Acoustic Blind Source Separation in Reverberant and Noisy Environments

Der Technischen Fakultät der
Friedrich-Alexander-Universität Erlangen-Nürnberg
zur Erlangung des Grades

Doktor-Ingenieur

vorgelegt von

Robert Aichner

Erlangen, 2007

Als Dissertation genehmigt von
der Technischen Fakultät der
Friedrich-Alexander-Universität
Erlangen-Nürnberg

Tag der Einreichung:	10. Mai 2007
Tag der Promotion:	24. Juli 2007
Dekan:	Prof. Dr.-Ing. Alfred Leipertz
Berichterstatter:	Prof. Dr.-Ing. Walter Kellermann
	Prof. Dr. Dr. Bastiaan Kleijn

Acknowledgments

I wish to express my sincere gratitude to my advisor Prof. Dr.-Ing. Walter Kellermann from the Friedrich-Alexander University in Erlangen, Germany, for giving me the opportunity to pursue my scientific interests at his research group and for his constant support, mentoring and feedback. I am grateful to Prof. Dr. Dr. Bastiaan Kleijn from the Royal Institute of Technology in Stockholm, Sweden for dedicating a large portion of his busy schedule to the review of this thesis and also for the possibility to spend two months as a visiting researcher at his laboratory. I want to thank Prof. Dr.-Ing. Wolfgang Koch and Prof. Dr.-Ing. Joachim Hornegger, all from the Friedrich-Alexander University in Erlangen for showing so much interest in my work and for participating in the defense of this thesis.

I am very indebted to my former colleague and office mate Herbert Buchner who significantly contributed to the successful completion of this thesis by the uncountable, fruitful, and very inspiring discussions I had with him. I am also grateful to all my other colleagues, who made this laboratory such an interesting and enjoyable place to work. I am thankful to the supportive staff, in particular to Mrs. Ursula Arnold and Mrs. Ute Hespelein for their help to cope with all the administrative tasks and to Mr. Manfred Lindner and Mr. Rüdiger Nägel for constructing the microphone array hardware. My appreciation also goes to all the students who have worked with me. Here, I am particularly grateful to Fei Yan for his commitment during the realization of a real-time blind source separation system.

I am also grateful to Prof. Dr. Shoji Makino from the NTT Communication Science Laboratories, Kyoto, Japan for giving me the opportunity to conduct research at his laboratory for my diploma thesis required by the University of Applied Sciences in Regensburg, Germany. He and his colleagues have sparked my interest in blind source separation and laid the foundations for this dissertation.

I wish to thank the European Union for partially funding this work by grants within the projects “Audio eNhancemnt In secured Telecommunications Applications (ANITA)” (FP5, IST-2001-34327) and “Hearing in the Communication Society (HEARCOM)” (FP6, Project 004171).

Finally, I want to thank my family and all of my friends for their continuous encouragement and for sharing many unforgettable moments with me. Last but not least I want to express my deepest gratitude to my wife Yuri who patiently supported and encouraged me throughout these years even during the early days we had to spend so far apart.

Contents

1 Introduction

In recent years large progress has been made in the field of seamless and hands-free acoustic human-machine interfaces, both, in basic research and in product-orientated development. In this area much effort is dedicated to terminals providing multimedia or telecommunication services which have to be designed to operate in various different scenarios due to the wide range of applications. These include, e.g., audio-/video-conferencing, hands-free telecommunication using car-kits or bluetooth headsets, dictation systems, or public information systems. In such applications the digital signal processing algorithms aim at the estimation of one desired source signal which may be superimposed by several interfering point sources such as competing speakers and possibly also by diffuse background noise such as car noise or speech babble in crowded environments. Additionally, as people want to be untethered and move freely, no close-talking microphones can be used. Thus, in environments with rigid walls also reflections of the desired and interfering signals are picked up which significantly complicate the problem of desired source signal recovery.

Until a few years ago, most acoustic human-machine interfaces offered only one microphone for audio signal acquisition which restricted the approaches to retrieve the desired source signal to single-channel algorithms such as [Bol79, EM84]. Even now, this topic continues to be an important research field as can be seen, e.g., in [BMC05, Sri05]. However, nowadays due to cheaper hardware costs, manufacturers also start to accommodate additional microphones in their products and thus, allow the applicability of multi-channel signal processing algorithms. Examples of products using microphone arrays can be found in several fields such as hands-free communication in cars [Per02], bluetooth headsets for cell phones [VTDM06, Bra07], arrays integrated in multimedia laptops [Mic05], or digital hearing aids [HCE+05].

Moving to more than one sensor allows, in addition to the temporal filtering, also spatial filtering of the acquired signals. This new degree of freedom is exploited by the traditional multi-channel or so-called array signal processing approaches which have originally been developed for narrowband signals as encountered in radar or sonar applications [JD93, Hay02]. Already several decades ago there have been attempts to apply these methods also to broadband signals such as speech. Since then, the area has matured and there are several methods available to enhance a desired acoustic signal corrupted by noise [BW01, Her05]. Typically, these so-called beamforming approaches assume that the positions of the sensors, i.e., the array geometry is known and try to add the desired signal coherently while the interfering signals are added incoherently. Thus, these

algorithms assume a single desired source whose position has to be known a-priori, or has to be estimated by an appropriate source localization algorithm. By applying linear adaptive filtering algorithms based on mean-square error minimization also a tracking of time-variant desired and interfering source positions is possible.

In several applications approaches are desirable where the aim is not only the estimation of a single desired source but the extraction and separation of several acoustic sources. One example are smart meeting rooms which are equipped with several microphones and cameras allowing for audio-/video-conferencing [Moo02]. The possibility to record a meeting allows post-processing such as speaker indexing and meeting transcription using automatic speech recognition, making it easier for people who miss a meeting to access relevant information [CRG+02]. As all participants are "desired signals", all speech signals have to be retrieved and for possibly overlapping speech segments speech separation techniques would be required. Another field for which it is desirable to separate several acoustic sources instead of extracting one desired source are surveillance applications. Additionally, in such scenarios the positions of the desired sources may be unknown so that approaches relying on less a-priori knowledge are desirable. Moreover, the array geometry may not always be known, e.g., if table-top microphones are used inside a meeting room. Another application where only inaccurate information about the sensor positions is available are digital hearing aids aiming at binaural processing of the data [PKRH04, ABZK07].

A possible solution to such problems are blind source separation (BSS) methods which do not require any information about the source and sensor positions. This lack of a-priori knowledge is compensated by exploiting not only the second-order statistical properties of the sensor signals as in linear adaptive filtering algorithms based on mean-square error minimization, but to process the observed signals based on their information content using information theoretic signal processing techniques. The underlying assumption for BSS which allows this point of view is that the source signals are mutually independent. By developing optimization criteria based on statistical descriptors such as entropy or measures determining the similarity of probability distributions, higher-order statistics can be incorporated into the adaptation algorithms. Additionally, also other source signal characteristics such as nonstationarity or nonwhiteness can be exploited. It was pointed out in a recent tutorial article [EP06] that this concept of applying information theoretic criteria to adaptive signal processing also yields improved results in related fields such as feature extraction, clustering, or system identification where up to now mean-square error approaches inherently based on second-order statistics are prevalent.

The concept of BSS can be traced back to the early 80's and since the early 90's it received an increasing interest in the signal processing community [JT00]. Most research was dealing with delayless superposition of the source signals and only since the mid 90's mixing systems accounting for reflections as encountered in acoustics have been considered

[Tor99]. Since then a vast amount of literature has been published on BSS in noiseless acoustic environments.

In contrast to most literature, we will address in this thesis convolutive blind source separation in the context of acoustic signals in both, reverberant *and* noisy environments. The main contribution of this thesis is twofold: First, it will be shown how the information theoretic criterion mutual information can be used to formulate a novel BSS optimization criterion for the first time exploiting all three signal properties nongaussianity, nonwhiteness, and nonstationarity. Based on this criterion a generic BSS framework will be presented. The benefit of the proposed framework is the unifying view on BSS algorithms. This allows to see on which approximations current state-of-the-art algorithms are based on and suggests new research directions to obtain novel algorithms based on less assumptions or more accurate approximations. Based on this point of view, several novel and efficient algorithms are derived and relationships to popular algorithms from the BSS literature are established. The second main contribution of this thesis is the presentation of several pre- and post-processing techniques to ensure noise-robust adaptation of these BSS algorithms for applying them in environments with high background noise. There, it will be shown how these extensions achieve a simultaneous suppression of the background noise in addition to the separation of the point sources.

The work presented in this thesis is structured as follows: In Chapter 2 we introduce the BSS model. After briefly describing the simplest case given by the instantaneous BSS model, we focus on the convolutive BSS model which can accommodate the fact that in acoustic environments also reflections of the original source signals are picked up by the sensors. Subsequently, the relationship of the BSS model to the fundamentals of acoustics is discussed. Then, the source signal properties which may be utilized in BSS approaches are examined and the ambiguities which arise due to the blindness of the BSS methods are addressed.

Based on the convolutive BSS model, a generic framework for BSS in reverberant environments is introduced in Chapter 3. First, the optimum solution for BSS and its consequences are discussed. Based on the distinction between broadband and narrowband optimization the BSS framework is introduced by the formulation of a generic broadband time-domain optimization criterion. A generic gradient-based algorithm is derived and several efficient novel and well-known algorithms are obtained by introducing certain approximations. Additionally, it is shown how broadband algorithms can be derived in the discrete Fourier transform (DFT) domain. These broadband algorithms behave equivalently to their time-domain counterparts and thus, do not exhibit the BSS ambiguities independently in each DFT bin as typical for narrowband algorithms. Moreover, by introducing selective approximations also efficient hybrid and purely narrowband algorithms can be derived. After addressing the different update strategies, experimental results in several reverberant rooms are given for the various BSS algorithms which have been

derived in Chapter 3.

In addition to the interfering point sources, in Chapter 4 also background noise is considered. Due to the diffuse sound field characteristics of several realistic noise types such as car noise or speech babble, the convolutive BSS model describing the superposition of several point sources cannot address the separation of the desired sources from the background noise. Therefore, several extensions to the generic BSS framework are discussed in Chapter 4. First several pre-processing methods are examined which allow a noise-robust adaptation of the BSS algorithms. Subsequently, post-processing approaches are investigated where single-channel postfilters are applied to each BSS output. Due to the background noise, the performance of BSS algorithms decreases so that the postfilter has to address both, the suppression of residual crosstalk from point source interferers and the reduction of background noise. Experimental results for both, pre- and post-processing algorithms are given.

Finally, this thesis is summarized and concluded in Chapter 5. In addition suggestions for future work are presented.

In the Appendices A and B mathematical operators are defined and several derivations are treated in detail. Moreover, in Appendix C all acoustic environments used in the experiments are described.

2 Acoustic Blind Source Separation Model

In blind source separation (BSS) a situation is considered where there are a number of signals emitted by some physical sources. These sources could be, e.g., different brain areas emitting electric signals or several speakers in the same room. Furthermore, it is assumed that there are several sensors which are located at different positions. Therefore, each sensor acquires a slightly different mixture of the original source signals. The goal of blind source separation is to recover the original source signals from this set of sensor signals. The term "blind" stresses the fact that the source signals and the mixing system are assumed to be unknown. The fundamental assumption necessary for applying BSS methods is that the original source signals are mutually statistically independent. In reality this assumption holds for a variety of signals, such as multiple speakers. Therefore, the problem of BSS refers to finding a demixing system whose outputs are statistically independent.

In the following we will first describe the different mixing models which are encountered in various applications. We will especially focus on the convolutive mixing system to model acoustic environments. Subsequently, the relationship between the convolutive model and the fundamentals of room acoustics is discussed. Furthermore, the source signal characteristics with respect to audio signals are examined and their possible utilization for BSS algorithms is explained. In the end of this chapter an overview of objective performance measures, which are useful in assessing BSS algorithms with respect to separation performance and signal quality, is given.

2.1 Instantaneous mixing model

The simplest BSS case deals with an instantaneous mixing model where no delayed versions of the source signals appear. It can be described as a set of observations $x_p(n)$, $p = 1, \ldots, P$, which are generated as a linear mixture of independent components $s_q(n)$, $q = 1, \ldots, Q$, by

$$x_p(n) = \sum_{q=1}^{P} h_{qp} s_q(n) + n_p(n), \tag{2.1}$$

where h_{qp} denote the scalar weights from each source to each sensor, $n_p(n)$ is a possible noise contribution at each sensor and n is the discrete-time index. In this instantaneous BSS case we are interested in finding a corresponding demixing system with the weights w_{pq}, which recover estimates $y_q(n)$, $q = 1, \ldots, Q$, of the original sources $s_q(n)$ from

$$y_q(n) = \sum_{p=1}^{P} w_{pq} x_p(n). \tag{2.2}$$

There are several applications where the instantaneous mixing model is applicable. E.g., in brain science BSS helps to identify underlying components of brain activity from recordings of brain activity as given by an electroencephalogram (EEG) (e.g., [CA02]) or in econometrics where BSS is used to find hidden factors from parallel financial time series (e.g., [HKO01]). Other fields are the extraction of independent features in image processing and improving the image quality, e.g., of astronomic observations (e.g., [Car03]). A comprehensive treatment of the instantaneous BSS case and of the respective algorithms can be found in [HKO01]. In this thesis we deal with BSS for acoustic environments and thus, the instantaneous mixing model is not appropriate as no delayed versions of the source signals are considered. Therefore, we will extend this model in the next section and show its relationship to room acoustics.

2.2 Convolutive mixing model

Extending the instantaneous mixing model by considering also delayed versions of the source signals $s_q(n)$ leads to a mixing system consisting of finite impulse response (FIR) filters instead of scalars. An M-tap mixing system is thus described by

$$x_p(n) = \sum_{q=1}^{Q} \sum_{\kappa=0}^{M-1} h_{qp,\kappa} s_q(n-\kappa) + n_p(n), \tag{2.3}$$

where $h_{qp,\kappa}$, $\kappa = 0, \ldots, M-1$ denote the coefficients of the FIR filter model from the q-th source to the p-th sensor. It should be noted that the source signals are assumed to be point sources so that the signal paths can be modeled as FIR filters. In addition to the source signals a noise signal $n_p(n)$ may be picked up by each sensor. Similar to instantaneous BSS, we are interested in finding a corresponding demixing system whose output signals $y_q(n)$ are described by

$$y_q(n) = \sum_{p=1}^{P} \sum_{\kappa=0}^{L-1} w_{pq,\kappa} x_p(n-\kappa). \tag{2.4}$$

The parameter L denotes the FIR filter length of the demixing filters $w_{pq,\kappa}$. The convolutive mixing model together with the demixing system is depicted as a block diagram

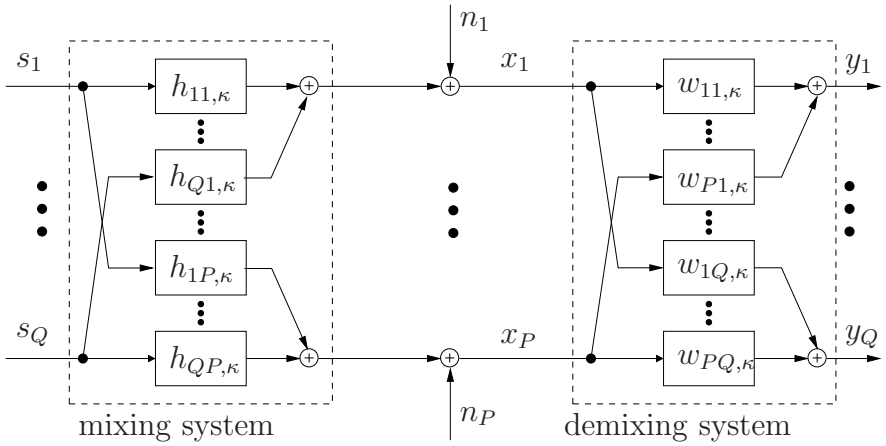


Figure 2.1: Convolutive MIMO model for BSS.

in Fig. 2.1. From this it is obvious that BSS can be classified as a blind multiple-input multiple-output (MIMO) technique.

Most commonly, BSS algorithms are developed under the assumption that the number Q of *simultaneously active* source signals $s_q(n)$ equals the number P of sensor signals $x_p(n)$. However, with the use of suitable techniques, the more general scenario with an arbitrary number of sources and sensors can always be reduced to the standard BSS model. The case that the sensors outnumber the sources is termed *overdetermined* BSS ($P > Q$). The main approach to simplify the separation problem in this case is to apply principle component analysis (PCA) [HKO01], to perform dimension reduction by extracting the first P components and then use standard BSS algorithms. The significantly more difficult case $P < Q$ is called *underdetermined* BSS or BSS with *overcomplete bases*. Mostly the sparseness of the sources in the time-frequency domain is used to determine clusters which correspond to the separated sources (e.g., [ZP01, Bof03, YR04]). Several researchers proposed methods to estimate the sparseness of the sources based on modeling the human auditory system and then subsequently applied time-frequency masking to separate the sources. This research field is termed computational auditory scene analysis (CASA) and a recent overview on the state-of-the-art can be found, e.g., in [Div05, WB06, BW05]. Another approach to the underdetermined case is to exploit the sparseness to eliminate only $Q - P$ sources and then apply again standard BSS algorithms [AMB+03].

Throughout this thesis, we therefore regard the standard BSS model where the number Q of *potentially simultaneously active source signals* $s_q(n)$ is equal to the number of sensor signals $x_p(n)$, i.e., $Q \leq P$. It should be noted that in contrast to other BSS algorithms we do not assume prior knowledge about the exact number of active sources. Thus, even if the algorithms will be derived for the case $Q = P$, the number of simultaneously active sources may change throughout the application of the BSS algorithm and only the

condition $Q \leq P$ has to be fulfilled.

The main focus of this thesis is on BSS for acoustic applications for which the convolutive model is appropriate. The shape of the filters $h_{qp,\kappa}$ of the mixing system determines the type of acoustic environment. This relationship will be discussed in the following sections and the connection to the fundamentals of room acoustics will be made. Moreover, it should be pointed out that convolutive mixtures can also be used, e.g., to model transmission paths in wireless communication scenarios (e.g., [HKO01, CA02]) and more recently they have been applied in brain science to the analysis of magnetoencephalography (MEG) signals [ASM03, DMH06].

2.2.1 Point sources in free-field environments

In principle any complex sound field can be considered as a superposition of numerous simple sound waves. The propagation of such sound waves in any homogeneous, dispersion-free, and lossless medium is governed by a differential equation called the wave equation. Homogeneity assures a constant propagation speed throughout space and time. Dispersion occurs in non-linear media, where the interaction with the medium depends on the amplitude and on the spectral content of the wave. A medium is lossless if the medium does not influence the amplitude attenuation of the propagating wave. Under these assumptions the wave equation is obtained by connecting the various acoustical quantities by a number of basic laws which finally yield [Pie91, Kut00]

$$\nabla^2 p(\vec{r}, t) = \frac{1}{c^2} \frac{\partial^2 p(\vec{r}, t)}{\partial t^2}. \tag{2.5}$$

The sound pressure $p(\vec{r}, t)$ measures the difference between the instantaneous pressure and the static pressure. These gas pressure variations occur under the influence of the sound wave and can be described as a function of the position of observation denoted by the vector [1] $\vec{r}$ and time t. The operator ∇^2 is the Laplacian sum of the second derivatives with respect to the three cartesian coordinates, i.e., the divergence of the gradient. The sound velocity c is given for dry air approximately as [Pie91]

$$c \approx 331 \frac{\mathrm{m}}{\mathrm{s}} + 0.6 T_{\mathrm{C}} \frac{\mathrm{m}}{\mathrm{s}\,^{\circ}\mathrm{C}}, \tag{2.6}$$

where T_{C} denotes the temperature in degrees Celsius and the sound velocity is given in meters per second. In this thesis we will assume a constant sound velocity of $c = 340\,\mathrm{m/s}$.

The mathematically simplest case considers sound propagation in a homogeneous lossless medium which is at rest and unbounded in all directions, i.e., the effects of any obstacles such as walls are neglected. Such a scenario is also termed free-field environment.

[1] Vectors describing directional quantities defined in a coordinate system using a certain metric are denoted as "physical" vectors by an arrow. A vector describing the concatenation of N elements to a general N-tuple is denoted in bold lower case.

In the following, two solutions of the wave equation for a monochromatic wave, i.e., one single harmonic with frequency ω are presented for the free-field environment. Due to the linearity of the wave equation, the principle of superposition holds. For the BSS scenario this is important as there are up to P simultaneously active sources. Moreover, more complicated wave fields, such as the propagation of wideband sources can be expressed as the Fourier integrals with respect to the temporal frequency ω [Pie91].

Monochromatic plane wave

One solution of the wave equation is the plane wave which describes a sound field where all acoustical quantities depend only on the time t and on one single direction. The general solution consists of waves moving in positive and negative directions, respectively, with speed c. If we assume that there is just one wave traveling away from the acoustic source then the solution of the wave equation is given for a monochromatic signal as

$$p(\vec{r}, t) = \hat{p} \cos(\omega t - \vec{k}^{\mathrm{T}} \vec{r}), \tag{2.7}$$

where $(\cdot)^{\mathrm{T}}$ denotes the transpose of a matrix or a vector. The plane wave exhibits a constant amplitude $\hat{p}$ and propagates in the direction determined by the wavenumber vector $\vec{k}$. The magnitude of $\vec{k}$ represents the number of cycles in radians per meter of length in the direction of propagation. The wavenumber magnitude $|\vec{k}| = \frac{2\pi}{\lambda_0}$ can thus be interpreted as the spatial frequency variable corresponding to the temporal frequency variable $\omega = \frac{2\pi}{T}$. Here, T determines the cycle duration and λ_0 is the wavelength[2]. The relation between temporal frequency ω and wavenumber magnitude $|\vec{k}|$ is given as

$$|\vec{k}| = \frac{\omega}{c}. \tag{2.8}$$

For a plane wave the points of equal amplitude are lying on planes which are defined by $\vec{k}^{\mathrm{T}} \vec{p} = a$, where a is a constant.

Acoustic sound fields can be described in reality by using a statistical model [Her05]. Therefore, a microphone at position $\vec{r}$ captures a sample function (realization) of the random process[3]. The realization is given for the plane wave sound field and the discrete-time case as

$$x(n) = \hat{p} \cos(\omega n T_{\mathrm{s}} - \vec{k}^{\mathrm{T}} \vec{r}) \tag{2.9}$$

with n being the discrete-time index and T_{s} denoting the temporal sampling period.

[2]It should be noted that the subscript in the definition of the wavelength λ_0 is used to avoid inconsistencies in the notation because λ will later denote the forgetting factor.

[3]All signals considered in this thesis are considered to be realizations of random processes. In some instances where the underlying random processes are required this will explicitly be pointed out.

Monochromatic spherical wave

Another common sound field is a spherically symmetric wave spreading out from a source in an unbounded, fluid or gaseous medium. The source is considered to be centered at the origin and to have perfect spherical symmetry insofar as the excitation of sound is concerned. Moreover, if only the waves moving away from the source are considered, then the solution of the wave equation for a monochromatic signal is given as

$$p(\vec{r},t) = \frac{\hat{p}}{|\vec{r}|}\cos(\omega t - |\vec{k}||\vec{r}|). \tag{2.10}$$

The wave fronts are spheres concentric to the spatial origin and, in contrast to a plane wave, the amplitude of the monochromatic spherical wave decreases hyperbolically with the distance of the observation, i.e., the radius $|\vec{r}|$. Generally the radiation of point sources is modeled by spherical waves if the circumference of the source is small compared to the wavelength and if the positions of observation are close to the source. The area close to the point source is often termed *near field* because the wave front of the propagating wave is perceptively curved with respect to the distance between the positions of the observations. Provided the distance $|\vec{r}|$ from the center is large compared with the wavelength λ, i.e., $|\vec{k}||\vec{r}| \gg 1$, then the wave field of point sources can be approximated by plane waves due to the decreasing curvature of the wave front. This is termed *far field approximation* and is shown in Fig. 2.2. The transition between the near field and far field of a point source depends on the maximum distance $d_{\max}$ between the observation positions. In literature (e.g., [Sko70, Teu05]) the minimum distance where the near field can be approximated by the far field is usually defined by the maximum tolerable phase error of 22.5° which is introduced by the far field approximation. Using this definition the far field can be assumed for

$$|\vec{r}| > \frac{2d_{\max}^2}{\lambda_0}. \tag{2.11}$$

Similarly to the sound field of a plane wave, the discrete-time sample function captured by the microphone at the distance $|\vec{r}|$ from the point source can be written for the monochromatic spherical wave as

$$x(n) = \frac{\hat{p}}{|\vec{r}|}\cos(\omega n T_s - |\vec{k}||\vec{r}|). \tag{2.12}$$

BSS mixing system for free-field environments

In the convolutive BSS model depicted in Fig. 2.1 the mixing system is described by FIR filters $h_{qp,\kappa}$ from each source s_q, $q = 1,\ldots,P$ to each sensor x_p, $p = 1,\ldots,P$. As shown above, the sound propagation of a point source in the near field can be described by a spherical wave and in the far field it can be approximated by a plane wave. This results for each source s_q in an FIR filter model $h_{qp,\kappa}$ consisting simply of a delay and attenuation

factor. The delay is determined by the distance $|\vec{r}_{qp}|$ between the q-th source and the p-th microphone. The delay in samples is given as $\Delta k = \frac{|\vec{r}_{qp}| \cdot f_\mathrm{s}}{c}$ where f_s denotes the sampling frequency. If the delay is not an integer value, then fractional delays have to be considered as discussed in [LVKL96]. Moreover, the unit impulse is attenuated by the distance $|\vec{r}_{qp}|$ between q-th source and p-th microphone. In BSS applications not the absolute delay but

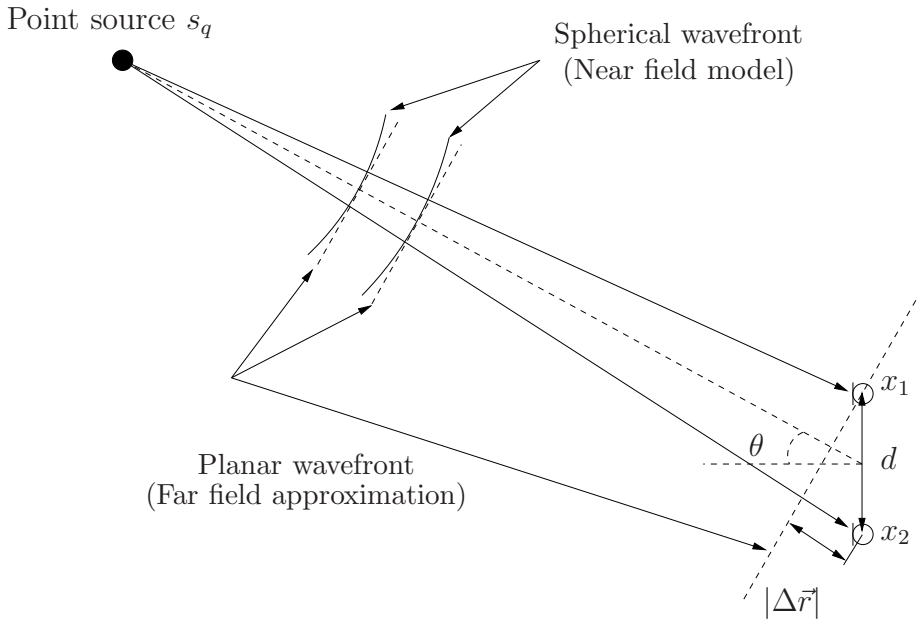


Figure 2.2: Far field and near field model.

the relative delay between different sensors is important. For two microphones x_1 and x_2 in the near field of the point source the relative delay can be determined using the distance difference $|\Delta\vec{r}| = |\vec{r}_{q1}| - |\vec{r}_{q2}|$. For the far field we can approximate the spherical waves by plane waves and thus the distance difference can be calculated by $|\Delta\vec{r}| = d \cdot \sin(\theta)$, where θ denotes the incident angle of the point source and d is the distance between the microphones (see Fig. 2.2).

Early works on BSS for acoustic signals have considered free-field environments. This was also termed BSS for delayed mixtures (see, e.g., [Tor96b, JRY00]).

2.2.2 Point sources in reverberant environments

The free-field model is often not appropriate in realistic environments because reverberation is encountered. Reverberation is caused by the fact that acoustic waves are reflected by room walls and other objects present in the room such that the signals recorded by the microphone array consist of a direct signal path and multiple delayed and attenuated versions. In general, due to the superposition principle even reverberant sound fields can be described by the wave equation. However, for complicated room geometries this requires considerable effort and may not be practical anymore. Therefore, we will present in the

following several variables for characterizing the reverberation. First, a global characterization of the room by a single parameter termed reverberation time is examined and then the notation of acoustic impulse responses, which can be measured experimentally and can be used to determine the mixing system of the BSS model in reverberant environments, are introduced. To describe the effect of the positioning of the point source and of the microphones on the perceived reverberation, several variables have been introduced in the literature and will be discussed.

Reverberation time

The reverberation is determined by the decay of the sound energy. Ideally, after switching off the sound source, the sound energy E decays exponentially according to [Sab22]

$$E(t) = E_0 \exp\left(\frac{A\ c\ \ln(1-\bar{\alpha}_{\mathrm{ab}})}{4V} t\right), \tag{2.13}$$

where E_0 is the initial energy, A is the area of all walls of the rooms, V is the volume of the room, and c is the velocity of sound. The average $\bar{\alpha}_{ab}$ of the individual absorption coefficients $\alpha_{\mathrm{ab},i}$ for the i-th wall is given as

$$\bar{\alpha}_{ab} = \frac{1}{A} \sum_i A_i \alpha_{\mathrm{ab},i}, \tag{2.14}$$

where A_i is the area of the i-th wall. The characteristic time constant in (2.13) is called reverberation time T_{60} and is a global characterization parameter of a reverberant room. The reverberation time was first defined in [Sab22] as the time needed for the sound pressure level to decay to $-60\,\mathrm{dB}$ of its original value E_0. From this definition together with (2.13) a formula to determine the reverberation time T_{60} in seconds was derived by Eyring in [Eyr30] and is given as

$$T_{60} = -\frac{24\ln(10)\ V}{A\ c\ \ln(1-\bar{\alpha}_{ab})}. \tag{2.15}$$

For small absorption coefficients $\alpha_{\mathrm{ab},i}$ the natural logarithm can be linearized and thus the formula for the reverberation time according to Sabine [Sab22] is obtained

$$T_{60} = \frac{24\ln(10)\ V}{c \sum_i A_i \alpha_{\mathrm{ab},i}}. \tag{2.16}$$

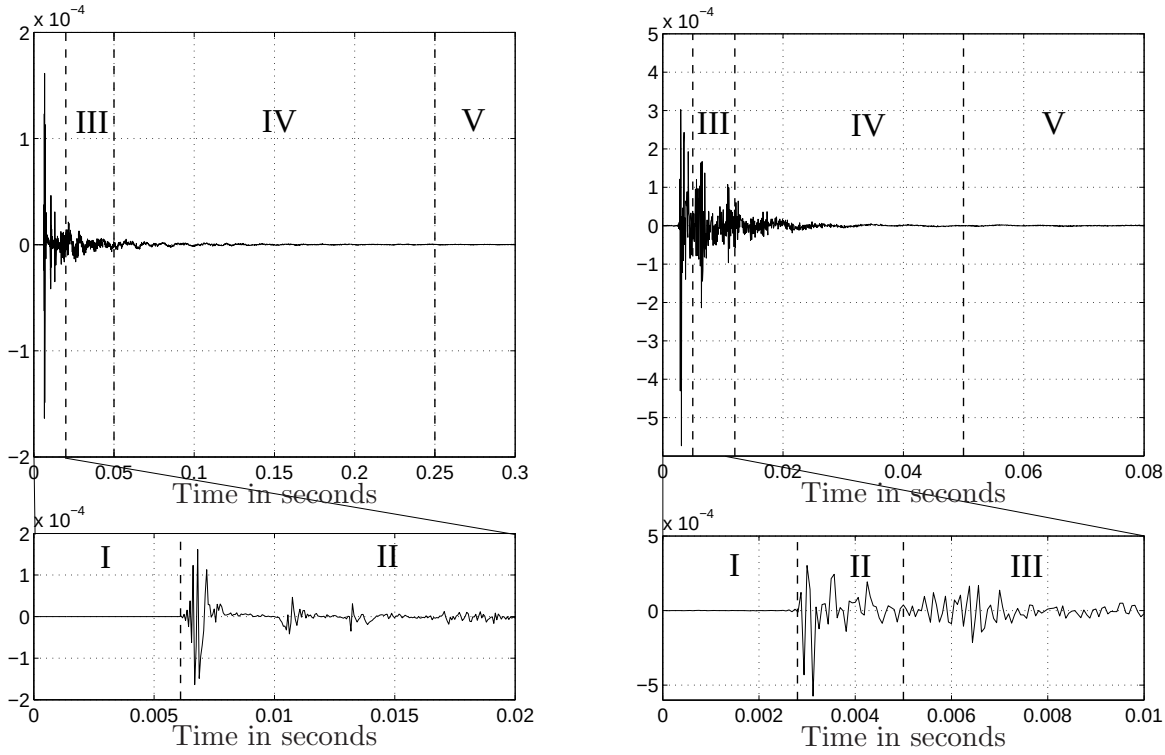
More advanced formulae for calculating the reverberation time have been described in [Neu01]. For office rooms, typical values for the reverberation time are in the order of 200-500 ms, whereas for a car T_{60} is typically smaller than 100 ms and T_{60} for a church or concert hall can be several seconds. It should also be noted that the reverberation time is frequency-dependent due to frequency-dependent absorption coefficients and decreases for higher frequencies. The reverberation time can also be determined experimentally which will be shown later in this section.

Acoustic impulse responses

The reverberant environment can be globally characterized using the reverberation time determined by the exponential decay of the sound energy. However, in reality only late reflections can be considered to decay according to (2.13). Therefore, for a more detailed investigation including also the early reflections, it is of interest to describe the acoustic path between two points (e.g. a point source and a microphone). The acoustic path can be modeled by a linear transfer function due to the linearity of the acoustic wave propagation. This model is in most cases well-justified even if in real-life additional phenomena occur, such as, e.g., diffraction or non-linear absorption [Cro98]. Since the positions of sources are not necessarily fixed, acoustic paths are generally time-varying.

In simulations the acoustic impulse responses are usually modeled using FIR filters. For rectangular enclosures they can be generated by the image method [AB79] which has been extended for microphone arrays in [Pet86]. The image method is an elegant and widely used method. Instead of tracing all reflections, mirror images of the sound source with respect to the room boundaries are created. From each image source, a direct path is traced towards the receiving microphone allowing for accurate modeling of the reflections. In [Bor84] the image model has been extended to arbitrary polyhedra. Nevertheless, for difficult room geometries such as the passenger compartment of a car a generation of the impulse responses by the image method is not feasible. In such cases the impulse responses can be measured by a loudspeaker-microphone system. The loudspeaker is placed at the source position and emits a white pseudo-random signal, e.g., maximum-length sequences [Sch79, RV89] to excite the room with equal power at all frequencies. The pseudo-random signal is picked up by the microphone and subsequently the acoustic impulse response can be calculated by a cross-correlation of the captured signal with the original sequence. In this thesis all experiments have been conducted with impulse responses measured by using this approach.

Two typical impulse responses are shown in Fig. 2.3 which were measured (a) in a reverberant room with dimensions 5.8 m $\times$ 5.9 m $\times$ 3.1 m where the speaker was located 2 m from the microphone and (b) in a car with the speaker at the driver seat and the microphone mounted at the interior mirror. The sampling frequency was $f_\mathrm{s} = 16$ kHz. Acoustic impulse responses consist typically of five parts [Mar95]. The first is called dead time, i.e., the time needed for the acoustic wave to propagate from the source to the microphone along the shortest direct acoustic path. The second period contains the direct path and the first set of early reflections. This period is characterized by single non-overlapping impulses which are caused by dominant reflections. This can be observed especially for the reverberant room in Fig. 2.3a where the first reflections from the room boundaries are clearly visible. The third phase is termed early reverberation and contains numerous overlapping reflections. The fourth phase is called late reverberation where the reverberation energy decays exponentially according to (2.13). Finally, in the fifth stage



(a) Reverberant room

(b) Car environment

Figure 2.3: Measured impulse responses consisting of five parts. I: dead time, II: direct path and early reflections III: early reverberation, IV: late reverberation, V: measurement noise.

the decay of the reverberation energy is buried under the constant noise level which is present in all stages and is determined by the measurement hardware.

Instead of calculating the reverberation time T_{60} according to (2.15) or (2.16) we can also estimate T_{60} from a measured acoustic impulse response. The energy decay curve can be calculated by integrating or, in the discrete-time case, by summing the squared impulse response over time according to [Sch65]

$$E_{\text{decay}}(n) = \sum_{\kappa=n}^{\infty} h^2(\kappa). \tag{2.17}$$

In Fig. 2.4 the corresponding energy decay curve of the acoustic impulse response measured in the car (Fig 2.3b) is shown. In real measurements it is often difficult to obtain a decay of the energy of 60 dB. Thus, standardized methods recommend to extrapolate the segment between -5 dB and -35 dB of the measured reverberation decay to 60 dB by linear least-squares regression [ISO97]. If a 30 dB decay range cannot be measured then a 20 dB range can be used. In Fig. 2.4 this extrapolation is indicated by the dashed line yielding a reverberation time of $T_{60} \approx 50$ ms.

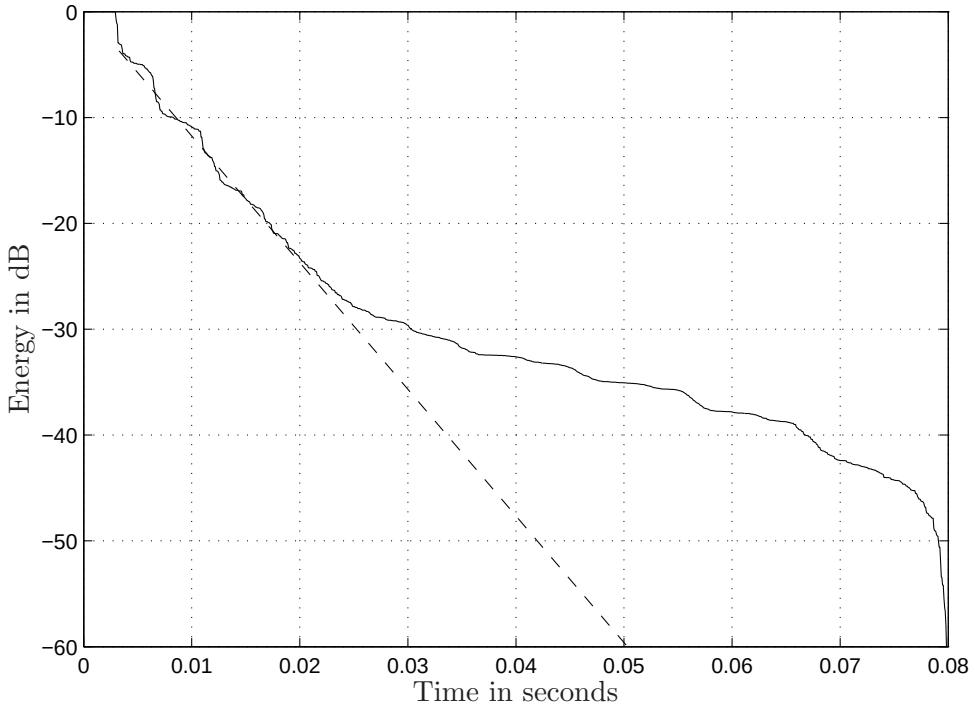


Figure 2.4: Energy decay curve of the acoustic impulse response measured in the car.

Definition D_{50}, clarity index C_{80}, and signal-to-reverberation ratio

The reverberation time T_{60} is considered as the most important objective quantity in room acoustics as it can be easily measured and as it does not significantly depend on the observer's position in the room. However, the reverberation time would only give a full description of the listening conditions in a room if the sound decay obeys strictly to the exponential law given in (2.13). However, especially the early reflections vary from one observation point to the other and hence, the exponential law is a crude approximation for early portions of an impulse response. Thus, for a full description of the prevailing listening conditions the reverberation time has to be supplemented by additional parameters [Kut00]. E.g., an important quantity is the critical delay time which separates useful from detrimental reflections, and which is in the range from 50 ms to 100 ms. This range is spanned by the different types of source signals as, e.g., if the sound signal is music instead of speech, our hearing is generally much less sensitive to reflections. For these reasons, objective criteria are necessary, which relate the reverberation at a certain position in a room with speech intelligibility or subjective sound perception.

Based on the impulse response $h_{qp,\kappa}$ between a source s_q and a microphone x_p, a measure called "definition" D_{50} was introduced which is defined as

$$D_{50} = \frac{\sum_{\kappa=0}^{n_{50}} h_{qp,\kappa}^2}{\sum_{\kappa=0}^{M-1} h_{qp,\kappa}^2} \cdot 100\%, \tag{2.18}$$

where n_{50} is the discrete-time index corresponding to 50 ms denoting the critical delay time for speech signals. In [Kut00] it was argued that there is a good correlation between D_{50} and the speech intelligibility (e.g., $D_{50} = 50\%$ corresponds to 90 % speech intelligiblity).

For music signals a similar quantity exists, which is widely accepted for a characterization of the transparency of music in concert halls and is termed clarity index C_{80}. There, the critical delay time of 80 ms is chosen higher compared to speech. According to [RAS75] it is defined for a impulse response $h_{qp,\kappa}$ of length M between a source s_q and a microphone x_p as

$$C_{80} = 10 \lg \frac{\sum_{\kappa=0}^{n_{80}} h_{qp,\kappa}^2}{\sum_{\kappa=n_{80}}^{M-1} h_{qp,\kappa}^2} \,\mathrm{dB}, \tag{2.19}$$

where n_{80} is the discrete-time index corresponding to 80 ms.

In addition to the previous two quantities D_{50} and C_{80}, which were defined in the context of room acoustics, another criterion measuring the ratio between direct sound and reverberation is the signal-to-reverberation ratio (SRR). In contrast to D_{50} and C_{80}, the SRR is a signal-dependent quantity and it is usually used in the signal processing literature for the evaluation of dereverberation approaches (see, e.g., [NG05]). It is measured in decibel (dB) and is defined for a signal s_q at a sensor x_p as

$$SRR_{p,s_q} = 10 \lg \frac{\sum_n \left(\sum_{\kappa=0}^{n_{50}} h_{qp,\kappa} s_q(n-\kappa)\right)^2}{\sum_n \left(\sum_{\kappa=n_{50}}^{M-1} h_{qp,\kappa} s_q(n-\kappa)\right)^2}. \tag{2.20}$$

It should be noted that in contrast to (2.20), in some literature the SRR is defined as the ratio of the direct signal path and the delayed signal paths, i.e., the critical delay time n_{50} is replaced by the discrete-time index of the direct signal path. However, this neglects the fact that the first reflections are considered as useful. Hence we will use the definition in (2.20) which accounts for these perceptual effects and we will use the SRR later to characterize the environments and setups used for the evaluation of the influence of the reverberation time and the source-sensor distance on the BSS algorithms in Section 3.6.6.

Critical distance r_h

As pointed out in the previous paragraph, the location of point source and microphone influences the shape of the early reflections and is thus important for the speech intelligibility and for the subjective quality of music signals.

Another quantity which is frequently used in room acoustics is describing the ratio of direct sound and reverberation and is called the critical distance. It is based on the fact that if we consider point sources in reverberant environments, then the sound at the listening position is composed of the direct sound from the source described by a spherical wave and the reverberant sound, which approximately constant in the room. The direct sound pressure level decreases inversely to the distance from the source as

described in Section 2.2.1 and will equal the reverberant sound pressure at a distance r_h. This equilibrium of direct and reverberant sound level is called critical distance or reverberation distance and is calculated as

$$r_\mathrm{h} = \sqrt{\frac{\bar{\alpha}_{ab} A}{16\pi}} \approx 0.1\mathrm{m}\sqrt{\frac{V/\mathrm{m}^3}{\pi T_{60}/\mathrm{s}}}. \tag{2.21}$$

The second term in (2.21) is an approximation which is obtained by expressing $\bar{\alpha}_{ab}$ using (2.14) and then inserting the formula for the reverberation time (2.16). Within the critical distance the direct sound from the point source dominates and outside the critical distance the reverberant sound field generated by the reflections from the walls prevails.

Many sound sources have a certain directivity which can be characterized by their directivity factor γ_s [Kut00] which is defined as the ratio of the maximum sound intensity and the sound intensity averaged over all directions. Then the maximum critical distance is increased and is given as

$$r_\mathrm{h} = \sqrt{\frac{\gamma_\mathrm{s}\bar{\alpha}_{ab} A}{16\pi}} \approx 0.1\mathrm{m}\sqrt{\frac{\gamma_\mathrm{s} V/\mathrm{m}^3}{\pi T_{60}/\mathrm{s}}}. \tag{2.22}$$

In [Mar95] the directivity factor for a human speaker has been estimated as $\gamma_\mathrm{s} \approx 1.44$ leading to an increase of the critical distance by a factor of 1.2 compared to a monopole emitting an isotropic sound field. Moreover, if directional microphones targeting at the point source are used instead of omnidirectional microphones, then the critical distance is further increased by the directivity factor γ_x of the sensor yielding

$$r_\mathrm{h} = \sqrt{\frac{\gamma_\mathrm{s}\gamma_\mathrm{x}\bar{\alpha}_{ab} A}{16\pi}} \approx 0.1\mathrm{m}\sqrt{\frac{\gamma_\mathrm{s}\gamma_\mathrm{x} V/\mathrm{m}^3}{\pi T_{60}/\mathrm{s}}}. \tag{2.23}$$

Using a hypercardiod microphone exhibiting a directivity factor $\gamma_\mathrm{x} = 4$ therefore doubles the critical distance [ZZ98]. Moreover, it should be noted that both, the directivity factor of the source and of the sensor are usually frequency-dependent.

The critical distance of a human speaker captured by an omnidirectional microphone in a reverberant room as described in Appendix C.2 with the dimensions (5.8 m $\times$ 5.9 m $\times$ 3.14 m) and reverberation time $T_{60} = 200$ ms can be determined according to (2.22) as $r_\mathrm{h} = 1.6$ m. For comparison, the critical distance in a car with an estimate of the acoustically relevant volume of $V = 1.3$ m^3 and a measured reverberation time of $T_{60} = 50$ ms leads to a critical distance of $r_\mathrm{h} = 0.35$ m [Mar95]. Thus, it can be seen that in realistic environments the area where the direct sound outweighs the reflected sound is very small and therefore, in many applications requiring audio signal capture, the desired source is located outside the critical distance. This shows that the adaptive BSS algorithms considered in this thesis need to explicitly take into account the reverberation by using the convolutive model (2.3), to be applicable to realistic environments.

2.2.3 Diffuse sound fields

In addition to point sources, in real-world scenarios also background noise may be present which often can be modeled as a diffuse sound field. In an ideally diffuse sound field the intensities of the incident sound are uniformly distributed over all possible directions. The phase relations between the sound waves are neglected as the phases of the sound waves are assumed to be uniformly distributed between 0 and 2π. Moreover, the average energy is the same at each point of the enclosure. The resulting 3-dimensional isotropic sound field is termed diffuse sound field [Kut00]. It can be modeled as the sound field created by statistically independent point sources which are uniformly distributed on a sphere and whose phases $\phi(\varphi, \theta)$ are uniformly distributed between 0 and 2π. The angles φ and θ describe the direction of the point source. If the radius of the sphere is $r \to \infty$, then the propagating waves from each point source picked up be the microphones x_p can be assumed to be plane waves. In Fig. 2.5 the diffuse sound field model is depicted where $d\tilde{A}$

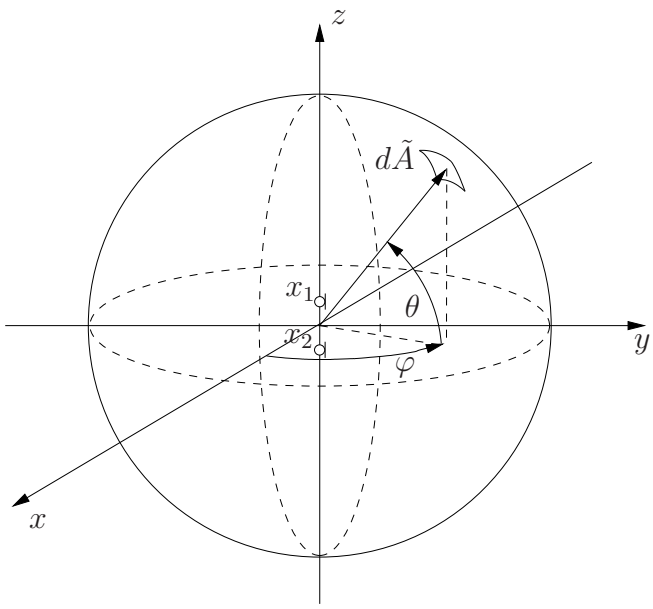


Figure 2.5: Diffuse sound field model.

denotes the area element of one point source. As shown in the next section, such diffuse sound fields can be described very well by the coherence function.

In realistic scenarios, the diffuse sound field is often used to model reverberation [Kut00]. As pointed out above, the assumption which allows this contemplation is that the direct sound and the reflections are assumed to be mutually incoherent, i.e., the phase relations between the sound waves are neglected and thus, a superposition of the sound waves only results in a summation of the sound intensities. However, the convolutive BSS mixing model accounts for the phase relations by the FIR filters describing the acoustic impulse responses. Also the demixing BSS system uses FIR filters of length L, so that

only the reflections exceeding the time-delay covered by L filter taps can be considered as being of diffuse nature. This case applies to highly reverberant environments such as, e.g., lecture rooms, or train stations. Additionally, independent of the reverberation time, diffuse sound fields occur if an infinite number of statistically independent and spatially distributed sound sources are active. This case is more relevant for this thesis as this allows to model background noise such as speech babble noise in a cafeteria, which is generated by a large number of background speakers, by the diffuse sound field. Moreover, it will be shown in Section 2.2.4.3 that exterior noise recorded in the passenger compartment of a car which is a superposition of many different sources such as, e.g., motor, wind, or street noise also exhibit diffuse sound field characteristics.

In the convolutive BSS model depicted in Fig. 2.1 we assumed that the number of simultaneously active point sources Q is equal to the number of sensors P. Due to the limited number of point sources Q in the BSS scenario, we thus cannot model the diffuse sound field by an infinite number of point sources. Therefore, they are included in the BSS model in Fig. 2.1 as noise components n_p, $p = 1, \ldots, P$ which are additively mixed to each microphone signal x_p. For diffuse noise fields, the noise components n_p may be correlated between the sensors as will be shown in the next section. Additionally, each n_p, $p = 1, \ldots, P$ may also contain sensor noise which is usually assumed independent across the different sensors.

2.2.4 Characterizing sound fields by the magnitude squared coherence function

The previous sections described different sound fields that are encountered in applications of BSS algorithms to acoustic environments. An adequate quantity to classify the sound field at the sensors is the magnitude-squared coherence (MSC) function $|\Gamma_{x_1x_2}(\omega)|^2$ [BP66, Car87]. With respect to Chapter 4, where BSS algorithms capable of dealing with different background noise fields will be investigated, we will in the following discuss the MSC and its estimation. The estimation procedures for the MSC were thoroughly examined in [Mar95] and we will briefly summarize the results presented in [Mar95] showing the effect of several parameters on the MSC estimation.

For two stationary discrete-time random processes X_1 and X_2 together with their realisations $x_1(n)$ and $x_2(n)$ the MSC is defined in the frequency domain using the discrete-time Fourier transform (DTFT) as

$$|\Gamma_{x_1x_2}(\omega)|^2 = \frac{|S_{x_1x_2}(\omega)|^2}{S_{x_1x_1}(\omega)S_{x_2x_2}(\omega)}, \tag{2.24}$$

where $S_{x_1x_2}(\omega)$ denotes the cross-power spectral density and $S_{x_1x_1}(\omega)$, $S_{x_2x_2}(\omega)$ are the auto-power spectral densities. Thus, the MSC describes the correlation of the two signals

$x_1(k)$ and $x_2(k)$ in the frequency domain. An important property of the MSC is that it can only attain values between zero and one, i.e.,

$$0 \leq |\Gamma_{x_1x_2}(\omega)|^2 \leq 1. \tag{2.25}$$

For $|\Gamma_{x_1x_2}(\omega)|^2 = 0$ the signals $x_1(n)$ and $x_2(n)$ are not correlated at the frequency ω. On the contrary, if the MSC is close to one, then the two signals are highly correlated. In the case that the two signals are related by a linear convolution or if the two input signals are the same, i.e., $x_1 \equiv x_2$, then the MSC is given as $|\Gamma_{x_1x_2}(\omega)|^2 = 1,\ \forall\, \omega$.

The MSC between two signals is not affected by linearly convolving these signals with arbitrary FIR filters. This can be seen by calculating the MSC $|\Gamma_{y_1y_2}(\omega)|^2$ of the random processes Y_1 and Y_2 (whose realisations y_1, y_2 are obtained by the linear convolution $y_q(n) = \sum_{\kappa=0}^{M-1} h_{q,\kappa}x_q(n-\kappa)$, $q = 1, 2$) which is given as

$$\begin{aligned} |\Gamma_{y_1y_2}(\omega)|^2 &= \frac{|S_{y_1y_2}(\omega)|^2}{S_{y_1y_1}(\omega)S_{y_2y_2}(\omega)} \\ &= \frac{|H_1(\omega)S_{x_1x_2}(\omega)H_2^*(\omega)|^2}{|H_1(\omega)|^2S_{x_1x_1}(\omega)S_{x_2x_2}(\omega)|H_2(\omega)|^2} \\ &= |\Gamma_{x_1x_2}(\omega)|^2, \end{aligned} \tag{2.26}$$

where $H_1(\omega)$ and $H_2(\omega)$ denote the DTFT of the impulse responses $h_1(n)$ and $h_2(n)$, respectively. This result shows that linear filtering does not affect the MSC.

2.2.4.1 Estimating the magnitude squared coherence function

Stationary and ergodic signals

The estimation of the MSC according to (2.24) based on the realisations x_1, x_2 of the random processes X_1, X_2 requires the estimation of the power spectral densities $S_{x_1x_2}(\omega)$, $S_{x_1x_1}(\omega)$, and $S_{x_2x_2}(\omega)$. A popular method to estimate power spectral densities for stationary and ergodic signals is the weighted overlapped-segment averaging technique sometimes referred to as Welch's method [Wel67]. There, the measured signals are decomposed into K overlapping segments of length R which are weighted by a window function $w_f(n)$ and then transformed by an R-point discrete Fourier transform (DFT) into the DFT domain. The weighted DFT of the signal $x_p(n)$ at the frequency bin ν and at the time segment m is given as

$$\underline{X}_p^{(\nu)}(m) = \sum_{n=0}^{R-1} x_p\left(m\frac{R}{\alpha} + n\right) w_f(n) e^{-j\frac{2\pi\nu n}{R}}, \tag{2.27}$$

where α denotes the overlap factor. In this thesis we use the convention that DFT-domain quantities are marked by an underline. Using (2.27) the modified cross-periodogram

between the signals $x_p(n)$ and $x_q(n)$ is obtained by

$$\begin{aligned}\underline{I}^{(\nu)}_{x_p x_q}(m) &= \frac{1}{R \cdot U}\underline{X}^{(\nu)}_p(m)\underline{X}^{(\nu)^*}_q(m) \\ &= \frac{1}{R \cdot U}\left(\sum_{n=0}^{R-1} x_p\left(m\frac{R}{\alpha}+n\right) w_f(n) e^{-j\frac{2\pi\nu n}{R}}\right) \\ &\quad \cdot \left(\sum_{n=0}^{R-1} x_q\left(m\frac{R}{\alpha}+n\right) w_f(n) e^{-j\frac{2\pi\nu n}{R}}\right)^* . \end{aligned} \tag{2.28}$$

The factor $R \cdot U$ normalizes the periodogram by the block length R and the energy of the window function $w_f(n)$ given as

$$U = \frac{1}{R}\sum_{n=0}^{R-1} w_f^2(n). \tag{2.29}$$

Outside the interval $0 \leq n \leq R-1$ the window function is equal to zero.

According to [Wel67] the estimate of the cross-power spectral density for the ν-th frequency bin is obtained by an average of the modified cross-periodograms over K segments

$$\underline{S}^{(\nu)}_{x_p x_q} = \frac{1}{K}\sum_{m=0}^{K-1} \underline{I}^{(\nu)}_{x_p x_q}(m). \tag{2.30}$$

It is shown in [Wel67, OSB98] that for the overlap factor $\alpha = 1$, i.e., no overlap, the variance of $\underline{S}^{(\nu)}_{x_p x_q}$ is inversely proportional to the number of periodograms averaged, and as K increases, the variance approaches zero. Moreover, if R increases then also the bias will approach zero and thus, the periodogram averaging provides an asymptotically unbiased, consistent estimate of $S_{x_1 x_2}(\omega)$. However, as is typical in statistical estimation problems, for a fixed data length there is a trade-off between bias and variance. To reduce the variance for a fixed data length, Welch considered overlapping segments and showed that if the overlap is one-half the window length, i.e., $\alpha = 2$, the variance is further reduced by almost a factor of two due to the doubling of the number of sections. Greater overlap does not continue to reduce the variance, because the segments become less and less independent as the overlap increases. Therefore, in practice it is usual to apply a Hamming or Hann window together with an overlap of successive blocks by 50%, i.e., the overlap factor $\alpha = 2$.

Based on the power-spectral density estimates we obtain an estimate of the magnitude-squared coherence function for the ν-th frequency bin

$$|\underline{\Gamma}^{(\nu)}_{x_1 x_2}|^2 = \frac{|\underline{S}^{(\nu)}_{x_1 x_2}|^2}{\underline{S}^{(\nu)}_{x_1 x_1}\underline{S}^{(\nu)}_{x_2 x_2}}. \tag{2.31}$$

In [CKN73, Car87] the bias and variance of this estimate has been investigated. Their empirical studies showed that the smallest bias and variance is obtained using an overlap

of 62.5%. However, with regard to the computational complexity they recommended an overlap of 50%.

In Section 2.2.4 it was shown in (2.26) that the MSC $|\Gamma_{x_1x_2}(\omega)|^2$ is not affected by an arbitrary linear processing of the signals. In [Mar95, Bit01] the effect of linear transformations $y_q(n) = \sum_{\kappa=0}^{M-1} h_{q,\kappa} x_q(n-\kappa)$ on the *estimated* MSC $|\underline{\Gamma}^{(\nu)}_{x_1x_2}|^2$ has been investigated for stationary and ergodic signals. Martin showed in [Mar95] that due to the windowing function and the finite length of the DFT the estimated MSC $|\underline{\Gamma}^{(\nu)}_{y_1y_2}|^2$ is not independent of the filter transfer functions and thus, only yields an estimate of the MSC $|\underline{\Gamma}^{(\nu)}_{x_1x_2}|^2$ of the original signals x_1 and x_2 which is biased towards zero. This bias due to the impulse responses $h_{1,\kappa}$ and $h_{2,\kappa}$ of length M is mitigated by using, e.g., a Hann window for the window function $w_f(n)$. Moreover, it is desirable that the ratio R/M is large, i.e., an observation interval R larger than the length M of the impulse response is chosen. This aspect is important when estimating the MSC of point sources in an acoustic environment as will be discussed in Section 2.2.4.2.

Nonstationary signals

The MSC estimate according to (2.31) is only defined for stationary signals. In acoustic signal processing, however, we are usually dealing with short-term stationary signals such as speech. For speech signals the period where stationarity can be assumed is only 5 ms to 20 ms [RS78]. Thus, the estimation of the power-spectral densities should ideally be performed in these short-term stationary periods. Usually a weighting function with an exponential forgetting factor γ is used for the estimation leading to [4]

$$\underline{S}^{(\nu)}_{x_px_q}(m) = (1-\gamma)\sum_{i=0}^{m} \gamma^{m-i}\underline{X}^{(\nu)}_p(i)\underline{X}^{(\nu)^*}_q(i). \tag{2.32}$$

This average can be expressed equivalently by a first-order recursive system yielding an estimate for the cross-power spectral density of the m-th block and the ν-th frequency bin

$$\underline{S}^{(\nu)}_{x_px_q}(m) = \gamma\underline{S}^{(\nu)}_{x_px_q}(m-1) + (1-\gamma)\underline{X}^{(\nu)}_p(m)\underline{X}^{(\nu)^*}_q(m). \tag{2.33}$$

The forgetting factor γ determines both, the temporal resolution and similarly to the parameter K in (2.30) also the variance of the estimate. The value of γ has to lie in the range $0 \leq \gamma < 1$ and is usually chosen as a value close to one. A rule of thumb is that a rectangular window with the length $R = (1+\gamma)/(1-\gamma)$ yields approximately the same estimate as the exponential window with forgetting factor γ. This has been proven for the estimation of short-term stationary auto-regressive Gaussian processes in [Bor85]. Thus,

[4] Due to the averaging over several short-term stationary blocks of data, the estimate will still be influenced by the nonstationarity of the signal. However, this is usually tolerated as the averaging reduces the bias of the estimate.

the MSC estimate for the m-th block and ν-th frequency bin is given as

$$|\underline{\Gamma}_{x_1x_2}^{(\nu)}(m)|^2 = \frac{|\underline{S}_{x_1x_2}^{(\nu)}(m)|^2}{\underline{S}_{x_1x_1}^{(\nu)}(m)\underline{S}_{x_2x_2}^{(\nu)}(m)}. \tag{2.34}$$

After averaging over K individual blocks we obtain the long-term estimate of the MSC.

$$|\bar{\underline{\Gamma}}_{x_1x_2}^{(\nu)}|^2 = \frac{1}{K}\sum_{i=0}^{K-1}|\underline{\Gamma}_{x_1x_2}^{(\nu)}(i)|^2. \tag{2.35}$$

2.2.4.2 Magnitude squared coherence of point sources

The magnitude squared coherence defined in (2.24) is given for a point source $s_q(n)$ as $|\Gamma_{s_qs_q}(\omega)| = 1$. In (2.26) it was shown that when using the DTFT the MSC is not affected by arbitrary filtering. However, it was pointed out in the previous section that a bias towards zero is introduced by the impulse response $h_{pq,\kappa}$ of length M between the point source $s_q(n)$ and the sensor $x_p(n)$ when estimating the MSC within finite time intervals by using the DFT of length R. For the choice of the observation interval R one usually has the two contradictory requirements that R should be less than or equal to the stationarity interval but, to avoid a bias of the MSC towards zero, it should also be larger than the FIR filter length M of the acoustic impulse responses. In realistic reverberant environments, where M may be very large, the ratio R/M is usually less than 1, i.e., a bias of the MSC cannot be avoided. Therefore, we will discuss in this section the MSC estimate of the point source at the sensors $x_1(n)$, $x_2(n)$ and the resulting bias for point sources in free-field and reverberant environments.

Point sources in free-field environments can be described by a spherical wave in the near field and by plane wave propagation in the far field as pointed out in Section 2.2.1. Therefore, the FIR filters $h_{pq,\kappa}$ modeling the propagation from the q-th source to the p-th microphone contain only a delay and attenuation factor. In [Car87] the influence of the relative delay between the two sensor signals $x_1(n)$ and $x_2(n)$ on the MSC estimate was studied for stationary signals. It was pointed out that due to the estimation using Welch's method and a rectangular window function the magnitude of the MSC is degraded in the mean by a constant factor which depends on the ratio of the relative delay τ between the observed signals and the DFT length R

$$\mathrm{E}\{|\underline{\Gamma}_{x_1x_2}^{(\nu)}|^2\} \approx \left(1 - \frac{|\tau|}{R}\right)^2 |\Gamma_{x_1x_2}(\omega)|^2. \tag{2.36}$$

This decrease of the MSC due to the delay τ and the finite DFT length R can also be observed for nonstationary signals such as speech. Fig. 2.6 shows the long-term estimate of the MSC (2.35) for a speech signal of length 20 sec obtained by using the recursive averaging procedure (2.33) with $\gamma = 0.9$, $\alpha = 2$, and choosing the Hann window for

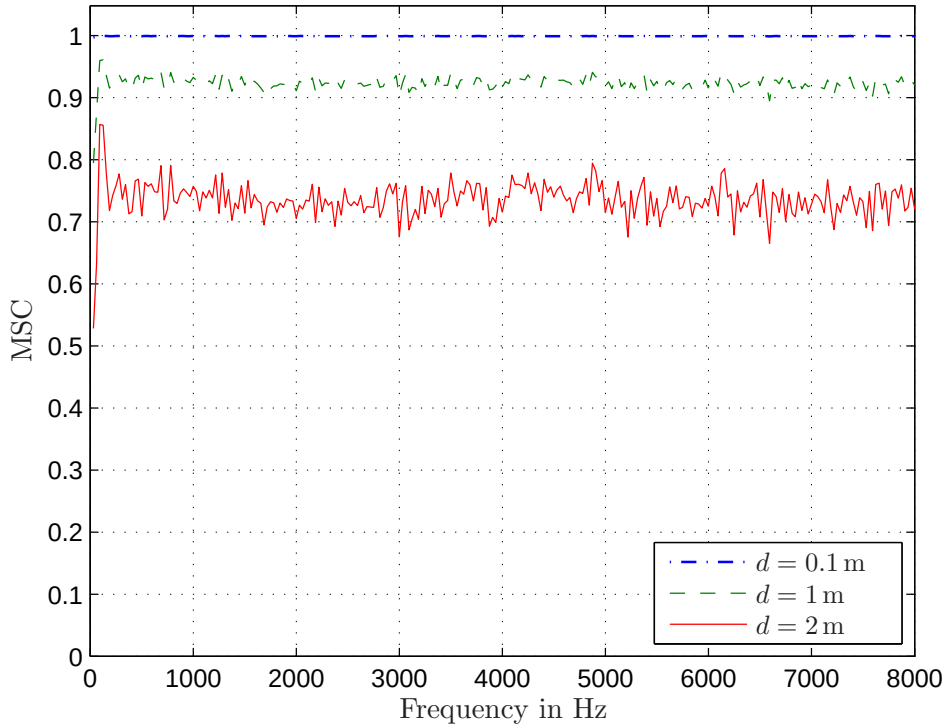


Figure 2.6: Estimate of MSC for a speech source in a free-field environment with different sensor spacing d.

$w_f(n)$. A free-field was assumed with the source located in the far field coming from a direction of $\theta = 70°$ and with the sampling frequency $f_\mathrm{s} = 16\,\mathrm{kHz}$. The sensor spacing d was varied from 0.1 m to 2 m. It can be seen that due to the finite DFT length $R = 512$ the delay between the two sensor signals introduces a bias towards zero for the MSC estimate. According to (2.36) the bias increases for larger sensor spacings as the relative time delay τ between the sensors increases which is confirmed by the experimental results obtained in Fig. 2.6.

For point sources in reverberant environments the bias of the MSC estimate depends on the power ratio between direct sound and reverberation. In [Mar95] it was pointed out that for a DFT length R smaller than the length M of the acoustic impulse responses the bias of the MSC $|\underline{\bar{\Gamma}}^{(\nu)}_{x_1x_2}|^2$ for a source s_q depends on the correlation of the acoustic impulse responses from source s_q to the sensors x_1 and x_2. For large correlation (i.e., strong direct path) the MSC is large and for small correlation (i.e., weak direct path and large reverberation) the MSC becomes small. In Fig. 2.7 the influence of the reverberation time on the bias of the long-term estimate of the MSC is shown. The long-term estimate of the MSC is obtained using (2.35) and (2.33) choosing $R = 512$, $\gamma = 0.9$, $\alpha = 2$, and a sensor spacing of $d = 20\,\mathrm{cm}$. The speech source was located at $\theta = 0°$ and 2 m

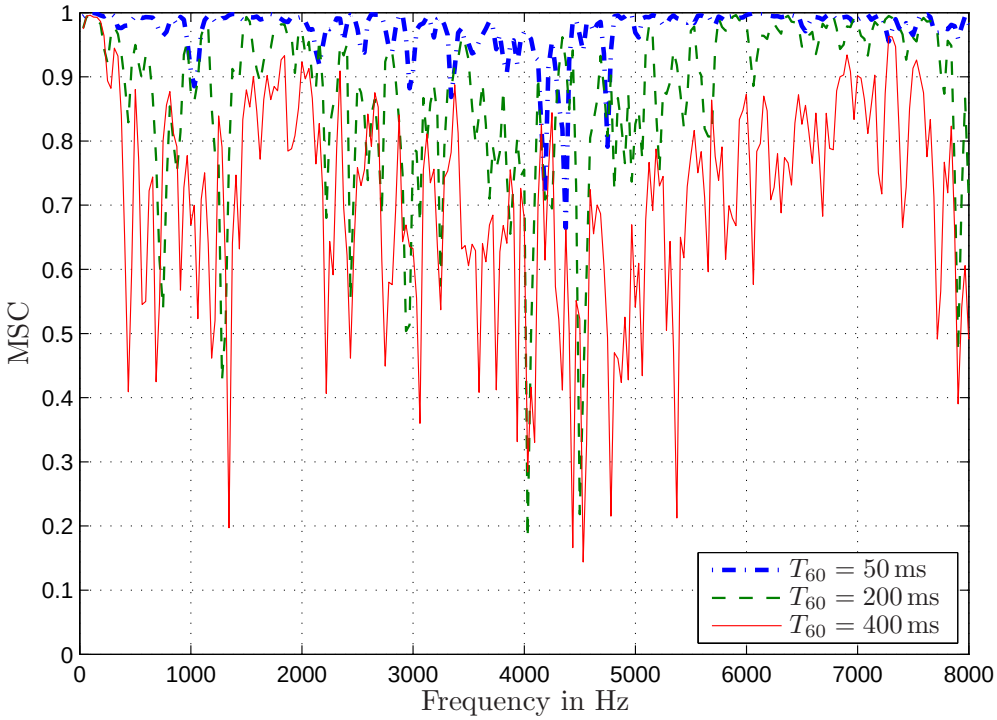


Figure 2.7: Estimate of MSC for a speech source at a distance of 2 m from the sensors in a reverberant environment with different reverberation times T_{60}.

distance to the sensors. Two different rooms as shown in the Appendices C.1 and C.2 with reverberation times T_{60} which can be modified within a range of 50 ms to 400 ms by retractable curtains have been investigated. The result shown in Fig. 2.7 confirms that the increase of reverberation increases also the bias of the MSC.

Additionally, the ratio of direct sound and reverberant sound is influenced by the distance of the point source to the sensors. An important quantity describing the distance where the direct sound component equals the reverberant sound is the critical distance r_h and was described in Section 2.2.2. For distances larger than r_h the reverberant sound outweighs the direct sound component, resulting in an increased bias of the MSC towards zero. This increased bias can be observed in Fig. 2.8 where the source at $\theta = 0°$ was placed at a distance of 0.25 m, 1 m, and 4 m from the two sensors which have a spacing of $d = 20\,\text{cm}$. The room is described in Appendix C.2 and has the dimensions $5.8\,\text{m} \times 5.9\,\text{m} \times 3.14\,\text{m}$ and exhibits a reverberation time of $T_{60} \approx 200\,\text{ms}$. The critical distance for a speech source is given in this room as $r_h \approx 1.6\,\text{m}$. From Fig. 2.8 it can be seen that for a distance larger than r_h a significant bias is introduced.

In [Mar95] it was shown that the use of directional microphones pointing to the source of interest yield a better estimate of the MSC as then also the critical distance is increased (see Section 2.2.2). However, in this thesis only omnidirectional microphones are used as no prior information about the source location is assumed to be available.

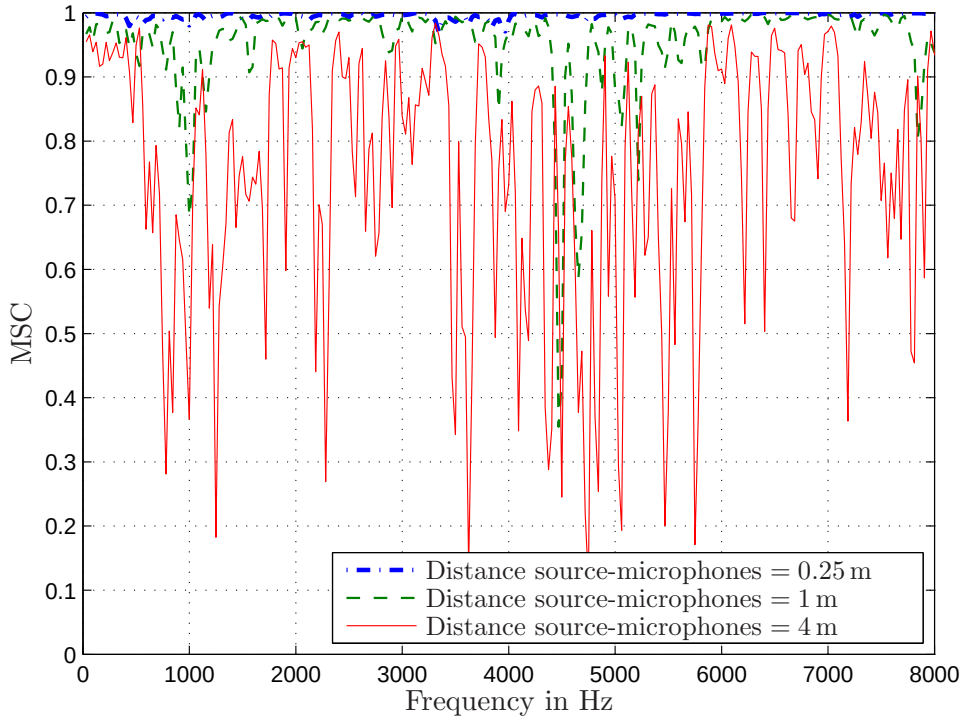


Figure 2.8: Estimate of MSC for a speech source in a reverberant environment ($T_{60} \approx 250\,\text{ms}$) with different distances between the speech source and the sensor array.

We can conclude that for the estimation of the MSC of point sources it is desirable to use large DFT window lengths R to avoid a bias towards zero. The bias becomes more severe the larger the distance between point source and sensors becomes. Additionally, also larger reverberation leads to an increased bias.

2.2.4.3 Magnitude squared coherence of diffuse sound fields

In Section 2.2.3 the properties of an diffuse sound field were discussed. It has been pointed out that the diffuse sound field can be modeled by an infinite number of statistically independent point sources uniformly distributed on a sphere (see also Fig. 2.5). Based on this model, the MSC between the microphone signals $x_1(n)$ and $x_2(n)$ in an ideally diffuse sound field has been derived in Appendix B.1 yielding

$$|\Gamma_{x_1x_2}(\omega)|^2 = \frac{\sin^2\left(\omega f_\text{s} d\, c^{-1}\right)}{\left(\omega f_\text{s} d\, c^{-1}\right)^2}, \tag{2.37}$$

where d denotes the distance between the microphones. This result assumes omnidirectional sensor characteristics and was first presented in [CWB+55]. Fig. 2.9 illustrates (2.37) for different sensor spacings d. In [Mar95, Elk01] the MSC has been investigated also for directional microphones. It should be noted, that (2.37) is only a necessary but not a sufficient condition for a diffuse sound field. Hence, it is possible to construct sound

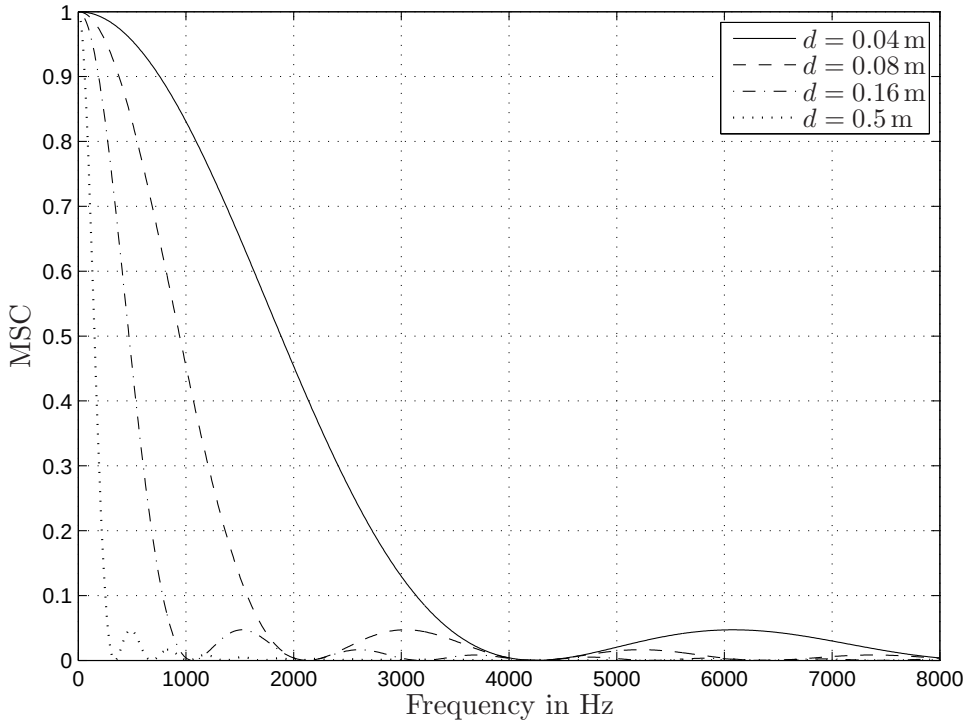
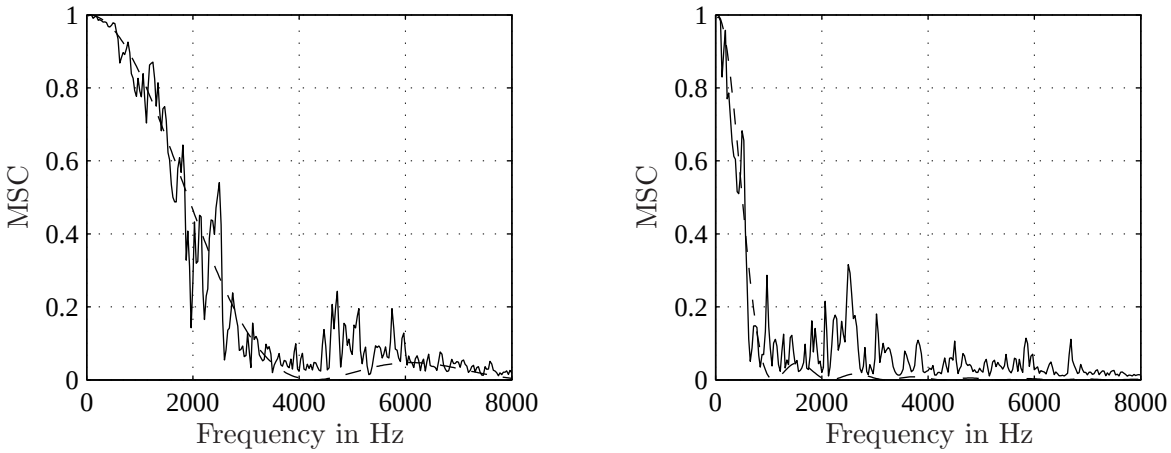


Figure 2.9: MSC $|\Gamma_{x_1x_2}(\omega)|^2$ between two sensors x_1 and x_2 in an ideally diffuse sound field for different sensor spacings d.

fields which have an MSC according to (2.37) but are not ideally diffuse [Däm57].

In Fig. 2.10 the estimated MSC of car noise is shown. According to Appendix C.4 a six-element omnidirectional microphone array was positioned in the passenger compartment at the interior mirror. Two different spacings of $d = 4\,\mathrm{cm}$ and $d = 16\,\mathrm{cm}$ have been chosen. The car noise was measured while driving through a suburban area. The long-term estimate of the MSC (2.35) for a signal length of 20 sec is obtained by using the recursive averaging procedure (2.33) with $\gamma = 0.9$, $\alpha = 2$, DFT length $R = 512$, and choosing the Hann window for $w_f(n)$. It can be seen that the MSC of the measured data (solid) corresponds very well to the $\sin(x)/x$ characteristic of the MSC of an ideal diffuse sound field (dashed) for both microphone spacings. Therefore, it can be concluded that the MSC of car noise can be approximated by the MSC of a diffuse sound field. However, as pointed out above this does not allow the conclusion of a diffuse sound field [Däm57, Mar95]. Additionally, in [Mar01b] it was shown experimentally that also office noise originating from computer fans and hard disk drives can be assumed to exhibit the MSC of a diffuse noise field.

In contrast to the scenario of a point source in a free-field or reverberant environment as examined in Section 2.2.4.2, the noise sources such as in car noise are spatially spread out, e.g., due to vibrating bodies and are thus coming from different directions. Hence, they cannot be modeled as one virtual point source which is picked up by the microphones



(a) d=4 cm (b) d=16 cm

Figure 2.10: MSC $|\Gamma_{x_1x_2}(\omega)|^2$ of car noise measured at two sensors x_1 and x_2 positioned at the interior mirror in a car compartment for different sensor spacings d.

after filtering with room impulse responses. Therefore, in contrast to point sources the MSC estimate is not biased towards zero and is largely independent of the length of the analysis frame as has been confirmed experimentally in [Mar95].

2.2.5 Effects of sensor imperfections and positioning

In this thesis omnidirectional microphones are used as no prior information on the spatial location of the desired sources is available. In conventional beamforming literature (e.g., [Doc03]) these microphones are generally considered to be perfect point receivers with ideal omni-directional properties and a flat frequency response equal to 1. Moreover, the microphone characteristics are assumed to be equal for all sensors in the microphone array. These assumptions are convenient for the beamformer design and thus sensor imperfections will have a detrimental effect on the beamformer characteristics. However, in reality the assumption of ideal and equal microphone characteristics are usually not fulfilled and therefore adaptive calibration schemes have been proposed (see e.g. [Buc02, Buc04, OK05]). The BSS algorithms investigated in this thesis are based on the convolutive model which allows for reverberant environments. Therefore, equalization of sensor imperfections such as, e.g., different frequency responses, phase differences, or shadowing effects as encountered in hearing aid applications can be incorporated into the FIR filters $h_{pq,\kappa}$ of the mixing model. As BSS algorithms have the advantage that no distinction between the component caused by the room acoustics and the component caused by the microphone characteristic has to be made, this avoids the necessity of a calibration procedure.

The microphones can be positioned in different microphone array configurations. The positioning of the microphones can be interpreted as a spatial sampling of the acoustic wave field. To avoid ambiguities in the representation of the acoustic waves, i.e., to avoid spatial aliasing (see, e.g., [vVB88, JD93]), the microphone distance d between two sensors has to fulfill

$$d \leq \frac{\lambda_{\min}}{2} \tag{2.38}$$

for the minimum wavelength $\lambda_{\min}$. This is the spatial analogon to the temporal sampling theorem. For discrete-time signals with sampling frequency f_{s}, this leads to the condition $d \leq c/f_{\mathrm{s}}$. In this thesis $f_{\mathrm{s}} = 16\,\mathrm{kHz}$ is chosen which would correspond to a maximum sensor spacing of $d \leq 2.1\,\mathrm{cm}$. This is an important aspect in the design of fixed and adaptive beamformers [vVB88, Her05]. On the other hand, the concept of BSS does in general not constrain the positioning of the sensors as these methods are assumed to be blind and therefore no information about the mixing system and the location of the sources and sensors is needed. Later in Section 3.1 it will be shown that the presented BSS algorithms do not rely on fulfilling the spatial sampling theorem and thus are applicable to arbitrary array configurations.

2.3 Source signal characteristics and their utilization in blind source separation

In this section we want to discuss the signal properties of acoustic source signals such as speech or music and subsequently discuss their utilization for BSS algorithms. The resulting variety of BSS algorithms based on different signal properties is one of the main motivations for the development of a generic BSS framework in Chapter 3 which allows a simultaneous utilization of all signal properties.

Basic signal properties of acoustic signals

The properties describing the signal statistics are given, e.g., in [Pap02] and will be discussed in the following in the context of acoustic signal processing:

Signal distribution. The signal distribution is described by the probability density function (pdf). Source signals such as speech or music exhibit a nongaussian pdf which can be described by a supergaussian density, i.e., it has a sharper peak and longer tails than the Gaussian pdf and is modeled e.g., by the Laplacian pdf.

Temporal dependencies. If the temporal samples of a signal are uncorrelated, then the signal is termed to be white. If in addition also the higher-order moments do

not depend on the samples, i.e., if the temporal samples are statistically independent, then the signal exhibits strict-sense whiteness. However, audio signals are in general showing temporal dependencies which are introduced, e.g., for speech signals by the vocal tract. Especially the second-order correlations have been investigated in detail in the literature on linear prediction (e.g., [DHP00]). For a sampling frequency of f_s =16 kHz the temporal correlation originating from the vocal tract spans 10 to 20 time lags. These temporal dependencies also affect the signal distribution which is then described by a multivariate pdf (see, e.g., [BS87, GZ03]).

Stationarity. In literature it is distinguished between strict-sense and wide-sense stationarity. The latter one assumes that the mean of the signal is constant and that the second-order correlation only depends on the time-difference and not on the absolute time instants. For strict-sense stationary signals also higher-order moments only depend on the time-difference. The majority of audio signals are considered in the literature as nonstationary signals and wide-sense or strict-sense stationarity is only assumed, e.g., for speech signals within a period of 5 ms to 20 ms [RS78].

Exploitable acoustic source signal properties for BSS criteria

The basic assumption for BSS algorithms is that the source signals $s_q(n)$, $q = 1, \ldots, Q$ are mutually statistically independent. The first BSS algorithms were derived for instantaneous mixtures and no temporal dependencies were taken into account so that originally the mutual statistical independence for temporally white signals given as

$$p_{s,Q}([s_1(n), \ldots, s_Q(n)]^\mathrm{T}) = \prod_{q=1}^{Q} p_{s_q,1}(s_q(n)), \tag{2.39}$$

was used. The variable $p_{s,Q}(\cdot)$ is the joint pdf of dimension Q for all source signals and $p_{s_q,1}(\cdot)$ is the univariate pdf for the q-th source. The dimension of the pdfs are denoted by the subscripts. For the case of convolutive mixtures (2.4) it was shown in [WFO93] that merely utilizing second-order statistics (SOS) by decorrelating the output signals $y_q(n)$ does not lead to a separation of the sources. This implies that we have to force the output signals to become statistically decoupled up to joint moments of a certain order by using additional conditions. This can be realized by exploiting one of the source signal properties discussed above leading to a formulation of several BSS criteria. They can be categorized into approaches exploiting the nongaussiantiy, nonwhiteness, or nonstationarity of the source signals and will be discussed in the following:

(a) **Nongaussianity**. It was stated above that the pdf of an acoustic source signal $s_q(n)$ is in general not Gaussian. Thus, the nongaussianity can be exploited by

using higher-order statistics (HOS) yielding a statistical decoupling of higher-order joint moments of the BSS output signals. BSS algorithms utilizing HOS are also termed independent component analysis (ICA) algorithms (e.g., in instantaneous BSS [Car89, JH91, CJH91, Com94, BS95] and convolutive BSS [Sma98]).

(b) **Nonwhiteness**. As audio signals exhibit temporal dependencies this can be exploited by the BSS criterion. Therefore, it can be assumed that the samples of *each* source signal are not independent along the time axis however, the signal samples from different sources are *mutually* independent. This leads to a generalization of (2.39) given as

$$p_{s,QC}([\mathbf{s}_1(n),\ldots,\mathbf{s}_Q(n)]^{\mathrm{T}}) = \prod_{q=1}^{Q} p_{s_q,C}(\mathbf{s}_q(n)), \tag{2.40}$$

where $\mathbf{s}_q(n)$ contains the C temporally dependent samples of the q-th source, $p_{s,QC}(\cdot)$ is the joint pdf of dimension QC over all sources, and $p_{s_q,C}(\cdot)$ is the multivariate pdf of dimension C of the q-th source. For a speech signal with a sampling frequency of f_{s} =16 kHz, the temporal dependencies introduced by the vocal tract span $C = 15,\ldots,20$ time-lags. The fundamental frequency of voiced sounds, which is in general between 50 Hz and 250 Hz, adds additional temporal dependencies within an interval of up to 20 ms, i.e., $C = 160$. Based on the assumption of mutual statistical independence for non-white sources (2.40) several algorithms can be found in the literature. Mainly the nonwhiteness is exploited using SOS by simultaneous diagonalization of output correlation matrices over multiple time-lags, (e.g., in instantaneous BSS [TLSH91, MS94, BAMCM97, WP97] and convolutive BSS [GC95, KJ00]). It should be noted that convolutive BSS algorithms which are based on the mutual statistical independence (2.39) for temporally white signals instead of (2.40) will aim at removing temporal dependencies and will for the case of audio signals therefore distort the separated output signals.

(c) **Nonstationarity**. The short-term correlations/dependencies of audio signals are in general assumed to be time-variant. Therefore, in most acoustic BSS applications nonstationarity of the source signals can be exploited by simultaneous diagonalization of short-term output correlation matrices at different time instants (e.g., in instantaneous BSS [WFO93, MOK95] and convolutive BSS [KMO98, PSV98, PS00]). The signals within the block, necessary for estimating the SOS correlation matrices, are usually assumed to be wide-sense stationary. A more detailed discussion of block-based estimation methods will be given in Section 3.3.5.

As a simultaneous exploitation of two or even all three signal properties leads to improved results we will present in the next chapter a generic framework which allows to explicitly incorporate nongaussianity, nonwhiteness, and nonstationarity. It should also be noted

that the latter two properties can already be utilized by using second-order correlation matrices. Thus, such algorithms are termed second-order statistics (SOS) BSS approaches.

2.4 Ambiguities in instantaneous and convolutive blind source separation

As the concept of BSS is solely based on the assumption of mutual independence of the source signals there arise some ambiguities. In instantaneous BSS the following indeterminacies appear [HKO01]:

- Scaling ambiguity: The estimated independent components can only be determined up to a scalar factor.
- Permutation ambiguity: The order of the independent components cannot be determined.

As both, the original source signals and the mixing system are unknown, a possible scaling and permutation of the source signals could always be undone by a different mixing system. Due to the impossibility to distinguish if the scaling and permutation occurred in the source signals or in the mixing system, these ambiguities cannot be resolved without using additional a-priori information if only the sensor signals are observed. Thus, the original sources can only be recovered up to an unknown scaling and permutation. In the convolutive BSS case the indeterminacies translate to:

- Filtering ambiguity: The estimated independent components can only be determined up to an arbitrary filtering operation.
- Permutation ambiguity: The order of the independent components cannot be determined.

Again the permutation ambiguity cannot be resolved without further a-priori information. However, if, e.g., the sensor positions are known, then the position of each separated source can be determined from the demixing system [BAS+05, ABWK06]. For some applications this may be sufficient for solving the permutation problem.

In general the scaling ambiguity translates for convolutive BSS to an arbitrary filtering of the output signals. Therefore, it has to be distinguished between blind source separation, where the goal is merely to separate the original source signals, and blind deconvolution of the mixing system (termed blind dereverberation for acoustic signals) with the more challenging task to recover the original source signals up to an arbitrary scaling factor and a constant delay. In acoustics it is difficult to distinguish between the

temporal correlations introduced by the vocal tract of the human speaker and the correlations originating from the reverberation of the room. Therefore, the multi-channel blind deconvolution (MCBD) algorithms which were originally designed for independent, identically distributed (i.i.d.) signals occurring in telecommunication applications are not applicable to solve the blind dereverberation problem (see also discussion in Section 3.2).

Even if we do not strive for solving the dereverberation problem, it is desirable to avoid the arbitrariness of the filtering operation in blind source separation. When discussing the optimum BSS solution in Section 3.1 it will be shown that, even in convolutive BSS, the filtering ambiguity reduces to a scaling ambiguity, if the optimum BSS demixing filter length is chosen. As will be discussed in Section 3.1.4, another popular approach to avoid the arbitrary filtering is to apply a constraint which minimizes the distortion introduced by the demixing system of the BSS algorithm. Thus, the q-th separated source y_q is constrained to be equal to the component of the desired source s_q picked up, e.g., at the q-th microphone. This is done by back-projecting the estimated sources to the sensors or by introducing a constrained optimization scheme [IM99, MN01].

2.5 Performance measures

In general, it is possible to distinguish between subjective and objective tests for assessing the performance of signal processing methods. In subjective testing, listening tests are conducted with a number of test persons which implies a considerable effort for the organization and performance of such tests. A discussion of different subjective test procedures can be found in [VHH98, Jek05]. To reduce the large effort of subjective evaluations it is desirable to substitute the listening tests by instrumental measuring methods (also termed objective measures) which usually compare processed and unprocessed signal in the time or frequency domain [QBC88, VHH98]. A useful objective measure should exhibit a high correlation with the results obtained from subjective evaluations. This can be done by developing objective measures which take perceptual aspects into account Mainly, the research efforts were directed towards the evaluation of speech quality in mobile telecommunications networks (e.g., [Möl00]) which resulted in two recommendations published by the International Telecommunication Union (ITU) [ITU01a, ITU01b]. Recently, there are also activities to use these methods for the evaluation of single-channel speech enhancement signal processing algorithms (see, e.g., [Hub03, RHK05]). However, as this is still an active research area, we will use in this thesis several established objective measures to assess the performance of BSS algorithms. To evaluate the BSS performance appropriately it has to be pointed out that the perceived quality of the BSS output signals is determined by three factors which have to be addressed individually:

(a) Suppression of interfering point sources

(b) Attenuation of background noise

(c) Distortion of the desired signal

In general, BSS algorithms focus on the suppression of interfering point sources and have only a limited capability of attenuating background noise. However, in Chapter 4 several extensions will be addressed which allow additional background noise suppression. Moreover, the application of BSS to commercially successful products also requires small signal distortions of the desired signals. To illustrate these three aspects and to define appropriate objective measures we decompose the output signals $y_q(n)$, $q = 1, \ldots, P$ of the BSS algorithm as

$$y_q(n) = y_{\mathrm{s}_r,q}(n) + y_{\mathrm{c},q}(n) + y_{\mathrm{n},q}(n), \tag{2.41}$$

where $y_{\mathrm{s}_r,q}$ is the component containing the desired source $s_r(n)$. In general, the desired source at the q-th output channel can be any of the source signals due to the permutation ambiguity. In the evaluation of the experiments in this thesis, the output channels have been manually reordered to avoid any permutation. In this no permutation of the output channels occurs and thus $r = q$. The component $y_{\mathrm{c},q}$ is the crosstalk component in the q-th output channel stemming from the remaining point sources that could not be suppressed by the BSS algorithm and $y_{\mathrm{n},q}$ denotes the contributions of the background noise at the q-th output. The microphone signals $x_q(n)$, $q = 1, \ldots, P$ can be decomposed analogously as

$$x_q(n) = x_{\mathrm{s}_r,q}(n) + x_{\mathrm{c},q}(n) + n_q(n), \tag{2.42}$$

where $x_{\mathrm{s}_r,q}(n)$ is the component stemming from the desired source signal $s_r(n)$ and $x_{\mathrm{c},q}(n)$ contains the contributions of the other interfering point sources. The background noise is denoted as $n_q(n)$.

Suppression of interfering point sources

The decomposition of the output signal into desired and interfering components allows to evaluate the interference suppression by calculating the signal-to-interference ratio (SIR) for each channel $q = 1, \ldots, P$ measured in decibel (dB) as

$$\begin{aligned} SIR_{y_q} &= 10 \lg \frac{\mathrm{E}\{y_{\mathrm{s}_r,q}^2\}}{\mathrm{E}\{y_{\mathrm{c},q}^2\}} \\ &\approx 10 \lg \frac{\sum_n y_{\mathrm{s}_r,q}^2(n)}{\sum_n y_{\mathrm{c},q}^2(n)}. \end{aligned} \tag{2.43}$$

As shown in (2.43) that the expectation operator $\mathrm{E}\{\cdot\}$ has to be replaced in practice by a time-average. This estimate of the SIR has only a weak correlation to the quality perceived by auditory measurements [VHH98]. Therefore, due to the nonstationarity of acoustic signals, in literature usually the segmental signal-to-interference ratio (SIR) is preferred.

It is based on time-varying local SIR estimates which are obtained by decomposing the signals into K_S segments/blocks of length N_S. The segment length N_S should be chosen according to the stationarity interval of the observed signals (e.g., 5 to 20 ms for speech). In this thesis we will use $N_\mathrm{S} = 256$ corresponding for a sampling frequency of $f_\mathrm{s} = 16\,\mathrm{kHz}$ to 16 ms. The SIR for the m-th segment is given as

$$SIR_{\mathrm{seg},y_q}(m) = 10\,\lg\left(\frac{\sum_{\kappa=1}^{N_\mathrm{S}} y_{\mathrm{s}_r,q}^2(\kappa + mN_\mathrm{S})}{\sum_{\kappa=1}^{N_\mathrm{S}} y_{\mathrm{c},q}^2(\kappa + mN_\mathrm{S})}\right) \tag{2.44}$$

and the segmental SIR is defined as the average over K_S segments

$$\overline{SIR}_{\mathrm{seg},y_q} = \frac{1}{K_\mathrm{S}} \sum_{m=1}^{K_\mathrm{S}} SIR_{\mathrm{seg},y_q}(m). \tag{2.45}$$

The segmental SIR is very sensitive to periods with low desired signal energy. This poses the problem that in desired signal pauses extremely large negative local SIR values will be encountered. This problem is alleviated by identifying silence periods and excluding them from segmental SIR calculations. Hence, a pause detection is necessary which can be realized, e.g., by comparing long-term and short-term energy of the signal [VHH98]. Another approach is to set a lower threshold and replace all frames with the local SIR below by the threshold [DHP00]. This prevents the measure from being overwhelmed by a few frames of silence. Similarly, frames with SIR_{seg,y_q} greater than 35 dB are not perceived by listeners as being significantly different from the clean desired signal [DHP00]. Therefore, an upper threshold of usually 35 dB is used to limit any unusually high SIR_{seg,y_q} measures. The two thresholds thereby prevent the final segmental SIR measure from being biased in either a positive or a negative direction from a few frames that do not contribute significantly to the overall signal quality. In this thesis the segmental SIR has been calculated according to [DHP00] using a lower threshold of -10 dB and an upper threshold of 35 dB.

