{W}}$ as the quadratic Wiener filter. Therefore, in a similar manner to (10), we obtain an unbiased estimator for a based on $\widetilde{\mathbf{h}}_{\mathrm{W}}$:

$$\hat{a}\left[\widetilde{\mathbf{h}}_{\mathrm{W}}\right] = \sqrt{\max\left\{\frac{\widetilde{\mathbf{h}}_{\mathrm{W}}^T\widetilde{\mathbf{y}} - \sigma^2\left[\widetilde{H}_{\mathrm{W}}(1) + \widetilde{H}_{\mathrm{W}}(4)\right]}{\widetilde{\mathbf{h}}_{\mathrm{W}}^T\mathbf{1}_4}, 0\right\}}, \tag{13}$$

where $\mathbf{1}_4$ is an "all-ones" vector of length 4.

While their structures exhibit some level of similarity, some key differences between $\mathbf{h}_{\mathrm{W}}$ and $\widetilde{\mathbf{h}}_{\mathrm{W}}$ should be addressed. For example, $\mathbf{h}_{\mathrm{W}}$ is, in general, designed to estimate a complex-valued variable, whereas $\widetilde{\mathbf{h}}_{\mathrm{W}}$ is designed to estimate a real and positive variable. In addition, $\mathbf{h}_{\mathrm{W}}$ only requires the second order-statistics of the noisy observations, but $\widetilde{\mathbf{h}}_{\mathrm{W}}$ takes advantage of their 4th-order statistics. As a result, the RMSE with $\hat{a}\left[\widetilde{\mathbf{h}}_{\mathrm{W}}\right]$ is expected to be potentially lower than with $\hat{a}\left[\mathbf{h}_{\mathrm{W}}\right]$. Note that the inversion of $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}$ requires more multiplication operations than $\mathbf{\Phi}_{\mathbf{y}}$, but when M is small, the additional complexity is insignificant.

We end this part by pointing out that both versions of the Wiener beamformers require the estimate of a to be known in advance. Since this is the value we wish to estimate, we will employ an iterative procedure in which every iteration consists of two steps: (a) deriving the appropriate beamformer for a given value of a and (b) using that beamformer and its corresponding estimator to generate a new estimate for a. It can be verified that due to the convex nature of the problem, the convergence of the beamformers, and thereby the estimate of a, is guaranteed. Summary of the magnitude estimation algorithm with the quadratic Wiener beamformer, given multiple noisy observations, is elaborated in Algorithm 1. We note that the estimation process with the linear Wiener beamformer is similar, but requires the following modifications: (a) equations (12) and (13) are replaced by (9) and (10), respectively, (b) lines 5 and 6 are omitted as $\mathbf{\Phi}_{\mathbf{y}}$ is computed directly from $\{\mathbf{y}_n\}_{n=1}^N$ in line 7, and (c) the expression $\mathbf{h}_{\mathrm{W}} \leftarrow [1\ 0]^T$ replaces line 9.

Algorithm 1 Magnitude Estimation with the Quadratic Wiener Beamformer

1: **Input:** $\{\mathbf{y}_n\}_{n=1}^N$, ▷ set of N noisy vectors
2: N_0, ▷ number of samples to estimate $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}$
3: I_0, ▷ number of iterations
4: a_0 ▷ initial guess for a
5: **for** n=1:N **do**
6: $\widetilde{\mathbf{y}}_n \leftarrow \mathbf{y}_n^* \otimes \mathbf{y}_n$ ▷ modify the observations
7: $\mathbf{\Phi}_{\widetilde{\mathbf{y}}} \leftarrow \frac{1}{N_0}\Sigma_{n=1}^{N_0}\widetilde{\mathbf{y}}_n\widetilde{\mathbf{y}}_n^H$
8: $\hat{a} \leftarrow a_0$ ▷ initialize $\hat{a}$
9: $\widetilde{\mathbf{h}}_{\mathrm{W}} \leftarrow [1\ 0\ 0\ 0]^T$ ▷ initialize $\widetilde{\mathbf{h}}_{\mathrm{W}}$
10: **for** i=1:I_0 **do**
11: obtain $\widetilde{\mathbf{h}}_{\mathrm{W}}$ using (12) ▷ update $\widetilde{\mathbf{h}}_{\mathrm{W}}$
12: **for** n=1:N **do**
13: obtain $\hat{a}_n$ using (13)
14: $\hat{a} \leftarrow \frac{1}{\#[\hat{a}_n>0]}\Sigma_{[\hat{a}_n>0]}\hat{a}_n$ ▷ update $\hat{a}$
15: **Output:** $\hat{a}$ ▷ desired signal magnitude estimate

B. Correlated Noise

The correlated noise case corresponds, for example, to directional interferences. That is, the same noise signal is received in both microphones but with a frequency-dependent phase difference. Hence, with two uncorrelated real directional interferences V_1 and V_2, (8) reduces to

$$\mathbf{y} = \begin{bmatrix} a \\ a \end{bmatrix} + \begin{bmatrix} V_1 \\ V_1 e^{-\jmath 2\pi f \delta \cos\theta_{\mathrm{i},1}/c} \end{bmatrix} + \begin{bmatrix} V_2 \\ V_2 e^{-\jmath 2\pi f \delta \cos\theta_{\mathrm{i},2}/c} \end{bmatrix}, \tag{14}$$

where $\theta_{\mathrm{i},1}$ and $\theta_{\mathrm{i},2}$ are the respective incident angles of V_1 and V_2. We assume $V_1, V_2 \sim \mathcal{N}\left(0, \sigma^2/2\right)$, and note that the aforementioned phase differences turn the problem from real to complex.

We turn our attention to the well-known beampattern, which exhibits the ULA response to a plane wave impinging from the direction θ. With a linear beamformer $\mathbf{h}$ of length M, the beampattern is defined by

$$\mathcal{B}_\theta\left[\mathbf{h}\right] = \mathbf{h}^H\mathbf{d}_\theta, \tag{15}$$

where the steering vector $\mathbf{d}_\theta$ is defined as in (3). It is well known that a linear beamformer of length $M = 2$ is only capable of placing a single zero in its beampattern (in addition to the distortionless constraint) [3]. Therefore, we cannot completely eliminate the two directional interferences simultaneously. Instead, we will use the linear Wiener beamformer, $\mathbf{h}_{\mathrm{W}}$, from the previous part and derive a corresponding magnitude estimator.

Recalling (7), we may define an analogous power beampattern with a quadratic beamformer $\widetilde{\mathbf{h}}$ of length $M^2 = 4$ by

$$\begin{aligned}\widetilde{\mathcal{B}}_\theta\left[\widetilde{\mathbf{h}}\right] &= \widetilde{\mathbf{h}}^H\widetilde{\mathbf{d}}_\theta \\ &= \begin{bmatrix} \widetilde{H}(3) \\ \widetilde{H}(1)+\widetilde{H}(4) \\ \widetilde{H}(2) \end{bmatrix}^H \begin{bmatrix} e^{\jmath 2\pi f\delta\cos\theta/c} \\ 1 \\ e^{-\jmath 2\pi f\delta\cos\theta/c} \end{bmatrix} \\ &= \widetilde{\mathbf{g}}^H \begin{bmatrix} 1 \\ e^{-\jmath 2\pi f\delta\cos\theta/c} \\ e^{-\jmath 4\pi f\delta\cos\theta/c} \end{bmatrix},\end{aligned} \tag{16}$$

where

$$\widetilde{\mathbf{h}} = \begin{bmatrix}\widetilde{H}(1) & \widetilde{H}(2) & \widetilde{H}(3) & \widetilde{H}(4)\end{bmatrix}^T, \tag{17}$$

$$\widetilde{\mathbf{g}} = e^{-\jmath 2\pi f\delta\cos\theta/c}\begin{bmatrix}\widetilde{H}(3) & \widetilde{H}(1)+\widetilde{H}(4) & \widetilde{H}(2)\end{bmatrix}^T. \tag{18}$$

We observe that the power beampattern of a beamformer $\widetilde{\mathbf{h}}$ of length 2 is mathematically equal to a linear beampattern of an alternative beamformer $\widetilde{\mathbf{g}}$ of length 3 whose elements

are formed by linear combinations of the elements of $\widetilde{\mathbf{h}}$. We deduce that $\widetilde{\mathbf{h}}$ is capable of placing two distinct nulls in its power beampattern.

Next, we will adapt the two versions of the Wiener beamformer and derive appropriate estimators. As $\mathbf{h}_{\mathrm{W}}$ depends merely on $\mathbf{\Phi}_{\mathbf{y}}$, whereas $\widetilde{\mathbf{h}}_{\mathrm{W}}$ depends on both $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}$ and $E\left(\widetilde{\mathbf{y}}\right)$, the linear beamformer remains the same as in (9), but the quadratic beamformer changes to

$$\begin{aligned}\widetilde{\mathbf{h}}_{\mathrm{W}} &= a^2\mathbf{\Phi}_{\widetilde{\mathbf{y}}}^{-1}E\left(\widetilde{\mathbf{y}}\right) \\ &= a^4\mathbf{\Phi}_{\widetilde{\mathbf{y}}}^{-1}\mathbf{1}_4 + \sigma^2\ \beta_\zeta,\end{aligned} \tag{19}$$

where

$$\beta_\zeta = \left[\begin{array}{cccc} 1 & \zeta/2 & \zeta^*/2 & 1 \end{array}\right]^T, \tag{20}$$

$$\zeta = e^{-\jmath 2\pi f\delta\cos\theta_{\mathrm{i},1}/c} + e^{-\jmath 2\pi f\delta\cos\theta_{\mathrm{i},2}/c}. \tag{21}$$

In addition, the magnitude estimators are adapted accordingly. We immediately have

$$\hat{a}\left[\mathbf{h}_{\mathrm{W}}\right] = \sqrt{\max\left\{\frac{|\mathbf{h}_{\mathrm{W}}^H\mathbf{y}|^2 - \sigma^2\left[\Re\{H_{\mathrm{W}}^*(1)H_{\mathrm{W}}(2)\zeta^*\} - \|\mathbf{h}_{\mathrm{W}}\|^2\right]}{|\mathbf{h}_{\mathrm{W}}^H\mathbf{1}_2|^2}, 0\right\}} \tag{22}$$

and

$$\hat{a}\left[\widetilde{\mathbf{h}}_{\mathrm{W}}\right] = \sqrt{\max\left\{\frac{\widetilde{\mathbf{h}}_{\mathrm{W}}^H\left[\widetilde{\mathbf{y}} - \sigma^2\beta_\zeta\right]}{\widetilde{\mathbf{h}}_{\mathrm{W}}^H\mathbf{1}_4}, 0\right\}}, \tag{23}$$

which can be both verified to be real and non-negative. We note that Algorithm 1 applies for the correlated noise case as well by appropriately modifying the expressions for the Wiener beamformers and the estimates of a.

