UNIVERSITÄT PADERBORN

FAKULTÄT FÜR
ELEKTROTECHNIK,
INFORMATIK UND
MATHEMATIK

Acoustic Beamforming in the Presence of Finite Sample Size Estimates and Sampling Offsets

Von der Fakultät für Elektrotechnik, Informatik und Mathematik der Universität Paderborn

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften (Dr.-Ing.)

genehmigte Dissertation
von

M.Sc. Tobias Gburrek

Erster Gutachter: Prof. Dr.-Ing. Reinhold Häb-Umbach
Zweiter Gutachter: Prof. Dr.-Ing. Sharon Gannot

Tag der mündlichen Prüfung: 26.02.2026

Paderborn 2026

Diss. EIM-E/398

Abstract

Recording acoustic signals with a set of spatially distributed devices promises great benefits for today's applications, e.g., transcribing what was said in a meeting. However, new challenges also emerge since the sampling processes of the devices' microphone signals are initially unsynchronized. They begin recording at different points in time and sample the microphone signals at slightly different rates. The resulting sampling time offsets (STOs) and sampling rate offsets (SROs) can have significant negative impacts on beamforming.

The thesis at hand analyzes the impacts of STOs and SROs on minimum variance distortionless response (MVDR) beamforming. To this end, a closed-form approximation of the signal-to-distortion ratio (SDR) at the beamformer output is derived that is based on a statistical model for the extraction of a target speaker's signal from a noisy and reverberant speech mixture. This closed-form approximation models the relationship between the size of the interval that is used to estimate the spatial covariance matrices (SCMs), necessary for calculation of the beamformer coefficients, and the SDR at the beamformer output. The analysis of the closed-form approximation of the SDR in combination with an experimental investigation reveals a trade-off between using a short SCM estimation interval to mitigate the negative effects of SROs and using a long SCM estimation interval to improve the quality of the SCM estimates. While this trade-off can be used to keep the drop in performance of the beamformer at an acceptable level by a careful choice of the size of the SCM estimation interval for small SROs, results show that large SROs always have to be compensated for to prevent a significant degradation of the beamforming performance. Further, it is shown that a minimization of the STOs by a typically used coarse time alignment of the signals is accurate enough for acoustic beamforming as downstream task.

Finally, a method for STO estimation is presented, which enables a synchronization that maintains the physical time differences of flight (TDOFs). Furthermore, an SRO estimator for dynamic scenarios with time-varying SROs as well as speaker changes is developed. Simulations show that the developed SRO estimator is able to outperform comparable SRO estimators that were designed for scenarios with time-constant SROs and without speaker changes.

Zusammenfassung

Die Aufzeichnung akustischer Signale mit räumlich verteilten Geräten verspricht große Vorteile für heutige Anwendungen, z.B. der Transkription dessen, was in einem Meeting gesagt wurde. Allerdings ergeben sich auch neue Herausforderungen, da die Abtastprozesse der Mikrofonsignale der Geräte üblicherweise zunächst nicht synchronisiert sind. Hierdurch beginnen die Geräte zu unterschiedlichen Zeitpunkten mit der Aufzeichnung und tasten die Mikrofonsignale mit leicht unterschiedlichen Raten ab. Die daraus resultierenden Versätze des Abtastzeitpunktes (engl. sampling time offsets (STOs)) und Abtastratenversätze (engl. sampling rate offsets (SROs)) können erhebliche negative Auswirkungen auf Beamforming haben.