If the input segmental SIR at the microphones $\overline{SIR}_{\mathrm{seg},x_q}$ at the q-th microphone

$$\begin{aligned}\overline{SIR}_{\mathrm{seg},x_q} &= \frac{1}{K_\mathrm{S}} \sum_{m=1}^{K_\mathrm{S}} SIR_{\mathrm{seg},x_q}(m) \\ &= \frac{1}{K_\mathrm{S}} \sum_{m=1}^{K_\mathrm{S}} \left(10\,\lg\left(\frac{\sum_{\kappa=1}^{N_\mathrm{S}} x_{\mathrm{s}_r,q}^2(\kappa + mN_\mathrm{S})}{\sum_{\kappa=1}^{N_\mathrm{S}} x_{\mathrm{c},q}^2(\kappa + mN_\mathrm{S})}\right)\right).\end{aligned} \tag{2.46}$$

is not equal to 0 dB then the $\overline{SIR}_{\mathrm{seg},y_q}$ alone is not sufficient in evaluating the separation performance of the BSS algorithm. It is rather the segmental SIR *improvement* in the q-th channel

$$\Delta\overline{SIR}_{\mathrm{seg},q} = \overline{SIR}_{\mathrm{seg},y_q} - \overline{SIR}_{\mathrm{seg},x_q} \tag{2.47}$$

which determines the capability of the BSS algorithm in suppressing interfering point sources. It should be noted that for closely-spaced microphone arrays $\overline{SIR}_{\text{seg},x_q}$ will be very similar for all microphones. However, for large spacings and if objects are between the microphones as, e.g., in binaural hearing aid applications, then $\overline{SIR}_{\text{seg},x_q}$ may differ significantly depending on the microphone location.

Each BSS output contains one desired output signal $y_{\text{s}_r,q}$ and thus, the segmental SIR improvement can be measured at each output $q = 1, \ldots, P$. An averaging over all channels is possible to obtain the average segmental SIR improvement

$$\Delta\overline{SIR}_{\text{seg}} = \frac{1}{P}\sum_{q=1}^{P}\Delta\overline{SIR}_{\text{seg},q}. \tag{2.48}$$

If it is desired to illustrate the convergence behavior of the BSS algorithms, then the quantities $\Delta SIR_{\text{seg},q}(m)$ or $\Delta SIR_{\text{seg}}(m)$

$$\Delta SIR_{\text{seg},q}(m) = SIR_{\text{seg},y_q}(m) - SIR_{\text{seg},x_q}(m) \tag{2.49}$$

$$\Delta SIR_{\text{seg}}(m) = \frac{1}{P}\sum_{q=1}^{P}\Delta SIR_{\text{seg},q}(m) \tag{2.50}$$

are plotted as a function of the segment index m instead of calculating the overall segmental SIR improvement.

Suppression of background noise

BSS is in general not aiming at suppressing background noise. Nevertheless, the influence of the BSS filters on the background noise should also be investigated. Moreover, the extensions presented in Chapter 4 leading to additional background noise suppression also have to be evaluated. Therefore, apart from the segmental signal-to-interference ratio we introduce also the segmental signal-to-noise ratio (SNR) which is defined for the q-th BSS output channel as

$$\begin{aligned}\overline{SNR}_{\text{seg},y_q} &= \frac{1}{K_{\text{S}}}\sum_{m=1}^{K_{\text{S}}} SNR_{\text{seg},y_q}(m) \\ &= \frac{1}{K_{\text{S}}}\sum_{m=1}^{K_{\text{S}}}\left(10\,\lg\left(\frac{\sum_{\kappa=1}^{N_{\text{S}}} y_{\text{s}_r,q}^2(\kappa+mN_{\text{S}})}{\sum_{\kappa=1}^{N_{\text{S}}} y_{\text{n},q}^2(\kappa+mN_{\text{S}})}\right)\right).\end{aligned} \tag{2.51}$$

The segmental SNR at the sensors is defined analogously to (2.46) as

$$\begin{aligned}\overline{SNR}_{\text{seg},x_q} &= \frac{1}{K_{\text{S}}}\sum_{m=1}^{K_{\text{S}}} SNR_{\text{seg},x_q}(m) \\ &= \frac{1}{K_{\text{S}}}\sum_{m=1}^{K_{\text{S}}}\left(10\,\lg\left(\frac{\sum_{\kappa=1}^{N_{\text{S}}} x_{\text{s}_r,q}^2(\kappa+mN_{\text{S}})}{\sum_{\kappa=1}^{N_{\text{S}}} n_q^2(\kappa+mN_{\text{S}})}\right)\right),\end{aligned} \tag{2.52}$$

and the segmental SNR improvement is calculated for each output as

$$\Delta\overline{SNR}_{\text{seg},q} = \overline{SNR}_{\text{seg},y_q} - \overline{SNR}_{\text{seg},x_q}. \tag{2.53}$$

Again it is possible to calculate the average over all output channels analogously to (2.48). Similarly, the quantities $\Delta SNR_{\text{seg},q}(m)$ or $\Delta SNR_{\text{seg}}(m)$ which are defined analogously to (2.49) and (2.50) can be plotted over the segment index m to depict the convergence behavior of the evaluated algorithms.

Distortion of the desired signal

The component in the q-th BSS output channel originating from the desired source $s_r(n)$ is given according to (2.41) as $y_{\text{s}_r,q}$. It should be again noted that only for the case that no permutation is present, the q-th source appears at the q-th BSS output (i.e., $r = q$). To evaluate the signal distortion introduced by the BSS algorithm the desired signal at the q-th BSS output is compared to the desired signal component $x_{\text{s}_r,p}$ at one of the microphones $p = 1, \ldots, P$ or even at all microphones. This is based on the assumption that the aim of BSS is to merely separate the original source signals and to preserve the spectral content of $x_{\text{s}_r,p}$, i.e., the desired source acquired at the q-th sensor. The filtering ambiguity which may lead to an arbitrary filtering would thus lead to a distortion of the desired signal.

A quantity which is reasonably well-correlated with the subjective perception of speech distortion is the logarithmic spectral distance [QBC88]. The unweighted log-spectral distance (SD) measures the Euclidean distance between short-time magnitude spectra in decibel (dB) and is defined for the desired source signal s_r at the q-th output as

$$SD_{s_r,q} = \frac{1}{K_S} \sum_{m=1}^{K_S} \sqrt{\frac{1}{R} \sum_{\nu=0}^{R-1} \left(20 \lg \frac{\left| \underline{Y}_{\text{s}_r,q}^{(\nu)} \right|}{\left| \underline{X}_{\text{s}_r,q}^{(\nu)} \right|} \right)^2} \tag{2.54}$$

where $\underline{Y}_{\text{s}_r,q}^{(\nu)}$ and $\underline{X}_{\text{s}_r,q}^{(\nu)}$ are the DFT-domain representations of $y_{\text{s}_r,q}$ and $x_{\text{s}_r,q}$, respectively. Alternatively, also the spectral envelope of the DFT-domain representation can be used which may be estimated by using linear prediction techniques [QBC88]. This has the benefit that small variations in the fine structure, which have little impact on the speech quality, are not considered in (2.54). To avoid a bias of the SD due to speech pauses, we only include the m-th segment in the arithmetic average if the input SIR and SNR are above -10 dB for this segment. Additionally, care should be taken that the time-domain signals $y_{\text{s}_r,q}$ and $x_{\text{s}_r,q}$ are properly time-aligned before computing the SD. From (2.54) it can be seen that a value of $SD_{s_r,q} = 0\,\text{dB}$ corresponds to no signal distortion.

2.6 Summary

In this chapter we introduced the instantaneous and convolutive BSS mixing model. The latter uses FIR filters to model the mixing and demixing process and is thus applicable to acoustic point sources in free-field and reverberant conditions.

For free-field environments it is possible to model point sources by plane waves and spherical waves. While plane waves can be used to describe signals in the far field of a microphone array, spherical waves can be applied to model the point source in the near field. Due to the superposition principle even reverberant environments could be modeled by the wave equation. To avoid the complexity of calculating the superposition of thousands of reflections usually acoustic impulse responses are used to model the acoustic path between a point source and a sensor. The acoustic impulse responses can be measured experimentally and allow the introduction of several quantities to describe the listening conditions in a room. The most commonly used parameter is the reverberation time T_{60} which is a global indicator for the reverberation of an enclosure. To obtain further information on the room acoustics and to establish a relationship also to speech intelligibility, the "definition" D_{50}, the signal-to-reverberation ratio (SRR), and the critical distance r_h have been introduced.

The diffuse sound field was presented to allow the modeling of several realistic background noises such as, e.g., car noise or speech babble. By utilizing the magnitude squared coherence (MSC) function it can be distinguished between sound fields originating from point sources where the MSC is equal to one and diffuse sound fields which exhibit a $\sin(x)/x$ MSC characteristic. The introduction of the MSC is also important as it will be shown in Section 3.4.3.4 that the MSC is closely related to BSS based on second-order statistics. The examination of the MSC estimation for point sources showed that in reverberant environments the room impulse responses introduce a bias towards zero. This bias can be mitigated by choosing the DFT length R much larger than the room impulse length M. However, in reverberant environments M is very large and thus this condition is usually not fulfilled. In such situations a larger bias occurs if the reverberation time T_{60}, the distance between sources and sensors or the sensor spacing is increased. On the other hand, for diffuse sound fields the MSC estimate is not biased and thus is largely independent of the room parameters.

It has been pointed out that BSS is robust against sensor imperfections and thus no sensor calibration schemes are needed. Moreover, the source signal properties nonwhiteness, nonstationarity, and nongaussianity and their utilization by BSS algorithms have been discussed.

In the end several performance measures have been introduced which allow a separate evaluation of the suppression of point sources, attenuation of background noise and distortion of the desired signal.

3 A Blind Source Separation Framework for Reverberant Environments

In the last chapter it was shown that BSS for acoustic environments requires the convolutive mixing model. Moreover, the three fundamental source signal properties nonwhiteness, nonstationarity, and nongaussianity and their utilization for BSS algorithms have been discussed. Based on the fundamentals in Chapter 2 we will present in this chapter a generic convolutive BSS framework which allows the simultaneous exploitation of the three signal properties and leads to efficient algorithms allowing real-time separation of multiple sources in reverberant environments. This treatment follows own previous publications at international conferences and journals [ABK03, BAK03b, BAK04a, BAK05a] and more references are given when appropriate.

This chapter is organized as follows: In Section 3.1 the optimum BSS solution will be discussed. From the requirement of perfect separation an optimum BSS demixing filter length is derived and the optimum demixing FIR filters are given. Then, in Section 3.2 we give a historical retrospect of convolutive BSS. This leads to a classification of BSS approaches according to the optimization scheme into broadband and narrowband algorithms. Broadband stands for a simultaneous optimization for all frequencies whereas narrowband optimization allows an independent application of the criterion for each frequency. A review of both approaches will be given and it will be emphasized on which signal properties the individual algorithms rely on. This is the basis for the presentation of a generic convolutive BSS framework in Section 3.3 which allows to combine all three signal properties into one optimization criterion operating in the time domain. Several novel efficient gradient and natural gradient algorithms are derived and links to well-known algorithms in the literature are developed. In Section 3.4 the broadband time-domain algorithms are formulated equivalently in the frequency domain leading to fast implementations. By introducing approximations, efficient algorithms can be derived from this framework and links to several well-known narrowband algorithms are established. Moreover, it will be shown that narrowband efficiency can be exploited by broadband algorithms if only *selective* narrowband approximations are introduced. Different update strategies are presented in Section 3.5 and a special emphasis is placed on the derivation of a block-online update rule which allows high separation performance at moderate com-

putational complexity. In the last section the experimental results of several algorithms are analyzed for different reverberant environments.

3.1 Optimum solution for blind source separation

3.1.1 Overall system matrix

In the convolutive BSS model (2.3) and (2.4) in Section 2.2 it can be seen that a concatenation of the FIR mixing filters $h_{qp,\kappa}$ of length M, i.e., $\kappa = 0, \ldots, M-1$ and the FIR demixing filters of length L, given as $w_{pq,\kappa}$, $\kappa = 0, \ldots, L-1$, occurs. The FIR filter taps of the mixing and demixing system can be captured by using vector notation as

$$\mathbf{h}_{qp} = [h_{qp,0}, \ldots, h_{qp,M-1}]^{\mathrm{T}}, \tag{3.1}$$

$$\mathbf{w}_{pq} = [w_{pq,0}, \ldots, w_{pq,L-1}]^{\mathrm{T}}, \tag{3.2}$$

where superscript $^{\mathrm{T}}$ denotes transposition of a vector or a matrix. The channel-wise FIR filters can be combined in matrix notation yielding the $QM \times P$ MIMO mixing matrix $\check{\mathbf{H}}$ and the $PL \times Q$ MIMO demixing matrix $\check{\mathbf{W}}$ defined as

$$\check{\mathbf{H}} = \begin{bmatrix} \mathbf{h}_{11} & \cdots & \mathbf{h}_{1P} \\ \vdots & \ddots & \vdots \\ \mathbf{h}_{Q1} & \cdots & \mathbf{h}_{QP} \end{bmatrix}, \tag{3.3}$$

$$\check{\mathbf{W}} = \begin{bmatrix} \mathbf{w}_{11} & \cdots & \mathbf{w}_{1Q} \\ \vdots & \ddots & \vdots \\ \mathbf{w}_{P1} & \cdots & \mathbf{w}_{PQ} \end{bmatrix}, \tag{3.4}$$

where Q denotes the number of sources and P stands for the number of microphones. It should be noted that in Section 3.1 we still distinguish between the number of sources and the number of microphones as this will give some further insight into the optimum BSS demixing filter length. From Section 3.2 on we will only consider the square case $Q = P$. Inserting the mixing equation (2.3) into the demixing equation (2.4) results in a convolution of the mixing and demixing FIR filters yielding the overall system FIR filters of length $M+L-1$. The overall system FIR filter from the q-th source to the r-th output of the demixing system $(q, r = 1, \ldots, Q)$ is denoted by the column vector $\mathbf{c}_{qr}$ given as

$$\mathbf{c}_{qr} = [c_{qr,0}, \ldots, c_{qr,M+L-2}]^{\mathrm{T}}, \tag{3.5}$$

and a combination of all channels yields the overall system matrix $\check{\mathbf{C}}$ of dimensions $Q(M+L-1) \times Q$ defined as

$$\check{\mathbf{C}} = \begin{bmatrix} \mathbf{c}_{11} & \cdots & \mathbf{c}_{1Q} \\ \vdots & \ddots & \vdots \\ \mathbf{c}_{Q1} & \cdots & \mathbf{c}_{QQ} \end{bmatrix}. \tag{3.6}$$

In general, a *convolution* between two sequences can also be *written as a matrix-vector product* if the matrix exhibits a special Toeplitz structure. By using this procedure we can express the FIR filter model $\mathbf{c}_{qr}$ from the q-th source to the r-th output as

$$\mathbf{c}_{qr} = \sum_{p=1}^{P} \mathbf{H}_{qp,L} \mathbf{w}_{pr} \tag{3.7}$$

with the $M + L - 1 \times L$ matrix $\mathbf{H}_{qp,L}$, containing the M filter taps in each column, given as

$$\mathbf{H}_{qp,L} = \begin{bmatrix} h_{qp,0} & 0 & \cdots & 0 \\ h_{qp,1} & h_{qp,0} & \ddots & \vdots \\ \vdots & h_{qp,1} & \ddots & 0 \\ h_{qp,M-1} & \vdots & \ddots & h_{qp,0} \\ 0 & h_{qp,M-1} & \ddots & h_{qp,1} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & h_{qp,M-1} \end{bmatrix}. \tag{3.8}$$

Matrices exhibiting the special Toeplitz structure given in (3.8) are termed Sylvester, convolution, or transmission matrices (see, e.g., [Eks73, MZB87, MK88]) and allow the expression of the convolution as a matrix-vector product. In the remainder of this thesis we will use the term Sylvester matrix. To ensure the equivalence between matrix-vector product and linear convolution, the width of the Sylvester matrix has to be adjusted to correspond to the length of the column vector in (3.7). Therefore, the width of the Sylvester matrix in (3.7) and (3.8) is indicated by the subscript L. The necessity of defining the width of the Sylvester matrix will become especially important when expressing a concatenation of several linear convolutions as matrix-vector products (see, e.g., Appendix A.2).

A combination of all channels yields the MIMO block-Sylvester matrix $\mathbf{H}_L$ of dimensions $Q(M + L - 1) \times PL$ given as

$$\mathbf{H}_L = \begin{bmatrix} \mathbf{H}_{11,L} & \cdots & \mathbf{H}_{1P,L} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{Q1,L} & \cdots & \mathbf{H}_{QP,L} \end{bmatrix}, \tag{3.9}$$

where the subscript L indicates again the width of the channel-wise Sylvester matrices. The concatenation of the mixing and demixing MIMO system results in the $Q(M + L - 1) \times Q$ overall system matrix $\check{\mathbf{C}}$ which was introduced in (3.6). The concatenation can be expressed as a matrix product by using the block-Sylvester matrix $\mathbf{H}_L$ given in (3.9) and the block matrix $\check{\mathbf{W}}$ defined in (3.4) leading to

$$\check{\mathbf{C}} = \mathbf{H}_L \check{\mathbf{W}}. \tag{3.10}$$

3.1.2 Optimum BSS solution and resulting optimum demixing filter length

In BSS the goal is the separation of the original sources at the outputs of the demixing system. In terms of the overall system matrix $\check{\mathbf{C}}$ this condition can be written as

$$\check{\mathbf{C}} - \mathrm{bdiag}\{\check{\mathbf{C}}\} = \mathrm{boff}\{\check{\mathbf{C}}\} \stackrel{!}{=} \mathbf{0}. \tag{3.11}$$

The operator bdiag applied on a block matrix consisting of several submatrices or vectors sets all submatrices or vectors on the off-diagonals to zero. Analogously, the boff operation sets all submatrices or vectors off the diagonal to zero (for more details on operators for block matrices see Appendix A.1). In (3.11) this corresponds to $\mathbf{c}_{qr} \stackrel{!}{=} \mathbf{0}$, $\forall\, q \neq r$. In the formulation of (3.11) it was assumed, without loss of generality, that no output channel order permutation is present. In this case the optimum solution yields the q-th source at the q-th output of the demixing system and all cross-channel terms $\mathbf{c}_{qr}$, $q \neq r$ are zero. It should be noted that in case of an output channel order permutation the vectors $\mathbf{c}_{qr}$ contained in the optimum $\check{\mathbf{C}}$ are permuted, but in each column and row there is at most one vector unequal to zero, which still guarantees perfect separation of the sources.

By inserting the definition of the overall system matrix (3.10) in (3.11) the optimum BSS solution can be expressed as

$$\mathrm{boff}\{\mathbf{H}_L\check{\mathbf{W}}\} \stackrel{!}{=} \mathbf{0}. \tag{3.12}$$

This relation allows us to derive a lower bound for the demixing FIR filter length L for which the optimum solution (3.11) can still be achieved. In Fig. 3.1 the matrix product $\check{\mathbf{C}} = \mathbf{H}_L\check{\mathbf{W}}$ is illustrated for the optimum solution in the case $Q = P = 3$. In general, for the q-th column $\mathbf{w}_{\mathrm{col},q} = [\mathbf{w}_{1q}^{\mathrm{T}}, \ldots, \mathbf{w}_{Pq}^{\mathrm{T}}]^{\mathrm{T}}$ of the demixing FIR filter matrix $\check{\mathbf{W}}$ a homogeneous linear system of equations can be established by picking the subset of Sylvester matrices $\mathbf{H}_{rp,L}$ with $r = 1, \ldots, Q$, $p = 1, \ldots, P$, and $r \neq q$. The subset for the q-th column $\mathbf{w}_{\mathrm{col},q}$ of the demixing filter matrix is obtained by removing the q-th row of submatrices and is denoted as $\mathbf{H}_{\mathrm{sub},q}$. In Fig. 3.1 the procedure to generate a homogeneous linear system is illustrated for $q = 1$. The two shaded areas show the first column $\mathbf{w}_{\mathrm{col},1}$ of the filter matrix $\check{\mathbf{W}}$ and the subset $\mathbf{H}_{\mathrm{sub},1}$ of $\mathbf{H}_L$, respectively. Thus, to determine the optimum demixing filter coefficients, for each column $\mathbf{w}_{\mathrm{col},q}$ a homogeneous linear system of equations

$$\mathbf{H}_{\mathrm{sub},q}\mathbf{w}_{\mathrm{col},q} = \mathbf{0} \tag{3.13}$$

has to be solved. The homogeneous linear system is obviously solved by the trivial solution $\mathbf{w}_{\mathrm{col},q} = \mathbf{0}$. Additional non-trivial solutions are obtained if the rank of the matrix $\mathbf{H}_{\mathrm{sub},q}$ is smaller than the number of elements of $\mathbf{w}_{\mathrm{col},q}$ [Har97].

In the following it is assumed that $\mathbf{H}_{\mathrm{sub},q}$ has full row rank. This assumption can be interpreted such that the FIR acoustic impulse responses contained in $\mathbf{H}_{\mathrm{sub},q}$ do not

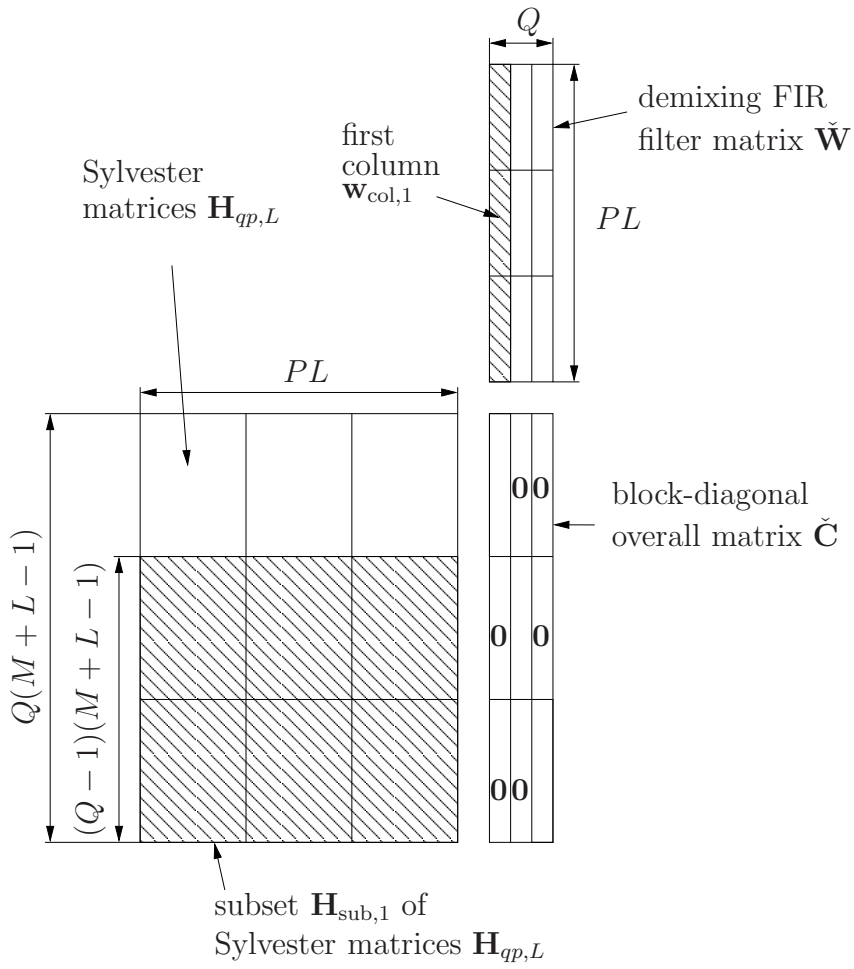


Figure 3.1: Illustration of the matrix product $\check{\mathbf{C}} = \mathbf{H}_L\check{\mathbf{W}}$ for the optimum solution $\mathrm{boff}\{\check{\mathbf{C}}\} \stackrel{!}{=} \mathbf{0}$ and $Q = P = 3$.

possess any common zeros in the z-domain. This assumption usually holds in practice and leads to the requirement that, for obtaining non-trivial solutions, the number of elements of $\mathbf{w}_{\mathrm{col},q}$ (i.e., the number of columns of $\mathbf{H}_{\mathrm{sub},q}$) have to outnumber the number of rows of $\mathbf{H}_{\mathrm{sub},q}$, i.e., $PL > (Q-1)(M+L-1)$. Solving for L yields the lower bound for the demixing filter length as

$$L > \frac{Q-1}{P-Q+1}(M-1). \tag{3.14}$$

If (3.14) is fulfilled then there exists an *infinite number of solutions* and the space spanned by the linearly independent solutions is called the solution space of the homogeneous linear system (3.13) or the null space of matrix $\mathbf{H}_{\mathrm{sub},q}$. The dimension of the null space of $\mathbf{H}_{\mathrm{sub},q}$, i.e., the number of linearly independent solutions, is given as the difference of the length of the vector $\mathbf{w}_{\mathrm{col},q}$ and the row rank of the matrix $\mathbf{H}_{\mathrm{sub},q}$, i.e., $PL - \mathrm{rank}(\mathbf{H}_{\mathrm{sub},q})$ [Har97]. Due to the full row rank, the rank is given as $\mathrm{rank}(\mathbf{H}_{\mathrm{sub},q}) = (Q-1)(M+L-1)$. If PL is

chosen to rank$(\mathbf{H}_{\text{sub},q}) + 1$, then the dimension of the null space is equal to one, i.e., *only one linearly independent solution* $\mathbf{w}_{\text{col},q}$ exists. This choice of L is denoted the optimum BSS filter length L_{opt} and is given as

$$L_{\text{opt}} = \frac{Q-1}{P-Q+1}(M-1)+1. \tag{3.15}$$

This choice means that only one linearly independent solution $\mathbf{w}_{\text{col},q}$ is possible and all other solutions in the null space are linearly dependent vectors obtained by multiplying $\mathbf{w}_{\text{col},q}$ with a scalar factor α_q. Hence, for L_{opt} the *arbitrary filtering* of the BSS output signals as discussed in Section 2.4 *reduces to* an *arbitrary scaling*.

It should be noted that the result for the optimum BSS filter length has also been derived in a different way independently in [Hof04, Hof05].

3.1.3 Optimum BSS demixing system and relationship to blind MIMO identification

In the previous section the optimum BSS filter length has been derived. Here, we discuss the optimum demixing system $\check{\mathbf{W}}_{\text{opt}}$ which can be obtained for the choice $L = L_{\text{opt}}$. For simplicity we will first start with the special case $Q = P = 2$ and then subsequently generalize the result to the square case for $P, Q > 2$.

Square case using two sources and two sensors ($Q = P = 2$).

According to (3.13) and Fig. 3.1 we have to solve for the case $Q = P = 2$ two homogeneous linear systems given as

$$\mathbf{H}_{11,L}\mathbf{w}_{12} + \mathbf{H}_{12,L}\mathbf{w}_{22} = \mathbf{0}, \tag{3.16}$$

$$\mathbf{H}_{21,L}\mathbf{w}_{11} + \mathbf{H}_{22,L}\mathbf{w}_{21} = \mathbf{0}. \tag{3.17}$$

The matrix-vector products in the equations (3.16), (3.17) represent convolutions of the FIR filters $\mathbf{h}_{qp}$ and $\mathbf{w}_{pq}$ which can also be written equivalently as a multiplication in the z-domain yielding [BAK05b, BAK07]

$$H_{11}(z)W_{12}(z) + H_{12}(z)W_{22}(z) = 0, \tag{3.18}$$

$$H_{21}(z)W_{11}(z) + H_{22}(z)W_{21}(z) = 0. \tag{3.19}$$

Due to the FIR filter structure the z-domain representations can be expressed by the zeros $z_{0H_{qp},\nu}$, $z_{0W_{pq},\mu}$ and the gains $A_{H_{qp}}$, $A_{W_{pq}}$ of the filters $H_{qp}(z)$ and $W_{pq}(z)$ respectively:

$$A_{H_{11}} \prod_{\nu=1}^{M-1}(z - z_{0H_{11},\nu}) A_{W_{12}} \prod_{\mu=1}^{L-1}(z - z_{0W_{12},\mu}) = -A_{H_{12}} \prod_{\nu=1}^{M-1}(z - z_{0H_{12},\nu}) A_{W_{22}} \prod_{\mu=1}^{L-1}(z - z_{0W_{22},\mu})$$

(3.20)

$$A_{H_{21}} \prod_{\nu=1}^{M-1} (z - z_{0H_{21},\nu}) A_{W_{11}} \prod_{\mu=1}^{L-1} (z - z_{0W_{11},\mu}) = -A_{H_{22}} \prod_{\nu=1}^{M-1} (z - z_{0H_{22},\nu}) A_{W_{21}} \prod_{\mu=1}^{L-1} (z - z_{0W_{21},\mu}) \tag{3.21}$$

As pointed out before, the matrices $\mathbf{H}_{\text{sub},q}$, $q = 1, 2$ are assumed to be full row rank and this translates in the z-domain to the assumption that $H_{11}(z)$ and $H_{12}(z)$ in (3.20) and $H_{21}(z)$ and $H_{22}(z)$ in (3.21) do not share common zeros. If no common zeros exist and if the optimum demixing filter length, given for the case $Q = P = 2$ as $L = M$, is chosen, then the equality in (3.20) can only hold if the zeros of the demixing filters are chosen as $z_{0W_{12},\mu} = z_{0H_{12},\mu}$ and $z_{0W_{22},\mu} = z_{0H_{11},\mu}$ for $\mu = 1, \ldots, M-1$. Analogously, the equality in (3.21) requires $z_{0W_{11},\mu} = z_{0H_{22},\mu}$ and $z_{0W_{21},\mu} = z_{0H_{12},\mu}$ for $\mu = 1, \ldots, M-1$. Additionally, to fulfill the equality, the gains of the demixing filters in (3.20) have to be chosen as $A_{W_{22}} = \alpha_2 A_{H_{11}}$ and $A_{W_{12}} = -\alpha_2 A_{H_{12}}$, where α_2 is an arbitrary scalar constant. Thus, the demixing filters are only determined up to a scalar factor α_2. Analogously, for the equality (3.21) the gains of the demixing filters are given as $A_{W_{11}} = \alpha_1 A_{H_{22}}$ and $A_{W_{21}} = -\alpha_1 A_{H_{21}}$ with the scalar constant α_1. In summary, this leads to the optimum demixing filter matrix $\check{\mathbf{W}}_{\text{opt}}$ given in the time domain as

$$\check{\mathbf{W}}_{\text{opt}} = \begin{bmatrix} \alpha_1 \mathbf{h}_{22} & -\alpha_2 \mathbf{h}_{12} \\ -\alpha_1 \mathbf{h}_{21} & \alpha_2 \mathbf{h}_{11} \end{bmatrix}, \tag{3.22}$$

where due to the scaling ambiguity each column of the block-adjoint is multiplied by an unknown scalar α_q.

It should be noted that the optimum solution for $Q = P = 2$, given by (3.22), performs blind MIMO system identification up to an arbitrary scalar constant. Thus, BSS algorithms aiming at the optimum solution (3.22) can be interpreted as an extension of single-input multi-output (SIMO) system identification approaches (see e.g., [XLTK95, GN95]) as was pointed out in [BAK05b, BAK07]. If additionally the geometrical information about the sensor locations is taken into account, then algorithms performing SIMO identification are often used for single-source localization (see, e.g., [Ben00]). Hence, the extension to MIMO system identification allows for BSS algorithms capable of performing multiple-source localization as has been demonstrated in [BAS+05, BAK07].

Additionally, (3.22) can be interpreted as a blind interference canceller for each BSS output channel. Due to the blind MIMO system identification property of the acoustic impulse responses between the sources and the sensors, the optimum BSS solution does not depend on the positioning of the sources or the sensors. Thus, BSS algorithms leading to (3.22) can be applied to arbitrary array geometries and hence do not suffer from spatial aliasing .

To see how the overall system matrix $\check{\mathbf{C}}$ behaves in the case of the optimum solution (3.22), the optimum demixing filters $\check{\mathbf{W}}_{\text{opt}}$ (with the optimum demixing filter length $L = M$) are inserted into (3.10) yielding

$$\begin{aligned}
\check{\mathbf{C}}_{\text{opt}} &= \mathbf{H}_M \check{\mathbf{W}}_{\text{opt}} \\
&= \begin{bmatrix} \mathbf{H}_{11,M} & \mathbf{H}_{12,M} \\ \mathbf{H}_{21,M} & \mathbf{H}_{22,M} \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \mathbf{h}_{22} & -\alpha_2 \mathbf{h}_{12} \\ -\alpha_1 \mathbf{h}_{21} & \alpha_2 \mathbf{h}_{11} \end{bmatrix} \\
&= \begin{bmatrix} \alpha_1(\mathbf{H}_{11,M}\mathbf{h}_{22} - \mathbf{H}_{12,M}\mathbf{h}_{21}) & \alpha_2(\mathbf{H}_{12,M}\mathbf{h}_{11} - \mathbf{H}_{11,M}\mathbf{h}_{12}) \\ \alpha_1(\mathbf{H}_{12,M}\mathbf{h}_{22} - \mathbf{H}_{22,M}\mathbf{h}_{21}) & \alpha_2(\mathbf{H}_{22,M}\mathbf{h}_{11} - \mathbf{H}_{21,M}\mathbf{h}_{12}) \end{bmatrix}.
\end{aligned} \tag{3.23}$$

Due to the Sylvester structure, each matrix-vector product in (3.23) denotes a linear convolution of two FIR filters. The linear convolution is commutative and thus, the order of the FIR filters in the matrix-vector product may be interchanged (e.g. $\mathbf{H}_{22,M}\mathbf{h}_{11} = \mathbf{H}_{11,M}\mathbf{h}_{22}$). Hence, (3.23) can be simplified to

$$\check{\mathbf{C}}_{\text{opt}} = \begin{bmatrix} \alpha_1(\mathbf{H}_{11,M}\mathbf{h}_{22} - \mathbf{H}_{12,M}\mathbf{h}_{21}) & \mathbf{0} \\ \mathbf{0} & \alpha_2(\mathbf{H}_{11,M}\mathbf{h}_{22} - \mathbf{H}_{12,M}\mathbf{h}_{21}) \end{bmatrix}. \tag{3.24}$$

From (3.24) it can be seen that as desired, the optimum solution achieves perfect separation because $\text{boff}\{\check{\mathbf{C}}_{\text{opt}}\} = \mathbf{0}$. However, the separated outputs of the overall system will be filtered versions of the original source signals due to the terms on the block-diagonal of $\check{\mathbf{C}}_{\text{opt}}$. These block-diagonal terms are column vectors representing FIR filters of length $2M - 1$. They differ only by a scalar constant, which shows again that for the case $L = L_{\text{opt}}$ the optimum solution does not lead to an arbitrary filtering, but only to an arbitrary scaling. Using the block determinant operator defined in the Appendix A.2 we can write (3.24) compactly as

$$\check{\mathbf{C}}_{\text{opt}} = \begin{bmatrix} \alpha_1 \text{bdet}_2\{\check{\mathbf{H}}\} & \mathbf{0} \\ \mathbf{0} & \alpha_2 \text{bdet}_2\{\check{\mathbf{H}}\} \end{bmatrix}. \tag{3.25}$$

General square case with more than two sources and sensors.

Now, we want to see how the above results generalize to the square case $Q = P$ with $P, Q > 2$. It can be seen that the optimum solution given in (3.22) for the case $Q = P = 2$ is the adjoint of the matrix $\check{\mathbf{H}}$ with its entries $\mathbf{h}_{qp}$ treated as scalar values. This operation is formalized in the Appendix A.2 by the introduction of the *block-adjoint operator* $\text{badj}_P(\cdot)$. There it is shown that the block-adjoint operator can be applied to block matrices with $Q = P$ and $P, Q \geq 2$. Due to the FIR filters contained in $\check{\mathbf{H}}$ the block-adjoint involves linear convolutions for $P, Q > 2$ and thus, in the general case the size of $\text{badj}_P\left\{\check{\mathbf{H}}\right\}$ is determined by the length of the convolutional product given as $P(M-1)+1 \times P$ (for more details on the block-adjoint operator see Appendix A.2). Hence, for the general

square case $Q = P$ and if the optimum BSS demixing filter length according to (3.15) is chosen, then the optimum demixing system $\check{\mathbf{W}}_{\text{opt}}$ is given as

$$\check{\mathbf{W}}_{\text{opt}} = \text{badj}_P \left\{\check{\mathbf{H}}\right\} \mathbf{\Lambda}_\alpha. \tag{3.26}$$

The diagonal matrix $\mathbf{\Lambda}_\alpha = \text{Diag}\{[\alpha_1, \ldots, \alpha_P]^{\text{T}}\}$ accounts for the scaling ambiguity. The operator $\text{Diag}\{\mathbf{a}\}$ denotes a square matrix with the elements of vector $\mathbf{a}$ on its main diagonal. To see that (3.26) really is the optimum solution for $P, Q \geq 2$, we examine the overall system matrix. Inserting the optimum solution $\check{\mathbf{W}}_{\text{opt}}$ given in (3.26) into (3.10) leads to

$$\begin{aligned} \check{\mathbf{C}}_{\text{opt}} &= \mathbf{H}\, \text{badj}_P \left\{\check{\mathbf{H}}\right\} \mathbf{\Lambda}_\alpha \\ &= \text{Bdiag}\left\{\text{bdet}_P\{\check{\mathbf{H}}\}, \ldots, \text{bdet}_P\{\check{\mathbf{H}}\}\right\}\ \mathbf{\Lambda}_\alpha, \end{aligned} \tag{3.27}$$

where for the second line the equation (A.10) in the Appendix A.2 has been used. The operator $\text{bdet}_P\{\check{\mathbf{H}}\}$ denotes the block determinant of the mixing system as defined in Appendix A.2. The block determinant $\text{bdet}_P\{\check{\mathbf{H}}\}$ has the dimensions $P(M-1)+1 \times 1$ and can be interpreted as an FIR filter of length $P(M-1)+1$. The result in (3.27) shows that $\text{boff}\{\check{\mathbf{C}}_{\text{opt}}\} = \mathbf{0}$, i.e., that perfect separation is achieved, for the optimum $\check{\mathbf{W}}_{\text{opt}}$ given in (3.26). On the other hand, (3.27) also shows that for the case $Q = P$ the original source signals will be filtered by the FIR filter of length $P(M-1)+1$ given by $\text{bdet}_P\{\check{\mathbf{H}}\}$. As the filter length is larger than the mixing filter length M this introduces additional reverberation at the BSS outputs. In the case $Q = P$ this problem becomes more severe if the number of sources and sensors P, Q increases. A remedy to this problem are additional constraints as discussed in the next section.

3.1.4 Constraining the optimum BSS solution to additionally minimize output signal distortions

In the previous section it was shown that perfect separation can be obtained if the adaptive BSS algorithm leads to the optimum demixing FIR filters according to (3.26). Then the *arbitrary* filtering reduces to an arbitrary scaling of the BSS output signals. Nevertheless, each source signal at the BSS output is still filtered by the FIR filter $\mathbf{c}_{qq}$ $(q = 1, \ldots, P)$ with $\mathbf{c}_{qq} = \text{bdet}_P\{\check{\mathbf{H}}\}$, $\forall\ q$.

As pointed out previously in Section 2.4, an approach to avoid filtering of the separated sources is to constrain, in addition to the block-offdiagonal of $\check{\mathbf{C}}$, also the block-diagonal of $\check{\mathbf{C}}$. A popular approach proposing such a constraint was presented in [MN01] and requires that the distortion caused by the demixing system should be minimized so that the q-th output signal y_q exhibits the same spectral envelope as each source signal s_q picked up at

the q-th sensor ($q = 1, \ldots, P$). This constraint corresponds to

$$\text{bdiag}\{\check{\mathbf{C}}\} \stackrel{!}{=} \text{bdiag}\{\check{\mathbf{H}}\}. \tag{3.28}$$

Usually, this constraint has to be incorporated in BSS optimization criteria by using the method of Lagrange multipliers or by using projection methods (see, e.g., [Fle81]).

3.1.5 Summary

In this section the optimum BSS solution was discussed. First we showed how the concatenation of the mixing and demixing MIMO systems can be expressed conveniently using matrix notation leading to an overall system matrix $\check{\mathbf{C}}$. Perfect separation is achieved if the block-offdiagonal elements of the matrix $\check{\mathbf{C}}$ are equal to zero (in the case of no output channel permutation). This requirement led to an equation yielding the optimum BSS demixing filter length. Subsequently, based on the optimum filter length, the optimum BSS demixing filters were given, first for the case $Q = P = 2$ and then for the more general case $Q = P$, $P, Q > 2$. It could be seen that this optimum solution results in MIMO identification of the mixing system reducing the arbitrary filtering to an arbitrary scaling of the BSS output signals. Moreover, this relationship allows for the application of BSS algorithms to multiple-source localization problems as was shown in [BAS+05, BAK07]. Finally, it was pointed out that in the case of the optimum demixing filters the BSS output signals will be filtered versions of the original sources. This filtering is not arbitrary and depends on the mixing system. For the case $Q = P$ considered in this thesis, the filtering becomes more severe if $P, Q > 2$. A remedy to this problem is the introduction of an additional constraint as shown in [MN01].

In the next sections we will first give an overview about the different BSS approaches in literature. Subsequently, in Section 3.3 we will introduce a novel optimization criterion leading to algorithms aiming at estimating the optimum BSS demixing filters as shown above.

3.2 Broadband versus narrowband optimization

In literature the various BSS algorithms are usually categorized into time-domain and frequency-domain algorithms. However, especially for BSS algorithms another important classification is based on the signal model which is used for the optimization scheme. One can distinguish the *narrowband signal model* where the BSS algorithms are designed in the DFT domain under the assumption that individual frequency bins of the input signals can be considered independently from each other and the *broadband signal model* where the optimization is performed for all frequency bins simultaneously. Comparing these two

categories with the traditional classification into time- and frequency domain[1] algorithms it can be seen that in total three approaches are conceivable:

(a) Optimization and implementation in the time domain and thus, inherently using the broadband signal model.

(b) Optimization in the time domain based on the broadband signal model together with the implementation in the DFT domain.

(c) Optimization and implementation in the DFT domain based on the narrowband signal model.

In this thesis all three approaches will be covered by firstly addressing (a) the optimization in the time-domain, resulting in a generic framework for convolutive BSS in Section 3.3. Subsequently, in Section 3.4 the framework will be extended to broadband algorithms in the DFT domain and thus, exploiting the efficiency of fast convolutions (b). This class is especially important as by the introduction of approximations also links to novel and well-known algorithms based on narrowband optimization schemes (c) can be developed (Section 3.4.3). There, the close connection between (b) and (c) becomes obvious in the derivation of novel hybrid algorithms obtained by using only *selective* approximations. Before deriving this novel framework a short overview of important work on convolutive BSS will be given and the various algorithms with respect to the three optimization schemes (a)-(c) will be classified. In the historical retrospect in [JT00] it was pointed out that the early works on source separation and the genesis of the concept itself can be traced back to the early 80's. Since the early 90's it received an increasing interest in the signal processing commmunity but most research was still dealing with instantaneous mixtures. However, with the application of BSS to acoustic mixtures in mind the research focused on BSS for convolutive mixtures since the mid 90's [Tor99]. Since then a vast amount of literature on convolutive BSS has been published and the major algorithmic milestones are now presented with respect to the optimization scheme.

(a) Optimization and implementation in the time domain

In early works feedback demixing systems were proposed [Tor96a, Tor96b, LBL97] resulting in infinite impulse response (IIR) filters. However, the drawback of IIR filters is that depending on the mixing system they may become unstable. Therefore, nowadays usually an FIR demixing system is used as illustrated in Fig. 2.1 in Section 2.2.

[1]It should be noted that the precise definition of the term frequency domain refers to the domain obtained by applying the discrete-time Fourier transform (DTFT). However, in adaptive signal processing literature (e.g., [Hay02]) the term frequency domain is usually used to denote the domain obtained by application of the discrete Fourier transform (DFT). Hence, in this thesis the term frequency domain refers to the application of the DFT unless otherwise noted.

Most algorithms which perform the optimization in the time domain and thus, inherently use the broadband signal model, introduce a temporal whitening of the original sources, e.g., [Lam96, ADCY97, KMO98, ZCA99, CACL99, DSM03]. The reason for the whitening is that these algorithms not only perform source separation but try to deconvolve the temporal correlation which is introduced by the mixing system. However, these algorithms cannot distinguish between correlation arising from the impulse responses of the mixing system and correlation introduced by the sources (e.g., by the means of the vocal tract) and therefore whiten the original source signals. This class is termed multichannel blind deconvolution (MCBD) algorithms. Due to the whitening effect they were initially proposed for independent, identically distributed (i.i.d.) telecommunication signals (e.g., [ADCY97, CACL99, JS03]). Later on, these algorithms have also been applied to acoustic signals (e.g. [MOTN03, DSM04b]). To prevent the whitening of the acoustic sources different heuristic counter-measures have been proposed. In [MN01] the minimal distortion principle has been presented which introduces a constraint minimizing the distortion caused by the demixing system of the MCBD algorithm. Another possibility is the temporal prewhitening of the acoustic mixtures by the application of linear prediction analysis filters prior to the MCBD algorithm [KZN03], leading to a mere spatial separation or by restoring the original temporal correlations after the MCBD algorithm by the use of linear prediction synthesis filters [SD01].

Also some algorithms based on time-domain optimization have been proposed which do not introduce a temporal whitening. In [GC95] an algorithm based on second-order statistics (SOS) has been presented but has been restricted to the adaptation of causal filters. In [KJ00, Joh04] algorithms based on the Frobenius norm have been proposed. In [BSBAM01] a criterion based on block-diagonalization of correlation matrices has been presented but no update procedure has been given. The same criterion was used in [ABK03, BAK03b, BAK05a] for the derivation of gradient and natural gradient algorithms.

(b) Implementation in the DFT domain with optimization based on the broadband model

Recently, also some algorithms were proposed performing the optimization based on the broadband model but implementing the algorithm in the DFT domain to exploit the efficiency of fast convolutions. In [JS03, DSM04a] two algorithms for MCBD based on higher-order statistics (HOS) were derived. BSS algorithms based on a second-order statistics (SOS) criterion were presented in [ABK03, BAK03b, BAK05a] and an extension to a HOS criterion was proposed in [BAK04a]. A more detailed discussion of such BSS approaches will be given in Section 3.4.

(c) Implementation in the DFT domain with optimization based on the narrowband model

Another very common approach is the independent optimization in each frequency bin of the DFT domain which is termed narrowband BSS. The advantage of the narrowband approach is that any BSS algorithm derived for instantaneous mixtures can now be applied in each frequency bin. Several algorithms have been proposed which utilize either HOS (e.g., [CSL95, Sma98]) or SOS (e.g., [WFO93, PSV98, WP99, PS00, FP01a, RR01]).

However, a serious drawback is that the scaling and permutation ambiguity in instantaneous BSS (see Section 2.4) appears in each frequency bin independently and has to be solved before applying the demixing filters to the sensor signals.

There exist several heuristic approaches to solve the permutation problem. In [IM99, AK00, Ane01] inter-frequency correlations of the estimated source signals are exploited to reorder the output channels. If the geometry of the sensor array is known and if the sensor spacing is small enough to avoid spatial aliasing (see Section 2.2.5), then it is possible to calculate the attenuation by the demixing filters for incident signals arriving from different angles. Plotting the attenuation over the angle yields a directivity pattern for each output channel and for each frequency bin. By comparing the minima of the directivity patterns it is possible to determine the direction of arrival (DOA) of the sources and thus, a reordering of the output channels is possible [KSK$^+$00, IM02]. In [SMAM04] a closed-form formula for calculating the DOA has been presented, leading to reduced computational complexity. It should be noted that for estimating the DOA it has been assumed that the sources are mixed in a free-field environment. In [SMAM05] these methods have been extended to avoid the necessity of the array geometry. However, the geometry has still to be constrained to avoid spatial aliasing [KBA06] or other extensions have to be incorporated to resolve the DOA ambiguity for the frequencies where spatial aliasing occurs [SAMM06].

The scaling ambiguity in each frequency bin can be removed by application of a constraint which reduces the distortion introduced by the demixing system of the BSS algorithm. This is done either by back-projecting the estimated sources to the sensors or by introducing a constrained optimization scheme in each frequency bin [IM99, MN01].

Another effect, which has to be accounted for, is that the circular convolution is only an approximation of the linear convolution. This can be mitigated by choosing the DFT length much larger than the demixing filter length (e.g., [Sma98, PS00]) or by applying spectral smoothing as in [SMdlKdR$^+$03].

Recently, several of these repair mechanisms have been combined to obtain a robust narrowband algorithm [SMAM04] which has also been used to separate several acoustic signals in reverberant environments [MSAM05].

3.3 Generic time-domain optimization criterion and algorithmic framework

In this section a time-domain optimization criterion is presented which is based on the introduction of a compact matrix notation. Due to the optimization in the time domain, it is inherently based on the broadband signal model. The proposed optimization criterion takes inherently all three fundamental signal properties, i.e., nonstationarity, nonwhiteness, and nongaussianity as discussed in Section 2.3, into account. So far, in literature at most two of these properties have been combined, e.g., nonstationarity and nonwhiteness by using second-order statistics (e.g., [KJ00, BAK03b]) or nonwhiteness and nongaussianity by introducing time-delayed decorrelation as a preprocessing step for an algorithm based on higher order statistics [LZOS98]. After introducing the optimization criterion, we derive the gradient and the natural gradient update equations based on multivariate probability density functions. After the derivation, several aspects which are important for the implementation of these algorithms will be addressed. First, different estimation techniques for the cross-relation matrices appearing in the updates will be discussed. Second, it will be shown that these update equations require a so-called "$\mathcal{S}$ylvester $\mathcal{C}$onstraint ($\mathcal{SC}$)" for which efficient implementations and the resulting appropriate initializations will be discussed. Then, several approximations are discussed which lead to efficient novel algorithms based on higher-order and second-order statistics. Moreover, this allows to derive several links to well-known BSS and MCBD algorithms following as special cases from the previous update equations. In the end of this section several regularization strategies are proposed to allow for a robust adaptation behavior even in adverse environments.

3.3.1 Matrix formulation

From the convolutive MIMO model illustrated in Fig. 2.1 in Section 2.2, it can be seen that the output signals $y_q(n)$ are obtained by convolving the input signals $x_p(n)$ with the demixing filter coefficients $w_{pq,\kappa}$, $\kappa = 0, \ldots, L-1$. To derive an algorithm which utilizes the nonwhiteness property of the source signals by taking into account $D-1$ time-lags, a memory containing the current and the last $D-1$ output signal values $y_q(n), \ldots, y_q(n-D+1)$ has to be introduced. The linear convolution yielding the D output signal values can be formulated using matrix-vector notation as

$$\mathbf{y}_q(n) = \sum_{p=1}^{P} \mathbf{W}_{pq}^{\mathrm{T}} \mathbf{x}_p(n), \tag{3.29}$$

with the column vectors $\mathbf{x}_p$ and $\mathbf{y}_q$ given as[2]

$$\mathbf{x}_p(n) = [x_p(n), \ldots, x_p(n-2L+1)]^{\mathrm{T}}, \tag{3.30}$$
$$\mathbf{y}_q(n) = [y_q(n), \ldots, y_q(n-D+1)]^{\mathrm{T}}. \tag{3.31}$$

Analogously to Section 3.1.1, the $2L \times D$ matrix $\mathbf{W}_{pq}$ has to exhibit a Sylvester structure that contains all L coefficients of the respective demixing filter in each column, which is needed for the matrix formulation of the linear convolution:

$$\mathbf{W}_{pq} = \begin{bmatrix} w_{pq,0} & 0 & \cdots & 0 \\ w_{pq,1} & w_{pq,0} & \ddots & \vdots \\ \vdots & w_{pq,1} & \ddots & 0 \\ w_{pq,L-1} & \vdots & \ddots & w_{pq,0} \\ 0 & w_{pq,L-1} & \ddots & w_{pq,1} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & w_{pq,L-1} \\ 0 & \cdots & 0 & 0 \\ \vdots & \cdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix}. \tag{3.32}$$

It can be seen that for the general case, $1 \leq D \leq L$, the last $L - D + 1$ rows of $\mathbf{W}_{pq}$ are padded with zeros to ensure compatibility with the length of $\mathbf{x}_p(n)$ with regard to a concise DFT-domain formulation in Section 3.4. Finally, to allow a convenient notation we combine all channels and thus, we can write (3.29) compactly as

$$\mathbf{y}(n) = \mathbf{W}^{\mathrm{T}}\mathbf{x}(n), \tag{3.33}$$

with

$$\mathbf{x}(n) = [\mathbf{x}_1^{\mathrm{T}}(n), \ldots, \mathbf{x}_P^{\mathrm{T}}(n)]^{\mathrm{T}}, \tag{3.34}$$
$$\mathbf{y}(n) = [\mathbf{y}_1^{\mathrm{T}}(n), \ldots, \mathbf{y}_P^{\mathrm{T}}(n)]^{\mathrm{T}}, \tag{3.35}$$
$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{11} & \cdots & \mathbf{W}_{1P} \\ \vdots & \ddots & \vdots \\ \mathbf{W}_{P1} & \cdots & \mathbf{W}_{PP} \end{bmatrix}, \tag{3.36}$$

with $\mathbf{W}$ exhibiting a block-Sylvester structure.

Analogously, we can describe the linear convolution, which models the mixing process, by a matrix-vector product. The sensor signal column vector $\mathbf{x}_p(n)$ of length $2L$ in the

[2]With respect to a concise DFT-domain formulation in Section 3.4 the vector $\mathbf{x}_p$ contains $2L$ sensor signal samples instead of the $L + D - 1$ samples required for the linear convolution ($1 \leq D \leq L$).

p-th channel can thus be expressed as

$$\mathbf{x}_p(n) = \sum_{q=1}^{P} \mathbf{H}_{qp,2L}^{\mathrm{T}} \mathbf{s}_q(n). \tag{3.37}$$

The source signal column vector containing $M + 2L - 1$ samples, necessary for the linear convolution with the mixing FIR filters of length M, is given as

$$\mathbf{s}_q(mL + j) = [s_q(mL + j), \ldots, s_q(mL + j - 2L - M + 2)]^{\mathrm{T}}, \tag{3.38}$$

and $\mathbf{H}_{qp,2L}$ is the $M+2L-1 \times 2L$ mixing matrix in Sylvester structure as defined in (3.8). The linearity condition of the convolution determines the width of the Sylvester matrix $\mathbf{H}_{qp,2L}$ which is indicated by the subscript $2L$. A combination for all sensor channels $p = 1, \ldots, P$ leads to

$$\mathbf{x}(n) = \mathbf{H}_{2L}^{\mathrm{T}} \mathbf{s}(n), \tag{3.39}$$

with

$$\mathbf{s}(n) = [\mathbf{s}_1^{\mathrm{T}}(n), \ldots, \mathbf{s}_P^{\mathrm{T}}(n)]^{\mathrm{T}}, \tag{3.40}$$

and the block-Sylvester matrix $\mathbf{H}_{2L}$ defined in (3.9).

The output signals $\mathbf{y}(n)$ can then be expressed as a concatenation of mixing and demixing system

$$\begin{aligned} \mathbf{y}(n) &= \mathbf{W}^{\mathrm{T}} \mathbf{H}_{2L}^{\mathrm{T}} \mathbf{s}(n) \\ &= \mathbf{C}^{\mathrm{T}} \mathbf{s}(n), \end{aligned} \tag{3.41}$$

where $\mathbf{C} = \mathbf{H}_{2L}\mathbf{W}$ denotes the $P(M+2L-1) \times PD$ overall system matrix. The channel-wise submatrices of $\mathbf{C}$ exhibit a Sylvester structure containing in each column the overall system FIR filters of length $M + L - 1$ which are given for the path from the q-th source to the r-th output as $\mathbf{c}_{qr} = [c_{qr,0}, \ldots, c_{qr,M+L-1}]^{\mathrm{T}}$ and which were already defined in (3.5).

The block-Sylvester matrix $\mathbf{C}$ is the Sylvester matrix counterpart to the overall system matrix $\check{\mathbf{C}}$ originally introduced in (3.6) in Section 3.1.1 which contains only the FIR filter taps without any Sylvester structure.

3.3.2 Optimization criterion

As pointed out before, we aim at an optimization criterion simultaneously exploiting the three signal properties nonstationarity, nonwhiteness, and nongaussianity (see Section 2.3).

To exploit the nonwhiteness property, a generic SOS algorithm for convolutive mixtures has been derived in [BAK03b, BAK05a] from an optimization criterion that explicitly contains correlation matrices that include several time-lags. The nonstationarity of the sources can be utilized by considering several short-time correlation matrices at

different time instants. Additionally, for exploiting nongaussianity, higher-order statistics are required. Higher-order approaches for BSS can be divided into three classes [HKO01]: maximum likelihood (ML) estimation [Car98], minimization of the mutual information (MMI) among the output signals [YA97], and maximization of the entropy (ME/"infomax") [BS95]. Although all of these HOS approaches lead to similar update rules, it was shown in [YA97] that MMI can be regarded as the most general one. The mutual information $\mathcal{I}$ is defined for P random processes $Y_1, Y_2, \ldots, Y_P$ as [CT91]

$$\mathcal{I}(Y_1, Y_2, \ldots, Y_P) = \mathrm{E}\left\{\log \frac{p_{y,P}([y_1, \ldots, y_P])}{\prod_{q=1}^{P} p_{y_q,1}(y_q)}\right\}, \tag{3.42}$$

where $p_{y,P}$ is the P-dimensional joint probability density function (pdf), $p_{y_q,1}$ denotes the univariate marginal pdfs, and $\mathrm{E}\{\cdot\}$ is the expectation operator. In acoustics we are dealing with signals exhibiting temporal dependencies. In instantaneous BSS algorithms and also early convolutive BSS algorithms the mutual information (3.42) was used as an optimization criterion. However, acoustic signals are in general temporally dependent so that for a separation only the mutual information between the output channels should be minimized without forcing the output signals to become also temporally independent. This requires a generalization of the mutual information which allows also for vectors, i.e., for multivariate pdfs in the denominator of (3.42). This generalization allows us to define the following optimization criterion which is termed "**TRI**ple-**N**-**I**ndependent component analysis for **CON**volutive mixtures" (**TRINICON**) [BAK03a] as it simultaneously accounts for the three fundamental properties **N**onwhiteness, **N**onstationarity, and **N**ongaussianity:

$$\begin{aligned}\mathcal{J}(m, \mathbf{W}) &= \sum_{i=0}^{\infty} \beta(i,m) \frac{1}{N} \sum_{j=0}^{N-1} \left\{\log \frac{\hat{p}_{y,PD}(\mathbf{y}(iL+j))}{\prod_{q=1}^{P} \hat{p}_{y_q,D}(\mathbf{y}_q(iL+j))}\right\} \\ &= \sum_{i=0}^{\infty} \beta(i,m) \tilde{\mathcal{J}}(i, \mathbf{W}),\end{aligned} \tag{3.43}$$

Instead of the true pdfs as used in the definition of the mutual information, the TRINICON optimization criterion is based on the estimates of the true pdfs. The variable $\hat{p}_{y_q,D}(\cdot)$ is the estimate or model of the multivariate probability density function (pdf) for channel q of dimension D and $\hat{p}_{y,PD}(\cdot)$ is the estimated joint pdf of dimension PD over all channels. The usage of pdfs allows to exploit the *nongaussianity* of the signals. Furthermore, the multivariate structure of the pdfs, which is given by the memory length D, i.e., the number of time-lags, models the *nonwhiteness* of the P signals with D chosen to $1 \leq D \leq L$. As we can observe at each BSS output only one realisation of a random process we also replaced the expectation operator of the mutual information (3.42) by a

short-time average using N time instants. For the algorithms treated in this thesis the averaging has to be done in general for $N > PD$ time instants as will be discussed in more detail in the respective sections. The block indices i, m refer to the blocks which are underlying to the statistical estimation of the multivariate pdfs. For each output signal block $\mathbf{y}_q$ containing D samples a sensor signal block of length $2L$ is required according to (3.34). The *nonstationarity* is taken into account by a weighting function $\beta(i, m)$ with the block indices i, m and with finite support. The weighting function is normalized according to $\sum_{i=0}^{\infty} \beta(i, m) = 1$, and allows offline, online, and block-online implementations of the algorithms. A detailed discussion of $\beta(i, m)$ will be given in Section 3.5. As an example, $\beta(i, m) = (1 - \lambda)\lambda^{m-i}$ for $0 \leq i \leq m$, and $\beta(i, m) = 0$ else, leads to an efficient online version allowing for tracking in time-variant environments [ABAM03].

The approach followed here is carried out with overlapping data blocks as the sensor signal blocks of length $2L$ are shifted only by L samples due to the time index iL in (3.43). Analogously to supervised block-based adaptive filtering [MAG95, BBK03], this increases the convergence rate and reduces the signal delay. If further overlapping is desired, then the time index iL in (3.43) is simply replaced by $i\frac{L}{\alpha}$. The overlap factor α with $1 \leq \alpha \leq L$ should be chosen suitably to obtain integer values for the time index. For clarity, we will omit the overlap factor and will point to it when necessary.

It should be noted that the optimization criterion (3.43) can be interpreted as the Kullback-Leibler divergence [Kul59, CT91] between the joint density of the output signals $\hat{p}_{y,PD}(\mathbf{y}(iL + j))$ and a source model pdf which, in the case of the BSS optimization criterion (3.43), is based on the assumption of mutually independent source signals and is therefore factorized with respect to the different sources. This point of view allowed a generalization of the TRINICON optimization criterion to different source models by replacing the denominator of (3.43) with a desired source model pdf $\hat{p}_{s,PD}(\mathbf{y}(iL + j))$ leading to [BAK04b]

$$\mathcal{J}_{\text{gen}}(m, \mathbf{W}) = \sum_{i=0}^{\infty} \beta(i, m) \frac{1}{N} \sum_{j=0}^{N-1} \left\{ \log \frac{\hat{p}_{y,PD}(\mathbf{y}(iL + j))}{\hat{p}_{s,PD}(\mathbf{y}(iL + j))} \right\}, \tag{3.44}$$

The BSS optimization criterion (3.43) is obtained from (3.44) by using the output channel pdf factorized with respect to the channels as the source model pdf

$$\hat{p}_{s,PD}(\mathbf{y}(iL + j)) \overset{\text{(BSS)}}{=} \prod_{q=1}^{P} \hat{p}_{y_q,D}(\mathbf{y}_q(iL + j)). \tag{3.45}$$

By additionally factorizing the output channel pdfs with respect to temporal dependencies

$$\hat{p}_{s,PD}(\mathbf{y}(iL + j)) \overset{\text{(MCBD)}}{=} \prod_{q=1}^{P} \prod_{d=1}^{D} \hat{p}_{y_q,1}(y_q(iL - d + j)), \tag{3.46}$$

we obtain a source model pdf which allows to cover also multi-channel blind deconvolution (MCBD) algorithms by this framework [BAK04b, ABK05]. MCBD algorithms have also been applied to BSS of acoustic signals in the literature (see Section 3.2) and thus, the relationships between the algorithms derived in this thesis and the approaches in literature will be discussed in detail in Section 3.3.7.

Moreover, it has been shown in [BAK04b, Buc] that also a partial factorization of $\hat{p}_{s,PD}(\mathbf{y}(iL+j))$ with respect to temporal dependencies is possible which was denoted as multi-channel blind partial deconvolution (MCBPD). This allows, e.g., for speech signals, to distinguish between the temporal correlation of the source signal introduced, e.g., by the vocal tract and the correlation due to the reverberation. This distinction is important for the design of dereverberation algorithms which should only minimize the influence of the room acoustics without affecting the quality of the audio signals. Dereverberation algorithms based on TRINICON are treated in detail in [Buc] and will not be covered in this thesis.

3.3.3 Gradient of the optimization criterion

In this thesis, algorithms based on first-order gradients are considered. The derivation of the gradient with respect to the demixing filter weights $w_{pq,\kappa}$ for $p,q = 1,\ldots,P$ and $\kappa = 0,\ldots,L-1$ can either be expressed elementwise for all P^2L elements or compactly in matrix notation. For the latter case we can use the notation already introduced in Section 3.1.1 where the demixing filter taps have been combined to a column vector $\mathbf{w}_{pq} = [w_{pq,0},\ldots,w_{pq,L-1}]^{\mathrm{T}}$ in (3.2). Furthermore, a combination of all channels led to the introduction of the $PL \times P$ matrix $\check{\mathbf{W}}$ in (3.4). Using the matrix $\check{\mathbf{W}}$ which contains all demixing filter elements, the gradient with respect to the demixing filter coefficients can be expressed in matrix notation as

$$\nabla_{\check{\mathbf{W}}}\tilde{\mathcal{J}}(m,\mathbf{W}) = \frac{\partial\tilde{\mathcal{J}}(m,\mathbf{W})}{\partial\check{\mathbf{W}}}. \tag{3.47}$$

In order to calculate the gradient (3.47), the TRINICON optimization criterion $\tilde{\mathcal{J}}(m,\mathbf{W})$ given in (3.43) has to be expressed in terms of the demixing filter coefficients $w_{pq,\kappa}$. This can be done by inserting the definition of the linear convolution $\mathbf{y} = \mathbf{W}^{\mathrm{T}}\mathbf{x}$ given in (3.33) into $\tilde{\mathcal{J}}(m,\mathbf{W})$ and subsequently transforming the output signal pdf $\hat{p}_{y,PD}(\mathbf{y}(iL+j))$ into the PD-dimensional input signal pdf $\hat{p}_{x,PD}(\cdot)$ using the Sylvester matrix $\mathbf{W}$, which is considered as a mapping matrix for this linear transformation [Pap02]. This is shown in detail in the Appendix B.2.1 and leads to an expression of the optimization criterion (3.43) with respect to the Sylvester matrix $\mathbf{W}$. To be able to take the derivative with respect to $\check{\mathbf{W}}$ instead of the Sylvester matrix $\mathbf{W}$, the chain rule for the derivative of a scalar function with respect to a matrix [Har97, PP06] is applied to the gradient (3.47) containing the derivative with respect to $\check{\mathbf{W}}$. Thus, (3.47) can be split up into the concatenation of the

gradient $\nabla_{\mathbf{W}}\tilde{\mathcal{J}}(m,\mathbf{W}) = \frac{\partial\tilde{\mathcal{J}}(m,\mathbf{W})}{\partial\mathbf{W}}$ with respect to the Sylvester matrix $\mathbf{W}$ and the partial derivative $\frac{\partial\mathbf{W}}{\partial\check{\mathbf{W}}}$. The resulting gradient $\nabla_{\mathbf{W}}\tilde{\mathcal{J}}(m,\mathbf{W})$ is a matrix of dimensions $2PL \times PD$ whereas the partial derivative $\frac{\partial\mathbf{W}}{\partial\check{\mathbf{W}}}$ results in a fourth order tensor. Thus, the application of the chain rule to (3.47) cannot be expressed in matrix notation but is given elementwise as

$$\frac{\partial\tilde{\mathcal{J}}(m,\mathbf{W})}{\partial\left[\mathbf{w}_{pq}\right]_g} = \sum_{k,j,r,s} \frac{\partial\tilde{\mathcal{J}}(m,\mathbf{W})}{\partial\left[\mathbf{W}_{rs}\right]_{k,j}} \cdot \frac{\partial\left[\mathbf{W}_{rs}\right]_{k,j}}{\partial\left[\mathbf{w}_{pq}\right]_g}. \tag{3.48}$$

In this formulation, the indices of boldface variables denote the channel-selective submatrices or vectors, and the indices $[\cdot]_{k,j}$ and $[\cdot]_g$ denote the elements of the respective submatrix or vector. Therefore, $[\mathbf{w}_{pq}]_g$ is the g-th element of the pq-th column vector in matrix $\check{\mathbf{W}}$, $g \in \{1,\ldots,L\}$, $p,q \in \{1,\ldots,P\}$ and $[\mathbf{W}_{rs}]_{k,j}$ is the k,j-th element of the rs-th Sylvester submatrix of $\mathbf{W}$, $k \in \{1,\ldots,2L\}$, $j \in \{1,\ldots,D\}$, $r,s \in \{1,\ldots,P\}$. From (3.48) it can be seen that first the gradient $\nabla_{\mathbf{W}}\tilde{\mathcal{J}}(m,\mathbf{W})$, i.e., the derivative of the TRINICON optimization criterion $\tilde{\mathcal{J}}(m,\mathbf{W})$ with respect to the elements $[\mathbf{W}_{rs}]_{k,j}$ of the Sylvester matrix $\mathbf{W}_{rs}$ is calculated. Subsequently, the derivative of the Sylvester matrix element $[\mathbf{W}_{rs}]_{k,j}$ with respect to the element $[\mathbf{w}_{pq}]_g$ of matrix $\check{\mathbf{W}}$ has to be calculated. By introducing the Kronecker delta defined as

$$\delta_{i,j} = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases} \tag{3.49}$$

the second term can be simplified due to the Sylvester structure of $\mathbf{W}_{rs}$ leading to

$$\begin{aligned} \frac{\partial\tilde{\mathcal{J}}(m,\mathbf{W})}{\partial\left[\mathbf{w}_{pq}\right]_g} &= \sum_{k,j,r,s} \frac{\partial\tilde{\mathcal{J}}(m,\mathbf{W})}{\partial\left[\mathbf{W}_{rs}\right]_{k,j}} \delta_{p,r}\delta_{q,s}\delta_{k,g+j-1} \\ &= \sum_{k,j} \frac{\partial\tilde{\mathcal{J}}(m,\mathbf{W})}{\partial\left[\mathbf{W}_{pq}\right]_{k,j}} \delta_{k,g+j-1}. \end{aligned} \tag{3.50}$$

Now the derivative (3.50) with respect to $[\mathbf{w}_{pq}]_g$ can be written for each element $g = 1,\ldots,L$ as

$$\begin{aligned} \frac{\partial\tilde{\mathcal{J}}(m,\mathbf{W})}{\partial\left[\mathbf{w}_{pq}\right]_1} &= \sum_{j=1}^{D} \frac{\partial\tilde{\mathcal{J}}(m,\mathbf{W})}{\partial\left[\mathbf{W}_{pq}\right]_{j,j}} \\ \frac{\partial\tilde{\mathcal{J}}(m,\mathbf{W})}{\partial\left[\mathbf{w}_{pq}\right]_2} &= \sum_{j=1}^{D} \frac{\partial\tilde{\mathcal{J}}(m,\mathbf{W})}{\partial\left[\mathbf{W}_{pq}\right]_{j+1,j}} \\ &\vdots \\ \frac{\partial\tilde{\mathcal{J}}(m,\mathbf{W})}{\partial\left[\mathbf{w}_{pq}\right]_L} &= \sum_{j=1}^{D} \frac{\partial\tilde{\mathcal{J}}(m,\mathbf{W})}{\partial\left[\mathbf{W}_{pq}\right]_{j+L-1,j}} \end{aligned} \tag{3.51}$$

Combining this elementwise description of the gradient with the compact matrix notation $\nabla_{\check{\mathbf{W}}}\tilde{\mathcal{J}}(m,\mathbf{W})$ is possible by introducing a new operator termed $\mathcal{S}$ylvester $\mathcal{C}$onstraint ($\mathcal{SC}$) [BAK07] which relates the gradient with respect to $\check{\mathbf{W}}$ and with respect to $\mathbf{W}$ as

$$\nabla_{\check{\mathbf{W}}}\tilde{\mathcal{J}}(m,\mathbf{W}) = \mathcal{SC}\left\{\nabla_{\mathbf{W}}\tilde{\mathcal{J}}(m,\mathbf{W})\right\}. \tag{3.52}$$

Thus, (3.52) represents the formulation of (3.51) in matrix notation and is illustrated for the pq-th submatrix of $\nabla_{\mathbf{W}}\tilde{\mathcal{J}}(m,\mathbf{W})$ in Fig. 3.2. It can be seen that the Sylvester

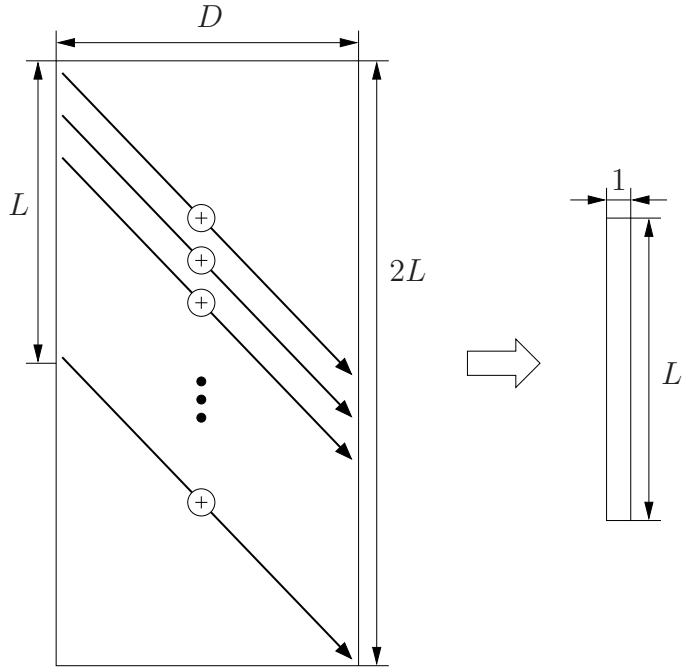


Figure 3.2: Illustration of the Sylvester constraint ($\mathcal{SC}$) for the gradient $\nabla_{\mathbf{W}_{pq}}\tilde{\mathcal{J}}(m,\mathbf{W})$ with respect to the pq-th submatrix $\mathbf{W}_{pq}$.

constraint defined by (3.51) corresponds (up to a scaling by the constant factor D) to an arithmetic average over the elements on each diagonal of the $2L \times D$ submatrices of the gradient $\nabla_{\mathbf{W}}\tilde{\mathcal{J}}(m,\mathbf{W})$. Thus, the $2PL \times PD$ gradient $\nabla_{\mathbf{W}}\tilde{\mathcal{J}}(m,\mathbf{W})$ will be reduced to the $PL \times P$ gradient $\nabla_{\check{\mathbf{W}}}\tilde{\mathcal{J}}(m,\mathbf{W})$. Later on in Section 3.3.6 efficient approximated versions of the Sylvester constraint $\mathcal{SC}$ will be given. Taking a look at the right-hand side of (3.52), we see that by using the Sylvester constraint operator we simplified the task as now the derivative of the TRINICON optimization criterion $\tilde{\mathcal{J}}(m,\mathbf{W})$ with respect to the Sylvester matrix $\mathbf{W}$ instead of $\check{\mathbf{W}}$ has to be taken. This is easier as the optimization criterion $\tilde{\mathcal{J}}(m,\mathbf{W})$ can be expressed in terms of the Sylvester matrix $\mathbf{W}$ by transforming the output signal pdf $\hat{p}_{y,PD}(\mathbf{y}(iL+j))$ into the PD-dimensional input signal pdf (see Appendix B.2.1). Thus, we are now able to calculate the derivative $\nabla_{\mathbf{W}}\tilde{\mathcal{J}}(m,\mathbf{W})$ at the right-hand side of (3.52) as shown in Appendix B.2.2. With the weighting function $\beta(i,m)$ introduced in the optimization criterion (3.43) and using (3.52) we can write the gradient

with respect to the demixing filters $\check{\mathbf{W}}$ as

$$\begin{aligned}\nabla_{\check{\mathbf{W}}}\mathcal{J}(m,\mathbf{W}) &= \mathcal{SC}\left\{\nabla_{\mathbf{W}}\mathcal{J}(m,\mathbf{W})\right\} \\ &= \mathcal{SC}\left\{\sum_{i=0}^{\infty}\beta(i,m)\nabla_{\mathbf{W}}\tilde{\mathcal{J}}(m,\mathbf{W})\right\} \end{aligned} \tag{3.53}$$

Inserting the result for $\nabla_{\mathbf{W}}\tilde{\mathcal{J}}(m,\mathbf{W})$ from Appendix B.2.2 in (3.53) finally yields the *generic HOS gradient* utilizing all three signal properties

$$\begin{aligned}\nabla_{\check{\mathbf{W}}}\mathcal{J}(m,\mathbf{W}) = \mathcal{SC}\Bigg\{\sum_{i=0}^{\infty}\beta(i,m)\frac{1}{N}\sum_{j=0}^{N-1}\Big\{&\mathbf{x}(iL+j)\boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{y}(iL+j)) \\ &-\mathbf{W}_{2PL\times PD}^{1_D0}\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}_{2PL\times PD}^{1_D0}\right)^{-1}\Big\}\Bigg\}, \end{aligned} \tag{3.54}$$

with the general weighting function $\beta(i,m)$ discussed in Section 3.5 and the *multivariate score function* $\boldsymbol{\Phi}(\mathbf{y})$ consisting of the stacked channel-wise multivariate score functions $\boldsymbol{\Phi}_q(\mathbf{y}_q)$, $q=1,\ldots,P$ defined as

$$\begin{aligned}\boldsymbol{\Phi}(\mathbf{y}(iL+j)) &= \left[\left(-\frac{\frac{\partial\hat{p}_{y_1,D}(\mathbf{y}_1(iL+j))}{\partial\mathbf{y}_1(iL+j)}}{\hat{p}_{y_1,D}(\mathbf{y}_1(iL+j))}\right)^{\mathrm{T}},\ldots,\left(-\frac{\frac{\partial\hat{p}_{y_P,D}(\mathbf{y}_P(iL+j))}{\partial\mathbf{y}_P(iL+j)}}{\hat{p}_{y_P,D}(\mathbf{y}_P(iL+j))}\right)^{\mathrm{T}}\right]^{\mathrm{T}} \\ &:= \left[\boldsymbol{\Phi}_1^{\mathrm{T}}(\mathbf{y}_1(iL+j)),\ldots,\boldsymbol{\Phi}_P^{\mathrm{T}}(\mathbf{y}_P(iL+j))\right]^{\mathrm{T}}. \end{aligned} \tag{3.55}$$

The $2PL\times PD$ window matrix $\mathbf{W}_{2PL\times PD}^{1_D0}$ is defined as

$$\mathbf{W}_{2PL\times PD}^{1_D0} = \mathrm{Bdiag}\left\{\mathbf{W}_{2L\times D}^{1_D0},\ldots,\mathbf{W}_{2L\times D}^{1_D0}\right\}, \tag{3.56}$$

$$\mathbf{W}_{2L\times D}^{1_D0} = \left[\mathbf{I}_{D\times D},\mathbf{0}_{D\times(2L-D)}\right]^{\mathrm{T}}. \tag{3.57}$$

The operator $\mathrm{Bdiag}\{\mathbf{A}_1,\ldots,\mathbf{A}_P\}$ denotes a block-diagonal matrix with submatrices $\mathbf{A}_1,\ldots,\mathbf{A}_P$ on its diagonal. For the description of window matrices (also appearing in the DFT-domain algorithms in Section 3.4) we use the following conventions:

- The lower index of a matrix denotes its dimensions.
- P-channel matrices (as indicated by the size in the lower index) are partitioned into P single-channel window matrices.
- The upper index describes the positions of ones and zeros. Unity submatrices are always located at the upper left ('10') or lower right ('01') corners of the respective single-channel window matrix. The size of these clusters is indicated in subscript (e.g., '1_D0').

The window matrix $\mathbf{W}_{2PL\times PD}^{1_D 0}$ appears due to the transformation of pdfs by the non-square Sylvester matrix $\mathbf{W}$ (see Appendix B.2.1).

With an iterative optimization procedure, the current demixing matrix is obtained by the recursive update equation

$$\check{\mathbf{W}}(m) = \check{\mathbf{W}}(m-1) - \mu\Delta\check{\mathbf{W}}(m), \tag{3.58}$$

where μ is a stepsize parameter, and $\Delta\check{\mathbf{W}}(m)$ is the update which is set equal to $\nabla_{\check{\mathbf{W}}}\mathcal{J}(m,\mathbf{W})$ for gradient descent adaptation. Due to the adaptation process, the coefficient matrix becomes time-variant.

Moreover, it should be noted that an extension to higher-order gradients is possible as shown in [Buc] and leads, e.g., for the second-order gradient to the appearance of fourth-order tensors in the coefficient update. In [Buc] it is shown that efficient algorithms can again be obtained by using similar approximations as shown later in Section 3.4.

3.3.4 Equivariance property and natural gradient update

It is known that stochastic gradient descent, i.e., $\Delta\check{\mathbf{W}}(m) = \nabla_{\check{\mathbf{W}}}\mathcal{J}(m,\mathbf{W})$ suffers from slow convergence in many practical problems due to statistical dependencies in the data being processed. In the BSS application the gradient and thus, the separation performance depends on the MIMO mixing system. This can be observed by considering the gradient $\nabla_{\mathbf{W}}\mathcal{J}(m,\mathbf{W})$ with respect to the Sylvester matrix $\mathbf{W}$ which is related to the gradient $\nabla_{\check{\mathbf{W}}}\mathcal{J}(m,\mathbf{W})$ by (3.53). The update of the overall system matrix $\mathbf{C}$ is then given as $\Delta\mathbf{C}(m) = \mathbf{H}_{2L}\nabla_{\mathbf{W}}\mathcal{J}(m,\mathbf{W})$. To see how the mixing system $\mathbf{H}_{2L}$ influences this update we pre-multiply $\Delta\mathbf{W}(m)$ by the Sylvester matrix $\mathbf{H}_{2L}$ leading to

$$\begin{aligned}\Delta\mathbf{C}(m) \;=\; \mathbf{H}_{2L}\sum_{i=0}^{\infty}\beta(i,m)\frac{1}{N}\sum_{j=0}^{N-1}\Big\{&\mathbf{x}(iL+j)\boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{y}(iL+j)) \\ &-\mathbf{W}_{2PL\times PD}^{1_D 0}\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}_{2PL\times PD}^{1_D 0}\right)^{-1}\Big\}.\end{aligned} \tag{3.59}$$

This way it can be easily seen that $\Delta\mathbf{C}(m)$ depends on the mixing system $\mathbf{H}_{2L}$ and therefore, on its conditioning.

Fortunately, a modification of the ordinary gradient, termed the *natural gradient* by Amari [Ama98] and the *relative gradient* by Cardoso [CL96, Car98] (which is equivalent to the natural gradient in the BSS application) has been developed that largely removes all effects of an ill-conditioned mixing matrix $\mathbf{H}_{2L}$, assuming an appropriate initialization of $\mathbf{W}$. The idea of the relative gradient is based on the equivariance property. Generally speaking, an estimator behaves equivariantly if it produces estimates that, under data transformation, are transformed in the same way as the data [CL96]. In the context of BSS the key property of equivariant estimators is that they exhibit uniform performance,

e.g., in terms of bias and variance, independently of the mixing system $\mathbf{H}_{2L}$. In [BAK03b, BAK05a] the natural/relative gradient has been extended to the case of Sylvester matrices $\mathbf{W}$ yielding

$$\nabla_{\mathbf{W}}^{\mathrm{NG}} \mathcal{J}(m, \mathbf{W}) = \sum_{i=0}^{\infty} \beta(i,m) \mathbf{W}(i) \mathbf{W}^{\mathrm{T}}(i) \nabla_{\mathbf{W}} \tilde{\mathcal{J}}(i, \mathbf{W}). \tag{3.60}$$

As we finally want to obtain the natural gradient with respect to the demixing filter weights $\check{\mathbf{W}}$ we have to apply the Sylvester constraint to (3.60) yielding

$$\nabla_{\check{\mathbf{W}}}^{\mathrm{NG}} \mathcal{J}(m, \mathbf{W}) = \mathcal{SC} \left\{ \sum_{i=0}^{\infty} \beta(i,m) \mathbf{W}(i) \mathbf{W}^{\mathrm{T}}(i) \nabla_{\mathbf{W}} \tilde{\mathcal{J}}(i, \mathbf{W}) \right\}. \tag{3.61}$$

Together with the expression (B.50) for $\nabla_{\mathbf{W}} \tilde{\mathcal{J}}(i, \mathbf{W})$ from Appendix B.2 this immediately leads to the following expression for the *HOS natural gradient*

$$\nabla_{\check{\mathbf{W}}}^{\mathrm{NG}} \mathcal{J}(m, \mathbf{W}) = \mathcal{SC} \left\{ \sum_{i=0}^{\infty} \beta(i,m) \mathbf{W}(i) \frac{1}{N} \sum_{j=0}^{N-1} \left\{ \mathbf{y}(iL+j) \boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{y}(iL+j)) - \mathbf{I} \right\} \right\}, \tag{3.62}$$

which is then used as update $\Delta\check{\mathbf{W}}(m)$ in (3.58).

In the derivation of the natural gradient for instantaneous mixtures, the fact that the demixing matrices form a so-called Lie group has played an important role [Ama98]. However, the block-Sylvester matrices $\mathbf{W}$ after (3.32), (3.36) do not form a Lie group (as they are generally not invertible). To see that the above formulation of the natural gradient is indeed justified, we pre-multiply the natural gradient with respect to the Sylvester matrix $\mathbf{W}$ given in (3.60) with $\mathbf{H}_{2L}$, which leads to

$$\begin{aligned} \Delta\mathbf{C}(m) &= \sum_{i=0}^{\infty} \beta(i,m) \mathbf{H}_{2L} \mathbf{W}(i) \frac{1}{N} \sum_{j=0}^{N-1} \left\{ \mathbf{y}(iL+j) \boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{y}(iL+j)) - \mathbf{I} \right\} \\ &= \sum_{i=0}^{\infty} \beta(i,m) \mathbf{C}(i) \frac{1}{N} \sum_{j=0}^{N-1} \left\{ \mathbf{y}(iL+j) \boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{y}(iL+j)) - \mathbf{I} \right\}. \end{aligned} \tag{3.63}$$

The matrix $\Delta\mathbf{C}(m)$ does not depend on the mixing matrix as it has been absorbed as an initial condition into $\mathbf{C}(0) = \mathbf{H}_{2L}\mathbf{W}(0)$. This leads to the desired uniform performance of (3.62) and proves the equivariance property of the natural gradient. Possible choices of the initial value $\mathbf{W}(0)$ will be discussed in Section 3.3.6.

Another well-known advantage of using the natural gradient is a reduction of the computational complexity of the update as the inversion of the $PD \times PD$ matrix $\mathbf{W}^{\mathrm{T}}\mathbf{W}^{1_D 0}_{2PL \times PD}$ in (3.54) does not need to be carried out in (3.62). Furthermore, it can be shown for specific pdfs (see Section 3.3.7) that instead of $N > PD$ for the gradient update the condition $N > D$ is sufficient for the averaging in the natural gradient update due to multivariate score functions $\boldsymbol{\Phi}(\mathbf{y})$, which lead only to the inversion of $D \times D$ channel-wise matrices.

The update in (3.62) represents a so-called holonomic algorithm as it imposes the constraint $\mathbf{y}(iL+j)\mathbf{\Phi}^{\mathrm{T}}(\mathbf{y}(iL+j)) = \mathbf{I}$ on the magnitudes of the recovered signals. However, when the source signals are nonstationary, these constraints may force a rapid change in the magnitude of the demixing matrix leading to numerical instabilities in some cases (see, e.g., [CA02]). By replacing $\mathbf{I}$ in (3.62) with the term $\mathrm{bdiag}\{\mathbf{y}(iL+j)\mathbf{\Phi}^{\mathrm{T}}(\mathbf{y}(iL+j))\}$ the constraint on the magnitude of the recovered signals can be avoided. This is termed the *nonholonomic natural gradient* algorithm which is given as

$$\nabla^{\mathrm{NG}}_{\check{\mathbf{W}}}\mathcal{J}(m,\mathbf{W}) = \mathcal{SC}\left\{\sum_{i=0}^{\infty}\beta(i,m)\mathbf{W}(i)\frac{1}{N}\sum_{j=0}^{N-1}\left\{\mathbf{y}(iL+j)\mathbf{\Phi}^{\mathrm{T}}(\mathbf{y}(iL+j)) - \mathrm{bdiag}\{\mathbf{y}(iL+j)\mathbf{\Phi}^{\mathrm{T}}(\mathbf{y}(iL+j))\}\right\}\right\}. \quad (3.64)$$

Here, the bdiag operator sets all channel-wise cross-terms to zero. Note that the nonholonomic property can also be directly taken into account by modifying the cost function as shown in [BAK03a].

Due to the reduced computational complexity and the improved convergence behavior, the remainder of this thesis will focus on algorithms based on the natural gradient. In particular the nonholonomic variant (3.64) will be chosen because of the nonstationary nature of acoustic signals.

3.3.5 Covariance versus correlation method

In the previous sections, the gradient update (3.54) and natural gradient updates (3.62), (3.64) of the optimization criterion have been derived. These updates exhibit short-time HOS cross-relation matrices which are estimated by time-averaging as

$$\mathbf{R}_{\mathbf{y\Phi(y)}}(i) = \frac{1}{N}\sum_{j=0}^{N-1}\mathbf{y}(iL+j)\mathbf{\Phi}^{\mathrm{T}}(\mathbf{y}(iL+j)), \quad (3.65)$$

$$\mathbf{R}_{\mathbf{x\Phi(y)}}(i) = \frac{1}{N}\sum_{j=0}^{N-1}\mathbf{x}(iL+j)\mathbf{\Phi}^{\mathrm{T}}(\mathbf{y}(iL+j)), \quad (3.66)$$

where $\mathbf{R}_{\mathbf{y\Phi(y)}}$ has dimensions $PD \times PD$ and $\mathbf{R}_{\mathbf{x\Phi(y)}}$ is a $2PL \times PD$ matrix. From linear prediction problems it is known that in principle, there are two basic methods for the block-based estimation of the SOS short-time auto- and cross-correlation matrices for nonstationary signals: the so-called *covariance method* and the *correlation method* [MG76, DHP00, VHH98, Hay02]. It should be emphasized that the terms covariance method and correlation method are not based upon the standard usage of the covariance

function as the correlation function with the means removed. In the following the two methods are first discussed for the case of SOS and then this distinction is analogously introduced for the cross-relation matrices occurring in HOS BSS algorithms.

Covariance method

Second-order statistics (SOS). As in the remainder of this thesis only natural gradient algorithms are considered, we show the difference between both estimation methods using the cross-relation matrix of the output signals $\mathbf{R}_{\mathbf{y}\boldsymbol{\Phi}(\mathbf{y})}$ defined in (3.65) and the SOS cross-correlation counterpart defined for each channel as

$$\mathbf{R}_{\mathbf{y}_p\mathbf{y}_q}(i) = \frac{1}{N}\sum_{j=0}^{N-1}\mathbf{y}_p(iL+j)\mathbf{y}_q^{\mathrm{T}}(iL+j). \tag{3.67}$$

In this thesis acoustic signals are considered which are real-valued quantities. However, with regard to the cross-spectral density matrices derived later in Section 3.4, we will formulate the estimation of the cross-relation matrices (3.73) for complex signals and thus, the transpose operator $^{\mathrm{T}}$ is replaced by the hermitian operator $^{\mathrm{H}}$ denoting transposition of the conjugate complex values. The cross-correlation matrix (3.67) can be written compactly in matrix notation

$$\mathbf{R}_{\mathbf{y}_p\mathbf{y}_q}(i) = \frac{1}{N}\mathbf{Y}_p(i)\mathbf{Y}_q^{\mathrm{H}}(i), \tag{3.68}$$

by defining the $D \times N$ Toeplitz matrix

$$\begin{aligned}\mathbf{Y}_p(i) &= [\mathbf{y}_p(iL),\ldots,\mathbf{y}_p(iL+N-1)] \\ &= \begin{bmatrix} y_p(iL) & \cdots & y_p(iL+N-1) \\ y_p(iL-1) & \ddots & y_p(iL+N-2) \\ \vdots & \ddots & \vdots \\ y_p(iL-D+1) & \cdots & y_p(iL-D+N) \end{bmatrix}.\end{aligned} \tag{3.69}$$

The definition of the cross-correlation matrix in (3.67), (3.68) describes the estimation method for the true correlation matrix $\mathrm{E}\{\mathbf{y}_p\mathbf{y}_q^{\mathrm{H}}\}$ which is commonly denoted as the covariance method (see e.g., [MG76]). In this case, the element of $\mathbf{R}_{\mathbf{y}_p\mathbf{y}_q}(i)$ in the u-th row and v-th column ($u, v \in \{0,\ldots,D-1\}$) is given as

$$r_{y_py_q}(i,u,v) = \frac{1}{N}\sum_{n=iL}^{iL+N-1} y_p(n-u)y_q^*(n-v). \tag{3.70}$$

where * denotes the conjugate complex operator. This leads in general to a cross-correlation matrix exhibiting no special structure as each element (3.70) of the matrix is computed by evaluating the signals at different time instants. Therefore the matrix

elements depend on the time-shift u and v of the respective signals. A combination of all channels leads to

$$\begin{aligned}\mathbf{R}_{\mathbf{yy}}(i) &= \frac{1}{N}\sum_{j=0}^{N-1}\mathbf{y}(iL+j)\mathbf{y}^{\mathrm{H}}(iL+j) \\ &= \frac{1}{N}\mathbf{Y}(i)\mathbf{Y}^{\mathrm{H}}(i),\end{aligned} \tag{3.71}$$

with

$$\mathbf{Y}(i) = [\mathbf{Y}_1^{\mathrm{T}}(i), \ldots, \mathbf{Y}_P^{\mathrm{T}}(i)]^{\mathrm{T}}. \tag{3.72}$$

Higher-order statistics (HOS). The covariance method which originates from linear prediction techniques also appears in the estimation of the HOS cross-relation matrices in the natural gradient algorithms considered in this thesis. The cross-relation matrix of the output signals $\mathbf{R}_{\mathbf{y}\boldsymbol{\Phi}(\mathbf{y})}$ defined in (3.65) is composed of the $D \times D$ channel-wise submatrices

$$\mathbf{R}_{\mathbf{y}_p\boldsymbol{\Phi}_q(\mathbf{y}_q)}(i) = \frac{1}{N}\sum_{j=0}^{N-1}\mathbf{y}_p(iL+j)\boldsymbol{\Phi}_q^{\mathrm{H}}(\mathbf{y}_q(iL+j)), \tag{3.73}$$

where the multivariate score function $\boldsymbol{\Phi}_q(\mathbf{y}_q(iL+j))$ of the q-th output channel is defined according to (3.55). Analogously, the cross-relation matrix (3.73) can be written compactly in matrix notation

$$\mathbf{R}_{\mathbf{y}_p\boldsymbol{\Phi}_q(\mathbf{y}_q)}(i) = \frac{1}{N}\mathbf{Y}_p(i)\mathbf{Y}_{\boldsymbol{\Phi}_q}^{\mathrm{H}}(i) \tag{3.74}$$

by using $\mathbf{Y}_p(i)$ defined in (3.69) and defining the $D \times N$ matrix

$$\begin{aligned}\mathbf{Y}_{\boldsymbol{\Phi}_q}(i) &= [\boldsymbol{\Phi}_q(\mathbf{y}_q(iL)), \ldots, \boldsymbol{\Phi}_q(\mathbf{y}_q(iL+N-1))] \\ &= \begin{bmatrix} \Phi_{q,0}(\mathbf{y}_q(iL)) & \cdots & \Phi_{q,0}(\mathbf{y}_q(iL+N-1)) \\ \Phi_{q,-1}(\mathbf{y}_q(iL)) & \ddots & \Phi_{q,-1}(\mathbf{y}_q(iL+N-1)) \\ \vdots & \ddots & \vdots \\ \Phi_{q,-D+1}(\mathbf{y}_q(iL)) & \cdots & \Phi_{q,-D+1}(\mathbf{y}_q(iL+N-1)) \end{bmatrix},\end{aligned} \tag{3.75}$$

where the elements of the multivariate score function $\boldsymbol{\Phi}_q(\mathbf{y}_q)$ for the q-th channel are given as $\boldsymbol{\Phi}_q(\mathbf{y}_q) = [\Phi_{q,0}(\mathbf{y}_q), \ldots, \Phi_{q,-D+1}(\mathbf{y}_q)]^{\mathrm{T}}$. It can be observed that the matrix $\mathbf{Y}_p(i)$ exhibits a Toeplitz structure whereas $\mathbf{Y}_{\boldsymbol{\Phi}_q}(i)$ is not necessarily Toeplitz as the multivariate score function $\boldsymbol{\Phi}(\cdot)$ is applied to each column. The definition of the cross-relation matrices in (3.73), (3.74) can be interpreted as an estimation according to the covariance method as the element in the u-th row and v-th column ($u, v \in \{0, \ldots, D-1\}$) of $\mathbf{R}_{\mathbf{y}_p\boldsymbol{\Phi}_q(\mathbf{y}_q)}(i)$ is given by

$$r_{y_p\Phi_q(\mathbf{y}_q)}(i,u,v) = \frac{1}{N}\sum_{n=iL}^{iL+N-1} y_p(n-u)\Phi_{q,-v}^*(\mathbf{y}_q(n)). \tag{3.76}$$

Analogously to the SOS case, this leads in general to a cross-relation matrix exhibiting no special structure as each element (3.76) of the matrix depends on the time-shift u of y_p and on the index v of the multivariate score function of the respective signals.