V. Simulations

Let us begin with the uncorrelated noise case. We set $a = 1$, $\phi_a = 0$, and generate $N = 10,000$ independent realizations of V_1 and V_2 drawn from two distinct probability distributions: normal, that is, $V_1, V_2 \sim \mathcal{N}\left(0, \sigma^2\right)$, and exponential, that is, $V_1, V_2 \sim \exp\left(1/\sigma\right)$. We note that the mean value is subtracted from each noise sample of the exponential distribution to form zero-mean samples. We use $N_0 = 500$ realizations to generate estimates of $\mathbf{\Phi}_{\mathbf{y}}$ and $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}$, respectively. Next, we perform the following iterative procedure for each of the beamformers which consists of $I_0 = 5$ iterations (although the simulations clearly indicated that $I_0 = 3$ iterations are enough). First, the beamformer is derived with the latest estimate of a fixed. Then, it is used to generate $10,000$ new estimates for a, out of which the positive estimates are averaged to acquire a single valid estimate. We note that both filters are initialized as identity filters and that the initial magnitude is $a_0 = 5$.

We repeat this experiment for varying values of the broadband input SNR from -20 dB to 20 dB, where it is defined by

$$\begin{aligned}\mathrm{iSNR} &= \frac{\int_f \phi_X(f)df}{\int_f \phi_{V_1}(f)df} \\ &= \frac{a^2}{\sigma^2}\end{aligned} \tag{24}$$

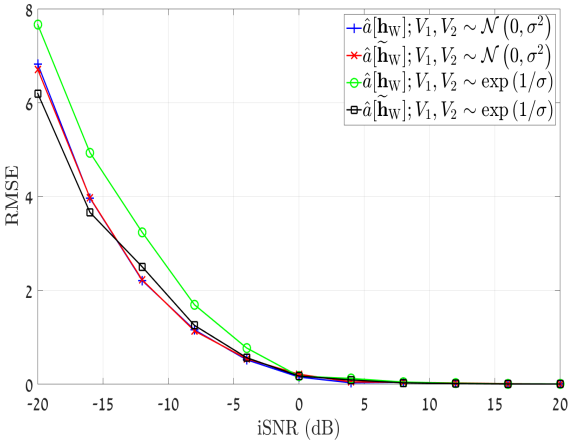


Fig. 1: Magnitude RMSE with the linear and quadratic Wiener beamformers, $\mathbf{h}_{\mathrm{W}}$ and $\widetilde{\mathbf{h}}_{\mathrm{W}}$, with two types of uncorrelated additive noise: normally- and (zero-mean) exponentially-distributed.

and employ the aforementioned RMSE defined by

$$\mathrm{RMSE}\left[\mathbf{h}_{\mathrm{W}}\right] = \sqrt{E\left|a - \hat{a}\left[\mathbf{h}_{\mathrm{W}}\right]\right|^2}, \tag{25}$$

$$\mathrm{RMSE}\left[\widetilde{\mathbf{h}}_{\mathrm{W}}\right] = \sqrt{E\left|a - \hat{a}\left[\widetilde{\mathbf{h}}_{\mathrm{W}}\right]\right|^2}, \tag{26}$$

as the performance measure. The RMSE values as a function for the input SNR are shown in Fig. 1. We observe that for high input SNRs, the RMSE converges to zero, with both estimators and noise distributions. For low input SNRs and normally-distributed noise, both estimators perform the same. This results from the fact that with normally-distributed noise the latent information in higher-order moments is limited. For example, the 3rd-order moment is strictly zero. On the contrary, with the exponentially-distributed noise, the RMSE with the quadratic Wiener beamformer is lower than with the linear Wiener beamformer, with the performance gap reducing as the input SNR increases.

We now turn to the correlated noise case. We maintain the same simulation settings of the uncorrelated noise case and generate samples according to the model in (14). We set $f = 4$ kHz, $\delta = 5$ mm, $\theta_{\mathrm{i},1} = 0^o$, and $\theta_{\mathrm{i},2} = 180^o$. We examine the RMSEs with the two Wiener beamformers and their respective beampatterns (power beampattern with $\widetilde{\mathbf{h}}_{\mathrm{W}}$). The results are depicted in Figs 2 and 3, respectively. We observe that the RMSE with $\widetilde{\mathbf{h}}_{\mathrm{W}}$ is significantly lower than with $\mathbf{h}_{\mathrm{W}}$, with the former achieving a practically zero RMSE for input SNRs higher than 0 dB. As before, for high input SNRs, the RMSE converges to zero with both beamformers. Addressing the beampatterns, we note that while $\mathbf{h}_{\mathrm{W}}$ exhibits a constant "all-pass" pattern, $\widetilde{\mathbf{h}}_{\mathrm{W}}$ exhibits an over-40 dB attenuation in both $\theta_{\mathrm{i},1}$ and $\theta_{\mathrm{i},2}$, whereas θ_{d} remains distortionless. Clearly, such a performance cannot be obtained using a linear beamformer of length $M = 2$.

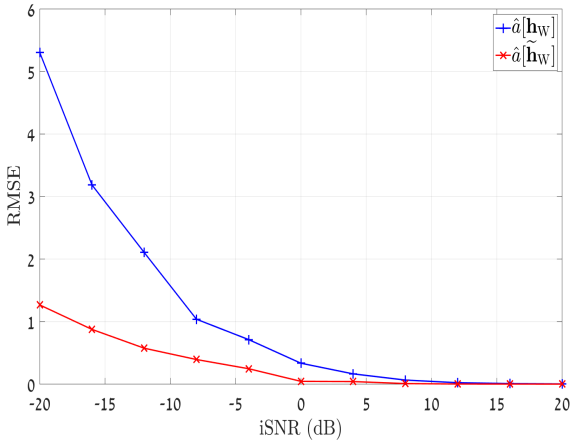
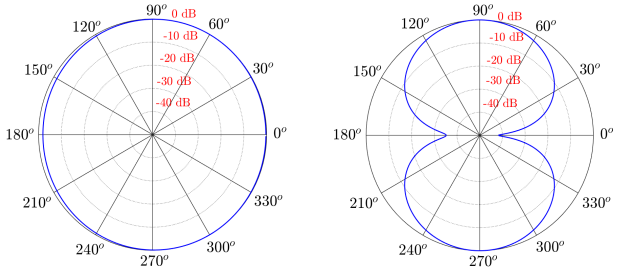


Fig. 2: Magnitude RMSE with the linear and quadratic Wiener beamformers, $\mathbf{h}_{\mathrm{W}}$ and $\widetilde{\mathbf{h}}_{\mathrm{W}}$, with two normally-distributed directional interferences.



(a) (b)

Fig. 3: Beampattern of the linear Wiener beamformer $\mathbf{h}_{\mathrm{W}}$ and a power beampattern of the quadratic Wiener beamformer $\widetilde{\mathbf{h}}_{\mathrm{W}}$, with two normally-distributed directional interferences at $\theta_{\mathrm{i},1} = 0^o$ and $\theta_{\mathrm{i},2} = 180^o$. (a) $\mathbf{h}_{\mathrm{W}}$ and (b) $\widetilde{\mathbf{h}}_{\mathrm{W}}$.

VI. Conclusions

We have introduced an optimal quadratic Wiener beamformer for a desired signal magnitude estimation which utilizes information from higher-order moments. To simplify the formulation, we assumed a two-microphone array, but the generalization to any array is straightforward. We developed an iterative algorithm for magnitude estimation based on a quadratic multichannel noise reduction approach, and addressed two types of additive noise: uncorrelated and correlated. For each noise type, we derived a quadratic version of the Wiener beamformer and a respective unbiased magnitude estimator, and compared their performances to the linear versions of the Wiener beamformer. With uncorrelated noise, we have shown that the quadratic magnitude estimator yields a lower RMSE with respect to the linear estimator in low input SNRs, in case the noise is exponentially distributed. When the noise is drawn from a normal distribution, both estimators perform equally. With correlated noise, we have shown that the quadratic beamformer eliminates two spatial directions, as opposed to a single direction with any linear beamformer. For low input SNRs, this resulted in a significantly lower RMSE using the quadratic beamformer.

References

[1] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Simon and Schuster, Inc., USA, 1992.

[2] H.L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*, Detection, Estimation, and Modulation Theory. Wiley, 2004.

[3] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*, Wiley-IEEE Press, 1st edition, 2018.

[4] J. Benesty, I. Cohen, and J. Chen, *Array Processing - Kronecker Product Beamforming*, Springer-Verlag, Switzerland, 2019.

[5] K. Paliwal and A. Basu, "A speech enhancement method based on kalman filtering," in *ICASSP '87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1987, vol. 12, pp. 177–180.

[6] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403 – 2418, 2001.

[7] I. Cohen and S. Gannot, *Spectral Enhancement Methods*, pp. 873–901, Springer Berlin Heidelberg, 2008.

[8] J. Benesty, J. Chen, and E. Habets, *Speech Enhancement in the STFT Domain*, Springer-Verlag Berlin Heidelberg, Berlin, 2012.

[9] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, Apr 1980.

[10] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec 1984.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr 1985.