Die vorliegende Arbeit untersucht die Auswirkungen von STOs und SROs auf minimum variance distortionless response (MVDR) Beamforming. Zu diesem Zweck wird eine geschlossene Näherung für das Signal-zu-Verzerrungs-Verhältnis (engl. signal-to-distortion ratio (SDR)) am Ausgang des Beamformers hergeleitet, die auf einem statistischen Modell für die Extraktion des Signals eines Zielsprechers aus einem verrauschten und hallbehafteten Sprachgemisch basiert. Diese geschlossene Näherung beschreibt unter anderem den Zusammenhang zwischen der Länge des Intervalls, das zur Schätzung der für die Berechnung der Beamformer-Koeffizienten erforderlichen räumlichen Kovarianzmatrizen (engl. spatial covariance matrices (SCMs)) verwendet wird, und dem SDR am Beamformer-Ausgang. Die Analyse der geschlossenen Näherung des SDRs in Kombination mit einer experimentellen Untersuchung zeigt, dass ein Kompromiss zwischen der Verwendung eines kurzen SCM-Schätzintervalls, um die negativen Effekte der SROs zu verringern, und der Verwendung eines langen SCM-Schätzintervalls, um die Qualität der SCM-Schätzungen zu verbessern, besteht. Während sich dieser Kompromiss bei kleinen SROs nutzen lässt, um durch eine sorgfältige Wahl der Länge des SCM-Schätzintervalls den Leistungsabfall des Beamformers auf einem akzeptablen Niveau zu halten, zeigen die Ergebnisse, dass große SROs stets kompensiert werden müssen, um eine signifikante Verschlechterung der Leistung des Beamformers zu vermeiden. Darüber hinaus wird gezeigt, dass eine Minimierung der STOs durch einen üblicherweise verwendeten groben Zeitabgleich der Signale hinreichend genau für akustisches Beamforming als nachfolgende Aufgabe ist.

Schließlich wird eine Methode zur STO-Schätzung vorgestellt, die eine Synchronisation ermöglicht, welche die physikalischen Laufzeitdifferenzen (engl. time differences of flight (TDOFs)) beibehält. Außerdem wird ein SRO-Schätzer für dynamische Szenarien mit zeitlich variierenden SROs sowie Sprecherwechseln entwickelt. Simulationen zeigen, dass der entwickelte SRO-Schätzer bessere Ergebnisse als vergleichbare SRO-Schätzer, die für Szenarien mit zeitlich konstanten SROs und ohne Sprecherwechsel entworfen wurden, erzielt.

Acknowledgments

First and foremost, I would like to thank my supervisor Prof. Dr.-Ing. Reinhold Haeb-Umbach for his support and his patience during the derivation of the theoretical framework used in this work. The encouragement to explore new topics beyond my original focus has allowed me to gain a comprehensive and multifaceted perspective on my field of research that I might never have developed otherwise. Without his guidance this thesis in its final form would not have been possible.

Further, I want to thank my fellow team members. Without our fruitful discussions, their detailed feedback and their input my work would not have reached its current depth and clarity. A special thanks goes out to Dr.-Ing. Joerg Schmalenstroeer, being the supervisor of the subproject in which I participated during the acoustic sensor network (ASN) project. His guidance, the opportunity to build on his original work and just the right amount of constructive criticism helped me to grow a lot as a researcher.

I would also like to thank the German Research Foundation (DFG) for funding the ASN project. Moreover, my thanks go out to the members of the ASN project for giving me a broad interdisciplinary perspective on my field of research.

Finally, a big thank you goes out to my family for all your support and encouragement throughout this journey. I could not have completed this work without you.

Contents

1 Introduction

Today, automatic speech recognition (ASR), i.e., the process of transcribing the (lexical) content of a speech signal, is a key technology for many applications in daily life. One common use case is the interaction with smart home assistants, such as Amazon Alexa or Google Assistant, which for example allow users to control lights and other devices using voice commands. In such situations, the speakers are typically several meters away from the microphones so that other sound emitting sources and reverberation have a significant impact on the recording quality. The resulting noisy and reverberant recording conditions do not only harm the listening experience but also are challenging for ASR. An additional challenge emerges in situations with natural conversations, like meetings, in which multiple speakers might be active concurrently. This makes ASR for such conversations even more difficult.