A combination of all channel-wise cross-relation matrices allows to express (3.65) conveniently in matrix notation as

$$\mathbf{R}_{\mathbf{y}\boldsymbol{\Phi}(\mathbf{y})}(i) = \frac{1}{N}\mathbf{Y}(i)\mathbf{Y}_{\boldsymbol{\Phi}}^{\mathrm{H}}(i), \tag{3.77}$$

with $\mathbf{Y}(i)$ defined in (3.72) and

$$\mathbf{Y}_{\boldsymbol{\Phi}}(i) = [\mathbf{Y}_{\boldsymbol{\Phi}_1}^{\mathrm{T}}(i), \ldots, \mathbf{Y}_{\boldsymbol{\Phi}_P}^{\mathrm{T}}(i)]^{\mathrm{T}}. \tag{3.78}$$

Correlation method

Second-order statistics (SOS). If the cross-correlation matrix is instead estimated using the correlation method then the cross-correlation elements do not depend on the individual time-shifts u, v but only on the relative time-lag $\tilde{v} = v - u$ ($\tilde{v} \in \{-D+1, \ldots, D-1\}$) of the signals $y_p(n-u)$ and $y_q(n-v)$). This leads to

$$\tilde{r}_{y_p y_q}(i, \tilde{v}) = \begin{cases} \frac{1}{N} \sum_{n=iL}^{iL+N-\tilde{v}-1} y_p(n+\tilde{v}) y_q^*(n) & \text{for } \tilde{v} \geq 0 \\ \frac{1}{N} \sum_{n=iL-\tilde{v}}^{iL+N-1} y_p(n+\tilde{v}) y_q^*(n) & \text{for } \tilde{v} < 0 \end{cases}, \tag{3.79}$$

where it can be seen that the number of summation elements decreases with increasing time-lag $\tilde{v}$. The usage of the correlation method leads to a *Toeplitz structure* of the $D \times D$ correlation matrix. To indicate the estimation by the correlation method, a tilde is used and thus, the Toeplitz correlation matrix is given as

$$\tilde{\mathbf{R}}_{\mathbf{y}_p\mathbf{y}_q}(i) = \begin{bmatrix} \tilde{r}_{y_p y_q}(i,0) & \cdots & \tilde{r}_{y_p y_q}(i,D-1) \\ \tilde{r}_{y_p y_q}(i,-1) & \ddots & \tilde{r}_{y_p y_q}(i,D-2) \\ \vdots & \ddots & \vdots \\ \tilde{r}_{y_p y_q}(i,-D+1) & \cdots & \tilde{r}_{y_p y_q}(i,0) \end{bmatrix}. \tag{3.80}$$

Analogously to the covariance method in (3.68) we can express the cross-correlation matrix (3.80) also for the correlation method as a matrix product

$$\tilde{\mathbf{R}}_{\mathbf{y}_p\mathbf{y}_q}(i) = \frac{1}{N}\tilde{\mathbf{Y}}_p(i)\tilde{\mathbf{Y}}_q^{\mathrm{H}}(i) \tag{3.81}$$

with the $D \times N + D - 1$ matrix

$$\tilde{\mathbf{Y}}_p(i) = \begin{bmatrix} y_p(iL) & \dots & y_p(iL+N-1) & 0 & \dots & 0 \\ 0 & y_p(iL) & \dots & y_p(iL+N-1) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \dots & \ddots & 0 \\ 0 & \dots & 0 & y_p(iL) & \dots & y_p(iL+N-1) \end{bmatrix}. \tag{3.82}$$

The combination of all channels leads to

$$\tilde{\mathbf{R}}_{\mathbf{yy}}(i) = \frac{1}{N}\tilde{\mathbf{Y}}(i)\tilde{\mathbf{Y}}^{\mathrm{H}}(i), \tag{3.83}$$

with

$$\tilde{\mathbf{Y}}(i) = [\tilde{\mathbf{Y}}_1^{\mathrm{T}}(i), \dots, \tilde{\mathbf{Y}}_P^{\mathrm{T}}(i)]^{\mathrm{T}}. \tag{3.84}$$

It can be seen from the definition of the matrix $\tilde{\mathbf{Y}}_p(i)$ that the correlation method only requires N signal values for estimating the cross-correlation matrix whereas the covariance method needs $N + D - 1$ values. This difference between both methods can be explained by the illustration in Fig. 3.3 where the signals used for the calculation of the element $r_{y_p y_q}(i, u, v)$ of the covariance method and $\tilde{r}_{y_p y_q}(i, \tilde{v})$ of the correlation method are depicted for $v = 0$, $u > 0$ and $\tilde{v} = v - u$, i.e., $\tilde{v} < 0$. It can be seen in Fig. 3.3a that for the

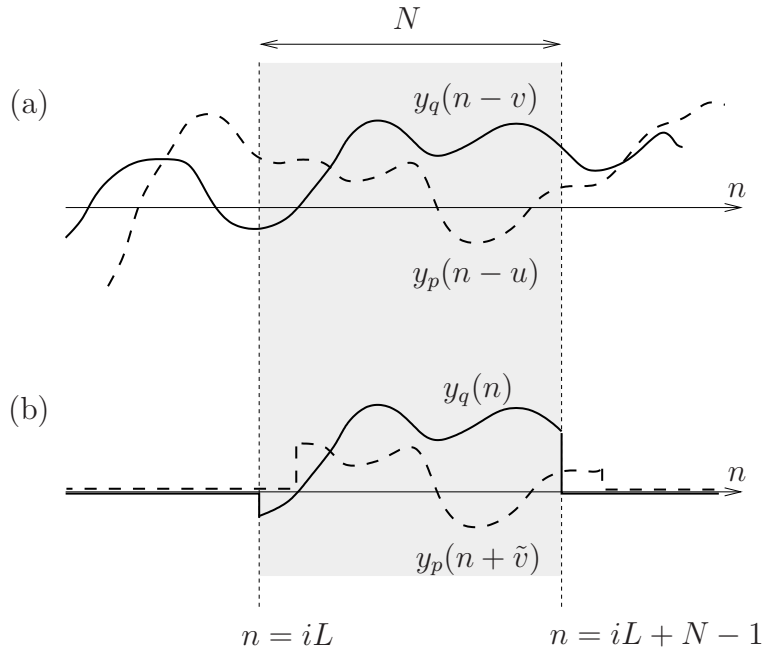


Figure 3.3: Illustration of the estimation of second-order cross-correlations by (a) covariance and (b) correlation method for $v = 0$, $u > 0$ and $\tilde{v} = v - u$, i.e., $\tilde{v} < 0$.

covariance method the signals are first shifted by u and v samples, respectively and then windowed by a window of length N. For the correlation method (Fig. 3.3b) the windowing

is applied to the signal before the relative shift of $\tilde{v}$ samples and thus, it can be seen that the number of summation elements decreases with an increasing absolute value of the relative shift $\tilde{v}$. The different windowing procedure leads to the fact that the correlation method only needs N signal values in contrast to the $N + D - 1$ for the covariance method. Moreover, it should be pointed out that the correlation method can be seen as an approximation of the more accurate covariance method which follows by assuming stationarity within each block i and compensating for the bias due to the nonuniform number of contributions to the sum. Therefore, the covariance method is sometimes also referred to as the cross-correlation function in the "nonstationary case" whereas the correlation method refers to the "stationary case" [DHP00]. Due to the Toeplitz structure of the cross-correlation matrices the correlation method exhibits a lower computational complexity and is thus used for the implementation of the SOS algorithms evaluated experimentally in Section 3.6.

Higher-order statistics (HOS). The correlation method can also be applied to the estimation of the cross-relation matrices $\mathbf{R}_{\mathbf{y}_p\mathbf{\Phi}_q(\mathbf{y}_q)}(i)$. In this case their elements do not depend on the individual indices u, v but only on the relative index $\tilde{v} = v - u$ ($\tilde{v} \in \{-D+1, \ldots, D-1\}$) of the signals $y_p(n-u)$ and $\Phi^*_{q,-v}(\mathbf{y}_q(n))$ leading to

$$\tilde{r}_{y_p\Phi_q(\mathbf{y}_q)}(i,\tilde{v}) = \begin{cases} \dfrac{1}{N}\displaystyle\sum_{n=iL}^{iL+N-\tilde{v}-1} y_p(n+\tilde{v})\Phi^*_{q,0}(\mathbf{y}_q(n)) & \text{for } \tilde{v} \geq 0 \\ \dfrac{1}{N}\displaystyle\sum_{n=iL-\tilde{v}}^{iL+N-1} y_p(n+\tilde{v})\Phi^*_{q,0}(\mathbf{y}_q(n)) & \text{for } \tilde{v} < 0 \end{cases}. \tag{3.85}$$

It can be seen that the number of summation elements decreases with increasing time-lag $\tilde{v}$. As the elements in (3.85) only depend on the relative time-delay $\tilde{v}$, this leads analogously to the SOS case to a *Toeplitz structure* of the $D \times D$ matrix $\mathbf{R}_{\mathbf{y}_p\mathbf{\Phi}_q(\mathbf{y}_q)}(i)$ which can then be expressed as

$$\tilde{\mathbf{R}}_{\mathbf{y}_p\mathbf{\Phi}_q(\mathbf{y}_q)}(i) = \begin{bmatrix} \tilde{r}_{y_p\Phi_q(\mathbf{y}_q)}(i,0) & \cdots & \tilde{r}_{y_p\Phi_q(\mathbf{y}_q)}(i,D-1) \\ \tilde{r}_{y_p\Phi_q(\mathbf{y}_q)}(i,-1) & \ddots & \tilde{r}_{y_p\Phi_q(\mathbf{y}_q)}(i,D-2) \\ \vdots & \ddots & \vdots \\ \tilde{r}_{y_p\Phi_q(\mathbf{y}_q)}(i,-D+1) & \cdots & \tilde{r}_{y_p\Phi_q(\mathbf{y}_q)}(i,0) \end{bmatrix}. \tag{3.86}$$

We can express the cross-relation matrix (3.86) as a matrix product

$$\tilde{\mathbf{R}}_{\mathbf{y}_p\mathbf{\Phi}_q(\mathbf{y}_q)}(i) = \frac{1}{N}\tilde{\mathbf{Y}}_p(i)\tilde{\mathbf{Y}}^{\mathrm{H}}_{\mathbf{\Phi}_q}(i), \tag{3.87}$$

with $\tilde{\mathbf{Y}}_p(i)$ defined in (3.82) and the $D \times N + D - 1$ matrix

$$\tilde{\mathbf{Y}}_{\mathbf{\Phi}_q}(i) = \begin{bmatrix} \Phi_{q,0}(\mathbf{y}_q(iL)) & \dots & \Phi_{q,0}(\mathbf{y}_q(iL+N-1)) & 0 & \dots & 0 \\ 0 & \ddots & \dots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \dots & \ddots & 0 \\ 0 & \dots & 0 & \Phi_{q,0}(\mathbf{y}_q(iL)) & \dots & \Phi_{q,0}(\mathbf{y}_q(iL+N-1)) \end{bmatrix}. \tag{3.88}$$

It can be seen in (3.88) that the same nonlinear weighting caused by the multivariate score function is applied to each diagonal of $\tilde{\mathbf{Y}}_{\mathbf{\Phi}_q}$. A combination of all channel-wise cross-relation matrices allows to express the estimation of $\tilde{\mathbf{R}}_{\mathbf{y}\mathbf{\Phi}(\mathbf{y})}$ via the correlation method conveniently in matrix notation as

$$\tilde{\mathbf{R}}_{\mathbf{y}\mathbf{\Phi}(\mathbf{y})}(i) = \frac{1}{N}\tilde{\mathbf{Y}}(i)\tilde{\mathbf{Y}}^{\mathrm{H}}_{\mathbf{\Phi}}(i), \tag{3.89}$$

with

$$\tilde{\mathbf{Y}}_{\mathbf{\Phi}}(i) = [\tilde{\mathbf{Y}}^{\mathrm{T}}_{\mathbf{\Phi}_1}(i), \dots, \tilde{\mathbf{Y}}^{\mathrm{T}}_{\mathbf{\Phi}_P}(i)]^{\mathrm{T}}. \tag{3.90}$$

3.3.6 Efficient Sylvester Constraint realizations and resulting initialization methods

In Section 3.3.3 the Sylvester constraint operator was introduced in (3.52) which resulted in averaging the elements on the diagonals as illustrated in Fig. 3.2. This is computationally expensive as all P^2LD elements of the natural gradient $\nabla^{\mathrm{NG}}_{\mathbf{W}}\mathcal{J}(m)$ have to be calculated in order to perform the averaging. However, it is possible to approximate the Sylvester constraint ($\mathcal{SC}$) by calculating only certain elements of the natural gradient $\nabla^{\mathrm{NG}}_{\mathbf{W}}\mathcal{J}(m)$ which neglects the averaging process but still allows us to obtain all P^2L relevant values for the filter updates. In the following two possibilities will be discussed for the nonholonomic natural gradient (3.64) which result in powerful efficient algorithms. Using the definition of the cross-relation matrix (3.65), the nonholonomic natural gradient (3.64) derived from the cost function (3.43) can be expressed as

$$\begin{aligned} \nabla^{\mathrm{NG}}_{\tilde{\mathbf{W}}}\mathcal{J}(m,\mathbf{W}) &= \sum_{i=0}^{\infty}\beta(i,m)\nabla^{\mathrm{NG}}_{\tilde{\mathbf{W}}}\tilde{\mathcal{J}}(i,\mathbf{W}) \\ &= \sum_{i=0}^{\infty}\beta(i,m)\mathcal{SC}\left\{\nabla^{\mathrm{NG}}_{\mathbf{W}}\tilde{\mathcal{J}}(i,\mathbf{W})\right\} \\ &= \sum_{i=0}^{\infty}\beta(i,m)\mathcal{SC}\left\{\mathbf{W}(i)\mathrm{boff}\left\{\mathbf{R}_{\mathbf{y}\mathbf{\Phi}(\mathbf{y})}(i)\right\}\right\}, \end{aligned} \tag{3.91}$$

where the operator $\text{boff}\{\mathbf{A}\} = \mathbf{A} - \text{bdiag}\{\mathbf{A}\}$, i.e., it sets all diagonal submatrices to zero. The Sylvester constraint is applied to the $2PL \times PD$ update $\nabla^{\text{NG}}_{\mathbf{W}}\tilde{\mathcal{J}}(i,\mathbf{W})$ which results from the matrix multiplication of the $2PL \times PD$ Sylvester matrix $\mathbf{W}(i)$ with the $PD \times PD$ block-offdiagonal matrix $\text{boff}\left\{\mathbf{R}_{\mathbf{y}\mathbf{\Phi}(\mathbf{y})}(i)\right\}$. This matrix multiplication can be expressed channel-wise and results for the pq-th update $\nabla^{\text{NG}}_{\mathbf{W}_{pq}}\tilde{\mathcal{J}}(i,\mathbf{W})$ of dimension $2L \times D$ in

$$\nabla^{\text{NG}}_{\mathbf{W}_{pq}}\tilde{\mathcal{J}}(i,\mathbf{W}) = \sum_{t=1,t\neq q}^{P} \mathbf{W}_{pt}(i)\mathbf{R}_{\mathbf{y}_t\mathbf{\Phi}_q(\mathbf{y}_q)}(i). \tag{3.92}$$

Based on this channel-wise matrix product two different approximations of the Sylvester constraint are now investigated.

Column Sylvester constraint $\mathcal{SC}_{\mathcal{C}}$

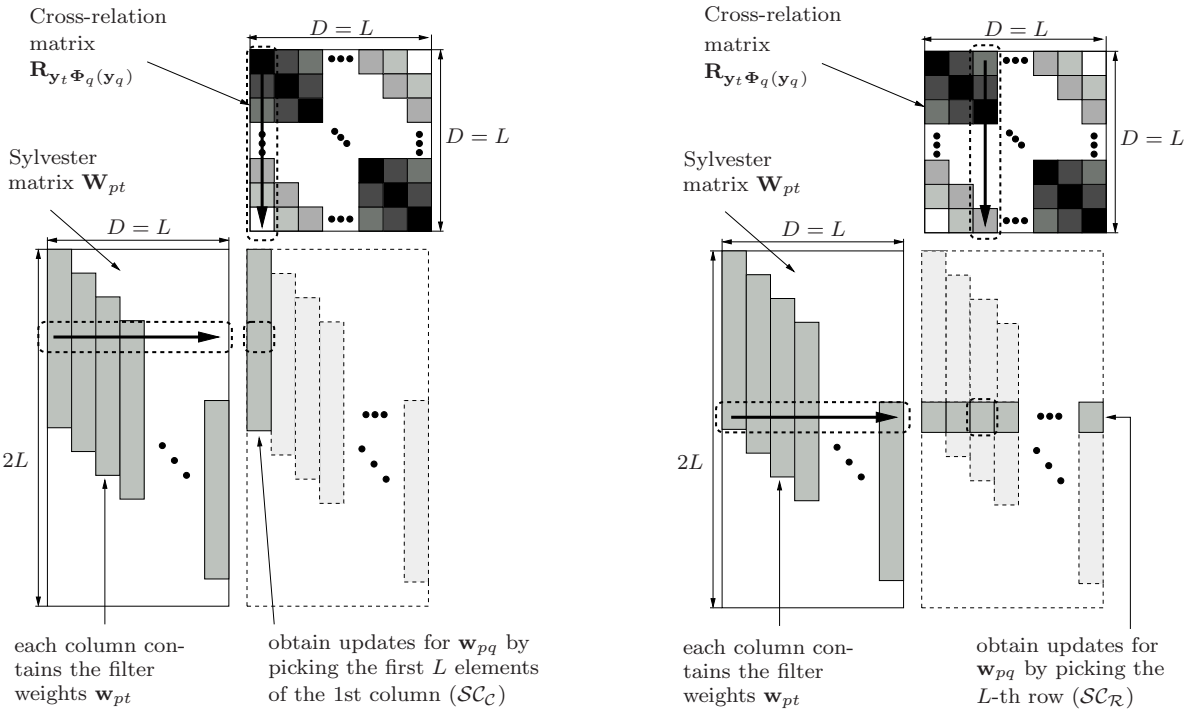
In Section 3.3.3 it was rigorously derived that the updates for the filter weights $\mathbf{w}_{pq}$ contained in the $PL \times P$ matrix $\check{\mathbf{W}}$ are obtained by the Sylvester constraint $\mathcal{SC}$ using the averaging operation depicted in Fig. 3.2. One possible approximation which still provides the updates for the filter weights $\mathbf{w}_{pq}$ is to compute only the *first L elements of the first column* of each $2L \times D$ submatrix $\nabla^{\text{NG}}_{\mathbf{W}_{pq}}\tilde{\mathcal{J}}(i,\mathbf{W})$ in (3.92). This neglects the averaging operation and thus reduces the matrix product to a matrix-vector product. This approach is termed *column Sylvester constraint* $\mathcal{SC}_{\mathcal{C}}$ and an illustration of $\mathcal{SC}_{\mathcal{C}}$ is given in Fig. 3.4a. There it can be seen that $D = L$ has to be chosen to be able to independently adjust all filter updates for all taps $\kappa = 0, \ldots, L-1$. Note that in general any other column could be chosen.

Furthermore, due to the Sylvester structure of $\mathbf{W}_{pt}(i)$ the matrix-vector product represents a convolution of the filter weight vector $\mathbf{w}_{pt}(i)$ with the first column of the cross-relation matrix $\mathbf{R}_{\mathbf{y}_t\mathbf{\Phi}_q(\mathbf{y}_q)}(i)$. By implementing the linear convolution expressed by the matrix-vector product as a fast convolution using fast Fourier transforms (FFTs) the computational complexity can thus be reduced to $\mathcal{O}(\log L)$. Exploiting this fact led to the efficient real-time implementation presented in [ABYK04, ABYK06].

Row Sylvester Constraint $\mathcal{SC}_{\mathcal{R}}$

If the number of time-lags D used to exploit the nonwhiteness property is chosen as $D = L$ then another possible approximation of the Sylvester constraint ($\mathcal{SC}$) is to compute only the *L-th row* of the update matrix $\nabla^{\text{NG}}_{\mathbf{W}_{pq}}\tilde{\mathcal{J}}(i,\mathbf{W})$. Thus, again the matrix product reduces to a matrix-vector product. This method is termed *row Sylvester constraint* $\mathcal{SC}_{\mathcal{R}}$ and is illustrated in Fig. 3.4b.

In the case of $\mathcal{SC}_{\mathcal{R}}$ the matrix-vector product can in general not be written as a convolution of two sequences as the matrix $\mathbf{R}_{\mathbf{y}_t\mathbf{\Phi}_q(\mathbf{y}_q)}(i)$ does not necessarily exhibit any special structure. However, if the correlation method (see Section 3.3.5) is used for estimating



(a) Sylvester constraint $\mathcal{SC}_\mathcal{C}$ (b) Sylvester constraint $\mathcal{SC}_\mathcal{R}$

Figure 3.4: Illustration of the channel-wise matrix-matrix product $\mathbf{W}_{pt}(i)\mathbf{R}_{\mathbf{y}_t\mathbf{\Phi}_q(\mathbf{y}_q)}(i)$ when using the Sylvester constraints $\mathcal{SC}_\mathcal{C}$ and $\mathcal{SC}_\mathcal{R}$.

the cross-relation matrices, then they exhibit a Toeplitz structure and are expressed as $\tilde{\mathbf{R}}_{\mathbf{y}_t\mathbf{\Phi}_q(\mathbf{y}_q)}(i)$. This approximation allows the expression of the matrix-vector product as the convolution of the filter weights $\mathbf{w}_{pq}$ with the two-sided cross-relation sequence.

Comparing the linear convolutions resulting from $\mathcal{SC}_\mathcal{C}$ and $\mathcal{SC}_\mathcal{R}$ shows that different cross-relation sequences are used for the convolution. In this discussion, it is assumed that for both Sylvester constraints the correlation method is used to estimate the cross-relation sequences. The version using $\mathcal{SC}_\mathcal{C}$ convolves the filter weights with the one-sided sequence of cross-relation elements $\tilde{r}_{y_t\Phi_q(\mathbf{y}_q)}(i,\tilde{v})$, $\tilde{v} = 0,\ldots,-L+1$ resulting in the filter update for the κ-th tap

$$\nabla^{\mathrm{NG}}_{w_{pq,\kappa}}\tilde{\mathcal{J}}(i,\mathbf{W}) \overset{(\mathcal{SC}_\mathcal{C})}{=} \sum_{\tilde{v}=0}^{\kappa} w_{pt,\tilde{v}}(i)\tilde{r}_{y_t\Phi_q(\mathbf{y}_q)}(i,\tilde{v}-\kappa). \tag{3.93}$$

The row Sylvester constraint $\mathcal{SC}_\mathcal{R}$ uses the two-sided sequence $\tilde{r}_{y_t\Phi_q(y_q)}(i,\tilde{v})$, $\tilde{v} = -L+1,\ldots,L-1$ resulting in a linear convolution

$$\nabla^{\mathrm{NG}}_{w_{pq,\kappa}}\tilde{\mathcal{J}}(i,\mathbf{W}) \overset{(\mathcal{SC}_\mathcal{R})}{=} \sum_{\tilde{v}=0}^{L-1} w_{pt,\tilde{v}}(i)\tilde{r}_{y_t\Phi_q(\mathbf{y}_q)}(i,\tilde{v}-\kappa). \tag{3.94}$$

These two implementation schemes have a crucial effect on the properties of the resulting

algorithm (see experiments in Section 3.6) and on the suitable initialization of $\mathbf{w}_{pq}$ as will be discussed in the following.

Appropriate initialization methods

When applying the previously derived BSS algorithms to realistic environments for $P = 2$, i.e., two sources and two microphones, we can distinguish two cases of acoustical scenarios (Fig. 3.5). From Section 3.1 we know that the demixing system contains the mixing filters in a rearranged order so that for each output channel blind interference cancellation is performed. Viewing BSS as blind interference cancellation allows to suggest possible initialization methods for both scenarios shown in Fig. 3.5. In the two-dimensional setup shown in Fig. 3.5a the two sources are located in two halfplanes and thus only causal FIR filters $\mathbf{w}_{pq}$ are needed to achieve interference cancellation as only delayed direct paths and reflections have to be modeled. Thus, the initialization of $\mathbf{w}_{pp}$ $(p = 1, 2)$ with a unit impulse at the first tap $w_{pp,0} = 1$ is sufficient while $w_{pt,\kappa} = 0$ for all κ and $t \neq p$. On the other hand, if the sources are in the same halfplane as in Fig. 3.5b, then for an

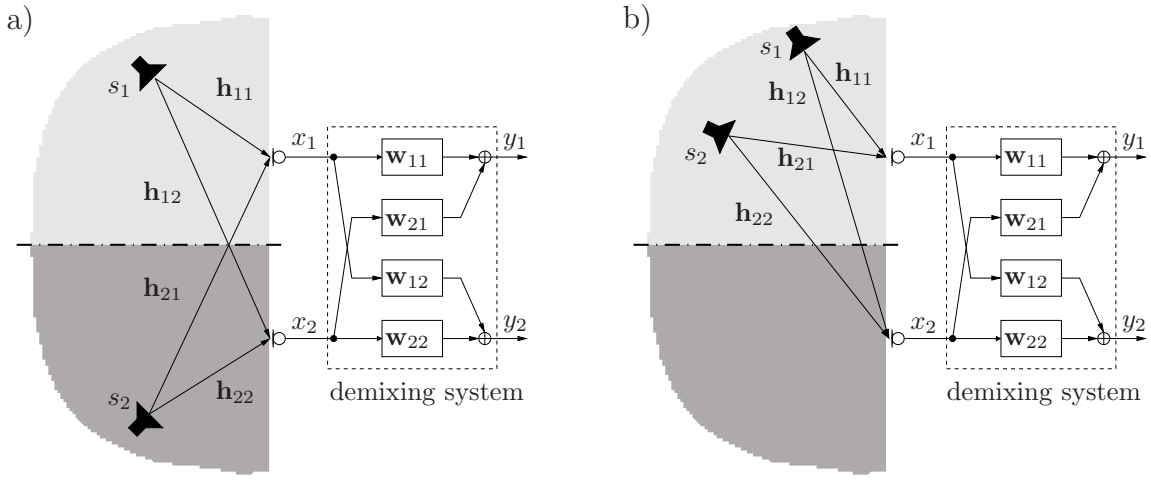


Figure 3.5: BSS setups for $P = 2$ requiring (a) only causal delays and (b) causal and acausal delays for the demixing FIR filters $\mathbf{w}_{pq}$.

initialization with $w_{pp,0} = 1$ (and if no permutation should be present, i.e., y_1 should be an estimate of s_1) a noncausal filter $\mathbf{w}_{21}$ is required. Analogously, if the sources are both in the other halfplane, $\mathbf{w}_{12}$ must be noncausal. For scenarios with $P > 2$ there will always be a need for noncausal filters. In practice noncausal filters can be implemented by initializing the adaptive filters $w_{pp,\kappa}$ with a shifted unit impulse. This initialization can be interpreted as shifting the point of origin which then allows both, causal and acausal filter taps. An appropriate shift is determined by the array geometry (i.e., for $P = 2$ by the distance between the sensors) and the resulting maximum possible delay

of the arriving signals between the sensors. This technique is also used in the design of fixed beamformers or in adaptive beamforming to be able to steer the spatial beam into all directions [vVB88]. As pointed out above, BSS can be seen as blind interference cancellation similar to conventional adaptive beamforming and thus, the approach of initialization with a shifted unit impulse is applicable.

The choice of initialization method also determines the suitable approximation of the Sylvester constraint $\mathcal{SC}$. In the case of causal mixtures (Fig. 3.5a), i.e., initialization with $w_{pp,0} = 1$, both Sylvester constraints $\mathcal{SC}_{\mathcal{C}}$ and $\mathcal{SC}_{\mathcal{R}}$ are possible. On the other hand for noncausal mixtures it is necessary to initialize with a shifted unit impulse such as, e.g., $w_{pp,L/2} = 1$. When evaluating the matrix product resulting from $\mathcal{SC}_{\mathcal{C}}$ (Fig. 3.4a) for the initialization $w_{pp,L/2} = 1$ it can be seen that all filter updates for the demixing filters $w_{pq,\kappa}$ for $0 \leq \kappa \leq L/2 - 1$ would be equal to zero, i.e., these filter coefficients could not be adapted. Thus, for the initialization with a shifted unit impulse only $\mathcal{SC}_R$ can be applied.

It can be concluded that if no a priori information about the location of the sources is available then the Sylvester constraint $\mathcal{SC}_{\mathcal{R}}$ together with an initialization using a shifted unit impulse should be applied due to its increased generality. Recall that then the correlation method is preferable for estimating the cross-correlation matrix because of its computational advantages over the covariance method. This applies also to the case for systems with $P > 2$. If it is known that the sources for a BSS system with $P = 2$ are in two halfplanes (e.g. car environment with array at the interior mirror), then also the slightly more robust (see [ABK05]) $\mathcal{SC}_{\mathcal{C}}$ can be used.

3.3.7 Approximations leading to known and novel algorithms

The natural gradient update (3.64) rule provides a very general basis for BSS of convolutive mixtures. However, to apply it to real-world scenarios, the multivariate score function (3.55) has to be estimated, i.e., we have to estimate P D-dimensional multivariate pdfs $\hat{p}_{y_q,D}(\mathbf{y}_q(iL + j))$, $q = 1, \ldots, P$. In general, this is a very challenging task, as it effectively requires estimation of all possible higher-order cumulants for a set of D output samples, where D may be on the order of several hundred or thousand in real acoustic environments.

In Section 3.3.7.1 we will present an efficient solution for the problem of estimating the multivariate pdfs by assuming so-called spherically invariant random processes (SIRPs). Moreover, efficient realizations based on second-order statistics will be derived in Section 3.3.7.2 by utilization of the multivariate Gaussian pdf. In Section 3.3.7.3 it will be shown how the approximation of multivariate pdfs by univariate pdfs leads to MCBD algorithms. Finally, relationships to well-known algorithms are shown.

3.3.7.1 Higher-order statistics realization based on multivariate pdfs

Early experimental measurements [Dav52] indicated that the pdf of speech signals in the time domain can be approximated by exponential distributions such as the Gamma or Laplacian pdf. Later on, a special class of multivariate pdfs based on the assumption of *spherically invariant random processes* (SIRPs) was introduced in [BS87] to model band-limited telephone speech. The SIRP model is representative for a wide class of stochastic processes [Yao73, Gol76, PSS00, Sel06] and it is very attractive since multivariate pdfs can be derived analytically from the corresponding univariate probability density function together with the correlation matrices covering multiple time-lags. The correlation matrices can be estimated from the data while for the univariate pdf appropriate models can be assumed or the univariate pdf can be estimated based on parameterized representations, such as the Gram-Charlier or Edgeworth expansions [Com94, HKO01].

The general model of a zero-mean non-white SIRP of D-th order for channel q is given by [BS87]

$$\hat{p}_{y_q,D}(\mathbf{y}_q(iL+j)) = \frac{1}{\sqrt{\pi^D \det(\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}(i))}} f_{y_q,D}\left(\mathbf{y}_q^{\mathrm{T}}(iL+j)\mathbf{R}^{-1}_{\mathbf{y}_q\mathbf{y}_q}(i)\mathbf{y}_q(iL+j)\right) \tag{3.95}$$

with the $D \times D$ correlation matrix $\mathbf{R}_{\mathbf{y}_p\mathbf{y}_q}$ defined in (3.68) and the function $f_{y_q,D}(\cdot)$ depending on the chosen univariate pdf. As the best known example, the multivariate Gaussian can be viewed as a special case of the class of SIRPs. The multivariate pdfs are completely characterized by the scalar function $f_{y_q,D}(\cdot)$ and $\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}$. Due to the quadratic form $\mathbf{y}_q^{\mathrm{T}}\mathbf{R}^{-1}_{\mathbf{y}_q\mathbf{y}_q}\mathbf{y}_q$, the pdf is spherically invariant which means for the bivariate case ($D = 2$) that independently of the choice of $f_{y_q,D}(\cdot)$ the bivariate pdfs based on the SIRP model exhibit ellipsoïdal or circular contour lines (see Fig. 3.6). The function $f_{y_q,D}(\cdot)$ is determined by the choice of the univariate pdf and can be calculated by using the so-called Meijer's G-functions as detailed in [Bre82, BS87].

In [Yao73] it was shown that the SIRP pdf can also be expressed as a random mixture of D-dimensional Gaussian pdfs. This is based on the decomposition of the column vector $\mathbf{y}_q$ of length D given for zero-mean random variables as

$$\mathbf{y}_q = \sqrt{z} \cdot \mathbf{v}, \tag{3.96}$$

where the column vector $\mathbf{v}$ of length D exhibits a multivariate Gaussian pdf with the correlation matrix $\mathrm{E}\{\mathbf{v}\mathbf{v}^{\mathrm{T}}\}$ and z is a non-negative scalar with distribution $\hat{p}_{z,1}(z)$. Then $\mathbf{y}_q$ is denoted as a *normal variance mixture* or *scale mixture of Gaussians*. The multivariate SIRP pdf is then obtained by calculating

$$\hat{p}_{y_q,D}(\mathbf{y}_q) = \int_0^\infty \hat{p}_{y_q|z,D}(\mathbf{y}_q|z)\hat{p}_{z,1}(z)\mathrm{d}z \tag{3.97}$$

which can be seen as an alternative to the computation of the SIRP pdf using (3.95) and Meijer's G-functions. It should be noted that the relationship of the correlation matrices of $\mathbf{y}_q$ and $\mathbf{v}$ is given as $\mathrm{E}\{\mathbf{y}_q\mathbf{y}_q^{\mathrm{T}}\} = z \cdot \mathrm{E}\{\mathbf{v}\mathbf{v}^{\mathrm{T}}\}$ [Yao73].

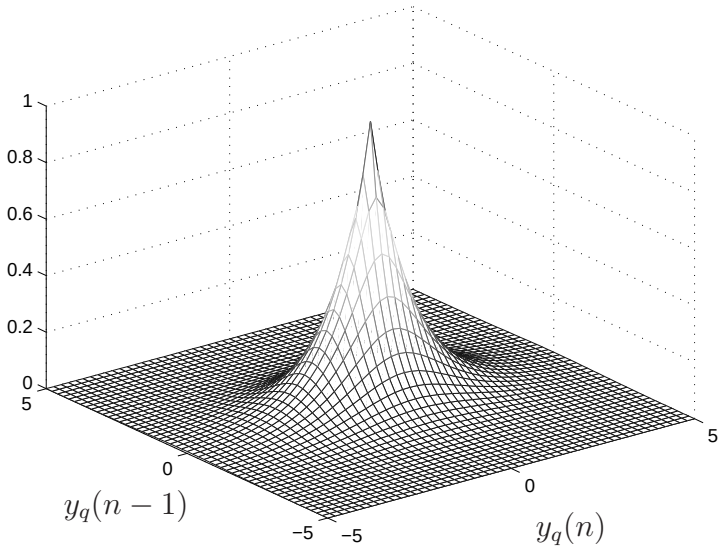


Figure 3.6: Illustration of a bivariate SIRP pdf (i.e., $D = 2$).

By introducing SIRPs into the BSS optimization criterion we obtain a considerably simplified expression for the multivariate score function (3.55) as first presented in [BAK03a]. After applying the chain rule to (3.95), the multivariate score function for the q-th channel can be expressed as

$$\begin{aligned}\boldsymbol{\Phi}_q(\mathbf{y}_q(iL+j)) &= -\frac{\frac{\partial \hat{p}_{y_q,D}(\mathbf{y}_q(iL+j))}{\partial \mathbf{y}_q(iL+j)}}{\hat{p}_{y_q,D}(\mathbf{y}_q(iL+j))} \\ &= \underbrace{2\left[-\frac{\frac{\partial f_{y_q,D}(u_q(iL+j))}{\partial u_q(iL+j)}}{f_{y_q,D}(u_q(iL+j))}\right]}_{:=\phi_{y_q,D}(u_q(iL+j))} \mathbf{R}^{-1}_{\mathbf{y}_q\mathbf{y}_q}(i)\mathbf{y}_q(iL+j), \end{aligned} \tag{3.98}$$

For convenience, we call the scalar function $\phi_{y_q,D}(u_q(iL+j))$ the *SIRP score* of channel q and the scalar argument given as the quadratic form is defined as

$$u_q(iL+j) = \mathbf{y}_q^{\mathrm{T}}(iL+j)\mathbf{R}^{-1}_{\mathbf{y}_q\mathbf{y}_q}(i)\mathbf{y}_q(iL+j). \tag{3.99}$$

From (3.98) it can be seen that the estimation of multivariate pdfs reduces to an estimation of the correlation matrix together with a computation of the SIRP score which can be determined by choosing suitable models for the multivariate SIRP pdf.

In [BS87, GZ03] it was shown that the *spherically symmetric multivariate Laplacian pdf* which exhibits Laplacian marginals is a good model for long-term properties of speech signals in the time-domain. A derivation of the multivariate Laplacian based on SIRPs

can be found in, e.g., [BS87, KKP01, EKL06, Sel06] and leads to

$$\hat{p}_{y_q,D}(\mathbf{y}_q(iL+j)) = \frac{1}{\sqrt{\pi^D \det\{\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}(i)\}}} \left(\frac{1}{\sqrt{2u_q(iL+j)}}\right)^{D/2-1} K_{D/2-1}\left(\sqrt{2u_q(iL+j)}\right), \tag{3.100}$$

and $K_\nu(\cdot)$ denotes the ν-th order modified Bessel function of the second kind. Comparing (3.95) and (3.100) shows that for the multivariate Laplacian SIRP pdf the function $f_{y_q,D}(u_q(iL+j))$ is given as

$$f_{y_q,D}(u_q(iL+j)) = \left(\frac{1}{\sqrt{2u_q(iL+j)}}\right)^{D/2-1} K_{D/2-1}\left(\sqrt{2u_q(iL+j)}\right). \tag{3.101}$$

The SIRP score for the multivariate Laplacian SIRP pdf can be straightforwardly derived by using the relation for the derivative of a ν-th order modified Bessel function of the second kind given as [AS72, Sel06]

$$\frac{\partial K_\nu\left(\sqrt{2u_q}\right)}{\partial\sqrt{2u_q}} = \frac{\nu}{\sqrt{2u_q}} K_\nu\left(\sqrt{2u_q}\right) - K_{\nu+1}\left(\sqrt{2u_q}\right), \tag{3.102}$$

and is obtained as

$$\phi_{y_q,D}(u_q(iL+j)) = \frac{1}{\sqrt{2u_q(iL+j)}} \frac{K_{D/2}\left(\sqrt{2u_q(iL+j)}\right)}{K_{D/2-1}\left(\sqrt{2u_q(iL+j)}\right)}. \tag{3.103}$$

It should be noted that a slightly different but equivalent formulation to (3.103) was given in [BAK03a]. In practical implementations the ν-th order modified Bessel function of the second kind $K_\nu(\sqrt{2u_q})$ may be approximated as [AS72]

$$K_\nu(\sqrt{2u_q}) = \sqrt{\frac{\pi}{2\sqrt{2u_q}}} e^{-\sqrt{2u_q}} \left(1 + \frac{4\nu^2-1}{8\sqrt{2u_q}} + \frac{(4\nu^2-1)(4\nu^2-9)}{2!(8\sqrt{2u_q})^2} + \dots\right). \tag{3.104}$$

Having derived the multivariate score function (3.98) for the SIRP model, we can now insert it into the generic HOS natural gradient update equation with its nonholonomic extension (3.64). Considering the fact that the auto-correlation matrices are symmetric so that $\left(\mathbf{R}^{-1}_{\mathbf{y}_q\mathbf{y}_q}\right)^{\mathrm{T}} = \mathbf{R}^{-1}_{\mathbf{y}_q\mathbf{y}_q}$ leads to the following expression for the *nonholonomic HOS-SIRP natural gradient*:

$$\nabla^{\mathrm{NG}}_{\check{\mathbf{W}}} \mathcal{J}(m) = \mathcal{SC}\left\{2\sum_{i=0}^{\infty} \beta(i,m)\mathbf{W}(i)\left\{\mathbf{R}_{\mathbf{y}\phi(\mathbf{y})}(i) - \mathrm{bdiag}\{\mathbf{R}_{\mathbf{y}\phi(\mathbf{y})}(i)\}\right\}\mathrm{bdiag}^{-1}\{\mathbf{R}_{\mathbf{yy}}(i)\}\right\} \tag{3.105}$$

with the second-order correlation matrix $\mathbf{R}_{\mathbf{yy}}$ defined in (3.71) and $\mathbf{R}_{\mathbf{y}\phi(\mathbf{y})}$ consisting of the channel-wise submatrices $\mathbf{R}_{\mathbf{y}_p\phi(\mathbf{y}_q)}$ given as

$$\mathbf{R}_{\mathbf{y}_p\phi(\mathbf{y}_q)}(i) = \frac{1}{N}\sum_{j=0}^{N-1} \mathbf{y}_p(iL+j)\phi_{y_q,D}(u_q(iL+j))\mathbf{y}_q^{\mathrm{T}}(iL+j). \tag{3.106}$$

The SIRP score $\phi_{y_q,D}(\cdot)$ of channel q in (3.106) is a scalar value function which causes a weighting of the correlation matrix and is defined in (3.98). With regard to a DFT-domain formulation we can formulate $\mathbf{R}_{\mathbf{y}_p\phi(\mathbf{y}_q)}$ analogously to (3.71) by using matrices instead of vectors and for a concise formulation we replace the transpose operator $(\cdot)^{\mathrm{T}}$ by the hermitian operator $(\cdot)^{\mathrm{H}}$ leading to

$$\mathbf{R}_{\mathbf{y}_p\phi(\mathbf{y}_q)}(i) = \frac{1}{N}\mathbf{Y}_p(i)\mathbf{\Lambda}_q^{\mathrm{H}}(i)\mathbf{Y}_q^{\mathrm{H}}(i), \tag{3.107}$$

where the nonlinear weighting by the multivariate score function is expressed using (3.99) as a diagonal $N \times N$ matrix given as

$$\mathbf{\Lambda}_q(i) = \phi_{y_q,D}\left(\mathrm{diag}\left\{\mathbf{Y}_q^{\mathrm{H}}(i)\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}^{-1}(i)\mathbf{Y}_q(i)\right\}\right), \tag{3.108}$$

where the operator $\mathrm{diag}\{\mathbf{A}\}$ sets all off-diagonal elements of matrix $\mathbf{A}$ to zero and the SIRP score function $\phi_{y_q,D}(\cdot)$ is applied element-wise to the diagonal matrix in its argument.

As only channel-wise submatrices have to be inverted in (3.105) it is sufficient to choose $N > D$ instead of $N > PD$ for the estimation of $\mathbf{R}_{\mathbf{yy}}(i)$. Moreover, from the update equation (3.105), it can be seen that the SIRP model leads to an inherent normalization by the auto-correlation submatrices. This becomes especially obvious if the update (3.105) is written explicitly for a 2-by-2 MIMO system leading to

$$\nabla_{\tilde{\mathbf{W}}}^{\mathrm{NG}}\mathcal{J}(m,\mathbf{W}) = \mathcal{SC}\left\{2\sum_{i=0}^{\infty}\beta(i,m)\mathbf{W}(i)\begin{bmatrix}\mathbf{0} & \mathbf{R}_{\mathbf{y}_1\phi(\mathbf{y}_2)}(i)\mathbf{R}_{\mathbf{y}_2\mathbf{y}_2}^{-1}(i)\\ \mathbf{R}_{\mathbf{y}_2\phi(\mathbf{y}_1)}(i)\mathbf{R}_{\mathbf{y}_1\mathbf{y}_1}^{-1}(i) & \mathbf{0}\end{bmatrix}\right\}. \tag{3.109}$$

The normalization is important as it provides good convergence even for correlated signals such as speech and also for a large number of filter taps. The normalization resembles the recursive least-squares (RLS) algorithm in supervised adaptive filtering where also the inverse of the auto-correlation matrix is computed [Hay02]. To obtain efficient implementations, the normalization by the computationally demanding inverse of the $D \times D$ matrix can be approximated in several ways as shown in Section 3.3.8 and 3.4.3.1. Additionally, for a real-time implementation also the argument u_q of the SIRP score function $\phi_{y_q,D}(u_q)$ has to be calculated efficiently by using suitable approximations as will be shown in Section 3.4.3.2. Moreover, it should be pointed out that the matrix multiplication in (3.109) with the Sylvester matrix $\mathbf{W}$ can be interpreted as a convolution (see Section 3.3.6) and thus, efficient implementations using fast convolutions are possible.

3.3.7.2 Second-order statistics realization based on the multivariate Gaussian pdf

Using the model of the multivariate Gaussian pdf leads to a second-order realization of the BSS algorithm utilizing the nonstationarity and the nonwhiteness of the source signals. The multivariate Gaussian pdf

$$\hat{p}_{y_q,D}(\mathbf{y}_q(iL+j)) = \frac{1}{\sqrt{(2\pi)^D \det(\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}(i))}} e^{-\frac{1}{2}\mathbf{y}_q^{\mathrm{T}}(iL+j)\mathbf{R}^{-1}_{\mathbf{y}_q\mathbf{y}_q}(i)\mathbf{y}_q(iL+j)} \tag{3.110}$$

is inserted in the expression for the multivariate score function (3.55) whose elements reduce to

$$\mathbf{\Phi}_q(\mathbf{y}_q(iL+j)) = \mathbf{R}^{-1}_{\mathbf{y}_q\mathbf{y}_q}(i)\mathbf{y}_q(iL+j). \tag{3.111}$$

Inserting (3.111) into the natural gradient update (3.62) yields the *generic SOS natural gradient*

$$\nabla^{\mathrm{NG}}_{\mathbf{W}} \mathcal{J}(m,\mathbf{W}) = \mathcal{SC}\left\{\sum_{i=0}^{\infty} \beta(i,m)\mathbf{W}(i)\left\{\mathbf{R}_{\mathbf{yy}}(i) - \operatorname{bdiag}\mathbf{R}_{\mathbf{yy}}(i)\right\}\operatorname{bdiag}^{-1}\mathbf{R}_{\mathbf{yy}}(i)\right\}. \tag{3.112}$$

Comparing (3.112) to the HOS-SIRP update (3.105) shows that due to the fact that only SOS are utilized, we obtain the same update with the nonlinearity (3.103) omitted, i.e., $\phi_{y_q,D}(u_q(iL+j)) = 1$, $q = 1,\ldots,P$. Therefore, the SOS natural gradient update also exhibits the inherent normalization by the auto-correlation matrices which leads to very robust convergence behavior in real-world environments. For the 2×2 case we can express (3.112) as

$$\nabla^{\mathrm{NG}}_{\mathbf{W}} \mathcal{J}(m,\mathbf{W}) = \mathcal{SC}\left\{\sum_{i=0}^{\infty} \beta(i,m)\mathbf{W}(i)\begin{bmatrix} \mathbf{0} & \mathbf{R}_{\mathbf{y}_1\mathbf{y}_2}(i)\mathbf{R}^{-1}_{\mathbf{y}_2\mathbf{y}_2}(i) \\ \mathbf{R}_{\mathbf{y}_2\mathbf{y}_1}(i)\mathbf{R}^{-1}_{\mathbf{y}_1\mathbf{y}_1}(i) & \mathbf{0} \end{bmatrix}\right\}. \tag{3.113}$$

Moreover, due to the inversion of channel-wise $D \times D$ submatrices, $N > D$ instead of $N > PD$ is again sufficient for the estimation of the correlation matrices. The estimation of the correlation matrices can be done according to the correlation or covariance method as outlined in Section 3.3.5. Furthermore, it is important to note that the matrix product of Sylvester matrices $\mathbf{W}_{pq}$ and the remaining matrices in the update equation (3.113) can be described by linear convolutions.

In Fig. 3.7 the structure of the cost function in the case of SOS and idealized/simplified mechanism of the adaptation update (3.112) is illustrated. By assuming the multivariate Gaussian pdf (3.110) and then minimizing $\mathcal{J}(m,\mathbf{W})$, all cross-correlations for D time-lags are reduced and thus the algorithm exploits nonwhiteness. Nonstationarity is utilized by minimizing the correlation matrices simultaneously for several blocks i. Ideally, the

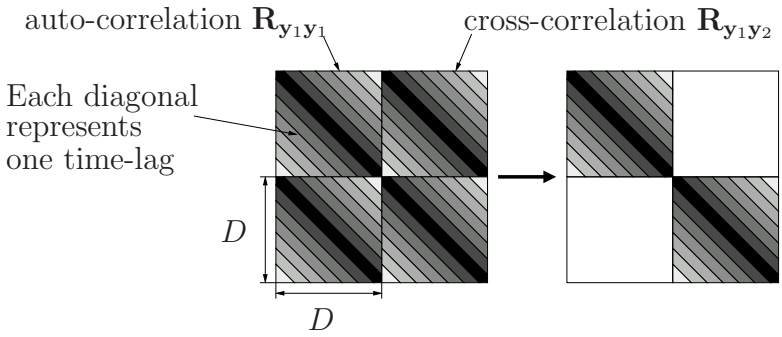


Figure 3.7: Illustration of the diagonalization of the correlation matrices performed by the natural gradient update (3.113) for the 2×2 case.

cross-correlations will be equal to zero upon convergence which causes the update term to be zero because then $\mathbf{R}_{\mathbf{yy}}(i) - \operatorname{bdiag} \mathbf{R}_{\mathbf{yy}}(i) = \mathbf{0}$.

Another interesting finding is that for both, the holonomic and nonholonomic versions of the HOS update (3.62), (3.64), the insertion of the Gaussian pdf (3.110) yields the same SOS BSS algorithm (3.112). This SOS BSS algorithm puts no constraint on the auto-correlation $\mathbf{R}_{\mathbf{y}_p\mathbf{y}_p}$ of the estimated sources. This explains its good performance for speech sources. According to the BSS terminology an algorithm with such a behavior would fall into the class of nonholonomic algorithms and thus, this shows that even if starting from a holonomic algorithm we can obtain nonholonomic algorithms if special pdfs such as the Gaussian multivariate pdf are used.

An alternative derivation of a SOS BSS algorithm leading to the same natural gradient update as given in (3.112) was presented in [BAK03b, BAK05a]. There, the derivation was based on a generalized version of the cost function used in [MOK95], which also simultaneously exploits nonwhiteness and nonstationarity of the sources.

3.3.7.3 Realizations based on univariate pdfs

If the output signals $y_q(n)$ are assumed to be temporally independent, then the D-dimensional multivariate pdf $\hat{p}_{y_q,D}(\mathbf{y}_q(iL+j))$ reduces to a product of univariate pdfs $\hat{p}_{y_q,1}(\cdot)$. Thereby, the multivariate score function (3.55) reduces to a univariate score function representing a scalar nonlinearity

$$\Phi_q(y_q(n)) = -\frac{\frac{\partial \hat{p}_{y_q,1}(y_q(n))}{\partial y_q(n)}}{\hat{p}_{y_q,1}(y_q(n))}, \tag{3.114}$$

which is applied individually to each element of $\mathbf{y}_q(iL+j) = [y_q(iL+j), \ldots, y_q(iL+j-D+1)]^{\mathrm{T}}$. The drawback of this simplified score function is that this approach can only be applied to temporally independent signals as encountered, e.g., in telecommunications. The resulting natural gradient algorithms obtained by inserting the scalar nonlinearity

(3.114) into (3.62) or (3.64) estimate the original source signals by blindly deconvolving the sensor signals and therefore, belong to the class of multi-channel blind deconvolution (MCBD) algorithms. Application of algorithms based on a scalar nonlinearity to temporally correlated signals results in a temporal whitening of the estimated separated output signals. As discussed in Section 3.2 several heuristic countermeasures have been proposed to mitigate this whitening effect.

The scalar score function in MCBD algorithms is chosen depending on the model of the source signals. In the literature, many heuristic nonlinearities have been proposed as realizations of the scalar score function. For supergaussian distributions, i.e., distributions with sharper peaks and longer tails than the Gaussian pdf, e.g., $\Phi(y_q(n)) = \tanh(a \cdot y_q(n))$ with the scaling parameter a or $\Phi(y_q(n)) = \text{sign}(y_q(n))$ are commonly used nonlinearities. For subgaussian distributions, e.g., $\Phi(y_q(n)) = y_q^3(n)$ is a popular choice. A discussion of the properties of various nonlinearities can be found, e.g., in [MD02]. Moreover, the score function can also be determined by estimating the univariate pdfs. This is done by using expansions for the pdf in the vicinity of a Gaussian pdf similar to the Taylor expansion. This allows to express the univariate pdfs in terms of the higher order moments which are then estimated. Two prominent examples which are usually used in this context are the Gram-Charlier or Edgeworth expansion [HKO01].

Using (3.114) several relationships between the generic HOS natural gradient update rule (3.62) and well-known MCBD algorithms in literature can be established [ABK05]. These links are obtained by the application of different implementations of the Sylvester constraint ($\mathcal{SC}$), the distinction between correlation and covariance method, and the different approximations of the multivariate pdfs. This altogether spans a whole tree of algorithms as depicted in Fig. 3.8. The most general algorithm is given as the generic HOS natural gradient algorithm (3.62) which is based on multivariate pdfs. A distinction with respect to the implementation of the Sylvester constraint $\mathcal{SC}$ leads to two branches which can again be split up with respect to the method used for the estimation of the cross-relation matrices. Approximating the multivariate pdfs by univariate ones, neglecting the nonstationarity, and using the Sylvester constraint $\mathcal{SC}_\mathcal{R}$ yields two block-based MCBD algorithms presented in [JS03, DSM04a]. By changing the block-based adaptation to a sample-by-sample algorithm, a link to the popular MCBD algorithm in [ADCY97] and to [DSM04b] can be established. It should be noted that also the nonholonomic extension of [ADCY97] presented in [CACL99] can be derived from the presented framework by using (3.64) instead of (3.62). By using the Sylvester constraint $\mathcal{SC}_\mathcal{C}$ a link to the MCBD algorithm in [ZCA99] is obtained. However, it should be noted that algorithms based on $\mathcal{SC}_\mathcal{C}$ are less general as only causal filters can be adapted and thus for MCBD algorithms only minimum-phase systems can be treated as was pointed out in [ZCA99].

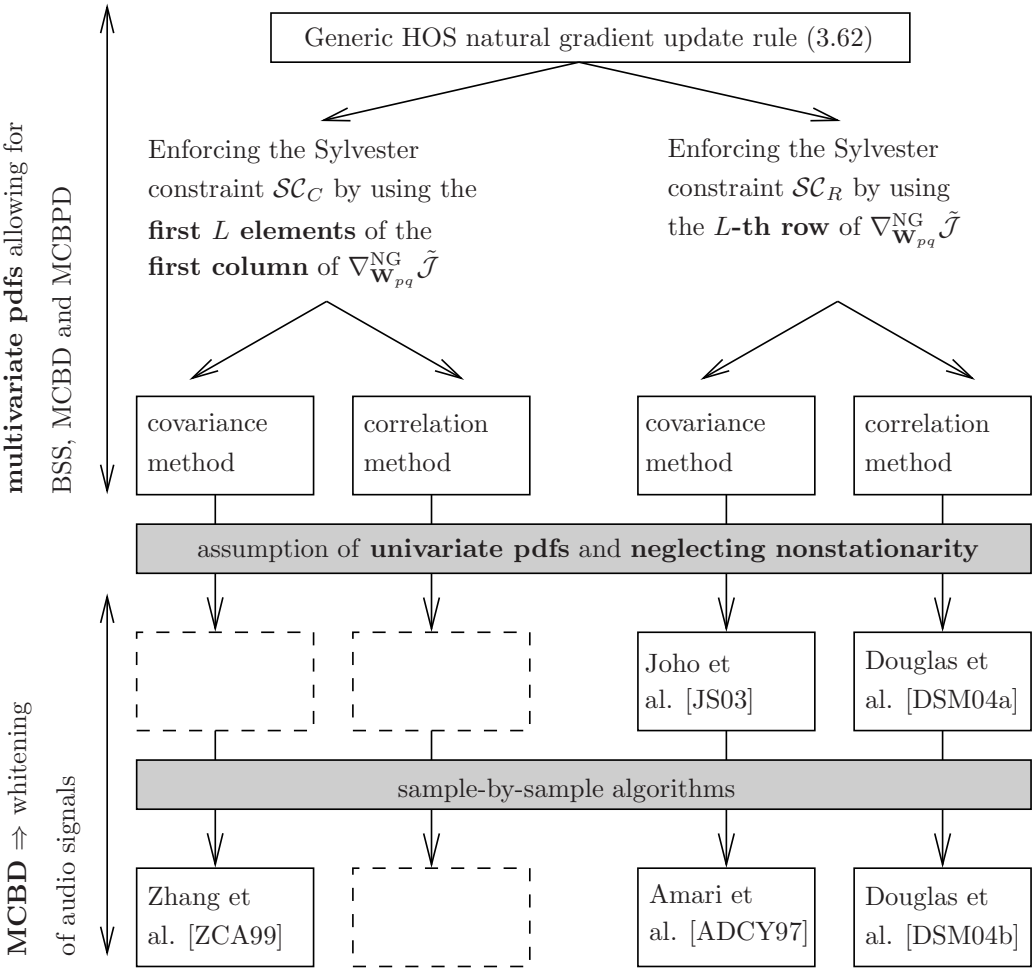


Figure 3.8: Overview of links between the generic HOS natural gradient update rule (3.62) and existing MCBD algorithms.

3.3.8 Efficient normalization and regularization strategies

In Section 3.3.7 it was pointed out that the usage of multivariate pdfs leads to an inherent normalization by the auto-correlation matrices. This is desirable as it guarantees good convergence of the adaptive filters even for large filter lengths and correlated input signals. On the other hand this poses the problem of large computational complexity due to the required matrix inversion of P matrices of size $D \times D$. The complexity is $\mathcal{O}(D^3)$ for using the covariance method and $\mathcal{O}(D^2)$ for the correlation method due to the Toeplitz structure involved. However, as D may be even larger than 1000 for realistic environments this is still prohibitive for a real-time implementation on regular PC platforms. Therefore, approximations are desirable which reduce the complexity with minimum degradation of the separation performance.

One possible solution is to approximate the auto-correlation matrices $\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}(i)$ by a

diagonal matrix, i.e., by the output signal powers

$$\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}(i) \approx \frac{1}{N}\sum_{j=0}^{N-1} \operatorname{diag}\left\{\mathbf{y}_q(iL+j)\mathbf{y}_q^{\mathrm{H}}(iL+j)\right\}. \tag{3.115}$$

This case corresponds to a simplification of the covariance method where the values on the diagonal are not identical. For the correlation method stationarity within each block i is assumed and thus, the correlation matrix which is then denoted as $\tilde{\mathbf{R}}_{\mathbf{y}_q\mathbf{y}_q}$ further simplifies to

$$\begin{aligned}\tilde{\mathbf{R}}_{\mathbf{y}_q\mathbf{y}_q}(i) &= \frac{1}{N}\sum_{n=iL}^{iL+N-1} |y_q(n)|^2\mathbf{I} \\ &:= \sigma_{y_q}^2(i)\mathbf{I},\end{aligned} \tag{3.116}$$

for $q = 1, \ldots, P$, where the operator $\operatorname{diag}\{\mathbf{A}\}$ sets all off-diagonal elements of matrix $\mathbf{A}$ to zero. This approximation is comparable to the well-known normalized least mean squares (NLMS) algorithm in supervised adaptive filtering approximating the RLS algorithm [Hay02]. It should be noted that the SOS natural gradient algorithm based on (3.112) together with the approximation (3.115) was also heuristically introduced for the case $D = L$ in [NSS02, AAM+02] as an extension of [KMO98] incorporating several time-lags. Further approximations of the normalizations exploiting the efficiency of computations in the DFT domain are presented in Section 3.4.3.1.

For blocks with speech pauses and low background noise the normalization by the auto-correlation matrix $\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}$ leads to the inversion of an ill-conditioned matrix or in the case of the approximations (3.115) and (3.116) to a division by very small output powers $\sigma_{y_q}^2$ or even by zero and thus the estimation of the filter coefficients becomes very sensitive. For a robust adaptation $\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}$ is replaced by a regularized version $\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q} + \delta_{y_q}\mathbf{I}$. The basic feature of the regularization is a compromise between fidelity to data and fidelity to prior information about the solution [CU94]. As the latter increases robustness but leads to biased solutions, in this thesis, similarly to supervised adaptive filtering [BBK03, BBK05], a dynamic regularization is used

$$\delta_{y_q} = \delta_{\max} e^{-\sigma_{y_q}^2/\sigma_0^2} \tag{3.117}$$

with two parameters $\delta_{\max}$ and σ_0^2. This exponential method provides a smooth transition between regularization for low output power $\sigma_{y_q}^2$ and data fidelity whenever the output power is large enough. Other popular strategies are the fixed regularization which simply adds a constant value to the output power and the approach of choosing the maximum out of the respective component $\sigma_{y_q}^2$ and a fixed threshold δ_{th}.

Moreover, it should be pointed out that especially for second-order BSS algorithms there exists another popular class of algorithms which completely lacks any normalization.

They are based on cost functions using the Frobenius norm $\|\mathbf{A}\|_{\mathrm{F}}^2 = \sum_{i,j} a_{ij}^2$ of a matrix $\mathbf{A} = (a_{ij})$ aiming at minimizing the cross-correlations. Several algorithms for instantaneous mixtures have been proposed (e.g., [HKO01, MS94, CS96]) and the Frobenius norm has also been suggested for convolutive mixtures, e.g., in [KJ00, Joh04]. In the latter case a possible cost function based on the Frobenius norm is given as [BAK04a, BAK05a]

$$\mathcal{J}_{\mathrm{F}}(m, \mathbf{W}) = \sum_{i=0}^{\infty} \beta(i,m) \left\| \mathbf{R}_{\mathbf{yy}}(i) - \mathrm{bdiag}\left\{\mathbf{R}_{\mathbf{yy}}(i)\right\} \right\|_{\mathrm{F}}^2 . \tag{3.118}$$

By using the trace operator $\mathrm{tr}\{\cdot\}$, the Frobenius norm can be expressed as $\|\mathbf{A}\|_{\mathrm{F}}^2 = \mathrm{tr}\{\mathbf{A}^{\mathrm{T}}\mathbf{A}\}$, and the derivative of the Frobenius norm with respect to $\mathbf{A}$ is given as $\frac{\partial\, \mathrm{tr}\{\mathbf{A}^{\mathrm{T}}\mathbf{A}\}}{\partial \mathbf{A}} = 2\mathbf{A}$ [Har97]. By using this relation and applying the chain rule it was shown in [BAK05a] that the natural gradient update is obtained as

$$\nabla_{\check{\mathbf{W}}}^{\mathrm{NG}} \mathcal{J}(m, \mathbf{W}) = \mathcal{SC}\left\{ 2 \sum_{i=0}^{\infty} \beta(i,m) \mathbf{W}(i) \mathbf{R}_{\mathbf{yy}}(i) \left(\mathbf{R}_{\mathbf{yy}}(i) - \mathrm{bdiag}\, \mathbf{R}_{\mathbf{yy}}(i) \right) \right\}. \tag{3.119}$$

This update equation differs from the more general equation (3.112) mainly in the lack of the inherent normalization expressed by the inverse matrices $\mathrm{bdiag}^{-1}\left\{\mathbf{R}_{\mathbf{yy}}\right\}$. Thus, (3.119) can be regarded as an analogon to the least-mean-square (LMS) algorithm [Hay02] in supervised adaptive filtering. Even if (3.119) can be efficiently implemented, many simulation results have shown that for large filter lengths L and nonstationary input signals, (3.119) is prone to instability, while algorithms exhibiting a normalization show a very robust convergence behavior in real acoustic environments which require a large filter length L [BAK05a].

3.3.9 Summary

In Section 3.3 the goal was to introduce a generic optimization criterion which allows to exploit all three signal properties: nongaussianity, nonstationarity, and nonwhiteness. The latter one was utilized by introducing a novel matrix formulation in Sect. 3.3.1 allowing for a memory of $D-1$ samples necessary for modeling temporal dependencies. In Sect. 3.3.2 the TRINICON optimization criterion, based on a generalization of the mutual independence and accounting for all three signal properties simultaneously, was presented. From the TRINICON optimization criterion several novel algorithms were derived by applying certain approximations as summarized in the flowchart in Fig. 3.9. First, the gradient and natural gradient update rule were derived in Sections 3.3.3 and 3.3.4, respectively. Subsequently, the block-based estimation of the higher-order cross-relation matrices, appearing in the update equations, was discussed. This distinction between different estimation methods is known from linear prediction problems and was applied to BSS algorithms in Section 3.3.5. Furthermore, a Sylvester constraint is necessary to

ensure that the gradient with respect to the Sylvester matrix $\mathbf{W}$ exhibits again a Sylvester structure. In Section 3.3.6 efficient versions of this Sylvester constraint $\mathcal{SC}$ and the resulting appropriate initializations are discussed. The generic update equations derived from the TRINICON optimization criterion require the estimation of high-dimensional multivariate pdfs which is a very challenging task. Therefore, in Section 3.3.7 several approximations as shown in Fig. 3.9 are discussed. An efficient solution to the estimation of multivariate pdfs is to assume spherically invariant random processes (Section 3.3.7.1) which only require the estimation of a univariate pdf together with the estimation of a correlation matrix including time lags. This leads to a novel HOS BSS approach which explicitly models the temporal dependencies of source signals such as speech. The multivariate Gaussian pdf as a special case of SIRP leads to an SOS BSS algorithm which simultaneously exploits nonwhiteness and nonstationarity. In Section 3.3.7.3 the temporal dependencies of the source signals are neglected and thus univariate pdfs are obtained in the generic update equations leading to a set of MCBD algorithms. This allows to establish relationships to several popular MCBD algorithms in literature. Finally, in Section 3.3.8 the normalization by the inverse auto-correlation matrices, which appears in algorithms based on multivariate pdfs is approximated by the inverse of a diagonal matrix. The normalization ensures good convergence when using temporally correlated and nonstationary signals such as speech and together with a regularization method leads to efficient and robust BSS algorithms.

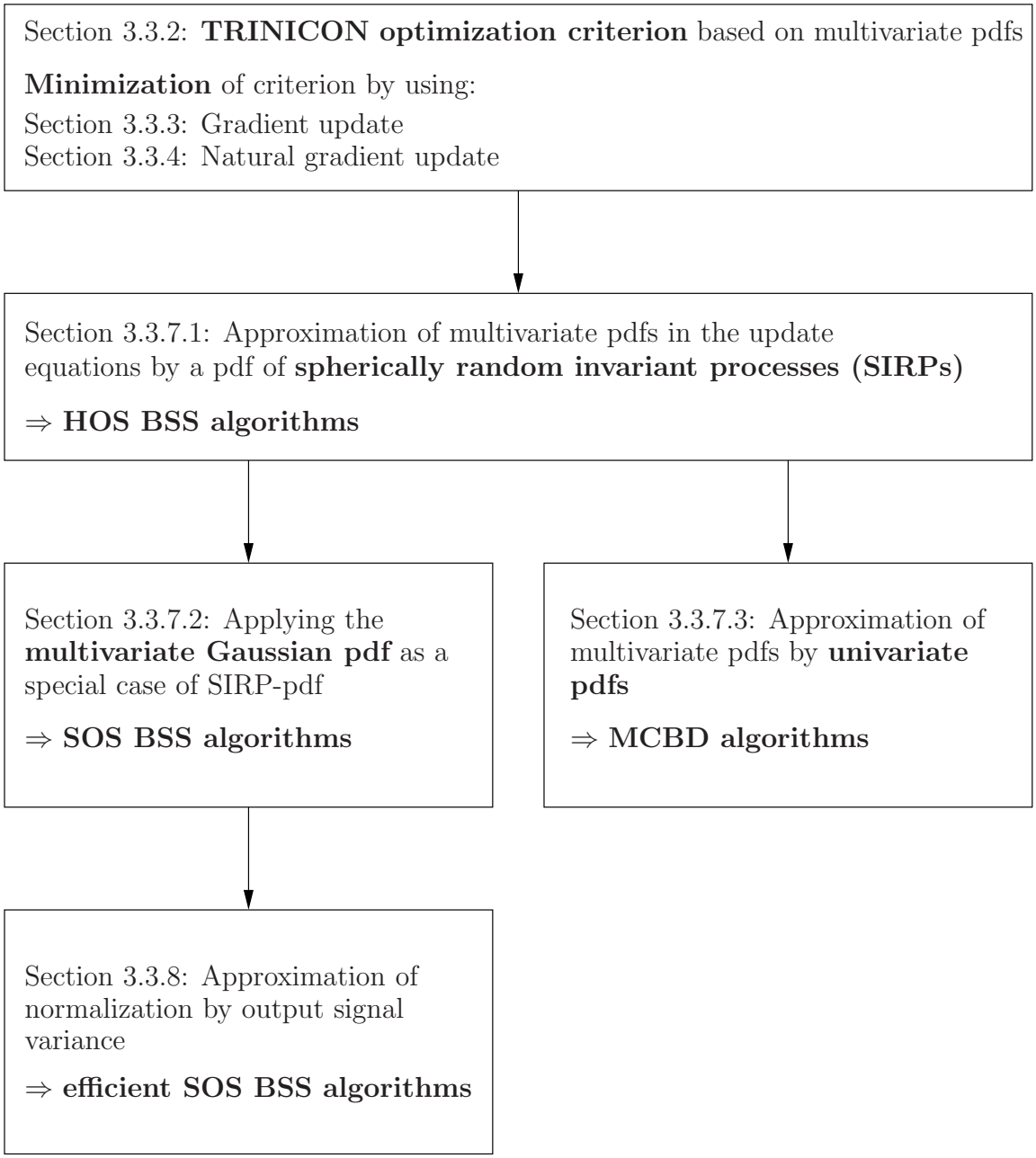


Figure 3.9: Flowchart showing the relations between the generic algorithms based on the TRINICON optimization criterion and its various approximations leading to efficient BSS and MCBD algorithms.

3.4 Broadband and narrowband DFT-domain algorithms

In the previous section a time-domain framework for convolutive BSS was introduced leading to several novel algorithms. However, due to the large matrices involved a straightforward implementation in the time domain is inefficient. Therefore, in this section the framework will be extended to the DFT domain which is attractive for reasons of computational efficiency. As already pointed out in Section 3.2 there are two possibilities to design algorithms in the DFT domain. The first approach is based on the broadband model which computes the matrix multiplications using fast convolution techniques based on the overlap-save principle. The second method utilizes the narrowband model which assumes a complete decoupling of different frequency bins. Thus, first the broadband and narrowband model will be discussed in Section 3.4.1. Then this knowledge is used in Section 3.4.2 to derive an equivalent formulation of the time-domain BSS algorithms in the DFT domain by applying the broadband model and thus, exploiting the computational savings resulting from the fast convolutions. Subsequently, in Section 3.4.3 selective approximations are introduced which allow to exploit the narrowband efficiency for broadband convolutive BSS leading to several novel and well-known algorithms. By completely decoupling the frequency bins using further approximations several algorithms based on the narrowband model are obtained as well. Finally, in Section 3.4.4 the results are summarized.

3.4.1 Broadband and narrowband signal model

For supervised adaptive filtering [Hay02] in the DFT domain it was shown in [KB03] that the distinction between broadband and narrowband signal model is important. In this thesis we will extend this examination to BSS algorithms which belong to the class of unsupervised adaptive filtering.

In (3.31) an output signal vector $\mathbf{y}_q(n) = [y_q(n), \ldots, y_q(n-D+1)]^{\mathrm{T}}$ containing a memory of $D-1$ past values with descending sample index n was introduced to incorporate $D-1$ time-lags into the BSS optimization criterion (3.43). Moreover, in the BSS optimization criterion (3.43) an additional $N-1$ values $y_q(n+1), \ldots, y_q(n+N-1)$ with an ascending sample index are required inside the summation necessary for the block-based estimation of the pdfs. The ascending index corresponds to successive time instants where for each time instant a linear convolution is performed. The linear convolution of the sensor signals $x_p(n)$ with the demixing FIR filters $w_{pq,\kappa}$, $\kappa = 0, \ldots, L-1$ is given for the output signal $y_q(n)$ at time instant n as $y_q(n) = \sum_{p=1}^{P} \sum_{\kappa=0}^{L-1} w_{pq,\kappa} x_p(n-\kappa)$. If a block processing procedure is used, then the linear convolution necessary for the block-based estimation of the pdfs can be expressed in matrix notation for a block of N output signal

samples $\bar{\mathbf{y}}_q(mL)$ as

$$\bar{\mathbf{y}}_q(mL) = \sum_{p=1}^{P} \mathbf{U}_p^{\mathrm{T}}(m)\mathbf{w}_{pq} \tag{3.120}$$

where m denotes the block index, and the block output signal vector $\bar{\mathbf{y}}_q(mL)$ and the weights $\mathbf{w}_{pq}$ of the MIMO filter taps from the p-th sensor channel to the q-th output channel are given by

$$\begin{aligned} \bar{\mathbf{y}}_q(mL) &= [y_q(mL), y_q(mL+1), \ldots, y_q(mL+N-1)]^{\mathrm{T}}, & (3.121)\\ \mathbf{w}_{pq} &= [w_{pq,0}, w_{pq,1}, \ldots, w_{pq,L-1}]^{\mathrm{T}}. & (3.122) \end{aligned}$$

The $N \times L$ matrix $\mathbf{U}_p^{\mathrm{T}}(m)$ contains the sensor signal samples and exhibits a Toeplitz structure

$$\mathbf{U}_p^{\mathrm{T}}(m) = \begin{bmatrix} x_p(mL) & \ldots & \ldots & x_p(mL-L+1) \\ x_p(mL+1) & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ x_p(mL+N-1) & \ldots & \ldots & x_p(mL+N-L) \end{bmatrix}. \tag{3.123}$$

Based on this matrix formulation the broadband and narrowband signal model will be discussed in the following.

Broadband signal model

To obtain an efficient computation of the linear convolution (3.120) in the DFT domain we can exploit the property that any circulant $R \times R$ matrix can be diagonalized by the DFT matrix $\mathbf{F}_R$ [GL96, p. 202], where R denotes the transformation length and the element in the i-th row and k-th column of the DFT matrix is given as $[\mathbf{F}_R]_{ik} = \frac{1}{\sqrt{R}} e^{-j2\pi ik/R}$, $i, k \in \{0, \ldots, R-1\}$ [3]. However, the Toeplitz matrix $\mathbf{U}_p^{\mathrm{T}}(m)$ is in general neither square nor does it exhibit a circulant structure. Nevertheless, we can utilize the diagonalization property of the DFT matrix by utilizing the key idea that any Toeplitz matrix can be "embedded" in a circulant [GL96, p. 202]. This means that an $R \times R$ circulant matrix can be generated by extending the Toeplitz matrix properly. This leads to an increased size of the matrix and therefore, several "window matrices" are necessary to ensure the original size of the matrix. This procedure yields the overlap-save algorithm [OSB98] in matrix notation which allows an exact implementation of the linear convolution. Thus, this approach is based on the broadband signal model and will now be examined in detail.

[3] Due to the factor $\frac{1}{\sqrt{R}}$ the DFT matrix is unitary, i.e., $\mathbf{F}_R^{\mathrm{H}}\mathbf{F}_R = \mathbf{F}_R^{-1}\mathbf{F}_R = \mathbf{I}$ which simplifies the notation of DFT-domain algorithms. The additional multiplication which is introduced by $\frac{1}{\sqrt{R}}$ causes this factor to be often omitted in implementations which then leads to $\mathbf{F}_R^{\mathrm{H}}\mathbf{F}_R = R \cdot \mathbf{F}_R^{-1}\mathbf{F}_R = R \cdot \mathbf{I}$.

For a DFT length R the $N \times L$ Toeplitz matrix $\mathbf{U}_p^{\mathrm{T}}(m)$ in (3.123) is first expanded to a $R \times R$ circulant matrix $\mathbf{C}_{\mathbf{U}_p}(m)$ given by

$$\mathbf{C}_{\mathbf{U}_p}(m) = \begin{bmatrix} x_p(mL+N-R) & x_p(mL+N-1) & \dots & x_p(mL+N-R+1) \\ x_p(mL+N-R+1) & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ x_p(mL+N-1) & x_p(mL+N-2) & \dots & x_p(mL+N-R) \end{bmatrix}. \tag{3.124}$$

It can be seen that the first column is circularly downshifted by one sample for each subsequent column and that the original Toeplitz matrix $\mathbf{U}_p^{\mathrm{T}}(m)$ can be found in the lower left corner. To obtain $\mathbf{U}_p^{\mathrm{T}}(m)$ from the circulant matrix $\mathbf{C}_{\mathbf{U}_p}(m)$ two windowing matrices have to be introduced leading to

$$\mathbf{U}_p^{\mathrm{T}}(m) = \mathbf{W}_{N\times R}^{01_N}\mathbf{C}_{\mathbf{U}_p}\mathbf{W}_{R\times L}^{1_L0}, \tag{3.125}$$

where $\mathbf{W}_{N\times R}^{01_N}$ and $\mathbf{W}_{R\times L}^{1_L0}$ denote window matrices given as

$$\mathbf{W}_{N\times R}^{01_N} = [\mathbf{0}_{N\times R-N}, \mathbf{I}_{N\times N}], \tag{3.126}$$

$$\mathbf{W}_{R\times L}^{1_L0} = [\mathbf{I}_{L\times L}, \mathbf{0}_{L\times R-L}]^{\mathrm{T}}. \tag{3.127}$$

For the description of window matrices we use the conventions as outlined already in Section 3.3.3:

- The lower index of a matrix denotes its dimensions.
- P-channel matrices (as indicated by the size in the lower index) are partitioned into P single-channel window matrices.
- The upper index describes the positions of ones and zeros. Unity submatrices are always located at the upper left ('10') or lower right ('01') corners of the respective single-channel window matrix. The size of these clusters is indicated in subscript (e.g., '01_D').

In Fig. 3.10 we illustrated (3.125) by showing the circulant $\mathbf{C}_{\mathbf{U}_p}(m)$ together with the window matrices which constrain the circulant to yield the Toeplitz matrix $\mathbf{U}_p^{\mathrm{T}}(m)$. Now we can exploit the key property that the circulant can be diagonalized by the DFT matrix so that we obtain a diagonal matrix

$$\underline{\mathbf{U}}_p(m) = \mathbf{F}_R\mathbf{C}_{\mathbf{U}_p}(m)\mathbf{F}_R^{-1}, \tag{3.128}$$

where $\mathbf{F}_R$ denotes the DFT matrix with the transformation length R and the diagonal matrix $\underline{\mathbf{U}}_p(m)$ can be expressed by the first column of $\mathbf{C}_{\mathbf{U}_p}(m)$,

$$\underline{\mathbf{U}}_p(m) = \mathrm{Diag}\left\{ \mathbf{F}_R \begin{bmatrix} x_p(mL+N-R) \\ \vdots \\ x_p(mL+N-1) \end{bmatrix} \right\}. \tag{3.129}$$

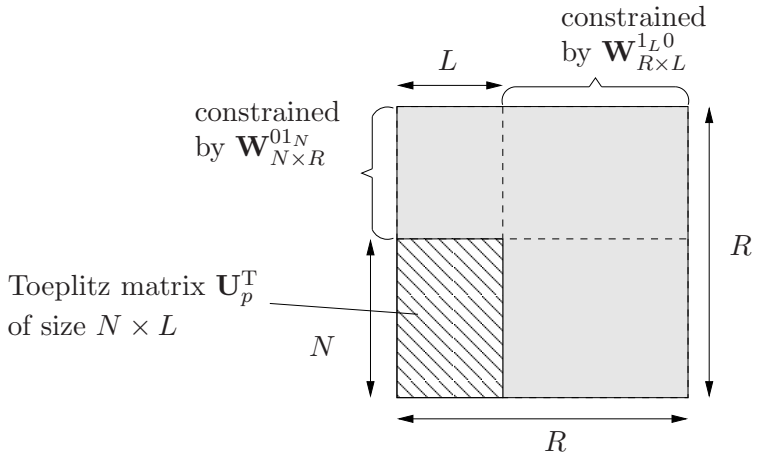


Figure 3.10: Illustration of (3.125) showing the relation between circulant matrix $\mathbf{C}_{\mathbf{U}_p}$ and Toeplitz matrix $\mathbf{U}_p^{\mathrm{T}}$.

The operator Diag$\{\mathbf{a}\}$ denotes a square matrix with the elements of vector $\mathbf{a}$ on its main diagonal. From (3.129) it can be seen that $\underline{\mathbf{U}}_p(m)$ has the DFT values of the first column of $\mathbf{C}_{\mathbf{U}_p}(m)$ on its main diagonal. In this thesis we use the convention that DFT-domain quantities are denoted by an underline. Relations (3.125) and (3.128) are the prerequisites necessary for formulating the linear convolution (3.120) equivalently in the DFT domain. Inserting (3.125) and (3.128) into (3.120) leads to

$$\bar{\mathbf{y}}_q(mL) = \sum_{p=1}^{P} \mathbf{W}_{N\times R}^{01_N} \mathbf{F}_R^{-1} \underline{\mathbf{U}}_p(m) \mathbf{F}_R \mathbf{W}_{R\times L}^{1_L 0} \mathbf{w}_{pq}. \tag{3.130}$$

By identifying the DFT representation $\underline{\mathbf{w}}_{pq}$ of the demixing filters $\mathbf{w}_{pq}$ padded with $R-L$ zeros as

$$\begin{aligned} \underline{\mathbf{w}}_{pq} &= \mathbf{F}_R[\mathbf{w}_{pq}^{\mathrm{T}}, \mathbf{0}_{R-L\times 1}]^{\mathrm{T}} \\ &= \mathbf{F}_R \mathbf{W}_{R\times L}^{1_L 0} \mathbf{w}_{pq}, \end{aligned} \tag{3.131}$$

we can express the linear filtering operation (3.120), yielding a block of N output samples, equivalently by using DFT variables as

$$\bar{\mathbf{y}}_q(mL) = \sum_{p=1}^{P} \mathbf{W}_{N\times R}^{01_N} \mathbf{F}_R^{-1} \underline{\mathbf{U}}_p(m) \underline{\mathbf{w}}_{pq}. \tag{3.132}$$

According to (3.132) the output of the linear convolution is obtained by multiplying the DFT values on the main diagonal of $\underline{\mathbf{U}}_p(m)$, containing the transformation of R input signal samples (3.129), with the DFT values $\underline{\mathbf{w}}_{pq}$ of the zero-padded FIR filter $\mathbf{w}_{pq}$. As $\underline{\mathbf{U}}_p$ is a diagonal matrix, the matrix-vector multiplication can be calculated as a scalar multiplication in each frequency bin. The result is transformed back into the time domain

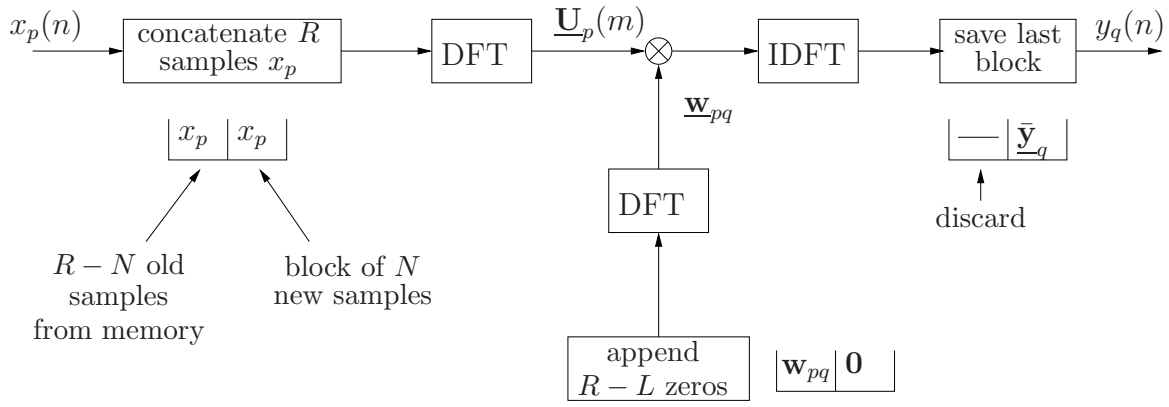


Figure 3.11: Illustration of the overlap-save method expressed in matrix notation in (3.132).

and the first $R-N$ samples are discarded due to the window matrix $\mathbf{W}_{N\times R}^{01_N}$. It should be pointed out that for an output signal block of length N and an FIR filter of length L the DFT length has to be chosen to $R \geq N+L-1$ to avoid circular convolution effects. For this choice (3.132) describes the *overlap-save algorithm in matrix notation*, resulting in an efficient equivalent implementation of the linear convolution. This is denoted as a DFT-implementation based on the broadband signal model. A visualization of (3.132) is given in Fig. 3.11. The matrix formulation in (3.132) is important as it allows to transform any broadband time-domain algorithm equivalently to the DFT-domain as will be shown in Section 3.4.2.

Narrowband signal model

The narrowband model is obtained if no output signal samples are discarded which means that the block length N is set to $N=R$ yielding an output signal vector $\bar{\mathbf{y}}_q(mL)$ of length R. Thus, circular convolution effects are tolerated and the linear convolution (3.132) is approximated by a circular convolution given as

$$\bar{\mathbf{y}}_q(mL) = \sum_{p=1}^{P} \mathbf{F}_R^{-1} \underline{\mathbf{U}}_p(m) \underline{\mathbf{w}}_{pq}. \tag{3.133}$$

It can be observed in (3.133) that due to the choice $N=R$ the window matrix $\mathbf{W}_{N\times R}^{01_N}$ reduces to an identity matrix. By approximating the window matrix $\mathbf{W}_{N\times R}^{01_N}$ it is assumed that the input signal $x_p(n)$ is a periodic signal with period R/κ, $\kappa \in \{1,2,\ldots,R\}$ so that the extension of the $N\times L$ input signal Toeplitz matrix $\mathbf{U}_p^{\mathrm{T}}(mL)$, defined in (3.123), to a $R\times R$ matrix would already exhibit a circulant structure [KB03]. This means that $x_p(n)$ is described by a finite Fourier series $x_p(n) = \sum_{i=0}^{R-1} c_i e^{j\frac{2\pi i n}{R}}$, i.e., by at most R complex exponentials. For audio signals, however, the input signals are in general not periodic and therefore, the DFT length R should be chosen much larger than the filter length L

to obtain a valid approximation and thus, to reduce the circular convolution effects. In [Gra72] it was shown, based on the properties of Toeplitz matrices and circulant matrices, that for the case $R \to \infty$ the broadband and narrowband signal model are equivalent. The equivalence is based on the Szegö theorem [GS58] and will be discussed in detail in Section 3.4.3 where the narrowband signal model is selectively introduced to obtain efficient algorithms. This selective approximation was also applied to supervised adaptive filtering algorithms in [BBK03, KB03, BBK05].

General method for designing efficient broadband convolutive BSS algorithms in the DFT domain

Based on the distinction between broadband and narrowband signal model, a general method for designing efficient broadband convolutive BSS algorithms can be given:

Step 1: A time-domain convolutive BSS update equation is expressed in terms of DFT variables using the broadband signal model (3.132). This is achieved by the following procedure:

At first, the property that a Toeplitz matrix can be embedded in a circulant matrix is exploited. Thus, the update equation has to be expressed in terms of Toeplitz matrices. These are then all replaced by square circulant matrices together with appropriate window matrices. The window matrices are necessary as the update equation only contains Toeplitz matrices and thus, the window matrices have to assure that a Toeplitz matrix is again obtained from a circulant matrix as shown in (3.125).

Secondly, the circulant matrices are diagonalized by the DFT matrix. This results in the desired DFT-domain representation of the time-domain quantities.

Step 2: To obtain efficient versions of the broadband DFT-domain algorithm the various constraints resulting from the application of Step 1 to all Toeplitz matrices are examined. A *selective approximation* of the constraints by selectively introducing the narrowband signal model (3.133) leads to efficient implementations without completely decoupling the individual DFT bins.

Instead of considering a broadband time-domain update equation in Step 1, it is also possible to use a broadband time-domain optimization criterion, such as, e.g., (3.43). Then, the optimization criterion is expressed analogously in terms of DFT matrices and window functions. Subsequently, the optimization criterion is minimized with respect to the *DFT-domain filter coefficients* $\underline{\mathbf{w}}_{pq}$ instead of the time-domain coefficients $\mathbf{w}_{pq}$. This procedure yields the same broadband DFT-domain formulation of the update equation as obtained in Step 1 of the procedure above. In both cases, it is especially important for BSS algorithms to selectively introduce the narrowband signal model in Step 2 only for some

of the Toeplitz matrices appearing in the update equations. If all window matrices are ignored then purely narrowband algorithms are obtained as will be shown in Section 3.4.3. The drawback is that, then, also the permutation ambiguity (see Section 2.4) appears in each frequency bin independently.

3.4.2 Equivalent formulation of broadband algorithms in the DFT domain

In the previous section the basis to describe broadband convolutive BSS algorithms in the DFT domain was given. In this section we will use this method to express the HOS and SOS realizations of the generic natural gradient algorithm (3.64) equivalently in the DFT domain.

First, an extended signal model based on matrix notation will be introduced which allows a compact notation for the linear convolution yielding the output signal samples. Then, the iterative update rule is given for the DFT-domain demixing matrix. Subsequently, the HOS realization (3.105) based on pdfs of multivariate spherically invariant random processes (SIRPs) and later on also the second-order statistics (SOS) realization (3.112) will be expressed equivalently in the DFT domain. The resulting expressions will be the basis for introducing selective narrowband approximations in Section 3.4.3.

3.4.2.1 Signal model expressed by Toeplitz matrices

The optimization criterion and the update equations in Section 3.3 were formulated using the signal model $\mathbf{y}(n) = \mathbf{W}^{\mathrm{T}}\mathbf{x}(n)$ given in (3.33) which allows the introduction of a memory of $D-1$ time-lags for each channel. The generation of these D output signal samples $y_q(n), y_q(n-1)\ldots, y_q(n-D+1)$ for each channel $q = 1, \ldots, P$ could be expressed conveniently as a matrix-vector product by expressing the demixing FIR filters $\mathbf{w}_{pq}$ as a Sylvester matrix $\mathbf{W}_{pq}$ defined in (3.32).

Additionally, the samples $y_q(n), \ldots, y_q(n+N-1)$ are used for the block-based estimation of the D-variate pdfs in the optimization criterion. It was shown in (3.120) that the linear convolution yielding these N output signal values can also be formulated in matrix-vector notation. Thus, in the optimization criterion (3.43) a total of $N+D-1$ output signal samples $y_q(n-D+1), \ldots, y_q(n+N-1)$ is required for each block and for each channel.

A combination of (3.33) and (3.120) is possible by introducing a compact signal model in matrix notation which provides all $N+D-1$ output signal samples required by the optimization criterion and is given as

$$\mathbf{Y}(m) = \mathbf{W}^{\mathrm{H}}\mathbf{X}(m). \tag{3.134}$$

It should be noted that due to the block-processing all signal quantities depend on the block index m. Moreover, to simplify the notation of the DFT-domain formulation we used the hermitian operator instead of the transpose operator in the definition of the signal model. This is permitted as we are only dealing with real-valued time-domain signals. The matrices in (3.134) are given as

$$\mathbf{Y}(m) = [\mathbf{Y}_1^{\mathrm{T}}(m), \ldots, \mathbf{Y}_P^{\mathrm{T}}(m)]^{\mathrm{T}}, \tag{3.135}$$

$$\mathbf{X}(m) = [\mathbf{X}_1^{\mathrm{T}}(m), \ldots, \mathbf{X}_P^{\mathrm{T}}(m)]^{\mathrm{T}}. \tag{3.136}$$

The channel-wise $D \times N$ submatrices $\mathbf{Y}_q(m)$ contain all $D+N-1$ output signal samples and were already defined in (3.69) as

$$\mathbf{Y}_p(m) = \begin{bmatrix} y_p(mL) & \cdots & y_p(mL+N-1) \\ y_p(mL-1) & \ddots & y_p(mL+N-2) \\ \vdots & \ddots & \vdots \\ y_p(mL-D+1) & \cdots & y_p(mL-D+N) \end{bmatrix}.$$

The $2L \times N$ submatrices $\mathbf{X}_p(m)$ contain $2L+N-1$ sensor signal samples [4] and are given as

$$\mathbf{X}_p(m) = \begin{bmatrix} x_p(mL) & \cdots & x_p(mL+N-1) \\ x_p(mL-1) & \ddots & x_p(mL+N-2) \\ \vdots & \ddots & \vdots \\ x_p(mL-2L+1) & \cdots & x_p(mL-2L+N) \end{bmatrix} \tag{3.137}$$

and the demixing filter matrix $\mathbf{W}$ consists of Sylvester submatrices as defined in (3.36). It should be noted that the resulting output signal matrix $\mathbf{Y}(m)$ was already used in Section 3.3.5 in the definition of the short-time HOS cross-relation and SOS cross-correlation matrices based on the covariance method.

As desired, all matrices in the signal model (3.134) exhibit a Toeplitz structure which allows for an equivalent formulation in the DFT domain according to the procedure presented in the previous section. Therefore, after formulating the iterative update rule in the DFT domain, we will express the various special cases derived in Section 3.3 using the output signal matrix $\mathbf{Y}(m)$ and then subsequently transform them to the DFT domain.

3.4.2.2 Iterative update rule in the DFT domain

To obtain a broadband update procedure in the DFT domain, the iterative time-domain update equation (3.58) has to be transformed to the DFT domain. All demixing filters

[4] Actually, already $L+D+N-2$ samples would be sufficient. However, with regard to a concise DFT-domain notation, the Sylvester matrices $\mathbf{W}_{pq}$ used in the signal model (3.33) were defined in (3.32) with appended rows of zeros leading to a dimension of $2L \times D$ instead of $L+D-1 \times D$. Thereby, the dimension of $\mathbf{X}_p$ is given as $2L \times N$ and thus, contains $2L+N-1$ sensor signal samples.

$\mathbf{w}_{pq}(m)$ can be expressed in the DFT domain by first appending $R-L$ zeros and then transforming the zero-padded filter by a DFT of length R. The DFT length R has to be chosen at least to $R \geq L$. Later on, when deriving the broadband DFT-domain algorithms it will be seen that depending on the estimation of the cross-relation matrices based on the covariance or correlation method different restrictions are imposed on the DFT length R. It was pointed out in (3.131) that the transformation of the demixing FIR filters is written in matrix notation as $\underline{\mathbf{w}}_{pq} = \mathbf{F}_R \mathbf{W}_{R\times L}^{1_L 0} \mathbf{w}_{pq}$. A combination of all channels leads to the DFT-domain demixing matrix

$$\underline{\check{\mathbf{W}}} = \begin{bmatrix} \underline{\mathbf{w}}_{11} & \cdots & \underline{\mathbf{w}}_{1P} \\ \vdots & \ddots & \vdots \\ \underline{\mathbf{w}}_{P1} & \cdots & \underline{\mathbf{w}}_{PP} \end{bmatrix}, \tag{3.138}$$

which combines all DFT-domain filters $\underline{\mathbf{w}}_{pq}$ obtained by zero-padding and transformation of the individual time-domain filters $\mathbf{w}_{pq}$. The relationship between the DFT-domain demixing matrix $\underline{\check{\mathbf{W}}}$ and the time-domain counterpart $\check{\mathbf{W}}$ (3.4) is given as

$$\underline{\check{\mathbf{W}}} = \mathbf{V}_{PL\times PR}^{\mathrm{H}} \check{\mathbf{W}}, \tag{3.139}$$

with the matrix $\mathbf{V}_{PL\times PR}^{\mathrm{H}}$ denoting the channel-wise transformation of the window matrix $\mathbf{W}_{R\times L}^{1_L 0}$ into the DFT domain

$$\mathbf{V}_{PL\times PR}^{\mathrm{H}} = \mathrm{Bdiag}\left\{\mathbf{F}_R \mathbf{W}_{R\times L}^{1_L 0}, \ldots, \mathbf{F}_R \mathbf{W}_{R\times L}^{1_L 0}\right\}. \tag{3.140}$$

Using the DFT-domain demixing matrix (3.138) the update procedure can be written in the DFT domain analogously to the time-domain update (3.58) as

$$\begin{aligned} \underline{\check{\mathbf{W}}}(m) &= \underline{\check{\mathbf{W}}}(m-1) - \mu \mathbf{V}_{PL\times PR}^{\mathrm{H}} \Delta\check{\mathbf{W}}(m) \\ &= \underline{\check{\mathbf{W}}}(m-1) - \mu \Delta\underline{\check{\mathbf{W}}}(m). \end{aligned} \tag{3.141}$$

In the following the various time-domain updates $\Delta\check{\mathbf{W}}$ derived in Section 3.3 will be expressed by using DFT-domain variables which are developed in the next section, so that together with (3.141) an update procedure solely based on DFT-domain quantities is obtained.

3.4.2.3 DFT representation of the Sylvester matrix W and the output signal Toeplitz matrices Y and $\tilde{\mathbf{Y}}$

The time-domain updates in Section 3.3.7 following from the generic natural gradient update (3.64) can be expressed in terms of the Sylvester matrix $\mathbf{W}$ and the output signal matrix $\mathbf{Y}$. Using the matrix $\mathbf{Y}$ corresponds to a block-based estimation of the cross-relation and cross-correlation matrices using the covariance method (see also Section 3.3.5). Especially for the algorithms based on SOS also the estimation of the cross-correlation matrices

using the correlation method is important as it leads to lower computational complexity. In this case the matrix $\mathbf{Y}$ is replaced by $\tilde{\mathbf{Y}}$ defined in (3.84). Thus, for an equivalent formulation of the time-domain algorithms in the DFT domain it is required that a transformation of all three matrices $\mathbf{W}$, $\mathbf{Y}$, and $\tilde{\mathbf{Y}}$ to the DFT domain is performed. This will be developed in the following.