[12] P.J. Wolfe and J.S. Godsill, "Efficient Alternatives to the Ephraim and Malah Suppression Rule for Audio Signal Enhancement," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 10, Sep 2003.

[13] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1741–1752, Aug 2007.

[14] Richard C. Hendriks, Heusdens Richard, and Jesper Jensen, "Log-spectral magnitude mmse estimators under super-gaussian densities.," 01 2009, pp. 1319–1322.

[15] Kuldip Paliwal, Kamil Wójcicki, and Belinda Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication*, pp. 450–475, 2010.

[16] Y. Wang and M. Brookes, "Model-based speech enhancement in the modulation domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 580–594, March 2018.

[17] G. Itzhak, J. Benesty, and I. Cohen, "Quadratic approach for single-channel noise reduction," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, 2020.

[18] G. Itzhak, J. Benesty, and I. Cohen, "Nonlinear kronecker product filtering for multichannel noise reduction," *Speech Communication*, vol. 114, pp. 49 – 59, 2019.

[19] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer-Verlag Berlin Heidelberg, 2008.

[20] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1218–1234, July 2006.

[21] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Inc., Boca Raton, FL, USA, 2nd edition, 2013.

Appendix B

Robust Differential Beamforming with Rectangular Arrays

Robust Differential Beamforming with Rectangular Arrays

Gal Itzhak[1], Israel Cohen[1], and Jacob Benesty[2]
[1]*Faculty of Electrical and Computer Engineering, Technion–Israel Institute of Technology*, Haifa 3200003, Israel
[2]*INRS-EMT, University of Quebec*, Montreal, QC H5A 1K6, Canada
galitz@campus.technion.ac.il, icohen@ee.technion.ac.il, benesty@emt.inrs.ca

***Abstract*—In this paper, we introduce a robust approach for rectangular differential beamforming. We present a 2-D multistage spatial mean operator which operates independently on the columns and rows of the observation signals of a uniform rectangular array (URA). The multistage approach enables high flexibility: two design parameters, Q_c and Q_r, set the number of mean stages in the array columns and rows, respectively. Then, we design a rectangular differential beamformer and apply it to the output of the spatial operator. We demonstrate that the first mean operation improves the white noise robustness of the resulting beamformer. We focus on the maximum directivity factor (MDF) and null-constrained maximum directivity factor (NCMDF) differential beamformers and analyze their performances in terms of both the white noise gain (WNG) and directivity factor (DF) measures. We show that the configuration (Q_c, Q_r) constitutes a useful mean to mitigate the white noise amplification of differential beamformers in low frequencies.**

***Index Terms*—Microphone arrays, uniform rectangular arrays (URAs), differential beamforming, robust beamforming, two-dimensional (2-D) arrays.**

I. Introduction

Communication and speech signals are often degraded by undesired noise which may severely deteriorate the functionality of systems that involve such signals. To attenuate the undesired noise, sensor arrays, or beamformers, are often employed. That is, an array consisting of multiple microphones is used to simultaneously capture samples in different locations in space. Among the extensively-studied field of array signal processing, differential microphone arrays (DMAs) are known to be particularly suitable for practical applications for the two following reasons: their physical size is small and their beampatterns tend to be frequency-invariant [1], [2]. As a result, DMAs have been widely studied and optimized over the years [3]–[6].

Recently, there has been a growing interest in differential uniform rectangular arrays (URAs) [7], [8]. Taking advantage of the rectangular geometry, such DMAs exhibit a better beam steering performance than uniform linear arrays (ULAs) and better directivity than uniform circular arrays (UCAs). In addition, URA beamformers may be decomposed into sub-beamformers by employing the Kronecker-product (KP) decomposition. This allows a significant design flexibility: the KP decomposition is not unique and each of the sub-beamformers may be independently designed with respect to a different criterion [7]. Nevertheless, high-directivity rectangular differential beamformers tend to be sensitive to white noise, in particular in low frequencies [9], [10].

This research was supported by the ISF-NSFC joint research program (grant No. 2514/17), and the Pazy Research Foundation.

In this paper, we introduce a robust approach for rectangular differential beamforming. We present a 2-D multistage spatial mean operator which operates independently on the columns and rows of the observation signals of a URA. Then, we design a rectangular differential beamformer and apply it to the output of the spatial operator. We show that the first mean operation improves the white noise robustness of the resulted beamformer in a controlled manner, by appropriately configuring the values of two design parameters, Q_c and Q_r.

Note that this work and the approach presented in [8] are closely related. However, the objective of the multistage differentials in [8] is to improve the array directivity in a controlled manner, which is attained at the expense of an increased sensitivity to white noise. By contrast, in this work, our objective is the opposite as we aim to improve the white noise robustness. We focus on the MDF and NCMDF differential beamformers and analyze their performances in terms of both the WNG and DF measures. We show that the configuration (Q_c, Q_r) constitutes a useful mean to mitigate the white noise amplification of differential beamformers in the low frequency range.

II. Signal Model

Consider a two-dimensional (2-D) microphone URA. Given the Cartesian coordinate system with microphone $(1, 1)$ as its origin, the URA is composed of M_{x} omnidirectional sensors along the x (negative) axis with a uniform interelement spacing equal to δ_{x} and M_{y} omnidirectional sensors along the y (negative) axis with a uniform interelement spacing equal to δ_{y}. We note that δ_{x} and δ_{y} are assumed to be small, to comply with the differential array settings. An illustration of the 2-D URA studied in this paper is depicted in Fig 1.

We assume that a farfield desired source signal (plane wave), on the same plane of the 2-D array, propagates from the azimuth angle, θ, in an anechoic acoustic environment at the speed of sound, i.e., $c = 340$ m/s, and impinges on the above described array. Then, the corresponding steering matrix (of size $M_{\mathsf{x}} \times M_{\mathsf{y}}$) is [1]:

$$\begin{aligned}\mathbf{D}_\theta(\omega) &= \left[\begin{array}{ccc} B_{\theta,1}(\omega)\,\mathbf{a}_\theta(\omega) & \cdots & B_{\theta,M_{\mathsf{y}}}(\omega)\,\mathbf{a}_\theta(\omega) \end{array}\right] \\ &= \mathbf{b}_\theta^T(\omega) \otimes \mathbf{a}_\theta(\omega), \end{aligned} \quad (1)$$

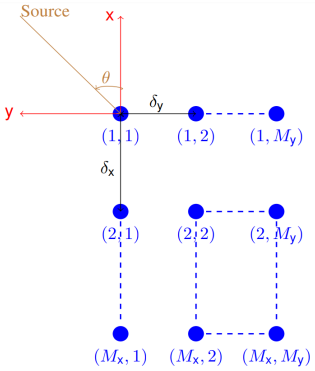


Fig. 1: Illustration of the studied rectangular microphone array.

where

$$\mathbf{a}_\theta(\omega) = \left[\begin{array}{cccc} 1 & e^{-\jmath\varpi_{\theta,\mathsf{x}}(\omega)} & \cdots & e^{-\jmath(M_\mathsf{x}-1)\varpi_{\theta,\mathsf{x}}(\omega)} \end{array}\right]^T \quad (2)$$

is the steering vector associated with the x axis,

$$\begin{aligned}\mathbf{b}_\theta(\omega) &= \left[\begin{array}{cccc} B_{\theta,1}(\omega) & B_{\theta,2}(\omega) & \cdots & B_{\theta,M_\mathsf{y}}(\omega) \end{array}\right]^T \\ &= \left[\begin{array}{cccc} 1 & e^{-\jmath\varpi_{\theta,\mathsf{y}}(\omega)} & \cdots & e^{-\jmath(M_\mathsf{y}-1)\varpi_{\theta,\mathsf{y}}(\omega)} \end{array}\right]^T \quad (3)\end{aligned}$$

is the steering vector associated with the y axis,

$$\varpi_{\theta,\mathsf{x}}(\omega) = \frac{\omega\delta_\mathsf{x}\cos\theta}{c},$$
$$\varpi_{\theta,\mathsf{y}}(\omega) = \frac{\omega\delta_\mathsf{y}\sin\theta}{c},$$

the superscript ^{T} denotes the transpose operator, $\otimes$ is the KP operator, $\jmath = \sqrt{-1}$ is the imaginary unit, $\omega = 2\pi f$ is the angular frequency, and $f > 0$ is the temporal frequency.

Exploiting (1), the observed signal matrix of size $M_\mathsf{x} \times M_\mathsf{y}$ of the URA can be expressed in the frequency domain as [2]:

$$\begin{aligned}\mathbf{Y}(\omega) &= \mathbf{X}(\omega) + \mathbf{V}(\omega) \\ &= \mathbf{D}_\theta(\omega) X(\omega) + \mathbf{V}(\omega), \quad (4)\end{aligned}$$

where $X(\omega)$ is the zero-mean desired source signal and $\mathbf{V}(\omega)$ is the zero-mean additive noise signal matrix.

It is also convenient to express (4) in a vector form. Defining the steering vector $\mathbf{d}_\theta(\omega)$ of length $M_\mathsf{x} \times M_\mathsf{y}$, which is formed by concatenating the columns of $\mathbf{D}_\theta(\omega)$, by:

$$\mathbf{d}_\theta = \mathbf{b}_\theta \otimes \mathbf{a}_\theta, \quad (5)$$

we have

$$\begin{aligned}\mathbf{y}(\omega) &= \left[\begin{array}{cccc} \mathbf{y}_1^T(\omega) & \mathbf{y}_2^T(\omega) & \cdots & \mathbf{y}_{M_\mathsf{y}}^T(\omega) \end{array}\right]^T \\ &= \mathbf{d}_\theta(\omega) X(\omega) + \mathbf{v}(\omega), \quad (6)\end{aligned}$$

where

$$\begin{aligned}\mathbf{y}_{m_\mathsf{y}}(\omega) &= \left[\begin{array}{cccc} Y_{m_\mathsf{y},1}(\omega) & Y_{m_\mathsf{y},2}(\omega) & \cdots & Y_{m_\mathsf{y},M_\mathsf{x}}(\omega) \end{array}\right]^T \\ &= B_{\theta,m_\mathsf{y}}(\omega)\mathbf{a}_\theta(\omega) X(\omega) + \mathbf{v}_{m_\mathsf{y}}(\omega), \quad (7)\end{aligned}$$

for $m_\mathsf{y} = 1, 2, \ldots, M_\mathsf{y}$. Dropping the dependence on ω to simplify the notation, we define the covariance matrix of $\mathbf{y}$ by:

$$\mathbf{\Phi}_\mathbf{y} = E\left(\mathbf{y}\mathbf{y}^H\right) = \phi_X \mathbf{d}_\theta \mathbf{d}_\theta^H + \mathbf{\Phi}_\mathbf{v}, \quad (8)$$

where $E(\cdot)$ denotes mathematical expectation, the superscript ^{H} is the conjugate-transpose operator, $\phi_X = E\left(|X|^2\right)$ is the variance of X, and $\mathbf{\Phi}_\mathbf{v} = E\left(\mathbf{v}\mathbf{v}^H\right)$ is the covariance matrix of $\mathbf{v}$. Assuming that the variance of the noise is approximately the same at all sensors, we can express (8) as:

$$\mathbf{\Phi}_\mathbf{y} = \phi_X \mathbf{d}_\theta \mathbf{d}_\theta^H + \phi_V \mathbf{\Gamma}_\mathbf{v}, \quad (9)$$

where ϕ_V is the variance of the noise at the reference microphone (i.e., the origin of the Cartesian coordinate system) and $\mathbf{\Gamma}_\mathbf{v} = \mathbf{\Phi}_\mathbf{v}/\phi_V$ is the pseudo-coherence matrix of the noise. From (9), we deduce that the input signal-to-noise ratio (SNR) is:

$$\text{iSNR} = \frac{\text{tr}\left(\phi_X \mathbf{d}_\theta \mathbf{d}_\theta^H\right)}{\text{tr}\left(\phi_V \mathbf{\Gamma}_\mathbf{v}\right)} = \frac{\phi_X}{\phi_V}, \quad (10)$$

where $\text{tr}(\cdot)$ denotes the trace of a square matrix.

III. Robust Differential Beamforming

Let us consider the signal model given in (7). We define the first-order forward spatial (unnormalized) mean of $\mathbf{y}_{m_\mathsf{y}}$ ($m_\mathsf{y} = 1, 2, \ldots, M_\mathsf{y}$) as:

$$\Sigma Y_{m_\mathsf{y},i} = Y_{m_\mathsf{y},i+1} + Y_{m_\mathsf{y},i} = Y_{m_\mathsf{y},(1),i}, \ i = 1, 2, \ldots, M_\mathsf{x} - 1, \quad (11)$$

where Σ is the forward spatial mean operator. Clearly, the forward spatial mean operator may be applied multiple times. in general, let $q = 0, 1, \ldots, Q_c$, with $1 \leq Q_c < M_\mathsf{x}$. Let us represent Σ in a vector/matrix form. By definition, we write $\mathbf{\Sigma}_{(0)} = \mathbf{I}_{M_\mathsf{x}}$,where $\mathbf{I}_{M_\mathsf{x}}$ is the $M_\mathsf{x} \times M_\mathsf{x}$ identity matrix. Therefore,

$$\mathbf{\Sigma}_{(0)}\mathbf{y}_{m_\mathsf{y}} = \mathbf{I}_{M_\mathsf{x}}\mathbf{y}_{m_\mathsf{y}} = \mathbf{y}_{m_\mathsf{y}}. \quad (12)$$

We define the qth-order forward spatial (unnormalized) mean of $\mathbf{y}_{m_\mathsf{y}}$ as:

$$\begin{aligned}\Sigma^q Y_{m_\mathsf{y},i} &= \Sigma^{q-1}\left(\Sigma Y_{m_\mathsf{y},i}\right) = \Sigma^{q-1} Y_{m_\mathsf{y},i+1} + \Sigma^{q-1} Y_{m_\mathsf{y},i} \\ &= \sum_{j=0}^{q} \binom{q}{j} Y_{m_\mathsf{y},i+j}, \quad (13)\end{aligned}$$

where $i = 1, 2, \ldots, M_\mathsf{x} - q$ and $\binom{q}{j}$ is the binomial coefficient. In a vector/matrix form, (13) is:

$$\mathbf{\Sigma}_{(Q_c)}\mathbf{y}_{m_\mathsf{y}} = \mathbf{y}_{m_\mathsf{y},(Q_c)}, \quad (14)$$

where

$$\mathbf{\Sigma}_{(Q_c)} = \left[\begin{array}{cccc} \mathbf{c}_{(Q_c)}^T & 0 & \cdots & 0 \\ 0 & \mathbf{c}_{(Q_c)}^T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{c}_{(Q_c)}^T \end{array}\right] \quad (15)$$

is a matrix of size $(M_{\mathsf{x}} - Q_c) \times M_{\mathsf{x}}$, with

$$\mathbf{c}_{(Q_c)} = \Bigg[\begin{pmatrix} Q_c \\ 0 \end{pmatrix} \quad \begin{pmatrix} Q_c \\ 1 \end{pmatrix} \quad \cdots \quad \begin{pmatrix} Q_c \\ Q_c - 1 \end{pmatrix} \quad 1 \Bigg]^T \tag{16}$$

being a vector of length $Q_c + 1$.

Now, substituting (7) into (13), it can be shown that:

$$\begin{aligned} \mathbf{\Sigma}_{(Q_c)} \mathbf{y}_{m_{\mathsf{y}}} &= B_{\theta,m_{\mathsf{y}}} \mathbf{\Sigma}_{(Q_c)} \mathbf{a}_\theta X + \mathbf{\Sigma}_{(Q_c)} \mathbf{v}_{m_{\mathsf{y}}} \\ &= B_{\theta,m_{\mathsf{y}}} \mu_{\theta,\mathsf{x}}^{Q_c} \mathbf{a}_{\theta,(Q_c)} X + \mathbf{v}_{m_{\mathsf{y}},(Q_c)} \\ &= \mathbf{y}_{m_{\mathsf{y}},(Q_c)}, \end{aligned} \tag{17}$$

where

$$\mu_{\theta,\mathsf{x}} = e^{-\jmath \varpi_{\theta,\mathsf{x}}} + 1, \tag{18}$$

$$\mathbf{a}_{\theta,(Q_c)} = \begin{bmatrix} 1 & e^{-\jmath \varpi_{\theta,\mathsf{x}}} & \cdots & e^{-\jmath (M_{\mathsf{x}} - Q_c - 1) \varpi_{\theta,\mathsf{x}}} \end{bmatrix}^T \tag{19}$$

is the steering vector of length $M_{\mathsf{x}} - Q_c$, and $\mathbf{v}_{m_{\mathsf{y}},(Q_c)} = \mathbf{\Sigma}_{(Q_c)} \mathbf{v}_{m_{\mathsf{y}}}$. In an analogous manner, equations (11)-(19) can be rewritten with the roles of x and y axes interchanged. That is, we may average over the rows of $\mathbf{Y}$ instead of over its columns. Define:

$$\mu_{\theta,\mathsf{y}} = e^{-\jmath \varpi_{\theta,\mathsf{y}}} + 1, \tag{20}$$

$$\mathbf{b}_{\theta,(Q_r)} = \begin{bmatrix} 1 & e^{-\jmath \varpi_{\theta,\mathsf{y}}} & \cdots & e^{-\jmath (M_{\mathsf{y}} - Q_r - 1) \varpi_{\theta,\mathsf{y}}} \end{bmatrix}^T, \tag{21}$$

with $1 \leq Q_r < M_{\mathsf{y}}$. Recalling the matrix form in (4), we may define:

$$\begin{aligned} \mathbf{Y}_{(Q_c,Q_r)} &= \mathbf{\Sigma}_{(Q_c)} \mathbf{Y} \mathbf{\Sigma}_{(Q_r)}^T \\ &= \mu_{\theta,\mathsf{x}}^{Q_c} \mu_{\theta,\mathsf{y}}^{Q_r} \left(\mathbf{b}_{\theta,(Q_r)}^T \otimes \mathbf{a}_{\theta,(Q_c)} \right) X \\ &\quad + \mathbf{\Sigma}_{(Q_c)} \mathbf{V} \mathbf{\Sigma}_{(Q_r)}^T. \end{aligned} \tag{22}$$

Applying the (column-wise) vectorization operator, $\mathrm{vec}[\cdot]$, to $\mathbf{Y}_{(Q_c,Q_r)}$, we obtain:

$$\begin{aligned} \mathbf{y}_{(Q_c,Q_r)} &= \mathrm{vec}\left[\mathbf{Y}_{(Q_c,Q_r)} \right] \\ &= \mu_{\theta,\mathsf{x}}^{Q_c} \mu_{\theta,\mathsf{y}}^{Q_r} \mathbf{d}_{\theta,(Q_c,Q_r)} X \\ &\quad + \left(\Sigma_{(Q_r)} \otimes \Sigma_{(Q_c)} \right) \mathbf{v}, \end{aligned} \tag{23}$$

where $\mathbf{d}_{\theta,(Q_c,Q_r)} = \mathbf{b}_{\theta,(Q_r)} \otimes \mathbf{a}_{\theta,(Q_c)}$ is a 2-D differential steering vector of length $(M_{\mathsf{x}} - Q_c)(M_{\mathsf{y}} - Q_r)$. We deduce that the $(M_{\mathsf{x}} - Q_c)(M_{\mathsf{y}} - Q_r) \times (M_{\mathsf{x}} - Q_c)(M_{\mathsf{y}} - Q_r)$ covariance matrix of $\mathbf{y}_{(Q_c,Q_r)}$ is:

$$\begin{aligned} \mathbf{\Phi}_{\mathbf{y}_{(Q_c,Q_r)}} &= \phi_X \left| \mu_{\theta,\mathsf{x}} \right|^{2Q_c} \left| \mu_{\theta,\mathsf{y}} \right|^{2Q_r} \\ &\quad \times \mathbf{d}_{\theta,(Q_c,Q_r)} \mathbf{d}_{\theta,(Q_c,Q_r)}^H \\ &\quad + \phi_V \left(\Sigma_{(Q_r)} \otimes \Sigma_{(Q_c)} \right) \mathbf{\Gamma}_{\mathbf{v}} \left(\Sigma_{(Q_r)} \otimes \Sigma_{(Q_c)} \right)^T. \end{aligned} \tag{24}$$