DFT representation of demixing filter matrix $\mathbf{W}(m)$

First, the channel-wise matrix $\mathbf{W}_{pq}(m)$ will be transformed into the DFT domain. Thus, the $2L \times D$ Toeplitz matrix $\mathbf{W}_{pq}(m)$ has to be extended to a circulant $\mathbf{C}_{\mathbf{W}_{pq}}(m)$ of size $R \times R$ with $R \geq 2L$ (note that in general $1 \leq D \leq L$). The relationship between the original matrix $\mathbf{W}_{pq}(m)$ and the circulant $\mathbf{C}_{\mathbf{W}_{pq}}(m)$ is given by

$$\mathbf{W}_{pq}(m) = \mathbf{W}_{2L\times R}^{1_{2L}0}\mathbf{C}_{\mathbf{W}_{pq}}(m)\mathbf{W}_{R\times D}^{1_D 0}, \tag{3.142}$$

where $\mathbf{W}_{2L\times R}^{1_{2L}0}$ and $\mathbf{W}_{R\times D}^{1_D 0}$ denote window matrices given as

$$\mathbf{W}_{2L\times R}^{1_{2L}0} = \left[\mathbf{I}_{2L\times 2L}, \mathbf{0}_{2L\times R-2L}\right], \tag{3.143}$$

$$\mathbf{W}_{R\times D}^{1_D 0} = \left[\mathbf{I}_{D\times D}, \mathbf{0}_{D\times R-D}\right]^{\mathrm{T}}. \tag{3.144}$$

An illustration of (3.142) showing the application of the window matrices to obtain the Toeplitz matrix $\mathbf{W}_{pq}(m)$ from the circulant matrix $\mathbf{C}_{\mathbf{W}_{pq}}(m)$ is given in Fig. 3.12 [5]. The key property of a circulant, namely that it can be diagonalized by the DFT matrix, is now exploited yielding

$$\mathbf{C}_{\mathbf{W}_{pq}}(m) = \mathbf{F}_R^{-1}\underline{\mathbf{W}}_{pq}(m)\mathbf{F}_R, \tag{3.145}$$

with

$$\underline{\mathbf{W}}_{pq}(m) = \mathrm{Diag}\left\{\mathbf{F}_R\left[\mathbf{w}_{pq}^{\mathrm{T}}(m), \mathbf{0}_{R-L\times 1}^{\mathrm{T}}\right]^{\mathrm{T}}\right\}. \tag{3.146}$$

Thus, $\underline{\mathbf{W}}_{pq}(m)$ has the DFT values of the first column of $\mathbf{C}_{\mathbf{W}_{pq}}(m)$ on its main diagonal. From (3.145) it can be seen that these values correspond to the DFT values of the zero-padded filter $\mathbf{w}_{pq}(m)$. By inserting (3.145) into (3.142) the relationship between the time-domain quantity $\mathbf{W}_{pq}(m)$ and the DFT-domain quantity $\underline{\mathbf{W}}_{pq}(m)$ is obtained

$$\mathbf{W}_{pq}(m) = \mathbf{W}_{2L\times R}^{1_{2L}0}\mathbf{F}_R^{-1}\underline{\mathbf{W}}_{pq}(m)\mathbf{F}_R\mathbf{W}_{R\times D}^{1_D 0}. \tag{3.147}$$

A combination of all channels yields the DFT-domain representation

$$\mathbf{W}(m) = \mathbf{V}_{2PL\times PR}\underline{\mathbf{W}}(m)\left(\mathbf{V}_{PD\times PR}^{1_D 0}\right)^{\mathrm{H}}, \tag{3.148}$$

with the channel-wise transformation of the window matrices given as

$$\mathbf{V}_{PD\times PR}^{1_D 0} = \mathrm{Bdiag}\left\{\mathbf{W}_{D\times R}^{1_D 0}\mathbf{F}_R^{-1}\ldots, \mathbf{W}_{D\times R}^{1_D 0}\mathbf{F}_R^{-1}\right\}, \tag{3.149}$$

$$\mathbf{V}_{2PL\times PR} = \mathrm{Bdiag}\left\{\mathbf{W}_{2L\times R}^{1_{2L}0}\mathbf{F}_R^{-1}\ldots, \mathbf{W}_{2L\times R}^{1_{2L}0}\mathbf{F}_R^{-1}\right\}. \tag{3.150}$$

[5] As can be seen in (3.32), the Sylvester matrix $\mathbf{W}_{pq}$ contains $L - D + 1$ rows of zeros. Thus, even in the case $D = L$ one row of zeros is retained and hence, strictly speaking, the hatched area does not hit the lower right corner of $\mathbf{W}_{pq}$ in Fig. 3.12.

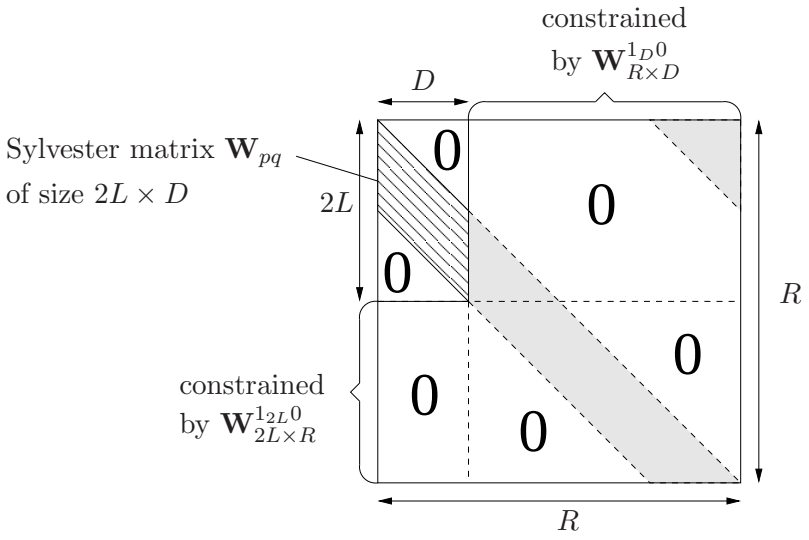


Figure 3.12: Illustration of (3.142) showing the relation between circulant matrix $\mathbf{C}_{\mathbf{W}_{pq}}$ and Toeplitz matrix $\mathbf{W}_{pq}$ for the case $D = L$.

DFT representation of output signal matrix $\mathbf{Y}(m)$

If the *covariance method* is used, then the output signal matrix $\mathbf{Y}(m)$ defined in (3.72) is required. Similarly to the demixing filter matrix $\mathbf{W}(m)$, the matrix $\mathbf{Y}(m)$ is transformed channel-wise into the DFT domain by first expressing the $D \times N$ Toeplitz matrix $\mathbf{Y}_p(m)$ defined in (3.69) in terms of the $R \times R$ circulant matrix $\mathbf{C}_{\mathbf{Y}_p}(m)$ resulting in

$$\mathbf{Y}_p(m) = \mathbf{W}_{D\times R}^{01_D} \mathbf{C}_{\mathbf{Y}_p}(m) \mathbf{W}_{R\times N}^{1_N 0} \tag{3.151}$$

with the window matrices

$$\mathbf{W}_{D\times R}^{01_D} = [\mathbf{0}_{D\times R-D}, \mathbf{I}_{D\times D}], \tag{3.152}$$

$$\mathbf{W}_{R\times N}^{1_N 0} = [\mathbf{I}_{N\times N}, \mathbf{0}_{N\times R-N}]^{\mathrm{T}}. \tag{3.153}$$

An illustration of (3.151) is given in Fig. 3.13. Utilizing again the diagonalization property of the DFT analogously to (3.145) yields the relation between the time-domain quantity $\mathbf{Y}_p$ and DFT-domain values $\underline{\mathbf{Y}}_p$

$$\mathbf{Y}_p(m) = \mathbf{W}_{D\times R}^{01_D} \mathbf{F}_R^{-1} \underline{\mathbf{Y}}_p(m) \mathbf{F}_R \mathbf{W}_{R\times N}^{1_N 0}. \tag{3.154}$$

The DFT-domain variable $\underline{\mathbf{Y}}_p(m)$ represents a diagonal matrix where the values on the diagonal are given as the DFT of the first column of $\mathbf{C}_{\mathbf{Y}_p}(m)$ (see Fig. 3.13). Therefore, $\underline{\mathbf{Y}}_p(m)$ is given as

$$\underline{\mathbf{Y}}_p(m) = \mathrm{diag}\left\{\mathbf{F}_R[0, \ldots, 0, y_p(mL - D + 1), \ldots, y_p(mL + N - 1)]^{\mathrm{T}}\right\}. \tag{3.155}$$

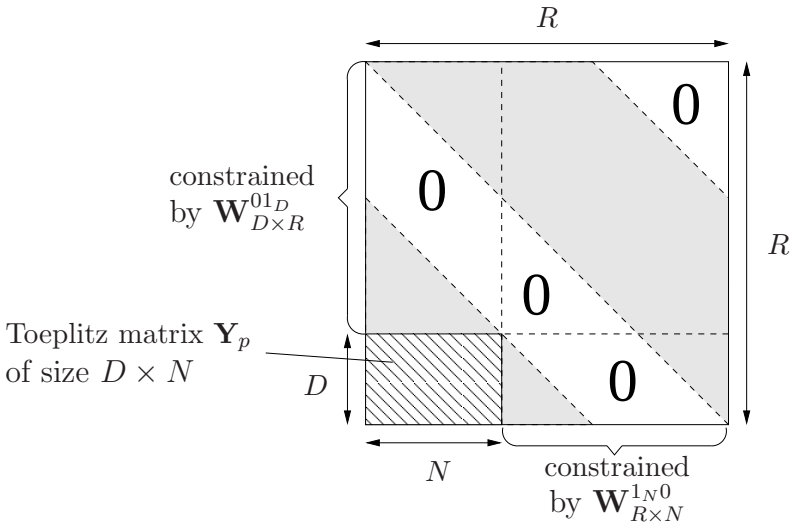


Figure 3.13: Illustration of (3.151) showing the relation between circulant matrix $\mathbf{C}_{\mathbf{Y}_p}$ and Toeplitz matrix $\mathbf{Y}_p$.

A combination of all channels $p = 1, \ldots, P$ results in

$$\mathbf{Y}(m) = \mathbf{V}_{PD\times PR}\underline{\mathbf{Y}}(m)\mathbf{F}_R\mathbf{W}_{R\times N}^{1_N 0}, \tag{3.156}$$

with the constraint matrix $\mathbf{V}_{PD\times PR}$ given as

$$\mathbf{V}_{PD\times PR} = \mathrm{Bdiag}\left\{\mathbf{W}_{D\times R}^{01_D}\mathbf{F}_R^{-1}\ldots, \mathbf{W}_{D\times R}^{01_D}\mathbf{F}_R^{-1}\right\}. \tag{3.157}$$

DFT representation of output signal matrix $\tilde{\mathbf{Y}}(m)$

If in contrast to the covariance method, the *correlation method* is used for estimating the cross-correlation matrices, then, instead of $\mathbf{Y}_p(m)$, the $D \times N + D - 1$ Toeplitz matrix $\tilde{\mathbf{Y}}_p(m)$ defined in (3.82) has to be expressed in terms of DFT quantities. Again, $\tilde{\mathbf{Y}}(m)$ is embedded into an $R \times R$ circulant matrix $\mathbf{C}_{\tilde{\mathbf{Y}}_p}$, which is subsequently diagonalized:

$$\begin{aligned}\tilde{\mathbf{Y}}_p(m) &= \mathbf{W}_{D\times R}^{01_D}\mathbf{C}_{\tilde{\mathbf{Y}}_p}(m)\mathbf{W}_{R\times N+D}^{1_{N+D}0} \\ &= \mathbf{W}_{D\times R}^{01_D}\mathbf{F}_R^{-1}\underline{\tilde{\mathbf{Y}}}_p(m)\mathbf{F}_R\mathbf{W}_{R\times N+D}^{1_{N+D}0},\end{aligned} \tag{3.158}$$

where

$$\mathbf{W}_{R\times N+D-1}^{1_{N+D-1}0} = [\mathbf{I}_{N+D-1\times N+D-1}, \mathbf{0}_{N+D-1\times R-N-D+1}]^{\mathrm{T}}. \tag{3.159}$$

The relation between circulant and Toeplitz matrix is illustrated in Fig. 3.14. The entries of the diagonal matrix $\underline{\tilde{\mathbf{Y}}}_p$ are given as the DFT of the first column of $\mathbf{C}_{\tilde{\mathbf{Y}}_p}$

$$\underline{\tilde{\mathbf{Y}}}_p(m) = \mathrm{Diag}\left\{\mathbf{F}_R[0, \ldots, 0, y_p(mL + N - 1), \ldots, y_q(mL), 0, \ldots, 0]^{\mathrm{T}}\right\}. \tag{3.160}$$

Comparing (3.155) with (3.160) it can be seen that in contrast to the covariance method where $N + D - 1$ values are transformed to the DFT domain, the correlation method

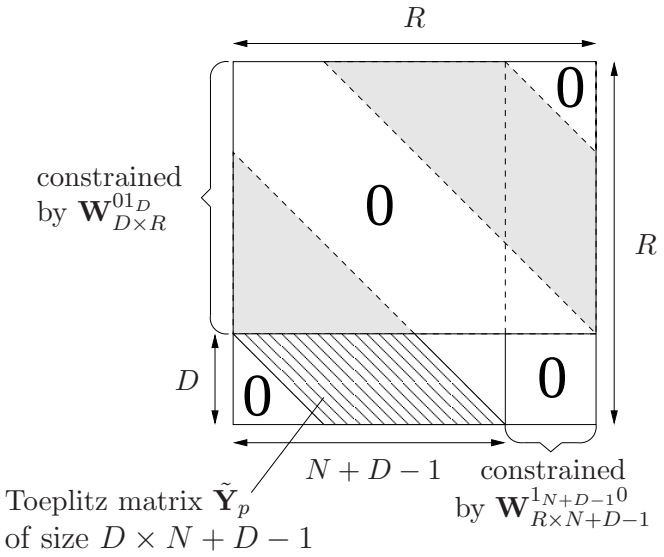


Figure 3.14: Illustration of (3.158) showing the relation between circulant matrix $\mathbf{C}_{\tilde{\mathbf{Y}}_p}$ and Toeplitz matrix $\tilde{\mathbf{Y}}_p$.

only requires N output signal values y_p. This was also pointed out in the discussion of the covariance and correlation method in Section 3.3.5. A combination of all channels $p = 1, \ldots, P$ yields

$$\tilde{\mathbf{Y}}(m) = \mathbf{V}_{PD\times PR}\underline{\tilde{\mathbf{Y}}}(m)\mathbf{F}_R\mathbf{W}^{1_{N+D-1}0}_{R\times N+D-1}, \tag{3.161}$$

where $\mathbf{V}_{PD\times PR}$ is defined in (3.157).

3.4.2.4 Higher-order statistics realization based on multivariate pdfs

After transforming the matrices $\mathbf{W}$ and $\mathbf{Y}$ in the previous section into the DFT domain, we will now show the expression of the HOS-SIRP natural gradient by these DFT-domain variables. In Section 3.3.4 the nonholonomic natural gradient (3.64) based on multivariate pdfs was derived. In Section 3.3.7.1 the SIRP model (3.95) was introduced to allow for an efficient estimation of multivariate pdfs. This led to the time-domain nonholonomic HOS-SIRP natural gradient given in (3.105) as

$$\Delta\check{\mathbf{W}}(m) = 2\sum_{i=0}^{\infty}\beta(i,m)\mathcal{SC}\left\{\mathbf{W}(i)\ \mathrm{boff}\{\mathbf{R}_{\mathbf{y}\phi(\mathbf{y})}(i)\}\ \mathrm{bdiag}^{-1}\{\mathbf{R}_{\mathbf{yy}}(i)\}\right\},$$

with the operator $\mathrm{boff}\{\mathbf{A}\} = \mathbf{A} - \mathrm{bdiag}\{\mathbf{A}\}$. The cross-correlation matrix $\mathbf{R}_{\mathbf{yy}}$ and the nonlinearly weighted cross-correlation matrix $\mathbf{R}_{\mathbf{y}\phi(\mathbf{y})}$ consist of the channel-wise $D \times D$ matrices defined in (3.68) and (3.107), respectively, as

$$\begin{aligned}\mathbf{R}_{\mathbf{y}_p\mathbf{y}_q}(i) &= \frac{1}{N}\mathbf{Y}_p(i)\mathbf{Y}_q^{\mathrm{H}}(i)\\ \mathbf{R}_{\mathbf{y}_p\phi(\mathbf{y}_q)}(i) &= \frac{1}{N}\mathbf{Y}_p(i)\mathbf{\Lambda}_q^{\mathrm{H}}(i)\mathbf{Y}_q^{\mathrm{H}}(i)\end{aligned}$$

where the $N \times N$ diagonal matrix $\mathbf{\Lambda}_q$ is given as

$$\mathbf{\Lambda}_q(i) = \phi_{y_q,D}\left(\operatorname{diag}\left\{\mathbf{Y}_q^{\mathrm{H}}(i)\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}^{-1}(i)\mathbf{Y}_q(i)\right\}\right),$$

and $\phi_{y_q,D}\left(\operatorname{diag}\{\mathbf{A}\}\right)$ is the scalar SIRP score function which is applied element-wise to the elements a_{jj} on the diagonal of $\mathbf{A}$. It was defined in (3.98) as

$$\phi_{y_q,D}\left(a_{jj}\right) = -\frac{\partial(\log f_{y_q,D}(a_{jj}))}{\partial a_{jj}} = -\frac{\frac{\partial f_{y_q,D}(a_{jj})}{\partial a_{jj}}}{f_{y_q,D}(a_{jj})}.$$

The $PD \times N$ matrix $\mathbf{Y}(i) = [\mathbf{Y}_1(i), \ldots, \mathbf{Y}_P(i)]^{\mathrm{T}}$ leads to an estimation of the cross-relation matrices $\mathbf{R}_{\mathbf{yy}}$ and $\mathbf{R}_{\mathbf{y}\phi(\mathbf{y})}$ via the covariance method (see also Section 3.3.5). Due to its generality, the covariance method will be used first for the derivation of the DFT-domain formulation. The differences between covariance and correlation method in the DFT domain, which are especially important for the implementation of SOS BSS algorithms, will be discussed in Section 3.4.2.5.

A formulation of the cross-correlation matrix $\mathbf{R}_{\mathbf{yy}}$ using DFT quantities can be achieved by inserting (3.154) in the definition of $\mathbf{R}_{\mathbf{y}_p\mathbf{y}_q}$ yielding

$$\mathbf{R}_{\mathbf{y}_p\mathbf{y}_q}(i) = \frac{1}{N}\mathbf{W}_{D\times R}^{01_D}\mathbf{F}_R^{-1}\underline{\mathbf{Y}}_p(i)\mathbf{G}_{R\times R}^{1_N0}\underline{\mathbf{Y}}_q^{\mathrm{H}}(i)\mathbf{F}_R\mathbf{W}_{R\times D}^{01_D}, \tag{3.162}$$

with the constraint matrix given as

$$\begin{aligned}\mathbf{G}_{R\times R}^{1_N0} &= \mathbf{F}_R\mathbf{W}_{R\times N}^{1_N0}\mathbf{W}_{N\times R}^{1_N0}\mathbf{F}_R^{-1} \\ &= \mathbf{F}_R\mathbf{W}_{R\times R}^{1_N0}\mathbf{F}_R^{-1}.\end{aligned} \tag{3.163}$$

A combination of all cross- and auto-correlations leads to

$$\mathbf{R}_{\mathbf{yy}}(i) = \frac{1}{N}\mathbf{V}_{PD\times PR}\underline{\mathbf{Y}}(i)\mathbf{G}_{R\times R}^{1_N0}\underline{\mathbf{Y}}^{\mathrm{H}}(i)\mathbf{V}_{PD\times PR}^{\mathrm{H}} \tag{3.164}$$

with $\mathbf{V}_{PD\times PR}$ defined in (3.157).

The nonlinearly weighted cross-correlation matrix was defined in (3.107) as $\mathbf{R}_{\mathbf{y}_p\phi(\mathbf{y}_q)} = \frac{1}{N}\mathbf{Y}_p\mathbf{\Lambda}_q^{\mathrm{H}}\mathbf{Y}_q^{\mathrm{H}}$. A DFT representation of $\mathbf{R}_{\mathbf{y}_p\phi(\mathbf{y}_q)}$ is obtained by again inserting (3.154) which yields

$$\mathbf{R}_{\mathbf{y}_p\phi(\mathbf{y}_q)}(i) = \frac{1}{N}\mathbf{W}_{D\times R}^{01_D}\mathbf{F}_R^{-1}\underline{\mathbf{Y}}_p(i)\mathbf{F}_R\mathbf{W}_{R\times N}^{1_N0}\mathbf{\Lambda}_q^{\mathrm{H}}(i)\mathbf{W}_{N\times R}^{1_N0}\mathbf{F}_R^{-1}\underline{\mathbf{Y}}_q^{\mathrm{H}}(i)\mathbf{F}_R\mathbf{W}_{R\times D}^{01_D}. \tag{3.165}$$

To allow for a concise formulation of all channels we define the channel-wise multiplication of $\mathbf{\Lambda}_q^{\mathrm{H}}$ and $\underline{\mathbf{Y}}_q^{\mathrm{H}}$ as

$$\underline{\mathbf{Y}}_\phi^{\mathrm{H}} = \left[\mathbf{F}_R\mathbf{W}_{R\times N}^{1_N0}\mathbf{\Lambda}_1^{\mathrm{H}}\mathbf{W}_{N\times R}^{1_N0}\mathbf{F}_R^{-1}\underline{\mathbf{Y}}_1^{\mathrm{H}}, \ldots, \mathbf{F}_R\mathbf{W}_{R\times N}^{1_N0}\mathbf{\Lambda}_P^{\mathrm{H}}\mathbf{W}_{N\times R}^{1_N0}\mathbf{F}_R^{-1}\underline{\mathbf{Y}}_P^{\mathrm{H}}\right]. \tag{3.166}$$

The transformation of the $N \times N$ diagonal matrix $\mathbf{\Lambda}_q(i)$ in (3.166) is given by inserting the time-frequency relations (3.154) and (3.162) leading to

$$\begin{aligned}\mathbf{\Lambda}_q(i) &= \phi_{y_q,D}\Big(\operatorname{diag}\Big\{\mathbf{W}_{N\times R}^{1_N 0}\mathbf{F}_R^{-1}\underline{\mathbf{Y}}_q^{\mathrm{H}}(i)\mathbf{F}_R\mathbf{W}_{R\times D}^{01_D}\cdot \\ &\quad\cdot\left(\frac{1}{N}\mathbf{W}_{D\times R}^{01_D}\mathbf{F}_R^{-1}\underline{\mathbf{Y}}_q^{\mathrm{H}}(i)\mathbf{G}_{R\times R}^{1_N 0}\underline{\mathbf{Y}}_q(i)\mathbf{F}_R\mathbf{W}_{R\times D}^{01_D}\right)^{-1}\mathbf{W}_{D\times R}^{01_D}\mathbf{F}_R^{-1}\underline{\mathbf{Y}}_q(i)\mathbf{F}_R\mathbf{W}_{R\times N}^{1_N 0}\Big\}\Big).\end{aligned} \tag{3.167}$$

In Section 3.4.3.2 it will be shown that by using suitable approximations an efficient computation of the nonlinearity $\mathbf{\Lambda}_q(i)$ can be achieved.

Using (3.148) and (3.164)-(3.166) the $PR \times P$ DFT-domain update $\Delta\underline{\check{\mathbf{W}}}$ is given for the HOS-SIRP algorithm as

$$\begin{aligned}\Delta\underline{\check{\mathbf{W}}}(m) &= 2\sum_{i=0}^{\infty}\beta(i,m)\mathbf{V}_{PL\times PR}^{\mathrm{H}}\mathcal{SC}\Big\{\mathbf{V}_{2PL\times PR}\underline{\mathbf{W}}(i)\left(\mathbf{V}_{PD\times PR}^{1_D 0}\right)^{\mathrm{H}}\mathbf{V}_{PD\times PR} \\ &\qquad \operatorname{boff}\left\{\underline{\mathbf{Y}}(i)\underline{\mathbf{Y}}_\phi^{\mathrm{H}}(i)\right\}\mathbf{V}_{PD\times PR}^{\mathrm{H}} \\ &\qquad \operatorname{bdiag}^{-1}\left\{\frac{1}{N}\mathbf{V}_{PD\times PR}\underline{\mathbf{Y}}(i)\mathbf{G}_{R\times R}^{1_N 0}\underline{\mathbf{Y}}^{\mathrm{H}}(i)\mathbf{V}_{PD\times PR}^{\mathrm{H}}\right\}\Big\}.\end{aligned} \tag{3.168}$$

It should be pointed out that in the derivation of the DFT-domain update (3.168) no approximations have been made and thus, $\Delta\underline{\check{\mathbf{W}}}$ could also be obtained by a channel-wise DFT of length R applied to the zero-padded $PL \times P$ time-domain update $\Delta\check{\mathbf{W}}$ given in (3.105). The advantage of the formulation in (3.168) is that all time-domain correlations and linear convolutions are expressed in an overlap-save structure using matrix notation and thus exploit the efficiency of fast convolutions. Due to the constraints $\mathbf{V}_{\ldots}$, $\mathbf{G}_{R\times R}^{1_N 0}$ (containing DFT, IDFT, and windowing operations) no circular convolution approximations are made and thus, the DFT bins are still coupled and the permutation ambiguity in each DFT bin is avoided. The formulation in (3.168) is the basis for *selective* narrowband approximations in Section 3.4.3 leading to efficient algorithms.

3.4.2.5 Second-order statistics realization based on the multivariate Gaussian pdf

In Section 3.3.7.2 it was shown that the usage of the multivariate Gaussian pdf leads to a second-order statistics algorithm given in (3.112), inherently exploiting the nonwhiteness property and exhibiting the same normalization by the auto-correlation matrices as the HOS-SIRP algorithm. Analogously to the previous section, we can express the SOS natural gradient update (3.112) in the DFT domain by using (3.141), (3.148), and (3.164) yielding

$$\begin{aligned}\Delta\underline{\check{\mathbf{W}}}(m) &= \sum_{i=0}^{\infty}\beta(i,m)\mathbf{V}_{PL\times PR}^{\mathrm{H}}\mathcal{SC}\Big\{\mathbf{V}_{2PL\times PR}\underline{\mathbf{W}}(i)\left(\mathbf{V}_{PD\times PR}^{1_D 0}\right)^{\mathrm{H}}\mathbf{V}_{PD\times PR}\\ &\qquad \mathrm{boff}\left\{\underline{\mathbf{Y}}(i)\mathbf{G}_{R\times R}^{1_N 0}\underline{\mathbf{Y}}^{\mathrm{H}}(i)\right\}\mathbf{V}_{PD\times PR}^{\mathrm{H}}\\ &\qquad \mathrm{bdiag}^{-1}\left\{\mathbf{V}_{PD\times PR}\underline{\mathbf{Y}}(i)\mathbf{G}_{R\times R}^{1_N 0}\underline{\mathbf{Y}}^{\mathrm{H}}(i)\mathbf{V}_{PD\times PR}^{\mathrm{H}}\right\}\Big\}, \qquad (3.169)\end{aligned}$$

which corresponds to an estimation of the cross-correlation matrices by the *covariance method*. Comparing (3.169) and (3.168) shows that for the algorithm based on second-order statistics the channel-wise nonlinearities $\mathbf{\Lambda}_q(i)$ are approximated as identity matrices.

If the correlation method is used instead, then the DFT-domain representation of the cross-correlation matrices is simplified as will be discussed in the following. Using the correlation method, the cross-correlation matrices are estimated in the time domain as $\tilde{\mathbf{R}}_{\mathbf{y}_p\mathbf{y}_q}(i) = \frac{1}{N}\tilde{\mathbf{Y}}_p(i)\tilde{\mathbf{Y}}_q^{\mathrm{H}}(i)$ with $\tilde{\mathbf{Y}}_q(i)$ defined in (3.82). The DFT representation of the time-domain Toeplitz matrix $\tilde{\mathbf{Y}}_q(i)$ is given in (3.161) as

$$\tilde{\mathbf{Y}}(i) = \mathbf{V}_{PD\times PR}\underline{\tilde{\mathbf{Y}}}(i)\mathbf{F}_R\mathbf{W}_{R\times N+D-1}^{1_{N+D-1}0}.$$

By approximating the window matrix $\mathbf{W}_{R\times N+D-1}^{1_{N+D-1}0}$ in (3.161) as an $R\times R$ identity matrix it can be seen in the lower right corner of the illustration in Fig. 3.14 that only columns of zeros are appended for each channel $q = 1,\ldots,P$ at the end of each matrix $\tilde{\mathbf{Y}}_q$, i.e., (3.161) is now for each channel of the form

$$\left[\tilde{\mathbf{Y}}_q(i), \mathbf{0}_{D\times R-N-D+1}\right] = \mathbf{W}_{D\times R}^{01_D}\mathbf{F}_R^{-1}\underline{\tilde{\mathbf{Y}}}_q(i)\mathbf{F}_R. \qquad (3.170)$$

These appended columns of zeros have no effect on the calculation of the correlation matrix $\tilde{\mathbf{R}}_{\mathbf{y}_p\mathbf{y}_q}$ as

$$\begin{aligned}\tilde{\mathbf{R}}_{\mathbf{y}_p\mathbf{y}_q}(i) &= \frac{1}{N}\tilde{\mathbf{Y}}_p(i)\tilde{\mathbf{Y}}_q^{\mathrm{H}}(i)\\ &= \frac{1}{N}\left[\tilde{\mathbf{Y}}_p(i), \mathbf{0}_{D\times R-N-D+1}\right]\left[\tilde{\mathbf{Y}}_q(i), \mathbf{0}_{D\times R-N-D+1}\right]^{\mathrm{H}} \qquad (3.171)\end{aligned}$$

and thus, the cross-correlation matrix in DFT representation can be written as

$$\tilde{\mathbf{R}}_{\mathbf{y}_p\mathbf{y}_q}(i) = \frac{1}{N}\mathbf{W}_{D\times R}^{01_D}\mathbf{F}_R^{-1}\underline{\tilde{\mathbf{Y}}}_p(i)\underline{\tilde{\mathbf{Y}}}_q^{\mathrm{H}}(i)\mathbf{F}_R\mathbf{W}_{R\times D}^{01_D}. \qquad (3.172)$$

A combination of all channels leads to

$$\tilde{\mathbf{R}}_{\mathbf{y}\mathbf{y}}(i) = \frac{1}{N}\mathbf{V}_{PD\times PR}\underline{\tilde{\mathbf{Y}}}(i)\underline{\tilde{\mathbf{Y}}}^{\mathrm{H}}(i)\mathbf{V}_{PD\times PR}^{\mathrm{H}}. \qquad (3.173)$$

Comparing (3.164) and (3.173) shows that by using the correlation method the constraint matrix $\mathbf{G}_{R\times R}^{01_N}$ is reduced to the identity matrix $\mathbf{I}_{R\times R}$. Therefore, by using the *correlation method* the SOS natural gradient in the DFT domain simplifies to

$$\Delta\underline{\check{\mathbf{W}}}(m) = \sum_{i=0}^{\infty}\beta(i,m)\mathbf{V}_{PL\times PR}^{\mathrm{H}}\mathcal{SC}\left\{\mathbf{V}_{2PL\times PR}\underline{\mathbf{W}}(i)\left(\mathbf{V}_{PD\times PR}^{1_{D}0}\right)^{\mathrm{H}}\mathbf{V}_{PD\times PR}\right.$$
$$\mathrm{boff}\left\{\underline{\tilde{\mathbf{Y}}}(i)\underline{\tilde{\mathbf{Y}}}^{\mathrm{H}}(i)\right\}\mathbf{V}_{PD\times PR}^{\mathrm{H}}$$
$$\left.\mathrm{bdiag}^{-1}\left\{\mathbf{V}_{PD\times PR}\underline{\tilde{\mathbf{Y}}}(i)\underline{\tilde{\mathbf{Y}}}^{\mathrm{H}}(i)\mathbf{V}_{PD\times PR}^{\mathrm{H}}\right\}\right\}. \qquad (3.174)$$

The two DFT-domain formulations (3.169) and (3.174) do not contain any approximations and thus, the time-domain correlations and linear convolutions are expressed equivalently in an overlap-save structure using matrix notation. To allow efficient implementations we will investigate in the next section the introduction of selective narrowband approximations which are applied, e.g., to the computationally demanding matrix inverse.

3.4.3 Selective approximations leading to well-known and novel algorithms

In this section we will discuss selective approximations leading to efficient DFT-domain implementations of the HOS and SOS natural gradient algorithms derived in (3.168), (3.169), and (3.174). The main computational complexity of these algorithms originates from the normalization by inverting the $D \times D$ auto-correlation matrices for each channel and from the calculation of the SIRP-based nonlinearity $\mathbf{\Lambda}_q$. In the following, suitable approximations addressing these two aspects will be presented. First, in Section 3.4.3.1 the inversion of the large $D \times D$ auto-correlation matrices is efficiently approximated by a narrowband normalization. This approximation is applied to the HOS and SOS DFT-domain updates in Sections 3.4.3.2 and 3.4.3.3. Moreover, further simplifications are discussed yielding several novel algorithms and also establishing links to popular state-of-the-art algorithms. Finally, in Section 3.4.3.4 the relationship of narrowband SOS-BSS and the magnitude squared coherence (MSC) is discussed, showing the link to other cost functions and to the estimation of the MSC presented in Chapter 2.

3.4.3.1 Narrowband normalization and regularization strategies

In literature it is very popular to apply BSS algorithms independently to each frequency bin in the DFT domain. This leads to an optimization based on the narrowband model as was pointed out in Section 3.2. Usually, these algorithms are computationally efficient due to the application of circular convolutions. However, the drawback is that they suffer from the scaling and permutation ambiguity in each DFT bin. By applying only selective narrowband approximations it is possible to exploit narrowband efficiency without encountering the scaling and permutation problem. The Szegö theorem [GS58] provides

the mathematical basis for the narrowband approximation and will now be applied to the normalization given by the inverse of the block-diagonal matrix, i.e., by the channel-wise inversion of the auto-correlation matrices in (3.168), (3.169), and (3.174). Subsequently, also a regularization method is presented which improves the robustness in realistic environments.

Approximation of the normalization based on the Szegö theorem

In the tutorial paper [Gra72] the Szegö theorem, originally introduced in [GS58], is formulated and proven for finite-order Toeplitz matrices. A finite-order Toeplitz matrix is defined as an $R \times R$ Toeplitz matrix where a finite D ($D < R$) exists such that all elements of the matrix with the row or column index greater than D are equal to zero. It was shown in [Gra72] that the $R \times R$ Toeplitz matrix of order D is asymptotically equivalent to the $R \times R$ circulant matrix generated from an appropriately complemented $D \times D$ Toeplitz matrix. If the two matrices are also of hermitian structure, then the Szegö theorem on the asymptotic eigenvalue distribution states:

1. The eigenvalues of both matrices lie between the same lower and upper bound.
2. The arithmetic means of the eigenvalues of both matrices are equal if the size R of both matrices approaches infinity.

Then, the eigenvalues of both matrices are said to be asymptotically equally distributed.

The Szegö theorem will now be applied to the inversion of the auto-correlation matrices which are needed in the HOS and SOS natural gradient updates (3.168), (3.169), and (3.174). At first we will consider the case that the auto-correlation matrices are estimated by the correlation method as $\tilde{\mathbf{R}}_{\mathbf{y}_q\mathbf{y}_q} = \frac{1}{N}\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q^{\mathrm{H}}$ yielding a Toeplitz structure. Later on, also the estimation by using the covariance method $\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q} = \frac{1}{N}\mathbf{Y}_q\mathbf{Y}_q^{\mathrm{H}}$ will be investigated.

Correlation method. In (3.172) the relationship between the $D \times D$ Toeplitz auto-correlation matrix $\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}$ and the DFT-domain quantity $\underline{\tilde{\mathbf{Y}}}_q$ is given for the case that the correlation method is used. We can rewrite (3.172) as

$$\tilde{\mathbf{R}}_{\mathbf{y}_q\mathbf{y}_q}(i) = \frac{1}{N}\mathbf{W}_{D\times R}^{01_D}\mathbf{C}_{\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q}(i)\mathbf{W}_{R\times D}^{01_D}, \tag{3.175}$$

with the $R \times R$ circulant matrix $\mathbf{C}_{\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q}$ given as

$$\mathbf{C}_{\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q}(i) = \mathbf{F}_R^{-1}\underline{\tilde{\mathbf{Y}}}_q(i)\underline{\tilde{\mathbf{Y}}}_q^{\mathrm{H}}(i)\mathbf{F}_R. \tag{3.176}$$

Thus, (3.175) shows the relationship between the $D \times D$ Toeplitz matrix $\tilde{\mathbf{R}}_{\mathbf{y}_q\mathbf{y}_q}$ and the $R \times R$ circulant matrix $\mathbf{C}_{\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q}$ generated from the Toeplitz matrix by extending and multiplying it with appropriate window matrices.

According to [Gra72] the circulant $\mathbf{C}_{\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q}$ and the D-th order Toeplitz matrix $\tilde{\mathbf{R}}_{\mathbf{y}_q\mathbf{y}_q}$ extended with zeros to the size $R \times R$ are asymptotically equivalent. Additionally, the Szegö theorem states that the eigenvalues of the $R \times R$ Toeplitz matrix generated by appending zeros to $\tilde{\mathbf{R}}_{\mathbf{y}_q\mathbf{y}_q}$ can be asymptotically approximated for $R \to \infty$ by the eigenvalues of $\mathbf{C}_{\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q}$ which are given as the elements on the main diagonal of the diagonal matrix $\underline{\tilde{\mathbf{Y}}}_q\underline{\tilde{\mathbf{Y}}}_q^{\mathrm{H}}$. The benefit of this approximation becomes clear if we consider the inverse of a circulant matrix. The inverse of a circulant matrix can be easily calculated by inverting its eigenvalues

$$\mathbf{C}^{-1}_{\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q}(i) = \mathbf{F}_R^{-1}\left(\underline{\tilde{\mathbf{Y}}}_q(i)\underline{\tilde{\mathbf{Y}}}_q^{\mathrm{H}}(i)\right)^{-1}\mathbf{F}_R. \tag{3.177}$$

By using the Szegö theorem we can now approximate the inverse of the Toeplitz matrix $\tilde{\mathbf{R}}_{\mathbf{y}_q\mathbf{y}_q}$ by the inverse of the circulant matrix (3.177) for $R \to \infty$:

$$\tilde{\mathbf{R}}^{-1}_{\mathbf{y}_q\mathbf{y}_q}(i) \approx N \cdot \mathbf{W}^{01_D}_{D\times R}\mathbf{F}_R^{-1}\left(\underline{\tilde{\mathbf{Y}}}_q(i)\underline{\tilde{\mathbf{Y}}}_q^{\mathrm{H}}(i)\right)^{-1}\mathbf{F}_R\mathbf{W}^{01_D}_{R\times D}. \tag{3.178}$$

This can also be denoted as *narrowband approximation* because the eigenvalues $\underline{\tilde{\mathbf{Y}}}_q\underline{\tilde{\mathbf{Y}}}_q^{\mathrm{H}}$ can easily be determined as the DFT of the first column of the circulant matrix $\mathbf{C}_{\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q}$. The inverse in (3.178) can now be efficiently implemented as a scalar inversion in each DFT bin because $\underline{\tilde{\mathbf{Y}}}_q\underline{\tilde{\mathbf{Y}}}_q^{\mathrm{H}}$ denotes a diagonal matrix. Moreover, it is important to note that the inverse of a circulant matrix is also circulant. Thus, after the windowing by $\mathbf{W}^{01_D}_{\dots}$ the resulting matrix $\tilde{\mathbf{R}}^{-1}_{\mathbf{y}_q\mathbf{y}_q}$ exhibits again a Toeplitz structure.

The error which is introduced by the narrowband approximation has been examined in [She85] for the case of stationary random processes. The error has been measured as the difference between the exact inversion of the Toeplitz matrix and the approximated inverse given in (3.178). The results obtained in [She85] show that for $R \gg D$ the narrowband approximation is well-justified.

In summary, (3.178) can be efficiently implemented as a DFT of the first column of $\mathbf{C}_{\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q}$ followed by a scalar inversion of the DFT-domain values and then applying the inverse DFT. After the windowing operation these values are then replicated to generate the Toeplitz structure of $\tilde{\mathbf{R}}^{-1}_{\mathbf{y}_q\mathbf{y}_q}$. This approach reduces the complexity of the matrix inversion from $\mathcal{O}(D^2)$ to $\mathcal{O}(R\log R)$.

Covariance method. If the covariance method is used instead to estimate the autocorrelation matrices then $\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}$ is given in (3.162) as

$$\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}(i) = \frac{1}{N}\mathbf{W}^{01_D}_{D\times R}\mathbf{F}_R^{-1}\underline{\mathbf{Y}}_q(i)\mathbf{G}^{1_N0}_{R\times R}\underline{\mathbf{Y}}_q^{\mathrm{H}}(i)\mathbf{F}_R\mathbf{W}^{01_D}_{R\times D},$$

where compared to the correlation method the additional constraint matrix $\mathbf{G}^{1_N0}_{R\times R}$ appears. The constraint matrix is defined in (3.163) as an DFT, windowing, and IDFT operation given as $\mathbf{G}^{1_N0}_{R\times R} = \mathbf{F}_R^{-1}\mathbf{W}^{1_N0}_{R\times R}\mathbf{F}_R$. Since the window matrix $\mathbf{W}^{1_N0}_{R\times R}$ is diagonal,

the constraint matrix $\mathbf{G}_{R\times R}^{1_N 0}$ is circulant which can be seen in the illustration in Fig. 3.15. In [BBK03, BBK05] such a constraint was examined in the context of supervised adaptive

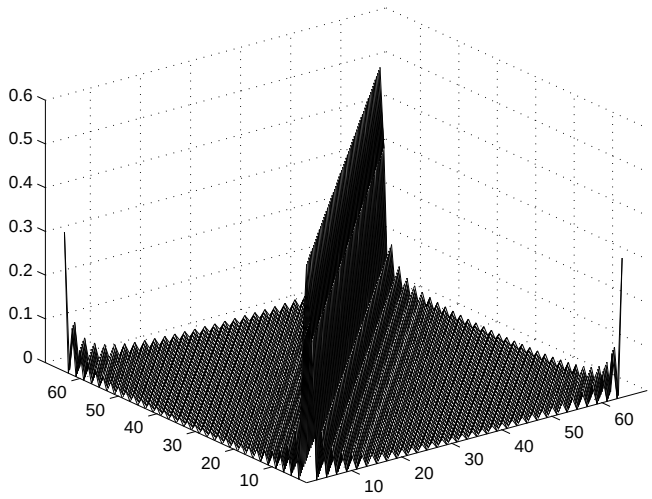


Figure 3.15: Illustration of the constraint matrix $|\mathbf{G}_{R\times R}^{1_N 0}|$ with $R = 64, N = 32$.

system identification and it was pointed out that the main diagonal is dominant in the mean-square sense. By neglecting the off-diagonals and the influence of the two isolated peaks on the lower left and upper right corner in Fig. 3.15, the constraint matrix can be approximated as a scaled identity matrix [BBK03, BBK05]

$$\mathbf{G}_{R\times R}^{1_N 0} \approx \frac{N}{R} \cdot \mathbf{I}_{R\times R}. \tag{3.179}$$

This can be interpreted as approximating the computation of the correlation $\mathbf{Y}_q\mathbf{Y}_q^{\mathrm{H}}$ in the time-domain as a circular convolution $\underline{\mathbf{Y}}_q\underline{\mathbf{Y}}_q^{\mathrm{H}}$ in the DFT domain. Thus, circular convolution effects originating from the narrowband computation are neglected and the computation of the time-domain auto-correlation matrices is considerably simplified to

$$\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}(i) = \frac{1}{N}\mathbf{W}_{D\times R}^{01_D}\mathbf{C}_{\mathbf{Y}_q\mathbf{Y}_q}(i)\mathbf{W}_{R\times D}^{01_D}, \tag{3.180}$$

with

$$\mathbf{C}_{\mathbf{Y}_q\mathbf{Y}_q}(i) = \frac{N}{R}\mathbf{F}_R^{-1}\underline{\mathbf{Y}}_q(i)\underline{\mathbf{Y}}_q^{\mathrm{H}}(i)\mathbf{F}_R. \tag{3.181}$$

In general $\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}$ does not exhibit any special structure if it is estimated using the covariance method. However, if the approximation (3.179) is used then, similar to the correlation method, $\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}$ is *Toeplitz even if it is estimated by the covariance method*. By using again the Szegö theorem, the inverse of the auto-correlation matrix $\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}$ can be

approximated by the inverse of the circulant matrix $\mathbf{C}_{\mathbf{Y}_q\mathbf{Y}_q}$ for $R \to \infty$ which yields for the covariance method

$$\mathbf{R}^{-1}_{\mathbf{y}_q\mathbf{y}_q}(i) \approx R \cdot \mathbf{W}^{01_D}_{D\times R}\mathbf{F}^{-1}_R \left(\underline{\mathbf{Y}}_q(i)\underline{\mathbf{Y}}^{\mathrm{H}}_q(i)\right)^{-1}\mathbf{F}_R\mathbf{W}^{01_D}_{R\times D}. \tag{3.182}$$

The inverse in (3.182) can now again be efficiently implemented as a scalar inversion.

This computationally efficient estimation of the inverse auto-correlation matrices based on the covariance method is also used in supervised frequency-domain adaptive filtering (FDAF) (e.g., [Shy92, BBK03, BBK05]). There, the narrowband approximation (3.179) allowed to derive low-complexity multi-channel RLS-type algorithms. Applications of such algorithms include multi-channel acoustic echo cancellation (e.g., [BBK03, BBK05]) or adaptive beamforming (e.g., [HBK03]).

Regularization of the matrix inverse

Prior to the inversion of the auto-correlation Toeplitz matrices according to (3.178) or (3.182) a regularization is necessary as in practice these matrices may be ill-conditioned. In [ABK06a] it was proposed to attenuate the off-diagonals of the auto-correlation Toeplitz matrices $\tilde{\mathbf{R}}_{\mathbf{y}_q\mathbf{y}_q}$ by multiplying them with the factor ρ ($0 \leq \rho \leq 1$):

$$\begin{aligned}\breve{\mathbf{R}}_{\mathbf{y}_q\mathbf{y}_q}(i) &= \rho\tilde{\mathbf{R}}_{\mathbf{y}_q\mathbf{y}_q}(i) + (1-\rho)\mathrm{diag}\left\{\tilde{\mathbf{R}}_{\mathbf{y}_q\mathbf{y}_q}(i)\right\} \\ &= \rho\tilde{\mathbf{R}}_{\mathbf{y}_q\mathbf{y}_q}(i) + (1-\rho)\sigma^2_{y_q}(i)\mathbf{I}\end{aligned} \tag{3.183}$$

It should be noted that for $\rho = 0$ the previous approximation of the normalization by the output signal variance (3.116) in Section 3.3.8 can be seen as a special case of the regularized version of the narrowband normalization.

The approximations in the previous section have led to a narrowband normalization which is characterized by an inversion of circulant matrices $\mathbf{C}_{\tilde{\mathbf{Y}}_p\tilde{\mathbf{Y}}_q}$, $\mathbf{C}_{\mathbf{Y}_p\mathbf{Y}_q}$ instead of Toeplitz matrices $\tilde{\mathbf{R}}_{\mathbf{y}_q\mathbf{y}_q}$. Thus, analogously to (3.183) it is desirable for the DFT-domain implementations to regularize $\mathbf{C}_{\tilde{\mathbf{Y}}_p\tilde{\mathbf{Y}}_q}$, $\mathbf{C}_{\mathbf{Y}_p\mathbf{Y}_q}$ instead of $\tilde{\mathbf{R}}_{\mathbf{y}_q\mathbf{y}_q}$ prior to inversion. This is shown here exemplarily for the circulant matrix based on estimation by the correlation method:

$$\breve{\mathbf{C}}_{\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q}(i) = \rho\mathbf{C}_{\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q}(i) + (1-\rho)\mathrm{diag}\left\{\mathbf{C}_{\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q}(i)\right\}. \tag{3.184}$$

In (3.176) it was pointed out that every circulant matrix can be expressed using the DFT and inverse DFT matrix and a diagonal matrix as $\mathbf{C}_{\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q} = \mathbf{F}^{-1}_R\underline{\tilde{\mathbf{Y}}}_q\underline{\tilde{\mathbf{Y}}}^{\mathrm{H}}_q\mathbf{F}_R$. The diagonal matrix $\underline{\tilde{\mathbf{Y}}}_q\underline{\tilde{\mathbf{Y}}}^{\mathrm{H}}_q$ contains the DFT values of the elements of the first column of the circulant matrix $\mathbf{C}_{\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q}$ on its diagonal. Thus, by applying the diag operator on $\mathbf{C}_{\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q}$ we can

write

$$\begin{aligned}\operatorname{diag}\left\{\mathbf{C}_{\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q}(i)\right\} &= r_{y_qy_q}(i,0)\cdot\mathbf{I}\\ &= \sigma^2_{y_q}(i)\cdot\mathbf{I}\\ &= \mathbf{F}_R^{-1}\,\sigma^2_{y_q}(i)\cdot\mathbf{I}\cdot\mathbf{F}_R,\end{aligned}\tag{3.185}$$

with the auto-correlation elements $r_{y_qy_q}(i,0)$ defined in (3.79). Thus, (3.184) can be simplified to a narrowband regularization in each frequency bin as

$$\begin{aligned}\breve{\mathbf{C}}_{\tilde{\mathbf{Y}}_p\tilde{\mathbf{Y}}_q}(i) &= \rho\mathbf{F}_R^{-1}\underline{\tilde{\mathbf{Y}}}_q(i)\underline{\tilde{\mathbf{Y}}}_q^{\mathrm{H}}(i)\mathbf{F}_R+(1-\rho)\sigma^2_{y_q}(i)\mathbf{I}\\ &= \mathbf{F}_R^{-1}\left(\rho\underline{\tilde{\mathbf{Y}}}_q(i)\underline{\tilde{\mathbf{Y}}}_q^{\mathrm{H}}(i)+(1-\rho)\sigma^2_{y_q}(i)\mathbf{I}\right)\mathbf{F}_R.\end{aligned}\tag{3.186}$$

The variance is added in (3.186) to each DFT bin which is equivalent to adding it to the diagonal of the time-domain matrix $\tilde{\mathbf{R}}_{\mathbf{y}_q\mathbf{y}_q}$ in (3.183). It should be noted that the regularization in (3.186) which has been derived based on broadband optimization can also be applied to algorithms based on narrowband optimization.

The narrowband normalization (3.178) or (3.182) together with the regularization (3.186) is a first step to decrease the computational complexity of the DFT-domain updates without much performance loss. In the following two sections the HOS and SOS DFT-domain updates are investigated in more detail and additional approximations are presented.

3.4.3.2 BSS based on higher-order statistics

In Section 3.3.7.1 a HOS-SIRP algorithm was presented and the resulting nonlinearity has been given exemplarily for the multivariate Laplacian SIRP pdf which is a good model for speech. The assumption of SIRPs considerably simplifies the estimation of multivariate pdfs as only second-order correlation matrices instead of higher-order moments have to be estimated. In Section 3.4.2.4 an equivalent representation of the HOS-SIRP algorithm in the DFT domain given by (3.168) has been discussed. However, for an efficient implementation of (3.168) not only the normalization has to be approximated as shown in the previous section, but also suitable approximations have to be introduced to efficiently calculate the nonlinearity $\mathbf{\Lambda}_q(i)$ in the DFT domain. In this Section we will show how efficient HOS algorithms can be obtained by selectively approximating the constraints originating from the broadband formulation (3.168) of the HOS-SIRP algorithm in the DFT domain. This will show relationships to several popular algorithms in the literature.

Narrowband algorithms based on multivariate pdfs

It has been shown in Section 3.4.3.1 that the narrowband normalization (3.182) leads to less computational complexity compared to the inversion of the time-domain auto-correlation matrices. This approximation is applied to the HOS-SIRP update equation

(3.168) expressed in the DFT domain by using the broadband signal model. Additionally, (3.168) can further be simplified by selectively introducing the narrowband signal model and thus, replacing linear convolutions by circular convolutions. Hence, the narrowband normalization together with the approximation of the constraint $\left(\mathbf{V}_{PD\times PR}^{1_D 0}\right)^{\mathrm{H}}\mathbf{V}_{PD\times PR}$ simplifies the HOS-SIRP update to

$$\Delta\underline{\check{\mathbf{W}}}(m) = \frac{2}{N}\sum_{i=0}^{\infty}\beta(i,m)\mathbf{V}_{PL\times PR}^{\mathrm{H}}\mathcal{SC}\left\{\mathbf{V}_{2PL\times PR}\underline{\mathbf{W}}(i)\mathrm{boff}\left\{\underline{\mathbf{Y}}(i)\underline{\mathbf{Y}}_{\phi}^{\mathrm{H}}(i)\right\}\right. \\ \left.\mathrm{diag}^{-1}\left\{\frac{1}{R}\underline{\mathbf{Y}}(i)\underline{\mathbf{Y}}^{\mathrm{H}}(i)\right\}\mathbf{V}_{PD\times PR}^{\mathrm{H}}\right\}, \tag{3.187}$$

where the nonlinearly weighted output signals $\underline{\mathbf{Y}}_{\phi}^{\mathrm{H}} = [\underline{\mathbf{Y}}_{\phi,1}^{\mathrm{H}},\ldots,\underline{\mathbf{Y}}_{\phi,P}^{\mathrm{H}}]$ are given as

$$\underline{\mathbf{Y}}_{\phi}^{\mathrm{H}} = \left[\mathbf{F}_R\mathbf{W}_{R\times N}^{1_N 0}\mathbf{\Lambda}_1^{\mathrm{H}}\mathbf{W}_{N\times R}^{1_N 0}\mathbf{F}_R^{-1}\underline{\mathbf{Y}}_1^{\mathrm{H}},\ldots,\mathbf{F}_R\mathbf{W}_{R\times N}^{1_N 0}\mathbf{\Lambda}_P^{\mathrm{H}}\mathbf{W}_{N\times R}^{1_N 0}\mathbf{F}_R^{-1}\underline{\mathbf{Y}}_P^{\mathrm{H}}\right]. \tag{3.188}$$

The matrices $\underline{\mathbf{W}}$ and $\underline{\mathbf{Y}}$ are composed of diagonal submatrices $\underline{\mathbf{W}}_{pq}$, $\underline{\mathbf{Y}}_q$ with the DFT bins on the diagonal. Due to the remaining constraint matrices $\mathbf{V}_{\ldots}$ and the $PD\times N$ matrix $\underline{\mathbf{Y}}_{\phi}$ a decoupling of the DFT bins, which would lead to an independent permutation and scaling ambiguity in each DFT bin, is prevented. Nevertheless, for reasons of computational complexity it would be desirable that also the channel-wise submatrices of $\underline{\mathbf{Y}}_{\phi}$ could be approximated as a diagonal matrices with the DFT bins on the diagonal. Then, the coupling would still be ensured by the constraints $\mathbf{V}_{\ldots}$ and also by the nonlinearity, which is still based on multivariate pdfs. However, the advantage would be that the matrix multiplications involving $\underline{\mathbf{W}}$, $\underline{\mathbf{Y}}$, and $\underline{\mathbf{Y}}_{\phi}$ could be performed element-wise for each DFT bin. These element-wise computations can be seen as a bin-wise decomposition of the $PR\times PR$ matrix $\underline{\mathbf{W}}$ or the $PR\times R$ matrix $\underline{\mathbf{Y}}$ into R smaller $P\times P$ matrices $\underline{\mathbf{W}}^{(\nu)}$ or $P\times 1$ column vectors $\underline{\mathbf{Y}}^{(\nu)}$ for each DFT bin $\nu = 1,\ldots,R-1$. An illustration of this decomposition is given in Fig. 3.16 exemplarily for $\underline{\mathbf{Y}}$.

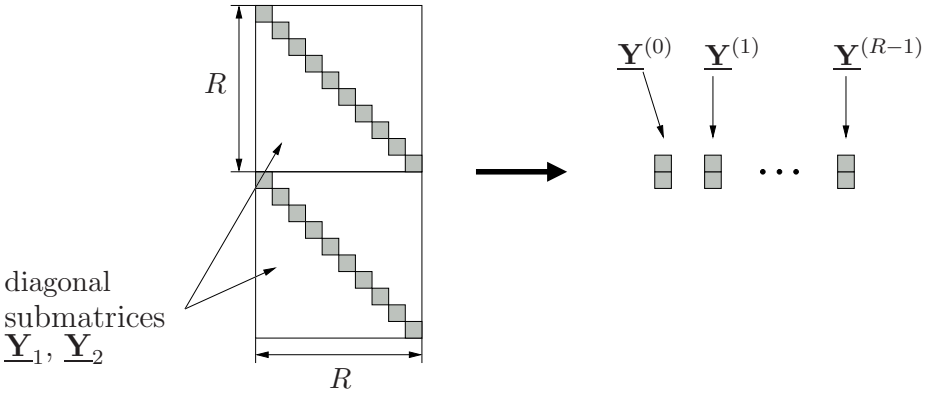


Figure 3.16: Illustration of the bin-wise decomposition of $\underline{\mathbf{Y}}$ for the case $P = 2$.

In the following we will discuss the approximations which also allow a bin-wise decomposition of the nonlinearly weighted output signals $\underline{\mathbf{Y}}_{\phi}$ and show how these approximations lead to several recently published algorithms. For that, we first examine the nonlinear weighting matrix $\mathbf{\Lambda}_q$ which is an $N \times N$ diagonal matrix originating from the assumption of multivariate SIRP pdfs of dimension D and is in (3.108) expressed in the time domain as

$$\mathbf{\Lambda}_q(i) = \phi_{y_q,D}\left(\mathrm{diag}\left\{\mathbf{Y}_q^{\mathrm{H}}(i)\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}^{-1}(i)\mathbf{Y}_q(i)\right\}\right). \tag{3.189}$$

In (3.167) this quantity was expressed equivalently in the DFT domain by using the time-frequency relations (3.154) and (3.162) leading to

$$\begin{aligned}\mathbf{\Lambda}_q(i) \;=\; & \phi_{y_q,D}\Big(\mathrm{diag}\Big\{\mathbf{W}_{N\times R}^{1_N 0}\mathbf{F}_R^{-1}\underline{\mathbf{Y}}_q^{\mathrm{H}}(i)\mathbf{F}_R\mathbf{W}_{R\times D}^{01_D} \\ & \left(\frac{1}{N}\mathbf{W}_{D\times R}^{01_D}\mathbf{F}_R^{-1}\underline{\mathbf{Y}}_q^{\mathrm{H}}(i)\mathbf{G}_{R\times R}^{1_N 0}\underline{\mathbf{Y}}_q(i)\mathbf{F}_R\mathbf{W}_{R\times D}^{01_D}\right)^{-1}\mathbf{W}_{D\times R}^{01_D}\mathbf{F}_R^{-1}\underline{\mathbf{Y}}_q(i)\mathbf{F}_R\mathbf{W}_{R\times N}^{1_N 0}\Big\}\Big).\end{aligned} \tag{3.190}$$

The inversion of the auto-correlation matrix can be approximated by the narrowband inverse as pointed out in (3.182) leading to the diagonal matrix $\underline{\mathbf{S}}_{\mathbf{y}_q\mathbf{y}_q}(i) = \frac{1}{R}\underline{\mathbf{Y}}_q(i)\underline{\mathbf{Y}}_q^{\mathrm{H}}(i)$ with the DFT bins on its diagonal. Due to the diagonal structure this allows for an element-wise inversion. Thus, $\mathbf{\Lambda}_q(i)$ can be simplified to

$$\mathbf{\Lambda}_q(i) \approx \phi_{y_q,D}\left(\mathrm{diag}\left\{\mathbf{W}_{N\times R}^{1_N 0}\mathbf{F}_R^{-1}\underline{\mathbf{Y}}_q^{\mathrm{H}}(i)\mathbf{G}_{R\times R}^{01_D}\underline{\mathbf{S}}_{\mathbf{y}_q\mathbf{y}_q}^{-1}(i)\mathbf{G}_{R\times R}^{01_D}\underline{\mathbf{Y}}_q(i)\mathbf{F}_R\mathbf{W}_{R\times N}^{1_N 0}\right\}\right), \tag{3.191}$$

where, by exploiting the fact that the product of the window matrices can be expressed as $\mathbf{W}_{R\times D}^{01_D}\mathbf{W}_{D\times R}^{01_D} = \mathbf{W}_{R\times R}^{01_D}$, the constraint matrix $\mathbf{G}_{R\times R}^{01_D}$ is defined as

$$\mathbf{G}_{R\times R}^{01_D} = \mathbf{F}_R\mathbf{W}_{R\times R}^{01_D}\mathbf{F}_R^{-1}. \tag{3.192}$$

The entries on the diagonal of $\underline{\mathbf{S}}_{\mathbf{y}_q\mathbf{y}_q}(i)$ correspond to estimates of the variance in each of the R DFT bins. In [Hir06, KEL06, KALL06, KALL07] it has been argued that the matrix $\underline{\mathbf{S}}_{\mathbf{y}_q\mathbf{y}_q}(i)$ may be further approximated as an identity matrix $\underline{\mathbf{S}}_{\mathbf{y}_q\mathbf{y}_q}(i) \approx \mathbf{I}_{R\times R}$. This assumption can be justified if in each DFT bin a prewhitening step (see, e.g., [HKO01]), i.e., a cross-channel decorrelation and normalization to unit variance, is applied. This additional approximation leads to

$$\begin{aligned}\mathbf{\Lambda}_q(i) \;&\approx\; \phi_{y_q,D}\left(\mathrm{diag}\left\{\mathbf{W}_{N\times R}^{1_N 0}\mathbf{F}_R^{-1}\underline{\mathbf{Y}}_q^{\mathrm{H}}(i)\mathbf{G}_{R\times R}^{01_D}\underline{\mathbf{Y}}_q(i)\mathbf{F}_R\mathbf{W}_{R\times N}^{1_N 0}\right\}\right) \\ &=\; \phi_{y_q,D}\left(\mathrm{diag}\left\{\mathbf{Y}_q^{\mathrm{H}}(i)\mathbf{Y}_q(i)\right\}\right),\end{aligned} \tag{3.193}$$

where the second line in (3.193) corresponds to the time-domain interpretation. As the HOS-SIRP algorithm is based on the covariance method, the entries on the diagonal of the $N\times N$ matrix $\mathrm{diag}\left\{\mathbf{Y}_q^{\mathrm{H}}(i)\mathbf{Y}_q(i)\right\}$ are not identical. Each of the N entries is determined by the summation over D output signal samples given as $\mathbf{y}_q^{\mathrm{H}}(iL+j)\mathbf{y}_q(iL+j)$, $j = 0,\ldots,N$.

This shows that in total $N+D-1$ output signal samples $y(iL-D+1),\ldots,y(iL+N-1)$ are taken into account. By assuming stationarity within this signal block, we can approximate the covariance method so that all values on the diagonal of diag $\left\{\mathbf{Y}_q^{\mathrm{H}}(i)\mathbf{Y}_q(i)\right\}$ are equal to the variance $\sigma_{y_q}^2$ given as

$$\begin{aligned}\sigma_{y_q}^2(i) &= \sum_{n=iL-D+1}^{iL+N-1} y_q^2(n) \\ &= \sum_{\nu=0}^{R-1} |\underline{Y}_q^{(\nu)}(i)|^2. \end{aligned} \tag{3.194}$$

For the DFT-domain representation in the second line Parseval's theorem[6] was invoked and $\underline{Y}_q^{(\nu)}$ denotes the element in the ν-th DFT bin given as the ν-th element on the diagonal of $\underline{\mathbf{Y}}_q$. Thus, the nonlinearity can be expressed as

$$\begin{aligned}\mathbf{\Lambda}_q(i) &\approx \phi_{y_q,D}\left(\sigma_{y_q}^2(i)\right)\cdot\mathbf{I}_{N\times N} \\ &= \phi_{y_q,D}\left(\sum_{\nu=0}^{R-1}|\underline{Y}_q^{(\nu)}(i)|^2\right)\cdot\mathbf{I}_{N\times N}. \end{aligned} \tag{3.195}$$

Inserting (3.195) in (3.188) leads to

$$\begin{aligned}\underline{\mathbf{Y}}_\phi^{\mathrm{H}}(i) &= \left[\mathbf{G}_{R\times R}^{1_N 0}\underline{\mathbf{Y}}_1^{\mathrm{H}}(i)\underline{\mathbf{\Lambda}}_1^{\mathrm{H}}(i),\ldots,\mathbf{G}_{R\times R}^{1_N 0}\underline{\mathbf{Y}}_P^{\mathrm{H}}(i)\underline{\mathbf{\Lambda}}_P^{\mathrm{H}}(i)\right] \\ &\approx \frac{N}{R}\underline{\mathbf{Y}}^{\mathrm{H}}(i)\underline{\mathbf{\Lambda}}^{\mathrm{H}}(i), \end{aligned} \tag{3.196}$$

where in the second line again the approximation (3.179) for the constraint matrix $\mathbf{G}_{R\times R}^{1_N 0}$ has been used and the diagonal matrices $\underline{\mathbf{\Lambda}}_q(i)$, $\underline{\mathbf{\Lambda}}(i)$ are defined as

$$\underline{\mathbf{\Lambda}}_q(i) = \phi_{y_q,D}\left(\sum_{\nu=0}^{R-1}|\underline{Y}_q^{(\nu)}(i)|^2\right)\cdot\mathbf{I}_{R\times R}, \tag{3.197}$$

$$\underline{\mathbf{\Lambda}}(i) = \mathrm{Bdiag}\left\{\underline{\mathbf{\Lambda}}_1(i),\ldots,\underline{\mathbf{\Lambda}}_P(i)\right\}. \tag{3.198}$$

We see now that the approximations have led to a simplified nonlinearly weighted matrix $\underline{\mathbf{Y}}_\phi^{\mathrm{H}}$ in (3.196) which consists of diagonal submatrices. Therefore, $\underline{\mathbf{Y}}_\phi^{\mathrm{H}}$ can be decomposed into the individual DFT bins. Nevertheless, the coupling between the DFT bins is ensured by the nonlinearity $\phi_{y_q,D}(\cdot)$ whose argument includes all DFT bins. The approximated nonlinearity $\underline{\mathbf{\Lambda}}(i)$ leads to the simplified update

$$\Delta\underline{\check{\mathbf{W}}}(m) = \frac{2}{R}\sum_{i=0}^{\infty}\beta(i,m)\mathbf{V}_{PL\times PR}^{\mathrm{H}}\mathcal{SC}\left\{\mathbf{V}_{2PL\times PR}\Delta\underline{\mathbf{W}}(i)\mathbf{V}_{PD\times PR}^{\mathrm{H}}\right\}, \tag{3.199}$$

[6]The DFT matrix $\mathbf{F}_R$ has been introduced as $[\mathbf{F}_R]_{ik} = \frac{1}{\sqrt{R}}e^{-j2\pi ik/R}$ and is a unitary matrix due to the scaling factor $\frac{1}{\sqrt{R}}$. For a unitary DFT matrix the Parseval theorem is given as $\sum_n |y_q(n)|^2 = \sum_{\nu=0}^{R-1}|\underline{Y}_q^{(\nu)}|^2$. It should be noted that if the scaling factor in the definition of the DFT matrix is omitted, then the Parseval theorem changes to $\sum_n |y_q(n)|^2 = \frac{1}{R}\sum_{\nu=0}^{R-1}|\underline{Y}_q^{(\nu)}|^2$.

with the $PR \times PR$ update defined as

$$\Delta\underline{\mathbf{W}}(i) = \underline{\mathbf{W}}(i)\text{boff}\left\{\underline{\mathbf{Y}}(i)\underline{\mathbf{Y}}^{\mathrm{H}}(i)\underline{\mathbf{\Lambda}}^{\mathrm{H}}(i)\right\}\text{diag}^{-1}\left\{\underline{\mathbf{S}}_{\mathbf{yy}}(i)\right\}. \tag{3.200}$$

All matrices in (3.200) consist of diagonal channel-wise submatrices with the DFT values on the diagonal. This means that (3.200) can easily be computed by element-wise multiplications instead of matrix multiplications. Thus, this can be seen as decomposing the large $PR \times PR$ or $PR \times P$ matrices for each DFT bin into R small $P \times P$ matrices or $P \times 1$ vectors as illustrated in Fig. 3.16.

The two constraint matrices $\mathbf{V}_{2PL\times PR}$ and $\mathbf{V}^{\mathrm{H}}_{PD\times PR}$ in (3.199) transform the DFT-domain matrix $\Delta\underline{\mathbf{W}}$ inside the Sylvester constraint $\mathcal{SC}$ into the time domain. For each channel this is done by the transformation $\mathbf{F}_R^{-1}\Delta\underline{\mathbf{W}}_{pq}(i)\mathbf{F}_R$. Multiplying a DFT-domain diagonal matrix on both sides with the IDFT and DFT matrices leads to a time-domain circulant matrix as was already discussed in Section 3.4.2.3 and illustrated for the Sylvester matrix $\mathbf{W}_{pq}$ in Fig. 3.12. For the DFT-domain update $\Delta\underline{\mathbf{W}}_{pq}$ the $R \times R$ circulant matrix is given as

$$\mathbf{C}_{\Delta\mathbf{W}_{pq}} = \mathbf{F}_R^{-1}\Delta\underline{\mathbf{W}}_{pq}(i)\mathbf{F}_R. \tag{3.201}$$

The circulant matrix is then constrained by the window matrices $\mathbf{W}^{1_{2L}0}_{2L\times R}$ and $\mathbf{W}^{1_D0}_{R\times D}$ to a Toeplitz matrix of size $2PL \times D$. This results in a channel-wise Toeplitz structure of $\mathbf{V}_{2PL\times PR}\Delta\underline{\mathbf{W}}(i)\mathbf{V}^{\mathrm{H}}_{PD\times PR}$ which considerably simplifies the Sylvester constraint $\mathcal{SC}$. In the illustration of the Sylvester constraint $\mathcal{SC}$ in Fig. 3.2 it was shown that the operator $\mathcal{SC}$ picks the first L diagonals in each channel and then performs an averaging of the values on each diagonal. Due to the channel-wise Toeplitz structure the values on each diagonal are already identical. Thus, we can discard the averaging operation and only need to ensure that in each channel the values on the first L diagonals are picked. This allows to simplify the Sylvester constraint in the update equation leading to the *Sylvester constraint for narrowband algorithms*

$$\Delta\underline{\check{\mathbf{W}}}(m) = \frac{2}{R}\sum_{i=0}^{\infty}\beta(i,m)\mathbf{G}^{1_L0}_{PR\times PR}\Delta\underline{\mathbf{W}}(i)\mathbf{L}_{\mathbf{I}}, \tag{3.202}$$

where the constraint matrix

$$\begin{aligned}\mathbf{G}^{1_L0}_{PR\times PR} &= \mathbf{V}^{\mathrm{H}}_{PL\times PR}\mathbf{V}_{2PL\times PR}\\ &= \text{Bdiag}\left\{\mathbf{F}_R\mathbf{W}^{1_L0}_{R\times R}\mathbf{F}_R^{-1},\ldots,\mathbf{F}_R\mathbf{W}^{1_L0}_{R\times R}\mathbf{F}_R^{-1}\right\},\end{aligned} \tag{3.203}$$

and $\mathbf{L}_{\mathbf{I}} = \text{Bdiag}\left\{\mathbf{1}_{R\times 1},\ldots,\mathbf{1}_{R\times 1}\right\}$ is a block-diagonal matrix consisting of column vectors $\mathbf{1}_{R\times 1}$ containing only ones. A multiplication on the right-hand side with matrix $\mathbf{L}_{\mathbf{I}}$ converts the DFT-domain $R \times R$ diagonal submatrices of $\Delta\underline{\mathbf{W}}$ to $R \times 1$ column vectors containing the R DFT values. The left-hand side multiplication with the constraint matrix $\mathbf{G}^{1_L0}_{PR\times PR}$ transforms the DFT-domain demixing filter coefficients for each channel to

the time domain, sets all values larger than the filter length L to zero, and then goes back to the DFT domain. This procedure ensures that the elements on the first L diagonals are used as filter coefficients as desired by the Sylvester constraint $\mathcal{SC}$.

This allows to finally write the natural gradient update for the *narrowband HOS algorithm based on multivariate SIRP pdfs* as:

$$\Delta\underline{\check{\mathbf{W}}}(m) = \frac{2}{R}\sum_{i=0}^{\infty}\beta(i,m)\mathbf{G}^{1_L0}_{PR\times PR}\underline{\mathbf{W}}(i)\mathrm{boff}\left\{\underline{\mathbf{Y}}(i)\underline{\mathbf{Y}}^{\mathrm{H}}(i)\underline{\mathbf{\Lambda}}^{\mathrm{H}}(i)\right\}\mathrm{diag}^{-1}\left\{\underline{\mathbf{S}}_{\mathbf{yy}}(i)\right\}\mathbf{L_I} \tag{3.204a}$$

$$\underline{\mathbf{\Lambda}}(i) = \mathrm{Bdiag}\left\{\underline{\mathbf{\Lambda}}_1(i),\ldots,\underline{\mathbf{\Lambda}}_P(i)\right\} \tag{3.204b}$$

$$\underline{\mathbf{\Lambda}}_q(i) = \phi_{y_q,D}\left(\sum_{\nu=0}^{R-1}|\underline{Y}_q^{(\nu)}(i)|^2\right)\cdot\mathbf{I}_{R\times R} \tag{3.204c}$$

It can be seen that the coupling between the DFT bins is ensured in two ways. First, the nonlinearity $\underline{\mathbf{\Lambda}}$ depends on the output signals in all DFT bins as can be seen in (3.204c) and second, the constraint matrix $\mathbf{G}^{1_L0}_{PR\times PR}$ contains a windowing operation in the time-domain which also leads to a coupling of the individual DFT bins. The latter procedure appears similarly in the well-known "constrained frequency-domain adaptive filtering" in the supervised case [Hay02, BBK03] and allows to avoid in a simple way an independent permutation and scaling in each DFT bin.

In [Hir06, KALL06, KALL07] a similar algorithm was derived which is based on the narrowband model and avoids the independent permutation and scaling in each DFT bin by assuming multivariate SIRP pdfs. Their algorithms can be obtained if the constraint $\mathbf{G}^{1_L0}_{PR\times PR}$ is further approximated as a scaled identity matrix and if the inverse is assumed to be $\mathrm{diag}^{-1}\left\{\underline{\mathbf{S}}_{\mathbf{yy}}(i)\right\}\approx\mathbf{I}_{PR\times PR}$ leading to

$$\Delta\underline{\check{\mathbf{W}}}(m)\propto\sum_{i=0}^{\infty}\beta(i,m)\underline{\mathbf{W}}(i)\mathrm{boff}\left\{\underline{\mathbf{Y}}(i)\underline{\mathbf{Y}}^{\mathrm{H}}(i)\underline{\mathbf{\Lambda}}^{\mathrm{H}}(i)\right\}\mathbf{L_I}. \tag{3.205}$$

The nonlinearity in [Hir06, KALL06, KALL07] was calculated according to (3.204c) and the SIRP score was based on an approximation of the multivariate Laplacian SIRP pdf (3.100) leading to

$$\phi_{y_q,D}\left(\sum_{\nu=0}^{R-1}|\underline{Y}_q^{(\nu)}(i)|^2\right)=\frac{1}{\sqrt{\sum_{\nu=0}^{R-1}|\underline{Y}_q^{(\nu)}(i)|^2}}. \tag{3.206}$$

So here only the joint power in the nonlinearity is used to avoid the internal permutation. Alternatively, the SIRP score (3.103) which is based on the exact multivariate Laplacian SIRP pdf and which was proposed in [BAK03a] can be used.

Narrowband algorithms based on univariate pdfs

Above we have seen that by using a multivariate SIRP pdf a coupling between the DFT bins can be retained leading to a score function $\phi_{y_q,D}(\cdot)$ where the argument depends on all DFT bins. This method prevents the independent permutation and scaling ambiguity in each DFT bin as has been pointed out in, e.g., [BAK04a, Hir06, KEL06]. On the other hand, the overwhelming majority of traditional narrowband approaches are based on univariate pdfs which do not provide any coupling between the DFT bins and thus, the permutation and scaling ambiguity have to be solved by different methods.

In the following we will show how also such narrowband algorithms can be derived from the broadband HOS-SIRP update (3.168) expressed in the DFT-domain. The assumption of a univariate pdf has consequences for the nonlinear weighting $\mathbf{\Lambda}_q$ contained in the output signals $\underline{\mathbf{Y}}_\phi$ which was given in (3.188). In (3.191) the nonlinear weighting $\mathbf{\Lambda}_q$ based on the D-variate SIRP pdf was expressed by DFT domain quantities together with several constraint matrices and the normalization already expressed as a narrowband inverse. If instead a univariate pdf for each DFT bin is chosen, then all constraint matrices $\mathbf{G}_{R\times R}^{01_D}$ are approximated as identity matrices because no coupling between the DFT bins is preserved. This means that the D-variate SIRP pdf defined in the time-domain in (3.95) as $\hat{p}_{y_q,D}(\mathbf{y}_q)$ is approximated by univariate pdfs $\hat{\underline{p}}_{\underline{Y}_q,1}^{(\nu)}(\underline{Y}_q^{(\nu)})$ for each DFT bin which are given as

$$\hat{\underline{p}}_{\underline{Y}_q,1}^{(\nu)}(\underline{Y}_q^{(\nu)}) = a \cdot f_{\underline{Y}_q,1}\left(\frac{|\underline{Y}_q^{(\nu)}|^2}{\sigma_{\underline{Y}_q}^{(\nu)^2}}\right), \tag{3.207}$$

where a is a normalization factor, $\sigma_{\underline{Y}_q}^{(\nu)^2}$ is an estimate of the variance for the q-th output signal in the ν-th DFT bin ($\nu = 0,\ldots,R-1$), and $f_{\underline{Y}_q,1}(\cdot)$ is a scalar function which depends on the chosen distribution. Thus, the nonlinear weighting in (3.191) reduces to

$$\mathbf{\Lambda}_q(i) = \mathbf{W}_{N\times R}^{1_N 0}\mathbf{F}_R^{-1}\underline{\phi}_{y_q,1}\left(\underline{\mathbf{Y}}_q^{\mathrm{H}}(i)\underline{\mathbf{S}}_{\mathbf{y}_q\mathbf{y}_q}^{-1}(i)\underline{\mathbf{Y}}_q(i)\right)\mathbf{F}_R\mathbf{W}_{R\times N}^{1_N 0}, \tag{3.208}$$

where $\underline{\phi}_{y_q,1}(\cdot)$ is the transformed score function resulting from the univariate pdf given in (3.207). The score function $\underline{\phi}_{y_q,1}(\cdot)$ is applied independently to the individual DFT bins of the diagonal DFT-domain matrix product. Inserting (3.208) into the definition of the nonlinearly weighted output signals $\underline{\mathbf{Y}}_\phi$ given in (3.188) leads for the q-th channel to

$$\underline{\mathbf{Y}}_{\phi,q}^{\mathrm{H}} = \mathbf{G}_{R\times R}^{1_N 0}\,\underline{\phi}_{y_q,1}\left(\underline{\mathbf{Y}}_q^{\mathrm{H}}(i)\underline{\mathbf{S}}_{\mathbf{y}_q\mathbf{y}_q}^{-1}(i)\underline{\mathbf{Y}}_q(i)\right)\mathbf{G}_{R\times R}^{1_N 0}\underline{\mathbf{Y}}_q^{\mathrm{H}}. \tag{3.209}$$

By further approximating the constraint matrices $\mathbf{G}_{R\times R}^{1_N 0}$ as scaled identity matrices we obtain the nonlinearly weighted DFT-domain output signals $\underline{\mathbf{Y}}_{\phi,q}$ given as the $R\times R$ matrix

$$\underline{\mathbf{Y}}_{\phi,q}^{\mathrm{H}}(i) \approx \underline{\phi}_{y_q,1}\left(\underline{\mathbf{Y}}_q^{\mathrm{H}}(i)\underline{\mathbf{S}}_{\mathbf{y}_q\mathbf{y}_q}^{-1}(i)\underline{\mathbf{Y}}_q(i)\right)\underline{\mathbf{Y}}_q^{\mathrm{H}}(i) \tag{3.210}$$

and the combination of all channels is denoted as

$$\underline{\mathbf{Y}}_{\phi}(i) = [\underline{\mathbf{Y}}_{\phi,1}(i), \ldots, \underline{\mathbf{Y}}_{\phi,P}(i)]^{\mathrm{T}}. \tag{3.211}$$

Inserting (3.211) into the update equation (3.187) leads to

$$\Delta\underline{\check{\mathbf{W}}}(m) = \frac{2}{N}\sum_{i=0}^{\infty}\beta(i,m)\mathbf{V}_{PL\times PR}^{\mathrm{H}}\mathcal{SC}\left\{\mathbf{V}_{2PL\times PR}\Delta\underline{\mathbf{W}}(i)\mathbf{V}_{PD\times PR}^{\mathrm{H}}\right\}, \tag{3.212}$$

with the narrowband update $\Delta\underline{\mathbf{W}}$ defined as

$$\Delta\underline{\mathbf{W}}(i) \propto \underline{\mathbf{W}}(i)\mathrm{boff}\left\{\underline{\mathbf{Y}}(i)\underline{\mathbf{Y}}_{\phi}^{\mathrm{H}}(i)\right\}\mathrm{diag}^{-1}\left\{\underline{\mathbf{S}}_{\mathbf{yy}}(i)\right\}. \tag{3.213}$$

Analogously to the previous section in (3.202), we can replace the Sylvester constraint $\mathcal{SC}$ by a multiplication with $\mathbf{L_I}$ and the constraint $\mathbf{G}_{PR\times PR}^{1_L 0}$ yielding

$$\Delta\underline{\check{\mathbf{W}}}(m) = \frac{2}{N}\sum_{i=0}^{\infty}\beta(i,m)\mathbf{G}_{PR\times PR}^{1_L 0}\Delta\underline{\mathbf{W}}(i)\mathbf{L_I}. \tag{3.214a}$$

As all matrices in the narrowband update $\Delta\underline{\mathbf{W}}$ are diagonal we can analogously to the illustration in Fig. 3.16 decompose the $PR \times PR$ matrices $\underline{\mathbf{W}}$, $\underline{\mathbf{S}}_{\mathbf{y}_q\mathbf{y}_q}$ and the $PR \times R$ matrices $\underline{\mathbf{Y}}$, $\underline{\mathbf{Y}}_{\phi}$ into smaller $P \times P$ matrices or $P \times 1$ vectors for each DFT bin $\nu = 0, \ldots, R-1$. This leads to a bin-wise update $\Delta\underline{\mathbf{W}}^{(\nu)}$ given for the ν-th DFT bin as

$$\Delta\underline{\mathbf{W}}^{(\nu)}(i) \propto \underline{\mathbf{W}}^{(\nu)}(i)\mathrm{boff}\left\{\underline{\mathbf{Y}}^{(\nu)}(i)\left(\underline{\mathbf{Y}}_{\phi}^{(\nu)}(i)\right)^{\mathrm{H}}\right\}\mathrm{diag}^{-1}\left\{\underline{\mathbf{S}}_{\mathbf{yy}}^{(\nu)}(i)\right\}, \tag{3.214b}$$

where the diagonal matrix $\mathrm{diag}\{\underline{\mathbf{S}}_{\mathbf{yy}}^{(\nu)}\}$ contains the variances $\left(\sigma_{\underline{Y}_q}^{(\nu)}\right)^2$ of the output channels on its diagonal. The bin-wise decomposition of $\underline{\mathbf{Y}}_{\phi} = [\underline{\mathbf{Y}}_{\phi,1}, \ldots, \underline{\mathbf{Y}}_{\phi,P}]^{\mathrm{T}}$ is given by the bin-wise decomposition of the score function $\underline{\phi}_{y_q,1}(\cdot)$ in (3.210) to $\underline{\phi}_{y_q,1}^{(\nu)}(\cdot)$ as

$$\underline{Y}_{\phi,q}^{(\nu)}(i) = \underline{\phi}_{y_q,1}^{(\nu)}\left(|\underline{Y}_q^{(\nu)}|^2/\sigma_{\underline{Y}_q}^{(\nu)^2}\right)\underline{Y}_q^{(\nu)}(i), \tag{3.214c}$$

$$\underline{\phi}_{y_q,1}^{(\nu)}\left(|\underline{Y}_q^{(\nu)}|^2/\sigma_{\underline{Y}_q}^{(\nu)^2}\right) = -\frac{\partial \log f_{\underline{Y}_q,1}\left(|\underline{Y}_q^{(\nu)}|^2/\sigma_{\underline{Y}_q}^{(\nu)^2}\right)}{\partial(|\underline{Y}_q^{(\nu)}|^2/\sigma_{\underline{Y}_q}^{(\nu)^2})}. \tag{3.214d}$$

The equations (3.214a)-(3.214d) define the *narrowband HOS algorithm based on univariate pdfs*. It can be seen from the narrowband update equation (3.214b) and the nonlinear weighting (3.214d) that all DFT bins are adapted independently. The only coupling between the DFT bins is given by the constraint matrix in (3.214a) which corresponds to a transformation of the filter coefficients back to the time domain, zeroing the last $R - L$

values, and transforming the result back to the DFT domain. This procedure appears similarly in the well-known "constrained frequency-domain adaptive filtering" in the supervised case [Hay02, BBK03]. In BSS, this theoretically founded mechanism largely eliminates the internal permutation problem in a simple way. It was first heuristically introduced for narrowband BSS algorithms in [Sma98], and also in [PSV98, PS00]. A more detailed experimental examination on this constraint was reported in [IM00]. However, due to the omission of the other constraints in the approximated gradients we will not perfectly remove the permutation ambiguity as observed experimentally in [IM00]. Traditional narrowband approaches also neglecting the constraint matrix in (3.214a) need additional measures for solving the permutation ambiguity (e.g., [IM99, SMAM03]).

For the implementation of the algorithm (3.214a)-(3.214d) the choice of a suitable nonlinearity $\underline{\phi}^{(\nu)}_{y_q,1}$ in (3.214c) remains to be discussed. It was shown in [GZ03] that the multivariate Laplacian pdf is a good model for speech in the time domain. Additionally, their experimental evaluations showed that the distribution of speech can also be modeled in transform domains by a Laplacian distribution. The Laplacian pdf is given for the ν-th DFT bin as

$$\hat{\underline{p}}^{(\nu)}_{\underline{Y}_q,1}(\underline{Y}^{(\nu)}_q) = a \cdot e^{-\left(\frac{|\underline{Y}^{(\nu)}_q|}{\sigma^{(\nu)}_{\underline{Y}_q}}\right)}, \tag{3.215}$$

where a denotes a normalization term. This leads to a nonlinearly weighted output signal $\underline{Y}^{(\nu)}_{\phi,q}$ given for the ν-th DFT bin as

$$\underline{Y}^{(\nu)}_{\phi,q}(i) = \frac{1}{2}\sigma^{(\nu)}_{\underline{Y}_q}(i)\frac{\underline{Y}^{(\nu)}_q(i)}{|\underline{Y}^{(\nu)}_q(i)|}. \tag{3.216}$$

Noting that the complex sign function is given as $\mathrm{sign}(\underline{Y}^{(\nu)}_q) = \underline{Y}^{(\nu)}_q/|\underline{Y}^{(\nu)}_q|$, we can express (3.216) as

$$\underline{Y}^{(\nu)}_{\phi,q}(i) = \frac{1}{2}\sigma^{(\nu)}_{\underline{Y}_q}(i) \cdot \mathrm{sign}(\underline{Y}^{(\nu)}_q(i)). \tag{3.217}$$

The sign-function is a popular nonlinearity for narrowband BSS applied to speech signals (see, e.g., [HKO01, MD02]). Another very popular univariate pdf for narrowband BSS is given as

$$\hat{\underline{p}}^{(\nu)}_{\underline{Y}_q,1}(\underline{Y}^{(\nu)}_q) = a - \cosh\left(|\underline{Y}^{(\nu)}_q|/\sigma^{(\nu)}_{\underline{Y}_q}\right), \tag{3.218}$$

where a is a constant. This leads to a nonlinearly weighted output signal $\underline{Y}^{(\nu)}_{\phi,q}$ given for the ν-th DFT bin as

$$\underline{Y}^{(\nu)}_{\phi,q}(i) = \frac{1}{2}\sigma^{(\nu)}_{\underline{Y}_q}(i)\tanh\left(\frac{|\underline{Y}^{(\nu)}_q|}{\sigma^{(\nu)}_{\underline{Y}_q}}\right)\frac{\underline{Y}^{(\nu)}_q(i)}{|\underline{Y}^{(\nu)}_q(i)|}. \tag{3.219}$$

The nonlinear weighting is in (3.219) only applied to the absolute value of the output signal as has been proposed in [SMAM02]. This is an improved version of the original proposition in [Sma98] to extend the information maximization approach [BS95] by applying the $\tanh(\cdot)$ to the real and imaginary part of the output signal separately.

It should be noted that both nonlinear weightings (3.217) and (3.219) contain the normalization by the output signal variance. However, in most of the narrowband BSS literature this normalization is omitted as usually a prewhitening step in each DFT bin is included which performs a cross-channel decorrelation and scales the signals to unit variance, i.e., $\sigma_{\underline{Y}_q}^{(\nu)^2} = 1, \forall\, \nu$.

3.4.3.3 BSS based on second-order statistics

In the SOS case we can apply the same approximation steps as discussed for the HOS case in Section 3.4.3.2. At first we apply the narrowband normalization explained in Section 3.4.3.1 to the SOS natural gradient update expressed in the DFT domain. For an estimation of the correlation matrices by the *covariance method* additionally the constraint $\mathbf{G}_{R\times R}^{1_N 0}$ is approximated according to (3.179) as $\mathbf{G}_{R\times R}^{1_N 0} \approx N/R \cdot \mathbf{I}_{R\times R}$ so that the DFT-domain update (3.169) simplifies to

$$\begin{aligned}\Delta\check{\underline{\mathbf{W}}}(m) \;=\; \sum_{i=0}^{\infty} \beta(i,m)\mathbf{V}_{PL\times PR}^{\mathrm{H}} \mathcal{SC}\left\{\mathbf{V}_{2PL\times PR}\underline{\mathbf{W}}(i)\mathbf{G}_{PR\times PR}^{1_D 0}\mathrm{boff}\left\{\underline{\mathbf{Y}}(i)\underline{\mathbf{Y}}^{\mathrm{H}}(i)\right\}\right.\\ \left.\mathbf{G}_{PR\times PR}^{1_D 0}\mathrm{diag}^{-1}\left\{\underline{\mathbf{Y}}(i)\underline{\mathbf{Y}}^{\mathrm{H}}(i)\right\}\mathbf{V}_{PD\times PR}^{\mathrm{H}}\right\}, \end{aligned} \tag{3.220}$$

with the constraint matrix $\mathbf{G}_{PR\times PR}^{1_D 0}$ given as

$$\mathbf{G}_{PR\times PR}^{1_D 0} = \mathrm{Bdiag}\left\{\mathbf{F}_R^{-1}\mathbf{W}_{R\times R}^{01_D}\mathbf{F}_R, \ldots, \mathbf{F}_R^{-1}\mathbf{W}_{R\times R}^{01_D}\mathbf{F}_R\right\}. \tag{3.221}$$

For an estimation of the correlation matrices by the *correlation method*, the matrices $\underline{\mathbf{Y}}$ are simply replaced by $\underline{\tilde{\mathbf{Y}}}$ as defined in (3.160) leading to

$$\begin{aligned}\Delta\check{\underline{\mathbf{W}}}(m) \;=\; \sum_{i=0}^{\infty} \beta(i,m)\mathbf{V}_{PL\times PR}^{\mathrm{H}} \mathcal{SC}\left\{\mathbf{V}_{2PL\times PR}\underline{\mathbf{W}}(i)\mathbf{G}_{PR\times PR}^{1_D 0}\mathrm{boff}\left\{\underline{\tilde{\mathbf{Y}}}(i)\underline{\tilde{\mathbf{Y}}}^{\mathrm{H}}(i)\right\}\right.\\ \left.\mathbf{G}_{PR\times PR}^{1_D 0}\mathrm{diag}^{-1}\left\{\underline{\tilde{\mathbf{Y}}}(i)\underline{\tilde{\mathbf{Y}}}^{\mathrm{H}}(i)\right\}\mathbf{V}_{PD\times PR}^{\mathrm{H}}\right\}, \end{aligned} \tag{3.222}$$

The algorithms (3.220) and (3.222) preserve most constraints and only the normalization is performed in a narrowband manner. Thus, narrowband efficiency is combined with avoiding the permutation and scaling ambiguity in each DFT bin due to the formulation of the remaining algorithm in a broadband manner. The constraints express the overlap-save

procedure which ensures that linear convolutions are computed. This hybrid algorithm has been proposed in [ABK06a, ABK06b] and the evaluations in Section 3.6 will show its good performance.

The narrowband approach is obtained by further approximating the constraint $\mathbf{G}_{PR\times PR}^{1_D 0} \approx D/R \cdot \mathbf{I}_{PR\times PR}$. This allows again a simplification of the Sylvester constraint $\mathcal{SC}$ as given in (3.202) and leads together with the DFT-domain update $\Delta\underline{\mathbf{W}}$ to the *narrowband natural gradient SOS BSS algorithm*

$$\Delta\underline{\check{\mathbf{W}}}(m) = \sum_{i=0}^{\infty} \beta(i,m)\mathbf{G}_{PR\times PR}^{1_L 0}\Delta\underline{\mathbf{W}}(i)\mathbf{L}_{\mathbf{I}}, \tag{3.223a}$$

$$\Delta\underline{\mathbf{W}}(i) \propto \underline{\mathbf{W}}(i)\mathrm{boff}\left\{\underline{\mathbf{Y}}(i)\underline{\mathbf{Y}}^{\mathrm{H}}(i)\right\}\mathrm{bdiag}^{-1}\left\{\underline{\mathbf{Y}}(i)\underline{\mathbf{Y}}^{\mathrm{H}}(i)\right\}. \tag{3.223b}$$

Analogously to the HOS narrowband algorithm in the previous section, we can decompose the DFT-domain update $\Delta\underline{\mathbf{W}}$ to a bin-wise formulation yielding

$$\Delta\underline{\mathbf{W}}^{(\nu)}(i) \propto \underline{\mathbf{W}}^{(\nu)}(i)\mathrm{boff}\left\{\underline{\mathbf{Y}}^{(\nu)}(i)\left(\underline{\mathbf{Y}}^{(\nu)}(i)\right)^{\mathrm{H}}\right\}\mathrm{bdiag}^{-1}\left\{\underline{\mathbf{Y}}^{(\nu)}(i)\left(\underline{\mathbf{Y}}^{(\nu)}(i)\right)^{\mathrm{H}}\right\}, \tag{3.224}$$

with $\nu = 0,\ldots,R-1$ denoting the DFT bin index. The complete decoupling of the DFT bins is again prevented by the constraint $\mathbf{G}_{PR\times PR}^{1_L 0}$ given in (3.223a). This constraint which consists of a channel-wise IDFT, windowing operation, and DFT was first heuristically proposed for SOS BSS algorithms in [PSV98, PS00]. The narrowband algorithm (3.223a), (3.223b) is equivalent to a natural gradient BSS algorithm proposed in [WP99] derived from a narrowband optimization criterion. Additionally, from this optimization criterion also a gradient version was derived in [WP99] and more recently Fancourt and Parra in [FP01a] obtained the same algorithm by using the magnitude-squared coherence (MSC) function as a criterion. This indicates that a relationship between the broadband TRINICON optimization criterion and the magnitude-squared coherence function can be established as will be shown in the following section.

3.4.3.4 Relationship of narrowband second-order BSS and the magnitude-squared coherence function

To obtain a link between the magnitude-squared coherence (MSC) function, we need to obtain first a SOS optimization criterion based on the broadband TRINICON optimization criterion (3.43). To this end, the joint densities $\hat{p}_{y,PD}(\cdot)$ and $\hat{p}_{y_q,D}(\cdot)$ in (3.43) are

assumed to be multivariate Gaussian pdfs (3.110) yielding

$$\tilde{\mathcal{J}}(i,\mathbf{W}) = \frac{1}{2N}\sum_{j=0}^{N-1}\left(\sum_{q=1}^{P}\left(\log\det\,\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}(i)+\mathbf{y}_q^{\mathrm{H}}(iL+j)\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}^{-1}(i)\mathbf{y}_q(iL+j)\right)\right.$$
$$\left.-\log\det\,\mathbf{R}_{\mathbf{yy}}(i)-\mathbf{y}^{\mathrm{H}}(iL+j)\mathbf{R}_{\mathbf{yy}}^{-1}(i)\mathbf{y}(iL+j)\right). \quad (3.225)$$

Exploiting the fact that the determinant of a block-diagonal matrix is given as [Har97]

$$\det\left\{\begin{bmatrix}\mathbf{A}_{11} & \mathbf{0} & \dots & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_{22} & & \mathbf{0}\\ \vdots & \vdots & \ddots & \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{A}_{PP}\end{bmatrix}\right\} = \det\{\mathbf{A_{11}}\}\cdot\det\{\mathbf{A_{22}}\}\cdot\ldots\cdot\det\{\mathbf{A}_{PP}\}, \quad (3.226)$$

we can rewrite the term $\sum_{q=1}^{P}\log\det\{\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}\} = \log\det\{\mathrm{bdiag}\{\mathbf{R}_{\mathbf{yy}}\}\}$. Using the matrix $\mathbf{Y}_q(i)$ defined in (3.69) which contains the N column vectors $\mathbf{y}_q(iL+j)$ for $j=0,\ldots,N-1$ and the correlation matrix $\mathbf{R}_{\mathbf{yy}}(i)=\frac{1}{N}\mathbf{Y}(i)\mathbf{Y}^{\mathrm{H}}(i)$ from (3.71) we can rewrite the quadratic form as

$$\begin{aligned}\frac{1}{N}\sum_{j=0}^{N-1}\mathbf{y}_q^{\mathrm{H}}(iL+j)\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}^{-1}(i)\mathbf{y}_q(iL+j) &= \mathrm{tr}\left\{\mathbf{Y}_q^{\mathrm{H}}(i)\left(\mathbf{Y}_q(i)\mathbf{Y}_q^{\mathrm{H}}(i)\right)^{-1}\mathbf{Y}_q(i)\right\}\\ &= \mathrm{tr}\left\{\mathbf{Y}_q(i)\mathbf{Y}_q^{\mathrm{H}}(i)\left(\mathbf{Y}_q(i)\mathbf{Y}_q^{\mathrm{H}}(i)\right)^{-1}\right\}\\ &= \mathrm{tr}\left\{\mathbf{I}_{D\times D}\right\} = D. \quad (3.227)\end{aligned}$$

For the second line of (3.227) we exploited the fact that the matrices inside the trace operator may be circularly shifted, i.e., $\mathrm{tr}\{\mathbf{ABC}\}=\mathrm{tr}\{\mathbf{CAB}\}$ [Har97]. Analogously, the quadratic form $\frac{1}{N}\sum_{j=0}^{N-1}\mathbf{y}^{\mathrm{H}}(iL+j)\mathbf{R}_{\mathbf{yy}}^{-1}(i)\mathbf{y}(iL+j)$ can be expressed using $\mathbf{Y}(i)$ defined in (3.72) which leads to the constant value PD. Using these results, the SOS optimization criterion (3.225) simplifies to

$$\mathcal{J}(m,\mathbf{W}) = \frac{1}{2}\sum_{i=0}^{m}\beta(i,m)\left(\log\det\{\mathrm{bdiag}\{\mathbf{R}_{\mathbf{yy}}\}\}-\log\det\{\mathbf{R}_{\mathbf{yy}}\}\right). \quad (3.228)$$

The same broadband optimization criterion was proposed in [ABK03, BAK03b, BAK05a] where it was not derived from the TRINICON optimization criterion (3.43) but motivated as the extension of the criterion proposed in [MOK95, KMO98] to incorporate D time lags.