We immediately obtain the WNG and DF between $\mathbf{y}_{(Q_c,Q_r)}$ and $\mathbf{y}$:

$$\begin{aligned} \mathcal{W}_{(Q_c,Q_r)} &= \frac{\left| \mu_{\theta,\mathsf{x}} \right|^{2Q_c} \left| \mu_{\theta,\mathsf{y}} \right|^{2Q_r} (M_{\mathsf{x}} - Q_c)(M_{\mathsf{y}} - Q_r)}{\mathrm{tr}\left(\left(\Sigma_{(Q_r)} \otimes \Sigma_{(Q_c)} \right) \left(\Sigma_{(Q_r)} \otimes \Sigma_{(Q_c)} \right)^T \right)} \\ &= \frac{\left| \mu_{\theta,\mathsf{x}} \right|^{2Q_c}}{\begin{pmatrix} 2Q_c \\ Q_c \end{pmatrix}} \times \frac{\left| \mu_{\theta,\mathsf{y}} \right|^{2Q_r}}{\begin{pmatrix} 2Q_r \\ Q_r \end{pmatrix}}, \end{aligned} \tag{25}$$

$$\mathcal{D}_{(Q_c,Q_r)} = \frac{\left| \mu_{\theta,\mathsf{x}} \right|^{2Q_c} \left| \mu_{\theta,\mathsf{y}} \right|^{2Q_r} (M_{\mathsf{x}} - Q_c)(M_{\mathsf{y}} - Q_r)}{\mathrm{tr}\left(\left(\Sigma_{(Q_r)} \otimes \Sigma_{(Q_c)} \right) \mathbf{\Gamma}_{\mathrm{d}} \left(\Sigma_{(Q_r)} \otimes \Sigma_{(Q_c)} \right)^T \right)}. \tag{26}$$

where $\mathbf{\Gamma}_{\mathrm{d}}$ is the pseudo-coherence matrix of the spherically-isotropic diffuse noise [8].

Next, we would like to apply a differential beamformer $\mathbf{w}_{(Q_c,Q_r)}$ of length $(M_{\mathsf{x}} - Q_c)(M_{\mathsf{y}} - Q_r)$ to the vector $\mathbf{y}_{(Q_c,Q_r)}$. Then, the beamformer output signal is:

$$Z_{(Q_c,Q_r)} = \mathbf{w}_{(Q_c,Q_r)}^H \mathbf{y}_{(Q_c,Q_r)} = X_{\mathrm{fd},(Q_c,Q_r)} + V_{\mathrm{rn},(Q_c,Q_r)}, \tag{27}$$

where $Z_{(Q_c,Q_r)}$ is the estimate of X,

$$X_{\mathrm{fd},(Q_c,Q_r)} = \mu_{\theta,\mathsf{x}}^{Q_c} \mu_{\theta,\mathsf{y}}^{Q_r} \left(\mathbf{w}^H \mathbf{d}_{\theta,(Q_c,Q_r)} \right) X \tag{28}$$

is the filtered desired signal, and:

$$V_{\mathrm{rn},(Q_c,Q_r)} = \mathbf{w}_{(Q_c,Q_r)}^H \mathbf{v}_{(Q_c,Q_r)} \tag{29}$$

is the residual noise, where $\mathbf{v}_{(Q_c,Q_r)} = \left(\Sigma_{(Q_r)} \otimes \Sigma_{(Q_c)} \right) \mathbf{v}$. Consequently, the variance of $Z_{(Q_c,Q_r)}$ is:

$$\begin{aligned} \phi_{Z_{(Q_c,Q_r)}} &= \mathbf{w}_{(Q_c,Q_r)}^H \mathbf{\Phi}_{\mathbf{y}_{(Q_c,Q_r)}} \mathbf{w}_{(Q_c,Q_r)} \\ &= \phi_{X_{\mathrm{fd},(Q_c,Q_r)}} + \phi_{V_{\mathrm{rn},(Q_c,Q_r)}}, \end{aligned} \tag{30}$$

where

$$\begin{aligned} \phi_{X_{\mathrm{fd},(Q_c,Q_r)}} &= \phi_X \left| \mu_{\theta,\mathsf{x}} \right|^{2Q_c} \left| \mu_{\theta,\mathsf{y}} \right|^{2Q_r} \\ &\quad \times \left| \mathbf{w}_{(Q_c,Q_r)}^H \mathbf{d}_{\theta,(Q_c,Q_r)} \right|^2, \end{aligned} \tag{31}$$

$$\phi_{V_{\mathrm{rn},(Q_c,Q_r)}} = \mathbf{w}_{(Q_c,Q_r)}^H \mathbf{\Phi}_{\mathbf{v}_{(Q_c,Q_r)}} \mathbf{w}_{(Q_c,Q_r)}, \tag{32}$$

and $\mathbf{\Phi}_{\mathbf{v}_{(Q_c,Q_r)}}$ is the correlation matrix of $\mathbf{v}_{(Q_c,Q_r)}$ which is given by:

$$\begin{aligned} \mathbf{\Phi}_{\mathbf{v}_{(Q_c,Q_r)}} &= \phi_V \left(\Sigma_{(Q_r)} \otimes \Sigma_{(Q_c)} \right) \mathbf{\Gamma}_{\mathbf{v}} \left(\Sigma_{(Q_r)} \otimes \Sigma_{(Q_c)} \right)^T \\ &= \phi_V \mathbf{\Gamma}_{\mathbf{v}_{(Q_c,Q_r)}}. \end{aligned} \tag{33}$$

Ultimately, it is clear that the distortionless constraint is given by:

$$\mathbf{w}_{(Q_c,Q_r)}^H \mathbf{d}_{\theta,(Q_c,Q_r)} = \mu_{\theta,\mathsf{x}}^{-Q_c} \mu_{\theta,\mathsf{y}}^{-Q_r}. \tag{34}$$

Now, let us relate the SNR gains corresponding to $\mathbf{w}_{(Q_c,Q_r)}$. It is clear from (30)-(32) that the WNG and DF are given by:

$$\begin{aligned} \mathcal{W}\left(\mathbf{w}_{(Q_c,Q_r)} \right) &= \left| \mu_{\theta,\mathsf{x}} \right|^{2Q_c} \left| \mu_{\theta,\mathsf{y}} \right|^{2Q_r} \\ &\quad \times \frac{\left| \mathbf{w}_{(Q_c,Q_r)}^H \mathbf{d}_{\theta,(Q_c,Q_r)} \right|^2}{\mathbf{w}_{(Q_c,Q_r)}^H \mathbf{\Xi}_{(Q_c,Q_r)} \mathbf{w}_{(Q_c,Q_r)}}, \end{aligned} \tag{35}$$

and the DF:

$$\mathcal{D}\left(\mathbf{w}_{(Q_c,Q_r)}\right) = \left|\mu_{\theta,\mathsf{x}}\right|^{2Q_c}\left|\mu_{\theta,\mathsf{y}}\right|^{2Q_r} \times \frac{\left|\mathbf{w}^H_{(Q_c,Q_r)}\mathbf{d}_{\theta,(Q_c,Q_r)}\right|^2}{\mathbf{w}^H_{(Q_c,Q_r)}\mathbf{\Gamma}_{\mathrm{d},(Q_c,Q_r)}\mathbf{w}_{(Q_c,Q_r)}}, \quad (36)$$

where

$$\mathbf{\Xi}_{(Q_c,Q_r)} = \left(\Sigma_{(Q_r)}\otimes\Sigma_{(Q_c)}\right)\left(\Sigma_{(Q_r)}\otimes\Sigma_{(Q_c)}\right)^T, \quad (37)$$

$$\mathbf{\Gamma}_{\mathrm{d},(Q_c,Q_r)} = \left(\Sigma_{(Q_r)}\otimes\Sigma_{(Q_c)}\right)\mathbf{\Gamma}_{\mathrm{d}}\left(\Sigma_{(Q_r)}\otimes\Sigma_{(Q_c)}\right)^T. \quad (38)$$

IV. Optimal Robust Differential Beamformers

Let us start by considering equation (36). The maximum DF (MDF) beamformer is derived from:

$$\min_{\mathbf{w}_{(Q_c,Q_r)}} \mathbf{w}^H_{(Q_c,Q_r)}\mathbf{\Gamma}_{\mathrm{d},(Q_c,Q_r)}\mathbf{w}_{(Q_c,Q_r)}$$
$$\text{s. t.} \quad \mathbf{w}^H_{(Q_c,Q_r)}\mathbf{d}_{\theta,(Q_c,Q_r)} = \mu^{-Q_c}_{\theta,\mathsf{x}}\mu^{-Q_r}_{\theta,\mathsf{y}}, \quad (39)$$

in which we considered the distortionless constraint. The solution is therefore given by:

$$\mathbf{w}_{\mathrm{MDF}(Q_c,Q_r)} = \frac{1}{\left(\mu^{Q_c}_{\theta,\mathsf{x}}\mu^{Q_r}_{\theta,\mathsf{y}}\right)^*} \times \frac{\mathbf{\Gamma}^{-1}_{\mathrm{d},(Q_c,Q_r)}\mathbf{d}_{\theta,(Q_c,Q_r)}}{\mathbf{d}^H_{\theta,(Q_c,Q_r)}\mathbf{\Gamma}^{-1}_{\mathrm{d},(Q_c,Q_r)}\mathbf{d}_{\theta,(Q_c,Q_r)}}. \quad (40)$$

We now turn to null-constrained version of $\mathbf{w}_{\mathrm{MDF}(Q_c,Q_r)}$. In practice, in order to give a desired shape to a beampattern or attenuate directional interferences, spatial null constraints may be required. For example, with $N = 2$ distinct null constraints (39) is transformed into:

$$\min_{\mathbf{w}} \mathbf{w}^H\mathbf{\Gamma}_{\mathrm{d},(Q_c,Q_r)}\mathbf{w}$$
$$\text{s. t.} \quad \mathbf{C}^H\left(\Sigma_{(Q_r)}\otimes\Sigma_{(Q_c)}\right)^T\mathbf{w}_{(Q_c,Q_r)} = \begin{bmatrix}1\\0\\0\end{bmatrix}. \quad (41)$$

where $\mathbf{C}$ is a constraint matrix of size $M_{\mathsf{x}}M_{\mathsf{y}}\times 3$:

$$\mathbf{C} = \begin{bmatrix}\mathbf{d}_\theta & \mathbf{d}_{\theta_1} & \mathbf{d}_{\theta_2}\end{bmatrix}, \quad (42)$$

whose first column is the steering vector in the direction of the desired signal and the remaining independent columns are the steering vectors in the directions of the desired nulls. The resulting null-constrained maximum MDF (NCMDF) beamformer is given by:

$$\mathbf{w}_{\mathrm{NCMDF}(Q_c,Q_r)} = \mathbf{\Gamma}^{-1}_{\mathrm{d},(Q_c,Q_r)}\left(\Sigma_{(Q_r)}\otimes\Sigma_{(Q_c)}\right)\mathbf{C} \times \left[\mathbf{C}^H\left(\Sigma_{(Q_r)}\otimes\Sigma_{(Q_c)}\right)^T\mathbf{\Gamma}^{-1}_{\mathrm{d},(Q_c,Q_r)} \times \left(\Sigma_{(Q_r)}\otimes\Sigma_{(Q_c)}\right)\mathbf{C}\right]^{-1}\begin{bmatrix}1\\0\\0\end{bmatrix}. \quad (43)$$

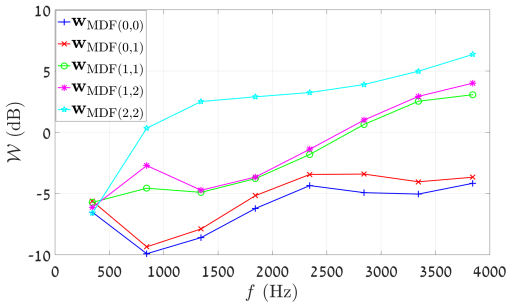
TABLE I: The WNG and DF in dB units between $\mathbf{y}_{(Q_c,Q_r)}$ and $\mathbf{y}$ for varying values of (Q_c, Q_r) and $f = 3$ kHz. Gray background color indicates optimal configurations further discussed in the paper. Simulation parameters: $\theta = 0^o$,$M_{\mathsf{x}} = 5$, $M_{\mathsf{y}} = 5$, $\delta_{\mathsf{x}} = 1$ cm and $\delta_{\mathsf{y}} = 1.5$ cm.

$\mathcal{W}_{(Q_c,Q_r)}$		Q_r					$\mathcal{D}_{(Q_c,Q_r)}$		Q_r				
		0	1	2	3	4			0	1	2	3	4
Q_c	0	0.0	3.0	4.3	5.1	5.6	Q_c	0	0.0	0.2	0.5	0.7	0.9
	1	2.7	5.7	6.9	7.7	8.3		1	-0.2	0.0	0.3	0.5	0.7
	2	3.6	6.6	7.9	8.6	9.2		2	-0.5	-0.2	0.0	0.3	0.5
	3	4.0	7.0	8.3	9.1	9.7		3	-0.7	-0.4	-0.2	0.0	0.3
	4	4.3	7.3	8.5	9.3	9.9		4	-0.9	-0.7	-0.4	-0.2	0.0

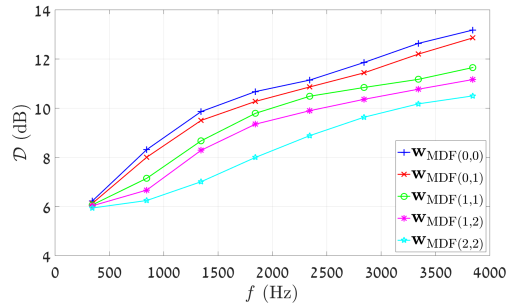
V. Simulations

For the purpose of the simulative part of the paper, let us assume $\theta = 0^o$ as well as the following URA: $M_{\mathsf{x}} = 5$, $M_{\mathsf{y}} = 5$, $\delta_{\mathsf{x}} = 1$ cm and $\delta_{\mathsf{y}} = 1.5$ cm. To begin with, it is valuable to evaluate the WNG and DF between $\mathbf{y}_{(Q_c,Q_r)}$ and $\mathbf{y}$. Recalling equations (25) and (26), we realize that the optimal configurations are scenario-dependent and are a function of the rectangular array structure and the desired signal incident angle. For example, $\mathcal{W}_{(Q_c,Q_r)}$ and $\mathcal{D}_{(Q_c,Q_r)}$ in our scenario are elaborated in Table I for $f = 3$ kHz. Analyzing the results, it is intuitively clear that averaging along the rows would be more beneficial than averaging along the columns. That is, the desired signal is, in fact, in broadside with respect to each ULA along the y axis. Therefore, applying the mean along the rows resembles the application of a series of delay-and-sum beamformers (which are known to be optimal in terms of the WNG) of length 2 to each two adjacent samples. In addition, we note that in case $Q_r \geq Q_c$ averaging over the rows yields either zero or small positive values of $\mathcal{D}_{(Q_c,Q_r)}$, in contrast to the the complementary case. Consequently, we will next focus on the five configurations of (Q_c, Q_r) which are marked in gray in Table I.

Next we investigate the WNG and DF performance of $\mathbf{w}_{\mathrm{MDF}(Q_c,Q_r)}$ and $\mathbf{w}_{\mathrm{NCMDF}(Q_c,Q_r)}$. We note that the latter is designed with two distinct nulls in $\theta_1 = 90^o$ and $\theta_2 = -70^o$. The results are depicted in Fig 2 and Fig 3, respectively. We observe that with both beamformers the WNG is improved upon a (Q_c, Q_r) configuration change, with a great accordance to the values of $\mathcal{W}_{(Q_c,Q_r)}$. In particular, it is important to accentuate the performance gap with $f = 1$ kHz, a relatively low frequency to which the human ear is highly sensitive, but in which high-directivity DMAs tend to exhibit significant white noise amplification. We observe that $\mathbf{w}_{\mathrm{MDF}(2,2)}$ is better than $\mathbf{w}_{\mathrm{MDF}(0,0)}$ by roughly 10 dB, whereas $\mathbf{w}_{\mathrm{NCMDF}(2,2)}$ is better than $\mathbf{w}_{\mathrm{NCMDF}(0,0)}$ by roughly 5 dB. On the contrary, we note that as the WNG improves the DF deteriorates, even though in the selected configurations $\mathcal{D}_{(Q_c,Q_r)}$ is always non-negative. The reason for that is clear- as the values of Q_c or Q_r increase, the length of $\mathbf{y}_{(Q_c,Q_r)}$ decreases. This results in smaller beamformers with less degrees of freedom which are optimized with respect to the array directivity.

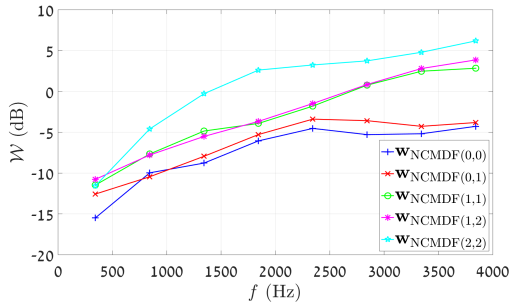


(a)

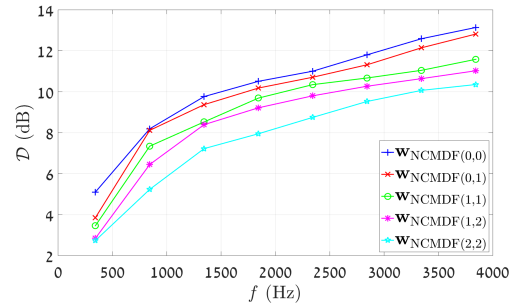


(b)

Fig. 2: WNG and DF measures with the robust MDF differential beamformers, $\mathbf{w}_{\mathrm{MDF}(Q_c,Q_r)}$, with varying values of (Q_c, Q_r). Simulation parameters: $M_{\mathsf{x}} = 5$, $M_{\mathsf{y}} = 5$, $\delta_{\mathsf{x}} = 1$ cm and $\delta_{\mathsf{y}} = 1.5$ cm. (a) WNG and (b) DF.



(a)



(b)

Fig. 3: WNG and DF measures with the robust NCMDF differential beamformers, $\mathbf{w}_{\mathrm{NCMDF}(Q_c,Q_r)}$, with varying values of (Q_c, Q_r). Simulation parameters: $M_{\mathsf{x}} = 5$, $M_{\mathsf{y}} = 5$, $\delta_{\mathsf{x}} = 1$ cm and $\delta_{\mathsf{y}} = 1.5$ cm. (a) WNG and (b) DF.

Nevertheless, even a mild selection of (Q_c, Q_r), for instance, $(1, 1)$, improves the WNG by approximately 5 dB, at the expense of about 1 dB degradation in the DF.

VI. Conclusions

We have introduced a robust approach for rectangular differential beamforming. We proposed to employ a 2-D multistage spatial mean operator which operates independently on the columns and rows of the observation signals of a URA. Through two design parameters, Q_c and Q_r, the multistage approach enables high flexibility. The design parameters correspond to the number of mean stages of the operator in the array columns and rows, respectively. Then, we design a rectangular differential beamformer and apply it to the output of the spatial operator. We showed that the first mean operation improves the robustness of the beamformer to white noise. We focused on the MDF and NCMDF differential beamformers and analyzed their performances in terms of both the WNG and DF measures. We showed that the configuration (Q_c, Q_r) constitutes a useful mean to mitigate the white noise amplification of rectangular differential beamformers in low frequencies, at the expanse of a minor degradation in the array directivity.