To see the link to the MSC we derive a narrowband optimization criterion from (3.228) by expressing it in the DFT domain and then approximating the constraint matrices as

identity matrices as shown in the previous sections. This leads to an optimization criterion

$$\mathcal{J}^{(\nu)}(m,\underline{\mathbf{W}}^{(\nu)}) = \frac{1}{2}\sum_{i=0}^{m}\beta(i,m)\left(\log\det\left\{\mathrm{diag}\{\underline{\mathbf{S}}_{\mathbf{yy}}^{(\nu)}(i)\}\right\} - \log\det\left\{\underline{\mathbf{S}}_{\mathbf{yy}}^{(\nu)}(i)\right\}\right) \tag{3.229}$$

applied independently to each DFT bin $\nu = 0,\ldots,R-1$. The $P\times P$ power-spectral density matrix in the ν-th DFT bin is defined as $\underline{\mathbf{S}}_{\mathbf{yy}}^{(\nu)} = \underline{\mathbf{Y}}^{(\nu)}\left(\underline{\mathbf{Y}}^{(\nu)}\right)^{\mathrm{H}}$ with its entries denoted as $\underline{S}_{y_py_q}^{(\nu)}$, $p,q\in\{1,\ldots,P\}$. The standard MSC was already defined and discussed in Section 2.2.4 and is given as

$$|\underline{\Gamma}_{y_py_q}^{(\nu)}(m)|^2 = \frac{|\underline{S}_{y_py_q}^{(\nu)}(m)|^2}{\underline{S}_{y_py_p}^{(\nu)}(m)\underline{S}_{y_qy_q}^{(\nu)}(m)} \tag{3.230}$$

with $p,q\in\{1,2\}$. It can be extended to more than two channels by the generalized coherence function [GC88]

$$|\underline{\Gamma}_{\mathbf{yy}}^{(\nu)}(m)|^2 = 1 - \frac{\det\{\underline{\mathbf{S}}_{\mathbf{yy}}^{(\nu)}(m)\}}{\prod_{p=1}^{P}\underline{S}_{y_py_p}^{(\nu)}(m)}, \tag{3.231}$$

which is valid for an arbitrary number P of channels and is equal to (3.230) for $P=2$. To show the relationship between the generalized coherence and (3.229), the optimization criterion has to be reformulated leading to

$$\begin{aligned}\mathcal{J}^{(\nu)}(m,\underline{\mathbf{W}}^{(\nu)}) &= \sum_{i=0}^{m}\beta(i,m)\left(\log\prod_{p=1}^{P}\underline{S}_{y_py_p}^{(\nu)}(i) - \log\det\{\underline{\mathbf{S}}_{\mathbf{yy}}^{(\nu)}(i)\}\right)\\ &= \sum_{i=0}^{m}\beta(i,m)\left(-\log\frac{\det\{\underline{\mathbf{S}}_{\mathbf{yy}}^{(\nu)}(i)\}}{\prod_{p=1}^{P}\underline{S}_{y_py_p}^{(\nu)}(i)}\right).\end{aligned} \tag{3.232}$$

A Taylor approximation

$$-\log(x) = (1-x) + \frac{(1-x)^2}{2} + \frac{(1-x)^3}{3} + \cdots$$

around $x=1$ for $0<x\leq 2$ finally yields

$$\mathcal{J}^{(\nu)}(m,\underline{\mathbf{W}}^{(\nu)}) = \sum_{i=0}^{m}\beta(i,m)\left(1-\frac{\det\{\underline{\mathbf{S}}_{\mathbf{yy}}^{(\nu)}(i)\}}{\prod_{p=1}^{P}\underline{S}_{y_py_p}^{(\nu)}(i)}\right), \tag{3.233}$$

which shows that the narrowband SOS optimization criterion can directly be seen as a minimization of the generalized coherence function [BAK03b].

Both, the MSC and the generalized coherence function satisfy the desirable property

$$0\leq|\underline{\Gamma}_{y_py_q}^{(\nu)}(m)|^2, |\underline{\Gamma}_{\mathbf{yy}}^{(\nu)}(m)|^2\leq 1, \tag{3.234}$$

which is very suitable for an optimization criterion as it directly translates into an inherent stepsize normalization of the corresponding update equation as can be seen, e.g., in (3.223b). Especially for colored signals such as speech, this normalization leads to improved performance compared to traditional BSS algorithms based on the Frobenius norm where usually a heuristic normalization is added (see, e.g., [PS00]). The performance gain due to the normalization has also been exploited in [FP01a] where the sum of all possible MSCs between the P output channels has been proposed as an optimization criterion for narrowband BSS. The criterion in [FP01a] is equivalent to (3.233) for $P = 2$, while for $P > 2$ the criterion (3.233) resulting from the broadband TRINICON optimization criterion is slightly more general.

The conclusion that the narrowband SOS BSS algorithm derived from the TRINICON framework is based on the generalized coherence allowed to establish the link to the algorithm in [FP01a]. Additionally, it shows that the influence of different parameters on the estimation of the magnitude-squared coherence as discussed in Section 2.2.4 also applies to the estimation of the narrowband SOS BSS optimization criterion and the algorithms following from it.

3.4.4 Summary

In Section 3.4 the aim was to formulate the broadband BSS algorithms derived from the TRINICON optimization criterion *equivalently* in the DFT domain and subsequently *introduce selective approximations* to obtain efficient algorithms. As shown in the flowchart in Fig. 3.17, this has been realized by firstly introducing the distinction between broadband and narrowband signal model in Section 3.4.1 based on the linear convolution in matrix notation which is expressed as a multiplication of a Toeplitz matrix with a vector. The broadband model yields the linear convolution expressed as an overlap-save procedure in matrix notation resulting in a multiplication of DFT-domain vectors or matrices together with so-called constraint matrices consisting of DFT, IDFT, and windowing operations. The narrowband model can be obtained from the broadband version by approximating the constraint matrices.

In Section 3.4.2 based on the two signal models a general procedure was given to express the broadband BSS update equations, which were derived in the time domain, equivalently in the DFT domain. The approach is to firstly express the time-domain variables appearing in the BSS update equations in terms of Toeplitz matrices (Section 3.4.2.1) and then express the Toeplitz matrices as DFT-domain quantities by using the broadband signal model (Section 3.4.2.3). Additionally, in Section 3.4.2.2 the iterative update rule was expressed in the DFT-domain. These were the prerequisites for expressing the higher-order and second-order statistics realizations of the generic TRINICON BSS algorithm equivalently in the DFT-domain in Sections 3.4.2.4 and 3.4.2.5.

The formulation of the algorithms in the DFT domain resulted in several constraint matrices which have been selectively approximated in Section 3.4.3 and led to recently published and also well-known algorithms from the literature. A selective approximation allows to gain computational efficiency but at the same time still avoids the permutation and scaling in each DFT bin. The main computational complexity arises from the normalization given for each channel by the inverse of the auto-correlation matrix. Therefore, a narrowband normalization together with a regularization scheme has been discussed in Section 3.4.3.1. Subsequently, by approximating additional constraints several efficient HOS and SOS BSS algorithms have been derived in Sections 3.4.3.2 and 3.4.3.3. The coupling between the DFT bins has either been assured by a last remaining constraint or by still considering multivariate pdfs instead of univariate pdfs for each DFT bin. This avoided the permutation and scaling ambiguity to a large extent. If the updates in the DFT bins are fully decoupled then the permutation and scaling problem have to be solved by applying other repair mechanisms developed for narrowband algorithms. Finally, in Section 3.4.3.4 a relationship between SOS narrowband BSS algorithms and the generalized coherence function have been given. This shows that the influence of different parameters on the estimation of the magnitude-squared coherence as discussed in Section 2.2.4 also applies to the estimation of the narrowband SOS BSS algorithms.

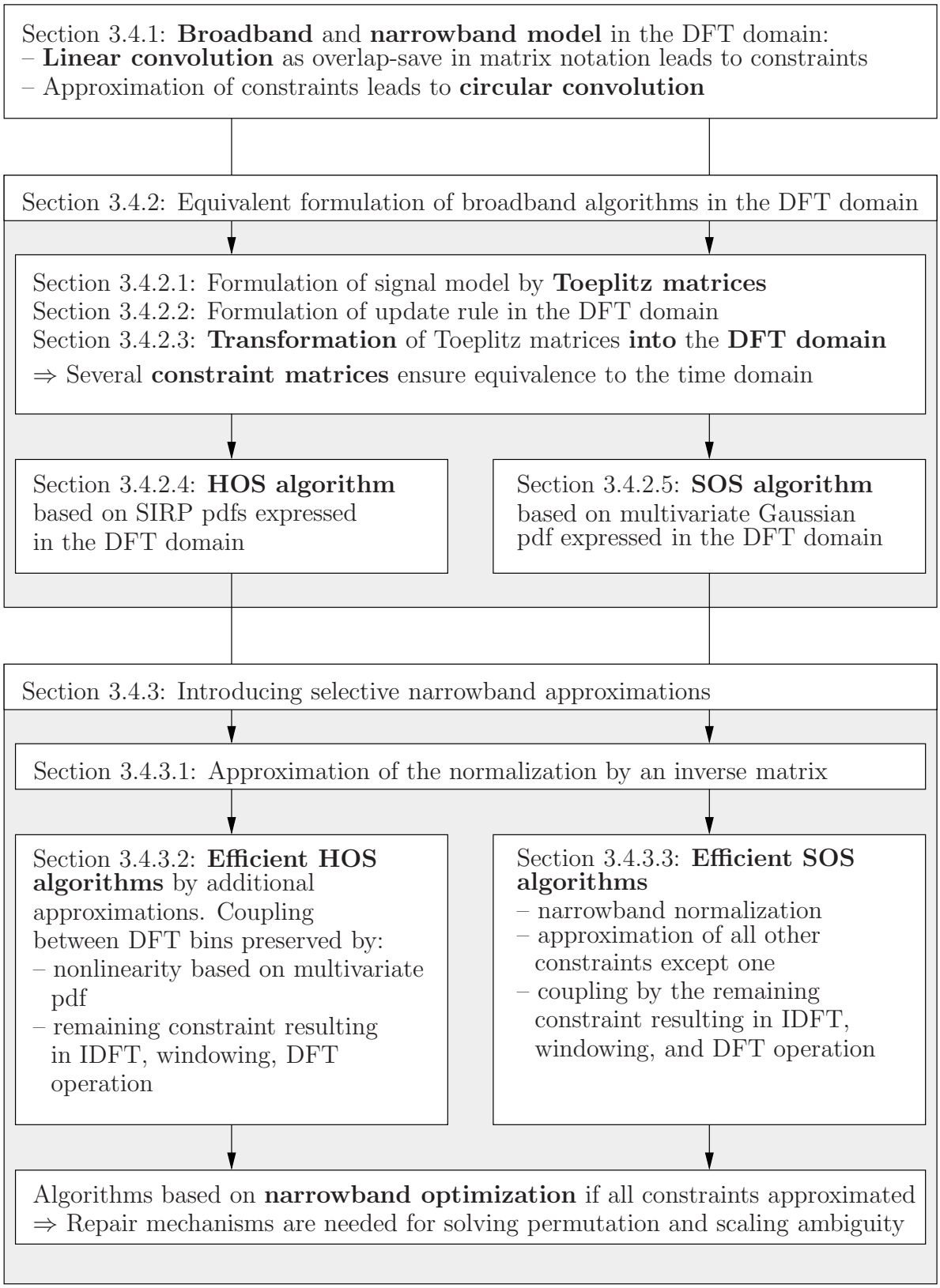


Figure 3.17: Flowchart showing the relations between the generic TRINICON update expressed in the DFT domain based on a broadband model and its various narrowband approximations leading to efficient BSS algorithms.

3.5 Algorithm formulation for different update strategies

In the TRINICON optimization criterion (3.43) a weighting function $\beta(i,m)$, with the block time indices i,m and with finite support normalized according to $\sum_{i=0}^{\infty}\beta(i,m)=1$, was introduced to allow different realizations of the algorithms. Based on (3.43) different update rules in the time-domain and DFT domain have been derived in the previous sections. The different algorithm adaptation possibilities will be shown in this section exemplarily for the iterative time-domain coefficient update given in (3.58) so that e.g. Newton-based methods are not considered here although they are possible (see [Buc]). All updates in this thesis are based on the natural gradient of the optimization criterion $\mathcal{J}(i,\mathbf{W})$ and thus the coefficient update in the m-th block can be expressed as

$$\begin{aligned}\check{\mathbf{W}}(m) &= \check{\mathbf{W}}(m-1)-\mu\Delta\check{\mathbf{W}}(m)\\ &= \check{\mathbf{W}}(m-1)-\mu\nabla^{\mathrm{NG}}_{\check{\mathbf{W}}}\mathcal{J}(i,\mathbf{W})\\ &= \check{\mathbf{W}}(m-1)-\mu\sum_{i=0}^{\infty}\beta(i,m)\nabla^{\mathrm{NG}}_{\check{\mathbf{W}}}\tilde{\mathcal{J}}(i,\mathbf{W}),\end{aligned} \tag{3.235}$$

where $\tilde{\mathcal{J}}(i,\mathbf{W})$ denotes the optimization criterion for the i-th block and was defined in (3.43). In the following we distinguish three different types of weighting functions $\beta(i,m)$ for offline, online, and block-online realizations [BAK04a].

3.5.1 Offline update

When realizing the algorithm as an offline or so-called batch algorithm, then $\beta(i,m)$ corresponds to a rectangular window (Fig. 3.18), which is described by

$$\beta(i,m)=\frac{1}{K_{\mathrm{sig}}}\epsilon_{0,(K_{\mathrm{sig}}-1)}(i), \tag{3.236}$$

where $\epsilon_{a,b}(i)$ is a rectangular window function, i.e., $\epsilon_{a,b}(i)=1$ for $a\le i\le b$, and $\epsilon_{a,b}(i)=0$ elsewhere. The entire signal is segmented into K_{sig} blocks, and then it is processed to estimate the demixing matrix $\mathbf{W}^{\ell}$, where the superscript ℓ denotes the current iteration. This leads to the coefficient update

$$\check{\mathbf{W}}^{\ell}=\check{\mathbf{W}}^{\ell-1}-\frac{\mu}{K_{\mathrm{sig}}}\sum_{i=0}^{K_{\mathrm{sig}}-1}\nabla^{\mathrm{NG}}_{\check{\mathbf{W}}}\tilde{\mathcal{J}}(i,\mathbf{W}), \tag{3.237}$$

where due to the offline processing the update does not depend on the block-time index anymore, but on the iteration index ℓ. Hence, the algorithm is generally visiting all the signal data repeatedly for each iteration ℓ and therefore, it usually achieves a better performance compared to its online counterpart.

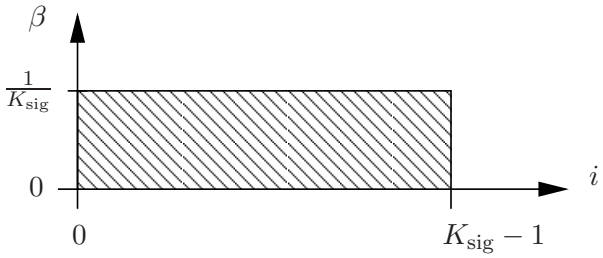


Figure 3.18: Weighting function $\beta(i,m)$ for offline implementation.

3.5.2 Online update

In time-variant environments an online implementation of (3.235) is required. An efficient realization can be achieved by using a weighting function with an exponential forgetting

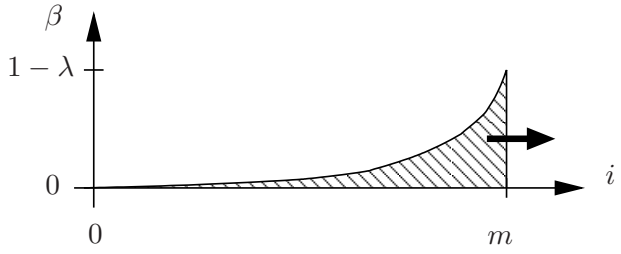


Figure 3.19: Weighting function $\beta(i,m)$ for online implementation.

factor λ (Fig. 3.19). It is defined by

$$\beta(i,m) = (1-\lambda)\lambda^{m-i}\epsilon_{0,m}(i), \tag{3.238}$$

where $0 \leq \lambda < 1$ and m denotes the current block. Inserting (3.238) in (3.235) yields

$$\check{\mathbf{W}}(m) = \check{\mathbf{W}}(m-1) - \mu \underbrace{(1-\lambda)\sum_{i=0}^{m}\lambda^{m-i}\nabla_{\check{\mathbf{W}}}^{\text{NG}}\tilde{\mathcal{J}}(i,\mathbf{W})}_{=\Delta\check{\mathbf{W}}(m)}. \tag{3.239}$$

Additionally, the update $\Delta\check{\mathbf{W}}(m)$ in (3.239) can be formulated recursively as

$$\begin{aligned}\Delta\check{\mathbf{W}}(m) &= (1-\lambda)\left(\lambda\sum_{i=0}^{m-1}\left(\lambda^{m-1-i}\nabla_{\check{\mathbf{W}}}^{\text{NG}}\tilde{\mathcal{J}}(i,\mathbf{W})\right) + \nabla_{\check{\mathbf{W}}}^{\text{NG}}\tilde{\mathcal{J}}(m,\mathbf{W})\right)\\ &= \lambda\Delta\check{\mathbf{W}}(m-1) + (1-\lambda)\nabla_{\check{\mathbf{W}}}^{\text{NG}}\tilde{\mathcal{J}}(m,\mathbf{W}).\end{aligned} \tag{3.240}$$

This reduces computational complexity and memory requirements since only the preceding update $\Delta\check{\mathbf{W}}(m-1)$ has to be saved. The recursive formulation thus leads to the following online coefficient update:

$$\check{\mathbf{W}}(m) = \check{\mathbf{W}}(m-1) - \mu \left(\lambda \Delta \check{\mathbf{W}}(m-1) + (1-\lambda) \nabla^{\mathrm{NG}}_{\check{\mathbf{W}}} \tilde{\mathcal{J}}(m, \mathbf{W}) \right) \tag{3.241}$$

It can be seen that the forgetting factor λ determines the memory of the algorithm update. A rule of thumb is that a rectangular window with the length $(1+\lambda)/(1-\lambda)$ yields approximately the same estimate as the exponential window with forgetting factor λ. This has been proven for the estimation of quasi-stationary auto-regressive Gaussian processes in [Bor85]. For the special case $\lambda = 0$ we have

$$\check{\mathbf{W}}(m) = \check{\mathbf{W}}(m-1) - \mu \nabla^{\mathrm{NG}}_{\check{\mathbf{W}}} \tilde{\mathcal{J}}(m, \mathbf{W}), \tag{3.242}$$

which corresponds to $\beta(i,m) = \delta(i-m)$.

3.5.3 Block-online update

The online adaptation presented in (3.241) allows for a real-time implementation, however, for rapidly time-variant mixing systems the convergence of the natural gradient algorithms may not be sufficient. Therefore, it is desirable to combine the improved convergence of offline adaptation with the online adaptation. Similarly to supervised adaptive filtering, where the weighting function $\beta(i,m) = (1-\lambda)\lambda^{m-i}\epsilon_{0,m}(i)$ allows to derive the recursive least-squares (RLS) algorithm, i.e., the online solution, from the corresponding offline least-squares (LS) solution [SS89, Hay02], we want to use the same methodology to obtain a *recursive* block-by-block solution based on the offline adaptation (where all data is required) given by

$$\check{\mathbf{W}}^{\ell}(m) = \check{\mathbf{W}}^{\ell-1}(m) - \mu \Delta \check{\mathbf{W}}^{\ell}(m), \quad \ell = 1, \ldots, \ell_{\max}. \tag{3.243}$$

The superscript ℓ denotes again the iteration number, μ is the stepsize and the update $\Delta \check{\mathbf{W}}^{\ell}(m)$ corresponds for the natural gradient to $\nabla^{\mathrm{NG}}_{\check{\mathbf{W}}} \mathcal{J}(m, \mathbf{W}^{\ell-1}(m))$. Here, the weighting function $\beta(i,m)$ is chosen as

$$\beta(i,m) = \frac{1-\lambda}{K} \sum_{m'=0}^{m} \lambda^{m-m'} \epsilon_{m'K, m'K+K-1}(i), \tag{3.244}$$

and is shown in Fig. 3.20. The horizontal axis shows the block index i with each block having a length of N samples. In the previously discussed offline and online adaptation methods the variable m denoted the current block of length N. When specifying $\beta(i,m)$ as given in (3.244), the current block m is of length $KL+N-L$ samples as it contains K subsequent blocks of length N with a blockshift of L samples each to allow the exploitation of the nonstationarity. As shown in Appendix B.3 we can derive an approximate recursive formulation of the offline update (3.243) by using the weighting function $\beta(i,m)$ given in (3.244). This leads to a so-called block-online method where an online update and

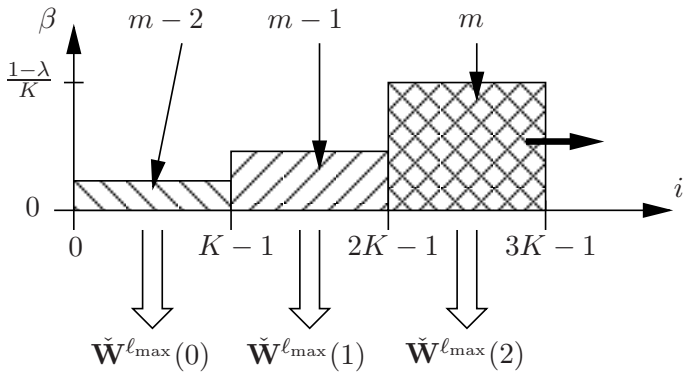


Figure 3.20: Weighting function $\beta(i,m)$ for block-online implementation at block time $m = 2$.

an offline update are combined similar to the approach in [MSAM03]. The advantage of this approach is that it allows a faster convergence and better tracking behavior than the online algorithm at moderate computational complexity.

According to Appendix B.3, the *offline part* is calculated iteratively for the current block m without exploiting any previous blocks (see Fig. 3.20) as,

$$\check{\mathbf{W}}^{\ell}(m) = \check{\mathbf{W}}^{\ell-1}(m) - \mu\frac{1}{K}\sum_{i=mK}^{mK+K-1}\nabla^{\mathrm{NG}}_{\check{\mathbf{W}}}\tilde{\mathcal{J}}(i,\mathbf{W}^{\ell-1}(i)). \tag{3.245}$$

where $\check{\mathbf{W}}^{\ell}(m)$ is the demixing filter matrix after ℓ iterations ($\ell = 1,\ldots,\ell_{\max}$) based on data of the m-th block. Equation (3.245) contains K update terms $\nabla^{\mathrm{NG}}_{\check{\mathbf{W}}}\tilde{\mathcal{J}}(i,\mathbf{W}^{\ell-1}(i))$ which are determined by using one of the algorithms derived in the previous section. This simultaneous optimization for K blocks allows to exploit the nonstationarity of the source signals as for each block the source statistics may change and thus, new support for the adaptation of the coefficient vector space may be added. A large number of iterations $\ell_{\max}$ allows a fast convergence of the natural gradient descent without introducing an additional algorithmic delay but at the cost of an increased computational complexity. In practice, the maximum number of iterations $\ell_{\max}$ is usually chosen to $5\ldots10$ iterations to keep the complexity at a moderate level. An analysis of the trade-off between convergence and maximum number of iterations can be found in [ASR+06] whose authors implemented a TRINICON-based algorithm together with the block-online adaptation procedure. Additionally, in Section 3.6.4 the influence of the number of offline iterations will be examined.

The demixing filter matrix $\check{\mathbf{W}}^{\ell_{\max}}(m)$ of the current block m which is obtained from the offline part after $\ell_{\max}$ iterations (see Fig. 3.20) is then used as input of the *online part* of the block-online algorithm which is written recursively as

$$\check{\mathbf{W}}(m) = \lambda \check{\mathbf{W}}(m-1) + (1-\lambda)\check{\mathbf{W}}^{\ell_{\max}}(m). \tag{3.246}$$

This yields the final demixing filter matrix $\check{\mathbf{W}}(m)$ of the current block m containing the filter weights $\mathbf{w}_{pq}(\mathrm{m})$ used for separation. The demixing filter weights $\mathbf{w}_{pq}(m)$ of the current block are then used as initial values for the offline algorithm (3.245) of the next block.

Analogously to supervised block-based adaptive filtering [MAG95], the approach followed here can also be carried out with overlapping data blocks in both, the online and offline part to further increase the convergence rate and to reduce the signal delay. The ratio of the number of previously used samples to the number of previously unseen samples is given by the overlap factors α_{off} for the offline part and α_{on} for the online part. These parameters are in the range $1 \leq \alpha_{\text{off}}, \alpha_{\text{on}} \leq L$ and should be chosen suitably to obtain integer values for the time index.

3.5.4 Adaptive stepsize techniques for block-online updates

In general the choice of the stepsize is very important for the performance of adaptive algorithms. In an offline processing scheme several trials can be run to establish the optimum stepsize. However, for an online procedure usually the stepsize has to be chosen very conservatively to prevent instability problems due to noise dependency, time-variance, etc. To make the adaptation more robust in real-world environments a stepsize control is desirable. In supervised adaptive filtering usually a closed-form solution for the stepsize is derived based on an observable reference signal. So far, in the BSS community there is little literature on this topic due to the absence of a reference signal. In the instantaneous mixing case some adaptive stepsize methods have been proposed (e.g., [DC98, SG00]) which are mainly relying on second order derivatives. However, for the convolutive mixing case such gradient stepsizes are computationally complex. In the neural networks community iterative methods for stepsize determination based on online measurements of the state of the adaptive system are more common and can be found in textbooks as, e.g., [CU94, Roj96]. We propose to use a simple but effective strategy for updating the stepsize based on a method presented in [VMR+88], [CU94, p. 146]. The procedure is to increase the stepsize if the value of the cost function $\tilde{\mathcal{J}}(i, \mathbf{W})$ is decreased compared to $\tilde{\mathcal{J}}(i-1, \mathbf{W})$ (indicating convergence) and to decrease it rapidly if the current value of $\tilde{\mathcal{J}}(i, \mathbf{W})$ exceeds the previous one $\tilde{\mathcal{J}}(i-1, \mathbf{W})$, by more than a pre-specified ratio (indicating divergence). In the latter case the current demixing filter update may be discarded ($\Delta\check{\mathbf{W}}(i) = 0$). After starting with a small stepsize $\mu(0)$ its modifications are described by

$$\mu(i+1) = \begin{cases} a \cdot \mu(i) & \text{if } \tilde{\mathcal{J}}(i, \mathbf{W}) < \tilde{\mathcal{J}}(i-1, \mathbf{W}) \quad , a > 1 \\ b \cdot \mu(i) & \text{if } \tilde{\mathcal{J}}(i, \mathbf{W}) \geq c \cdot \tilde{\mathcal{J}}(i-1, \mathbf{W}) \quad , b < 1, c > 1 \\ \mu(i-1) & \text{otherwise} \end{cases} \tag{3.247}$$

where for our application the values $a = 1.1$, $b = 0.5$, $c = 1.3$ provided robust behavior for all evaluated parameter settings. Moreover, to avoid instabilities, in practice the adaptive stepsize should be restricted to a finite range $[\mu_{\min}, \mu_{\max}]$. In (3.247) the cost function $\tilde{\mathcal{J}}$ defined in (3.43) has to be evaluated for each block i. It can be seen in (3.43) that in general this involves the estimation of multivariate pdfs. However, in practice the efficient algorithms based on approximations of the generic TRINICON-based nonholonomic natural gradient algorithm (3.64) will be used. Thus, different quantities which allow the assessment of the separation performance, such as the cross-correlation sequences between the output channels may be used instead of evaluating the original cost function. In the experiments in Section 3.6 we use the Frobenius norm of the cross-correlation matrices to obtain a scalar value, which is used to replace $\tilde{\mathcal{J}}$ for the decision process in (3.247).

It may be also desirable to use a frequency-dependent adaptive stepsize $\mu^{(\nu)}(m)$. This stepsize can be calculated for every frequency bin $\nu = 0, \ldots, R-1$ according to the algorithm given in (3.247) where $\tilde{\mathcal{J}}$ has to be replaced by the narrowband DFT-domain cost function $\tilde{\mathcal{J}}^{(\nu)}$ or by some other suitable quantity such as the magnitude squared coherence (MSC) between the output channels. For applying the bin-dependent stepsize, the update $\Delta\check{\mathbf{W}}(m)$ is transformed by an DFT, multiplied in each frequency bin by $\mu^{(\nu)}(m)$ and transformed back into the time-domain using an IDFT.

Furthermore, more sophisticated schemes which apply individual adaptive stepsizes to different filters are possible. This can be useful if, e.g., only one speaker is moving and thus, only a few demixing filters $\check{\mathbf{W}}_{pq}(m)$ are highly affected.

3.6 Experimental results

After introducing a generic BSS framework leading to various algorithms in time domain and DFT domain in the previous sections, we will now experimentally evaluate the different approaches. Firstly, in Section 3.6.2 the effect of different implementations of the Sylvester constraint on the separation performance will be investigated. Then the approximation of the more accurate covariance method used for estimating the correlation matrices by the efficient correlation method is examined in Section 3.6.3. Subsequently, in Section 3.6.4 the block-online update procedure is analyzed and the effect of the offline iterations and the adaptive stepsize is discussed. The investigations performed in these three sections are the basis for the comparison of the various efficient realizations of the generic BSS algorithm. Both, higher-order statistics and second-order statistics algorithms will be discussed in Section 3.6.5. Then, in the last section the performance of one selected high-performance algorithm will be evaluated in several realistic environments with different reverberation times and for various source-sensor distances.

3.6.1 Experimental setup

For the experiments several acoustic environments are used which are described in detail in Appendix C. They were chosen to reflect different realistic application scenarios with reverberation times ranging from $T_{60} = 50\,\text{ms}$ as, e.g., in a car environment, up to $T_{60} = 850\,\text{ms}$ for large lecture or conference rooms. In all environments two omnidirectional microphones with a spacing of $d = 20\,\text{cm}$ are used. Different source positions (which are described in the following sections and also in Appendix C) are considered and to allow the calculation of the SIR, we have measured the acoustic impulse responses from the source positions to the microphones by using the method described in [Sch79, RV89]. The height of the loudspeakers used to measure the impulse responses and the height of the microphones was approximately 140 cm. The impulse responses were then convolved with two dry source signals with a length of 10 sec recorded from a male and female speaker, respectively. The sampling rate was chosen to 16 kHz.

3.6.2 Sylvester constraint $\mathcal{SC}$ and its efficient implementations

In the derivation of the generic gradient BSS update equation in Section 3.3.3 it was shown in (3.52) that a Sylvester constraint operator $\mathcal{SC}$ is necessary. It was illustrated in Fig. 3.2 that the operator $\mathcal{SC}$ performs an averaging of the values on the diagonals of the demixing matrix update in each channel. Additionally, to obtain more efficient algorithms, it was proposed in Section 3.3.6 that instead of the averaging operation only the first column or the L-th row in each channel of the demixing matrix update may be picked. This was denoted as column Sylvester constraint $\mathcal{SC}_{\mathcal{C}}$ and row Sylvester constraint $\mathcal{SC}_{\mathcal{R}}$, respectively.

Here, we compare the performance of the different Sylvester constraints by using the generic SOS natural gradient algorithm introduced in (3.112). The adaptation will be performed in an offline manner according to (3.237). The correlation matrices were estimated by using the more accurate covariance method. The generic SOS algorithm still has a high complexity due to the large matrix computations involved in the update equation. However, in contrast to the efficient broadband and narrowband realizations in the DFT domain, no approximations are made. Thus, to obtain a reasonable computational complexity we perform these experiments in a low reverberant room with $T_{60} = 50\,\text{ms}$ so that we can choose a moderate demixing filter length $L = 256$. More reverberant environments will be considered when evaluating the efficient algorithms. The number of sources and microphones was chosen to $P = 2$ and the source-sensor distance was 1 m for all positions. Two different setups were examined: (a) sources positioned at $\pm 70^\circ$ and (b) sources positioned at $+45^\circ$, $+90^\circ$ (see also Appendix C.1 for layout of the room and positions of sources and sensors). As pointed out in Section 3.3.6 this allows for different initializations. In the case of $P = 2$ and two-sided setups such as (a) only causal delays are

needed in the demixing filters and thus, for the case (a) the demixing filters are initialized as $w_{pp,1} = 1$ for $p = 1, 2$. For one-sided setups such as in (b) or for $P > 2$ in general also acausal delays are necessary. This can be achieved by initializing the demixing filter in (b) with a shifted unit impulse. The minimum shift of the unit impulse is determined by by the maximum relative time delay between the sensors and is given as $f_s d/c$. In our case we chose the initialization for scenario (b) as $w_{pp,15} = 1$ for $p = 1, 2$. The parameter D which determines the number of time-lags considered in the correlation matrices is chosen to $D = L = 256$ and the block length determining the number of samples used for the estimation of the correlation matrices is $N = 512$. The auto-correlation matrices are regularized before inversion by multiplying the off-diagonal values with a factor $\rho = 0.75$ according to (3.183). Additionally, to prevent a division by zero, i.e., to cope with speech pauses, a constant value $\delta_{y_q} = 10^{-4}$ is added to the main diagonal of each auto-correlation matrix.

In Fig. 3.21 the segmental signal-to-interference ratio (SIR) improvement $\Delta\overline{SIR}_{\text{seg}}$ for two different scenarios and the source signals described in Section 3.6.1 is shown. Each curve represents the average of the channel-wise SIR improvements $\Delta\overline{SIR}_{\text{seg},q}$ according to (2.48). In the results for the two-sided setup depicted in Fig. 3.21(a) it can be seen that the original Sylvester constraint operator $\mathcal{SC}$ achieves the highest separation performance. Approximating $\mathcal{SC}$ by the row Sylvester constraint $\mathcal{SC}_{\mathcal{R}}$ or the column Sylvester constraint $\mathcal{SC}_{\mathcal{C}}$ does not lead to a significant degradation of the separation performance. In the one-sided scenario (b) the original Sylvester constraint operator $\mathcal{SC}$ exhibits again the highest separation performance. The approximation by $\mathcal{SC}_{\mathcal{R}}$ still achieves excellent separation. The application of the column Sylvester constraint $\mathcal{SC}_{\mathcal{C}}$ is not recommendable for one-sided setups or for the case $P > 2$ because no acausal delays can be adapted as discussed in Section 3.3.6. Hence, the use of $\mathcal{SC}_{\mathcal{C}}$ in setup (b) leads to the suppression of one source signal, i.e., one output of the BSS system contains a separated source with $\Delta\overline{SIR}_{\text{seg}} = 18\,\text{dB}$. The other output, however, still contains the mixture of both sources (i.e, $\Delta\overline{SIR}_{\text{seg}} = 0\,\text{dB}$) and thus, the average of both channels leads to the curve plotted in Fig. 3.21(b). It should be noted, that due to the permutation ambiguity it can in general not be predicted which source signal will be separated.

The results of these experiments show that the row Sylvester constraint $\mathcal{SC}_{\mathcal{R}}$ is a suitable approximation of the more complex Sylvester constraint $\mathcal{SC}$. It was also confirmed that the column Sylvester constraint can only be applied to the special case of a two-sided scenario with $P = 2$. To ensure the generality of the algorithms we will thus use the row Sylvester constraint $\mathcal{SC}_{\mathcal{R}}$ for the remaining experimental evaluation.

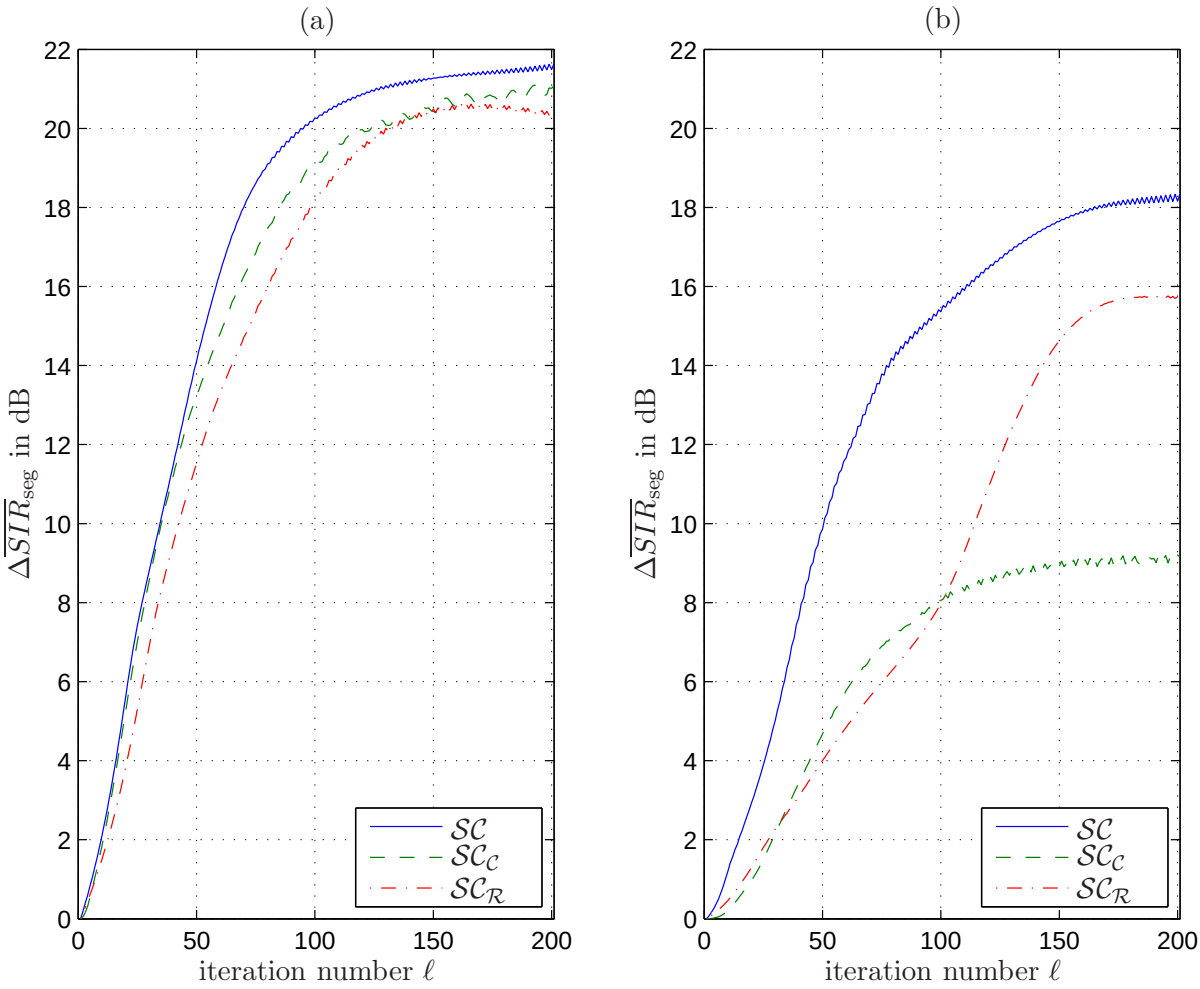


Figure 3.21: Segmental SIR improvement $\Delta\overline{SIR}_{\text{seg}}$ for the offline generic SOS algorithm for different Sylvester constraints. The positions of the $P = 2$ sources are given in the two scenarios as (a) $\pm 70°$ and (b) $+45°$, $+90°$.

3.6.3 Block-based estimation using covariance or correlation method

The correlation matrices appearing in the previously derived BSS algorithms can be estimated based on the covariance or correlation method as discussed in Section 3.3.5. The differences between these block-based estimation methods will be evaluated using again the generic SOS natural gradient algorithm (3.112) and the offline update procedure (3.237). The Sylvester constraint $\mathcal{SC}_{\mathcal{R}}$ is used for determining the demixing filter weights as it is applicable to all possible scenarios and according to the experimental results of the previous section it yields almost the same separation performance as the Sylvester constraint $\mathcal{SC}$ with less computational complexity. We use the same scenarios as in the previous section and the initialization is again given as $w_{pp,1} = 1$ for scenario (a) and

$w_{pp,15} = 1$, $p = 1, 2$ for scenario (b). The parameters of the algorithm are set to the same values as in the previous section and both setups (a) and (b) are examined.

The results in Fig. 3.22 show that for both setups the BSS algorithms based on covariance and correlation method converge to the same segmental SIR improvement. In

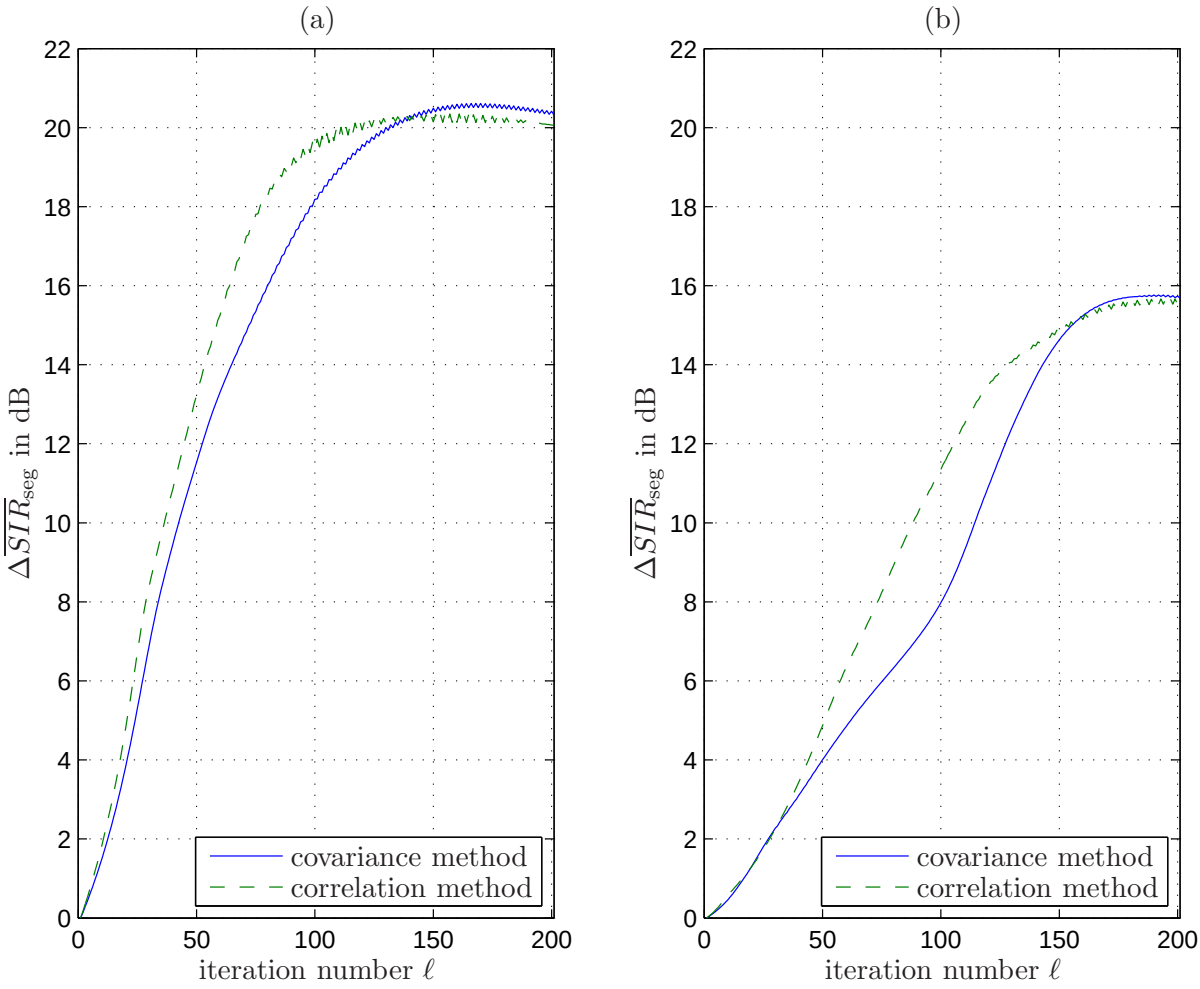


Figure 3.22: Segmental SIR improvement $\Delta\overline{SIR}_{\text{seg}}$ for the offline generic SOS algorithm using a block-based estimation of correlation matrices based on correlation and covariance method. The positions of the sources are given in the two scenarios as (a) $\pm 70°$ and (b) $+45°$, $+90°$.

general, the covariance method has the potential of giving better performance as no stationarity is assumed within the signal block of length $N = 512$. However, the estimation by the covariance method usually requires a larger regularization of the correlation matrix as the values on the main diagonal are not constant and "holes" on the main diagonal, i.e., small values lead to an ill-conditioned matrix. Thus, the regularization with a constant value added to the main diagonal of the auto-correlation matrices is chosen to $\delta_{y_q} = 10^{-4}$ for the covariance method and $\delta_{y_q} = 10^{-5}$ for the correlation method. The smaller regularization in case of the correlation method leads to a faster convergence as can be observed

in Fig. 3.22.

From the results in Fig. 3.22 we can conclude that for speech signals the correlation method is an appropriate approach for the estimation of the correlation matrices. This has the benefit that computational complexity is reduced due to the Toeplitz structure of the correlation matrices. Similar results have been observed in linear prediction where the correlation method is the preferred approach due to lower computational complexity and increased stability (see, e.g., [MG76, BNR96]).

3.6.4 Block-online adaptation and adaptive stepsize

In the previous sections the offline update has been used for the adaptation of the BSS algorithm as it provides an upper bound for the segmental SIR improvement. For real-time applications a block-online update procedure has been proposed in Section 3.5.3 which improves convergence by performing an offline update with $\ell_{\max}$ iterations for K subsequent blocks. To evaluate the effect of the iteration number $\ell_{\max}$ on the SIR improvement we use again the generic SOS natural gradient algorithm (3.112) together with the estimation of the correlation matrices by the correlation method. Additionally we use in the following, due to its increased versatility, the Sylvester constraint $\mathcal{SC}_{\mathcal{R}}$ together with the initialization $w_{pp,15} = 1$, $p = 1, 2$. Analogously to the previous experiments, the demixing filter length is chosen to $L = 256$, the block length $N = 512$, and the regularization is given as $\rho = 0.75$ and $\delta_{y_q} = 10^{-5}$. To exploit the nonstationarity of the signals we simultaneously use $K = 8$ blocks of length N for the offline iterations given in (3.245) and the forgetting factor is set to $\lambda = 0.2$. To illustrate the convergence behavior of the BSS algorithm over time, we use the segmental SIR improvement $\Delta SIR_{\text{seg}}(m)$ which depends on the block index m and is defined in (2.50). Fig. 3.23 shows the resulting segmental SIR improvements for the low reverberant room with $T_{60} = 50\,\text{ms}$ and the sources positioned at $\pm 70^\circ$ when varying the number of offline iterations $\ell_{\max} = 1, 5, 10$. It can be seen that the increase of $\ell_{\max}$ leads to a faster initial convergence. Especially for time-variant mixing systems a fast convergence time is crucial to be able to adapt to changes in the environment. However, for the choice $\ell_{\max} = 5, 10$ it can be seen that the stepsize $\mu = 0.01$ becomes too large when adapting towards the correct demixing filters and thus, the segmental SIR improvement is limited ($\ell_{\max} = 5$) or even decreases again ($\ell_{\max} = 10$). This is the motivation for introducing an adaptive stepsize according to Section 3.5.4 which is able to choose μ from a given interval $[\mu_{\min}, \mu_{\max}]$.

In (3.247) an update rule for an adaptive stepsize was given based on comparing the value of the cost function for subsequent demixing filter coefficients. The values of the cost function can be interpreted as an indicator for convergence or divergence of the filter weights. As the cross-correlations, which are already calculated during the coefficient update, also allow to assess the separation performance, they can be used

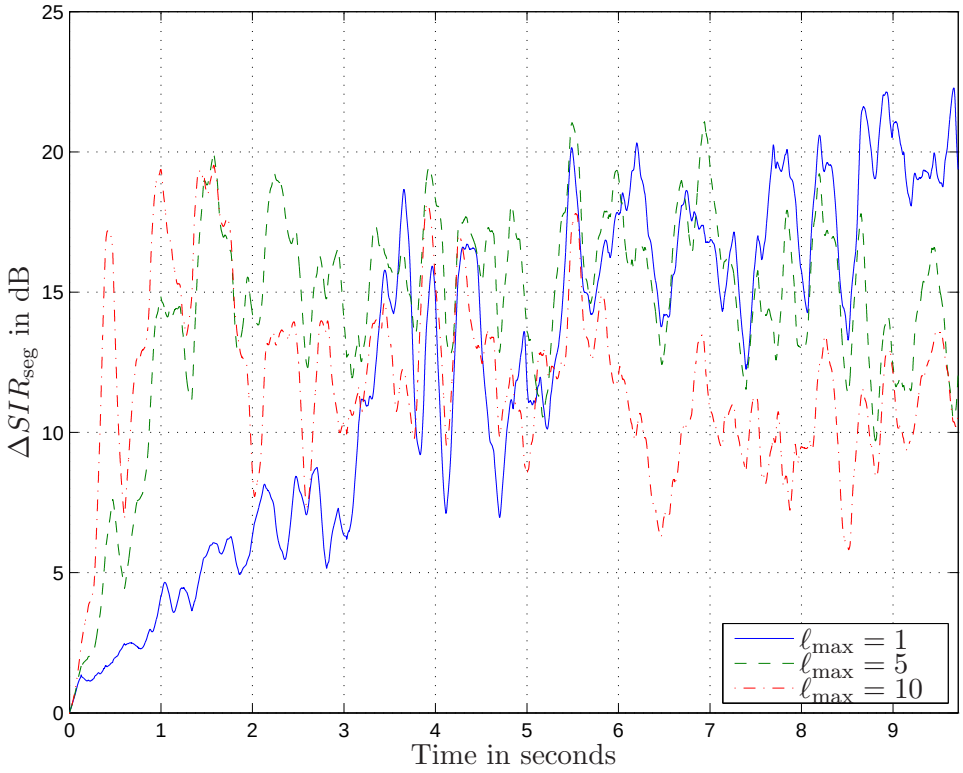


Figure 3.23: Segmental SIR improvement ΔSIR_{seg} for the block-online generic SOS algorithm for different number of offline iterations $\ell_{\max}$.

instead of the cost function in (3.247). Therefore, the Frobenius norm of the cross-correlations averaged over K blocks and over all possible output combinations given as $\frac{1}{K(P^2-P)} \sum_{i=1}^{K} \sum_{p,q} \|\mathbf{R}_{\mathbf{y}_p\mathbf{y}_q}\|_{\text{F}}$, $p \neq q$ is used as an indicator for convergence or divergence determining the value of the adaptive stepsize. The parameters in (3.247) are chosen as $a = 1.1$, $b = 0.5$, $c = 1.3$, and the range of the stepsize is given as $[0.0001, 0.01]$. Fig. 3.24 exemplarily shows the advantage of the adaptive stepsize for $\ell_{\max} = 5$. The algorithm behaves the same during the initial convergence phase. In the adaptive case the stepsize μ is reduced after the initial convergence which allows a finer adaptation of the filter weights and thus leads to a higher segmental SIR improvement compared to a fixed stepsize. Additionally, in real-time implementations the fixed stepsize would have to be set very conservatively to avoid divergence of the demixing filter weights. In the case of an adaptive stepsize this is avoided by an automatic reduction of the stepsize so that the upper limit $\mu_{\max}$ may be chosen larger than a conservative fixed stepsize.

We can conclude that it is recommendable to introduce a moderate number of offline iterations (e.g, $\ell_{\max} = 5$) to increase the convergence of the BSS algorithm. Moreover, an adaptive stepsize should be incorporated to increase the robustness and to allow for higher stepsizes compared to a fixed stepsize.

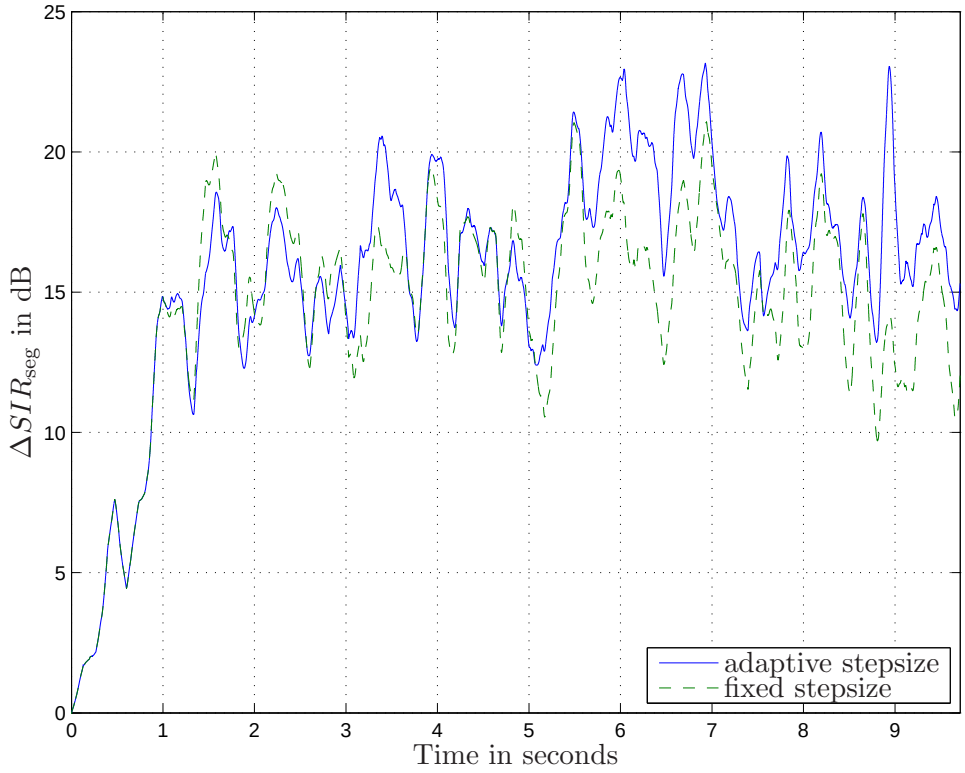


Figure 3.24: Segmental SIR improvement ΔSIR_{seg} for the block-online generic SOS algorithm with $\ell_{\max} = 5$ iterations using fixed and adaptive stepsizes.

3.6.5 Comparison of different HOS and SOS realizations

In this section several different HOS and SOS realizations resulting from the generic HOS natural gradient algorithm (3.64) will be compared. Based on the results of the previous sections we will use for all examined algorithms the Sylvester constraint $\mathcal{SC}_{\mathcal{R}}$ together with the initialization $w_{pp,15} = 1$, $p = 1, 2$ and use the correlation method for the estimation of the correlation matrices. Additionally, the block-online update with $\ell_{\max} = 5$ together with the adaptive stepsize procedure is used to allow online processing of the sensor signals while maintaining fast convergence. The parameters of the block-online update are chosen as in the previous section and the stepsize has been maximized for each algorithm up to the stability margin. The separation performance of the various BSS algorithms is shown for a living room scenario with a reverberation time $T_{60} = 200\,\text{ms}$. Two different setups are examined with the sources placed at a distance of 1 m at either $-20°$, $40°$ or $20°$, $80°$ (see also Appendix C.2 for layout of the room and positions of sources and sensors). To address the reverberation, the demixing filter length has been chosen to $L = 1024$ taps. The nonwhiteness is exploited by the memory of $D = L = 1024$ introduced in the multivariate pdfs and in the correlation matrices. For accurate estimation of these quantities a block length of $N = 2048$ was chosen.

The algorithms which are evaluated include the computationally complex higher-order and second-order statistics generic algorithms as well as the efficient algorithms obtained by applying several approximations to the generic algorithms. A list of all algorithms is given in Table 3.1. It should be noted that all algorithms were implemented efficiently in the DFT-domain either in a broadband or narrowband manner.

Second order statistics (SOS)	
(A)	Generic broadband SOS algorithm (3.174) based on the multivariate Gaussian pdf.
(B)	Broadband SOS algorithm based on multivariate Gaussian pdf (3.222) with normalization approximated as a narrowband inverse.
(C)	Broadband SOS algorithm based on multivariate Gaussian pdf (3.174) with normalization approximated as a scaling by the output signal variance (3.116).
(D)	Narrowband SOS algorithm based on multivariate Gaussian pdf (3.223a), (3.223b) where the coupling between the DFT bins is ensured by one remaining constraint matrix.

Higher order statistics (HOS)	
(E)	Broadband HOS algorithm (3.168) based on the model of a Laplacian SIRP pdf (3.100).
(F)	Narrowband HOS algorithm (3.204) based on a multivariate SIRP pdf. The coupling between the DFT bins is ensured by one remaining constraint matrix and by the argument of the multivariate score function (3.206).
(G)	Narrowband HOS algorithm (3.214) based on univariate pdfs with the score function chosen according to (3.219). The coupling between the DFT bins is ensured by one remaining constraint matrix.

Table 3.1: List of algorithms evaluated in the reverberant living room scenario.

The results for the SOS algorithms (A)-(D) can be found in Fig. 3.25. It can be seen that the generic broadband SOS algorithm (A) provides the best performance in both setups. However, its high computational complexity prevents a real-time implementation on current state-of-the-art hardware platforms for such large demixing filter lengths. Therefore approximations are needed which minimally affect the separation performance but result in computationally efficient algorithms. The main complexity in the second-order statistics algorithms is caused by the inverse of the auto-correlation matrix for each output channel. This inverse is approximated in the broadband algorithm (B) by a narrowband inverse which leads to a scalar inversion in each DFT bin. The algorithm (B) can be implemented in real-time on regular PC hardware and it can be seen in Fig. 3.25 that for both setups the separation performance is only slightly reduced. In the broadband algorithm (C) the normalization is further simplified by using the variance of each

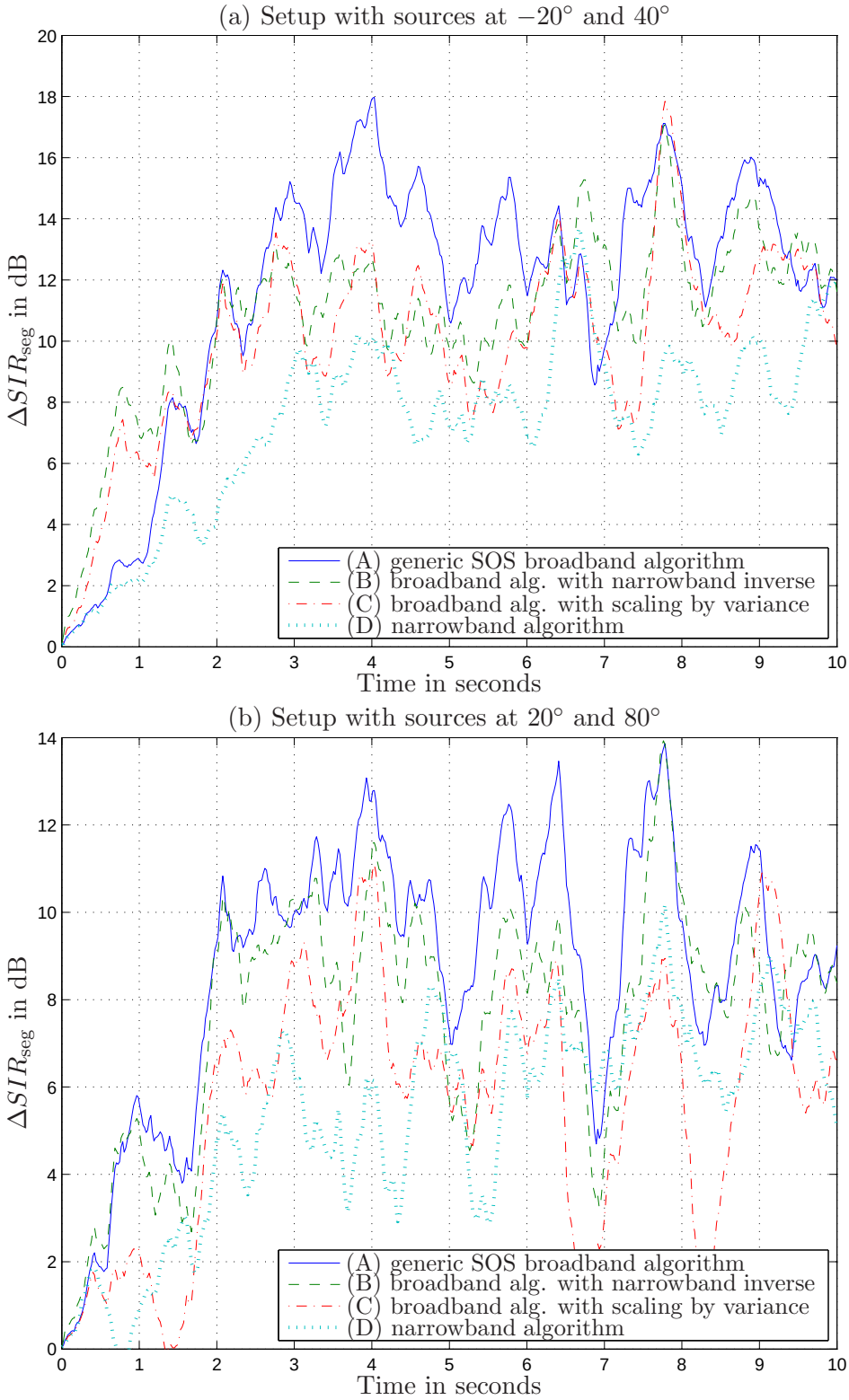


Figure 3.25: Segmental SIR improvement $\Delta\overline{SIR}_{\text{seg}}$ for the second-order statistics algorithms (A)-(D) evaluated in the living room scenario ($T_{60} = 200\,\text{ms}$) for two source position setups.

output signal. This means that the normalization is not frequency-dependent anymore. The experiments show that this is not so critical in scenarios where only causal filters have to be adapted. However, for the setup with the sources placed at 20° and 80° a clear reduction of the separation performance can be observed. In the narrowband algorithm (D) all constraint matrices except one are approximated. This means that the narrowband normalization is done analogously to algorithm (B), however, due to discarding all constraint matrices except for one, the complete decoupling of the DFT bins is only prevented by the last remaining constraint matrix. Thus, the algorithm (D) already suffers from the permutation and scaling problem occurring in each DFT bin. This explains the inferior separation performance compared to the broadband algorithms (A)-(C).

In Fig. 3.26 the results for the HOS algorithms (E)-(G) are depicted. The generic HOS broadband algorithm (E) exhibits the highest performance in both setups. Similar to the generic SOS broadband algorithm, the high computational complexity of (E) requires that certain approximations are made to obtain efficient algorithms. In Section 3.4.3.2 two possibilities have been shown to obtain more efficient algorithms. Approximating several constraint matrices leads to a narrowband algorithm (F) where the coupling between the DFT bins is still given by the argument of the score function which is based on a multivariate SIRP pdf and additionally by one remaining constraint matrix. A different example how to approximate the generic HOS algorithm is the algorithm (G) which is based on score functions resulting from univariate pdfs. Then, the coupling between the DFT bins is solely ensured by the last remaining constraint matrix. It can be seen that both algorithms (F) and (G) exhibit similar performance and show also a reduced separation performance compared to the broadband generic HOS algorithm. One reason for this behavior is that in the generic broadband algorithm the argument of the nonlinear SIRP score function is given by the quadratic form u_q defined in (3.99) which contains the inverse of the auto-correlation matrix. This ensures that in the mean the argument of the SIRP score function is well-scaled and thus, the generic HOS algorithm exhibits good performance. However, due to the approximations of the quadratic form leading to the algorithms (F) and (G), the scaling is not as accurate anymore and thus, the separation performance decreases.

When comparing the results for the SOS and HOS algorithms in Figs. 3.25 and 3.26 it can be seen that the usage of higher-order statistics, i.e., the exploitation of the nongaussianity of the sources yields higher separation performance for the narrowband algorithms (F), (G) compared to (D). For the generic broadband algorithms, however, for the given data the performance does not increase when using HOS. This can be attributed to the SIRP score $\phi_{y_q,D}(u_q)$ which was obtained from the multivariate Laplacian SIRP pdf in (3.103) and which exhibits a pole for $u_q \to 0$. This required a careful regularization of algorithm (E) and thus, much higher separation performance compared to the SOS generic algorithm was difficult to obtain. Instead of introducing a regularization it was

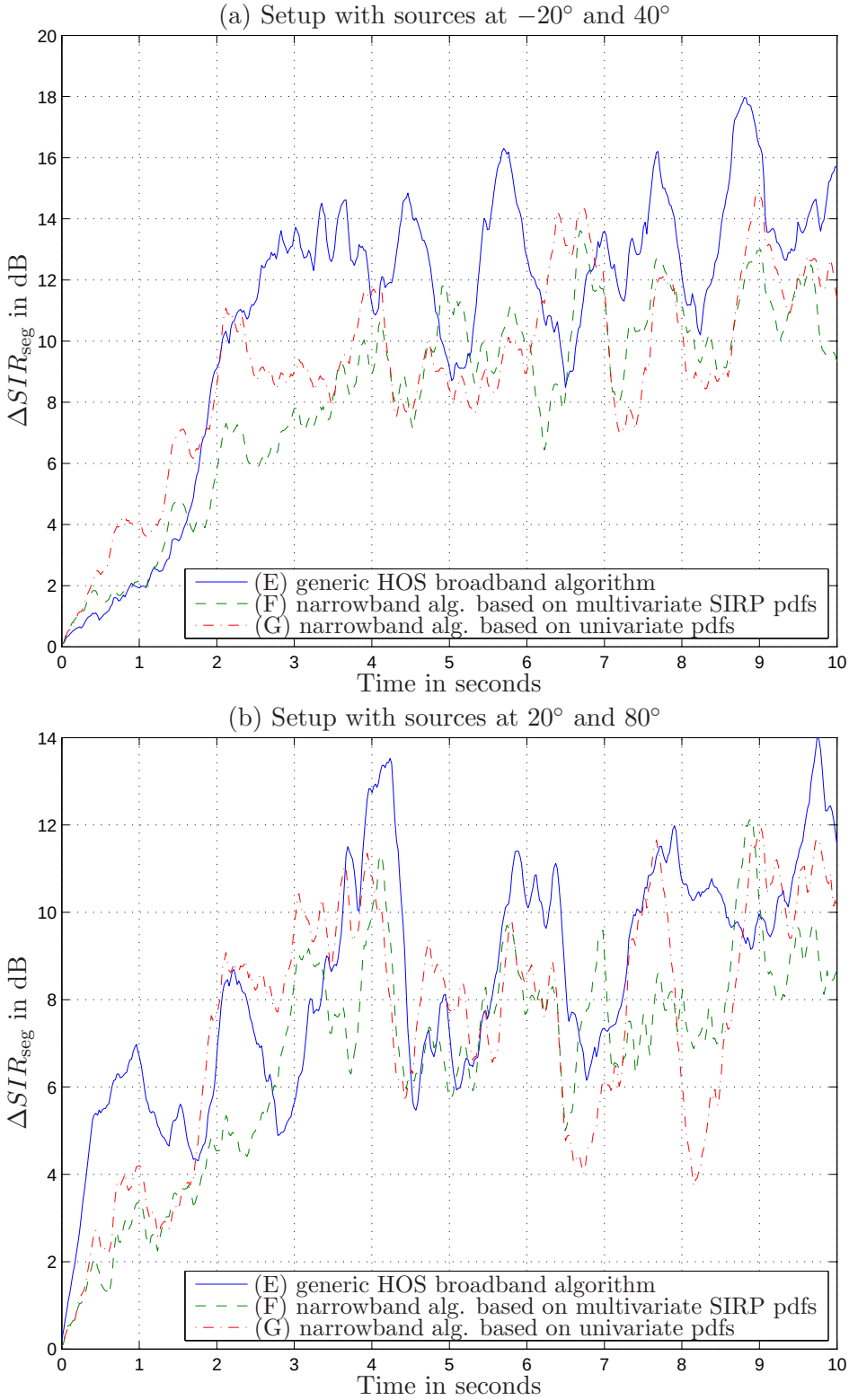


Figure 3.26: Segmental SIR improvement $\Delta\overline{SIR}_{\text{seg}}$ for the higher-order statistics algorithms (E)-(G) evaluated in the living room scenario ($T_{60} = 200\,\text{ms}$) for two source position setups.

shown in [Buc] that another possibility is to remove the nonlinearity, i.e., $\phi_{y_q,D}(u_q) = 1$ for small arguments $u_q \to 0$ so that for small excitations the multivariate Gaussian pdf is used instead. This procedure, was derived by using the theory of robust statistics [Hub81] and avoids poles in the nonlinearity. In that case the generic higher-order statistics algorithm (E) outperforms the second-order counterpart (A). This concept was also applied successfully for the case of instantaneous BSS in [CD06]. In [Buc] the theory of robust statistics is extended to multivariate pdfs where even further separation improvements are expected. Moreover, similar to the second-order case it would also be desirable to introduce less approximations than in (F) or (G) to obtain efficient and high-performing versions of the generic higher-order statistics algorithm (E).

3.6.6 Influence of reverberation time and source-sensor distance

In Chapter 2 it was pointed out that reverberant environments can be characterized by different quantities. The reverberation time T_{60}, which is determined by measuring the sound decay, is a global characterization of the room but does not fully describe the listening conditions. Especially the early reflections vary from one point to another so that the positions of the sources and sensors and in particular the source-sensor distance influence the perceived quality of the sound signal. One quantity which can measure the effects of different source positions is the signal-to-reverberation ratio (SRR) which was defined in (2.20) as the ratio of useful reflections arriving earlier than 50 ms to the later reflections which are perceived as reverberation.

To examine the applicability of the BSS algorithms in different reverberant environments, we will evaluate in this section the influence of the reverberation time and the source-sensor distance on the separation performance. For this purpose we will use the algorithm (B) from Table 3.1 which is given as the broadband SOS algorithm using a narrowband normalization as derived in (3.222). In the comparison in the previous section this algorithm showed the best results among all efficient algorithms. The filter length and the memory length are chosen as $L = D = 1024$ and the block length is $N = 2048$. For the narrowband normalization, the DFT length is $R = 4096$ and the regularization prior to inversion is given by $\rho = 0.75$, $\delta_{y_q} = 10^{-5}$. We again use the block-online update with $\ell_{\max} = 5$ offline iterations and the adaptive stepsize lies within the interval $[0.0001, 0.05]$.

First, the BSS algorithm is applied in three different environments: the living room scenario with closed and open curtains and the lecture room (see Appendices C.2 and C.3). The reverberation times for these rooms are given as $T_{60} = 200\,\text{ms}$, 400 ms, and 850 ms and the source-sensor distance is 1 m. For each environment the signal-to-reverberation ratio SRR_{q,s_q} was calculated according to (2.20) for each source s_q, $q = 1, 2$ and then averaged over both sources. This leads to the signal-to-reverberation ratios of 17.7 dB, 12.6 dB, and 7.4 dB for the environments with the reverberation times $T_{60} = 200\,\text{ms}$, 400 ms,

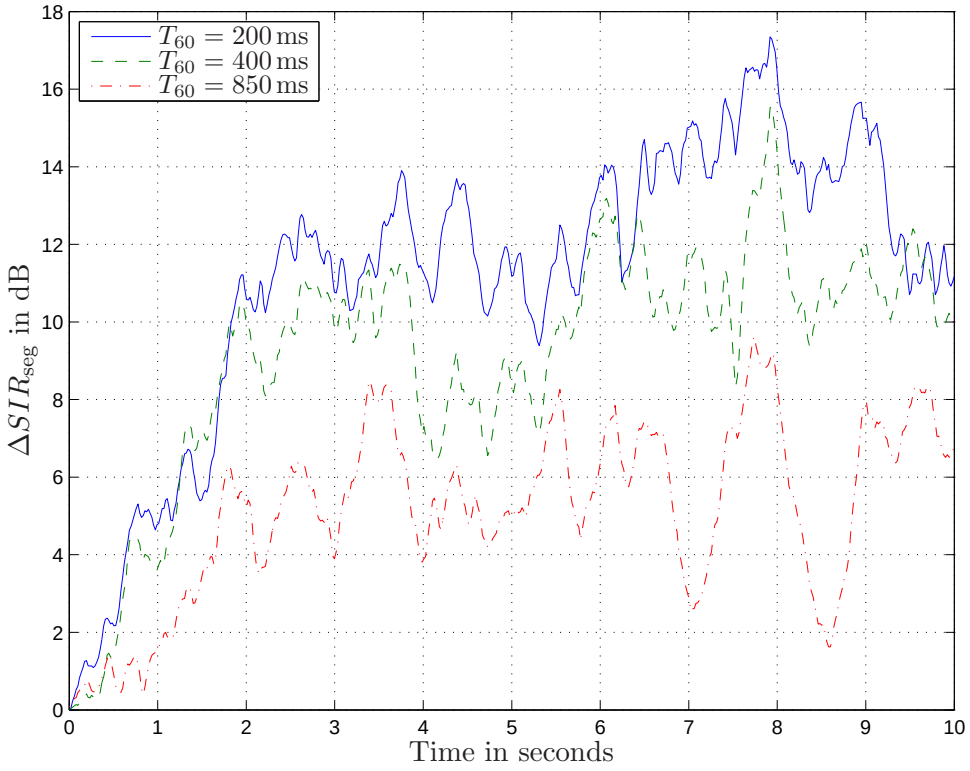


Figure 3.27: Segmental SIR improvement $\Delta\overline{SIR}_{\text{seg}}$ for different reverberation times T_{60}.

and 850 ms, respectively. Subsequently, a comparison of the separation performance for different source-sensor distances of 1 m, 2 m, and 4 m is carried out in the living room scenario with closed curtains. In terms of the signal-to-reverberation ratio, the different distances lead to of 17.7 dB, 13.1 dB, and 12.2 dB, respectively. The direction of arrival (DOA) of the sources is chosen for all experiments in this section to $\pm 20°$.

In Fig. 3.27 the separation results for different reverberation times are depicted. It can be seen that a high separation can be achieved in the living room scenario for both cases, curtains opened or closed. However, for highly reverberant environments like the lecture room the segmental SIR decreases because the demixing filter length L is too short to cover all reflections. Although an increase of L would allow to address the large reverberation, this would mean that also the block length N has to be increased as in general $N > L$ samples are required to estimate the correlation matrices. On the other hand, speech is a nonstationary process and thus, the block length should be chosen as small as possible. These two contradictory requirements may be addressed by a partitioned adaptive filtering scheme as previously proposed in supervised adaptive filtering [SP90], however, this is outside the scope of this work.

The influence of the source-sensor distance is shown in Fig. 3.28. It can be seen that the further the sources are away from the sensors, i.e., the lower the signal-to-reverberation

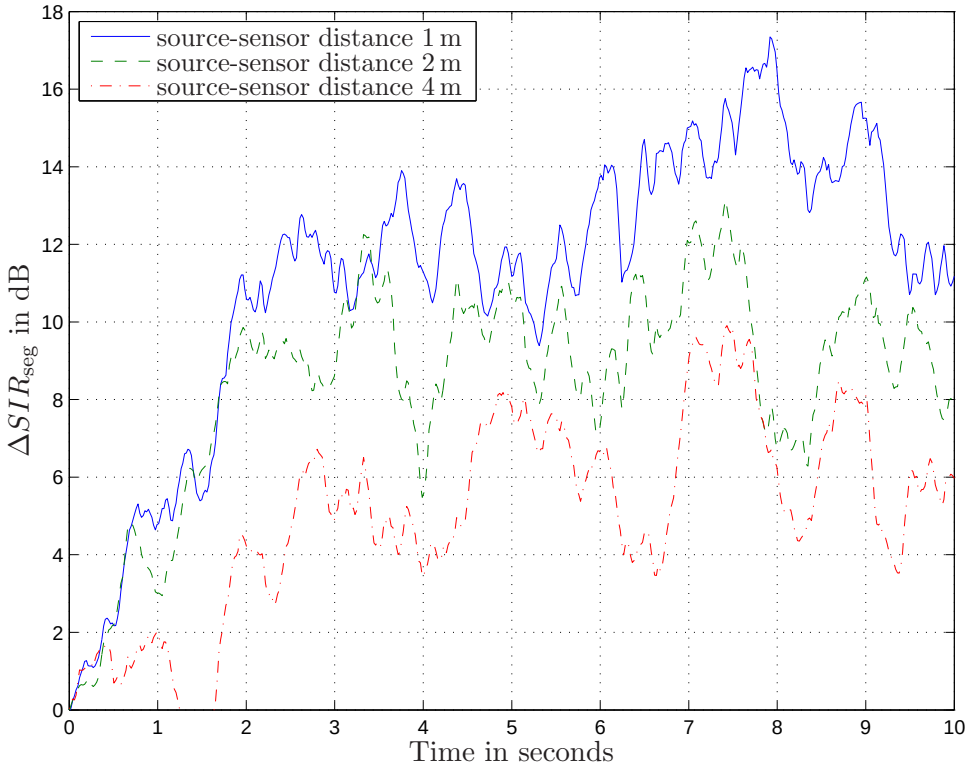


Figure 3.28: Segmental SIR improvement $\Delta\overline{SIR}_{\text{seg}}$ for different source-sensor distances.

ratio is, the harder it is to achieve high separation performance. This behavior was also observed in Chapter 2 for the estimation of the magnitude-squared coherence (MSC) for point sources in reverberant environments. The estimation of the MSC exhibits a bias which depends on the ratio of the observation length to the acoustic impulse response length. Moreover, it was shown that the bias is increased for larger reverberation times T_{60} and larger source-sensor distances. Ideally, for the MSC estimation the observation length would have to be much larger than the acoustic impulse response length. This analogous behavior of the demixing filter estimation and the MSC estimation is not surprising as it was shown in Section 3.4.3.4 that the narrowband SOS BSS algorithm can be derived from the generalized coherence which is in the case $P = 2$ equal to the MSC. Due to this analogy, an increasing observation length, i.e. block length N, should also yield better results for large source-sensor distances. However, as pointed out already above this requirement can usually not be fulfilled as the nonstationary nature of acoustic signals such as speech prevents the choice of large block lengths. Thus, here again a BSS scheme based on partitioned adaptive filtering may yield improved results and would therefore be an interesting topic of future research.

3.7 Summary

In this chapter we presented a framework for BSS in reverberant environments which is termed TRINICON ("TRIple-N Independent component analysis for CONvolutive mixtures") and provides a unified view on convolutive BSS algorithms. This framework allows to derive novel algorithms and to establish relationships to existing state-of-the-art algorithms.

First, in Section 3.1 the optimum BSS solution was discussed and the resulting optimum demixing filter length was derived. It could be seen that BSS aims at multi-channel blind system identification and thus, can be interpreted as a blind interference cancellation technique.

Then, in Section 3.2 the broadband and narrowband signal models were introduced which are the basis of the different optimization schemes used to derive BSS algorithms. Broadband optimization is obtained if the derivation is performed in the time domain or if the optimization is carried out for all DFT bins simultaneously. This preserves the coupling between the DFT bins and has the advantage that the permutation and scaling ambiguity do not appear in each DFT bin independently. On the other hand, narrowband optimization is obtained if the optimization is done in each DFT bin independently. Narrowband BSS algorithms are usually computationally more efficient, however, they need additional measures to provide a consistent assignment of the sources to the output channels and a consistent scaling for all DFT bins. Based on this categorization an overview about the BSS literature was given in Section 3.2.

The TRINICON BSS framework was developed in Section 3.3 using a time-domain optimization criterion based on the generalization of the mutual information for temporally dependent signals. The term "TRIple-N" accounts for the fact that all three signal properties nongaussianity, nonstationarity, and nonwhiteness can be exploited. The minimization of the optimization criterion was achieved by using gradient descent and natural gradient descent iterative update procedures. The derivation of the gradient and natural gradient led to generic BSS algorithms which are characterized by high-dimensional multivariate pdfs. In general, the estimation of these multivariate pdfs is a very challenging task as all corresponding higher-order cumulants have to be determined. Fortunately, signals such as speech can be modeled as spherically invariant random processes (SIRPs) which lead to multivariate pdfs solely determined by the second-order correlation matrix and the univariate marginal pdf. This approximation resulted in a higher-order BSS algorithm realization and a generic second-order BSS algorithm based on the multivariate Gaussian pdf as a special case of a SIRP pdf. Several further approximations were discussed leading to novel algorithms and to relationships to state-of-the-art BSS algorithms from the literature.

The higher-order and second-order BSS algorithm realizations in the time domain

were the starting point in Section 3.4 to extend the framework to the DFT domain and investigate broadband and narrowband DFT-domain algorithms. First, the broadband and narrowband signal models introduced in Section 3.2 were expressed by using matrix notation. This yielded in the broadband case the well-known overlap-save algorithm which implements a linear convolution in the DFT domain. In the narrowband case the constraint originating from the overlap-save structure is approximated and thus a circular convolution is obtained. The broadband signal model in matrix notation was then applied to express the time-domain update equations equivalently in the DFT domain. This resulted in broadband DFT-domain algorithms which exhibit several constraint matrices consisting of IDFT, windowing, and DFT operations. Selective approximations of these constraint matrices were introduced to obtain computationally more efficient novel algorithms, but to still preserve the coupling between the DFT bins so that the permutation and scaling ambiguities do not appear independently in each DFT bin. This allowed, e.g., to avoid computationally complex matrix inverses and led to more efficient methods to calculate the HOS score functions originating from the multivariate SIRP pdfs. Moreover, relationships to several well-known algorithms from the BSS literature could be established. It was also shown that algorithms based on narrowband optimization can be derived from the TRINICON framework if all constraint matrices are removed.

In Section 3.5 the general weighting function introduced in the TRINICON optimization criterion was discussed. The weighting function allows the implementation of several different iterative update procedures necessary for the iterative gradient or natural gradient descent. First, the well-known offline update, which iterates over the whole signal data, and the online update, which can be efficiently implemented in a recursive manner, were discussed. The online update has the advantage that it can track time-variant systems and is thus well-suited for real-time implementations. However, in rapidly time-variant systems the convergence of the online algorithms is sometimes not sufficient, so that a combination of online and offline update was introduced. This so-called block-online approach includes an online part which ensures that the method is applicable to real-time systems by continuously processing new blocks. Additionally, every time for the current block an iterative offline processing is performed. This improves convergence and only moderately increases computational complexity. Moreover, to achieve more robust adaptation, an adaptive stepsize technique has been proposed.

The algorithms and update procedures developed in Sections 3.3 - 3.5 were evaluated experimentally in reverberant environments in Section 3.6. First, the different methods to implement the Sylvester constraint and the block-based estimation of the correlation matrices were investigated using the offline generic second-order BSS algorithm. Based on these results the same algorithm was used to examine the block-online update procedure and the benefit of the adaptive stepsize. It was shown that already a moderate number of offline iterations leads to a drastic increase of the algorithm convergence and that the

application of an adaptive stepsize leads to higher separation performance. Subsequently, the various higher-order and second-order BSS algorithms derived in Sections 3.3 and 3.4 have been evaluated. It was shown that broadband algorithms, where only selective narrowband approximations were introduced, yielded better separation performance compared to their narrowband counterparts where the coupling between the DFT bins was ensured only by a last remaining constraint. In the end, one of the most promising BSS algorithms was additionally evaluated in different reverberant environments to examine the influence of the reverberation time and the source-sensor distance on the separation performance.

4 Extensions for Blind Source Separation in Noisy Environments

In Chapter 3 only noiseless reverberant environments were considered with the maximum number of simultaneously active point sources Q assumed to be equal to the number of sensors P. However, in realistic scenarios, in addition to the point sources to be separated, also some background noise will usually be present. It was pointed out in Section 2.2.3 that several types of background noise (e.g., car noise, babble noise, etc.) can be described by a diffuse sound field which can be modeled by an infinite number of statistically independent point sources distributed on a sphere. Due to the assumption that $Q = P$, the BSS algorithms treated in Chapter 3 cannot model an infinite or in practice at least very large number of noise point sources. Therefore, in the block diagram of the BSS model in Fig. 2.1 the noise signals were not modeled as point sources but as contributions $n_1, \ldots, n_P$ to the sensor signals $x_1, \ldots, x_P$. Hence, the BSS framework derived in Chapter 3 does not aim at the reduction of diffuse background noise, but the noise has rather to be regarded as a perturbation of the BSS signal model. Thus, in general BSS faces two different challenges in noisy environments:

1. The *adaptation* of the demixing BSS filters should be *robust against* the *noise signals* $n_1, \ldots, n_P$ to ensure high separation performance of the desired point sources $s_1, \ldots, s_P$. This means that the signal-to-interference ratio (SIR) should not deteriorate compared to the noiseless case.

2. The *noise* contribution contained in the separated BSS output signals *should be suppressed*, i.e., the signal-to-noise ratio (SNR) should be maximized.

Both requirements must be met if BSS should be attractive for noisy environments.

According to the literature (see e.g., [CA02, HKO01] and references therein), it has been tried to address the first point by developing noise-robust BSS algorithms. However, so far this has been considered only for the instantaneous BSS case. Additionally, several assumptions are usually imposed on the noise signals allowing to generate optimization criteria which are not affected by the noise signals. In [CC01] the assumption of spatially correlated but temporally uncorrelated noise allowed the formulation of a joint diagonalization criterium which only exploits correlations for time-lags unequal to zero. Other noise-robust instantaneous BSS algorithms for the case of temporally correlated

but spatially uncorrelated noise were presented in [Car04]. However, in the case of convolutive BSS for acoustic signals these assumptions for the noise signals are too restrictive. As pointed out above, in realistic scenarios *temporally correlated* background noise at the sensors can often be described by a diffuse sound field leading to signals which are also *spatially correlated for low frequencies* (see Section 2.2.4.3). This does not allow the application of the above-mentioned algorithms to the noisy convolutive BSS problem.

Another class of noise-robust instantaneous BSS algorithms aims at merely using higher-order statistics (HOS), e.g., by using fourth-order cumulants, to be immune against Gaussian noise signals [Car92, CA02, HKO01]. This could also be applied to background noise described by diffuse sound fields which can be modeled as a superposition of an infinite number of statistically independent sources. Due to the central limit theorem the distribution of the superpositioned sources, i.e., of the noise signals observed at the sensors, will approach the Gaussian distribution. However, by neglecting the second-order correlations and relying merely on HOS, these algorithms become very sensitive to outliers and thus, they are not useful in practice.

A more promising approach for increasing the robustness of the BSS adaptation are pre-processing methods. In Section 4.1 we will describe single-channel and multi-channel methods in order to remove the bias of the second-order correlation matrices caused by the noise. This will lead to a better performance of the BSS algorithms discussed in Chapter 3. Moreover, it will be pointed out how the information of additional sensors could be exploited by using subspace techniques to achieve noise suppression for the input signals of the BSS algorithm.

Another approach to the noisy convolutive BSS problem is the application of post-processing methods to the outputs of the BSS system. Without pre-processing the separation performance of the BSS algorithms will decrease in noisy environments. Therefore, the post-processing technique has to aim at the suppression of both, background noise and residual crosstalk from interfering point sources which could not be canceled by the BSS demixing filters. This will be discussed in detail in Section 4.2.

Finally, in Section 4.3 the results of the different techniques will be summarized.

4.1 Pre-processing for noise-robust adaptation

From the literature only few pre-processing approaches for BSS in noisy environments are known. If the number of sensors P is equal to the number of sources Q, as considered in this thesis, then usually so-called bias-removal techniques are used which aim at estimating and subtracting the contribution of the noise in the sensor signal itself or in the second-order correlation matrix and possibly also in the higher-order relation matrix of the sensor signals. These techniques will be discussed in Section 4.1.1. If more sensors than sources are available, i.e., $P > Q$, then also subspace techniques can be used as a pre-processing

step to achieve a suppression of the background noise. Even if we restricted ourselves in this thesis to the case $P = Q$ we will, for the sake of completeness, briefly summarize the history of subspace approaches, their application to BSS, and outline possible directions of future research in Section 4.1.2.

4.1.1 Bias-removal techniques

The signal model in matrix-vector notation (3.33) yields the BSS output signals $\mathbf{y}(n)$ containing D output signal samples for each of the $Q = P$ channels. If background noise $n_p(n)$ is superimposed at each sensor $p = 1, \ldots, P$, the signal model can be decomposed as

$$\begin{aligned} \mathbf{y}(n) &= \mathbf{W}^{\mathrm{T}}\mathbf{x}(n) \\ &= \mathbf{W}^{\mathrm{T}}\left(\mathbf{H}_{2L}^{\mathrm{T}}\mathbf{s}(n) + \mathbf{n}(n)\right) \end{aligned} \tag{4.1}$$

where the background noise samples are contained in the column vectors

$$\mathbf{n}(n) = [\mathbf{n}_1^{\mathrm{T}}(n), \ldots, \mathbf{n}_P^{\mathrm{T}}(n)]^{\mathrm{T}}, \tag{4.2}$$

$$\mathbf{n}_p(n) = [n_p(n), \ldots, n_p(n - 2L + 1)]^{\mathrm{T}}. \tag{4.3}$$

It can be seen from the noisy signal model (4.1) that the second-order correlation matrix $\mathbf{R}_{\mathbf{yy}}$ and also the higher-order relation matrix $\mathbf{R}_{\mathbf{y\Phi(y)}}$ will contain a bias due to the background noise. In this chapter we will use the second-order statistics BSS algorithm given in (3.222) as it yielded the best results of all efficient algorithms presented in the previous chapter. Thus, we will focus here only on second-order statistics. The background noise $\mathbf{n}$ and the point-source signals $\mathbf{s}$ are assumed to be mutually uncorrelated so that the second-order correlation matrix $\mathbf{R}_{\mathbf{yy}}$ defined in (3.71) can be decomposed as

$$\mathbf{R}_{\mathbf{yy}}(n) = \mathbf{W}^{\mathrm{T}}\left(\mathbf{H}_{2L}^{\mathrm{T}}\mathbf{R}_{\mathbf{ss}}(n)\mathbf{H}_{2L} + \mathbf{R}_{\mathbf{nn}}(n)\right)\mathbf{W} \tag{4.4}$$

with the source correlation matrix $\mathbf{R}_{\mathbf{ss}}$ and noise correlation matrix $\mathbf{R}_{\mathbf{nn}}$ defined as

$$\mathbf{R}_{\mathbf{ss}}(n) = \frac{1}{N}\sum_{j=0}^{N-1}\mathbf{s}(n+j)\mathbf{s}^{\mathrm{T}}(n+j), \tag{4.5}$$

$$\mathbf{R}_{\mathbf{nn}}(n) = \frac{1}{N}\sum_{j=0}^{N-1}\mathbf{n}(n+j)\mathbf{n}^{\mathrm{T}}(n+j). \tag{4.6}$$

To remove the bias introduced by the background noise it is possible to either aim at estimating and subsequently removing the noise component in (4.1), e.g., by using single-channel noise reduction techniques, or to estimate and remove the noise correlation matrix $\mathbf{R}_{\mathbf{nn}}$. The latter approach is already known from the literature on instantaneous BSS.

There, usually spatially and temporally uncorrelated Gaussian noise is assumed, i.e., $\mathbf{R_{nn}}$ is a diagonal matrix (see, e.g., [CDA98, DCA98, HKO01, CA02]). Moreover, most approaches assume $\mathbf{R_{nn}}$ is known a-priori and stationary. However, in realistic scenarios usually temporally correlated background noise is present at the sensors. The noise can often be described by a diffuse sound field, leading to noise signals which are also spatially correlated for low frequencies (see Section 2.2.4.3). Additionally, background noise is in general nonstationary and its stochastic properties can at best be assumed slowly time-variant which thus requires a continuous estimation of the correlation matrix $\mathbf{R_{nn}}$ based on short-time stationarity according to (4.6). The following bias removal techniques, aiming at the noise signal $\mathbf{n}$ or the noise correlation matrix $\mathbf{R_{nn}}$, will be examined under these conditions.

4.1.1.1 Single-channel noise reduction

If the estimation and suppression of the noise components $\mathbf{n}_p$ is desired for each sensor signal $\mathbf{x}_p$ individually, then for each channel $p = 1, \ldots, P$ a single-channel noise reduction algorithm can be used. The estimation and suppression of background noise using one channel is already a long-standing research topic. In general, all algorithms consist of two main building blocks: the estimation of the noise contribution and the computation of a weighting rule to suppress the noise and enhance the desired signal. An overview of various methods can be found, e.g., in [HS04, BMC05].

In all well-known noise estimation methods in the literature usually the noise power spectral density (psd) is estimated without recovering the phase of the clean signal but using the phase of the noisy signal instead. This is motivated by the fact that for power spectral density estimates the Wiener filter, which only modifies the signal amplitude, is optimal. Additionally, it was shown in [WL82] that the human perception of speech is not much affected by a modification of the phase of the clean signal. However, for BSS algorithms the relative phase of the signals acquired by different microphones is crucial as this information is implicitly used to suppress signals depending on their different direction of arrival. To evaluate the importance of amplitude and phase for pre-processing techniques applied to BSS algorithms, we will in the following generate pre-processed sensor signals by using the DFT-domain amplitude of the clean mixture signals and the phase of the noisy mixture signals. This corresponds to an optimum single-channel speech enhancement algorithm which perfectly estimates the amplitude of the clean mixture signal and thus, suppresses the background noise completely. These signals are then used as inputs for the second-order statistics BSS algorithm given in (3.222).

For this experiment we use two noisy scenarios. The first one is a car environment which is described in Appendix C.4. A pair of omnidirectional microphones with a spacing of 20 cm was used. The long-term SNR was adjusted to 0 dB which is a realistic value commonly encountered inside car compartments. Analogously to the BSS experiments in

Chapter 3 a male and a female speaker were convolved with the acoustic impulse response measured for the driver and co-driver positions. The second scenario corresponds to the cocktail party problem which is usually described by the task of listening to one desired point source in the presence of speech babble noise consisting of the utterances of many other speakers. The long term statistics of speech babble are well described by a diffuse sound field, however, there may also be several other distinct noise point sources present. In our experiments we simulated such a cocktail party scenario inside a living room environment (see Appendix C.2) where speech babble noise was generated by a circular loudspeaker array with a diameter of 3 m. The two omnidirectional microphones with a spacing of 20 cm were placed in the center of the loudspeaker array from which 16 speech signals were reproduced to simulate the speech babble noise. Additionally, two distinct point sources at a distance of 1 m and at the angles of 0° and -80° were used to simulate the desired and one interfering point source, respectively. For more details on the layout of the environment, see Appendix C.2. The long-term input SNR at the microphones has been adjusted for the living room scenario to 10 dB. This is realistic, as due to the speech-like spectrum of the background noise the microphone signals exhibiting higher SNR values are perceptually already as annoying as those with significantly lower SNR values for lowpass car noise.

Due to the perfect estimation of the clean signal amplitude the background noise is almost inaudible. However, the results in Fig. 4.1 show that due to the noisy phase for the car environment no improvement in terms of separation performance can be obtained. Similarly, for the cocktail party scenario only a small improvement in terms of separation of the point sources can be achieved. Further experiments also indicated that when using a realistic state-of-the-art noise reduction algorithm as, e.g., proposed in [Mar01a], then also the improvements shown in Fig. 4.1b disappear. Therefore, it is concluded that pre-processing by single-channel noise reduction algorithms only suppresses the background noise, but does not improve the degraded separation performance of the subsequent BSS algorithm. To improve the separation, it is crucial that both, amplitude and phase of the clean mixture signals are estimated. This usually requires multi-channel methods as presented in the next section.

4.1.1.2 Multi-channel bias removal

To also account for the phase contribution of the background noise there are some methods initially proposed for instantaneous BSS which aim at estimating and subsequently removing the noise correlation matrix $\mathbf{R}_{\mathbf{nn}}$. For convolutive BSS there have only been a few approaches proposed: In [HZ05] the special case of spatio-temporally white noise was addressed and has been extended to the diffuse noise case in [HZ06]. There, stationarity of the noise was assumed and the preceding noise-only segments have been used for the estimation of the correlation matrix. Already earlier in [ABK04, ABYK06] a similar

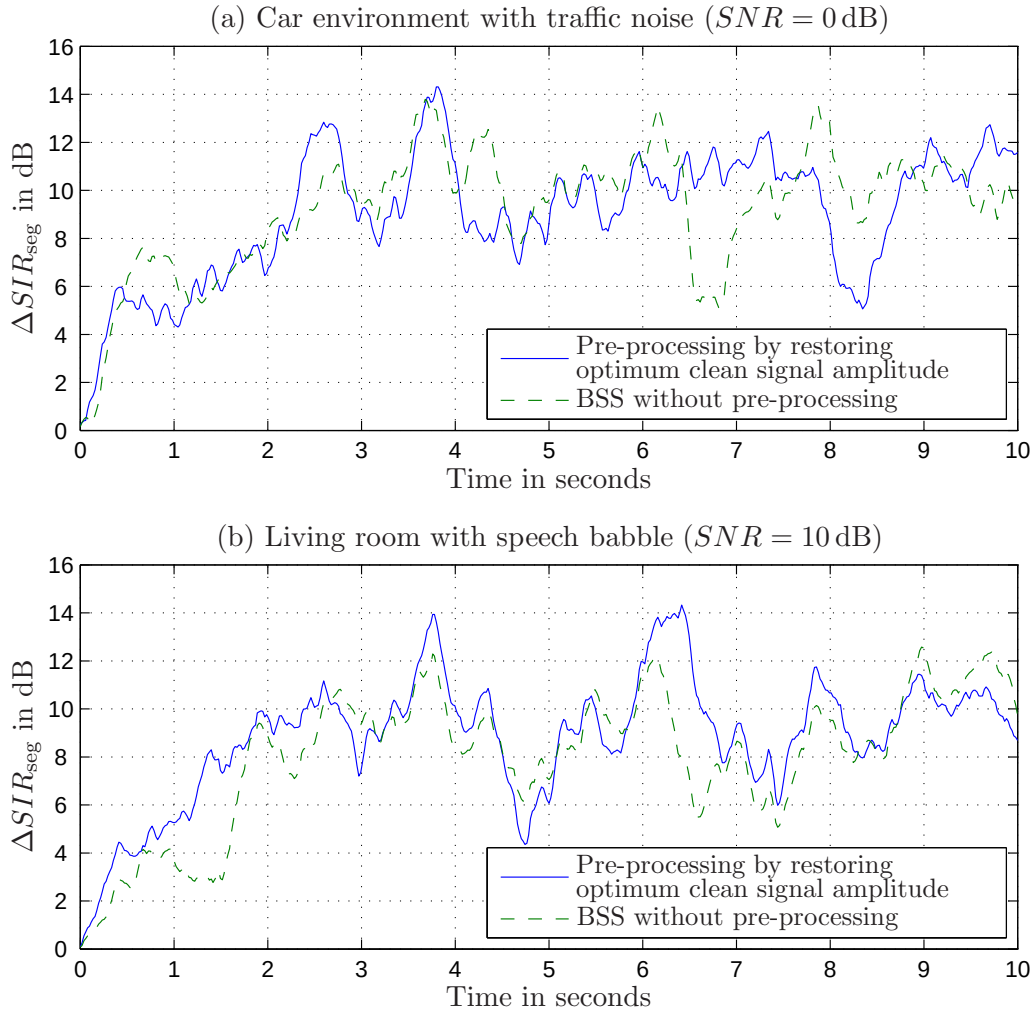


Figure 4.1: Segmental SIR improvement $\Delta SIR_{\mathrm{seg}}$ depicted over time for two noisy environments. Speech separation results are shown for the BSS outputs adapted with the noisy mixtures and for BSS with pre-processing by restoring the magnitude of the clean mixture signals but with the phase of the noisy mixtures.

procedure was proposed where the minimum statistics approach [Mar01a] was used for the estimation of the noise characteristics. This method operates in the DFT domain and is based on the observation that the power of a noisy speech signal frequently decays to the power of the background noise. Hence by tracking the minima an estimate for the auto-power spectral density of the noise is obtained. However, due to the spatial correlation not only the auto- but also the cross-power spectral densities of the noisy signal x_p and the background noise n_p are required. They are estimated and averaged recursively for each DFT bin whenever we detect a minimum (i.e. speech pause) of the noisy speech signals. Thus, for slowly time-varying noise statistics this method gives an accurate estimate of the noise spectral density matrix used for the bias removal. This has been applied as a pre-processing technique for the second-order statistics BSS algorithm.