References

[1] H. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*, ser. Detection, Estimation, and Modulation Theory. Wiley, 2004.

[2] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*. Singapore: Wiley-IEEE Press, 2018.

[3] E. D. Sena, H. Hacihabiboglu, and Z. Cvetkovic, "On the Design and Implementation of Higher Order Differential Microphones," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 162–174, 2012.

[4] F. Borra, A. Bernardini, F. Antonacci, and A. Sarti, "Efficient implementations of first-order steerable differential microphone arrays with arbitrary planar geometry," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1755–1766, 2020.

[5] G. Itzhak, J. Benesty, and I. Cohen, "On the design of differential kronecker product beamformers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1397–1410, 2021.

[6] J. Benesty, I. Cohen, and J. Chen, *Array Beamforming with Linear Difference Equations*. Springer, 2021.

[7] G. Huang, J. Benesty, J. Chen, and I. Cohen, "Robust and steerable kronecker product differential beamforming with rectangular microphone arrays," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 211–215.

[8] G. Itzhak, J. Benesty, and I. Cohen, "Multistage approach for rectangular differential beamforming," *Submitted to IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

[9] X. Wu and H. Chen, "Directivity Factors of the First-Order Steerable Differential Array With Microphone Mismatches: Deterministic and Worst-Case Analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 300–315, 2016.

[10] J. Jin, G. Huang, X. Wang, J. Chen, J. Benesty, and I. Cohen, "Steering study of linear differential microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 158–170, 2021.

Bibliography

[AFJ10] Y. I. Abramovich, G. J. Frazer, and B. A. Johnson. Iterative Adaptive Kronecker MIMO Radar Beamformer: Description and Convergence Analysis. *IEEE Transactions on Signal Processing*, 58(7):3681–3691, 2010.

[BBAS19] F. Borra, A. Bernardini, F. Antonacci, and A. Sarti. Uniform linear arrays of first-order steerable differential microphones. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):1906–1918, 2019.

[BC11] J. Benesty and J. Chen. *Optimal Time-Domain Noise Reduction Filters; A Theoretical Study*. Springer-Verlag Berlin Heidelberg, 2011.

[BC13] J. Benesty and J. Chen. *Study and Design of Differential Microphone Arrays*. Springer-Verlag Berlin Heidelberg, 2013.

[BCB18] Y. Buchris, I. Cohen, and J. Benesty. Frequency-Domain Design of Asymmetric Circular Differential Microphone Arrays. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(4):760–773, 2018.

[BCB19] Y. Buchris, I. Cohen, and J. Benesty. On the design of time-domain differential microphone arrays. *Applied Acoustics*, 148:212–222, 2019.

[BCC15] J. Benesty, J. Chen, and I. Cohen. *Design of Circular Differential Microphone Arrays*. Springer, Switzerland, 2015.

[BCC18] J. Benesty, I. Cohen, and J. Chen. *Fundamentals of Signal Enhancement and Array Signal Processing*. Wiley-IEEE Press, 2018.

[BCC19] J. Benesty, I. Cohen, and J. Chen. *Array Processing - Kronecker Product Beamforming*. Springer-Verlag, Switzerland, 2019.

[BCH08] J. Benesty, J. Chen, and Y. Huang. *Microphone Array Signal Processing*. Springer-Verlag Berlin Heidelberg, 2008.

[BCH12] J. Benesty, J. Chen, and E. A. P. Habets. *Speech Enhancement in the STFT Domain*. Springer-Verlag Berlin Heidelberg, 2012.

[BCHC09] J. Benesty, J. Chen, Y. Huang, and I. Cohen. *Noise Reduction in Speech Processing*. Springer-Verlag Berlin Heidelberg, 2009.

[BCHD07] J. Benesty, J. Chen, Y. Huang, and J. Dmochowski. On microphone-array beamforming from a MIMO acoustic signal processing perspective. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1053–1065, March 2007.

[BIB14] M.R. Bai, J.G. IH, and J. Benesty. *Acoustic Array Systems: Theory, Implementation, and Application*. Wiley-IEEE Press, 2014.

[Buc02] M. Buck. Aspects of first-order differential microphone arrays in the presence of sensor imperfections. *European Transactions on Telecommunications*, 13:115–122, 2002.

[Cap69] J. Capon. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418, Aug. 1969.

[CB01] I. Cohen and B. Berdugo. Speech enhancement for non-stationary noise environments. *Signal Processing*, 81(11):2403–2418, 2001.

[CBC19] I. Cohen, J. Benesty, and J. Chen. Differential kronecker product beamforming. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(5):892–902, 2019.

[CBP14] J. Chen, J. Benesty, and C. Pan. On the design and implementation of linear differential microphone arrays. *The Journal of the Acoustical Society of America*, 136(6):3097–3113, 2014.

[CHRJ09] Richard C. Hendriks, Heusdens Richard, and Jesper Jensen. Log-spectral magnitude MMSE estimators under super-gaussian densities. pages 1319–1322, Jan. 2009.

[Coh02] I. Cohen. Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. *IEEE Signal Processing Letters*, 9(4):113–116, Apr. 2002.

[Coh05a] I. Cohen. Relaxed statistical model for speech enhancement and a priori SNR estimation. *IEEE Transactions on Speech and Audio Processing*, 13(5):870–881, Sept. 2005.

[Coh05b] I. Cohen. Speech enhancement using super-gaussian speech models and noncausal a priori SNR estimation. *Speech Communication*, 47(3):336 – 350, 2005.

[Coh06] I. Cohen. Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models. *Signal Processing*, 86(4):698–709, 2006.

[EHHJ07] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen. Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(6):1741–1752, Aug. 2007.

[EM84] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6):1109–1121, Dec. 1984.

[EM85] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2):443–445, Apr. 1985.

[GJ82] L. Griffiths and C. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, 30(1):27–34, Jan. 1982.

[HB12] Y. A. Huang and J. Benesty. A multi-frame approach to the frequency-domain single-channel noise reduction problem. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1256–1269, May 2012.

[HBC17] G. Huang, J. Benesty, and J. Chen. Design of robust concentric circular differential microphone arrays. *The Journal of the Acoustical Society of America*, 141(5):3236–3249, 2017.

[HBCC20a] G. Huang, J. Benesty, J. Chen, and I. Cohen. Robust and steerable kronecker product differential beamforming with rectangular microphone arrays. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 211–215, 2020.

[HBCC20b] G. Huang, J. Benesty, I. Cohen, and J. Chen. A simple theory and new method of differential beamforming with uniform linear microphone arrays. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1079–1093, 2020.

[HBLC14] G. Huang, J. Benesty, T. Long, and J. Chen. A family of maximum SNR filters for noise reduction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):2034–2047, Dec. 2014.

[HCB18] G. Huang, J. Chen, and J. Benesty. On the design of differential beamformers with arbitrary planar microphone array geometry. *The Journal of the Acoustical Society of America*, 144(1):66–70, 2018.

[HCB20a] G. Huang, J. Chen, and J. Benesty. Design of planar differential microphone arrays with fractional orders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:116–130, 2020.

[HCB+20b] G. Huang, J. Chen, J. Benesty, I. Cohen, and X. Zhao. Steerable differential beamformers with planar microphone arrays. *EURASIP Journal on Audio, Speech, and Music Processing*, 2020, 2020.

[ICB21] G. Itzhak, I. Cohen, and J. Benesty. Robust differential beamforming with rectangular arrays. In *Proc. 29th European Signal Processing Conference, EUSIPCO-2021*, Aug 2021.

[JD92] D. H. Johnson and D. E. Dudgeon. *Array Signal Processing: Concepts and Techniques*. Simon and Schuster, Inc., USA, 1992.

[Lac71] R. T. Lacoss. Data adaptive spectral analysis methods. *Geophysics*, 36(4):661–675, 1971.

[Loi13] P. C. Loizou. *Speech Enhancement: Theory and Practice*. CRC Press, Inc., Boca Raton, FL, USA, 2nd edition, 2013.

[Mar02] R. Martin. Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors. In *Proceedings of the 27th IEEE International Conference Acoustics Speech Signal Processing, ICASSP-02*, volume 1, pages 253—256, May 2002.

[Mar05] R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *Speech and Audio Processing, IEEE Transactions on*, 13:845–856, Oct. 2005.

[MB03] R. Martin and C. Breithaupt. Speech enhancement in the DFT domain using Laplacian speech priors. In *Proceedings of the 8th International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 87–90, Sept. 2003.

[MM80] R. McAulay and M. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(2):137–145, Apr. 1980.

[Ols46] H. F. Olson. Gradient microphones. *The Journal of the Acoustical Society of America*, 17(3):192–198, 1946.

[PLV+19] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Chang, and T. Sainath. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219, 2019.

[PW19] A. Pandey and D. Wang. A new framework for cnn-based speech enhancement in the time domain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(7):1179–1188, 2019.

[RdAM16] L. N. Ribeiro, A. L. F. de Almeida, and J. C. M. Mota. Tensor beamforming for multilinear translation invariant arrays. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2966–2970, 2016.

[SBA10] M. Souden, J. Benesty, and S. Affes. A study of the LCMV and MVDR noise reduction filters. *IEEE Transactions on Signal Processing*, 58(9):4925–4935, Sept. 2010.

[SHC12] E. De Sena, H. Hacihabiboglu, and Z. Cvetkovic. On the Design and Implementation of Higher Order Differential Microphones. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):162–174, 2012.

[TJCB16] V. Tavakoli, J. Jensen, M. Christensen, and J. Benesty. A framework for speech enhancement with ad hoc microphone arrays. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(6):1038–1051, 2016.