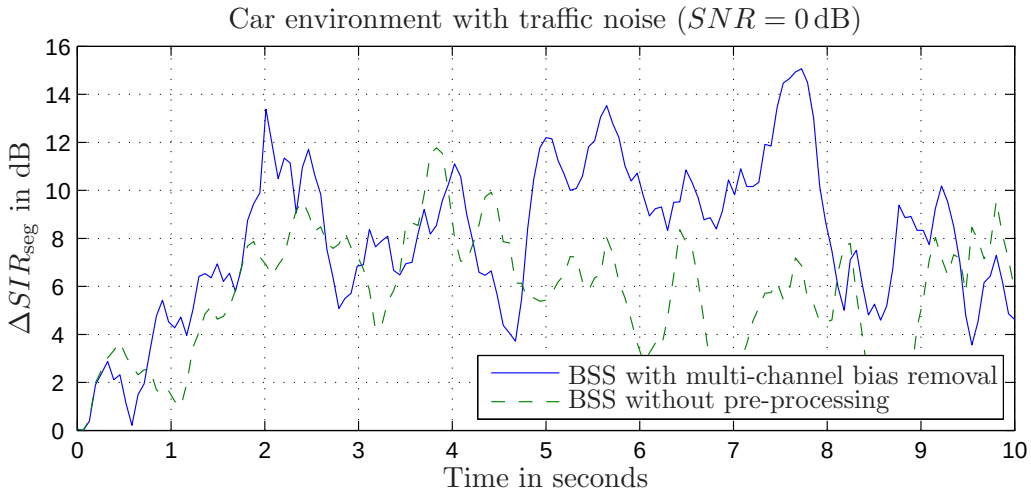


Figure 4.2: Segmental SIR improvement $\Delta SIR_{\mathrm{seg}}$ depicted over time for the noisy car environment. Speech separation results are shown for the BSS outputs adapted with the noisy mixtures and for BSS with pre-processing by multi-channel bias removal.

As the pre-processing is done in each DFT bin independently, we also used the narrowband second-order statistics BSS algorithm (3.223) which allows an easier integration of the pre-processing technique. In Fig. 4.2 the separation of the two point sources is depicted in terms of the segmental SIR improvement. It can be seen that the pre-processing slightly improves the separation performance for the noisy car environment. In the cocktail party scenario this approach did not achieve good results as the noise statistics are more time-variant and due to only few speech pauses of the point sources the noise psd cannot be estimated very well. Multi-channel bias removal approaches do not achieve any background noise reduction as they merely aim at providing a better estimate of the correlation matrix of the point sources which is then used for the adaptation of the demixing filter weights. To additionally suppress the background noise, this approach would have to be complemented by a post-processing technique. Note also that due to fewer speech pauses it is more difficult to estimate the noise correlation matrix for multiple active speakers compared to a single speaker as typically encountered in single-channel speech enhancement applications. Therefore, the estimation of the noise contribution may be done more reliably after the BSS stage where already a partial suppression of the interfering point sources is achieved. This will be investigated in detail for the post-processing approach discussed in Section 4.2

4.1.2 Subspace methods

Subspace methods are an attractive alternative to bias removal methods if more sensors than microphones are available, i.e., $P > Q$. Even if the main scope of this thesis is on the case $P = Q$, we will briefly review the history of subspace approaches and their

application to BSS algorithms and present some new ideas for future work.

Originally, subspace methods were proposed in [DBC91] for single-channel speech enhancement. There, it was shown that by using the singular value decomposition of a Toeplitz matrix containing several sensor signal samples, one could decompose the sensor signal into a signal-plus-noise subspace and a noise subspace. This decomposition is possible if the clean signal can be modeled by a low-rank model, which is known to be appropriate for speech. Similar results are obtained by the eigenvalue decomposition of the single-channel sensor signal correlation matrix [ET95]. After the decomposition, the noise subspace is removed and the clean speech signal is estimated from the remaining signal-plus-noise subspace. The resulting temporal filtering can be interpreted as an adaptive extraction of the most important formants of the speech signal, thereby reducing the amount of noise [DM02]. Moreover, it can be shown that this algorithm is analogous to the frequency-domain Wiener filter with the DFT replaced by an adaptive transform which projects the input speech into the signal subspace (see, e.g., [JC05]). Thus, as typical for single-channel speech enhancement algorithms, there is always a trade-off between noise reduction and signal distortion. A good review of single-channel approaches based on subspace techniques can be found, e.g., in [JC05].

More recently, this principle has been extended to multi-microphone approaches in [JC01, DM02]. The introduction of multiple sensors allows, additionally to the temporal filtering used in the single-channel case, also for spatial filtering. This means that in addition to the temporal samples now also the spatial degrees of freedom can be used to decompose the noisy sensor signals into a signal-plus-noise subspace and a noise subspace. This is achieved either by means of a generalized singular value decomposition of the stacked Toeplitz sensor signal matrices of all channels or the eigenvalue decomposition of the spatio-temporal sensor signal correlation matrices. Subsequently, the noise subspace is discarded by dimensionality reduction. Moreover, in [DM02] it is proposed to use multi-channel Wiener filtering for a further reduction of the noise in the signal-plus-noise subspace leading to the optimum minimum mean-squared error (MMSE) estimator. Similarly to BSS, the approach in [DM02] does not require any information about the positions of the desired sources nor is it affected by microphone tolerances. However, as the filtering is based on the multi-channel Wiener filter, an adaptation control is required to estimate the noise correlation matrix which is usually assumed to be slowly time-variant. The spatial directivity pattern presented in [DM02] showed that the solution obtained by the subspace approach can be interpreted as a beamformer whose adaptation relies on an adaptation control but does not require any geometrical information.

All subspace approaches discussed so far are based on the decomposition of time-domain quantities. Alternatively, one could also apply subspace methods in transform domains such as the DFT domain. In [AMAM00, AIO$^+$03] a DFT-domain subspace approach was proposed as a pre-processing step for a narrowband BSS algorithm. By

assuming $P > Q$ and by having a-priori information about how many point sources Q are present, the dimensionality is reduced in each DFT bin from P to Q. This method is computationally efficient as in each DFT bin only an eigenvalue decomposition of small $P \times P$ spatial correlation matrices is performed instead of using larger spatio-temporal correlation matrices. However, due to the independent bin-wise decomposition, no coupling between the DFT bins is retained and thus, it is not ensured that a permutation of the output channels of the subspace method is avoided. For narrowband BSS algorithms this is not critical as they are also applied independently to each DFT bin so that the permutation problem has to be resolved anyway. But due to this problem, narrowband subspace methods are not applicable to broadband algorithms without additional repair mechanisms which restore a consistent order of the output signals of the pre-processing scheme. Nevertheless, a combination of subspace methods and broadband BSS algorithms, which showed the best performance in the previous chapter, would be desirable.

A suitable subspace pre-processing scheme for broadband BSS algorithms could be developed by using a decomposition of a time-domain sensor signal matrix or sensor signal correlation matrix into the signal-plus-noise and the noise subspace analogously to [DM02]. In contrast to [DM02] only the noise subspace should be removed without additionally filtering the signal-plus-noise subspace. This means that only a dimension reduction from P sensor signals to Q signals in the signal-plus-noise subspace should be carried out. The filtering of the signal-plus-noise subspace, which usually requires an adaptation control, is then performed by the subsequent BSS algorithm. Hence, the adaptation control becomes unnecessary and the only required a-priori information is the number of the point sources Q, i.e., the number of dimensions to be retained. By this procedure, all point sources are preserved and the information of additional sensors is exploited to suppress background noise signals for the broadband algorithms considered in this thesis. This can be interpreted as a concatenation of a beamformer reducing the dimensionality from P to Q and a BSS system. Due to the focus of this dissertation on BSS models assuming $P = Q$, the discussion of this topic is not deepened here but may be considered as a future research topic.

4.2 Post-processing for suppression of residual crosstalk and background noise

In Section 4.1 several pre-processing approaches have been discussed. It could be seen that for the case $P = Q$ only multi-channel bias removal methods achieved some noise robustness of the BSS algorithm. For this a reliable voice activity detection is crucial but it might be difficult to realize this in environments with several speech point sources so that in such cases post-processing methods are a preferable alternative. Post-processing

methods have the advantage that the BSS system already achieves a suppression of the interfering point sources so that in each BSS output channel only some remaining interference of the other point sources is present. As will be shown later, this simplifies the estimation of the quantities required by the post-processing method. A suitable post-processing scheme is given by a single-channel postfilter $g_{q,\kappa}$ applied to each BSS output channel as shown in Fig. 4.3. The motivation of using a single-channel postfilter for each

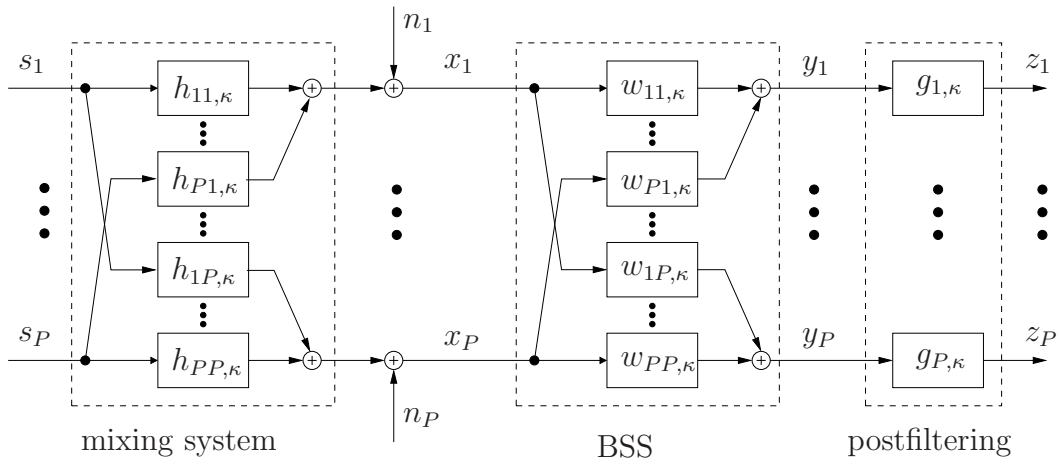


Figure 4.3: Noisy BSS model combined with postfiltering.

BSS output channel is twofold:

Firstly, it is desired that the remaining background noise is reduced at the BSS output channels. The background noise, usually described by a diffuse sound field, cannot be suppressed by the BSS algorithm as the demixing filters of the BSS system act for each output channel as a blind adaptive interference canceller aiming at the suppression of the interfering point sources (see Section 3.1.3). It is known from adaptive beamforming (e.g., [MMS98, SBM01]) that in environments with diffuse noise the concatenation of an adaptive interference canceller with a single-channel postfilter can improve the noise reduction.

Secondly, BSS algorithms are in noisy environments usually not able to converge to the optimum solution due to the bias introduced by the background noise. Moreover, moving point sources or an insufficient demixing filter length, which only partly covers the existing room reverberation, may lead to reduced signal separation performance and thus, to the presence of residual crosstalk from interfering point sources at the BSS output channels. In such situations, the single-channel postfilter should be designed such that it provides also additional separation performance. Analogously, similar considerations have led to a single-channel postfilter in acoustic echo cancellation which was first proposed in [MA95, AF95].

The reduced separation quality due to an insufficient demixing filter length in realistic

environments was the motivation of several single-channel postfilter approaches that have been previously proposed in the BSS literature [MASM02b, MASM02a, VL03, VRM04, CJLK04, PPSK06]. Nevertheless, a comprehensive treatment of the simultaneous suppression of residual crosstalk and background noise is still missing and will be presented in the following sections. We will first discuss in Section 4.2.1 the advantages of the implementation of the single-channel postfilter in the DFT domain and will introduce a spectral gain function requiring the power spectral density (psd) estimates of the residual crosstalk and background noise. Then, the signal model for the residual crosstalk and the background noise will be discussed in Section 4.2.2 allowing to point out the relationships to previous post-processing approaches. The chosen signal model will lead to the derivation of a novel residual crosstalk psd estimation and additionally the estimation of the background noise will be addressed. Subsequently, experimental results will be presented which illustrate the improvements that can be obtained by the application of single-channel postfilters both, in terms of SIR and SNR.

4.2.1 Spectral weighting function for a single-channel postfilter

In Chapter 2, the decomposition of the BSS output signals $y_q(n)$, $q = 1, \ldots, P$ was already given in (2.41) for the q-th channel as

$$y_q(n) = y_{\mathrm{s}_r,q}(n) + y_{\mathrm{c},q}(n) + y_{\mathrm{n},q}(n), \tag{4.7}$$

where $y_{\mathrm{s}_r,q}$ is the component containing the desired source $s_r(n)$. As a possible permutation of the separated sources at the BSS outputs, i.e., $r \neq q$ does not affect the post-processing approach we will simplify the notation and denote in the following the desired signal component in the q-th channel as $y_{\mathrm{s},q}$. The quantity $y_{\mathrm{c},q}$ is the residual crosstalk component from the remaining point sources that could not be suppressed by the BSS algorithm and $y_{\mathrm{n},q}$ denotes the contribution of the background noise.

From single-channel speech enhancement (e.g., [Mar05]) or from the literature on single-channel postfiltering for beamforming (e.g., [SBM01]) it is well-known that it is beneficial to utilize the DFT-domain representation of the signals and estimate the single-channel postfilter in the DFT domain. Thus, N_{post} samples are combined to an output signal block which is, after applying a windowing operation, transformed by the DFT of length $R_{\mathrm{post}} \geq N_{\mathrm{post}}$ yielding the DFT-domain representation of the output signals as

$$\underline{Y}_q^{(\nu)}(m) = \underline{Y}_{\mathrm{s},q}^{(\nu)}(m) + \underline{Y}_{\mathrm{c},q}^{(\nu)}(m) + \underline{Y}_{\mathrm{n},q}^{(\nu)}(m) \tag{4.8}$$

where $\nu = 0, \ldots, R_{\mathrm{post}} - 1$ is the index of the DFT bin and m denotes the block time index. The advantage is that in the DFT domain speech signals are sparser, i.e., we can find regions in the time-frequency plane where the individual speech sources do not overlap (see e.g., [BAM03, YR04]). This property is often exploited in underdetermined blind

source separation where there are more simultaneously active sources than sensors (e.g., [Div05, WB06]). Here, this sparseness is used for the estimation of the quantities necessary for the implementation of the spectral gain function. A block diagram showing the main building blocks of a DFT-based postfilter is given in Fig. 4.4. There it can already

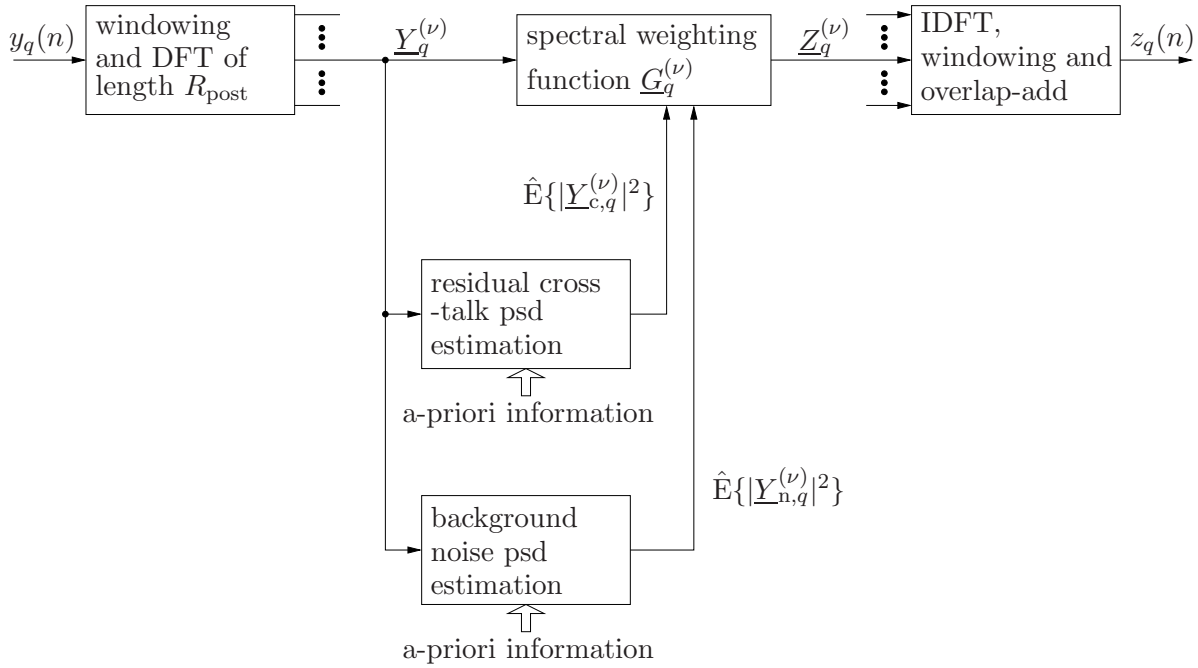


Figure 4.4: DFT-based single-channel postfiltering depicted for the ν-th DFT bin in the q-th channel.

be seen that analogously to single-channel speech enhancement or post-filtering applied to beamforming or acoustic echo cancellation, the DFT bins are treated in a narrowband manner as all computations are carried out independently in each DFT bin. Because of the narrowband treatment we have to ensure that circular convolution effects, appearing due to the signal modification by the spectral weighting, are not audible. Thus, the enhanced output signal z_q, which is the estimate $\hat{y}_{\text{s},q}$ of the clean desired source component, is computed by means of an inverse DFT using a weighted overlap-add method including a tapered analysis and synthesis windows as suggested in [GL84, MC99]. This is in contrast to the BSS algorithms treated in Chapter 3 where the linear convolution of the sensor signals with the estimated FIR demixing system is implemented without approximations equivalently in the DFT domain by the overlap-save method. In Chapter 3, selective narrowband approximations have only been made in the adaptation process of the demixing filters to obtain efficient BSS algorithms.

According to Fig. 4.4 a spectral gain function $\underline{G}_q^{(\nu)}$ in the ν-th DFT bin aiming at simultaneous suppression of residual crosstalk and background noise has to be derived. The output signal of the post-processing scheme is the estimate of the clean desired source

signal $\underline{Z}_q^{(\nu)} = \hat{\underline{Y}}_{\mathrm{s},q}^{(\nu)}$ and is given as

$$\underline{Z}_q^{(\nu)}(m) = \underline{G}_q^{(\nu)}(m)\underline{Y}_q^{(\nu)}(m). \tag{4.9}$$

According to [AZBK06] we choose in this thesis to minimize the mean-squared error $\mathrm{E}\left\{\left(\underline{Z}_q^{(\nu)}(m) - \underline{Y}_{\mathrm{s},q}^{(\nu)}(m)\right)^2\right\}$ with respect to $\underline{G}_q^{(\nu)}$. This leads to the ν-th bin of the well-known Wiener filter for the q-th channel given as (see, e.g., [Mar05, DHP00])

$$\underline{G}_q^{(\nu)}(m) = \frac{\mathrm{E}\{|\underline{Y}_{\mathrm{s},q}^{(\nu)}(m)|^2\}}{\mathrm{E}\{|\underline{Y}_q^{(\nu)}(m)|^2\}}. \tag{4.10}$$

With the assumption that the desired signal component, the interfering signal components and the background noise in the q-th channel are all mutually uncorrelated, (4.10) can be expressed as

$$\underline{G}_q^{(\nu)}(m) = \frac{\mathrm{E}\{|\underline{Y}_{\mathrm{s},q}^{(\nu)}(m)|^2\}}{\mathrm{E}\{|\underline{Y}_{\mathrm{s},q}^{(\nu)}(m)|^2\} + \mathrm{E}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)|^2\} + \mathrm{E}\{|\underline{Y}_{\mathrm{n},q}^{(\nu)}(m)|^2\}}. \tag{4.11}$$

From this equation it can be seen that for regions with desired signal *and* residual crosstalk or background noise components the output signal spectrum is reduced, whereas in regions without crosstalk or background noise the signal is passed through. On the one hand this fulfills the requirement that an undisturbed desired source signal passes through the Wiener filter without any distortion. On the other hand, if crosstalk or noise is present, the magnitude spectrum of the noise or crosstalk attains a shape similar to that of the desired source signal, so that noise and crosstalk are therefore partially masked by the desired source signal. This effect was already exploited in postfiltering for acoustic echo cancellation aiming at the suppression of residual echo. There, this effect has been termed "echo shaping" [Mar96]. Moreover, it can be observed in (4.11) that if the BSS system achieves the optimum solution, i.e., the residual crosstalk in the q-th channel $\underline{Y}_{\mathrm{c},q}^{(\nu)} = 0$, then (4.11) reduces to the well-known Wiener filter for a signal with additive noise used in single-channel speech enhancement. To realize (4.11) in a practical system, the ensemble average $\mathrm{E}\{\cdot\}$ has to be estimated and thus, it is usually replaced by a time average $\hat{\mathrm{E}}\{\cdot\}$. Thereby, the Wiener filter is approximated by

$$\underline{G}_q^{(\nu)}(m) \approx \frac{\hat{\mathrm{E}}\{|\underline{Y}_q^{(\nu)}(m)|^2\} - \hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)|^2\} - \hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{n},q}^{(\nu)}(m)|^2\}}{\hat{\mathrm{E}}\{|\underline{Y}_q^{(\nu)}(m)|^2\}}, \tag{4.12}$$

where $\hat{\mathrm{E}}\{|\underline{Y}_q^{(\nu)}|^2\}$, $\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}|^2\}$, and $\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{n},q}^{(\nu)}|^2\}$ are the psd estimates of the BSS output signal, residual crosstalk, and background noise, respectively. Due to the reformulation in (4.12) the unobservable desired signal psd $\mathrm{E}\{|\underline{Y}_{\mathrm{s},q}^{(\nu)}(m)|^2\}$ does not have to be estimated. However, the main difficulty is still to obtain reliable estimates of the unobservable residual

crosstalk and background noise psds. A novel method for this estimation process leading to high noise reduction with little signal distortion will be shown in the next section. Moreover, an estimate of the observable BSS output signal psd is required. The psd estimates can be used to implement spectral weighting algorithms other than the Wiener filter as described, e.g., in [HS04, Mar05].

4.2.2 Estimation of residual crosstalk and background noise

In this section a model for the residual crosstalk and background noise is introduced. Subsequently, based on the residual crosstalk model an estimation procedure will be given which relies on an adaptation control. Different adaptation control strategies will be outlined. Moreover, also the estimation of the background noise psd will be discussed.

4.2.2.1 Model of residual crosstalk and background noise

We restricted our scenario to the case that the number of microphones equals the maximum number of simultaneously active point sources. Therefore, the BSS algorithm is able to provide an estimate of one separated point source at each output y_q. As pointed out above, due to movement of sources or long reverberation, the BSS algorithm might not converge fast enough to the optimum solution and thus, some residual crosstalk from point source interferers, denoted in the DFT domain by $\underline{Y}_{\mathrm{c},q}^{(\nu)}$, remains in the BSS output. To obtain a good estimate of the residual crosstalk psd $\mathrm{E}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}|^2\}$ as needed for the post-filter in the q-th channel, we first need to set up an appropriate model.

In Fig. 4.5a the concatenation of the mixing and demixing is expressed by the overall system matrix $\check{\mathbf{C}} = \mathbf{H}_L\check{\mathbf{W}}$ which was introduced in (3.10). The entries $\mathbf{c}_{qr}$ of $\check{\mathbf{C}}$ denote the path from the q-th source to the r-th output. For simplicity, we have depicted the case $Q = P = 2$ in Fig. 4.5. As can be seen in Fig. 4.5a, the crosstalk component $y_{\mathrm{c},1}(n)$ of the first output channel is determined in the case $Q = P = 2$ by the source signal $s_2(n)$ and the filter $\mathbf{c}_{21}$. However, as neither the original source signals nor the overall system matrix are observable, the crosstalk component $y_{\mathrm{c},1}(n)$ is expressed in Fig. 4.5b in terms of the desired source signal component $y_{\mathrm{s},2}(n)$ at the second output. This residual crosstalk model could be used if a good estimate of $y_{\mathrm{s},2}(n)$ is provided by the BSS system, i.e., if the source in the second channel is well-separated.

It should also be noted that even if $y_{\mathrm{s},2}(n)$ is available, then this model does not allow a perfect estimation of the residual crosstalk $y_{\mathrm{c},1}(n)$. This is due to the fact that for a perfect replica of $y_{\mathrm{c},1}(n)$ based on the input signal $y_{\mathrm{s},2}(n)$, the filter $\mathbf{b}_{21}$ has to model the combined system of $\mathbf{c}_{21}$ and the inverse of $\mathbf{c}_{22}$. However, $\mathbf{c}_{22}$ is in general a non-minimum phase FIR filter and thus, cannot be inverted in an exact manner by a single-input single-output system as was shown in [MK88]. Hence, analogously to single-channel blind dereverberation approaches, it is only possible to obtain an optimum filter

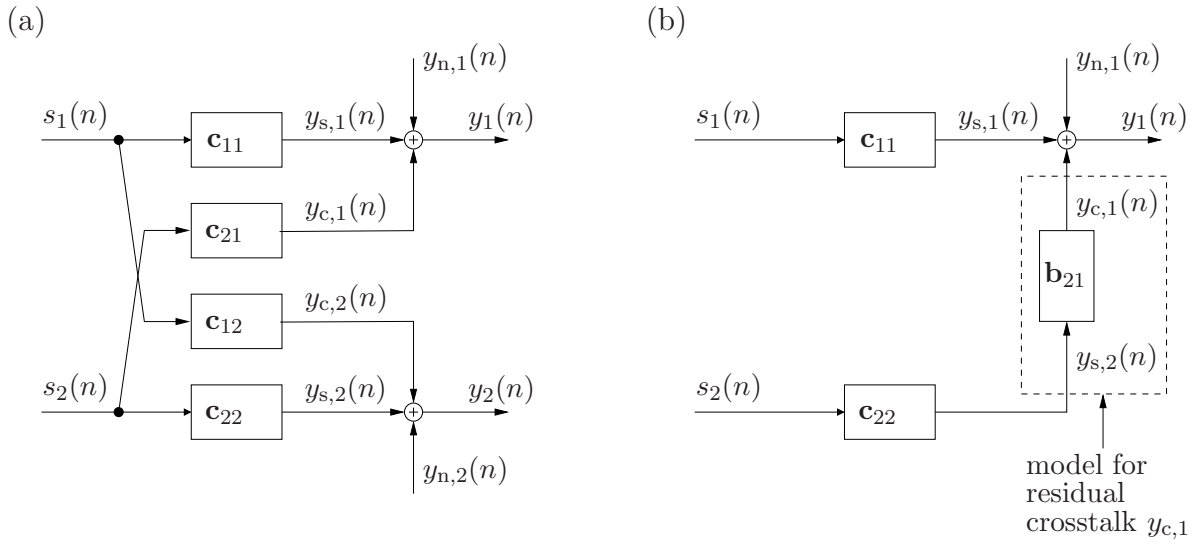


Figure 4.5: (a) Representation of mixing and demixing system for the case $P = 2$ by using the overall system FIR filters $\mathbf{c}_{qr}$. (b) Resulting model for the residual crosstalk $y_{\mathrm{c},1}(n)$.

$\mathbf{b}_{21}$ in the least-squares sense [MK88]. We will see in the following that due to the usage of additional a-priori information this model is nevertheless suitable for the estimation of the residual-cross talk psd.

The model in Fig. 4.5b requires the desired source signal component $y_{\mathrm{s},2}(n)$ in the second BSS output. However, in practice it cannot be assumed that the BSS system always achieves perfect source separation. Especially in the initial convergence phase or with moving sources, there is some residual crosstalk remaining in all outputs. Therefore, we have to modify the residual crosstalk model so that only observable quantities are used. Hence, in Fig. 4.6 the desired signal component $y_{\mathrm{s},i}$ for the i-th channel is replaced by the signal $\breve{y}_{i,q}$ which denotes the BSS output signal of the i-th channel but without any interfering crosstalk components from the q-th point source (i.e., desired source s_q). This means that the overall filters $\mathbf{c}_{qi}$ from the q-th source to the i-th output ($i = 1, \ldots, P$, $i \neq q$) are assumed to be zero. In practice, this condition is fulfilled by an adaptation control which determines time-frequency points where the desired source s_q is inactive. This a-priori information about desired source absence is important for a good estimation of the residual crosstalk psd and thus, for achieving additional residual crosstalk cancellation. A detailed discussion of the adaptation control will be given in Section 4.2.2.3. Due to the bin-wise application of the single-channel postfilter we will in the following formulate the model in the DFT domain. Consequently, the model for the residual crosstalk in the q-th channel based on observable quantities is expressed for the ν-th DFT bin ($\nu = 1, \ldots, R_{\mathrm{post}}$)

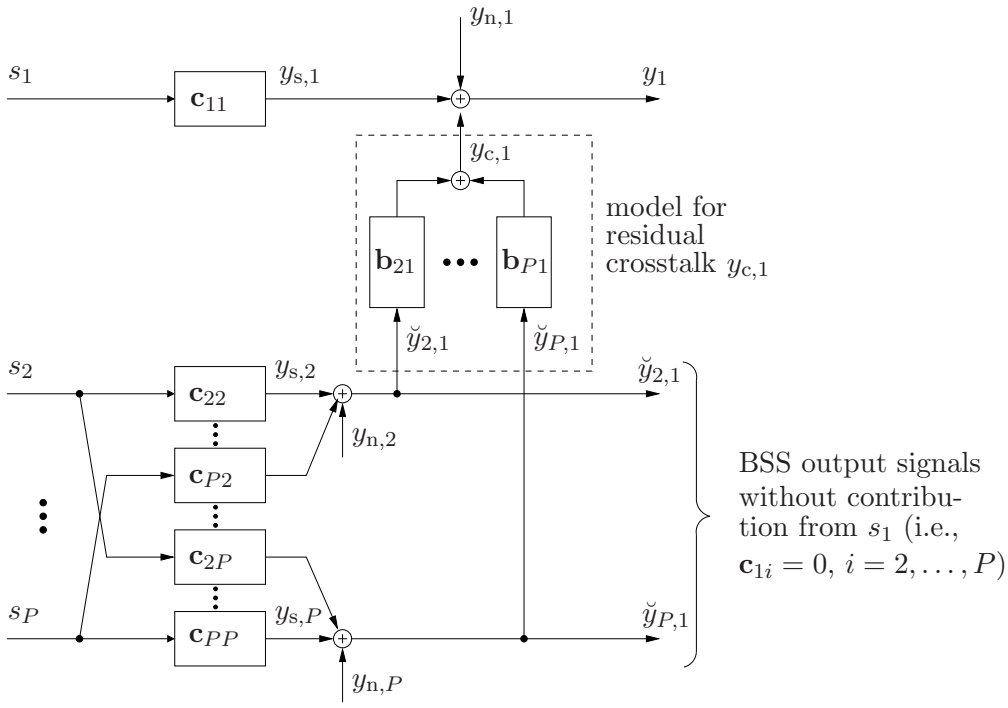


Figure 4.6: Model of the residual crosstalk component $y_{\mathrm{c},q}$ contained in the q-th BSS output channel y_q illustrated for the first channel, i.e., $q = 1$. In contrast to Fig. 4.5b this model is solely based on observable quantities.

as

$$\begin{aligned} \underline{Y}_{\mathrm{c},q}^{(\nu)}(m) &= \sum_{i=1, i\neq q}^{P} \breve{\underline{Y}}_{i,q}^{(\nu)}(m)\underline{B}_{i,q}^{(\nu)}(m) \\ &= \breve{\underline{\mathbf{y}}}_{q}^{(\nu)^{\mathrm{T}}}(m)\underline{\mathbf{b}}_{q}^{(\nu)}(m), \end{aligned} \tag{4.13}$$

where $\breve{\underline{Y}}_{i,q}^{(\nu)}$ and $\underline{B}_{i,q}^{(\nu)}$ are the DFT-domain representations of $\breve{y}_{i,q}$ and $\mathbf{b}_{iq}$, respectively. The variable $\breve{\underline{\mathbf{y}}}_{q}^{(\nu)}$ is the $P-1$ dimensional DFT-domain column vector containing $\breve{\underline{Y}}_{i,q}^{(\nu)}$ for $i = 1, \ldots, P$, $i \neq q$, and $\underline{\mathbf{b}}_{q}^{(\nu)}$ is the column vector containing the unknown filter weights $\underline{B}_{i,q}^{(\nu)}$ for $i = 1, \ldots, P$, $i \neq q$.

It should be pointed out that the adaptation control only ensures that the desired source s_q is absent in the i-th BSS output channel $\breve{\underline{Y}}_{i,q}^{(\nu)}$. However, the background noise $\underline{Y}_{\mathrm{n},i}^{(\nu)}$ is still present in the i-th BSS output channel as can also be seen in Fig. 4.6. If the background noise is spatially correlated between the q-th and i-th BSS output channel, then the coefficient $\underline{B}_{i,q}^{(\nu)}$ would not only model the leakage from the separated source in i-th channel, but $\underline{B}_{i,q}^{(\nu)}$ would also be affected by the spatially correlated background noise. However, as an additional measure, the noise psd $\mathrm{E}\{|\underline{Y}_{\mathrm{n},q}^{(\nu)}|^2\}$ is estimated individually in each channel by one of the noise estimation methods known from single-channel speech en-

hancement. Therefore, if the background noise is already included in the residual crosstalk model, this would lead to an overestimation of the noise psd. In Chapter 2 the character of the background noise such as car or babble noise was examined and the correlation of the noise sources between the sensors was evaluated using the magnitude-squared coherence (MSC). It was concluded that the MSC of such background noise exhibits the same characteristics as a diffuse sound field leading to strong spatial correlation for low frequencies but to very small spatial correlation at higher frequencies. The model of the residual crosstalk is based on the BSS output signals and hence, it is of interest how the BSS system changes the MSC of the noise signals. In Fig. 4.7a and Fig. 4.7b the MSC

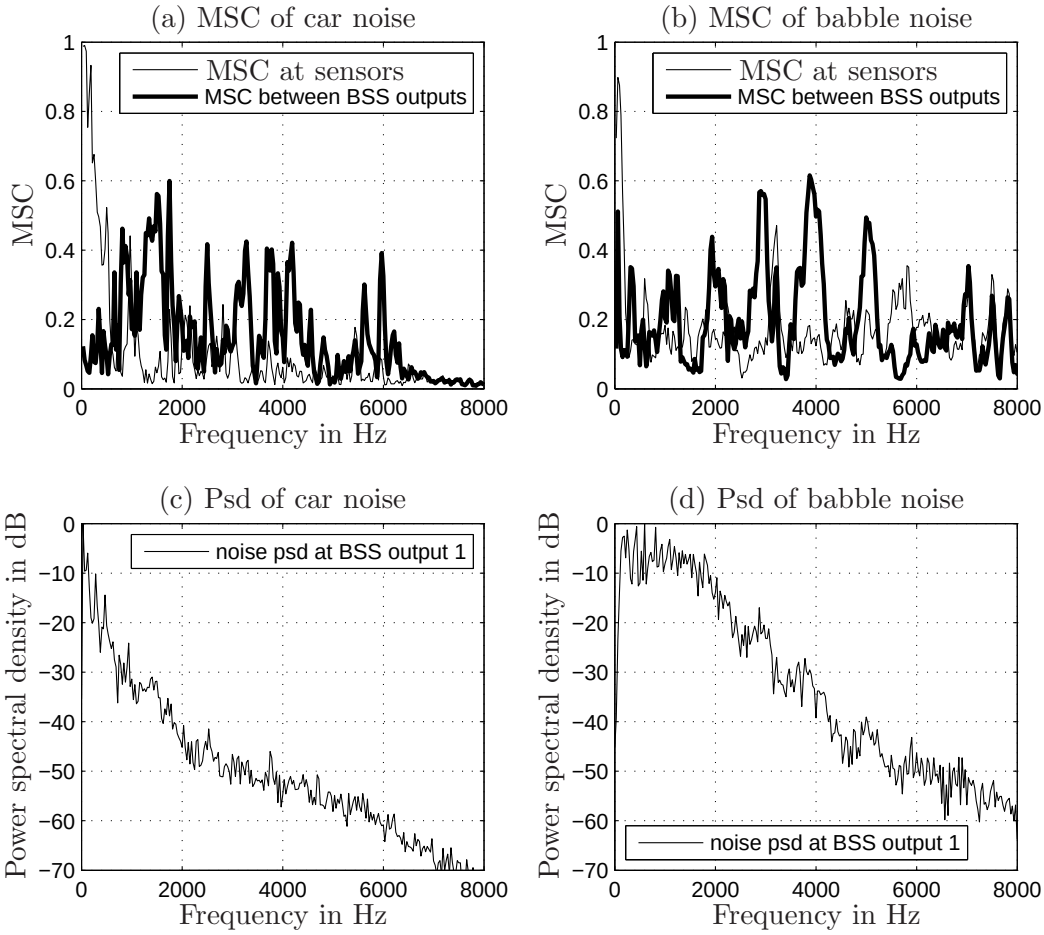


Figure 4.7: Magnitude-squared coherence (MSC) of the car (a) and speech babble (b) background noise between the sensors and between the BSS outputs. The long-term noise psds of car noise and speech babble are shown in (c) and (d), respectively.

of car noise and babble noise, which was estimated recursively according to (2.33), (2.35) with the parameters $R = 512$, $\gamma = 0.9$, $\alpha = 2$, is plotted. For the car scenario described

in the Appendix C.4 a two-microphone array with a spacing of 4 cm was mounted at the interior mirror and the driver and co-driver were speaking simultaneously. Then the block-online BSS algorithm given in (3.222) was applied. The same experiment was performed with two sources in the reverberant room described in the Appendix C.2 where the babble noise was generated by a circular loudspeaker array using 16 individual speech signals. As pointed out in the previous chapter, the BSS algorithm tries to achieve source separation by aiming at mutual independence of the BSS output signals. From Fig. 4.7 it can be seen that in the presence of background noise this also leads to a spatial decorrelation of the noise signals at the BSS outputs. The car noise which is dominant at low frequencies (see Fig. 4.7c) has a MSC close to zero at these frequencies (see Fig. 4.7a). Only at higher frequencies, where the noise signal has much less energy, a larger MSC can be observed. The reduction of the MSC for the relevant frequencies can analogously also be observed for the babble noise. This observation shows that the background noise is spatially decorrelated at the BSS outputs and thus, confirms that the model for the residual crosstalk introduced in (4.13) is valid even in the case of background noise. This also justifies the independent estimation of the background noise in each channel and thus, we can apply noise estimation methods previously derived for single-channel speech enhancement algorithms. The residual cross-talk, however, is correlated across the output channels. These characteristics of residual cross-talk and background noise will be exploited in the next section to derive suitable estimation procedures.

After introducing the residual crosstalk model and validating it for the case of existing background noise, we briefly discuss the relationships to the models used in previous publications on post-processing for BSS. In [PPSK06] a Wiener-based approach for residual crosstalk cancellation is presented for the case $P = 2$. There, a very simple model is used where all coefficients $\underline{B}_{i,q}^{(\nu)}$ $(i, q = 1, 2,\, i \neq q)$ are assumed to be equal to one. Similarly, in [VRM04] one constant factor was chosen for all $\underline{B}_{i,q}^{(\nu)}$. A model closer to (4.13), but based on magnitude spectra, was given in [CJLK04] which was then used for the implementation of a spectral subtraction rule. In contrast to the estimation method presented in the next section, the frequency-dependent coefficients of the model were learned by a modified least-mean-squares (LMS) algorithm. In [MASM02b] and [MASM02a] more sophisticated models were proposed allowing for time-delays or FIR filtering in each DFT bin. The model parameters were estimated by exploiting correlations between the channels or by using an NLMS algorithm. In all of these single-channel approaches the information of the multiple channels is only exploited to estimate the psds necessary for the spectral weighting rule. Alternatively, if also the phase of $\underline{Y}_{\mathrm{c},q}^{(\nu)}$ is estimated, then it is also possible to directly subtract the estimate of the crosstalk component $\underline{Y}_{\mathrm{c},q}^{(\nu)}$ from the q-th channel. This was proposed in [LNT04] resulting in an adaptive noise canceller (ANC) structure [WGM$^+$75]. The ANC was adapted by a leaky LMS algorithm [Gre98] which includes a variable step size to allow also for strong desired signal activity without the necessity of

an adaptation control.

The background noise component in the q-th channel $\underline{Y}_{\mathrm{n},q}^{(\nu)}$ is usually assumed to be more stationary than the desired signal component $\underline{Y}_{\mathrm{s},q}^{(\nu)}$. This assumption is necessary for the noise estimation methods known from single-channel speech enhancement which will be used to estimate the noise psd $\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{n},q}^{(\nu)}|^2\}$ in each channel and which are briefly discussed in the next section.

4.2.2.2 Estimation of residual crosstalk and background noise power spectral densities

After introducing the residual crosstalk model (4.13) we need to estimate the psds $\mathrm{E}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}|^2\}$ of the residual crosstalk and $\mathrm{E}\{|\underline{Y}_{\mathrm{n},q}^{(\nu)}|^2\}$ of the background noise for evaluating (4.12). To obtain an estimation procedure based on observable quantities we first calculate the cross-power spectral density vector $\underline{\mathbf{s}}_{\breve{\mathbf{y}}_q Y_{\mathrm{c},q}}^{(\nu)}$ between $\underline{\breve{\mathbf{y}}}_q^{(\nu)}$ and $\underline{Y}_{\mathrm{c},q}^{(\nu)}$ given as

$$\begin{aligned}\underline{\mathbf{s}}_{\breve{\mathbf{y}}_q Y_{\mathrm{c},q}}^{(\nu)}(m) &= \hat{\mathrm{E}}\{\underline{\breve{\mathbf{y}}}_q^{(\nu)^*}(m)\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)\} \\ &= \hat{\mathrm{E}}\{\underline{\breve{\mathbf{y}}}_q^{(\nu)^*}(m)\underline{\breve{\mathbf{y}}}_q^{(\nu)^{\mathrm{T}}}(m)\}\underline{\mathbf{b}}_q^{(\nu)}(m) \\ &=: \underline{\mathbf{S}}_{\breve{\mathbf{y}}_q\breve{\mathbf{y}}_q}^{(\nu)}(m)\underline{\mathbf{b}}_q^{(\nu)}(m), \end{aligned} \tag{4.14}$$

where in the first step $\underline{\mathbf{b}}_q^{(\nu)}$ was assumed to be slowly time-varying. Using (4.13), the power spectral density estimate $\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}|^2\}$ can be expressed as

$$\begin{aligned}\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)|^2\} &= \hat{\mathrm{E}}\{\underline{Y}_{\mathrm{c},q}^{(\nu)^H}(m)\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)\} \\ &= \underline{\mathbf{b}}_q^{(\nu)^H}(m)\underline{\mathbf{S}}_{\breve{\mathbf{y}}_q\breve{\mathbf{y}}_q}^{(\nu)}(m)\underline{\mathbf{b}}_q^{(\nu)}(m). \end{aligned} \tag{4.15}$$

Solving (4.14) for $\underline{\mathbf{b}}_q^{(\nu)}$ and inserting it into (4.15) leads to

$$\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)|^2\} = \underline{\mathbf{s}}_{\breve{\mathbf{y}}_q Y_{\mathrm{c},q}}^{(\nu)^H}(m)\left(\underline{\mathbf{S}}_{\breve{\mathbf{y}}_q\breve{\mathbf{y}}_q}^{(\nu)}(m)\right)^{-1}\underline{\mathbf{s}}_{\breve{\mathbf{y}}_q Y_{\mathrm{c},q}}^{(\nu)}(m). \tag{4.16}$$

As $\underline{Y}_{\mathrm{c},q}^{(\nu)}$, $\underline{Y}_{\mathrm{s},q}^{(\nu)}$, and $\underline{Y}_{\mathrm{n},q}^{(\nu)}$ in Fig. 4.6 are assumed to be mutually uncorrelated, $\underline{\mathbf{s}}_{\breve{\mathbf{y}}_q Y_{\mathrm{c},q}}^{(\nu)}$ can also be estimated as the cross-power spectral density $\underline{\mathbf{s}}_{\breve{\mathbf{y}}_q Y_q}^{(\nu)}$ between $\underline{\breve{\mathbf{y}}}_q^{(\nu)}$ and q-th output of the BSS system $\underline{Y}_q^{(\nu)}$ leading to the final estimation procedure:

$$\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)|^2\} = \underline{\mathbf{s}}_{\breve{\mathbf{y}}_q Y_q}^{(\nu)^H}(m)\left(\underline{\mathbf{S}}_{\breve{\mathbf{y}}_q\breve{\mathbf{y}}_q}^{(\nu)}(m)\right)^{-1}\underline{\mathbf{s}}_{\breve{\mathbf{y}}_q Y_q}^{(\nu)}(m). \tag{4.17}$$

Thus, the power spectral density of the residual crosstalk for the q-th channel can be efficiently estimated in each DFT bin $\nu = 0, \ldots, R-1$ by computing the $1 \times P-1$ cross-power spectral density vector $\underline{\mathbf{s}}_{\breve{\mathbf{y}}_q Y_q}^{(\nu)}$ between input and output of the model shown in Fig. 4.6 and calculating the $P-1 \times P-1$ cross-power spectral density matrix $\underline{\mathbf{S}}_{\breve{\mathbf{y}}_q\breve{\mathbf{y}}_q}^{(\nu)}$ of the inputs. One possible implementation for estimating this expectation is given by

an exponentially weighted average $\hat{\mathrm{E}}\{a(m)\} = (1-\gamma)\sum_i \gamma^{m-i} a(i)$, where $a(m)$ is the quantity to be averaged. The advantage is that this can also be formulated recursively leading to

$$\underline{\mathbf{S}}^{(\nu)}_{\breve{\mathbf{y}}_q\breve{\mathbf{y}}_q}(m) = \gamma \underline{\mathbf{S}}^{(\nu)}_{\breve{\mathbf{y}}_q\breve{\mathbf{y}}_q}(m-1) + (1-\gamma)\underline{\breve{\mathbf{y}}}^{(\nu)^*}_q(m)\underline{\breve{\mathbf{y}}}^{(\nu)^{\mathrm{T}}}_q(m), \tag{4.18}$$

$$\underline{\mathbf{s}}^{(\nu)}_{\breve{\mathbf{y}}_q Y_q}(m) = \gamma \underline{\mathbf{s}}^{(\nu)}_{\breve{\mathbf{y}}_q Y_q}(m-1) + (1-\gamma)\underline{\breve{\mathbf{y}}}^{(\nu)^*}_q(m)\underline{Y}^{(\nu)^{\mathrm{T}}}_q(m). \tag{4.19}$$

In summary, the power spectral density of the residual crosstalk for the q-th channel can be efficiently estimated in each DFT bin $\nu = 0, \ldots, R-1$ using (4.17) together with the recursive calculation of the $P-1 \times P-1$ cross-power spectral density matrix (4.18) and the $P-1 \times 1$ cross-power spectral density matrix vector (4.19). It should be noted that such an estimation technique has also been used to determine a post-filter for residual echo suppression in the context of acoustic echo cancellation (AEC) [TGSB97, EMV02a]. However, the method presented in [TGSB97, EMV02a] is different in two ways: Firstly, in contrast to BSS where several interfering point sources may be active, the AEC post-filter was derived for a single channel, i.e., the residual echo originates from only one point source and thus all quantities in (4.17) reduce to scalar values. Secondly, in the AEC problem a reference signal for the echo is available. In BSS however, $\underline{\breve{\mathbf{y}}}^{(\nu)}_q$ is not immediately available as it can only be estimated if the desired source signal in the q-th channel is currently inactive. Strategies how to determine such time intervals are discussed in the next section.

The estimation of the psd of the background noise $\hat{\mathrm{E}}\{|\underline{Y}^{(\nu)}_{\mathrm{n},q}|^2\}$ is already a long-standing research topic in single-channel speech enhancement and an overview on the various methods can be found, e.g., in [HS04]. Usually it is assumed that the noise psd is at least more stationary than the desired speech psd. The noise estimation can be performed during speech pauses, which have to be detected properly by a voice activity detector. As voice activity detection algorithms are rather unreliable in low SNR conditions, several methods have been proposed which can track the noise psd continuously. One of the most prominent methods is the minimum statistics approach proposed in [Mar94]. This method is based on the observation that the power of a noisy speech signal frequently decays to the power of the background noise. Hence, by tracking the minima the power spectral density of the noise is obtained. In [Mar01a] an improved version was proposed which includes an optimal smoothing of the noise psd together with a bias correction and which will be applied in the experiments in Section 4.2.3. Other methods providing continuous noise psd estimates can be found, e.g., in [CB01, Coh03, HS04].

4.2.2.3 Adaptation control based on SIR estimation

In the previous sections it was shown that the estimation of the residual crosstalk power spectral density in the q-th channel is only possible at time instants when the desired point source of the q-th channel is inactive. As pointed out already above, speech signals can be assumed to be sufficiently sparse in the time-frequency domain so that even in reverberant environments regions can be found where one or more sources are inactive (see, e.g., [BAM03, YR04]). This fact will be exploited by constructing a DFT-based adaptation control necessary for the estimation of the residual cross-talk psd. In this section we will first briefly review an adaptation control approach which is already known from the literature on post-processing for BSS. Due to the similarity of the adaptation control necessary for estimating the residual crosstalk and the control necessary for adaptive beamformers applied to acoustic signals, also the existing approaches in the beamforming literature will be briefly summarized. A sophisticated bin-wise adaptation control originally proposed in [HTK03, HNK04] in the context of adaptive beamforming will then be applied in a slightly modified version to the post-processing scheme.

In general, all adaptation controls aim at estimating the SIR in the time domain or in a bin-wise fashion in the DFT domain. For the latter, the SIR estimate is given for the ν-th DFT bin as the ratio of the desired signal psd and the psd of the interfering signals. Thus, the SIR estimate at the q-th BSS output is given as

$$\widehat{\underline{SIR}}_q^{(\nu)}(m) = \frac{\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{s},q}^{(\nu)}(m)|^2\}}{\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)|^2\}}. \tag{4.20}$$

For the case of a BSS system with two output channels ($P = 2$) together with the assumption that the number of simultaneously active point sources $Q \leq P$, a simple SIR estimate is given by approximating the desired signal component $\underline{Y}_{\mathrm{s},q}^{(\nu)}$ with the BSS output signal $\underline{Y}_q^{(\nu)}$ of the q-th channel and approximating the interfering signal component by the BSS output signal of the other channel. This yields, e.g., for the approximated SIR estimate in the first BSS output channel

$$\widehat{\underline{SIR}}_1^{(\nu)}(m) \approx \frac{\hat{\mathrm{E}}\{|\underline{Y}_1^{(\nu)}(m)|^2\}}{\hat{\mathrm{E}}\{|\underline{Y}_2^{(\nu)}(m)|^2\}}. \tag{4.21}$$

This approximation is justified if the BSS system already provides enough separation performance so that the BSS output signals can be seen as estimates of the point sources. In [MASM02b, MASM02a] the time-average $\hat{\mathrm{E}}\{\cdot\}$ in (4.21) has been approximated by taking the instantaneous psd values and the resulting approximated SIR was used successfully as a decision variable for controlling the estimation of the residual crosstalk. If $\widehat{\underline{SIR}}_1^{(\nu)}(m) < 1$, then the crosstalk $\underline{Y}_{\mathrm{c},1}^{(\nu)}$ was estimated and for $\widehat{\underline{SIR}}_1^{(\nu)}(m) > 1$ the crosstalk of the second channel $\underline{Y}_{\mathrm{c},2}^{(\nu)}$ was determined. In [AZBK06] this adaptation control was refined by the introduction of a safety margin Υ to improve reliability. By comparing (4.21)

to a fixed threshold Υ it is ensured that a certain SIR value $\widehat{\underline{SIR}}_1^{(\nu)}(m) < \Upsilon$ has to be attained to allow the conclusion that the desired signal is absent and thus, allow estimation of the residual crosstalk $\underline{Y}_{\mathrm{c},1}^{(\nu)}$. The safety margin Υ has to be chosen between $0 < \Upsilon \leq 1$ and was set in [AZBK06] to $\Upsilon = 0.9$. For an extension of this mechanism to $P, Q > 2$ a suitable approximation for $\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)|^2\}$ in the SIR estimate (4.20) is important. In [AZBK06] it was suggested for $P, Q > 2$ to use the maximum psd of the remaining channels $\hat{\mathrm{E}}\{|\underline{Y}_i^{(\nu)}(m)|^2\}$, $i \neq q$. For increasing P, Q this requires a very careful choice of Υ. In such scenarios, it is advantageous to replace the fixed threshold Υ by adaptive thresholding. As we will see in the following, such sophisticated adaptation controls were treated in the beamforming literature and will now be applied to the BSS post-processing scheme.

If adaptive beamformers, such as the generalized sidelobe canceller (GSC) [GJ82], are applied to acoustic signals, then usually an adaptation control is required for the adaptive filters aiming at interference cancellation. Analogously to the residual crosstalk estimation procedure discussed in Section 4.2.2.2, the adaptation of the adaptive interference canceller has to be stalled in the case of a strong desired signal. This analogy allows to apply the approaches in the literature on adaptive beamforming to the post-processing of the BSS output signals. To control the adaptation of beamformers, a correlation-based method was proposed in [GZ92] and recently in a modified form also in [HSLF06]. Another approach relies on the comparison of the outputs of a fixed beamformer with the main lobe steered towards the desired source and a complementary beamformer which steers a spatial null towards the desired source [HS99]. The ratio of the output signal powers, which constitutes an estimate of the SIR is then compared to a threshold to decide if the adaptation should be stopped. As both methods were suggested in the time-domain, this corresponds to a full-band adaptation control, so that in case of a strong desired signal the adaptation is stopped for all DFT bins. It has been pointed out before that speech signals are sparse in the DFT domain and thus, better performance of the adaptation algorithm can be expected when using a bin-wise adaptation control. This was the motivation in [HTK03, HNK04] for transferring the approach based on two fixed beamformer outputs to the DFT domain leading to a frequency-dependent SIR estimate. Instead of a fixed threshold Υ, additionally an adaptive threshold $\underline{\Upsilon}_q^{(\nu)}(m)$ for each channel and DFT bin has been proposed leading to a more robust decision. The application of this adaptation control to the estimation of the residual crosstalk, which is required for the post-processing algorithm, will be discussed in the following.

In [HTK03, HNK04] the estimate $\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{s},q}^{(\nu)}(m)|^2\}$ of the desired signal required for the SIR estimate (4.20) is obtained by a delay-and-sum beamformer. This requires an array of several microphones which should have a spacing that is sufficiently large to allow the suppression of the interfering signals also at low frequencies. Moreover, the positions of the microphones are assumed to be known. This is in contrast to the BSS application where the sensors can be arbitrarily positioned and where there might be only a small

number of sensors available (e.g., $P = 2$). Therefore, instead of a fixed beamformer output we will use the q-th BSS output signal psd $\hat{\mathrm{E}}\{|\underline{Y}_q^{(\nu)}(m)|^2\}$ as an estimate of the desired signal psd $\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{s},q}^{(\nu)}(m)|^2\}$.

The estimate of the interfering signal components required for the SIR estimate (4.20) are obtained in [HTK03, HNK04] by a complementary beamformer which places a spatial null towards the desired source. The difference to the procedure in [HS99] is that this is done in a bin-wise manner. In our application we will use the psd of a complementary BSS signal $\underline{\bar{Y}}_q^{(\nu)}$ which is obtained analogously to [HTK03, HNK04] as

$$\hat{\mathrm{E}}\{|\underline{\bar{Y}}_q^{(\nu)}(m)|^2\} = \hat{\mathrm{E}}\{|\underline{X}_q^{(\nu)}(m)|^2\} - \hat{\mathrm{E}}\{|\underline{Y}_q^{(\nu)}(m)|^2\}. \tag{4.22}$$

Similarly to the approach in [HTK03, HNK04], it is assumed here that the filtering due to the BSS demixing system is approximately linear phase and that the BSS output signal and the microphone signal have been properly time-aligned before subtracting their psd estimates. It should be noted that due to the usage of a broadband BSS algorithm, the permutation at the BSS output signals is not frequency-dependent. Therefore, a possible permutation of the BSS output channels has no effect on the calculation of the complementary BSS signal.

Usually recursive averaging is used for the time-average indicated by the operator $\hat{\mathrm{E}}\{\cdot\}$ which leads to the psd estimates

$$\underline{S}_{x_q x_q}^{(\nu)}(m) = \gamma \underline{S}_{x_q x_q}^{(\nu)}(m-1) + (1-\gamma)|\underline{X}_q^{(\nu)}(m)|^2, \tag{4.23}$$

$$\underline{S}_{y_q y_q}^{(\nu)}(m) = \gamma \underline{S}_{y_q y_q}^{(\nu)}(m-1) + (1-\gamma)|\underline{Y}_q^{(\nu)}(m)|^2, \tag{4.24}$$

necessary for the estimation of the SIR in the q-th BSS output channel. The SIR estimate (4.20) can thus be expressed as

$$\widehat{\underline{SIR}}_q^{(\nu)}(m) \approx \frac{\underline{S}_{y_q y_q}^{(\nu)}(m)}{\underline{S}_{x_q x_q}^{(\nu)}(m) - \underline{S}_{y_q y_q}^{(\nu)}(m)}. \tag{4.25}$$

The SIR estimate (4.25) is then compared to an adaptive threshold $\underline{\Upsilon}_q^{(\nu)}(m)$. If $\widehat{\underline{SIR}}_q^{(\nu)}(m) < \underline{\Upsilon}_q^{(\nu)}(m)$, then the absence of the desired signal in the q-th channel can be assumed. The adaptive threshold is given as the minimum of SIR estimate $\widehat{\underline{SIR}}_q^{(\nu)}(m)$ which is determined for each DFT bin by taking into account the last D_Υ blocks [Mar01a]. In practice D_Υ must be large enough to bridge any peak of desired signal activity but short enough to track the nonstationary SIR variations in case of absence of the desired signal. Here, we choose an interval equivalent to a time period of 1.5 s. Moreover, for small variations $\left|\left(\widehat{\underline{SIR}}_q^{(\nu)}(m) - \underline{\Upsilon}_q^{(\nu)}(m)\right)/\underline{\Upsilon}_q^{(\nu)}(m)\right| \leq \Delta\Upsilon$ the threshold $\underline{\Upsilon}_q^{(\nu)}(m)$ is updated immediately. In Fig. 4.8 the SIR estimate $\widehat{\underline{SIR}}_q^{(\nu)}$ and the adaptive threshold $\underline{\Upsilon}_q^{(\nu)}$ determined by minimum tracking are illustrated for the DFT bin corresponding to 1 kHz.

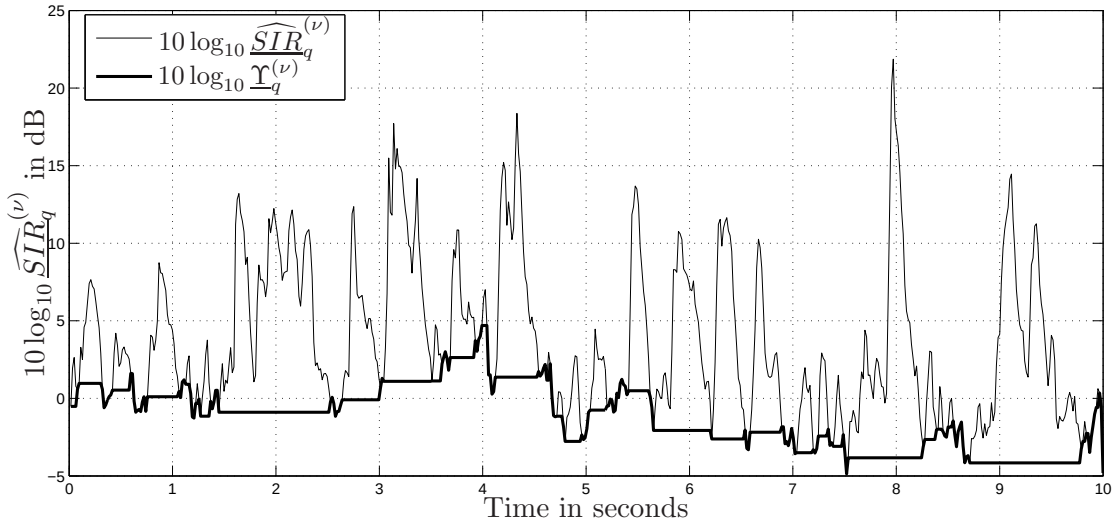


Figure 4.8: Estimate $10\log_{10}\widehat{\underline{SIR}}_q^{(\nu)}$ of the SIR and adaptive threshold $\underline{\Upsilon}_q^{(\nu)}$ determined by minimum tracking illustrated for the DFT bin corresponding to 1 kHz.

The results are based on the output signals of the BSS system applied to the car environment described in Appendix C.4. It can be seen that due to the parameter $\Delta\Upsilon = 0.3$ the threshold follows small changes of the SIR estimate immediately. Moreover, it should be pointed out that the SIR estimate in Fig. 4.8 exhibits high positive values due to the good convergence of the BSS algorithm. This is the reason why even in speech pauses of the desired signal, the SIR estimate does rarely exhibit negative SIR values.

In Fig. 4.9 the decision of the adaptation control is illustrated for the first output channel of the BSS system applied to the car environment ($P = Q = 2$). The desired component, residual crosstalk, and background noise component at the first BSS output are depicted in (a)-(c). The decision of the adaptation control is obtained by estimating the SIR according to (4.25) solely based on observable quantities. Especially due to the existence of background noise $y_{\mathrm{n},q}$ this leads to a biased SIR. Nevertheless, the adaptation control is very robust due to the adaptive threshold $\underline{\Upsilon}_1^{(\nu)}$ based on minimum tracking and the parameter $\Delta\Upsilon = 0.3$ which allows for small variation of the threshold. This can be seen, when comparing the results of the adaptation control with the true SIR illustrated in (e) which is estimated based on unobservable quantities according to (4.20). In case of high SIR values $\widehat{\underline{SIR}}_1^{(\nu)}$, the desired signal in the first channel is present and the residual crosstalk psd $\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{c},2}^{(\nu)}(m)|^2\}$ of the other channel is estimated. Vice versa, a low SIR in the first channel allows to adapt $\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{c},1}^{(\nu)}(m)|^2\}$.

In case that the adaptation control stalls the estimation of the residual crosstalk for the ν-th DFT bin in one of the P BSS output channels, the residual crosstalk estimate from the previous block has to be used. As speech is a nonstationary process and therefore, the statistics of the residual crosstalk are quickly time-varying, this would deteriorate

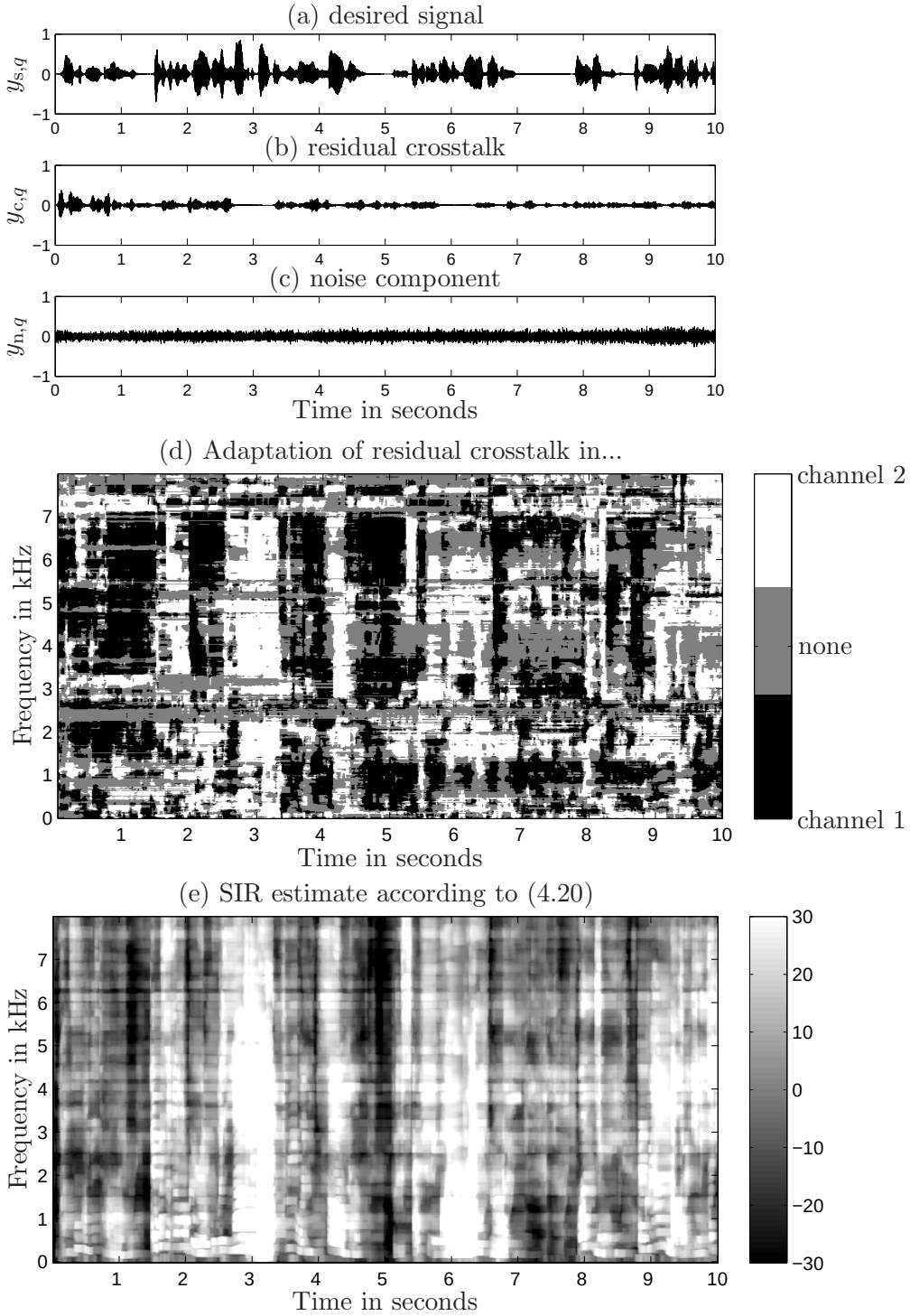


Figure 4.9: BSS output signal components for the car environment (Appendix C.4) with an input SNR at the sensors of 0 dB showing the desired signal (a), residual crosstalk (b) and background noise (c) in the first channel. Based on the SIR estimate (4.25) and the adaptive threshold $\underline{\Upsilon}_1^{(\nu)}$ the decision of the adaptation control is shown in (d). For comparison, the SIR (4.20) computed for the true signal components in the first channel is illustrated in (e).

1.	Estimate $\underline{S}^{(\nu)}_{x_q x_q}(m)$ and $\underline{S}^{(\nu)}_{y_q y_q}(m)$ according to (4.23) and (4.24)
2.	Estimate $\widehat{\underline{SIR}}_q^{(\nu)}(m)$ according to (4.25)
3.	Estimate $\hat{\mathrm{E}}\{\lvert\underline{Y}^{(\nu)}_{\mathrm{n},q}(m)\rvert^2\}$ by minimum statistics algorithm
4.	Tracking of minima of $\widehat{\underline{SIR}}_q^{(\nu)}(m)$: If $\lvert(\widehat{\underline{SIR}}_q^{(\nu)}(m) - \underline{\Upsilon}_q^{(\nu)}(m))/\underline{\Upsilon}_q^{(\nu)}(m)\rvert \leq \Delta\Upsilon$ Replace all values of $\underline{\Upsilon}_q^{(\nu)}(i)$ inside the buffer, i.e., $\underline{\Upsilon}_q^{(\nu)}(i) = \widehat{\underline{SIR}}_q^{(\nu)}(m)$, $i = m, \ldots, m - D_\Upsilon + 1$ If $\widehat{\underline{SIR}}_q^{(\nu)}(m)$ is the minimum of $\underline{\Upsilon}_q^{(\nu)}(m-i)$, $i = 0, \ldots, D_\Upsilon - 1$ Set current value of buffer $\underline{\Upsilon}_q^{(\nu)}(m) = \widehat{\underline{SIR}}_q^{(\nu)}(m)$
5.	If minimum is detected, i.e., $\widehat{\underline{SIR}}_q^{(\nu)}(m) \leq \underline{\Upsilon}_q^{(\nu)}(m)$: Calculate residual crosstalk $\hat{\mathrm{E}}\{\lvert\underline{Y}^{(\nu)}_{\mathrm{c},q}(m)\rvert^2\}$ according to (4.17) Compute postfilter (4.12) for residual crosstalk and noise
6.	If no minimum is detected, i.e., $\widehat{\underline{SIR}}_q^{(\nu)}(m) > \underline{\Upsilon}_q^{(\nu)}(m)$: Compute postfilter (4.26) for noise only

Table 4.1: Adaptation control and application of the postfilter for the q-th BSS output channel and ν-th DFT bin.

the performance of the postfilter $\underline{G}_q^{(\nu)}$. On the other hand, as pointed out above, the minimum statistics algorithm can provide continuous noise psd estimates even in periods with desired signal activity. Therefore, for those time instants where the estimate of residual crosstalk cannot be updated, i.e., where the desired source signal is dominant, a postfilter

$$\underline{G}^{(\nu)}_{\mathrm{n},q}(m) = \frac{\hat{\mathrm{E}}\{\lvert\underline{Y}_q^{(\nu)}(m)\rvert^2\} - \hat{\mathrm{E}}\{\lvert\underline{Y}^{(\nu)}_{\mathrm{n},q}(m)\rvert^2\}}{\hat{\mathrm{E}}\{\lvert\underline{Y}_q^{(\nu)}(m)\rvert^2\}} \tag{4.26}$$

merely aiming at suppression of the background noise is applied.

In Table 4.1 the adaptation control and the resulting application of the postfilters is outlined for the q-th BSS output channel.

4.2.3 Experimental results

In the evaluation of the postfiltering algorithm summarized in Table 4.1 the same two noisy scenarios have been considered as in Section 4.1 and their description is briefly summarized. The first one is a car environment which is described in Appendix C.4. A pair of omnidirectional microphones with a spacing of 20 cm was used with a male and female speaker at the driver and co-driver positions, respectively. The long-term SNR was adjusted to 0 dB which is a realistic value commonly encountered inside car compartments. The second scenario corresponds to the cocktail party problem which is usually described

by the task of listening to one desired point source in the presence of speech babble noise consisting of the utterances of many other speakers. Speech babble is well described by a diffuse sound field, however, there may also be several other distinct noise point sources present. In our experiments we simulated such a cocktail party scenario inside a living room environment (see Appendix C.2) where speech babble noise was generated by a circular loudspeaker array with a diameter of 3 m. The two omnidirectional microphones with a spacing of 20 cm were placed in the center of the loudspeaker array from which 16 speech signals were reproduced to simulate the speech babble noise. Additionally, two distinct point sources at a distance of 1 m and at the angles of 0° and -80° were used to simulate the desired and one interfering point source, respectively. For more details on the layout of the environment, see Appendix C.2. The long-term input SNR at the microphones has been adjusted for the living room scenario to 10 dB. This is realistic, as due to the speech-like spectrum of the background noise the microphone signals which exhibit higher SNR values are perceptually already as annoying as those with significantly lower SNR values for lowpass car noise.

The second-order statistics BSS algorithm with the narrowband normalization given in (3.222) is applied to the two noisy scenarios. To evaluate the performance two measures have been used: The segmental SIR which is defined as the ratio of the signal power of the desired signal to the signal power of the residual crosstalk stemming from point source interferers and the segmental SNR defined as the ratio of the signal power of the desired signal to the signal power of the possibly diffuse background noise. In both cases, the SIR and SNR improvement due to the application of the postfilter is measured and averaged over both channels. The segmental SIR improvement $\Delta SIR_{\text{seg}}(m)$ defined in (2.50) is plotted as a function of the block index m to illustrate the convergence effect of the BSS system. The channel-averaged segmental SNR improvement $\Delta\overline{SNR}_{\text{seg}}$ is given as the average over all blocks. To assess the desired signal distortion, the unweighted log-spectral distance (SD) defined in (2.54) has been measured between the input and the output of the postfilter. The DFT length for computing the SD is usually set to be small so that speech can be assumed stationary. In our experiments we used $R = 256$ and set K_S large enough to cover the whole signal length.

To reduce artifacts such as, e.g., musical noise, the postfilter (4.12) is usually calculated using an adaptive oversubtraction factor $\xi_q^{(\nu)}$ as proposed in [BSM79]. Moreover, negative gains of the postfilters are set to zero. Hence in the experiments the postfilter

$$\underline{G}_q^{(\nu)}(m) = \frac{\max\left[\left(\hat{\mathrm{E}}\{|\underline{Y}_q^{(\nu)}(m)|^2\} - \xi_q^{(\nu)}\left(\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{n},q}^{(\nu)}(m)|^2\} + \hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)|^2\}\right)\right), 0\right]}{\hat{\mathrm{E}}\{|\underline{Y}_q^{(\nu)}(m)|^2\}}. \quad (4.27)$$

was used. For the post-processing algorithm, $\gamma = 0.9$ and a DFT length of $R_{\text{post}} = 2048$ was chosen. The block length N_{post} was equal to the DFT length and an overlap factor $\alpha = 4$ was used. The parameters of the adaptation control are given as $\Delta\Upsilon = 0.3$ and

$D_\Upsilon = 94$ corresponding to a period of 1.5 s over which the minimum is tracked.

In Fig. 4.10 the results for the separation of the two speech point sources can be seen. For both scenarios the separation performance of the combined system of BSS and single-

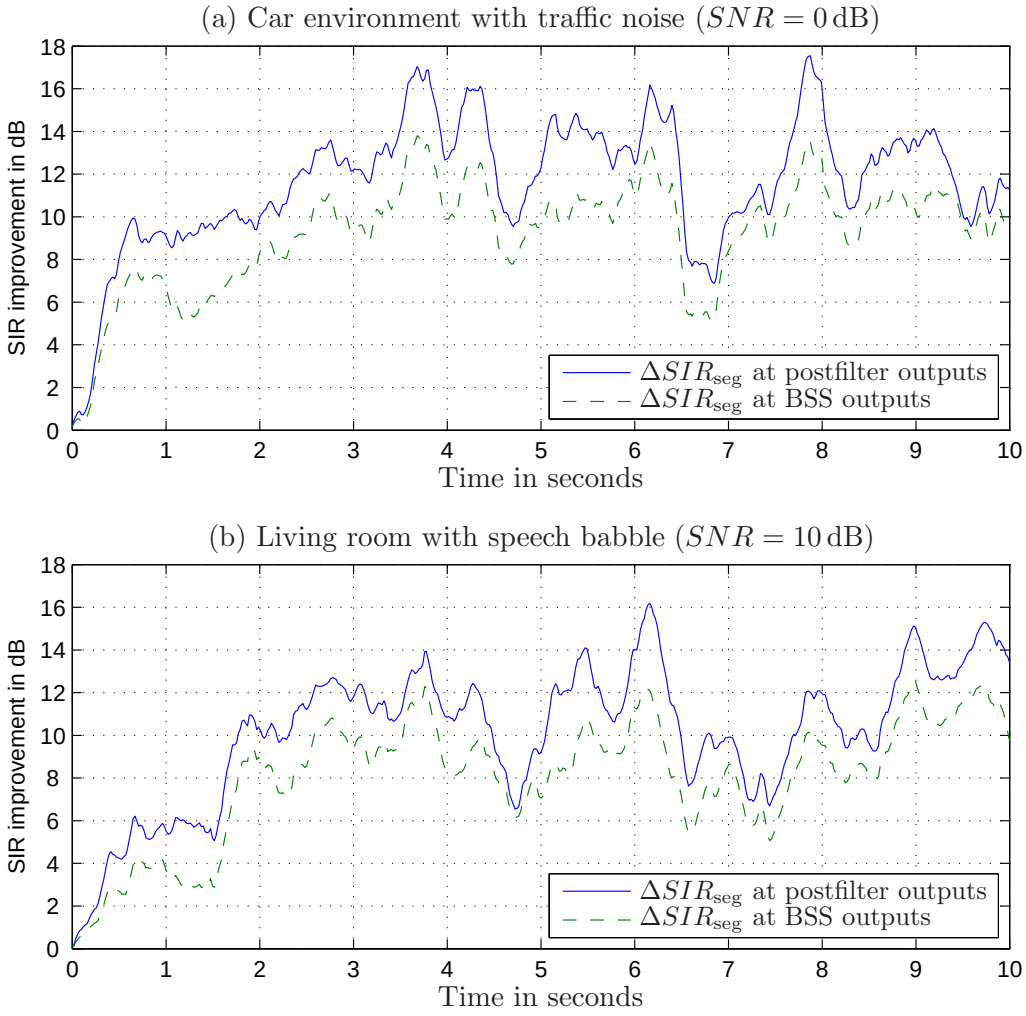


Figure 4.10: Segmental SIR improvement ΔSIR_{seg} depicted over time for two environments containing two speech point-source and additional background noise: (a) car compartment with background noise consisting of car and traffic noise ($SNR = 0$ dB) and (b) living room scenario with speech babble background noise from 16 speakers ($SNR = 10$ dB). Speech separation results are shown for BSS outputs and postfilter outputs.

channel postfilter (solid) outperforms the BSS performance (dashed). In contrast to the BSS system which possesses an inherent adaptation control implied by the normalization term in the update equation, the postfilter relies on a-priori information provided by the adaptation control. Hence, it is possible to accurately estimate the residual crosstalk at the BSS outputs and therefore, to further improve the speech separation performance. The reduced absolute level of the SIR improvement in the cocktail party scenario, i.e., in the reverberant living room (Fig. 4.10b) is due to longer reverberation and especially due

	$\Delta\overline{SNR}_{\text{seg}}$ at BSS outputs	$\Delta\overline{SNR}_{\text{seg}}$ at post-filter outputs	SD at postfilter outputs
Car scenario	3.0 dB	4.9 dB	1.0 dB
Cocktail party scenario	0.2 dB	1.3 dB	1.6 dB

Table 4.2: Segmental SNR and unweighted log-spectral distortion for both scenarios

to the background babble noise which exhibits a speech-like long-term spectrum.

Moreover, in the both scenarios also the background noise could be partially suppressed. In Table 4.2 the segmental SNR averaged over all output channels of the BSS system and of the postfilter is shown. It can be observed that the postfilter achieves an additional SNR gain. As the car noise is more stationary compared to the speech babble noise, the minimum statistics algorithm can better estimate the noise psd and thus a higher SNR improvement can be achieved by the postfilter.

To assess the speech quality, the unweighted log-spectral distortion between the desired signal at the input and output of the post-filter was calculated and averaged over both output channels. The small values in Table 4.2 indicate that the quality of the desired signal is preserved. This was also confirmed by informal listening tests where additionally no musical noise was observed.

4.3 Summary

In this chapter the convolutive BSS framework presented in Chapter 3 for reverberant environments has been extended to the case of noisy mixtures. As realistic background noise types often have a diffuse sound field characteristic it is not possible to address their suppression or separation from the point sources by the standard convolutive BSS model. Therefore, different pre- and post-processing methods have been investigated to complement the BSS algorithms. These methods have to achieve two different goals: First, the adaptation of the BSS filters should be robust against background noise and second, the background noise should be suppressed at the outputs.

In Section 4.1 this was addressed by pre-processing techniques which have to estimate and remove the bias introduced by the background noise components. This can be done either by an estimation of the noise contribution at the sensors or by an estimation of the noise correlation matrix. The former can be achieved by applying single-channel speech enhancement techniques to each sensor signal individually. All well-known single-channel approaches only aim at recovering the clean speech signal magnitude and do not aim at the estimation of the phase. This is motivated by the fact that only noise power spectral density estimates are available so that a Wiener filter, which only modifies the magnitude of the noisy signal is optimal. Moreover, human perception is not affected very much by a

modification of the clean signal phase. However, for BSS algorithms the relative phase of the signals acquired by the different microphones is very important. This was confirmed by the experiments which showed that only a restoration of the clean sensor signal magnitude does not lead to improved results. Hence, a multi-channel bias removal technique was investigated which aims at the estimation of the noise correlation matrix. This has been achieved by using an adaptation control based on jointly evaluating multiple sensor signals which provided information on the time instants when only noise is present. This method showed some improvement in the separation performance of the BSS algorithm. However, for multiple speakers less speech pauses are present and thus the estimation of the noise correlation matrix can be performed at fewer time instants. Moreover, the pre-processed correlation matrix is only used for the adaptation of the demixing system so that in contrast to the single-channel pre-processing method another technique for a suppression of the background noise would be required. Due to these reasons, post-processing methods are a favorable alternative. For completeness, also subspace methods which are applicable to the case of more sensors than sources, i.e., $P > Q$ have been briefly discussed.

In Section 4.2 post-processing methods based on single-channel postfiltering have been investigated. As the BSS performance will deteriorate in the presence of noise, these methods have to account for both, suppression of the residual crosstalk from the interfering point sources and suppression of the background noise. A novel estimation procedure has been presented for the estimation of the residual crosstalk power spectral density (psd) necessary for the Wiener-based weighting rule of the postfilter. An adaptation control known from adaptive beamforming has been applied in a modified form which made the estimation of the residual crosstalk psd quite robust. In contrast to the pre-processing techniques, the adaptation control utilizes the BSS outputs where the interfering point sources are already partially suppressed. For the background noise psd estimation a conventional noise estimator from single-channel speech enhancement has been used. The resulting postfilter has led to an increased separation performance and also to a partial suppression of the background noise in both, a noisy car environment and a cocktail party scenario where babble noise is encountered.

5 Summary and Conclusions

In recent years a lot of research has been devoted to convolutive BSS algorithms for acoustic signals. There are several reasons why there are many efforts to apply the technique of BSS to acoustic human-machine interfaces which is an area where fixed and adaptive beamforming schemes are still predominant. One reason is that BSS approaches only rely on the assumption of mutual independence of the source signals and do not need additional a-priori information, such as the array geometry or the positions of the desired and interfering sources. Moreover, different frequency characteristics of the individual microphones do not affect the performance of BSS algorithms. Another reason is that in several applications such as surveillance of public spaces, it is desirable to simultaneously focus on several different point sources instead of extracting one desired source as common in beamforming. Furthermore, the mean-squared error approaches, commonly encountered in adaptive beamforming, are inherently based on second-order statistics. In contrast to these methods, BSS algorithms can be based on information theoretic criteria which allow to incorporate also higher-order statistics into the adaptation algorithms. Due to these advantages, BSS techniques for acoustic signals have received a great deal of attention in the signal processing community in the last few years.

The topic of this thesis has been aiming at BSS for acoustic signals and the main achievements can be described as follows. We presented several important special cases of a generic BSS framework which was termed TRINICON ("TRIple-N Independent component analysis for CONvolutive mixtures") in [BAK03a]. This framework allows a unified view on convolutive BSS algorithms leading to novel algorithms and showing also relationships to popular state-of-the-art algorithms. Here, we presented some approximations which lead to highly efficient algorithms while still preserving the superior properties of the general framework relative to other known algorithms. A second major contribution is that, beyond existing BSS literature, we addressed in this thesis the application of BSS to reverberant *and* noisy environments by presenting several pre- and post-processing techniques in a coherent treatment. The algorithms allow to maintain a high separation performance of the BSS algorithms also in noisy scenarios and additionally, are capable of suppressing the undesired diffuse background noise which cannot be treated by the convolutive BSS model.

To achieve these results, we first introduced a broadband time-domain optimization criterion based on a generalization of the mutual information measure. This criterion is based on the mutual independence of the source signals but allows temporal dependencies

within each source signal. By using multivariate probability density functions (pdfs) in the criterion it is possible to account for the nonwhiteness of the source signals. This allowed us to exploit all three usable (i.e., "TRIple-N") signal properties nongaussianity, nonwhiteness and nonstationarity by a broadband time-domain criterion.

Subsequently, several broadband iterative gradient descent and natural gradient descent BSS algorithms have been derived from the TRINICON optimization criterion. The estimation of the multivariate pdfs in the update equations has been tackled by assuming spherically invariant random processes (SIRPs) which are known to be a good model for speech signals. This has considerably simplified the implementation of the update equations. Moreover, by using the multivariate Gaussian pdf as a special case of a SIRP pdf, efficient BSS algorithms solely based on second-order statistics have been obtained.

All these considerations have so far been carried out in the time domain. To allow efficient DFT-domain implementations, the generic broadband time-domain update equations have been formulated equivalently in the DFT domain. This equivalence has been achieved by using a matrix notation which allows to express the resulting Toeplitz matrices in terms of circulant matrices together with window matrices. Subsequently, the circulant matrices have been transformed to the DFT domain. The rigorous application of this procedure yielded equivalent broadband update equations expressed by DFT-domain quantities. Due to the broadband nature, several constraint matrices appear in the update equations which ensure a coupling between the DFT bins. This is in contrast to narrowband optimization where each DFT bin is considered independently and where then also the scaling and permutation ambiguities occur in each DFT bin independently.

The broadband DFT-domain formulation was the starting point to introduce selective narrowband approximations, i.e., to selectively discard some constraint matrices. This allows to compute, e.g., a matrix inverse efficiently by approximating it as a scalar inversion in each DFT bin. However, at the same time, several constraint matrices are retained which ensures that there is still some coupling between the DFT bins. Such hybrid algorithms still avoid that the BSS ambiguities appear independently in each DFT bin, but already offer a reduced computational complexity. This procedure of introducing selective approximations also allowed to establish several links to popular state-of-the-art BSS algorithms. These relationships support the claim that the TRINICON framework allows a unified view on convolutive BSS algorithms.

Up to this point, the framework presented in this work was based on the noiseless convolutive BSS model which only allows for point sources but does not account for (possibly diffuse) background noise. To ensure the applicability of the previously derived BSS algorithms also in noisy environments, either a noise-robust optimization criterion has to be found or the algorithms have to be complemented by pre- or post-processing approaches. As the former is difficult for realistic background noises, we focused on pre- and post-processing. The aim of these methods is twofold: First, they have to maintain

the separation performance and second, they have to suppress also the background noise which the BSS algorithms cannot cope with. It was shown that pre-processing methods are only of limited applicability due to the difficult task of developing a robust adaptation control and due to the fact that it is crucial to restore not only the magnitude spectra but also the phase of the noiseless mixture signals. In contrast, post-processing methods can achieve a better simultaneous suppression of background noise and residual cross-talk. A method based on single-channel postfiltering has been presented and several links to existing approaches have been shown. It has been demonstrated that this technique considerably improves the separation performance and also reduces the background noise.

This thesis also pointed out some starting points of future research work. In [Buc] also partitioned adaptive filtering is considered for the TRINICON framework. Partitioned filters allow to address the two contradicting requirements of a small block length demanded by the nonstationarity of the acoustic signals and a large demixing filter length needed to cover all reflections of the reverberant environments. Thus, further work may include the development of efficient BSS algorithms based on partitioned adaptive filtering. Moreover, it was shown in the context of postfiltering for acoustic echo cancellation in [EMV01, EMV02b, EMV02a] that the partitioning results in better estimation of the power spectral densities required for the postfilter. Thus, by using partitioning it can be expected that the good results for the BSS postfilter are even improved. Furthermore, the partitioning would also allow for low-delay implementations as desired e.g., for binaural hearing aid applications.

Another rewarding future research topic may be the development of other suitable approximations to efficiently compute the nonlinearity in the BSS algorithms based on higher-order statistics. Moreover, the examination of more robust nonlinearities, e.g. based on robust statistics [Hub81] are a promising research topic as has been shown recently for the instantaneous BSS in [CD06] and for broadband convolutive BSS in [Buc].

Last but not least, in certain applications such as video-/audio conferencing, the number of sensors may outnumber the number of active sources. In such a case subspace approaches, as briefly discussed in Section 4.1.2, are a promising but not yet fully explored technique to utilize the additional sensors. Moreover, if information about the array geometry is available, then adaptive beamforming approaches exploiting information-theoretic criteria rather than the conventional mean-square error-based criteria constitute another option. This would ideally allow to exploit all three signal properties nongaussianity, nonwhiteness and nonstationarity and the adaptation could be done without an adaptation control as usually required in adaptive beamforming. First approaches addressing this problem have been presented in [FP01b, PA02, KMG$^+$07, Buc].

A Operators for Block Matrices and Block-Sylvester Matrices

A.1 Operators generating diagonal and block-diagonal matrices

In this section operators for generating diagonal or block-diagonal matrices are defined. As illustrated in Fig. A.1, a diagonal matrix is obtained if all off-diagonals are set to zero by applying the operator diag{**A**} to the square matrix **A**. Similarly, we can generate a

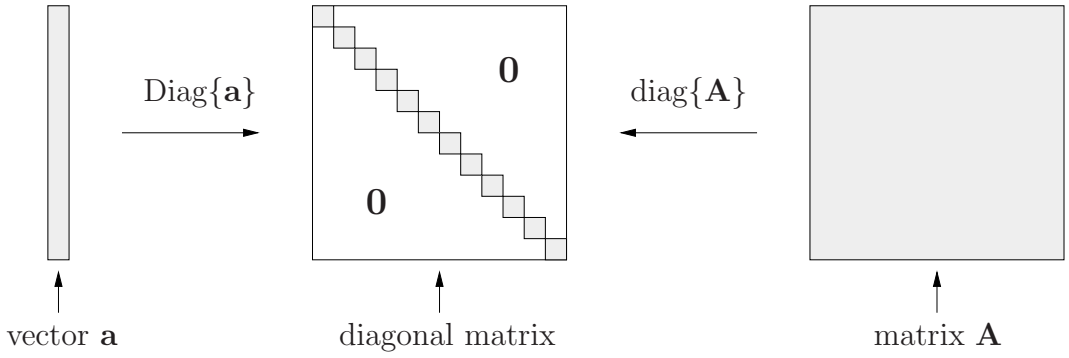


Figure A.1: Illustration of the diag and Diag operators.

diagonal square matrix from a given vector **a** by applying the operator Diag{**a**}.

In this thesis usually MIMO systems are considered and therefore, often block matrices occur which consist of several submatrices. These submatrices usually correspond to the different channels of the MIMO systems. To simplify the notation in this thesis, a counterparts to the diag and Diag operators are necessary, which take the block structure into account. Fig. A.2 illustrates a block matrix **A**, which exhibits P columns and P rows, consisting of the submatrices $\mathbf{A}_{pq}$. It should be noted that the submatrices are not required to be square matrices. As a special case, the submatrices may even correspond to row or column vectors. As depicted in Fig. A.2, the bdiag{**A**} operator yields a block-diagonal matrix by setting all off-diagonal submatrices $\mathbf{A}_{pq}$, $p \neq q$ of the block matrix **A** to zero. The block-diagonal matrix can also be generated by using the operator Bdiag$\{\mathbf{A}_{11}, \ldots, \mathbf{A}_{PP}\}$ leading to a block-diagonal matrix with the submatrices $\mathbf{A}_{pp}$, $p = 1, \ldots, P$ on its main diagonal.

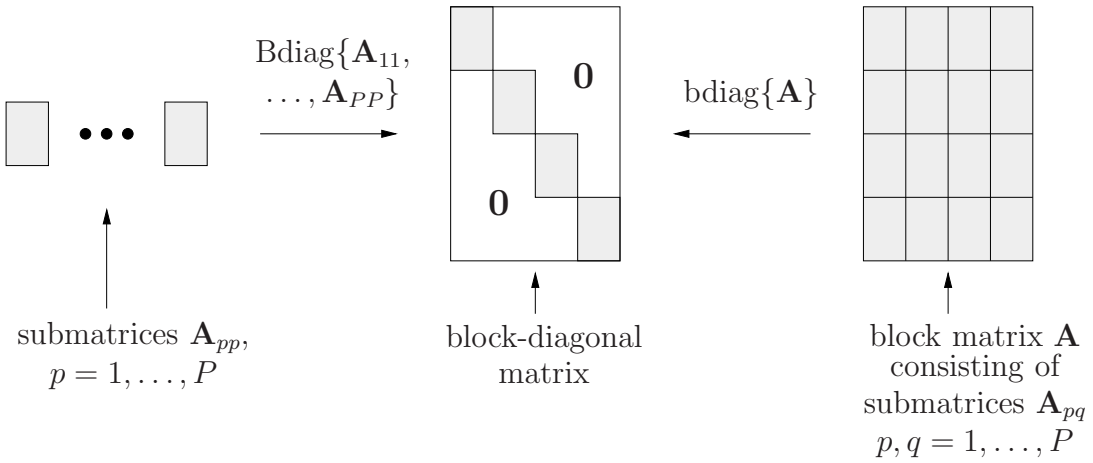


Figure A.2: Illustration of the bdiag and Bdiag operators for the case $P = 4$.

A.2 Block-determinant and block-adjoint operators

According to unpublished previous work [BA05] we will define in this section the block-determinant and the block-adjoint operators for block matrices consisting of column vectors which contain FIR filters. These operators are required for the derivation of the BSS optimum solution in Section 3.1. The block matrices are defined as a compact representation of a MIMO system consisting of FIR filters. In Section 3.1.1 this MIMO representation is used to model the acoustic impulse responses from the sources to the microphones which leads according to (3.3) to a mixing system given as the $QM \times P$ matrix

$$\check{\mathbf{H}} = \begin{bmatrix} \mathbf{h}_{11} & \cdots & \mathbf{h}_{1P} \\ \vdots & \ddots & \vdots \\ \mathbf{h}_{Q1} & \cdots & \mathbf{h}_{QP} \end{bmatrix}. \tag{A.1}$$

The column vectors $\mathbf{h}_{qp} = [h_{qp,0}, \ldots, h_{qp,M-1}]^{\mathrm{T}}$ contain the acoustic impulse responses modeled as FIR filters of length M. To allow a concatenation of MIMO systems containing FIR filters also a block-Sylvester matrix $\mathbf{H}_L$ is defined in (3.9) as

$$\mathbf{H}_L = \begin{bmatrix} \mathbf{H}_{11,L} & \cdots & \mathbf{H}_{1P,L} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{Q1,L} & \cdots & \mathbf{H}_{QP,L} \end{bmatrix}, \tag{A.2}$$

where each submatrix exhibits a Sylvester structure as defined in (3.8) and the subscript $_L$ denotes the width of each channel-wise Sylvester matrix. The advantage is that due to the Sylvester structure of the submatrices a convolution of two sequences can be written as a matrix-vector product. Thus, a convolution of two FIR MIMO systems can be written as the multiplication of a matrix in block-Sylvester structure (A.2) with a block matrix containing the individual FIR filters as column vectors (A.1). This representation is used

in (3.10) in Section 3.1.1 when determining the overall system matrix $\check{\mathbf{C}}$ resulting as a concatenation of mixing MIMO system $\mathbf{H}_L$ and demixing MIMO system $\check{\mathbf{W}}$.

In the following we define the block determinant and the block-adjoint operator for block matrices exemplarily for the mixing system $\check{\mathbf{H}}$ by using the two representations given in (A.1) and (A.2). It is assumed that the block matrices exhibit the same number of rows and columns of submatrices ($Q = P$), which means that they represent $P \times P$ MIMO systems containing P^2 FIR filters of a certain length.

Block-determinant operator $\mathrm{bdet}_P\{\cdot\}$

The block determinant $\mathrm{bdet}_P\{\cdot\}$ of a block matrix $\check{\mathbf{H}}$, containing the P^2 FIR mixing filters of length M, will be first discussed for the special case $Q = P = 2$. In general, the subscript $_P$ of the block determinant operator denotes the number of rows and columns of its argument, i.e., of the block matrix. In the case $Q = P = 2$, $\mathrm{bdet}_P\{\check{\mathbf{H}}\}$ is defined as

$$\mathrm{bdet}_2\{\check{\mathbf{H}}\} = \mathbf{H}_{11,M}\mathbf{h}_{22} - \mathbf{H}_{12,M}\mathbf{h}_{21}, \tag{A.3}$$

where, due to the Sylvester structure of the $2M-1 \times M$ matrices $\mathbf{H}_{11,M}$ and $\mathbf{H}_{12,M}$, each matrix-vector product denotes a linear convolution resulting in an FIR filter of length $2M-1$. Thus, the dimension of $\mathrm{bdet}_2\{\check{\mathbf{H}}\}$ is determined by the resulting FIR filter length as $2M-1 \times 1$. Equation (A.3) shows that the block determinant is defined analogously to the ordinary *determinant in linear algebra with the scalar multiplications replaced by* matrix-vector products which represent *linear convolutions*. As the convolution is commutative, the order of the FIR filters in the matrix-vector products can be interchanged, i.e, $\mathbf{H}_{11,M}\mathbf{h}_{22}$ is equal to $\mathbf{H}_{22,M}\mathbf{h}_{11}$ and thus, (A.3) may also be expressed as

$$\mathrm{bdet}_2\{\check{\mathbf{H}}\} = \mathbf{H}_{22,M}\mathbf{h}_{11} - \mathbf{H}_{21,M}\mathbf{h}_{12}. \tag{A.4}$$

Analogously to the ordinary determinant in linear algebra (see, e.g., [Har97]), the block determinant can be generalized straightforwardly for block matrices with $Q = P$ and $P, Q > 2$. In linear algebra determinants of square matrices larger than 2×2 are expressed by the cofactors of the matrix [Har97]. This procedure can be extended to block determinants as follows: First a submatrix $\check{\mathbf{H}}_{\mathrm{sub},ij}$ is obtained by removing the subvectors of the block matrix $\check{\mathbf{H}}$ in the i-th row and j-th column. The block determinant of this $(P-1)M \times (P-1)$ submatrix $\mathrm{bdet}_{P-1}\{\check{\mathbf{H}}_{\mathrm{sub},ij}\}$ is called the *block-minor* of the subvector $\mathbf{h}_{ij}$. The *block-cofactor* of $\mathbf{h}_{ij}$ is then given as the block-minor multiplied by the scalar factor $(-1)^{i+j}$. Using the definition of the block-cofactors the block determinant can in the general case be expressed as

$$\mathrm{bdet}_P\{\check{\mathbf{H}}\} = \sum_{j=1}^{P} (-1)^{i+j} \mathbf{H}_{ij,(P-1)(M-1)+1} \mathrm{bdet}_{P-1}\{\check{\mathbf{H}}_{\mathrm{sub},ij}\}. \tag{A.5}$$

In general, the size of $\mathrm{bdet}_P\{\check{\mathbf{H}}\}$ for $Q = P$ and $P, Q > 2$ is given as $P(M-1)+1 \times 1$ due to the linear convolutions performed inside the block determinant.

To illustrate the calculation of the block determinant we show the procedure exemplarily for the case $Q = P = 3$:

$$\begin{aligned}\mathrm{bdet}_3\{\check{\mathbf{H}}\} &= \mathbf{H}_{11,2M-1}\mathrm{bdet}_2\left\{\begin{bmatrix}\mathbf{h}_{22} & \mathbf{h}_{23}\\ \mathbf{h}_{32} & \mathbf{h}_{33}\end{bmatrix}\right\} - \mathbf{H}_{12,2M-1}\mathrm{bdet}_2\left\{\begin{bmatrix}\mathbf{h}_{21} & \mathbf{h}_{23}\\ \mathbf{h}_{31} & \mathbf{h}_{33}\end{bmatrix}\right\} + \\ &\quad + \mathbf{H}_{13,2M-1}\mathrm{bdet}_2\left\{\begin{bmatrix}\mathbf{h}_{21} & \mathbf{h}_{22}\\ \mathbf{h}_{31} & \mathbf{h}_{32}\end{bmatrix}\right\} \\ &= \mathbf{H}_{11,2M-1}(\mathbf{H}_{22,M}\mathbf{h}_{33} - \mathbf{H}_{23,M}\mathbf{h}_{32}) - \mathbf{H}_{12,2M-1}(\mathbf{H}_{21,M}\mathbf{h}_{33} - \mathbf{H}_{23,M}\mathbf{h}_{31}) + \\ &\quad + \mathbf{H}_{13,2M-1}(\mathbf{H}_{21,M}\mathbf{h}_{32} - \mathbf{H}_{22,M}\mathbf{h}_{31}) \end{aligned} \tag{A.6}$$

It should again be pointed out that in contrast to the regular determinant operator which yields a scalar value, the block determinant is a vector containing a sum of linear convolutions. This can be seen in (A.6) where each element of the sum consists of two linear convolutions which are written as a matrix-vector product by exploiting the Sylvester structure of the submatrices $\mathbf{H}_{qp,\ldots}$. As the mixing FIR filters contained in $\check{\mathbf{H}}$ all have the length M, the result of the block determinant for $Q = P = 3$ is a column vector of length $2M-1$.

Block-adjoint operator $\mathrm{badj}_P\{\cdot\}$

For any $P \times P$ matrix $\mathbf{A} = \{a_{ij}\}$, the $P \times P$ matrix whose ji-th element is the cofactor of a_{ij} is called the cofactor matrix of $\mathbf{A}$. The transpose of this matrix is called the *adjoint matrix* of $\mathbf{A}$ and is denoted by the symbol $\mathrm{adj}\{\mathbf{A}\}$ [Har97]. For block matrices containing $P \times P$ submatrices or subvectors we can analogously define a *block-adjoint* operator $\mathrm{badj}_P\{\cdot\}$. In the previous paragraph the block-cofactor of $\mathbf{h}_{ij}$ was defined as $(-1)^{i+j}\mathrm{bdet}_{P-1}\{\check{\mathbf{H}}_{\mathrm{sub},ij}\}$. Using the block-cofactors the block adjoint can be defined for the general case $Q = P$, $P, Q \geq 2$ as

$$\mathrm{badj}\{\check{\mathbf{H}}\}_P = \begin{bmatrix} (-1)^2\mathrm{bdet}_{P-1}\{\check{\mathbf{H}}_{\mathrm{sub},11}\} & \ldots & (-1)^{1+P}\mathrm{bdet}_{P-1}\{\check{\mathbf{H}}_{\mathrm{sub},1P}\} \\ \vdots & \ddots & \vdots \\ (-1)^{P+1}\mathrm{bdet}_{P-1}\{\check{\mathbf{H}}_{\mathrm{sub},P1}\} & \ldots & (-1)^{2P}\mathrm{bdet}_{P-1}\{\check{\mathbf{H}}_{\mathrm{sub},PP}\} \end{bmatrix}^{\mathrm{T}} \tag{A.7}$$

Due to the linear convolution inside the block determinant the dimensions of the block-adjoint $\mathrm{badj}_P\{\check{\mathbf{H}}\}$ are given as $P(P-1)(M-1)+1 \times P$. For the special case $Q = P = 2$ the block-adjoint for the block matrix $\check{\mathbf{H}}$ is given as

$$\mathrm{badj}_2\{\check{\mathbf{H}}\} = \begin{bmatrix} \mathbf{h}_{22} & -\mathbf{h}_{12} \\ -\mathbf{h}_{21} & \mathbf{h}_{11} \end{bmatrix}. \tag{A.8}$$

An important property of the adjoint of a $P \times P$ matrix $\mathbf{A}$ is [Har97]

$$\begin{aligned} \mathbf{A}\,\mathrm{adj}\{\mathbf{A}\} &= \det\{\mathbf{A}\}\mathbf{I}_{P\times P} \\ &= \mathrm{Diag}\left\{\det\{\mathbf{A}\},\ldots,\det\{\mathbf{A}\}\right\}, \end{aligned} \tag{A.9}$$

where the operator $\det\{\mathbf{A}\}$ denotes the determinant of $\mathbf{A}$ and $\mathbf{I}_{P\times P}$ is the $P \times P$ identity matrix.

For a block matrix $\mathbf{H}_{(P-1)(M-1)+1}$ consisting of P rows each containing P Sylvester submatrices of size $(P(M-1)+1)\times((P-1)(M-1)+1)$, this property can be expressed analogously as

$$\mathbf{H}_{(P-1)(M-1)+1}\,\mathrm{badj}_P\{\check{\mathbf{H}}\} = \mathrm{Bdiag}\left\{\mathrm{bdet}_P\{\check{\mathbf{H}}\},\ldots,\mathrm{bdet}_P\{\check{\mathbf{H}}\}\right\}, \tag{A.10}$$

which results in a block-diagonal matrix with the block determinants of $\check{\mathbf{H}}$ on its block-diagonal.

B Derivations

B.1 Derivation of the magnitude-squared coherence function for diffuse sound fields

The definition of the magnitude-squared coherence (MSC) of a diffuse sound field was first given in [CWB+55] but the derivation was only sketched in a few steps. The MSC of diffuse sound fields is used in many contexts (e.g., beamforming literature [BW01]) but a detailed derivation can hardly be found in any paper. Therefore, a complete derivation is given in this section following the steps outlined in [Mar95].

In Fig. 2.5 it was shown that the diffuse sound field can be modeled as statistically independent sound sources which are uniformly distributed on a sphere and emit monochromatic plane waves. The microphone array consisting of omnidirectional microphones is located in the spatial origin. The monochromatic plane wave (2.7) of the point source located at the area element $\tilde{A}$ and with frequency $\omega_0 = 2\pi f_0 f_\mathrm{s}^{-1}$ and amplitude $\hat{p} = 1$ generates a signal q_1 at the microphone x_1. After sampling with the sampling frequency f_s this signal is given for the time instant n and frequency ω_0 as

$$q_1(n, \omega_0, \varphi, \theta) = \cos(\omega_0 n + \phi(\varphi, \theta)) d\tilde{A}, \tag{B.1}$$

where $\phi(\varphi, \theta)$ is an arbitrary phase which depends on the direction of the incident wave and is uniformly distributed between 0 and 2π. To allow for aliasing-free reconstruction of the magnitude squared coherence of time-discrete signals only frequencies in the range $0 \leq \omega_0 < \pi$ are considered. The integration over all area elements $d\tilde{A}$ on the sphere leads to the microphone signal $x_1(n)$

$$\begin{aligned} x_1(n) &= \int_{\tilde{A}} \cos(\omega_0 n + \phi(\varphi, \theta)) d\tilde{A} && \text{(B.2)} \\ &= \int_0^{2\pi} \int_0^{\pi} \cos(\omega_0 n + \phi(\varphi, \theta)) \sin(\theta) d\theta d\varphi. && \text{(B.3)} \end{aligned}$$

The signal generated from the plane wave emitted at the area element $\tilde{A}$ at the second microphone is given analogously to (B.1) as

$$q_2(n, \omega_0, \varphi, \theta) = \cos\left(\omega_0 n + \phi(\varphi, \theta) - \frac{\omega_0 f_\mathrm{s} d \sin(\theta)}{c}\right) d\tilde{A}. \tag{B.4}$$

It should be noted that compared to (B.1) a phase difference has to be considered which depends on the distance of the microphones (see also Fig. 2.2). The second microphone signal $x_2(n)$ is calculated analogously by integrating over the source signals from all directions leading to

$$x_2(n) = \int_{\tilde{A}} \cos\left(\omega_0 n + \phi(\varphi,\theta) - \frac{\omega_0 f_s d \sin(\theta)}{c}\right) d\tilde{A} \tag{B.5}$$

$$= \int_0^{2\pi}\int_0^{\pi} \cos\left(\omega_0 n + \phi(\varphi,\theta) - \frac{\omega_0 f_s d \sin(\theta)}{c}\right) \sin(\theta) d\theta d\varphi. \tag{B.6}$$

The definition of the cross-correlation function of two signals $x(n)$ and $y(n)$ is given as

$$r_{xy}(\nu) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{i=-N}^{N} x(i+\nu) y(i) \tag{B.7}$$

and thus can determine the cross-correlation between the microphone signal $x_1(n)$ and the signal $q_2(n, \omega_0, \bar{\varphi}, \bar{\theta})$ arising from *one* area element on the sphere $d\bar{A}$ yielding

$$r_{x_1 q_2(\bar{\varphi},\bar{\theta})}(\nu) = \lim_{N\to\infty} \int_0^{2\pi}\int_0^{\pi} \frac{d\bar{A}}{2N+1} \sum_{n=-N}^{N} \cos(\omega_0 n + \omega_0 \nu + \phi(\varphi,\theta)) \cdot \cos\left(\omega_0 n + \phi(\bar{\varphi},\bar{\theta}) - \frac{\omega_0 f_s d \sin(\bar{\theta})}{c}\right) \sin(\theta) d\theta d\varphi. \tag{B.8}$$

Now the trigonometric addition formula

$$cos(x)cos(y) = \frac{1}{2}(\cos(x-y) + cos(x+y)) \tag{B.9}$$

is applied to (B.8). The model for the diffuse sound field assumes that the signals originating from different areas on the sphere exhibit a phase $\phi(\varphi,\theta)$ which is uniformly distributed from 0 to 2π. Thus, the cross-correlations between signals originating from different areas on the sphere disappear due to the averaging over these components and therefore, the cross-correlation is only for the coherent, i.e., for components arising from the area $\bar{A}$, not equal to zero [Kut00]. Hence, (B.8) together with (B.9) reduces to

$$r_{x_1 q_2(\bar{\varphi},\bar{\theta})}(\nu) = \frac{1}{2} \cos\left(\omega_0 \nu + \frac{\omega_0 f_s d \sin(\bar{\theta})}{c}\right) d\bar{A}. \tag{B.10}$$

To obtain the cross-correlation function of the two microphone signals $x_1(n)$ and $x_2(n)$ we have to integrate over all area elements on the sphere leading to

$$r_{x_1 x_2}(\nu) = \frac{1}{2} \int_{\tilde{A}} \cos\left(\omega_0 \nu + \frac{\omega_0 f_s d \sin(\bar{\theta})}{c}\right) d\tilde{A}. \tag{B.11}$$

The auto-correlation function of the microphone signal x_p, $p = 1, 2$ results from (B.11) as

$$r_{x_p x_p}(\nu) = \frac{1}{2} \int_{\tilde{A}} \cos\left(\omega_0 \nu\right) d\tilde{A}. \tag{B.12}$$

The power spectral densities which are necessary to determine the magnitude squared coherence function in (2.24) are given by

$$S_{x_p x_q} = \sum_{\nu=-\infty}^{\infty} r_{x_p x_q}(\nu) \exp(-j\omega\nu). \tag{B.13}$$

The transformation relation of a cosine with frequency $\omega_\circ$ is given as [OSB98]

$$\begin{aligned} &\sum_{\nu=-\infty}^{\infty} \cos(\omega_0\nu + \psi) \exp(-j\omega\nu)|_{\omega=\omega_0} \\ &= \pi \sum_{k=-\infty}^{\infty} (\exp(j\psi)\delta(\omega - \omega_0 + 2\pi k) + \exp(-j\psi)\delta(\omega + \omega_0 + 2\pi k))|_{\omega=\omega_0} \\ &= \pi \exp(j\psi). \end{aligned} \tag{B.14}$$

This yields an expression for the cross-power spectral density for the frequency range $0 \leq \omega < \pi$

$$\begin{aligned} S_{x_1 x_2}(\omega) &= \frac{1}{2} \int_0^{2\pi} \int_0^{\pi} \sum_{\nu=-\infty}^{\infty} \cos\left(\omega\nu + \frac{\omega f_\mathrm{s} d \sin(\theta)}{c}\right) \exp(-j\omega\nu) \sin(\theta) d\theta d\varphi \\ &= \frac{\pi}{2} \int_0^{2\pi} \int_0^{\pi} \exp\left(j\omega f_\mathrm{s} d \, \sin(\theta) c^{-1}\right) \sin(\theta) d\theta d\varphi \\ &= \pi^2 \int_0^{\pi} \exp\left(j\omega f_\mathrm{s} d \, \sin(\theta) c^{-1}\right) \sin(\theta) d\theta. \end{aligned} \tag{B.15}$$

The solution of the integral in (B.15) can be found in Section 3.915 in [GR65, p. 482] and thus (B.15) reduces to

$$S_{x_1 x_2}(\omega) = \frac{2\pi^2}{j\omega f_\mathrm{s} d \, c^{-1}} \sinh\left(j\omega f_\mathrm{s} d \, c^{-1}\right). \tag{B.16}$$

The hyperbolic sine function is defined for a complex argument as

$$\sinh(j \cdot a) = \frac{e^{ja} - e^{-ja}}{2} = \sin(a), \tag{B.17}$$

and therefore the cross-power spectral density (B.16) simplifies to

$$S_{x_1 x_2}(\omega) = \frac{2\pi^2}{j\omega f_\mathrm{s} d \, c^{-1}} \sin\left(\omega f_\mathrm{s} d \, c^{-1}\right). \tag{B.18}$$

The auto-power spectral density for the p-th microphone signal is given analogously to (B.15) as

$$\begin{aligned} S_{x_p x_p}(\omega) &= \pi^2 \int_0^{\pi} \sin(\theta) d\theta \\ &= 2\pi^2 \end{aligned} \tag{B.19}$$

Inserting (B.18) and (B.19) in the definition of the magnitude squared coherence (2.24) finally yields

$$|\Gamma_{x_1 x_2}(\omega)|^2 = \frac{\sin^2\left(\omega f_\mathrm{s} d \, c^{-1}\right)}{\left(\omega f_\mathrm{s} d \, c^{-1}\right)^2}. \tag{B.20}$$

B.2 Derivation of the gradient of the time-domain optimization criterion

We will present in this section a detailed derivation of the gradient $\nabla_{\mathbf{W}}\tilde{\mathcal{J}}(m,\mathbf{W})$ based on unpublished previous work [BA03]. To be able to calculate the gradient we will show in Section B.2.1 how the output signal pdf can be transformed by the block-Sylvester matrix $\mathbf{W}$ so that the cost function $\tilde{\mathcal{J}}(m,\mathbf{W})$ can be expressed in terms of $\mathbf{W}$. Subsequently, in Section B.2.2 the derivation of the gradient is given.

B.2.1 Transformation of the output signal pdf by a block-Sylvester matrix

In general, the joint statistics of a vector $\mathbf{y}$ of length K_y, which is related by a reversibly unambiguous and differentiable mapping $\mathbf{g}(\cdot)$ to a vector $\mathbf{x}$ of length K_x given as

$$\mathbf{y} = \mathbf{g}(\mathbf{x}), \tag{B.21}$$

can be determined in terms of the joint statistics of $\mathbf{x}$ [Pap02]. The mapping $\mathbf{g}(\mathbf{x})$ is a set of K_y functions $g_1(\mathbf{x}),\ldots,g_{K_\mathrm{y}}(\mathbf{x})$. For the case that both vectors are of equal length, i.e., $K_\mathrm{y} = K_\mathrm{x}$, the transformed joint pdf is given as

$$\hat{p}_{y,K_\mathrm{y}}(\mathbf{y}) = \frac{\hat{p}_{x,K_\mathrm{x}}(\mathbf{x})}{|J(\mathbf{x})|}, \tag{B.22}$$

where

$$J(\mathbf{x}) = \det\left\{\begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_{K_\mathrm{x}}} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_{K_\mathrm{y}}}{\partial x_1} & \cdots & \frac{\partial g_{K_\mathrm{y}}}{\partial x_{K_\mathrm{x}}} \end{bmatrix}\right\} \tag{B.23}$$

is the Jacobian of the transformation $\mathbf{g}(\mathbf{x})$.

If the length of the vector $\mathbf{y}$ is less than the length of $\mathbf{x}$, i.e., $K_\mathrm{y} < K_\mathrm{x}$, then we could first form the joint pdf $\hat{p}_{y\tilde{x},K_\mathrm{x}}(\mathbf{y},\tilde{\mathbf{x}})$ of the vector $\mathbf{y}$ and of $K_\mathrm{x}-K_\mathrm{y}$ elements $\tilde{\mathbf{x}}$ of the vector $\mathbf{x}$ so that both multivariate pdfs $\hat{p}_{y\tilde{x},K_\mathrm{x}}(\mathbf{y},\tilde{\mathbf{x}})$ and $\hat{p}_{x,K_\mathrm{x}}(\mathbf{x})$ have the same dimensionality. The transformation is then given as

$$\hat{p}_{y\tilde{x},K_\mathrm{x}}(\mathbf{y},\tilde{\mathbf{x}}) = \frac{\hat{p}_{x,K_\mathrm{x}}(\mathbf{x})}{|J(\mathbf{x})|}. \tag{B.24}$$

Subsequently, the joint pdf $\hat{p}_{y\tilde{x},K_\mathrm{x}}(\mathbf{y},\tilde{\mathbf{x}})$ has to be reduced to $\hat{p}_{y,K_\mathrm{y}}(\mathbf{y})$ by integration for $\tilde{\mathbf{x}}$ leading to [Pap02]

$$\hat{p}_{y,K_\mathrm{y}}(\mathbf{y}) = \frac{1}{|J(\mathbf{x})|}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\hat{p}_{x,K_\mathrm{x}}(\mathbf{x})\mathrm{d}\tilde{\mathbf{x}}. \tag{B.25}$$

These well-known relationships are now used to express the PD-dimensional joint output pdf $\hat{p}_{y,PD}(\mathbf{y}(n))$ in terms of the $2PL$-dimensional joint input pdf $\hat{p}_{x,2PL}(\mathbf{x}(n))$. The linear relation between the input and output signals was given in (3.33) as

$$\mathbf{y}(n) = \mathbf{W}^{\mathrm{T}}\mathbf{x}(n), \tag{B.26}$$

with $\mathbf{x}(n)$ and $\mathbf{y}(n)$ containing the stacked channel-wise input and output vectors given as

$$\mathbf{x}(n) = [\mathbf{x}_1^{\mathrm{T}}(n), \ldots, \mathbf{x}_P^{\mathrm{T}}(n)]^{\mathrm{T}}, \tag{B.27}$$

$$\mathbf{y}(n) = [\mathbf{y}_1^{\mathrm{T}}(n), \ldots, \mathbf{y}_P^{\mathrm{T}}(n)]^{\mathrm{T}}, \tag{B.28}$$

and the $2PL \times PD$ demixing matrix

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{11} & \cdots & \mathbf{W}_{1P} \\ \vdots & \ddots & \vdots \\ \mathbf{W}_{P1} & \cdots & \mathbf{W}_{PP} \end{bmatrix} \tag{B.29}$$

which is termed block-Sylvester since each $2L \times D$ submatrix $\mathbf{W}_{pq}$ exhibits a Sylvester structure. The transformation matrix $\mathbf{W}$ is not quadratic since $D \leq L$ and thus, (B.25) has to be applied. This requires first to extend the output signal vector to a length of $2PL$ by appending values of the input signal vector $\mathbf{x}(n)$. As we are dealing with MIMO systems, it is convenient to extend the vector $\mathbf{y}(n)$ in a channel-wise manner so that the linear relation (B.26) is modified to

$$[\mathbf{y}_1^{\mathrm{T}}(n), \tilde{\mathbf{x}}_1^{\mathrm{T}}(n), \ldots, \mathbf{y}_P^{\mathrm{T}}(n), \tilde{\mathbf{x}}_P^{\mathrm{T}}(n)]^{\mathrm{T}} = \tilde{\mathbf{W}}^{\mathrm{T}}\mathbf{x}(n), \tag{B.30}$$

where $\tilde{\mathbf{x}}_p(n)$ denotes the $2L-D$ last elements of the p-th input channel vector $\mathbf{x}_p(n)$. The new quadratic transformation matrix $\tilde{\mathbf{W}}$ of dimensions $2PL \times 2PL$ is given as

$$\tilde{\mathbf{W}} = \begin{bmatrix} \mathbf{W}_{11} & \begin{bmatrix} \mathbf{0}_{D\times 2L-D} \\ \mathbf{I}_{2L-D\times 2L-D} \end{bmatrix} & \ldots & \mathbf{W}_{1P} & \mathbf{0}_{2L\times 2L-D} \\ \vdots & & \ddots & & \vdots \\ \mathbf{W}_{P1} & \mathbf{0}_{2L\times 2L-D} & \ldots & \mathbf{W}_{PP} & \begin{bmatrix} \mathbf{0}_{D\times 2L-D} \\ \mathbf{I}_{2L-D\times 2L-D} \end{bmatrix} \end{bmatrix}. \tag{B.31}$$

Using the relationship for vectors of different lengths given in (B.25) leads to

$$\hat{p}_{y,PD}(\mathbf{y}(n)) = \frac{1}{|\det\{\tilde{\mathbf{W}}\}|} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \hat{p}_{x,2PL}(\mathbf{x}(n)) \mathrm{d}\tilde{\mathbf{x}}_1 \cdot \ldots \cdot \mathrm{d}\tilde{\mathbf{x}}_P. \tag{B.32}$$

Defining the column vector $\breve{\mathbf{x}}(n)$ of length PD obtained by removing the last $2L-D$ values of $\mathbf{x}(n)$ in each channel, we can write (B.32) as

$$\hat{p}_{y,PD}(\mathbf{y}(n)) = \frac{1}{|\det\{\tilde{\mathbf{W}}\}|} \hat{p}_{x,PD}(\breve{\mathbf{x}}(n)). \tag{B.33}$$

This equation can be further simplified if we use the fact that the value of a determinant does not change if any rows or columns are exchanged. Therefore, the matrix $\tilde{\mathbf{W}}$ can be reordered to result in a block-triangular matrix as shown in Fig. B.1. It is known that

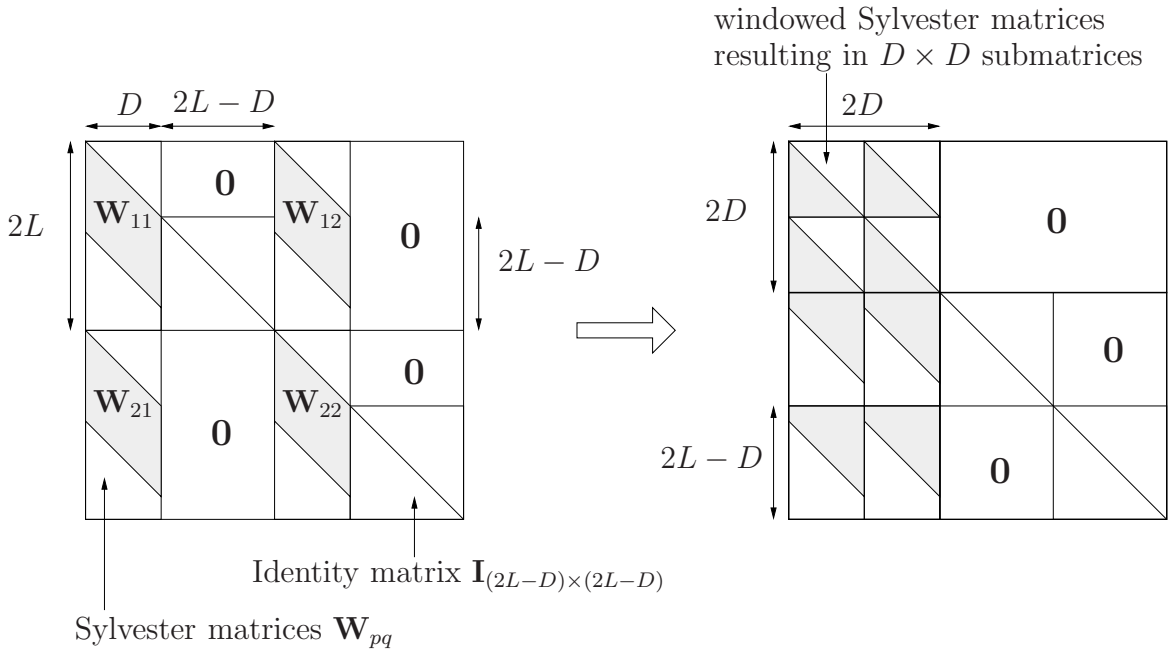


Figure B.1: Illustration of $\tilde{\mathbf{W}}$ defined in (B.31) shown for the case $P = 2$ and subsequent reordering of the columns and rows of $\tilde{\mathbf{W}}$ for simplified calculation of the determinant.

the determinant of a block-triangular matrix can be calculated as [Har97]

$$\det\left\{\begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & & \mathbf{0} \\ \vdots & \vdots & \ddots & \\ \mathbf{A}_{P1} & \mathbf{A}_{P2} & \dots & \mathbf{A}_{PP} \end{bmatrix}\right\} = \det\{\mathbf{A_{11}}\} \cdot \det\{\mathbf{A_{22}}\} \cdot \ldots \cdot \det\{\mathbf{A}_{PP}\}. \tag{B.34}$$

Thus, the determinant of the $2PL \times 2PL$ matrix $\tilde{\mathbf{W}}$ can be replaced by the determinant of the $PD \times PD$ matrix (upper left corner of the reordered matrix in the illustration in Fig. B.1). This $PD \times PD$ matrix can be obtained by constraining the size of the $2PL \times PD$ block-Sylvester matrix $\mathbf{W}$. This can be done by multiplying each channel of $\mathbf{W}$ with a window matrix defined in (3.57) as

$$\mathbf{W}_{2L\times D}^{1_D 0} = \left[\mathbf{I}_{D\times D}, \mathbf{0}_{D\times(2L-D)}\right]^{\mathrm{T}}. \tag{B.35}$$

A combination of all channels leads to a window matrix defined in (3.56) as

$$\mathbf{W}_{2PL\times PD}^{1_D 0} = \mathrm{Bdiag}\left\{\mathbf{W}_{2L\times D}^{1_D 0}, \ldots, \mathbf{W}_{2L\times D}^{1_D 0}\right\}. \tag{B.36}$$

Using this window matrix, (B.33) can finally be expressed as

$$\hat{p}_{y,PD}(\mathbf{y}(n)) = \frac{1}{\left|\det\left\{\left(\mathbf{W}_{2PL\times PD}^{1_D 0}\right)^{\mathrm{T}}\mathbf{W}\right\}\right|}\hat{p}_{x,PD}(\breve{\mathbf{x}}(n)). \tag{B.37}$$

B.2.2 Derivation of the gradient update

Using the result (B.37) of the previous section we can write the optimization criterion $\tilde{\mathcal{J}}(i,\mathbf{W})$ given in (3.43) as

$$\tilde{\mathcal{J}} = -\frac{1}{N}\sum_{j=0}^{N-1}\left\{\sum_{q=1}^{P}\log\left(\hat{p}_{y_q,D}(\mathbf{y}_q(iL+j))\right) + \log\left(\left|\det\left\{\left(\mathbf{W}_{2PL\times PD}^{1_D 0}\right)^{\mathrm{T}}\mathbf{W}\right\}\right|\right)\right\} + \text{const.} \tag{B.38}$$

where the constant corresponds to the estimate of the entropy of the input signals $\log(\hat{p}_{x,PD}(\breve{\mathbf{x}}))$. This quantity does not depend on the demixing system $\mathbf{W}$ and therefore, does not have to be taken into account for the optimization. To obtain the gradient $\nabla_{\mathbf{W}}\tilde{\mathcal{J}}$, we take the derivative of (B.38) with respect to the Sylvester matrix $\mathbf{W}$. The derivative of the first term in the braces is obtained by applying the chain rule, rather than transforming the pdfs. The derivative can be expressed element-wise with respect to $[\mathbf{W}_{rs}]_{kj}$ where the indices rs denote the channel-wise submatrix $\mathbf{W}_{rs}$ and $[\cdot]_{kj}$ are the element indices within the submatrices. Then, the chain rule can be expressed element-wise as

$$-\sum_q \frac{\partial \log \hat{p}_{y_q,D}(\mathbf{y}_q)}{\partial\,[\mathbf{W}_{rs}]_{kj}} = -\sum_q \frac{\partial \log \hat{p}_{y_q,D}(\mathbf{y}_q)}{\partial\,[\mathbf{y}_q]_g}\cdot\frac{\partial\,[\mathbf{y}_q]_g}{\partial\,[\mathbf{W}_{rs}]_{kj}}. \tag{B.39}$$

Writing the linear relation (B.26) between input and output signals element-wise yields

$$[\mathbf{y}_q]_g = \sum_{p,h}[\mathbf{W}_{pq}]_{hg}\,[\mathbf{x}_p]_h\,. \tag{B.40}$$

Inserting (B.40) into (B.39) leads to

$$\begin{aligned}
-\sum_q \frac{\partial \log \hat{p}_{y_q,D}(\mathbf{y}_q)}{\partial\,[\mathbf{W}_{rs}]_{kj}} &= -\sum_q \frac{\partial \log \hat{p}_{y_q,D}(\mathbf{y}_q)}{\partial\,[\mathbf{y}_q]_g}\cdot\sum_{p,h}\frac{\partial}{\partial\,[\mathbf{W}_{rs}]_{kj}}[\mathbf{W}_{pq}]_{hg}\,[\mathbf{x}_p]_h \\
&= -\sum_q \frac{\partial \log \hat{p}_{y_q,D}(\mathbf{y}_q)}{\partial\,[\mathbf{y}_q]_g}\cdot\sum_{p,h}[\mathbf{x}_p]_h\,\delta_{r,p}\delta_{s,q}\delta_{h,k}\delta_{g,j} \\
&= -\frac{\partial \log \hat{p}_{y_s,D}(\mathbf{y}_s)}{\partial\,[\mathbf{y}_s]_j}\,[\mathbf{x}_r]_k\,,
\end{aligned} \tag{B.41}$$

where the Kronecker delta $\delta_{a,b}$ was defined in (3.49). Combining (B.41) for all elements $r,s\in\{1,\ldots,P\}$, $j\in\{1,\ldots,D\}$, $k\in\{1,\ldots,2L\}$, and using matrix notation allows to finally write for the derivative of the first term

$$-\sum_q \frac{\log\left(\hat{p}_{y_q,D}(\mathbf{y}_q)\right)}{\partial\mathbf{W}} = \mathbf{x}\cdot\mathbf{\Phi}^{\mathrm{T}}(\mathbf{y}), \tag{B.42}$$

where $\mathbf{\Phi}(\mathbf{y})$ denotes a $PD \times 1$ column vector given as the concatenation of the score functions of each channel yielding

$$\begin{aligned}\mathbf{\Phi}(\mathbf{y}) &= \left[\left(-\frac{\partial \log \hat{p}_{y_1,D}(\mathbf{y}_1)}{\partial \mathbf{y}_1}\right)^{\mathrm{T}}, \ldots, \left(-\frac{\partial \log \hat{p}_{y_P,D}(\mathbf{y}_P)}{\partial \mathbf{y}_P}\right)^{\mathrm{T}}\right]^{\mathrm{T}} \\ &= \left[\left(-\frac{\frac{\partial \hat{p}_{y_1,D}(\mathbf{y}_1)}{\partial \mathbf{y}_1}}{\hat{p}_{y_1,D}(\mathbf{y}_1)}\right)^{\mathrm{T}}, \ldots, \left(-\frac{\frac{\partial \hat{p}_{y_P,D}(\mathbf{y}_P)}{\partial \mathbf{y}_P}}{\hat{p}_{y_P,D}(\mathbf{y}_P)}\right)^{\mathrm{T}}\right]^{\mathrm{T}}. \end{aligned} \tag{B.43}$$

Analogously, the derivative of the second term in the braces of (B.38) can be expressed by the chain rule as

$$\begin{aligned}-\frac{\partial}{\partial \mathbf{W}} \log\left(\left|\det\left\{\left(\mathbf{W}_{2PL\times PD}^{1_D 0}\right)^{\mathrm{T}} \mathbf{W}\right\}\right|\right) = -\frac{1}{\left|\det\left\{\left(\mathbf{W}_{2PL\times PD}^{1_D 0}\right)^{\mathrm{T}} \mathbf{W}\right\}\right|} \cdot \\ \cdot \frac{\partial\left|\det\left\{\left(\mathbf{W}_{2PL\times PD}^{1_D 0}\right)^{\mathrm{T}} \mathbf{W}\right\}\right|}{\partial \det\left\{\left(\mathbf{W}_{2PL\times PD}^{1_D 0}\right)^{\mathrm{T}} \mathbf{W}\right\}} \cdot \frac{\partial \det\left\{\left(\mathbf{W}_{2PL\times PD}^{1_D 0}\right)^{\mathrm{T}} \mathbf{W}\right\}}{\partial \mathbf{W}}.\end{aligned} \tag{B.44}$$

The derivative of the absolute value is given by the signum function $\mathrm{sign}(\cdot)$

$$\frac{\partial\left|\det\left\{\left(\mathbf{W}_{2PL\times PD}^{1_D 0}\right)^{\mathrm{T}} \mathbf{W}\right\}\right|}{\partial \det\left\{\left(\mathbf{W}_{2PL\times PD}^{1_D 0}\right)^{\mathrm{T}} \mathbf{W}\right\}} = \mathrm{sign}\left(\det\left\{\left(\mathbf{W}_{2PL\times PD}^{1_D 0}\right)^{\mathrm{T}} \mathbf{W}\right\}\right), \tag{B.45}$$

with the signum function $\mathrm{sign}(x)$ defined as

$$\mathrm{sign}(x) = \begin{cases} 0 & \text{for } x = 0 \\ \frac{x}{|x|} & \text{for } x \neq 0 \end{cases} \tag{B.46}$$

The remaining derivative in (B.44) can be computed by using the following matrix derivative (see e.g. [CA02]) given as

$$\frac{\partial \mathbf{AXB}}{\partial \mathbf{X}} = \det\{\mathbf{AXB}\}\, \mathbf{A}^{\mathrm{T}} (\mathbf{B}^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}})^{-1} \mathbf{B}^{\mathrm{T}}. \tag{B.47}$$

Thus, we obtain

$$\frac{\partial \det\left\{\left(\mathbf{W}_{2PL\times PD}^{1_D 0}\right)^{\mathrm{T}} \mathbf{W}\right\}}{\partial \mathbf{W}} = \det\left\{\left(\mathbf{W}_{2PL\times PD}^{1_D 0}\right)^{\mathrm{T}} \mathbf{W}\right\} \mathbf{W}_{2PL\times PD}^{1_D 0} \left(\mathbf{W}^{\mathrm{T}} \mathbf{W}_{2PL\times PD}^{1_D 0}\right)^{-1}. \tag{B.48}$$

Inserting (B.45) and (B.48) into (B.44) and noting that $\mathrm{sign}(x) \cdot x = |x|$ we finally obtain the derivative of the second term

$$-\frac{\partial}{\partial \mathbf{W}} \log\left(\left|\det\left\{\left(\mathbf{W}_{2PL\times PD}^{1_D 0}\right)^{\mathrm{T}} \mathbf{W}\right\}\right|\right) = -\mathbf{W}_{2PL\times PD}^{1_D 0} \left(\mathbf{W}^{\mathrm{T}} \mathbf{W}_{2PL\times PD}^{1_D 0}\right)^{-1}. \tag{B.49}$$

By using (B.42) and (B.49) we can express the gradient of the optimization criterion (B.38) with respect to the Sylvester matrix $\mathbf{W}$ as

$$\nabla_{\mathbf{W}}\tilde{\mathcal{J}}(i,\mathbf{W}) = \frac{1}{N}\sum_{j=0}^{N-1}\left\{\mathbf{x}(iL+j)\mathbf{\Phi}^{\mathrm{T}}(\mathbf{y}(iL+j)) - \mathbf{W}_{2LP\times DP}^{1D0}\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}_{2LP\times DP}^{1D0}\right)^{-1}\right\}. \tag{B.50}$$

B.3 Derivation of the block-online update

A numerical offline optimization for $\check{\mathbf{W}}$ is given for any choice of $\beta(i,m)$ by

$$\check{\mathbf{W}}^{\ell}(m) = \check{\mathbf{W}}^{\ell-1}(m) - \mu\Delta\check{\mathbf{W}}^{\ell}(m). \tag{B.51}$$

The update $\Delta\check{\mathbf{W}}^{\ell}(m)$ is for gradient adaptation set to $\nabla_{\check{\mathbf{W}}}\mathcal{J}(m,\mathbf{W}^{\ell-1}(m))$ and for natural gradient adaptation to $\nabla_{\check{\mathbf{W}}}^{\mathrm{NG}}\mathcal{J}(m,\mathbf{W}^{\ell-1}(m))$. As in this thesis only natural gradient algorithms are considered, we choose only natural gradient adaptation leading to

$$\begin{aligned}\check{\mathbf{W}}^{\ell}(m) &= \check{\mathbf{W}}^{\ell-1}(m) - \mu\nabla_{\check{\mathbf{W}}}^{\mathrm{NG}}\mathcal{J}(m,\mathbf{W}^{\ell-1}(m)) \\ &= \check{\mathbf{W}}^{\ell-1}(m) - \mu\sum_{i=0}^{\infty}\beta(i,m)\nabla_{\check{\mathbf{W}}}^{\mathrm{NG}}\tilde{\mathcal{J}}(i,\mathbf{W}^{\ell-1}(i))\end{aligned} \tag{B.52}$$

From (B.52) a recursive block-online algorithm can be derived by inserting the block-online weighting function given in (3.244) as

$$\beta(i,m) = \frac{1-\lambda}{K}\sum_{m'=0}^{m}\lambda^{m-m'}\epsilon_{m'K,m'K+K-1}(i),$$

yielding

$$\begin{aligned}\check{\mathbf{W}}^{\ell}(m) &= \check{\mathbf{W}}^{\ell-1}(m) - \mu\frac{1-\lambda}{K}\sum_{i=0}^{\infty}\sum_{m'=0}^{m}\lambda^{m-m'}\epsilon_{m'K,m'K+K-1}(i)\nabla_{\check{\mathbf{W}}}^{\mathrm{NG}}\tilde{\mathcal{J}}(i,\mathbf{W}^{\ell-1}(i)) \\ &= \check{\mathbf{W}}^{\ell-1}(m) - \mu(1-\lambda)\sum_{m'=0}^{m}\lambda^{m-m'}\underbrace{\frac{1}{K}\sum_{i=m'K}^{m'K+K-1}\nabla_{\check{\mathbf{W}}}^{\mathrm{NG}}\tilde{\mathcal{J}}(i,\mathbf{W}^{\ell-1}(i))}_{=:\nabla_{\check{\mathbf{W}}}^{\mathrm{NG}}\tilde{\mathcal{J}}_{\mathrm{K}}(m',\mathbf{W}^{\ell-1}(m'))}.\end{aligned} \tag{B.53}$$

The quantity $\nabla_{\check{\mathbf{W}}}^{\mathrm{NG}}\tilde{\mathcal{J}}_{\mathrm{K}}(m',\mathbf{W}^{\ell-1}(m'))$ contains an average over K update terms $\nabla_{\check{\mathbf{W}}}^{\mathrm{NG}}\tilde{\mathcal{J}}(i,\mathbf{W}^{\ell-1}(i))$. This simultaneous optimization for K blocks allows to exploit the nonstationarity of the source signals as for each block the source statistics change and thus new conditions are generated. The iterative offline update (B.51) can also be expressed in an explicit manner

$$\check{\mathbf{W}}^{\ell}(m) = \check{\mathbf{W}}^{0}(m) - \mu\sum_{k=1}^{\ell_{\max}}\Delta\check{\mathbf{W}}^{k}(m). \tag{B.54}$$

Expressing (B.53) analogously in an explicit manner leads to

$$\check{\mathbf{W}}^{\ell}(m) = \check{\mathbf{W}}^{0}(m) - \mu \sum_{k=1}^{\ell_{\max}} (1-\lambda) \sum_{m'=0}^{m} \lambda^{m-m'} \nabla_{\check{\mathbf{W}}}^{\mathrm{NG}} \tilde{\mathcal{J}}_{\mathrm{K}}(m', \mathbf{W}^{k-1}(m')). \tag{B.55}$$

It can be seen in (B.55) that for every block m a processing of all the data up to the block m is required. Analogous to the derivation of the RLS algorithm from the Newton-Raphson method in [SS89, p. 329] and [BBGK06] we need to introduce an approximation to produce an implementable recursive algorithm. Thus, analogously to [SS89, BBGK06] it is assumed that the Sylvester matrix $\mathbf{W}^{k-1}(m'-1)$, generated from the filter weights contained in $\check{\mathbf{W}}^{k-1}(m'-1)$, is the minimum point of $\nabla_{\mathbf{W}}^{\mathrm{NG}} \mathcal{J}_{\mathrm{K}}(m'-1, \mathbf{W}^{k-1}(m'-1))$ allowing the following approximation:

$$\nabla_{\check{\mathbf{W}}}^{\mathrm{NG}} \mathcal{J}_{\mathrm{K}}(m'-1, \mathbf{W}^{k-1}(m'-1)) = \mathbf{0} \tag{B.56}$$

Using (B.56) we can reduce (B.55) to

$$\check{\mathbf{W}}^{\ell}(m) = \check{\mathbf{W}}^{0}(m) - \mu \sum_{k=1}^{\ell_{\max}} (1-\lambda) \nabla_{\check{\mathbf{W}}}^{\mathrm{NG}} \mathcal{J}_{\mathrm{K}}(m, \mathbf{W}^{k-1}(m)). \tag{B.57}$$

Adding at the right-hand side of (B.57) the term $\lambda\check{\mathbf{W}}^{0}(m) - \lambda\check{\mathbf{W}}^{0}(m) = \mathbf{0}$ yields

$$\check{\mathbf{W}}^{\ell}(m) = \lambda\check{\mathbf{W}}^{0}(m) + (1-\lambda) \left(\check{\mathbf{W}}^{0}(m) - \mu \sum_{k=1}^{\ell_{\max}} \nabla_{\check{\mathbf{W}}}^{\mathrm{NG}} \mathcal{J}_{\mathrm{K}}(m, \mathbf{W}^{k-1}(m)) \right). \tag{B.58}$$

To obtain a block-online procedure allowing a combination of offline and online update, the algorithm will increase the block index m every $\ell_{\max}$-th iteration by one block. Therefore, we define $\check{\mathbf{W}}(m) := \check{\mathbf{W}}^{\ell}(m)$, $\check{\mathbf{W}}(m-1) := \check{\mathbf{W}}^{0}(m)$. Additionally, the offline update $\Delta\check{\mathbf{W}}(m)$ for the m-th block is defined as

$$\begin{aligned} \Delta\check{\mathbf{W}}(m) &:= \check{\mathbf{W}}^{0}(m) - \mu \sum_{k=1}^{\ell_{\max}} \nabla_{\check{\mathbf{W}}}^{\mathrm{NG}} \mathcal{J}_{\mathrm{K}}(m, \mathbf{W}^{k-1}(m)) \\ &= \check{\mathbf{W}}(m-1) - \mu \sum_{k=1}^{\ell_{\max}} \nabla_{\check{\mathbf{W}}}^{\mathrm{NG}} \mathcal{J}_{\mathrm{K}}(m, \mathbf{W}^{k-1}(m)). \end{aligned} \tag{B.59}$$

With these definitions (B.58) can be expressed as a recursive online update

$$\check{\mathbf{W}}(m) = \lambda\check{\mathbf{W}}(m-1) + (1-\lambda)\Delta\check{\mathbf{W}}(m). \tag{B.60}$$

The offline update (B.59) containing $\ell_{\max}$ iterations can also be written recursively for $\ell = 1, \ldots, \ell_{\max}$ as

$$\begin{aligned}\check{\mathbf{W}}^{\ell}(m) &= \check{\mathbf{W}}^{\ell-1}(m) - \mu\nabla^{\mathrm{NG}}_{\check{\mathbf{W}}}\mathcal{J}_{\mathrm{K}}(m, \mathbf{W}^{\ell-1}(m)) \\ &= \check{\mathbf{W}}^{\ell-1}(m) - \mu\frac{1}{K}\sum_{i=mK}^{mK+K-1}\nabla^{\mathrm{NG}}_{\check{\mathbf{W}}}\tilde{\mathcal{J}}(i, \mathbf{W}^{\ell-1}(i)). \end{aligned} \tag{B.61}$$

Thus, the final block-online procedure consists of an online part given by the recursive formulation (B.60) and an offline part given by the iterative procedure (B.61) which is computed for $\ell = 1, \ldots, \ell_{\max}$.

C Acoustic Environments Used in the Experiments

In this appendix we will describe the acoustic environments used for the evaluations presented in this thesis. The BSS algorithms derived from the generic framework presented in Chapter 3 were evaluated in several rooms with varying reverberation time. First, in Section C.1 the low reverberation chamber used for examining the different BSS algorithms is shown. Subsequently, more realistic environments such as a living room or lecture room scenario (Sections C.2 and C.3) are introduced. Additionally, setups with different types of background noise are needed to evaluate the noisy BSS algorithms presented in Chapter 4. Babble noise was simulated by a circular loudspeaker array placed inside the living room scenario (Section C.2). Moreover, recordings inside a car have been carried out to address the case of car noise (Section C.4).

C.1 Low reverberation chamber

In Sections 3.6.2 - 3.6.4 different aspects of the generic SOS natural gradient algorithm (3.112) were examined. Due to the high complexity of the generic algorithm a low reverberant room was desired to allow for a short demixing filter length and thus, keep the complexity at a moderate level. For this reason, the low reverberation chamber depicted in Fig. C.1 was chosen. The reverberation time of this room was determined from the energy decay curve of a measured impulse response according to Section 2.2.2 and is given as $T_{60} = 50\,\text{ms}$. An omnidirectional microphone pair with a spacing of $d = 20\,\text{cm}$ has been used for the measurements. Four different source positions with the angles $\theta = -70°, 45°, 70°, 90°$ have been simulated by loudspeakers and the acoustic impulse responses between sources and microphones have been measured. The source-sensor distance was chosen to 1 m.

C.2 Living room

A more realistic scenario is given by the room shown in Fig. C.2. This room is used as a multimedia room at the Chair of Multimedia Communications and Signal Processing at the University of Erlangen-Nuremberg. The dimensions are similar to a typical living room

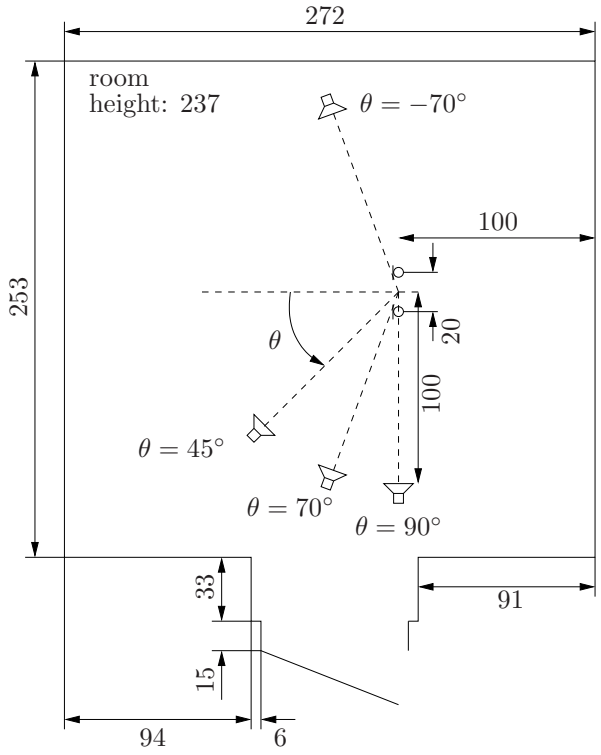


Figure C.1: Setup for the evaluations in the low reverberation chamber (lengths/distances in centimeters).

and due to the retractable curtains the reverberation time can be modified. In this thesis we use two conditions: all curtains closed and all curtains opened. For the former one the reverberation time was determined to be $T_{60} \approx 200\,\text{ms}$ and by opening the curtains, thereby exposing concrete walls and windows leading to hard reflections, the reverberation time increases to $T_{60} \approx 400\,\text{ms}$. An omnidirectional microphone pair with a spacing of 20 cm was placed in the center of a circular loudspeaker array with a radius of 1.5 m. The loudspeaker array comprises 48 loudspeakers and is usually used for wavefield synthesis research [STKR04]. In this thesis the loudspeaker array is used for the simulation of diffuse speech babble noise to evaluate the noisy BSS algorithms in Chapter 4. Therefore, the impulse responses from each loudspeaker to each microphone have been measured for both setups: curtains opened and closed. For the simulation of diffuse speech babble noise only every third loudspeaker is used as already 16 loudspeakers emitting 16 different speech signals produce a magnitude-squared coherence (MSC) function which is characteristic for a diffuse sound field. Moreover, the impulse responses to different point source positions have been measured for various angles $\theta = -20°, 0°, 20°, 40°, 80°$, at a distance of 1 m and in the case of $\theta = \pm 20°$ additionally for the source-sensor distances 2 m and 4 m.

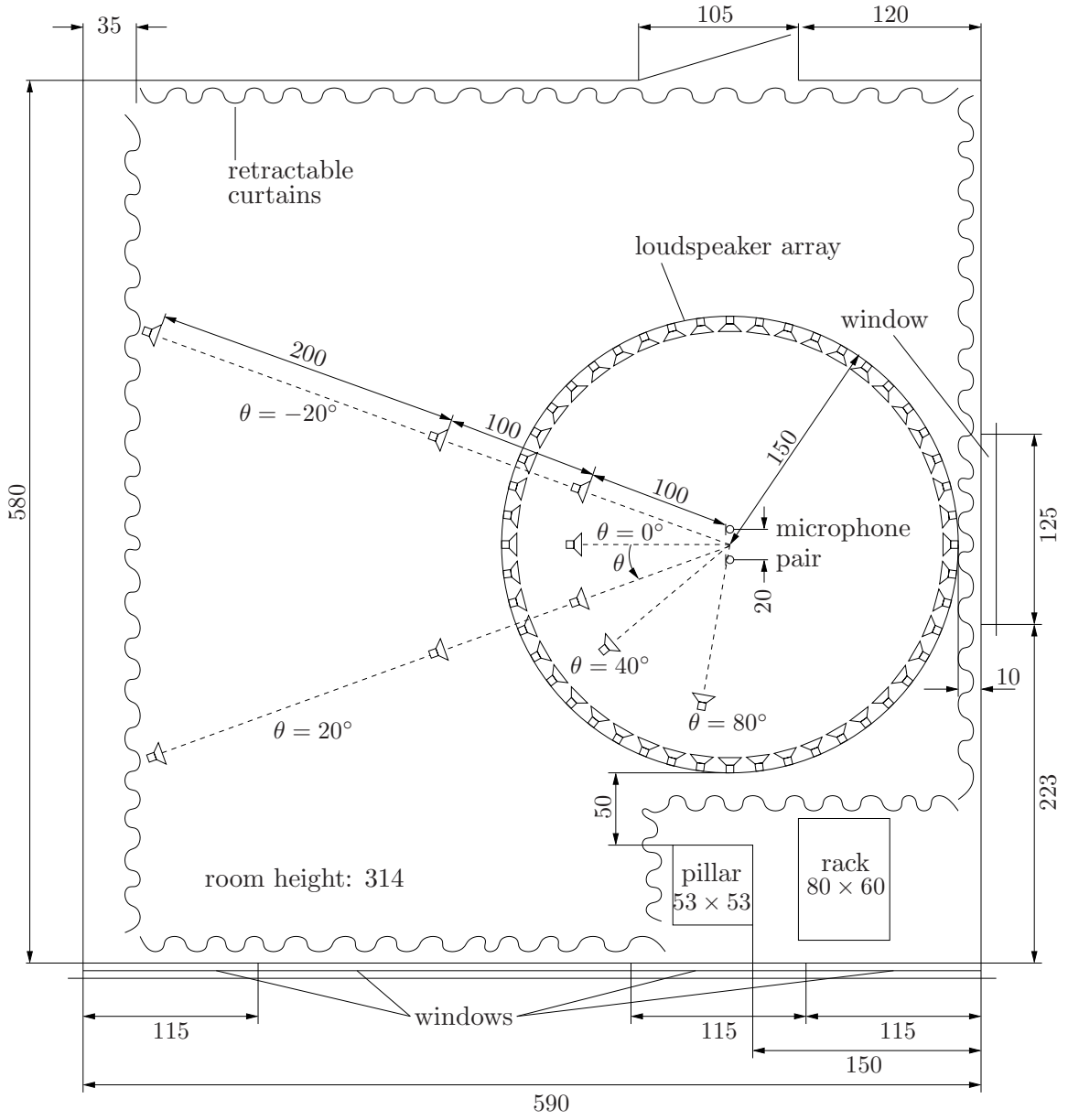


Figure C.2: Setup for the evaluations in a living room scenario (lengths/distances in centimeters).

C.3 Lecture room

To examine the effect of reverberation on the BSS algorithms in Section 3.6.6 also a very reverberant lecture room depicted in Fig. C.3 was used. The room exhibits a large reverberation time of $T_{60} \approx 850\,\text{ms}$. An omnidirectional microphone pair with a spacing of $d = 20\,\text{cm}$ has been used for the measurements. The impulse responses between the sensors and the point source positions have been measured for the angles $\theta = \pm 20°$ at a distance of 1 m.

C.4 Car environment

To evaluate the performance of BSS algorithms also in a car environment, a microphone pair with two omnidirectional microphones with a spacing of 20 cm was mounted at the interior mirror. The acoustic impulse responses between the sources and the sensors have been measured for the driver and co-driver positions. The reverberation time for these positions was determined as $T_{60} = 50\,\text{ms}$. Additionally, car background noise was recorded while driving through a suburban area at a speed of approximately 100 km/h.

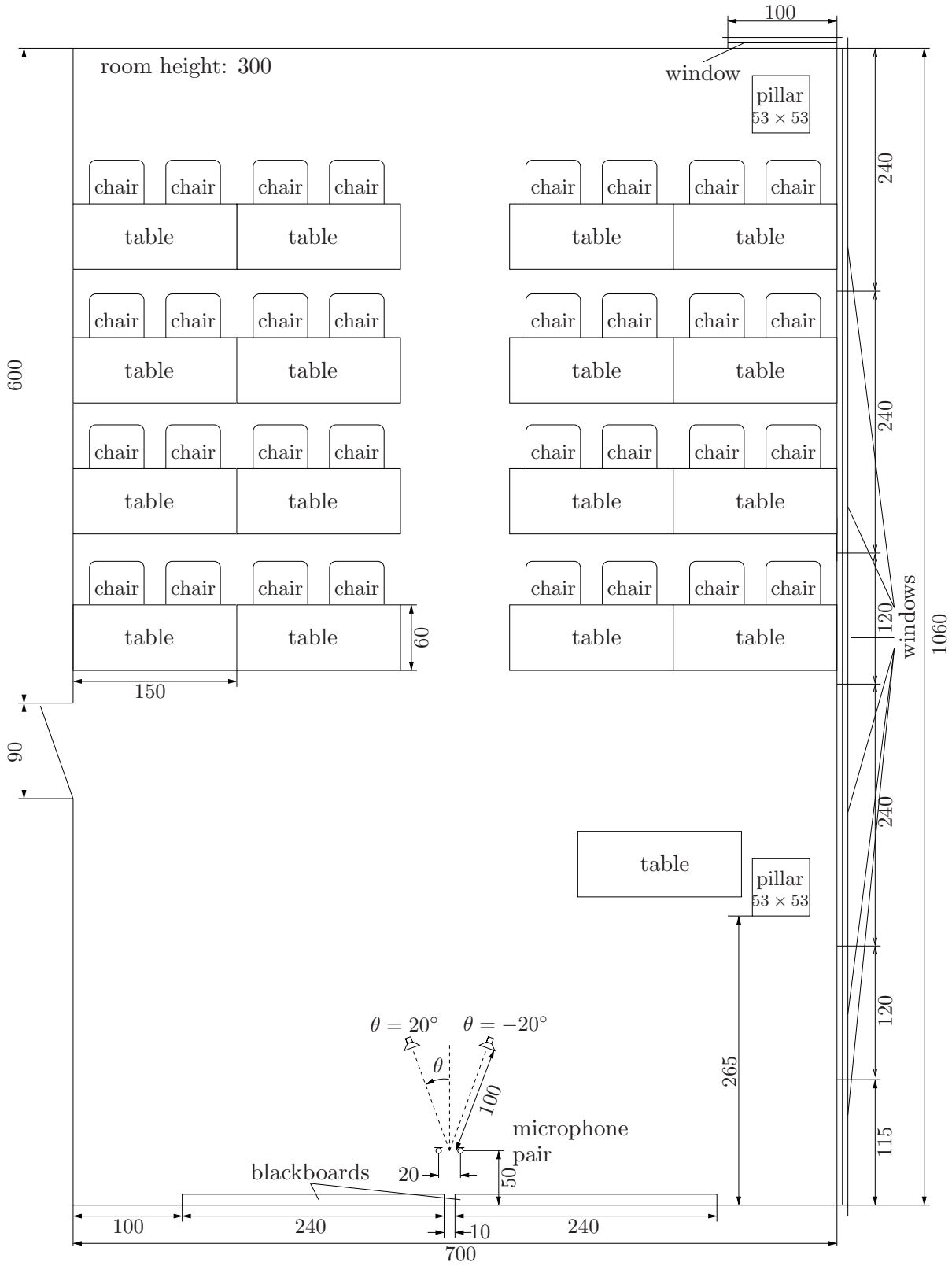


Figure C.3: Setup for the evaluations in a lecture room scenario (dimensioning in centimeters).

D Real-Time Implementation of Broadband BSS Algorithms

The work presented in this thesis has also led to a real-time implementation of two efficient broadband BSS algorithms introduced in Section 3.3 and evaluated experimentally in Section 3.6.5. Both real-time algorithms are based on second-order statistics and differ by the type of normalization. In the following, we will briefly describe each algorithm and give a summary using pseudo-code. The algorithms were implemented on a conventional PC system using the block-online adaptation procedure. In addition to the pseudo code an estimate of the computational complexity in terms of the number of arithmetic operations per sample will be given for different parameter settings.

D.1 BSS algorithm using a normalization based on diagonal matrices in the time domain

Algorithm summary

To achieve a fast convergence, both real-time algorithms have been implemented using the block-online adaptation procedure presented in in Section 3.5.3. This approach consists of an online part combined with an offline part. The online processing allows for a continuous processing of new samples and returns the current demixing filter coefficients $\check{\mathbf{W}}(m)$ according to

$$\check{\mathbf{W}}(m) = \lambda \check{\mathbf{W}}(m-1) + (1-\lambda)\check{\mathbf{W}}^{\ell_{\max}}(m), \tag{D.1}$$

where m denotes the block index and λ is the forgetting factor. The variable $\check{\mathbf{W}}^{\ell_{\max}}$ denotes the demixing filter weights obtained from the offline part after $\ell_{\max}$ iterations of the update rule

$$\check{\mathbf{W}}^{\ell}(m) = \check{\mathbf{W}}^{\ell-1}(m) - \mu \frac{1}{K} \sum_{i=mK}^{mK+K-1} \nabla^{\mathrm{NG}}_{\check{\mathbf{W}}} \tilde{\mathcal{J}}(i, \mathbf{W}^{\ell-1}(i)). \tag{D.2}$$

Both BSS algorithms are based on second-order statistics using the natural gradient

$$\nabla^{\mathrm{NG}}_{\check{\mathbf{W}}} \tilde{\mathcal{J}}(i, \mathbf{W}^{\ell-1}(i)) = \mathcal{SC}_{\mathcal{R}} \left\{ \mathbf{W}^{\ell-1}(i) \left(\tilde{\mathbf{R}}_{\mathbf{yy}}(i) - \operatorname{bdiag} \tilde{\mathbf{R}}_{\mathbf{yy}}(i) \right) \operatorname{bdiag}^{-1} \tilde{\mathbf{R}}_{\mathbf{yy}}(i) \right\}, \tag{D.3}$$

which was derived in Section 3.3.7.2. For increased versatility, the Sylvester constraint $\mathcal{SC}_{\mathcal{R}}$ is applied as it also allows for initializations leading to acausal filter taps as discussed in Section 3.3.6. The cross-correlation matrices are estimated using the computationally efficient correlation method given as

$$\tilde{\mathbf{R}}_{\mathbf{yy}}(i) = \frac{1}{N}\tilde{\mathbf{Y}}(i)\tilde{\mathbf{Y}}^{\mathrm{H}}(i), \tag{D.4}$$

with $\tilde{\mathbf{Y}}(i)$ defined in (3.82). By using the correlation method, all submatrices $\tilde{\mathbf{R}}_{\mathbf{y}_p\mathbf{y}_q}(i)$ exhibit Toeplitz structure

$$\tilde{\mathbf{R}}_{\mathbf{y}_p\mathbf{y}_q}(i) = \begin{bmatrix} \tilde{r}_{y_py_q}(i,0) & \cdots & \tilde{r}_{y_py_q}(i,D-1) \\ \tilde{r}_{y_py_q}(i,-1) & \ddots & \tilde{r}_{y_py_q}(i,D-2) \\ \vdots & \ddots & \vdots \\ \tilde{r}_{y_py_q}(i,-D+1) & \cdots & \tilde{r}_{y_py_q}(i,0) \end{bmatrix}, \tag{D.5}$$

where the elements are defined as

$$\tilde{r}_{y_py_q}(i,\tilde{v}) = \begin{cases} \dfrac{1}{N}\displaystyle\sum_{n=iL}^{iL+N-\tilde{v}-1} y_p(n+\tilde{v})y_q^*(n) & \text{for } \tilde{v} \geq 0 \\ \dfrac{1}{N}\displaystyle\sum_{n=iL-\tilde{v}}^{iL+N-1} y_p(n+\tilde{v})y_q^*(n) & \text{for } \tilde{v} < 0 \end{cases}, \tag{D.6}$$

and depend on the relative time-lag $\tilde{v} \in \{-D+1,\ldots,D-1\}$ of the signals $y_p(n)$ and $y_q(n)$.

The complexity of the inverses of the autocorrelation matrices $\tilde{\mathbf{R}}_{\mathbf{y}_q\mathbf{y}_q}$, $q = 1,\ldots,P$ contained in $\mathrm{bdiag}^{-1}\,\tilde{\mathbf{R}}_{\mathbf{yy}}$ in (D.3) prohibits a real-time implementation. Therefore, in Chapter 3 different approximations of the normalization were proposed. One possibility is to approximate the auto-correlation matrices by diagonal matrices, i.e., by the output signal variances $\sigma_{y_q}^2$ and additionally ensure the robustness by adding a regularization parameter δ_{y_q} leading to

$$\begin{aligned} \tilde{\mathbf{R}}_{\mathbf{y}_q\mathbf{y}_q}(i) &= \frac{1}{N}\sum_{n=iL}^{iL+N-1} y_q^2(n)\mathbf{I} + \delta_{y_q} \\ &= (\sigma_{y_q}^2(i) + \delta_{y_q})\mathbf{I}. \end{aligned} \tag{D.7}$$

The algorithm (D.3) with the normalization approximated by (D.7) together with the block-online update rule (D.1), (D.2) constitutes the first real-time algorithm. An experimental evaluation of this algorithm can be found in Section 3.6.5 where it was denoted as algorithm (C) in Table 3.1. The pseudo code of the implementation of this real-time algorithm is briefly summarized in Table D.1.

Table D.1: Pseudo code of the efficient block-online broadband BSS algorithm implementation approximating the normalization as a diagonal time-domain matrix. The pseudo code is exemplarily shown for the update $\Delta\mathbf{w}_{11}(m)$ for the case $P = Q = 2$ and the memory length $D = L$.

Online part:	
1.	Get $KL+N$ new samples $x_p(mKL/\alpha), \ldots, x_p((m+1)KL/\alpha + N - 1)$ of the sensors x_p, $p = 1, 2$ for each block $m = 0, 1, 2, \ldots$ and with the overlap factor α
Offline part:	
	Compute for each iteration $\ell = 1, \ldots, \ell_{\max}$:
2.	Compute output signals $y_q(mKL), ..., y_q((m+1)KL + N - L - 1)$, $q = 1, 2$ by convolving x_p with filter weights $\check{\mathbf{W}}^{\ell-1}$ from previous iteration
3.	Generate K blocks of N samples $[y_q(iL), \ldots, y_q(iL + N - 1)]$ with offline block index $i = mK, \ldots, mK + K - 1$ to exploit nonstationarity
	Compute for each block $i = mK, \ldots, mK + K - 1$:
4.	Calculate the signal energy of each block m $\sigma^2_{y_1}(i) = \tilde{r}_{y_1 y_1}(i, 0) = \sum_{n=iL}^{iL+N-1} y_1^2(n)$
5.	Compute cross-correlation matrix $\tilde{\mathbf{R}}_{\mathbf{y}_2\mathbf{y}_1}(i)$ by $\tilde{r}_{y_2 y_1}(i, \tilde{v})$ for $\tilde{v} = -L + 1, \ldots, L - 1$ according to (D.6)
6.	Normalize by elementwise division with regularized signal energy $\tilde{r}_{y_2 y_1}(i, \tilde{v})/(\sigma^2_{y_1}(i) + \delta_{y_1})$ for $\tilde{v} = -L + 1, \ldots, L - 1$
7.	Compute the matrix product $\mathbf{W}_{12}^{\ell-1}(m) \frac{\tilde{\mathbf{R}}_{\mathbf{y}_2\mathbf{y}_1}(i)}{\sigma^2_{y_1}(i)+\delta_{y_1}}$ as a convolution according to Fig. 3.4b (Sylvester constraint $\mathcal{SC}_{\mathcal{R}}$). Thus, the natural gradient w.r.t. each filter weight $w^{\ell}_{11,\kappa}$, $\kappa = 0, \ldots, L - 1$ is calculated as: $\nabla^{\mathrm{NG}}_{w_{11,\kappa}} \tilde{\mathcal{J}}(i, \mathbf{W}^{\ell-1}(i)) = \sum_{n=0}^{L-1} w^{\ell-1}_{12,n}(m) \tilde{r}_{y_2 y_1}(i, n - \kappa)/(\sigma^2_{y_1}(i) + \delta_{y_1})$
8.	Compute Steps 4-7 analogously for the other channels
9.	Compute update equation (D.2) for the offline part: $\check{\mathbf{W}}^{\ell}(m) = \check{\mathbf{W}}^{\ell-1}(m) - \mu \frac{1}{K} \sum_{i=mK}^{mK+K-1} \nabla^{\mathrm{NG}}_{\check{\mathbf{W}}} \tilde{\mathcal{J}}(i, \mathbf{W}^{\ell-1}(i))$
Online part:	
10.	Compute the recursive update (D.1) of the online part yielding the demixing filter $\check{\mathbf{W}}(m)$ used for separation: $\check{\mathbf{W}}(m) = \lambda \check{\mathbf{W}}(m-1) + (1-\lambda)\check{\mathbf{W}}^{\ell_{\max}}(m)$
11.	Use the demixing filters $\check{\mathbf{W}}(m)$ as the initial filters for the offline part $\check{\mathbf{W}}^0(m+1) = \check{\mathbf{W}}(m)$

Computational complexity

In the following we discuss the computational complexity of the algorithm summarized in Table D.1 in terms of arithmetic operations, i.e., the number of real multiplications and real additions. Divisions are usually counted as multiplications, assuming inverted constants and subtraction is addition by negated number. Thereby, each complex multiplication is realized by 4 real multiplications and 2 real additions and each complex addition is realized by 2 real additions. Moreover, the discrete Fourier transform of length R is computed using the FFT routine devised by [SJHB87] which requires $2R\log_2 R - \frac{3R}{2} - 4$ operations.

Table D.2 shows the computational complexity and in Fig. D.1 the complexity is illustrated as a function of filter length L and for $P = 2, 3, 4$ source and sensor signals. Additionally, the dependency of the complexity on the number K of simultaneously processed blocks to exploit the nonstationarity is illustrated by comparing $K = 4$ (solid) to $K = 8$ (dashed). The overlap factor of the online part (Step 1 in Table D.1) has been chosen as $\alpha = K$ to ensure a blockshift of L samples independent of the choice of K. The number of iterations and the block length have been chosen analogously to the experiments in Section 3.6.5 as $\ell_{\max} = 5, N = 2L$. The curves illustrate that, essentially, the complexity depends logarithmically on the filter length L, linearly on the number of blocks K, but quadratically on the number of channels P. For comparison it should be

Table D.2: Computational complexity for the block-online broadband SOS BSS algorithm implementation using the normalization based on diagonal matrices in the time domain.

	Arithmetic OPs for K blocks and P channels
Compute offline part for each iteration ℓ:	
Perform filtering of x_p with $\mathbf{W}^{\ell-1}$:	
FFT of demixing filter with FFT length $KL+N$	$P^2(2(KL+N)\log_2[KL+N] - 3(KL+N)/2 - 4)$
FFT of sensor signals x_p	$P(2(KL+N)\log_2[KL+N] - 3(KL+N)/2 - 4)$
Compute convolution in DFT domain	$(4P^2+P)(KL+N+2)$
IFFT to obtain time-domain signals y_q	$P(2(KL+N)\log_2[KL+N] - 3(KL+N)/2 - 4)$
Compute for each block $i = 1, \ldots, K$:	
Calculate scaling factor $\sigma^2_{y_q}(i)$ after (D.7)	$P(N + (K-1)(4L+2))$
Calculate cross-correlations $\tilde{r}_{y_p y_q}(i, \tilde{\nu})$ for $\tilde{\nu} = -L+1, \ldots, L-1$:	
FFT of output signals y_q with length $2N$	$PK(4N\log_2[2N] - 3N - 4)$
Compute cross-power spectral densities	$3(P^2-P)K(N+1)$
IFFT to obtain cross-correlations	$(P^2-P)K(4N\log_2[2N] - 3N - 4)/2$
normalize $\tilde{r}_{y_p y_q}(i, \tilde{\nu})$ using $\sigma^2_{y_q}(i)$	$2(P^2-P)KL$
Calculate matrix product as convolution:	
FFT of demixing filters $w_{pq,\kappa}$ of length $2L$	$P^2(4L\log_2[2L] - 3L - 4)$
FFT of cross-correlations of length $2L$	$(P^2-P)K(4L\log_2[2L] - 3L - 4)/2$
Compute convolution in DFT domain	$P^2(8P-10)K(L+1)$
IFFT	$P^2K(4L\log_2[2L] - 3L - 4)$
offline update rule (D.2)	$2P^2L + P^2(K-1)L$
online update rule (D.1)	$3P^2L$

noted that the well-known (single-channel) NLMS algorithm used in supervised adaptive filtering [Hay02] has a complexity of $4L + 7$ arithmetic operations. Thus, e.g., the complexity of the two-channel BSS algorithm for $K = 4$ and $L \approx 2000$ corresponds, according to Fig. D.1 approximately to that of a single-channel NLMS algorithm for the same filter length.

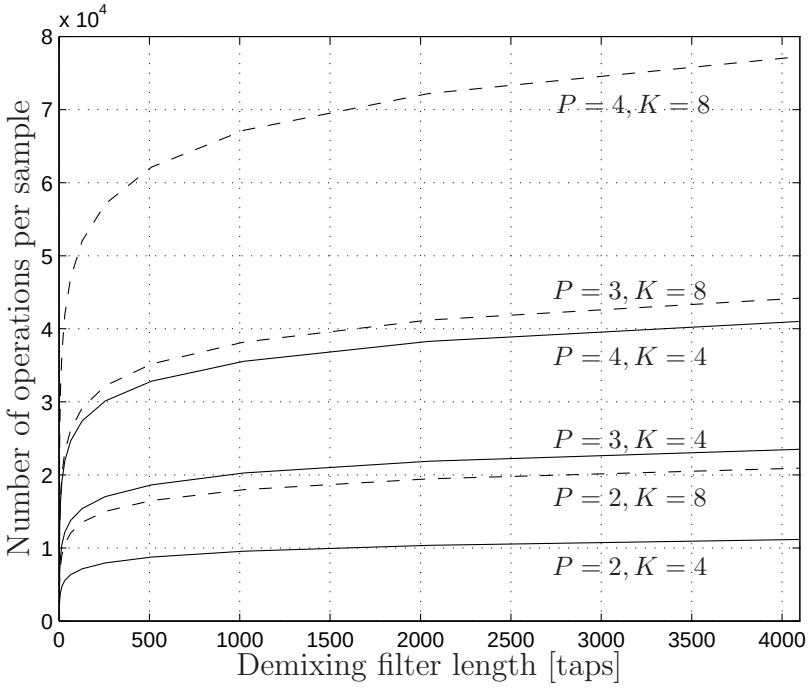


Figure D.1: Computational complexity for various filter lengths L and number of channels ($P = 2, 3, 4$) for simultaneous processing of $K = 4$ (solid) and $K = 8$ (dashed) blocks, respectively.

D.2 BSS algorithm using a narrowband normalization

Algorithm summary

In Section 3.4.3.1 a more sophisticated narrowband normalization was presented which was derived by expressing the BSS algorithms equivalently in the DFT domain and subsequently introducing suitable approximations. This allowed to express the inverse of the Toeplitz matrix $\tilde{\mathbf{R}}_{\mathbf{y}_q\mathbf{y}_q}$ as an inverse of a circulant matrix $\mathbf{C}_{\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q}$ leading to

$$\tilde{\mathbf{R}}^{-1}_{\mathbf{y}_q\mathbf{y}_q}(i) \approx N \cdot \mathbf{W}^{01_D}_{D\times R}\mathbf{C}^{-1}_{\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q}\mathbf{W}^{01_D}_{R\times D}, \tag{D.8}$$

with the circulant given as

$$\mathbf{C}_{\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q}(i) = \mathbf{F}_R^{-1}\underline{\tilde{\mathbf{Y}}}_q(i)\underline{\tilde{\mathbf{Y}}}_q^{\mathrm{H}}(i)\mathbf{F}_R. \tag{D.9}$$

The diagonal matrix $\underline{\tilde{\mathbf{Y}}}_p$ contains the DFT-domain values of the BSS output signal samples y_q and is defined as

$$\underline{\tilde{\mathbf{Y}}}_p(i) = \mathrm{Diag}\left\{\mathbf{F}_R[0,\ldots,0,y_p(iL+N-1),\ldots,y_q(iL),0,\ldots,0]^{\mathrm{T}}\right\}. \tag{D.10}$$

As was shown in Section 3.4.3.1, to obtain improved results in realistic environments a regularization term should be added to $\mathbf{C}_{\tilde{\mathbf{Y}}_q\tilde{\mathbf{Y}}_q}$ before inversion leading to

$$\breve{\mathbf{C}}_{\tilde{\mathbf{Y}}_p\tilde{\mathbf{Y}}_q}(i) = \mathbf{F}_R^{-1}\left(\rho\underline{\tilde{\mathbf{Y}}}_q(i)\underline{\tilde{\mathbf{Y}}}_q^{\mathrm{H}}(i) + (1-\rho)\sigma_{y_q}^2(i)\mathbf{I}\right)\mathbf{F}_R. \tag{D.11}$$

Inserting (D.11) in (D.8) finally yields the narrowband normalization

$$\tilde{\mathbf{R}}_{\mathbf{y}_q\mathbf{y}_q}^{-1}(i) \approx N\cdot\mathbf{W}_{D\times R}^{01_D}\mathbf{F}_R^{-1}\left(\rho\underline{\tilde{\mathbf{Y}}}_q(i)\underline{\tilde{\mathbf{Y}}}_q^{\mathrm{H}}(i) + (1-\rho)\sigma_{y_q}^2(i)\mathbf{I}\right)^{-1}\mathbf{F}_R\mathbf{W}_{R\times D}^{01_D}. \tag{D.12}$$

Thus, the second real-time implementation is given by (D.3) together with the normalization by (D.12) and the block-online update rule (D.1), (D.2). This algorithm corresponds to algorithm (B) in Table 3.1 in Section 3.6.5 where it was experimentally evaluated. The pseudo code of the algorithm is summarized in Table D.3.

Computational complexity

Compared to the previous algorithm, the complexity is increased due to several DFT and IDFT operations appearing due to the application of the improved normalization as summarized in Step 5-10 of Table D.3. The computational complexity of the algorithm using the narrowband normalization is described in Table D.4. Moreover, we illustrate again the dependency of the complexity on the number K of simultaneously processed blocks in Fig. D.2 by comparing $K = 4$ (solid) to $K = 8$ (dashed). The parameters $\alpha = K$, $\ell_{\max} = 5$, and $N = 2L$ have been chosen analogously to the previous section. The curves illustrate that, similar to the first real-time algorithm, the complexity depends again logarithmically on the filter length L, linearly on the number of blocks K, but quadratically on the number of channels P.

Table D.3: Pseudo code of the efficient block-online broadband BSS algorithm implementation using a narrowband normalization. The pseudo code is exemplarily shown for the update $\Delta\mathbf{w}_{11}(m)$ for the case $P = Q = 2$ and the memory length $D = L$.

Online part:	
1.	Get $KL + N$ new samples $x_p(mKL/\alpha), \ldots, x_p((m+1)KL/\alpha + N - 1)$ of the sensors x_p, $p = 1, 2$ for each block $m = 0, 1, 2, \ldots$ and with the overlap factor α
Offline part:	
	Compute for each iteration $\ell = 1, \ldots, \ell_{\max}$:
2.	Compute output signals $y_q(mKL), ..., y_q((m+1)KL + N - L - 1)$, $q = 1, 2$ by convolving x_p with filter weights $\check{\mathbf{W}}^{\ell-1}$ from previous iteration
3.	Generate K blocks of N samples $[y_q(iL), \ldots, y_q(iL + N - 1)]$ with offline block index $i = mK, \ldots, mK + K - 1$ to exploit nonstationarity
	Compute for each block $i = mK, \ldots, mK + K - 1$:
4.	Compute cross-correlation matrix $\tilde{\mathbf{R}}_{\mathbf{y}_2\mathbf{y}_1}(i)$ by $\tilde{r}_{y_2y_1}(i, \tilde{v})$ for $\tilde{v} = -L + 1, \ldots, L - 1$ according to (D.6)
5.	Calculate the values on the diagonal of $\underline{\tilde{\mathbf{Y}}}_1$ by computing the DFT of length R of the i-th output signal block of length N of Step 3
6.	Calculate the signal energy of each block i as $\sigma^2_{y_1}(i) = \sum_{n=iL}^{iL+N-1} y_1^2(n)$
7.	Calculate $\underline{\tilde{\mathbf{Y}}}_1^H \underline{\tilde{\mathbf{Y}}}_1$ by scalar multiplication in each DFT bin and perform narrowband regularization according to (D.11) by using the signal energy $\sigma^2_{y_1}$: $\underline{\mathbf{S}}_{\mathbf{y}_1\mathbf{y}_1}(i) = \rho\underline{\tilde{\mathbf{Y}}}_1(i)\underline{\tilde{\mathbf{Y}}}_1^H(i) + (1-\rho)\sigma^2_{y_1}(i)\mathbf{I}$
8.	Perform scalar inversion of the DFT-domain values on the main diagonal of $\underline{\mathbf{S}}_{\mathbf{y}_1\mathbf{y}_1}(i)$ as given in (D.12) and apply the inverse DFT to the resulting vector. As the inverse of a circulant yields again a circulant, the vector represents the first column of the circulant matrix $\mathbf{C}^{-1}_{\tilde{\mathbf{Y}}_1\tilde{\mathbf{Y}}_1}(i)$
9.	In (D.8) the circulant matrix $\mathbf{C}^{-1}_{\tilde{\mathbf{Y}}_1\tilde{\mathbf{Y}}_1}(i)$ is constrained to yield the approximation of the inverse of the Toeplitz matrix $\tilde{\mathbf{R}}^{-1}_{\mathbf{y}_1\mathbf{y}_1}(i)$. Matrix $\tilde{\mathbf{R}}^{-1}_{\mathbf{y}_1\mathbf{y}_1}(i)$ can be generated by picking the first L and last $L - 1$ values of the resulting vector from Step 8
10.	Compute the matrix product $\tilde{\mathbf{R}}_{\mathbf{y}_2\mathbf{y}_1}(i)\tilde{\mathbf{R}}^{-1}_{\mathbf{y}_1\mathbf{y}_1}(i)$ in (D.3) by fast convolution techniques exploiting the Toeplitz structure of both matrices. The result $\mathbf{A}_{\mathbf{y}_2\mathbf{y}_1}(i)$ of the matrix product may be approximated due to complexity reasons by calculating only the entries $[a(i,0), \ldots, a(i,-L+1)]$ in the first column and the entries $[a(i,0), \ldots, a(i,L-1)]$ in the first row and generate a Toeplitz structure from these values.
11.	Compute the matrix product $\mathbf{W}_{12}^{\ell-1}(m)\mathbf{A}_{\mathbf{y}_2\mathbf{y}_1}(i)$ as a convolution using Sylvester constraint $\mathcal{SC}_{\mathcal{R}}$. Thus, the natural gradient w.r.t. each filter weight $w^{\ell}_{11,\kappa}$, $\kappa = 0, \ldots, L-1$ is calculated as: $\nabla^{\mathrm{NG}}_{w_{11,\kappa}}\tilde{\mathcal{J}}(i, \mathbf{W}^{\ell-1}(i)) = \sum_{n=0}^{L-1} w^{\ell-1}_{12,n}(m)a(i, n-\kappa)$
14.	Compute Steps 4-11 analogously for the other channels
12.	Compute update equation (D.2) for the offline part: $\check{\mathbf{W}}^{\ell}(m) = \check{\mathbf{W}}^{\ell-1}(m) - \mu\frac{1}{K}\sum_{i=mK}^{mK+K-1}\nabla^{\mathrm{NG}}_{\check{\mathbf{W}}}\tilde{\mathcal{J}}(i, \mathbf{W}^{\ell-1}(i))$
Online part:	
13.	Compute the recursive update (D.1) of the online part yielding the demixing filter $\check{\mathbf{W}}(m)$ used for separation: $\check{\mathbf{W}}(m) = \lambda\check{\mathbf{W}}(m-1) + (1-\lambda)\check{\mathbf{W}}^{\ell_{\max}}(m)$
14.	Use demixing filters $\check{\mathbf{W}}(m)$ as initial filters for offline part: $\check{\mathbf{W}}^0(m+1) = \check{\mathbf{W}}(m)$

Table D.4: Computational complexity for the block-online broadband SOS BSS algorithm implementation using the narrowband normalization.

	Arithmetic OPs for K blocks and P channels
Compute offline part for each iteration ℓ:	
Perform filtering of x_p with $\mathbf{W}^{\ell-1}$:	
FFT of demixing filter with FFT length $KL+N$	$P^2(2(KL+N)\log_2[KL+N]-3(KL+N)/2-4)$
FFT of sensor signals x_p	$P(2(KL+N)\log_2[KL+N]-3(KL+N)/2-4)$
Compute convolution in DFT domain	$(4P^2+P)(KL+N+2)$
IFFT to obtain time-domain signals y_q	$P(2(KL+N)\log_2[KL+N]-3(KL+N)/2-4)$
Compute for each block $i=1,\ldots,K$:	
Calculate cross-correlations $\tilde{r}_{y_p y_q}(i,\tilde{v})$	
for $\tilde{v}=-L+1,\ldots,L-1$:	
FFT of output signals y_q with length $2N$	$PK(4N\log_2[2N]-3N-4)$
Compute cross-power spectral densities	$3(P^2-P)K(N+1)$
IFFT to obtain cross-correlations	$(P^2-P)K(4N\log_2[2N]-3N-4)/2$
Compute normalization term:	
Compute DFT values on the diagonal of $\underline{\tilde{\mathbf{Y}}}_q$	$PK(2R\log_2[R]-3R/2-4)$
Calculate regularization term $\sigma^2_{y_q}(i)$	$P(N+(K-1)(4L+2))$
Narrowband computations in first $\frac{R}{2}+1$ DFT bins:	
Calculate $\underline{\tilde{\mathbf{Y}}}_q\underline{\tilde{\mathbf{Y}}}_q^H$	$3PK(R/2+1)$
Compute regularization	$3PK(R/2+1)$
Compute scalar inversion	$PK(R/2+1)$
Compute IFFT of length R	$PK(2R\log_2[R]-3/2R-4)$
Compute normalization by fast convolution:	
FFT of cross-correlations with FFT length R	$(P^2-P)K(2R\log_2[R]-3R/2-4)/2$
FFT of normalization term with FFT length R	$PK(2R\log_2[R]-3R/2-4)$
Compute convolution in DFT domain	$6(P^2-P)K(R/2+1)$
IFFT to obtain normalized cross-correlation values	$(P^2-P)K(2R\log_2[R]-3R/2-4)$
Calculate matrix product of $\mathbf{W}^{\ell-1}_{pq}(m)$ with normalized cross-correlation as convolution:	
FFT of demixing filters $w_{pq,\kappa}$ of length $2L$	$P^2(4L\log_2[2L]-3L-4)$
FFT of $[a(i,-L+1),\ldots,a(i,L+1)]$ of length $2L$	$(P^2-P)K(4L\log_2[2L]-3L-4)/2$
Compute convolution in DFT domain	$P^2(8P-10)K(L+1)$
IFFT	$P^2K(4L\log_2[2L]-3L-4)$
offline update rule (D.2)	$2P^2L+P^2(K-1)L$
online update rule (D.1)	$3P^2L$

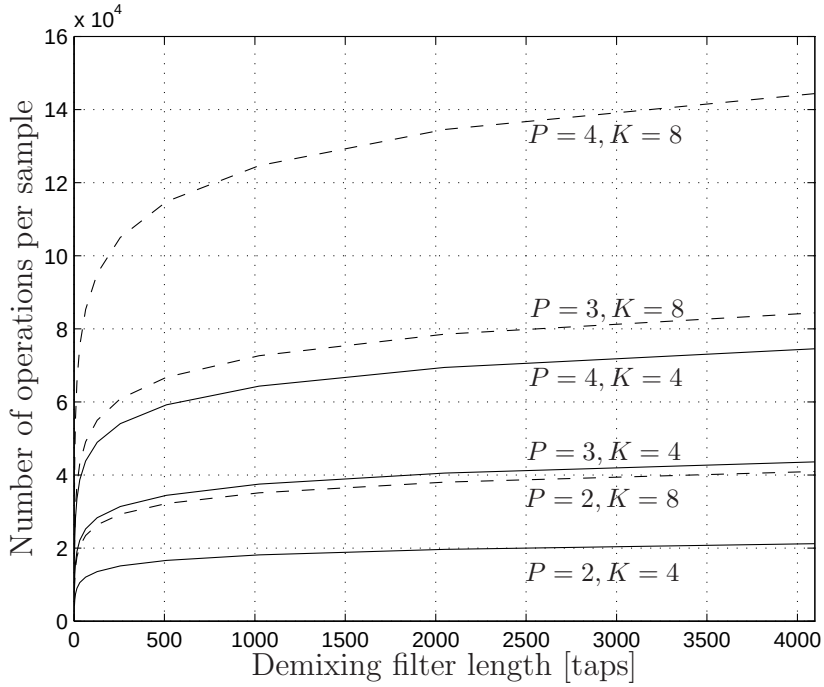


Figure D.2: Computational complexity for various filter lengths L and number of channels ($P = 2, 3, 4$) for simultaneous processing of $K = 4$ (solid) and $K = 8$ (dashed) blocks, respectively.

E Notations

E.1 Conventions

In this thesis we use lower case boldface for vectors describing the concatenation of several elements and upper case boldface denotes matrices. Vectors denoting directional quantities are instead denoted by an arrow. Underlined quantities $\underline{(\cdot)}$ represent vectors or matrices in the DFT domain. An underlined quantity with a superscript $\underline{(\cdot)}^{(\nu)}$ denotes the ν-th DFT bin of the corresponding DFT-domain quantity $\underline{(\cdot)}$ containing all DFT bins.