[VT04] H.L. Van Trees. *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. Detection, Estimation, and Modulation Theory. Wiley, New York, 2004.

[WC16] X. Wu and H. Chen. Directivity Factors of the First-Order Steerable Differential Array With Microphone Mismatches: Deterministic and Worst-Case Analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2):300–315, 2016.

[WG03] P.J. Wolfe and J.S. Godsill. Efficient Alternatives to the Ephraim and Malah Suppression Rule for Audio Signal Enhancement. *EURASIP Journal on Advances in Signal Processing*, 2003(10), Sept. 2003.

[WOM33] J. Weinberger, H.F. Olson, and F. Massa. A uni-directional ribbon microphone. *The Journal of the Acoustical Society of America*, 5(2):139–147, 1933.

[XDDL15] Y. Xu, J. Du, L. Dai, and C. Lee. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):7–19, 2015.

[YZW+20] C. Yu, R. E. Zezario, S. Wang, J. Sherman, Y. Hsieh, X. Lu, H. Wang, and Y. Tsao. Speech enhancement based on denoising autoencoder with multi-branched encoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2756–2769, 2020.

[ZWW19] Y. Zhao, Z. Wang, and D. Wang. Two-stage deep learning for noisy-reverberant speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(1):53–62, 2019.

מעצבי אלומה דו-מימדיים מרובי דרגות הנגזרים על פי גישה זו ומדגים כי הביצועים שלהם עדיפים על אלה של מעצבי אלומה חד-מימדיים מרובי דרגות אשר הוצעו בגישות קודמות. עדיפות הביצועים הינה ביחס ליכולת הניחות של רעשי רקע והדהודים מהסביבה והן ביחס למובנות ולאיכות של אותות הדיבור המתקבלים לאחר הסינון במסננים אלה. הדבר בולט במיוחד כאשר כיוון ההגעה של האות הרצוי למערך הינו רחוק מהכיוון המקביל לכיוון הישר עליו מצויים המיקרופונים במערך הלינארי.

כי הגישה הריבועית מציגה ביצועיים טובים יותר מאשר הגישה הלינארית, במיוחד כאשר יחס האות לרעש הוא נמוך או כאשר מספר המיקרופונים במערך הוא קטן.

הנושא השני בו נוגע מחקר זה הינו גישה ריבועית לסינון רעשים חד-ערוצי, בה נעשה שימוש בתכונת ההתאמה של מקדמי אותות דיבור בייצוגם במרחב הזמן-תדר. בעוד הניסוח המתמטי דומה לגישה הריבועית הרב-ערוצית, העקרונות והתכלית של השיטה החד-ערוצית שונים באופן מהותי. למשל, בעוד שבגישה הרב-ערוצית המסנן אופטימלי מתקבל על ידי מינימיזציה של ריבוע ההספק של הדגימות המסוננות, ניתן להסתכל על המסנן הריבועי האופטימאלי בגישה החד-ערוצית כהכללה של המסנן הלינארי האופטימאלי; המסנן הריבועי הממקסם את יחס האות לרעש במוצא נגזר ומנותח כמקרה פרטי. אנו מראים במחקר שמסנן זה משיג ביצועים טובים יותר מהמסנן הלינארי המקביל לפי קריטריון יחס האות לרעש במוצא, ועשוי להגיע (באופן תיאורטי בלבד) ליחס אות לרעש שאיננו חסום, תחת הנחה שסטטיסטיקת הרעש מסדר שני ידועה בצורת מדויקת. בנוסף, אנו מראים כי אותות הדיבור המסוננים בעזרת המסנן הריבועי בעלי מדדים גבוהים יותר של איכות ומובנות, בעיקר כאשר יחס האות לרעש של האות הדגום נמוך.

הנושא השלישי שנידון בתזה זו קשור בתכן של מערכי מיקרופונים הפרשיים (דיפרנציאליים). המחקר מציע גישה גמישה לתכן של מעצבי אלומה הפרשיים מרובי דרגות על ידי שימוש בפירוק בעזרת מכפלת קרונקר של מערך המיקרופונים לשני תתי מערכים הניתנים לתכנון ואופטימיזציה ביחס לקריטריונים שונים ובאופן בלתי תלוי. גישה זו מתאפיינת בשלושה פרמטרים, או דרגות חופש, אשר תוך בחירה שלהם ניתן לקבל מעצבי אלומה השייכים לגישות קודמות בספרות כמקרים פרטיים. כחלק מהמחקר, אנו מראים כי גישת התכן החדשה עשויה להוביל לביצועים טובים יותר מאשר הגישות הקודמות, כתלות בבחירה מתאימה של אותם הפרמטרים.

הנושא הרביעי עוסק בתכן של מעצבי אלומה הפרשיים מנקודת מבט הנותנת דגש על היכולת להסיט את כיוון האלומה. אנו מציגים גישה לתכן מרובה דרגות של מעצב אלומה באמצעות מערך הפרשי דו-מימדי. בצעד הראשון על-פי גישה זו, אנו משתמשים באופרטור הפרשי דו-מימדי הפועל באופן בלתי תלוי על השורות והעמודות של המערך הדו-מימדי. פעולה זו יוצרת מטריצת הפרשים דו-מימדית התלויה בשני פרמטרים הקובעים את מספר דרגות ההפרש בעמודות ובשורות המערך בהתאמה. בשלב השני, אנו מתכננים מעצב אלומה דו-מימדי ומבצעים באמצעותו סינון על מטריצת ההפרשים שהתקבלה בשלב הראשון. המחקר מוכיח כי השלב הראשון משפר באופן משמעותי את כיווניות המערך תוך פגיעה ברגישות שלו אל מול רעש לבן. תכונה זו קשורה באופן הדוק לבחירת הפרמטרים, אשר נבחרים באמצעות אופטימיזציה לפי כיוון ההגעה של האות הרצוי למערך. המחקר מציג ארבעה

תקציר

המחקר המוצג במסגרת תזה זו כולל תכן של מעצבי אלומה ריבועיים ומרובי דרגות שתכונותיהן עשויות להיות חיוניות עבור מגוון רחב של יישומים מעשיים: מערכות תקשורת, זיהוי אוטומטי של אותות דיבור, ממשקי אדם-מכונה ועוד.

בעיית סינון הרעשים וההפרעות מאות דגום רועש הינה בעיה יסודית בתחום של עיבוד אותות אקוסטיים, אשר בכדי להתמודד עימה הוצגו בספרות לאורך השנים מספר רב של שיטות תכן ואלגוריתמים מסוגים שונים. בצורה גסה, ניתן לחלק שיטות אלה לשתי משפחות עיקריות: שיטות סינון רעשים חד-ערוציות, העושות שימוש במיקרופון בודד, ושיטות סינון רעשים רב-ערוציות, העושות שימוש במערך מיקרופונים. שיטות סינון חד-ערוציות בדרך-כלל מצויות בהתקני תקשורת בעלי גודל פיזי קטן ועלות נמוכה, בעוד ששיטות סינון רב-ערוציות בדרך-כלל מצויות במערכות מורכבות יותר ועושות שימוש במספר חיישנים בו זמנית על-מנת לנצל מידע מרחבי לשם שחזור האות הרצוי ולדכא מפריעים כיווניים במרחב.

התזה המוצגת כאן דנה בארבעה נושאי מחקר. הנושא הראשון הינו תכן מעצבי אלומה ריבועיים. באופן המקובל ביותר בתחום, סינון מרחבי מתקבל בדרך-כלל על ידי הכפלת מסנן לינארי בעל מקדמיים קומפלקסיים בוקטור של דגימות רועשות על-מנת לקבל שיערוך של הייצוג (הקומפלקסי) של האות במרחב בו מתרחש העיבוד (למשל, מרחב התדר או מרחב הזמן-תדר). גישות אלה מסתמכות באופן בלעדי על הסטטיסטיקה מסדר שני של הדגימות הרועשות, על אף שבאופן מעשי עשוי להמצא מידע חיוני רב בסדרים גבוהים יותר. במחקר המבוצע בתזה זו, אנו מתמקדים בשיערוך של הספק האות הרצוי אשר ידוע בכעל משמעות רבה יותר ביחס לפאזה שלו ביישומים רבים (למשל, עבור אותות דיבור). כחלק מהמחקר, בעיית השיערוך מנוסחת בצורה שונה התואמת את ההתמקדות בהספק האות הרצוי ומוצעת גישה של סינון ריבועי ממנה ניתן לגזור מעצבי אלומה שונים. בניגוד לגישות קודמות, בניסוח שלנו קיים שימוש ישיר בסטטיסטיקה מסדר גבוה. המחקר מציג ניתוח אנאליטי של בעיה מפושטת וכן נערכות סימולציות מקיפות בנוכחות רעש מסוגים שונים. אנו מראים

המחקר בוצע בהנחייתם של פרופסור ישראל כהן ופרופסור ג'ייקוב בניסטי בפקולטה להנדסת חשמל ומחשבים.

תודות

ברצוני להביע את תודתי הכנה למנחה האחראי שלי, פרופ' ישראל כהן, על ההנחיה הצמודה לכל אורך הדרך. העצות שלו, התמיכה הבלתי-פוסקת ורעיונותיו המועילים תרמו בצורה רבה ומשמעותית למחקר שלי.

בנוסף, ברצוני להודות לפרופ' ג'ייקוב בניסטי, אשר היווה מנחה משני ויועץ, על הרעיונות המקוריים והיצירתיים שלו, ועל ההערות המועילות שלו כמומחה בתחום.

לבסוף, ברצוני להודות למשפחתי על ההבנה ועל העזרה במהלך המסע של לימודי הדוקטורט שלי.

תכן וניתוח של מעצבי אלומה ריבועיים ומרובי דרגות

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר

דוקטור לפילוסופיה

גל יצחק

הוגש לסנט הטכניון — מכון טכנולוגי לישראל

טבת תשפ"ב חיפה דצמבר 2021

תכן וניתוח של מעצבי אלומה ריבועיים ומרובי דרגות

גל יצחק