E.2 Abbreviations and Acronyms

ANC	adaptive noise canceller
BSS	blind source separation
CASA	computational auditory scene analysis
DFT	discrete Fourier transform
DTFT	discrete-time Fourier transform
FDAF	frequency-domain adaptive filtering
FFT	fast Fourier transform
FIR	finite impulse response
GSC	generalized sidelobe canceller
HOS	higher-order statistics
IDFT	inverse discrete Fourier transform
i.i.d.	independent identically distributed
IIR	infinite impulse response
ITU	International Telecommunication Union
LMS	least-mean-squares
MCBD	multichannel blind deconvolution
MCBPD	multichannel blind partial deconvolution
ME	maximum entropy
MIMO	multiple-input multiple-output
ML	maximum likelihood
MMI	minimum mutual information
MMSE	minimum mean-squared error

MSC	magnitude-squared coherence
NLMS	normalized least-mean-squares
PCA	principle component analysis
pdf	probability density function
RLS	recursive least-squares
SD	spectral distance
SIR	signal-to-interference ratio
SIRP	spherically invariant random process
SNR	signal-to-noise ratio
SRR	signal-to-reverberation ratio
SOS	second-order statistics
TRINICON	triple-N-independent component analysis for convolutive mixtures

E.3 Mathematical Symbols

Operators

$*$	convolution operator
$(\cdot)^{\mathrm{T}}$	transpose of $(\cdot)$
$(\cdot)^{*}$	conjugate complex of $(\cdot)$
$(\cdot)^{\mathrm{H}}$	hermitian, i.e., conjugate transpose of $(\cdot)$
$(\cdot)^{-1}$	inverse operation of $(\cdot)$
$\lvert\cdot\rvert$	absolute value of $(\cdot)$ (scalar)
$\lVert\cdot\rVert_{\mathrm{F}}$	Frobenius (or Euclidean) norm of $(\cdot)$ (vector or matrix)
$\mathrm{E}\{\cdot\}$	expectation operator
$\hat{\mathrm{E}}\{\cdot\}$	estimate of the expectation by a time-average
$\nabla_{\mathbf{W}}$	gradient with respect to $\mathbf{W}$
$\nabla_{\mathbf{W}}^{\mathrm{NG}}$	natural gradient with respect to $\mathbf{W}$
∇^2	Laplacian operator
$\partial/\partial x$	partial derivative with respect to x
$\delta_{i,j}$	Kronecker delta
$\mathrm{tr}\{\cdot\}$	trace of a matrix, i.e., sum of elements on main diagonal
$\mathrm{rank}\{\cdot\}$	rank of the argument
$\mathrm{diag}\{\mathbf{A}\}$	operator setting all off-diagonal values of matrix $\mathbf{A}$ to zero (see Appendix A)
$\mathrm{Diag}\{\mathbf{a}\}$	operator generating a square matrix with the elements of vector $\mathbf{a}$ on its main diagonal (see Appendix A)

$\mathrm{bdiag}\{\mathbf{A}\}$	operator setting all block-off-diagonal submatrices of matrix $\mathbf{A}$ to zero (see Appendix A)
$\mathrm{Bdiag}\{\mathbf{A}_{11}, \ldots, \mathbf{A}_{PP}\}$	operator generating a block-diagonal square matrix with the elements $\mathbf{A}_{11}, \ldots, \mathbf{A}_{PP}$ on its main block-diagonal (see Appendix A)
$\det\{\cdot\}$	determinant of a square matrix
$\mathrm{bdet}_P\{\cdot\}$	block-determinant of a partitioned matrix with P^2 submatrices (see Appendix A)
$\mathrm{badj}_P\{\cdot\}$	block-adjoint of a partitioned matrix with P^2 submatrices (see Appendix A)
$\mathcal{O}(\cdot)$	order
$\mathcal{SC}\{\cdot\}$	Sylvester constraint
$\mathcal{SC}_\mathcal{C}\{\cdot\}$	column Sylvester constraint
$\mathcal{SC}_\mathcal{R}\{\cdot\}$	row Sylvester constraint

Symbols

$\mathbf{0}_{\ldots\times\ldots}$	zero matrix with the size indicated in the subscript
α	overlap factor
α_q	arbitrary scaling factor due to the scaling ambiguity
$\alpha_{\mathrm{ab},i}$	absorbtion coefficient of the i-th wall
$\bar{\alpha}_{\mathrm{ab}}$	average of the individual absorbtion coefficients
$\beta(i,m)$	general weighting function
$\delta(n)$	unit impulse at position $n = 0$
δ_{y_q}	dynamical regularization for the q-th BSS output signal
$\delta_{\max}$	maximum value of the dynamical regularization
$\Delta\mathbf{C}$	update of the overall MIMO system matrix $\mathbf{C}$
$\Delta SIR_{\mathrm{seg},q}(m)$	segmental signal-to-interference ratio improvement for the q-th channel calculated at the m-th block of length N_S
$\Delta SIR_{\mathrm{seg}}(\mathrm{m})$	segmental signal-to-interference ratio improvement averaged over all channels calculated at the m-th block of length N_S
$\Delta\overline{SIR}_{\mathrm{seg},q}$	segmental signal-to-interference ratio improvement for the q-th channel calculated at the m-th block of length N_S
$\Delta\overline{SIR}_{\mathrm{seg}}$	segmental signal-to-interference ratio improvement averaged over all channels calculated at the m-th block of length N_S
$\Delta\overline{SNR}_{\mathrm{seg},q}$	segmental signal-to-noise ratio improvement for the q-th channel calculated at the m-th block of length N_S
$\Delta\overline{SNR}_{\mathrm{seg}}$	segmental signal-to-noise ratio improvement averaged over all channels calculated at the m-th block of length N_S
$\Delta\check{\mathbf{W}}$	update of the MIMO demixing matrix $\check{\mathbf{W}}$
$\Delta\underline{\check{\mathbf{W}}}$	update of the MIMO DFT-domain demixing matrix $\underline{\check{\mathbf{W}}}$
$\epsilon_{a,b}(m)$	rectangular window function

γ	forgetting factor for recursive estimation
γ_s	directivity factor of a source
γ_x	directivity factor of a sensor
$\lvert\Gamma_{x_p x_q}(\omega)\rvert^2$	magnitude squared coherence between the signals x_p and x_q at frequency ω
$\lvert\Gamma_{y_p y_q}(\omega)\rvert^2$	magnitude squared coherence between the signals y_p and y_q at frequency ω
$\lvert\underline{\Gamma}^{(\nu)}_{x_p x_q}(m)\rvert^2$	magnitude squared coherence between the signals x_p and x_q at the ν-th DFT bin and m-th block
$\lvert\bar{\underline{\Gamma}}^{(\nu)}_{x_p x_q}\rvert^2$	magnitude squared coherence between the signals x_p and x_q at the ν-th DFT bin averaged over K blocks
$\lvert\Gamma_{\mathbf{yy}}(\omega)\rvert^2$	generalized coherence function taking into account all BSS output channels
λ	forgetting factor for online or block-online adaptation
λ_0	wavelength of a monochromatic wave
$\mathbf{\Lambda}_\alpha$	diagonal matrix containing the arbitrary scaling factors for each BSS output channel
$\mathbf{\Lambda}_q(m)$	diagonal matrix containing the nonlinear weighting of $\mathbf{R}_{\mathbf{y}_p\phi(\mathbf{y}_q)}(m)$ based on a multivariate SIRP pdf
$\mathbf{\Lambda}(m)$	diagonal matrix containing all matrices $\mathbf{\Lambda}_q(m)$
$\underline{\mathbf{\Lambda}}_q(m)$	DFT-domain representation of the nonlinear weighting matrix $\mathbf{\Lambda}_q(m)$
$\underline{\mathbf{\Lambda}}(m)$	DFT-domain matrix of size $PR \times PR$ containing $\underline{\mathbf{\Lambda}}_q(m)$ on the block-diagonal
μ	stepsize
$\mu(m)$	time-dependent stepsize given for the m-th block
$\mu^{(\nu)}(m)$	time- and frequency-dependent stepsize given for the ν-th DFT bin and the m-th block
ν	DFT bin index
ω	continuous frequency
φ	elevation angle
$\phi_{y_q,D}(\cdot)$	SIRP score of the q-th BSS output channel
$\underline{\phi}_{y_q,1}(\cdot)$	DFT-domain univariate score of the q-th BSS output channel
$\Phi_{q,i}(\mathbf{y}_q)$	i-th element of the multivariate score function $\mathbf{\Phi}_q(\mathbf{y}_q)$ of dimension D for the q-th BSS output channel
$\mathbf{\Phi}_q(\mathbf{y}_q)$	multivariate score function of dimension D for the q-th BSS output channel
$\mathbf{\Phi}(\mathbf{y})$	multivariate score function consisting of the stacked channel-wise score functions $\mathbf{\Phi}_q(\mathbf{y}_q)$
ρ	factor for attenuating the offdiagonal values of a correlation matrix prior to inversion
$\sigma^2_{y_q}(m)$	variance of the q-th BSS output signal in the m-th block

$\sigma^{(\nu)^2}_{\underline{Y}_q}$	estimate of the variance of the q-th BSS output signal in the ν-th DFT bin
σ_0	parameter determining the decay of the dynamical regularization
θ	azimuthal angle of the point source measured counter-clockwise w.r.t. the axis perpendicular to the microphone array axis
Υ	saftey margin parameter for the adaptation control of the post-processing approach
$\underline{\Upsilon}^{(\nu)}_q(m)$	adaptive threshold for the adaptation control of the post-processing approach given for the m-th block and ν-th DFT bin
$\xi^{(\nu)}_q$	oversubtraction factor for the q-th channel and the ν-th DFT bin
A	area of all walls of the room
A_i	area of the i-th wall of the room
$A_{W_{pq}}$	gain of the filter $W_{pq}(z)$
$A_{H_{qp}}$	gain of the filter $H_{qp}(z)$
$\underline{B}^{(\nu)}_{p,q}$	coefficient modelling the residual crosstalk in the q-th BSS output channel based on the signal $\underline{\breve{Y}}^{(\nu)}_{p,q}$
$\mathbf{b}_{pq}$	vector containing the filter coefficients modelling the contribution of the p-th BSS output channel to the residual cross-talk $y_{\mathrm{c},q}$ in the q-th BSS output channel
$\underline{\mathbf{b}}_q$	DFT-domain vector containing the coefficients $\underline{B}^{(\nu)}_{p,q}$ for all $p = 1, \ldots, P$, $p \neq q$
c	sound velocity
$c_{qr,\kappa}$	κ-th FIR filter tap of the overall system FIR filter from the q-th source to the r-th output
$\mathbf{c}_{qr}$	vector of length $M+L-1$ containing the taps $c_{qr,\kappa}$ of the overall system FIR filter from the q-th source to the r-th output
C	number of samples for which temporal dependencies of the source signals exist
C_{80}	clarity index, a measure for subjective music perception in reverberant environments
$\breve{\mathbf{C}}$	MIMO overall system matrix of dimensions $Q(M+L-1) \times Q$ containing the overall system FIR filters $\mathbf{c}_{qr}$ of length $M+L-1$
$\mathbf{C}$	MIMO block-Sylvester overall system matrix of dimensions $Q(M+2L-1) \times QD$
$\mathbf{C}_{\mathbf{U}_p}$	circulant $R \times R$ matrix generated from the sensor signal matrix $\mathbf{U}_p$
$\mathbf{C}_{\mathbf{W}_{pq}}$	circulant $R \times R$ matrix generated from the MIMO demixing matrix $\mathbf{W}_{pq}$
$\mathbf{C}_{\mathbf{X}_p}$	circulant $R \times R$ matrix generated from the sensor signal matrix $\mathbf{X}_p$
$\mathbf{C}_{\mathbf{Y}_p}$	circulant $R \times R$ matrix generated from the output signal matrix $\mathbf{Y}_p$
$\mathbf{C}_{\tilde{\mathbf{Y}}_p}$	circulant $R \times R$ matrix generated from the output signal matrix $\tilde{\mathbf{Y}}_p$
$\mathbf{C}_{\mathbf{Y}_p\mathbf{Y}_p}$	circulant $R \times R$ matrix generated from the matrix product $\mathbf{Y}_p\mathbf{Y}_p^{\mathrm{H}}$
$\mathbf{C}_{\tilde{\mathbf{Y}}_p\tilde{\mathbf{Y}}_p}$	circulant $R \times R$ matrix generated from the matrix product $\tilde{\mathbf{Y}}_p\tilde{\mathbf{Y}}_p^{\mathrm{H}}$

d	distance between a pair of sensors
D	memory leading to D output signal samples concatenated in the vector $\mathbf{y}_q(n)$
D_{50}	"definition", a measure for subjective sound perception in reverberant environments
$E_{\text{decay}}(n)$	energy decay at discrete time n
$E(t)$	sound energy at continous time t
E_0	initial sound energy
f_{s}	sampling frequency
$f_{y_q,D}(\cdot)$	scalar function determined by the univariate pdf chosen for the multivariate SIRP pdf
$f_{\underline{Y}_q,D}(\cdot)$	scalar function determined by the univariate pdf chosen for narrowband BSS algorithms
$\mathbf{F}_R$	DFT matrix of size $R \times R$
$g_{q,\kappa}$	κ-th filter tap of the postfilter for the q-th BSS output channel
$\underline{G}_q^{(\nu)}(m)$	DFT-domain postfilter for the ν-th DFT bin and m-th block aiming at cancellation of residual crosstalk and background noise in the q-th BSS output signal
$\underline{G}_{\text{n},q}^{(\nu)}(m)$	DFT-domain postfilter for the ν-th DFT bin and m-th block aiming at cancellation of background noise in the q-th BSS output signal
$\mathbf{G}_{\ldots\times\ldots}^{\ldots}$	channel-wise constraint matrix consisting of the matrix product of DFT matrix, window matrix and IDFT matrix, with dimensions indicated in the subscript and the type of window matrix indicated in the superscript
$h_{qp,\kappa}$	κ-th FIR filter tap of the FIR filter from the q-th source to the p-th microphone
$H_{qp}(z)$	z-domain representation of the mixing FIR filter $\mathbf{h}_{qp}$
$\mathbf{h}_{qp}$	vector of length M containing the taps of the FIR filter from the q-th source to the p-th microphone
$\check{\mathbf{H}}$	MIMO mixing matrix of dimensions $QM \times P$ containing the FIR filters $\mathbf{h}_{qp}$ of length M
$\mathbf{H}_{qp,L}$	Sylvester mixing matrix of dimensions $M+L-1 \times L$ containing the M filter taps of the mixing filter $\mathbf{h}_{qp}$
$\mathbf{H}_L$	MIMO block-Sylvester mixing matrix of dimensions $Q(M+L-1) \times PL$ containing the submatrices $\mathbf{H}_{qp,L}$
$\mathbf{H}_{\text{sub},q}$	submatrix obtained by removing q-th row of submatrices from $\mathbf{H}_L$
$\mathbf{I}$	identity matrix
$I_{x_p x_q}^{(\nu)}(m)$	modified cross-periodogram between signals $x_p(n)$ and $x_q(n)$ for the ν-th DFT bin and m-th block
$\mathcal{J}(m,\mathbf{W})$	TRINICON BSS optimization criterion for the m-th block including a general weighting function
$\mathcal{J}^{(\nu)}(m,\underline{\mathbf{W}}^{(\nu)})$	narrowband optimization criterion for the ν-th DFT bin

$\mathcal{J}_{\text{gen}}(m, \mathbf{W})$	generalized TRINICON optimization criterion
$\tilde{\mathcal{J}}(i, \mathbf{W})$	optimization criterion for the i-th block without the general weighting function
$\vec{k}$	wavenumber vector
K	number of blocks
K_{S}	number of blocks for the estimation of segmental SIR and SNR
K_{sig}	number of blocks available for the entire signal using offline adaptation
K_ν	ν-th order modified Bessel function of the second kind
ℓ	superscript index denoting the iteration number for offline or block-online adapation
$\ell_{\max}$	maximum iteration number for block-online adapation
L	filter length of the demixing system FIR filters
L_{opt}	optimum BSS filter length
$\mathbf{L_I}$	block-diagonal matrix consisting of column vectors containing only ones
m	discrete-time block index
M	filter length of the mixing system FIR filters
n	discrete-time index
N	block length for the BSS algorithms
N_{S}	block length for the estimation of segmental SIR and SNR
N_{post}	block length for the postfiltering algorithm
$n_p(n)$	background noise signal at the p-th sensor as a function of the discrete time n
n_{50}	critical delay time for speech signals
n_{80}	critical delay time for music signals
$\mathbf{n}_p(n)$	vector containing $2L$ background noise signal samples at the p-th sensor
$\mathbf{n}(n)$	background noise signal vector of length $2PL$ containing all vectors $\mathbf{n}_p(n)$
$p(\vec{r}, t)$	propagating wave observed at position $\vec{p}$ at time t
$\hat{p}$	amplitude of the propagating wave
$p_{s_q,1}(\cdot)$	univariate pdf for the q-th source
$p_{s,Q}(\cdot)$	Q-dimensional joint pdf of all sources
$p_{s_q,C}(\cdot)$	multivariate pdf of the q-th source capturing time-dependencies of C samples
$p_{s,QC}(\cdot)$	joint pdf of dimension QC taking C samples of all Q source signals into account
$\hat{p}_{y_q,D}(\cdot)$	D-dimensional pdf for the q-th BSS output channel
$\hat{p}_{y,PD}(\cdot)$	PD-dimensional joint pdf over all BSS output channels
$\hat{\underline{p}}^{(\nu)}_{\underline{Y}_q,1}(\cdot)$	univariate pdf of the q-th BSS output channel in the ν-th DFT bin
P	number of microphones
Q	number of simultaneously active sources

$\vec{r}$	observation position vector
$\lvert\vec{r}_{qp}\rvert$	distance between q-th source and p-th microphone
r_h	critical distance
$r_{y_p y_q}(i,u,v)$	element in the u-th row and v-th column of the cross-correlation matrix $\mathbf{R}_{\mathbf{y}_p\mathbf{y}_q}$ estimated using the covariance method
$\tilde{r}_{y_p y_q}(i,\tilde{v})$	element for the $\tilde{v}$-th time-lag of the cross-correlation matrix $\tilde{\mathbf{R}}_{\mathbf{y}_p\mathbf{y}_q}$ estimated using the correlation method
R	DFT length for the BSS algorithms
R_{post}	DFT length for the postfiltering algorithm
$\mathbf{R}_{\mathbf{nn}}(m)$	cross-correlation matrix of size $2PL \times 2PL$ between all P background noise signals estimated using the covariance method
$\mathbf{R}_{\mathbf{ss}}(m)$	cross-correlation matrix of size $P(2L+M-1)\times P(2L+M-1)$ between all P source signals estimated using the covariance method
$\mathbf{R}_{\mathbf{y}_p\mathbf{y}_q}(m)$	cross-correlation matrix of size $D \times D$ between the p-th and q-th BSS output signals estimated using the covariance method
$\mathbf{R}_{\mathbf{yy}}(m)$	cross-correlation matrix of size $PD \times PD$ between all P BSS output signals estimated using the covariance method
$\tilde{\mathbf{R}}_{\mathbf{y}_p\mathbf{y}_q}(m)$	cross-correlation matrix of size $D \times D$ between the p-th and q-th BSS output signals estimated using the correlation method
$\tilde{\mathbf{R}}_{\mathbf{yy}}(m)$	cross-correlation matrix of size $PD \times PD$ between all P BSS output signals estimated using the correlation method
$\mathbf{R}_{\mathbf{x}\boldsymbol{\Phi}(\mathbf{y})}(m)$	higher-order statistics cross-relation matrix of size $2PL \times PD$ between the P microphone signals and the P BSS output signals
$\mathbf{R}_{\mathbf{y}_p\phi(\mathbf{y}_q)}(m)$	higher-order statistics cross-relation matrix of size $D \times D$ between the p-th and q-th BSS output signal based on multivariate SIRPs
$\mathbf{R}_{\mathbf{y}\phi(\mathbf{y})}(m)$	higher-order statistics cross-relation matrix of size $PD \times PD$ based on multivariate SIRPs
$\mathbf{R}_{\mathbf{y}_p\boldsymbol{\Phi}_q(\mathbf{y}_q)}(m)$	higher-order statistics cross-relation matrix of size $D \times D$ between the p-th and q-th BSS output signals
$\mathbf{R}_{\mathbf{y}\boldsymbol{\Phi}(\mathbf{y})}(m)$	higher-order statistics cross-relation matrix of size $PD \times PD$ between all P BSS output signals
$s_q(n)$	q-th source signal as a function of the discrete time n
$\underline{\mathbf{s}}^{(\nu)}_{\breve{\mathbf{y}}_q Y_{\mathrm{c},q}}(m)$	DFT-domain vector of size $P-1\times 1$ containing the power spectral densities between all channels contained in $\underline{\breve{\mathbf{y}}}^{(\nu)}_q$ and the residual crosstalk $\underline{Y}^{(\nu)}_{\mathrm{c},q}$ in the q-th BSS output channel
$\underline{\mathbf{s}}^{(\nu)}_{\breve{\mathbf{y}}_q Y_q}(m)$	DFT-domain vector of size $P-1\times 1$ containing the power spectral densities between all channels contained in $\underline{\breve{\mathbf{y}}}^{(\nu)}_q$ and the q-th BSS output signal $\underline{Y}^{(\nu)}_q$
$\underline{S}_{x_p x_q}(\omega)$	cross-power spectral density between signals x_p and x_q at frequency ω
$\underline{S}^{(\nu)}_{x_p x_q}(m)$	cross-power spectral density between signals x_p and x_q at ν-th DFT bin and m-th block

$\underline{S}^{(\nu)}_{y_p y_q}(m)$	ν-th element on the diagonal of $\underline{\mathbf{S}}_{\mathbf{y}_p\mathbf{y}_q}(m)$ denoting the cross-power spectral density between the p-th and q-th BSS output channels for the ν-th DFT bin
$\underline{\mathbf{S}}_{\mathbf{y}_p\mathbf{y}_q}(m)$	DFT-domain diagonal matrix of size $R\times R$ containing the DFT values of the cross-power spectral density between the p-th and q-th BSS output channels on the diagonal
$\underline{\mathbf{S}}_{\mathbf{yy}}(m)$	DFT-domain matrix of size $PR \times PR$ containing all matrices $\underline{\mathbf{S}}_{\mathbf{y}_p\mathbf{y}_q}(m)$
$\underline{\mathbf{S}}^{(\nu)}_{\breve{\mathbf{y}}_q\breve{\mathbf{y}}_q}(m)$	DFT-domain matrix of size $P-1\times P-1$ containing the power spectral densities between all channels contained in $\underline{\breve{\mathbf{y}}}^{(\nu)}_q$
$SD_{\mathrm{s}_r,q}$	unweighted log-spectral distance for the desired signal s_r at the q-th channel
$SIR_{\mathrm{seg},x_q}(m)$	segmental signal-to-interference ratio for the q-th sensor calculated at the m-th block of length N_S
$\overline{SIR}_{\mathrm{seg},x_q}$	segmental signal-to-interference ratio for the q-th sensor averaged over K_S blocks of length N_S
SIR_{y_q}	signal-to-interference ratio for the q-th BSS output channel
$SIR_{\mathrm{seg},y_q}(m)$	segmental signal-to-interference ratio for the q-th BSS output channel calculated at the m-th block of length N_S
$\overline{SIR}_{\mathrm{seg},y_q}$	segmental signal-to-interference ratio for the q-th BSS output averaged over K_S blocks of length N_S
$\widehat{\underline{SIR}}_q$	estimate of the signal-to-interference ratio for the q-th BSS output needed for the postfilter adaptation control
$\overline{SNR}_{\mathrm{seg},x_q}$	segmental signal-to-noise ratio for the q-th sensor averaged over K_S blocks of length N_S
$\overline{SNR}_{\mathrm{seg},y_q}$	segmental signal-to-noise ratio for the q-th BSS output averaged over K_S blocks of length N_S
SRR_{p,s_q}	signal-to-reverberation ratio for the source signal s_q at the p-th sensor
t	continous observation time
T_{60}	reverberation time
T_C	room temperature
T_s	sampling period
u_q	argument of the function $f_{y_q,D}(u_q)$
U	energy of the window function $w_f(n)$
$\mathbf{U}_p$	sensor signal matrix of size $N \times L$ in Toeplitz structure
V	volume of the room
$\mathbf{V}^\mathrm{H}_{\ldots\times\ldots}$	channel-wise transformation of a window matrix in the DFT domain with the size of the matrix indicated in the subscript
$w_f(n)$	window function as a function of the discrete time n
$w_{pq,\kappa}$	κ-th FIR filter tap of the FIR filter from the p-th sensor to the q-th output
$W_{pq}(z)$	z-domain representation of the demixing FIR filter $\mathbf{w}_{pq}$

$\mathbf{w}_{\mathrm{col},q}$	q-th column of the demixing FIR filter matrix $\check{\mathbf{W}}$
$\mathbf{w}_{pq}$	vector of length L containing the taps of the demixing FIR filter from the p-th microphone to the q-th output
$\underline{\mathbf{w}}_{pq}$	vector of length R containing the DFT-domain representation of the demixing FIR filter $\mathbf{w}_{pq}$
$\check{\mathbf{W}}$	MIMO demixing matrix of dimensions $PL \times Q$ containing all FIR filters $\mathbf{w}_{pq}$ of length L
$\check{\mathbf{W}}_{\mathrm{opt}}$	optimum MIMO demixing matrix of dimensions $PL \times Q$
$\mathbf{W}_{pq}$	Sylvester demixing matrix of dimensions $2L \times D$ containing the L filter taps of the demixing filter $\mathbf{w}_{pq}$ in each column
$\mathbf{W}$	MIMO demixing matrix of dimensions $2PL \times PD$ in block-Sylvester structure containing all Sylvester matrices $\mathbf{W}_{pq}$
$\mathbf{W}^{\cdots}_{\ldots\times\ldots}$	Window matrix with dimensions indicated in subscript and the position of ones and zeros indicated in the superscript
$\underline{\mathbf{W}}_{pq}$	DFT-domain diagonal demixing matrix of dimensions $R \times R$ containing the filter coefficients $\underline{\mathbf{w}}_{pq}$ on the main diagonal
$\underline{\mathbf{W}}$	DFT-domain MIMO demixing matrix of dimensions $PR \times PR$ containing all matrices $\underline{\mathbf{W}}_{pq}$
$\underline{\mathbf{W}}^{(\nu)}$	DFT-domain MIMO demixing matrix of dimensions $P \times P$ for the ν-th DFT bin
$\underline{\check{\mathbf{W}}}$	DFT-domain MIMO demixing matrix of dimensions $PR \times PQ$ containing all column vectors $\underline{\mathbf{w}}_{pq}$
$x_p(n)$	p-th sensor signal as a function of the discrete time n
$x_{\mathrm{s}_r,p}(n)$	desired signal component at the p-th sensor containing the desired source signal $s_r(n)$
$x_{\mathrm{c},p}(n)$	crosstalk signal at the p-th sensor as a function of the discrete time n
$\mathbf{x}_p(n)$	vector containing $2L$ samples from the p-th sensor
$\mathbf{x}(n)$	sensor signal vector of length $2PL$ containing all vectors $\mathbf{x}_p(n)$
$\mathbf{X}_p(m)$	Toeplitz matrix of of size $2L \times N$ containing the p-th sensor signal
$\mathbf{X}(m)$	Toeplitz matrix of of size $2PL \times N$ containing all matrices $\mathbf{X}_p$
$\underline{X}_p^{(\nu)}(m)$	DFT-domain representation of p-th sensor signal for the ν-th DFT bin and m-th block
$y_q(n)$	q-th BSS output signal as a function of the discrete time n
$y_{\mathrm{s}_r,q}(n)$	desired signal component at the q-th BSS output containing the desired source signal $s_r(n)$
$y_{\mathrm{c},q}(n)$	residual crosstalk signal at the q-th BSS output as a function of the discrete time n
$y_{\mathrm{n},q}(n)$	background noise signal at the q-th BSS output as a function of the discrete time n
$\breve{y}_{p,q}(n)$	p-th BSS output signal as a function of the discrete time n without the contribution of the desired source in the q-th BSS output channel

$\mathbf{y}_q(n)$	vector containing D output signal samples of the q-th BSS output channel
$\bar{\mathbf{y}}_q(n)$	vector containing N output signal samples of the q-th BSS output channel
$\mathbf{y}(n)$	output signal vector of length PD containing all vectors $\mathbf{y}_q(n)$
$\breve{\underline{\mathbf{y}}}_q^{(\nu)}(m)$	DFT-domain vector containing $\breve{\underline{Y}}_{p,q}^{(\nu)}(m)$ for all $p = 1, \ldots, P$, $p \neq q$
$\mathbf{Y}_q(m)$	output signal matrix of size $D \times N$ containing $N + D - 1$ output signal samples of the q-th BSS output channel
$\mathbf{Y}(m)$	output signal matrix of size $PD \times N$ containing all matrices $\mathbf{Y}_q(m)$
$\mathbf{Y}_\phi(m)$	matrix containing the channel-wise multiplication of $\mathbf{\Lambda}_q$ and $\underline{\mathbf{Y}}_q$
$\underline{Y}_q^{(\nu)}(m)$	ν-th element on the diagonal of $\underline{\mathbf{Y}}_q(m)$ denoting the DFT-domain representation of the q-th output signal for the ν-th DFT bin and m-th block
$\underline{Y}_{\mathrm{s},q}^{(\nu)}(m)$	DFT-domain representation of the desired source component of the q-th output signal for the ν-th DFT bin and m-th block
$\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)$	DFT-domain representation of the residual crosstalk component of the q-th output signal for the ν-th DFT bin and m-th block
$\underline{Y}_{\mathrm{n},q}^{(\nu)}(m)$	DFT-domain representation of the noise component of the q-th output signal for the ν-th DFT bin and m-th block
$\bar{\underline{Y}}_q^{(\nu)}(m)$	complementary BSS signal in the ν-th DFT bin at the m-th block
$\breve{\underline{Y}}_{p,q}^{(\nu)}(m)$	DFT-domain representation of the p-th output signal $\breve{y}_{p,q}$ for the ν-th DFT bin and m-th block
$\underline{\mathbf{Y}}_q(m)$	DFT-domain diagonal output signal matrix of dimensions $R \times R$ containing the DFT-domain representation of $N + D - 1$ output signal samples on the main diagonal
$\underline{\mathbf{Y}}_{\phi,q}(m)$	DFT-domain matrix of dimensions $R \times R$ containing the DFT-domain representation of the q-th nonlinearly weighted BSS output signal
$\underline{\mathbf{Y}}(m)$	DFT-domain output signal matrix of size $PR \times R$ containing all matrices $\underline{\mathbf{Y}}_q(m)$
$\underline{\mathbf{Y}}_\phi(m)$	DFT-domain matrix of dimensions $PR \times R$ containing all matrices $\underline{\mathbf{Y}}_{\phi,q}(m)$
$\underline{\mathbf{Y}}^{(\nu)}(m)$	DFT-domain output signal vector of dimensions $P \times 1$ for the ν-th DFT bin
$\tilde{\mathbf{Y}}_q(m)$	output signal matrix of size $D \times N + D - 1$ containing N output signal samples of the q-th BSS output channel
$\tilde{\mathbf{Y}}(m)$	output signal matrix of size $PD \times N + D - 1$ containing all matrices $\tilde{\mathbf{Y}}_q(m)$
$\tilde{\underline{\mathbf{Y}}}_q(m)$	DFT-domain diagonal output signal matrix of dimensions $R \times R$ containing the DFT-domain representation of N output signal samples on the main diagonal
$\tilde{\underline{\mathbf{Y}}}(m)$	DFT-domain output signal matrix of size $PR \times R$ containing all matrices $\tilde{\underline{\mathbf{Y}}}_q(m)$

$z_q(n)$	output signal of the post-processing algorithm in the q-th channel as a function of the discrete time n
$\underline{Z}_q^{(\nu)}(m)$	DFT-domain representation of the q-th output signal of the post-processing scheme for the ν-th DFT bin and m-th block
$z_{0H_{qp},\nu}$	ν-th zero of the filter $H_{qp}(z)$
$z_{0W_{pq},\nu}$	ν-th zero of the filter $W_{pq}(z)$

F Titel, Inhaltsverzeichnis, Einleitung und Zusammenfassung

The following german translations of the title (Section F.1), the table of contents (Section F.2), the introduction (Section F.3), and the summary (Section F.4) are a mandatory requirement for the dissertation at the Faculty of Engineering of the University of Erlangen-Nuremberg.

F.1 Titel

Akustische blinde Quellentrennung in verhallten und störbehafteten Umgebungen.

F.2 Inhaltsverzeichnis

F.3 Einleitung

In den letzten Jahren wurden auf dem Gebiet der akustischen Mensch-Maschine Schnittstelle sowohl in der Grundlagenforschung als auch in der Produktentwicklung große Fortschritte erzielt. Viele Bemühungen auf diesem Gebiet sind der Entwicklung von Endgeräten für Multimedia- oder Telekommunikationsdienste gewidmet. Durch die vielfältigen Einsatzbereiche müssen diese Endgeräte ihre Funktionalität in den unterschiedlichsten Szenarien unter Beweis stellen. Mögliche Anwendungsgebiete umfassen z.B. Audio-/Videokonferenzsysteme, Freisprecheinrichtungen in Kraftfahrzeugen oder durch Bluetooth Kopfhörer, Diktiersysteme, oder öffentliche Informationssysteme. Die digitalen Signalverarbeitungsalgorithmen zielen in diesen Anwendungen darauf ab ein gewünschtes Quellensignal zu schätzen. Dieses Signal kann jedoch von mehreren punktförmigen Störquellen wie z.B. konkurrierenden Sprechern überlagert werden. Möglicherweise sind auch zusätzliche räumlich diffuse Hintergrundgeräusche vorhanden, welche z.B. von Kraftfahrzeugen oder der Überlagerung vieler Sprachsignale z.B. in einer Kantine verursacht werden. Da sich die Endgerätebenutzer wünschen, möglichst natürlich und räumlich ungebunden das Gerät benutzen zu können, ist eine Verwendung von Nahbesprechungsmikrophonen ausgeschlossen. Dies erhöht die Komplexität der Schätzung des gewünschten Quellensignals bedeutend, da damit auch Reflexionen des Wunschsignals und der Störsignale aufgenommen werden.

Bis vor einigen Jahren enthielten die meisten akustischen Mensch-Maschine Schnittstellen nur ein Mikrophon zur Audiosignalaufnahme. Dies beschränkte die Ansätze zur Wiedergewinnung des gewünschten Signals auf einkanalige Geräuschreduktionsalgorithmen wie z.B. [Bol79, EM84]. Auch heute ist dieses Thema weiterhin ein wichtiges Forschungsgebiet wie z.B. in [BMC05, Sri05] gesehen werden kann. Durch gefallene Hardwarekosten beginnen jedoch heutzutage die Hersteller ihre Produkten mit mehreren Mikrophonen auszustatten und ermöglichen damit die Anwendung von mehrkanaligen Signalverarbeitungsalgorithmen. Beispiele von Produkten, welche Mikrophongruppen benutzen,

können auf verschiedenen Gebieten beobachtet werden wie z.B. bei Freisprecheinrichtungen im Pkw [Per02], Bluetooth Kopfhöhrern für Mobiltelefone [VTDM06, Bra07], integrierten Mikrophongruppen in Multimedia Laptops [Mic05] oder digitalen Hörgeräten [HCE+05].

Die Benutzung von mehr als einem Sensor erlaubt zusätzlich zur zeitlichen Filterung auch eine räumliche Filterung der aufgenommenen Signale. Dieser neue Freiheitsgrad wird durch die traditionellen Mehrkanal- oder sogenannten "Array"-Signalverabeitungsansätze ausgenutzt. Diese Algorithmen wurden ursprünglich für schmalbandige Signale entwickelt wie sie z.B. in Radar- oder Sonaranwendungen anzutreffen sind [JD93, Hay02]. Bereits vor mehreren Jahrzehnten gab es erste Versuche diese Methoden auf breitbandige Signale wie z.B. Sprache anzuwenden. Seit diesen Anfängen hat sich das Gebiet weiterentwickelt und es sind verschiedene Methoden verfügbar um ein akustisches gestörtes Signal zu verbessern [BW01, Her05]. Typischerweise nehmen diese sogenannten "beamforming"-Ansätze an dass die Positionen der Sensoren, d.h. die Geometrie der Sensorgruppe bekannt ist und versuchen dann eine kohärente Überlagerung des Wunschsignals zu erreichen, während die Störsignale inkohärent überlagert werden. Dies bedeutet dass diese Algorithmen das Vorhandensein einer einzigen Wunschquelle annehmen deren Position a-priori bekannt sein muss oder von geeigneten Quellenlokalisierungsalgorithmen geschätzt werden muss. Durch die Anwendung linearer adaptiver Filteralgorithmen, basierend auf der Minimierung des mittleren quadratischen Fehlers, ist es zusätzlich möglich die Bahnen der zeitvarianten Wunschquelle und Störquellen zu verfolgen.

In mehreren Anwendungen sind Ansätze wünschenswert welche anstatt der Extraktion einer gewünschten Quelle eine Trennung von mehreren akustischen Quellen zum Ziel haben. Ein Beispiel sind intelligente Besprechungsräume welche mit mehreren Mikrophonen und Kameras ausgestattet sind und Audio-/Videokonferenzen erlauben [Moo02]. Die Möglichkeit eine Besprechung aufzuzeichnen erlaubt eine Nachverarbeitung wie z.B. die Indizierung der Sprecher oder Transkription der Besprechung durch automatische Spracherkennung und erleichtert damit den Zugang zu wichtigen Informationen für Personen welche nicht an der Besprechung teilnehmen konnten [CRG+02]. Da alle Teilnehmer "Wunschquellen" sind, müssen alle Sprachsignale wiedergewonnen werden und für evtl. überlappende Sprachsegmente wären Methoden zur Quellentrennung notwendig. Ein anderes Gebiet in dem eine Trennung von mehreren akustischen Quellen anstatt der Extraktion einer gewünschten Quelle bevorzugt ist, stellt die Sicherheits- und Überwachungstechnik dar. Außerdem sind in solchen Anwendungen oftmals die Positionen der Wunschquellen unbekannt, sodass Ansätze welche auf weniger a-priori Information angewiesen sind wünschenswert sind. Zudem ist es möglich dass in einigen Anwendungen die Geometrie der Mikrophongruppe nicht bekannt ist wie z.B. in einem Besprechungsraum bei auf der Tischplatte aufgestellten Mikrophonen. Ein weiteres Anwendungsgebiet welches nur ungenaue Informationen über die Position der Sensoren zur Verfügung stellt, ist die binaurale

Verarbeitung der Mikrophonsignale von digitalen Hörgeräten [PKRH04, ABZK07].

Eine mögliche Lösung solcher Probleme sind Methoden zur blinden Quellentrennung (engl. "blind source separation (BSS)") welche keine Informationen über Quellen- und Sensorpositionen benötigen. Dieses fehlende a-priori Wissen wird dadurch kompensiert dass die beobachteten Signale basierend auf ihrem Informationsgehalt durch informationstheoretische Signalverarbeitungsalgorithmen verarbeitet werden. Dies ist im Gegensatz zu den linearen adaptiven Filteralgorithmen welche auf der Minimierung des mittleren quadratischen Fehlers basieren und damit nur die statistischen Eigenschaften zweiter Ordnung der Sensorsignale ausnutzen. Die der BSS zugrunde liegende Annahme welche diese Sichtweise erlaubt ist die wechselseitige statistische Unabhängigkeit der Quellensignale. Durch die Formulierung von Optimierungskriterien basierend auf statistischen Größen wie Entropie oder Abstandsmaßen welche die Ähnlichkeit von Wahrscheinlichkeitsdichten bestimmen, kann Statistik höherer Ordnung in die Adaptationsalgorithmen integriert werden. Zusätzlich können andere Signaleigenschaften wie Instationarität oder zeitliche Abhängigkeiten (sogenannte "Nichtweißheit") der Quellensignale ausgenutzt werden. In einem kürzlich erschienenen Überblicksartikel [EP06] wurde darauf hingewiesen, dass dieses Konzept der Anwendung von informationstheoretischen Kriterien auf die adaptive Signalverarbeitung auch in benachbarten Gebieten, wie z.B. Merkmalsextraktion, Clustering oder Systemidentifikation verbesserte Ergebnisse liefert. Jedoch sind in diesen Gebieten heutzutage immer noch Verfahren vorherrschend, welche auf dem mittleren quadratischen Fehler und damit inhärent auf Statistik zweiter Ordnung basieren.

Das Konzept der blinden Quellentrennung kann bis zum Anfang der 80'er Jahre zurückverfolgt werden und seit dem Anfang der frühen 90'er stieß es auch in der Signalverarbeitungsgemeinde auf wachsendes Interesse [JT00]. Damals beschäftigten sich die meisten Forschungsarbeiten mit der verzögerungsfreien Überlagerung der Quellensignale und erst seit Mitte der 90'er Jahre werden Mischsysteme betrachtet welche Reflexionen, wie in der Akustik vorkommend, betrachten [Tor99]. Seitdem wurde eine umfangreiche Anzahl von Forschungsarbeiten auf dem Gebiet der akustischen blinden Quellentrennung in geräuschlosen Umgebungen veröffentlicht.

Im Gegensatz zu der überwiegenden Literatur werden wir in dieser Dissertation die blinde Quellentrennung in verhallten *und* störbehafteten Umgebungen betrachten. Der Hauptbeitrag dieser Dissertation ist von zweifacher Natur: Zum Einen wird gezeigt wie das informationstheoretische Kriterium der Transinformation benutzt werden kann um zum ersten Mal alle drei Signaleigenschaften Nicht-Gaußheit, Instationarität und Nichtweißheit, d.h. zeitliche Abhängigkeiten auszunutzen. Basierend auf diesem Kriterium wird ein generelles Konzept zur blinden Quellentrennung präsentiert. Der Nutzen des vorgeschlagenen Konzepts ist die vereinheitlichte Sicht auf BSS Algorithmen. Dies erlaubt zu erkennen auf welchen Näherungen derzeitige Algorithmen basieren und zeigt damit vielversprechende neue Forschungsrichtungen auf um neue Algorithmen mit wenigeren oder

zutreffenderen Näherungen zu erhalten. Basierend auf dieser Betrachtungsweise werden mehrere neue und effiziente Algorithmen hergeleitet und Beziehungen zu bekannten Algorithmen der BSS Literatur hergestellt. Der zweite Hauptbeitrag dieser Dissertation ist die Präsentation von mehreren Vor- und Nachverarbeitungstechniken um eine geräuschrobuste Adaption der BSS Algorithmen zu gewährleisten, damit die BSS Algorithmen auch in Umgebungen mit starken Hintergrundgeräuschen angewendet werden können. Außerdem wird gezeigt, wie diese Erweiterungen eine Unterdrückung des Hintergrundgeräuschs bei gleichzeitiger Trennung der Punktquellen erreichen.

Die Arbeit welche in dieser Dissertation präsentiert wird ist folgendermaßen strukturiert: In Kapitel 2 wird das BSS Modell eingeführt. Nachdem kurz der einfachste Fall, gegeben durch das verzögerungsfreie BSS Modell, beschrieben wird, konzentrieren wir uns auf das BSS Modell mit Faltungsmixturen. Dieses Modell berücksichtigt, dass in akustischen Umgebungen auch Reflexionen der ursprünglichen Quellensignale durch die Sensoren aufgenommen werden. Anschließend wird der Zusammenhang des BSS Modells mit den Grundlagen der Akustik diskutiert. Danach werden die Signaleigenschaften untersucht, welche von den BSS Ansätzen genutzt werden können. Außerdem werden die durch die Blindheit der BSS Methoden auftretenden Mehrdeutigkeiten adressiert.

Basierend auf dem BSS Modell für Faltungsmixturen wird in Kapitel 3 ein allgemeines Konzept zur blinden Quellentrennung in verhallten Umgebungen eingeführt. Zuerst werden die optimale BSS Lösung und deren Auswirkungen diskutiert. Basierend auf der Unterscheidung zwischen breitbandiger und schmalbandiger Optimierung wird das allgemeine BSS Konzept durch die Formulierung eines generischen breitbandigen Zeitbereichsoptimierungskriterium eingeführt. Ein generischer gradientenbasierter Algorithmus wird hergeleitet und es wird gezeigt wie mehrere effiziente neue und altbekannte Algorithmen durch das Einführen bestimmter Näherungen erhalten werden können. Außerdem wird die Herleitung breitbandiger Algorithmen im diskreten Fourier Transformationsbereich (DFT-Bereich) präsentiert. Diese breitbandigen Algorithmen sind äquivalent zu ihren Gegenstücken im Zeitbereich und weisen daher, im Gegensatz zu den rein schmalbandigen Algorithmen, die BSS Mehrdeutigkeiten nicht unabhängig in jedem DFT Frequenzband auf. Außerdem können durch selektive Näherungen auch effiziente Hybridalgorithmen und reine Schmalbandalgorithmen hergeleitet werden. Nach der Behandlung der verschiedenen Aktualisierungsstrategien werden die in Kapitel 3 hergeleiteten verschiedenen BSS Algorithmen in mehreren verhallten Räumen experimentell untersucht.

Zusätzlich zu den punktförmigen Störquellen werden in Kapitel 4 auch Hintergrundgeräusche betrachtet. Das BSS Modell für Faltungsmixturen beschreibt eine Überlagerung von mehreren Punktquellen und kann daher durch den diffusen Schallfeldcharakter mehrerer realistischer Geräuscharten, wie z.B. Fahrzeuggeräuschen oder ein Gewirr von Stimmen, keine Trennung von gewünschten Punktquellen und Hintergrundgeräuschen bewirken. Deshalb werden in Kapitel 4 mehrere Erweiterungen des allgemeinen BSS Kon-

zepts diskutiert. Zuerst werden mehrere Vorverarbeitungsmethoden betrachtet welche eine geräuschrobuste Adaption der BSS Algorithmen erlauben. Anschließend werden Nachverarbeitungsansätze untersucht, in denen einkanalige Nachfilter in jedem BSS Ausgangskanal angewendet werden. Durch die Hintergrundgeräusche verringert sich die Trennungsleistung der BSS Algorithmen, sodass das Nachfilter sowohl eine Unterdrückung des restlichen Übersprechens von punktförmigen Störquellen, als auch eine Verringerung des Hintergrundgeräuschs bewirken muss. Die vorgestellten Vor- und Nachverarbeitungsalgorithmen werden danach experimentell untersucht.

Abschließend wird die Dissertation in Kapitel 5 zusammengefaßt. Außerdem werden Schlußfolgerungen und Vorschläge für zukünftige Arbeiten diskutiert.

In den Anhängen A und B werden mathematische Operatoren definiert und mehrere Herleitungen detailliert ausgeführt. Aßerdem sind im Anhang C alle in den Experimenten benutzten Umgebungen beschrieben.

F.4 Zusammenfassung and Schlussfolgerungen

In den letzten Jahren gab es viele Forschungsarbeiten auf dem Gebiet der blinden Quellentrennung (engl. “blind source separation (BSS)”) für Faltungsmixturen im Bereich der akustischen Signalverarbeitung. Hier ist vor allem das Gebiet der akustischen Mensch-Maschine Schnittstelle von Interesse, welches noch immer von der Anwendung von festen und adaptiven “Beamformern” geprägt ist. Allerdings gibt es mehrere Gründe, warum es in den letzten Jahren viele Bemühungen gab, BSS auf dieses Gebiet anzuwenden. Ein Grund ist, dass BSS Ansätze nur auf der Annahme der wechselseitigen statistischen Unabhängigkeit der Quellensignale basieren und keine zusätzliche a-priori Information, wie z.B. die Geometrie der Sensorgruppe oder die Positionen der punktförmigen Wunsch- oder Störquellen, benötigen. Außerdem wird die Leistungsfähigkeit eines BSS Algorithmus durch unterschiedliche Frequenzgänge der einzelnen Mikrophone nicht beeinträchtigt. Ein weiterer Grund ist, dass es in mehreren Anwendungen, wie z.B. bei der Überwachung von öffentlichen Plätzen, gewünscht ist, mehrere unterschiedliche Punktquellen zu verfolgen, anstatt eine gewünschte Punktquelle zu extrahieren wie es normalerweise beim “Beamforming” der Fall ist. Außerdem zielen die Ansätze, welche beim adaptiven “Beamforming” verwendet werden, darauf ab den mittleren quadratischen Fehler zu minimieren und basieren daher von Natur aus auf Statistik zweiter Ordnung. Im Gegensatz dazu können BSS Algorithmen informationstheoretische Maße benutzen welche es erlauben, Statistik höherer Ordnung in die Adaptionsalgorithmen zu integrieren. Aufgrund dieser Vorteile wurde den BSS Ansätzen in den letzten Jahren eine große Beachtung in der Signalverarbeitungsgemeinde zuteil.

Das Thema dieser Dissertation beschäftigt sich mit der blinden Quellentrennung von akustischen Signalen und die Errungenschaften dieser Arbeit können wie folgt beschrie-

ben werden. Ein Hauptbeitrag dieser Arbeit ist die Präsentation mehrerer wichtiger Spezialfälle eines allgemeinen BSS Konzepts welches in [BAK03a] TRINICON ("TRIple-N Independent component analysis for CONvolutive mixtures") genannt wurde. Dieses Konzept erlaubt eine vereinheitlichte Betrachtungsweise der BSS Algorithmen für Faltungsmixturen und ermöglicht die Herleitung neuer Algorithmen und zeigt außerdem Verbindungen zu bekannten Algorithmen aus der BSS Literatur. In dieser Arbeit haben wir einige Näherungen präsentiert, welche zu höchst effizienten Algorithmen geführt haben während gleichzeitig, im Gegensatz zu anderen bekannten Algorithmen, die überlegenen Eigenschaften des allgemeinen Konzepts beibehalten wurden. Ein zweiter Hauptbeitrag ist die in sich stimmige und über die existierende BSS Literatur hinausgehende Betrachtung der Anwendung von BSS auf verhallte *und* störbehaftete Umgebungen. Dies wurde ermöglicht durch eine Erweiterung des allgemeinen Konzepts mit einigen Vor- und Nachverarbeitungsmethoden welche die hohe Trennungsleistung der BSS Algorithmen auch in störbehafteten Umgebungen erhalten. Zusätzlich erlauben diese Methoden auch eine Unterdrückung des unerwünschten Hintergrundrauschens, welches durch das BSS Modell für Faltungsmixturen nicht behandelt werden kann.

Um diese Resultate zu erreichen wurde zuerst ein breitbandiges Zeitbereichsoptimierungskriterium, basierend auf einer Verallgemeinerung der Transinformation, eingeführt. Dieses Kriterium basiert auf der statistischen Unabhängigkeit der Quellensignale, erlaubt aber zeitliche Abhängigkeiten innerhalb jedes Quellensignals. Durch die Verwendung von multivariaten Wahrscheinlichkeitsdichtefunktionen (engl. "probability density functions (pdfs)") innerhalb des Kriteriums, ist es möglich die zeitlichen Abhängigkeiten, d.h., die Nichtweißheit der Quellensignale zu berücksichtigen. Dies erlaubt uns eine Ausnutzung aller drei (d.h., "TRIple-N") Signaleigenschaften Nichtgaußheit, Nichtweißheit und Instationarität durch ein breitbandiges Zeitbereichskriterium.

Anschließend wurden, ausgehend von dem TRINICON Optimierungskriterium, mehrere breitbandige iterative BSS Algorithmen hergeleitet, welche auf dem Gradientenabstieg und natürlichem Gradientenabstieg basieren. Die Schätzung der multivariaten pdfs in den Aktualisierungsgleichungen wurde durch die Annahme von sphärischen rotationsinvarianten Zufallsprozessen (engl. "spherically invariant random processes (SIRPs)"), welche bekanntermaßen ein gutes Modell für Sprachsignale darstellen, ermöglicht. Dieses Modell vereinfacht die Implementierung der Aktualisierungsgleichungen erheblich. Außerdem konnten durch die Benutzung der multivariaten Gauß-Dichte als einen Spezialfall einer SIRP pdf effiziente und ausschließlich auf Statistik zweiter Ordnung basierende BSS Algorithmen erhalten werden.

Alle diese Betrachtungen wurden bis hierher im Zeitbereich ausgeführt. Um effiziente Implementierungen im DFT Bereich zu erhalten wurden die generischen Zeitbereichsaktualisierungsgleichungen äquivalent im DFT Bereich formuliert. Diese Äquivalenz wurde durch eine Matrixschreibweise erreicht, welche es erlaubt, die resultierenden Töplitz-

Matrizen durch zirkulante Matrizen zusammen mit Fenstermatrizen darzustellen. Anschließend wurden die zirkulanten Matrizen in den DFT Bereich transformiert. Die strenge Anwendung dieser Prozedur lieferte äquivalente breitbandige Aktualisierungsgleichungen ausgedrückt durch DFT-Bereichsgrößen. Durch die breitbandige Natur tauchen mehrere Matrizen (engl. "constraint matrices") auf, welche eine Kopplung zwischen den DFT Frequenzbändern erzwingen. Dies steht im Gegensatz zur schmalbandigen Optimierung bei der jedes Frequenzband unabhängig betrachtet wird und wodurch auch die Skalierungs- und Permutationsmehrdeutigkeiten in jedem Frequenzband unabhängig auftreten.

Die breitbandige DFT-Bereichsimplementierung war der Startpunkt für die Einführung selektiver Schmalbandnäherungen, d.h. der selektiven Entfernung einiger dieser "Constraint"-Matrizen. Dies erlaubt es z.B. eine Matrixinverse effizient durch eine skalare Inversion in jedem DFT Frequenzband zu berechnen. Durch die gleichzeitige Beibehaltung einiger "Constraint"-Matrizen ist jedoch weiterhin eine Kopplung zwischen den DFT Frequenzbändern gesichert. Damit vermeiden solche Hybridalgorithmen, dass die BSS Mehrdeutigkeiten in jedem Frequenzband unabhängig auftreten und erzielen dabei eine gleichzeitige Reduktion der Rechenkomplexität. Durch die Einführung selektiver Näherungen konnten außerdem einige Verbindungen zu bekannten BSS Algorithmen in der Literatur aufgezeigt werden. Diese Zusammenhänge unterstützen den Anspruch einer vereinheitlichten Sichtweise auf BSS Algorithmen für Faltungsmixturen durch das allgemeine TRINICON Konzept.

Bis zu diesem Punkt basierte das allgemeine Konzept, welches in dieser Arbeit präsentiert wurde, auf einem BSS Modell für Faltungsmixturen welches nur Punktquellen, jedoch keine (möglicherweise diffusen) Hintergrundgeräusche erlaubt. Um die Anwendung der vorher hergeleiteten BSS Algorithmen auch in störbehafteten Umgebungen sicherzustellen, muss entweder ein störrobustes Optimierungskriterium gefunden werden, oder die Algorithmen müssen durch Vor- oder Nachverarbeitungsansätze ergänzt werden. Da eine Formulierung von störrobusten Kriterien für realistische Hintergrundgeräusche schwierig ist, konzentrierten wir uns in dieser Arbeit auf Vor- und Nachverarbeitung. Das Ziel dieser Methoden ist von zweifacher Art: Erstens müssen sie sicherstellen, dass die Trennungsleistung der BSS Algorithmen erhalten bleibt, und zweitens müssen sie die Hintergrundgeräusche, welche der BSS Algorithmus nicht adressieren kann, unterdrücken. Es wurde gezeigt, dass Vorverarbeitungsmethoden nur von beschränkter Anwendbarkeit sind, wegen der schwierigen Aufgabe eine robuste Adaptationssteuerung zu entwickeln und wegen der Tatsache, dass eine Wiederherstellung nicht nur des Betragsspektrums, sondern auch der Phase der geräuschfreien überlagerten Punktquellensignale entscheidend ist. Im Gegensatz dazu können Nachverarbeitungsmethoden eine bessere gleichzeitige Unterdrückung von Hintergrundgeräuschen und restlichem Übersprechen erreichen. Eine auf einkanaliger Nachfilterung basierende Methode wurde präsentiert und mehrere Verbindungen zu existierenden Ansätzen wurden aufgezeigt. Durch Experimente wurde bestätigt, das die-

ser Ansatz eine erhebliche Verbesserung der Trennungsleistung erzielt und außerdem die Hintergrundgeräusche reduziert.

Diese Dissertation zeigt auch einige Startpunkte für zukünftige Forschungsarbeiten auf. Ein zukünftiges Themengebiet könnte die Entwicklung von BSS Algorithmen basierend auf partitionierter adaptiver Filterung sein. Dies würde es erlauben, die zwei widersprüchlichen Forderungen nach einer kurzen Blocklänge, wegen der Instationarität der akustischen Signale, und nach einer großen Entmischfilterlänge, zur Abdeckung aller Reflexionen in verhallten Umgebungen, zu erfüllen. Außerdem wurde auf dem Gebiet der Nachfilterung für akustische Echokompensation in [EMV01, EMV02b, EMV02a] gezeigt, dass eine Partitionierung in einer besseren Schätzung der spektralen Leistungsdichten, welche für das Nachfilter benötigt werden, resultiert. Deshalb kann durch eine Partitionierung auch eine Verbesserung der bereits sehr guten Ergebnisse für ein BSS Nachfilter erwartet werden. Zusätzlich würde eine Partitionierung auch eine Echtzeitimplementierung mit sehr geringer Verzögerung erlauben, wie sie z.B. bei Hörgeräten erwünscht ist.

Ein weiteres lohnendes zukünftiges Forschungsthema könnte die Entwicklung anderer geeigneter Näherungen darstellen, welche die Nichtlinearität innerhalb der auf Statistik höherer Ordnung basierenden BSS Algorithmen effizient berechnen. Außerdem ist die Untersuchung von robusteren Nichtlinearitäten, z.B. basierend auf robuster Statistik [Hub81] ein vielversprechendes Forschungsthema wie vor kurzem für den Schmalbandfall in [CD06] und für den Breitbandfall in [Buc] gezeigt wurde.

Nicht zuletzt stehen in gewissen Anwendungen wie z.B. Video- oder Audiokonferenzsystemen mehr Sensoren zur Verfügung als aktive Quellen vorhanden sind. In diesem Fall stellen die in Abschnitt 4.1.2 kurz diskutierten Unterraumverfahren eine vielversprechende Methode dar, die Information mehrerer Sensoren auszunutzen. Außerdem ist es möglich, dass Information über die Geometrie der Sensorgruppe vorliegt. In diesem Fall wäre es wünschenswert diese a-priori Informationen auszunutzen, jedoch unter Verwendung der in dieser Dissertation benutzten informationstheoretischen Kriterien anstatt der konventionellen adaptiven “Beamforming”-Ansätze. Damit könnten idealerweise weiterhin alle drei Signaleigenschaften Nichtgaußheit, Nichtweissheit und Instationarität ausgenutzt werden und es könnte auf eine Adaptionssteuerung, wie gemeinhin bei adaptiven “Beamformern” benötigt, verzichtet werden. Erste Ansätze, welche dieses Problem adressieren, wurden in [FP01b, PA02, KMG$^+$07] präsentiert.

Bibliography

[AAM+02] R. Aichner, S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari. Time-domain blind source separation of non-stationary convolved signals by utilizing geometric beamforming. In *Proc. IEEE Int. Workshop Neural Networks for Signal Processing (NNSP)*, pages 445–454, Martigny, Switzerland, September 2002.

[AB79] J. Allen and D. Berkley. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Amer.*, 65(4):943–950, April 1979.

[ABAM03] R. Aichner, H. Buchner, S. Araki, and S. Makino. On-line time-domain blind source separation of nonstationary convolved signals. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 987–992, Nara, Japan, April 2003.

[ABK03] R. Aichner, H. Buchner, and W. Kellermann. Comparison and a theoretical link between time-domain and frequency-domain blind source separation. In *Proc. 29th Annual German Conference on Acoustics (DAGA)*, pages 178–179, Aachen, Germany, March 2003.

[ABK04] R. Aichner, H. Buchner, and W. Kellermann. Convolutive blind source separation for noisy mixtures. In *Proc. of Joint Meeting of the German and the French Acoustical Societies (CFA/DAGA 2004)*, pages 583–584, Strasbourg, France, March 2004.

[ABK05] R. Aichner, H. Buchner, and W. Kellermann. On the causality problem in time-domain blind source separation and deconvolution algorithms. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 5, pages 181–184, Philadelphia, PA, USA, March 2005.

[ABK06a] R. Aichner, H. Buchner, and W. Kellermann. Exploiting narrowband efficiency for broadband convolutive blind source separation. *EURASIP Journal on Applied Signal Processing*, 2007:1–9, September 2006.

[ABK06b] R. Aichner, H. Buchner, and W. Kellermann. A novel normalization and regularization scheme for broadband convolutive blind source separation. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 527–535, Charleston, SC, USA, March 2006.

[ABWK06] R. Aichner, H. Buchner, S. Wehr, and W. Kellermann. Robustness of acoustic multiple-source localization in adverse environments. In *Proc. of 7th ITG-Fachtagung Sprachkommunikation*, Kiel, Germany, April 2006.

[ABYK04] R. Aichner, H. Buchner, F. Yan, and W. Kellermann. Real-time convolutive blind source separation based on a broadband approach. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 833–840, Granada, Spain, September 2004.

[ABYK06] R. Aichner, H. Buchner, F. Yan, and W. Kellermann. A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments. *Signal Processing*, 86(6):1260–1277, June 2006.

[ABZK07] R. Aichner, H. Buchner, M. Zourub, and W. Kellermann. Multi-channel source separation preserving spatial information. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 1, pages 5–8, Honolulu, HI, USA, April 2007.

[ADCY97] S.-I. Amari, S.C. Douglas, A. Cichocki, and H.H. Yang. Multichannel blind deconvolution and equalization using the natural gradient. In *Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications*, pages 101–104, Paris, France, April 1997.

[AF95] B. Ayad and G. Faucon. Acoustic echo and noise cancelling for hands-free communication systems. In *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 91–94, Røros, Norway, June 1995.

[AIO+03] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki. Combined approach of array processing and independent component analysis for blind separation of acoustic signals. *IEEE Trans. Speech Audio Processing*, 11(3):204–215, May 2003.

[AK00] J. Anemüller and B. Kollmeier. Amplitude modulation decorrelation for convolutive blind source separation. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 215–220, Helsinki, Finland, June 2000.

[Ama98] S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276, 1998.

[AMAM00] F. Asano, Y. Motomura, H. Asoh, and T. Matsui. Effect of PCA filter in blind source separation. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, Helsinki, Finland, June 2000.

[AMB+03] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada. Blind separation of more speech than sensors with less distortion by combining sparseness and ICA. In *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 271–274, Kyoto, Japan, September 2003.

[Ane01] J. Anemüller. *Across-frequency processing in convolutive blind source separation*. PhD thesis, Universität Oldenburg, Oldenburg, Germany, 2001.

[AS72] M. Abramowitz and I.A. Stegun, editors. *Handbook of Mathematical Functions*. Dover Publications, Inc., New York, USA, 1972.

[ASM03] J. Anemüller, T.J. Sejnowski, and S. Makeig. Complex independent component analysis of frequency-domain electroencephalographic data. *Neural Networks*, 16:1311–1323, 2003.

[ASR+06] F. Antonacci, D. Saiu, P. Russo, A. Sarti, M. Tagliasacchi, and S. Tubaro. Experimental evaluation of a localization algorithm for multiple acoustic sources in reverberating environments. In *Proc. Eur. Signal Processing Conf. (EUSIPCO)*, Florence, Italy, September 2006.

[AZBK06] R. Aichner, M. Zourub, H. Buchner, and W. Kellermann. Post-processing for convolutive blind source separation. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 5, pages 37–40, Toulouse, France, May 2006.

[BA03] H. Buchner and R. Aichner. Derivation of the gradient of the time-domain optimization criterion, 2003. Unpublished notes.

[BA05] H. Buchner and R. Aichner. Operators for block Sylvester matrices and their application to the derivation of the optimum BSS solution, 2005. Unpublished notes.

[BAK03a] H. Buchner, R. Aichner, and W. Kellermann. Blind source separation algorithms for convolutive mixtures exploiting nongaussianity, nonwhiteness, and nonstationarity. In *Proc. Int. Workshop on Acoustic Echo*

and Noise Control (IWAENC), pages 275–278, Kyoto, Japan, September 2003.

[BAK03b] H. Buchner, R. Aichner, and W. Kellermann. A generalization of a class of blind source separation algorithms for convolutive mixtures. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, volume 1, pages 945–950, Nara, Japan, April 2003.

[BAK04a] H. Buchner, R. Aichner, and W. Kellermann. Blind source separation for convolutive mixtures: A unified treatment. In J. Benesty and Y. Huang, editors, *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, pages 255–293. Kluwer Academic Publishers, Boston, 2004.

[BAK04b] H. Buchner, R. Aichner, and W. Kellermann. TRINICON: A versatile framework for multichannel blind signal processing. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 3, pages 889–892, Montreal, Canada, May 2004.

[BAK05a] H. Buchner, R. Aichner, and W. Kellermann. A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics. *IEEE Trans. Speech Audio Processing*, 13(1):120–134, January 2005.

[BAK05b] H. Buchner, R. Aichner, and W. Kellermann. Relation between blind system identification and convolutive blind source separation. In *Proc. Joint Workshop on Hands-Free Communication and Microphone Arrays*, pages d3–d4, Piscataway, NJ, USA, March 2005.

[BAK07] H. Buchner, R. Aichner, and W. Kellermann. TRINICON-based blind system identification with application to multiple-source localization and separation. In S. Makino, T.-W. Lee, and S. Sawada, editors, *Blind Speech Separation*. Springer, Berlin/Heidelberg, 2007.

[BAM03] A. Blin, S. Araki, and S. Makino. Blind source separation when speech signals outnumber sensors using a sparseness-mixing matrix estimation (SMME). In *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 211–214, Kyoto, Japan, September 2003.

[BAMCM97] A. Belouchrani, K. Abed-Meraim, J.F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *IEEE Trans. Signal Processing*, 45(2):434–444, February 1997.

[BAS+05] H. Buchner, R. Aichner, J. Stenglein, H. Teutsch, and W. Kellermann. Simultaneous localization of multiple sound sources using blind adaptive MIMO filtering. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 3, pages 97–100, Philadelphia, PA, USA, March 2005.

[BBGK06] H. Buchner, J. Benesty, T. Gänsler, and W. Kellermann. Robust extended multidelay filter and double-talk detector for acoustic echo cancellation. *IEEE Trans. Speech Audio Processing*, 14(5):1633–1644, September 2006.

[BBK03] H. Buchner, J. Benesty, and W. Kellermann. Multichannel frequency-domain adaptive filtering with application to acoustic echo cancellation. In Y. Huang and J. Benesty, editors, *Adaptive signal processing: Application to real-world problems*, pages 95–128. Springer, Berlin, January 2003.

[BBK05] H. Buchner, J. Benesty, and W. Kellermann. Generalized multichannel frequency-domain adaptive filtering: Efficient realization and application to hands-free speech communication. *Signal Processing*, 85:549–570, 2005.

[Ben00] J. Benesty. Adaptive eigenvalue decomposition for passive acoustic source localization. *J. Acoust. Soc. Amer.*, 107:384–391, January 2000.

[Bit01] J. Bitzer. *Mehrkanalige Geräuschunterdrückungssyteme - eine vergleichende Analyse*. PhD thesis, Universität Bremen, Bremen, September 2001. In German.

[BMC05] J. Benesty, S. Makino, and J. Chen, editors. *Speech Enhancement*. Springer, Berlin, 2005.

[BNR96] T.P Barnwell III, K. Nayebi, and C.H. Richardson. *Speech Coding: A Computer Laboratory Textbook*. Digital Signal Processing Laboratory Series. Georgia Tech, 1996.

[Bof03] P. Bofill. Underdetermined blind source separation of delayed sound sources in the frequency domain. *Neurocomputing*, 55(1):627–641, 2003.

[Bol79] S.F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-27(2):113–120, April 1979.

[Bor84] J. Borish. Extension of the image model to arbitrary polyhedra. *J. Acoust. Soc. Amer.*, 75(6):1827–1836, 1984.

[Bor85] B. Borat. Second-order equivalence of rectangular and exponential windows in least-squares estimation of gaussian autoregressive processes. *IEEE Trans. Acoust., Speech, Signal Processing*, 33(4):1209–1212, October 1985.

[BP66] J. Bendat and A. Piersol. *Measurement and analysis of random data*. John Wiley & Sons, New York, 1966.

[Bra07] J. Bradley. Gennum releases the nX6000 bluetooth headset, January 2007. Gennum Corporation, Press release.

[Bre82] H. Brehm. Description of spherically invariant random processes by means of G-functions. In D. Jungnickel and K. Vedder, editors, *Lecture Notes in Mathematics*, pages 39–73. Springer, Berlin, 1982.

[BS87] H. Brehm and W. Stammler. Description and generation of spherically invariant speech-model signals. *Signal Processing*, 12:119–141, 1987.

[BS95] A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[BSBAM01] H. Bousbie-Salah, A. Belouchrani, and K. Abed-Meraim. Blind separation of non-stationary sources using joint block diagonalization. In *Proc. IEEE Int. Workshop on Statistical Signal Processing*, pages 448–451, 2001.

[BSM79] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, pages 208–211, April 1979.

[Buc] H. Buchner. *Broadband Adaptive MIMO Filtering: A Unified Treatment and Applications to Acoustic Human-Machine Interfaces*. PhD thesis, Universität Erlangen-Nürnberg, Erlangen, Germany. To appear.

[Buc02] M. Buck. Aspects of first-order differential microphone arrays in the presence of sensor imperfections. *European Transactions on Telecommunications*, 13(2):115–122, March-April 2002.

[Buc04] M. Buck. *Mehrkanalige Systeme zur Geräuschunterdrückung für Sprachanwendungen unter Berücksichtigung von Mikrofoneigenschaften*. PhD thesis, Universität Ulm, Ulm, March 2004. In German.

[BW01] M. Brandstein and D. Ward, editors. *Microphone Arrays*. Springer, 2001.

[BW05] G.J. Brown and D. Wang. Separation of speech by computational auditory scene analysis. In J. Benesty, S. Makino, and J. Chen, editors, *Speech Enhancement*, pages 371–402. Springer, Berlin, 2005.

[CA02] A. Cichocki and S.-I. Amari. *Adaptive Blind Signal and Image Processing*. John Wiley & Sons, Chichester, 2002.

[CACL99] S. Choi, S.-I. Amari, A. Cichocki, and R.-W. Liu. Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 371–376, Aussois, France, January 1999.

[Car87] G.C. Carter. Coherence and time delay estimation. *Proceedings of the IEEE*, 75(2):236–255, February 1987.

[Car89] J.-F. Cardoso. Source separation using higher-order moments. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 4, pages 2109–2112, Glasgow, Scotland, May 1989.

[Car92] J.-F. Cardoso. Iterative techniques for blind source separation using only fourth order cumulants. In *Proc. Eur. Signal Processing Conf. (EUSIPCO)*, pages 739–742, Brussels, Belgium, August 1992.

[Car98] J.-F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 9(10):2009–2025, October 1998.

[Car03] J.-F. Cardoso. Independent component analysis of the cosmic microwave background. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 1111–1116, Nara, Japan, April 2003.

[Car04] J.-F. Cardoso. Optimization issues in noisy Gaussian ICA. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 41–48, Granada, Spain, September 2004.

[CB01] I. Cohen and B. Berdugo. Speech enhancement for non-stationary noise environments. *Signal Processing*, 81:2403–2418, 2001.

[CC01] S. Choi and A. Cichocki. Correlation matching approach to source separation in the prescence of spatially correlated noise. In *Proc. IEEE Int. Symp. Signal Processing and Applications (ISSPA)*, pages 272–275, Kuala Lumpur, Malaysia, August 2001.

[CD06] J.-C. Chao and S. Douglas. A simple and robust FASTICA algorithm using the Huber M-estimator cost function. In *Proc. IEEE Int. Conf.*

Acoustics, Speech, Signal Processing (ICASSP), volume 5, pages 685–688, Toulouse, France, May 2006.

[CDA98] A. Cichocki, S. Douglas, and S.-I. Amari. Robust techniques for independent component analysis (ICA) with noisy data. *Neurocomputing*, 22:113–129, 1998.

[CJH91] P. Comon, C. Jutten, and J. Herault. Blind separation of sources, part II: Problems statement. *Signal Processing*, 24:11–20, 1991.

[CJLK04] C. Choi, G.-J. Jang, Y. Lee, and S.R. Kim. Adaptive cross-channel interference cancellation on blind source separation outputs. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 857–864, Granada, Spain, September 2004.

[CKN73] G.C. Carter, C.H. Knapp, and A.H. Nuttall. Estimation of the magnitude-squared coherence function via overlapped fast fourier transform processing. *IEEE Trans. Audio Electroacoust.*, AU-21(4):337–344, August 1973.

[CL96] J.-F. Cardoso and B.H. Laheld. Equivariant adaptive source separation. *IEEE Trans. Signal Processing*, 44(12):3017–3030, December 1996.

[Coh03] I. Cohen. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Trans. Speech Audio Processing*, 11(5):466–475, September 2003.

[Com94] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, April 1994.

[CRG+02] R. Cutler, Y. Rui, A. Gupta, J.J. Cadiz, I. Tashev, L.-W. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg. Distributed meetings: A meeting capture and broadcasting system. In *Proc. Int. Conf. on Multimedia*, pages 503–512, Juan-les-Pins, France, 2002.

[Cro98] M.J. Crocker, editor. *Handbook of Acoustics*. John Wiley & Sons, New York, 1998.

[CS96] J.-F. Cardoso and A. Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM Journal of Matrix Analysis and Applications*, 17(1):161–167, January 1996.

[CSL95] V. Capdevielle, C. Serviere, and J.L. Lacoume. Blind separation of wide-band sources in the frequency domain. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 3, pages 2080–2083, Detroit, MI, May 1995.

[CT91] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.

[CU94] A. Cichocki and R. Unbehauen. *Neural Networks for Optimization and Signal Processing*. John Wiley & Sons, Chichester, 1994.

[CWB+55] R.K. Cook, R.V. Waterhouse, R.D. Berendt, S. Edelman, and M.C. Thompson, Jr. Measurement of correlation coefficients in reverberant sound fields. *J. Acoust. Soc. Amer.*, 27(6):1072–1077, November 1955.

[Däm57] P. Dämmig. Zur Messung der Diffusität von Schallfeldern durch Korrelation. *Acustica*, 7:387ff., 1957. In German.

[Dav52] W.B. Davenport. An experimental study of speech wave propability distribution. *J. Acoust. Soc. Amer.*, 24(4):390–399, 1952.

[DBC91] M. Dendrinos, S. Bakamidis, and G. Carayannis. Speech enhancement from noise: A regenerative approach. *Speech Communication*, 10:45–57, 1991.

[DC98] S. Douglas and A. Cichocki. Adaptive step size techniques for decorrelation and blind source separation. In *Proc. 32nd Asilomar Conf. on Signals, Systems, and Computers*, pages 1191–1195, Pacific Grove, CA, USA, November 1998.

[DCA98] S.C. Douglas, A. Cichocki, and S.-I. Amari. A bias removal technique for blind source separation with noisy measurements. *Electronic Letters*, 34(14):1379–1380, July 1998.

[DHP00] J.R. Deller, J.H.L. Hansen, and J.G. Proakis. *Discrete-Time Processing of Speech Signals*. IEEE Press, New York, 2000.

[Div05] P. Divenyi, editor. *Speech Separation by Humans and Machines*. Kluwer Academic Publishers, MA, USA, 2005.

[DM02] S. Doclo and M. Moonen. GSVD-based optimal filtering for single and multimicrophone speech enhancement. *IEEE Trans. Signal Processing*, 50(9):2230–2244, September 2002.

[DMH06] M. Dyrholm, S. Makeig, and L.K. Hansen. Model structure selection in convolutive mixtures. In *Proc. 6th Int. Conf. on Independent Component Analysis and Blind Signal Separation*, volume 3889 of *LNCS*, pages 74–81. Springer, 2006.

[Doc03] S. Doclo. *Multi-microphone noise reduction and dereverberation techniques for speech applications*. PhD thesis, Katholieke Universiteit Leuven, Leuven, May 2003.

[DSM03] S. Douglas, H. Sawada, and S. Makino. Natural gradient multichannel blind deconvolution and equalization using causal FIR filters. In *Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers*, pages 197–201, Pacific Grove, CA, USA, November 2003.

[DSM04a] S. Douglas, H. Sawada, and S. Makino. A causal frequency-domain implementation of a natural gradient multichannel blind deconvolution and source separation algorithm. In *Proc. Int. Congr. on Acoustics*, volume 1, pages 85–88, Kyoto, Japan, April 2004.

[DSM04b] S. Douglas, H. Sawada, and S. Makino. Natural gradient multichannel blind deconvolution and source separation using causal FIR filters. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 5, pages 477–480, Montreal, Canada, May 2004.

[EKL06] T. Eltoft, T. Kim, and T.-W. Lee. On the multivariate Laplace distribution. *IEEE Signal Processing Lett.*, 13(5):300–303, May 2006.

[Eks73] M.P. Ekstrom. A spectral characterization of the ill-conditioning in numerical deconvolution. *IEEE Trans. Audio Electroacoust.*, AU-21(4):344–348, August 1973.

[Elk01] G.W. Elko. Spatial coherence functions for differential microphones in isotropic noise fields. In M. Brandstein and D. Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*, pages 61–85. Springer, Berlin, 2001.

[EM84] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-32(6):1109–1121, December 1984.

[EMV01] G. Enzner, R. Martin, and P. Vary. On spectral estimation of residual echo in hands-free telephony. In *Proc. Int. Workshop on Acoustic*

Echo and Noise Control (IWAENC), pages 211–214, Darmstadt, Germany, September 2001.

[EMV02a] G. Enzner, R. Martin, and P. Vary. Partitioned residual echo power estimation for frequency-domain acoustic echo cancellation and postfiltering. *Eur. Trans. Telecommun.*, 13(2):103–114, 2002.

[EMV02b] G. Enzner, R. Martin, and P. Vary. Unbiased residual echo power estimation for hands-free telephony. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 2, pages 1893–1896, Orlando, FL, USA, May 2002.

[EP06] D. Erdogmus and J.C. Principe. From linear adaptive filtering to nonlinear information processing. *IEEE Signal Processing Magazine*, pages 14–33, November 2006.

[ET95] Y. Ephraim and H.L. Van Trees. A signal subspace approach for speech enhancement. *IEEE Trans. Speech Audio Processing*, 3(4):251–266, July 1995.

[Eyr30] C.F. Eyring. Reverberation time in "dead" rooms. *J. Acoust. Soc. Amer.*, 1(2A):217–241, January 1930.

[Fle81] R. Fletcher. *Practical Methods of Optimization*. Volume 2: Constrained Optimization. John Wiley & Sons, 1981.

[FP01a] C.L. Fancourt and L. Parra. The coherence function in blind source separation of convolutive mixtures of non-stationary signals. In *Proc. IEEE Int. Workshop Neural Networks for Signal Processing (NNSP)*, pages 303–312, 2001.

[FP01b] C.L. Fancourt and L. Parra. The generalized sidelobe decorrelator. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2001.

[GC88] H. Gish and D. Cochran. Generalized coherence. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 5, pages 2745–2748, April 1988.

[GC95] S. Van Gerven and D. Van Compernolle. Signal separation by symmetric adaptive decorrelation: Stability, convergence and uniqueness. *IEEE Trans. Signal Processing*, 43(7):1602–1612, July 1995.

[GJ82] L.J. Griffiths and C.W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. on Antennas and Propagation*, AP-30(1):27–34, January 1982.

[GL84] D.W. Griffin and J.S. Lim. Signal estimation from modified short-time fourier transform. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-32(2):236–243, April 1984.

[GL96] G.H. Golub and C.F. Van Loan. *Matrix Computations*. The John Hopkins University Press, Baltimore, ML, 3rd edition, 1996.

[GN95] M.I. Gürelli and C.L. Nikias. EVAM: An eigenvector-based algorithm for multichannel blind deconvolution of input colored signals. *IEEE Trans. Signal Processing*, 43(1):134–149, January 1995.

[Gol76] J. Goldman. Detection in the presence of spherically symmetric random vectors. *IEEE Trans. Inform. Theory*, 22(1):52–59, January 1976.

[GR65] I.S. Gradshteyn and I.M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, New York, 1965.

[Gra72] R.M. Gray. On the asymptotic eigenvalue distribution of Toeplitz matrices. *IEEE Trans. Inform. Theory*, 18(6):725–730, 1972.

[Gre98] J.E. Greenberg. Modified LMS algorithms for speech processing with an adaptive noise canceller. *IEEE Trans. Speech Audio Processing*, 6(4):338–351, 1998.

[GS58] U. Grenander and G. Szegö. *Toeplitz Forms and Their Applications*. University California Press, Berkeley, California, 1958.

[GZ92] J.E. Greenberg and P.M. Zurek. Evaluation of an adaptive beamforming method for hearing aids. *J. Acoust. Soc. Amer.*, 91(3):1662–1676, March 1992.

[GZ03] S. Gazor and W. Zhang. Speech propability distribution. *IEEE Signal Processing Lett.*, 10(7):204–207, July 2003.

[Har97] D. A. Harville. *Matrix Algebra from a Statistician's Perspective*. Springer, Berlin, Germany, 1997.

[Hay02] S. Haykin. *Adaptive Filter Theory*. Prentice-Hall, Englewood Cliffs, NJ, 4th edition, 2002.

[HBK03] W. Herbordt, H. Buchner, and W. Kellermann. An acoustic human-machine front-end for multimedia applications. *EURASIP Journal on Applied Signal Processing*, 1(2003):21–31, January 2003.

[HCE+05] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass. Signal processing in high-end hearing aids: State of the art, challenges, and future trends. *EURASIP Journal on Applied Signal Processing*, 18(2005):2915–2929, 2005.

[Her05] W. Herbordt. *Sound capture for human/machine interfaces - Practical aspects of microphone array signal processing*, volume 315 of *Lecture Notes in Control and Information Sciences*. Springer, Berlin, Germany, 2005.

[Hir06] A. Hiroe. Solution of permutation problem in frequency domain ICA, using multivariate probability density functions. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 601–608, Charleston, SC, USA, March 2006.

[HKO01] A. Hyvaerinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[HNK04] W. Herbordt, S. Nakamura, and W. Kellermann. Multi-channel estimation of the power spectral density of noise for mixtures of non-stationary signals. *IEICE Technical Reports*, 2004(131):211–216, December 2004.

[Hof04] M. Hofbauer. On the FIR inversion of an acoustical convolutive mixing system: Properties and limitations. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 643–651, Granada, Spain, September 2004.

[Hof05] M. Hofbauer. *Optimal Linear Separation and Deconvolution of Acoustical Convolutive Mixtures*. PhD thesis, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland, May 2005.

[HS99] O. Hoshuyama and A. Sugiyama. An adaptive microphone array with good sound quality using auxiliary fixed beamformers and its DSP implementation. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, pages 949–952, Phoenix, Arizona, USA, March 1999.

[HS04] E. Hänsler and G. Schmidt. *Acoustic Echo and Noise Control: A Practical Approach*. John Wiley & Sons, Hoboken, NJ, 2004.

[HSLF06] T.P. Hua, A. Sugiyama, R. Le Bouquin Jeannes, and G. Faucon. Estimation of the signal-to-interference ratio based on normalized cross-correlation with symmetric leaky blocking matrices in adaptive microphone arrays. In *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 1–4, Paris, France, September 2006.

[HTK03] W. Herbordt, T. Trini, and W. Kellermann. Robust spatial estimation of the signal-to-interference ratio for non-stationary mixtures. In *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 247–250, Kyoto, Japan, September 2003.

[Hub81] P.J. Huber. *Robust Statistics*. John Wiley & Sons, New York, 1981.

[Hub03] R. Huber. *Objective assessment of audio quality using an auditory processing model*. PhD thesis, Universität Oldenburg, Oldenburg, Germany, December 2003.

[HZ05] R. Hu and Y. Zhao. Adaptive decorrelation filtering algorithm for speech source separation in uncorrelated noises. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 1, pages 1113–1115, Philadelphia, PA, USA, May 2005.

[HZ06] R. Hu and Y. Zhao. Fast noise compensation for speech separation in diffuse noise. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 5, pages 865–868, Toulouse, France, May 2006.

[IM99] S. Ikeda and N. Murata. A method of ICA in time-frequency-domain. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 365–371, January 1999.

[IM00] M.Z. Ikram and D.R. Morgan. Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 2, pages 1041–1044, Istanbul, Turkey, June 2000.

[IM02] M.Z. Ikram and D.R. Morgan. A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 1, pages 881–884, Orlando, FL, USA, May 2002.

[ISO97] ISO 3382. *Acoustics – Measurement of the reverberation time of rooms with reference to other acoustical parameters*. Geneve, 1997.

[ITU01a] ITU-R. Recommendation BS.1387: Method for objective measurements of perceived audio quality, July 2001.

[ITU01b] ITU-T. Recommendation P.862: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, February 2001.

[JC01] F. Jabloun and B. Champagne. A multi-microphone signal subspace approach for speech enhancement. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, pages 205–208, Salt Lake City, Utah, USA, May 2001.

[JC05] F. Jabloun and B. Champagne. Signal subspace techniques for speech enhancement. In J. Benesty, S. Makino, and J. Chen, editors, *Speech Enhancement*, pages 135–159. Springer, Berlin, 2005.

[JD93] D. H. Johnson and D. E. Dudgeon. *Array signal processing: Concepts and techniques*. Signal Processing Series. Prentice-Hall, 1993.

[Jek05] U. Jekosch. *Voice and Speech Quality Perception*. Springer, Berlin, Germany, 2005.

[JH91] C. Jutten and J. Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[Joh04] M. Joho. Blind signal separation of convolutive mixtures: A time-domain joint-diagonalization approach. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 578–585, Granada, Spain, September 2004.

[JRY00] A. Jourjine, S. Rickard, and Ö. Yilmaz. Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 5, pages 2985–2988, Istanbul, Turkey, June 2000.

[JS03] M. Joho and P. Schniter. Frequency domain realization of a multichannel blind deconvolution algorithm based on the natural gradient. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 543–548, Nara, Japan, April 2003.

[JT00] C. Jutten and A. Taleb. Source separation: From dusk till dawn. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 15–26, Helsinki, Finland, June 2000.

[KALL06] T. Kim, H. Attias, S.-Y. Lee, and T.-W. Lee. Frequency domain blind source separation exploiting higher-order dependencies. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 5, pages 677–680, Toulouse, France, May 2006.

[KALL07] T. Kim, H. Attias, S.-Y. Lee, and T.-W. Lee. Blind source separation exploiting higher-order frequency dependencies. *IEEE Trans. Audio, Speech and Language Processing*, 15(1):70–79, 2007.

[KB03] W. Kellermann and H. Buchner. Wideband algorithms versus narrowband algorithms for adaptive filtering in the DFT domain. In *Proc. Asilomar Conference on Signals, Systems, and Computers*, volume 2, pages 1278–1282, Pacific Grove, CA, USA, November 2003.

[KBA06] W. Kellermann, H. Buchner, and R. Aichner. Separating convolutive mixtures with TRINICON. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 5, pages 961–964, Toulouse, France, May 2006.

[KEL06] T. Kim, T. Eltoft, and T.-W. Lee. Independent vector analysis: An extension of ICA to multivariate components. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 175–172, Charleston, SC, USA, March 2006.

[KJ00] B.S. Krongold and D.L. Jones. Blind source separation of nonstationary convolutively mixed signals. In *Proc. IEEE Workshop on Statistical Signal and Array Processing (SSAP)*, pages 53–57, Pocono Manor, PA, USA, August 2000.

[KKP01] S. Kotz, T. Kozubowski, and K Podgorski. *The Laplace Distribution and Generalizations*. Birkhäuser Verlag, 2001.

[KMG+07] K. Kumatani, U. Mayer, T. Gehrig, E. Stoimenov, J. McDonough, and M. Wolfel. Minimum mutual information beamforming for simultaneous active speakers. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Honolulu, HI, USA, April 2007.

[KMO98] M. Kawamoto, K. Matsuoka, and N. Ohnishi. A method of blind separation for convolved non-stationary signals. *Neurocomputing*, 22:157–171, 1998.

[KSK+00] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura. Evaluation of blind signal separation method using directivity pattern under

reverberant conditions. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 5, pages 3140–3143, Istanbul, Turkey, June 2000.

[Kul59] S. Kullback. *Information Theory and Statistics*. John Wiley & Sons, 1959.

[Kut00] H. Kuttruff. *Room Acoustics*. Spon Press, London, 4th edition, 2000.

[KZN03] K. Kokkinakis, V. Zarzoso, and A.K. Nandi. Blind separation of acoustic mixtures based on linear prediction analysis. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 343–348, Nara, Japan, 2003.

[Lam96] R.H. Lambert. *Multichannel Blind Deconvolution: FIR Matrix Algebra and Separation of Multipath Mixtures*. PhD thesis, University of Southern California, CA, USA, 1996.

[LBL97] T.-W. Lee, A.J. Bell, and R.H. Lambert. Blind separation of delayed and convolved sources. In *Advances in Neural Information Processing Systems*, volume 9, pages 758–764. MIT Press, Cambridge, 1997.

[LNT04] S.Y. Low, S. Nordholm, and R. Tognieri. Convolutive blind signal separation with post-processing. *IEEE Trans. Speech Audio Processing*, 12(5):539–548, September 2004.

[LVKL96] T.I. Laasko, V. Välimäki, M. Karjalainen, and U.K. Laine. Splitting the unit delay. *IEEE Signal Processing Magazine*, pages 30–60, January 1996.

[LZOS98] T.-W. Lee, A. Ziehe, R. Orglmeister, and T. Sejnowski. Combining time-delayed decorrelation and ICA: Towards solving the cocktail party problem. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, pages 1249–1252, Seattle, WA, May 1998.

[MA95] R. Martin and J. Altenhöner. Coupled adaptive filters for acoustic echo control and noise reduction. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, pages 3043–3046, Detroit, MI, May 1995.

[MAG95] E. Moulines, O.A. Amrane, and Y. Grenier. The generalized multidelay adaptive filter: structure and convergence analysis. *IEEE Trans. Signal Processing*, 43(1):14–28, January 1995.

[Mar94] R. Martin. Spectral subtraction based on minimum statistics. In *Proc. Eur. Signal Processing Conf. (EUSIPCO)*, pages 1182–1185, September 1994.

[Mar95] R. Martin. *Freisprecheinrichtungen mit mehrkanaliger Echokompensation und Störgeräuschreduktion*. PhD thesis, RWTH Aachen, Aachen, Germany, June 1995. In German.

[Mar96] R. Martin. The echo shaping approach to acoustic echo control. *Speech Communication*, 20:181–190, 1996.

[Mar01a] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Processing*, 9(5):504–512, July 2001.

[Mar01b] R. Martin. Small microphone arrays with postfilters for noise and acoustic echo reduction. In M. Brandstein and D. Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*, pages 255–279. Springer, Berlin, 2001.

[Mar05] R. Martin. Statistical methods for the enhancement of noisy speech. In J. Benesty, S. Makino, and J. Chen, editors, *Speech Enhancement*, pages 43–65. Springer, Berlin, 2005.

[MASM02a] R. Mukai, S. Araki, H. Sawada, and S. Makino. Removal of residual crosstalk components in blind source separation using LMS filters. In *Proc. IEEE Int. Workshop Neural Networks for Signal Processing (NNSP)*, pages 435–444, Martigny, Switzerland, September 2002.

[MASM02b] R. Mukai, S. Araki, H. Sawada, and S. Makino. Removal of residual crosstalk components in blind source separation using time-delayed spectral subtraction. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 2, pages 1789–1792, Orlando, FL, USA, May 2002.

[MC99] R. Martin and R.V. Cox. New speech enhancement techniques for low bit rate speech coding. In *Proc. IEEE Workshop on Speech Coding*, pages 165–167, 1999.

[MD02] H. Mathis and S.C. Douglas. On the existence of universal nonlinearities for blind source separation. *IEEE Trans. Signal Processing*, 50(5):1007–1016, May 2002.

[MG76] J.D. Markel and A.H. Gray. *Linear Prediction of Speech*. Springer, Berlin, 1976.

[Mic05] Microsoft Corporation. Microphone array support in Windows Vista, August 2005. White Paper.

[MK88] M. Miyoshi and Y. Kaneda. Inverse filtering of room acoustics. *IEEE Trans. Acoust., Speech, Signal Processing*, 36(2):145–152, February 1988.

[MMS98] C. Marro, Y. Mahieux, and K.U. Simmer. Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. *IEEE Trans. Speech Audio Processing*, 6(3):240–259, May 1998.

[MN01] K. Matsuoka and S. Nakashima. Minimal distortion principle for blind source separation. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 722–727, San Diego, CA, USA, December 2001.

[MOK95] K. Matsuoka, M. Ohya, and M. Kawamoto. Neural net for blind separation of nonstationary signals. *IEEE Trans. Neural Networks*, 8(3):411–419, 1995.

[Möl00] S. Möller. *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic Publishers, Boston, USA, 2000.

[Moo02] D.C. Moore. The IDIAP smart meeting room. Technical report, Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), Martigny, Switzerland, November 2002.

[MOTN03] K. Matsuoka, Y. Ohba, Y. Toyota, and S. Nakashima. Blind separation for convolutive mixture of many voices. In *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 279–282, Kyoto, Japan, September 2003.

[MS94] L. Molgedey and H.G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 72:3634–3636, 1994.

[MSAM03] R. Mukai, H. Sawada, S. Araki, and S. Makino. Robust real-time blind source separation for moving speakers using blockwise ICA and residual crosstalk subtraction. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 975–980, Nara, Japan, April 2003.

[MSAM05] R. Mukai, H. Sawada, S. Araki, and S. Makino. Real-time blind source separation and DOA estimation using small 3-D microphone array. In *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 45–48, Eindhoven, Netherlands, September 2005.

[MZB87] F. Milinazzo, C. Zala, and I. Barrodale. On the rate of growth of condition numbers for convolution matrices. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-35(4):471–475, April 1987.

[Neu01] R.O. Neubauer. Existing reverberation time formulae - A comparison with computer simulated reverberation times. In *Proc. International Congress on Sound and Vibration*, pages 805–812, HongKong, China, July 2001.

[NG05] P. Naylor and N.D. Gaubitch. Speech dereverberation. In *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Eindhoven, Netherlands, September 2005.

[NSS02] T. Nishikawa, H. Saruwatari, and K. Shikano. Comparison of time-domain ICA, frequency-domain ICA and multistage ICA for blind source separation. In *Proc. Eur. Signal Processing Conf. (EUSIPCO)*, volume 2, pages 15–18, September 2002.

[OK05] P. Oak and W. Kellermann. A calibration algorithm for robust generalized sidelobe cancelling beamformers. In *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 97–100, Eindhoven, Netherlands, September 2005.

[OSB98] A.V. Oppenheim, R.W. Schafer, and J.R. Buck. *Discrete-Time Signal Processing*. Signal Processing Series. Prentice-Hall, Upper Saddle River, NJ, 2nd edition, 1998.

[PA02] L. Parra and C. Alvino. Geometric source separation: Merging convolutive source separation with geometric beamforming. *IEEE Trans. Speech Audio Processing*, 10(6):352–362, September 2002.

[Pap02] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, Boston, 4th edition, 2002.

[Per02] A. Persterer. AKG supplies array microphones for new Mercedes-Benz E-class, 2002. AKG Acoustics, Press release.

[Pet86] P.M. Peterson. Simulating the response of multiple microphones to a single acoustic source in a reverberant room. *J. Acoust. Soc. Amer.*, 80(5):1527–1529, 1986.

[Pie91] A.D. Pierce. *Acoustics: An Introduction to Its Physical Principles and Applications*. Acoustical Society of America, New York, 1991.

[PKRH04] M.S. Pedersen, U. Kjems, K.B. Rasmussen, and L.K. Hansen. Semi-blind source separation using head-related transfer functions. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 5, pages 713–716, Montreal, Canada, May 2004.

[PP06] K.B. Petersen and M.S. Pedersen. The matrix cookbook. Technical report, Technical University of Denmark, February 2006.

[PPSK06] K.S. Park, J.S. Park, K.S. Son, and H.T. Kim. Postprocessing with Wiener filtering technique for reducing residual crosstalk in blind source separation. *IEEE Signal Processing Lett.*, 13(12):749–751, December 2006.

[PS00] L. Parra and C. Spence. Convolutive blind source separation of non-stationary sources. *IEEE Trans. Speech Audio Processing*, 8(3):320–327, May 2000.

[PSS00] L. Parra, C. Spence, and P. Sajda. Higher-order statistical properties arising from the non-stationarity of natural signals. In *Advances in Neural Information Processing Systems*, volume 13, pages 786–792. MIT Press, Cambridge, 2000.

[PSV98] L. Parra, C. Spence, and B. De Vries. Convolutive blind source separation based on multiple decorrelation. In *Proc. IEEE Int. Workshop Neural Networks for Signal Processing (NNSP)*, pages 23–32, September 1998.

[QBC88] T.P. Quackenbush, T.P. Barnwell, and M.A. Clements. *Objective measures of speech quality.* Prentice-Hall, Englewood Cliffs, NJ, 1988.

[RAS75] W. Reichardt, A. Alim, and W. Schmidt. Definition and Messgrundlage eines objektiven Masses zur Ermittlung der Grenze zwischen brauchbarer and unbrauchbarer Durchsichtigkeit bei Musikdarbietung. *Acustica*, 32:126–137, 1975. In German.

[RHK05] T. Rhodenburg, V. Hohmann, and B. Kollmeier. Objective perceptual quality measures for the evaluation of noise reduction schemes. In *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 169–172, Eindhoven, Netherlands, September 2005.

[Roj96] R. Rojas. *Neural Networks.* Springer, Berlin, 1996.

[RR01] K. Rahbar and J.P. Reilly. Blind source separation of convolved sources by joint approximate diagonalization of cross-spectral density matrices.

In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 5, pages 2745–2748, Salt Lake City, Utah, USA, May 2001.

[RS78] L.R. Rabiner and R.W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, 1978.

[RV89] D. D. Rife and J. Vanderkooy. Transfer-function measurement with maximum-length sequences. *J. Audio Eng. Soc.*, 37(6):419–444, June 1989.

[Sab22] W.C. Sabine. *Collected Papers on Acoustics*. Harvard University Press, Cambridge, 1922.

[SAMM06] H. Sawada, S. Araki, R. Mukai, and S. Makino. Solving the permutation problem of frequency-domain BSS when spatial aliasing occurs with wide sensor spacing. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 5, pages 77–80, Toulouse, France, May 2006.

[SBM01] K. U. Simmer, J. Bitzer, and C. Marro. Post-filtering techniques. In M. Brandstein and D. Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*, pages 39–60. Springer, Berlin, 2001.

[Sch65] M.R. Schroeder. New method of measuring reverberation time. *J. Acoust. Soc. Amer.*, 37(3):409–412, June 1965.

[Sch79] M.R. Schroeder. Integrated-impulse method measuring sound decay without using impulses. *J. Acoust. Soc. Amer.*, 66(2):497–500, August 1979.

[SD01] X. Sun and S. Douglas. A natural gradient convolutive blind source separation algorithm for speech mixtures. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, San Diego, CA, USA, December 2001.

[Sel06] I.W. Selesnick. Laplace random vectors, gaussian noise, and the generalized incomplete gamma function. In *Int. Conf. on Image Processing (ICIP)*, pages 2097–2100, Atlanta, USA, October 2006.

[SG00] N.N. Schraudolph and X. Giannakopoulos. Online independent component analysis with local learning rate adaptation. In S.A. Solla, T.K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 789–795. MIT Press, Cambridge, 2000.

[She85] P. J. Sherman. Circulant approximations of the inverses of Toeplitz matrices and related quantities with applications to stationary random processes. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-33(6):1630–1632, December 1985.

[Shy92] J.J. Shynk. Frequency-domain and multirate adaptive filtering. *IEEE Signal Processing Magazine*, pages 14–37, January 1992.

[SJHB87] H.V. Sorensen, D.L. Jones, M.T. Heideman, and C.S. Burrus. Real-valued fast Fourier transform algorithms. *IEEE Trans. Acoust., Speech, Signal Processing*, 35(6):849–863, June 1987.

[Sko70] M. Skolnik, editor. *Radar Handbook*. McGraw-Hill, 1970.

[Sma98] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22:21–34, 1998.

[SMAM02] H. Sawada, R. Mukai, S. Araki, and S. Makino. Polar coordinate based nonlinear function for frequency-domain blind source separation. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 1, pages 1001–1004, Orlando, FL, USA, May 2002.

[SMAM03] H. Sawada, R. Mukai, S. Araki, and S. Makino. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 505–510, Nara, Japan, April 2003.

[SMAM04] H. Sawada, R. Mukai, S. Araki, and S. Makino. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Trans. Speech Audio Processing*, 12(5):530–538, September 2004.

[SMAM05] H. Sawada, R. Mukai, S. Araki, and S. Makino. Frequency-domain blind source separation without array geometry information. In *Proc. Joint Workshop on Hands-Free Communication and Microphone Arrays*, pages d13–d14, Piscataway, NJ, USA, March 2005.

[SMdlKdR+03] H. Sawada, R. Mukai, S. de la Kethulle de Ryhove, S. Araki, and S. Makino. Spectral smoothing for frequency-domain blind source separation. In *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 311–314, Kyoto, Japan, September 2003.

[SP90] J.-S. Soo and K.K. Pang. Multidelay block frequency domain adaptive filter. *IEEE Trans. Acoust., Speech, Signal Processing*, 38(2):373–376, February 1990.

[Sri05] Sriram Srinivasan. *Knowledge-Based Speech Enhancement*. PhD thesis, Royal Institute of Technology (KTH), October 2005.

[SS89] H.V. Söderström and P. Stoica. *System Identification*. Int. Series in Systems and Control Engineering. Prentice-Hall, Upper Saddle River, NJ, 1989.

[STKR04] S. Spors, H. Teutsch, A. Kuntz, and R. Rabenstein. Sound field synthesis. In J. Benesty and Y. Huang, editors, *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, pages 323–344. Kluwer Academic Publishers, Boston, 2004.

[Teu05] H. Teutsch. *Wavefield Decomposition Using Microphone Arrays and Its Application to Acoustic Scene Analysis*. PhD thesis, Universität Erlangen-Nürnberg, Erlangen, Germany, October 2005.

[TGSB97] V. Turbin, A. Gilloire, P. Scalart, and C. Beaugeant. Using psychoacoustic criteria in acoustic echo cancellation algorithms. In *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 53–56, London, UK, September 1997.

[TLSH91] L. Tong, R.-W. Liu, V.C. Soon, and Y.-F. Huang. Indeterminacy and identifiability of blind identification. *IEEE Trans. on Circuits and Systems*, 38(5):499–509, May 1991.

[Tor96a] K. Torkkola. Blind separation of convolved sources based on information maximization. In *Proc. IEEE Int. Workshop Neural Networks for Signal Processing (NNSP)*, pages 423–432, Kyoto, Japan, September 1996.

[Tor96b] K. Torkkola. Blind separation of delayed sources based on information maximization. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 6, pages 3510–3513, Atlanta, Georgia, 1996.

[Tor99] K. Torkkola. Blind separation for audio signals - are we there yet? In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 239–244, Aussois, France, January 1999.

[VHH98] P. Vary, U. Heute, and W. Hess. *Digitale Sprachsignalverarbeitung*. B.G. Teubner, Stuttgart, 1998. In German.

[VL03] E. Visser and T.-W. Lee. Speech enhancement using blind source separation and two-channel energy based speaker detection. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 1, pages 836–839, HongKong, April 2003.

[VMR+88] T.P. Vogl, J.K. Mangis, A.K. Rigler, W.T. Zink, and D.L. Allcon. Accelerating the convergence of the back propagation method. *Biological Cybernetics*, 59:257–263, 1988.

[VRM04] J.-M. Valin, J. Rouat, and F. Michaud. Microphone array post-filter for separation of simultaneous non-stationary sources. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 1, pages 221–224, Montreal, Canada, May 2004.

[VTDM06] E. Visser, J. Toman, T. Davis, and B. Momeyer. Patent WO2006/028587 A3: Headset for separation of speech signals in a noisy environment, March 2006.

[vVB88] B. van Veen and K. Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE Trans. Acoust., Speech, Signal Processing*, pages 4–24, April 1988.

[WB06] D. Wang and G.J. Brown, editors. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. John Wiley & Sons, 2006.

[Wel67] P.D. Welch. 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.*, AU-15(2):70–73, June 1967.

[WFO93] E. Weinstein, M. Feder, and A. Oppenheim. Multi-channel signal separation by decorrelation. *IEEE Trans. Speech Audio Processing*, 1(4):405–413, October 1993.

[WGM+75] B. Widrow, J. Glover, J. MacCool, J. Kautnitz, C. Williams, R. Hearn, J. Zeidler, E. Dong, and R. Goodlin. Adaptive noise cancelling: principles and applications. *Proc. IEEE*, 63:1692–1716, 1975.

[WL82] D. Wang and J. Lim. The unimportance of phase in speech enhancement. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-30(4):679–681, August 1982.

[WP97] H.-C. Wu and J.C. Principe. A unifying criterion for blind source separation and decorrelation: simultaneous diagonalization of correlation matrices. In *Proc. IEEE Int. Workshop Neural Networks for Signal Processing (NNSP)*, pages 496–505, Amelie Island, Fl, USA, 1997.

[WP99] H.-C. Wu and J.C. Principe. Simultaneous diagonalization in the frequency domain (SDIF) for source separation. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 245–250, Aussois, France, December 1999.

[XLTK95] G. Xu, H. Liu, L. Tong, and T. Kailath. A least-squares approach to blind channel identification. *IEEE Trans. Signal Processing*, 43(12):2982–2993, December 1995.

[YA97] H.H. Yang and S.-I. Amari. Adaptive on-line learning algorithms for blind source separation - maximum entropy and minimum mutual information. *Neural Computation*, 9:1457–1482, 1997.

[Yao73] K. Yao. A representation theorem and its applications to spherically-invariant random processes. *IEEE Trans. Inform. Theory*, 19(5):600–608, September 1973.

[YR04] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. Signal Processing*, 52(7):1830–1847, July 2004.

[ZCA99] L.-Q. Zhang, A. Cichocki, and S.-I. Amari. Geometrical structures of FIR manifold and their application to multichannel blind deconvolution. In *Proc. IEEE Int. Workshop Neural Networks for Signal Processing (NNSP)*, pages 303–312, Madison, WI, USA, August 1999.

[ZP01] M. Zibulevsky and B.A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation*, 13:863–882, 2001.

[ZZ98] M. Zollner and E. Zwicker. *Elektroakustik*. Springer, Berlin, 3rd edition, 1998. In German.

Index