A common way to remedy the challenging recording conditions is to employ speech enhancement methods before applying an ASR system. In this context, acoustic beamforming, i.e., spatial filtering, has been shown to deliver decent results for extracting a target speaker's signal from a noisy and reverberant speech mixture. Acoustic beamforming is a technique that emphasizes the signal of a target source emitted at a particular position by combining the signals from multiple microphones while suppressing the signals of other sources at different positions. In a simplified view, beamforming is based on the fact that signals emitted at different positions arrive at the microphones with different time delays. These time delays are also referred to as time differences of flight (TDOFs). Often, compact microphone arrays, e.g., being a part of smart speakers or video conferencing systems, are employed for acoustic beamforming. Alternatively, it could be utilized that many devices, such as smartphones, tablets and laptops, are equipped with at least one microphone. By combining multiple of such devices, a distributed microphone array can be realized, that exhibits a more diverse spatial distribution of the microphones than a single compact device. A possible application scenario for this could be meeting recognition, i.e., the process of transcribing what has been said in a meeting, utilizing the signals which are recorded with the devices of the participants.

Recording setups with spatially distributed devices offer the advantage of an improved signal enhancement via acoustic beamforming over compact microphone arrays. For example, signals emitted at different positions arrive at the microphones at significantly different points in time compared to the case with a compact microphone array. However, multiple distributed recording devices also come with the drawback that the recording processes of the different devices are independent of each other, leading to sampling time offsets (STOs) and sampling rate offsets (SROs) between the signals. STOs arise from the fact that the device start their recording process at different points in time. SROs correspond to small differences in the sampling rates at which the microphone signals of the different devices are sampled. In

previous works [1]–[6], it was shown that STOs and SROs can lead to a significant degradation of the performance of an acoustic beamformer. Thus, typically a compensation for STOs and SROs is employed before applying beamforming.

Many beamforming algorithms rely on spatial covariance matrices (SCMs), which capture the statistics of the microphone signals and are usually estimated from the microphone signals. The SCMs model the differences of the point in time at which signals emitted by a source arrive at the microphones among other aspects, like reverberation. On the one hand, it was shown in previous works [7]–[11] that the quality of the SCM estimates and, therefore, the performance of the beamformer improves as the sample size used for SCM estimation, i.e., the length of the observation to obtain the estimates from, grows. On the other hand, the negative impact of SROs on the quality of the SCM estimates needed to calculate the beamforming coefficients becomes stronger as the sample size used for SCM estimation increases [5], since the signals continuously drift apart due to the SROs.

Previous works [5], [7]–[11] investigated the impact of using a finite sample size for SCM estimation on beamforming and the effect of SROs on beamforming only in isolation from each other. The effect of estimating the SCMs based on a finite sample size was only considered for radio frequency (RF) transmissions of signals with time-constant power before. Hence, these results cannot be directly transferred to the problem at hand with speech signals that show a time-varying power. While the degradation of the performance of a beamformer due to SROs has already been studied before [4], [5], it was not set in connection with the sample size utilized for estimating the SCMs.

The central research question addressed in the thesis at hand can be formulated as follows: Does a trade-off exist between choosing a large sample size for SCM estimation in order to mitigate the impact of SCM estimation from a finite sample size on beamforming and choosing a small sample size to diminish the impact of SROs on beamforming? If such a trade-off exists without causing a significant degradation in beamforming performance, the need for SRO estimation and SRO compensation before beamforming could be rendered unnecessary. The question stated above is systematically answered at hand of the extraction of a target speaker's signal from a noisy and reverberant speech mixture via minimum variance distortionless response (MVDR) beamforming. To approach this question, first the effect of SCM estimation from a finite set of samples and the effect of SROs on beamforming are separately analyzed from a theoretical perspective. The aim of this step is to establish a theoretical understanding how a finite sample size and SROs individually affect the quality of the estimated SCMs and how this affects the performance of the beamformer. Then, this thesis investigates the interplay between the effects of estimating the SCMs from a finite sample size on beamforming and the effects SROs on beamforming within a unified theoretical model. This step makes it possible to explore whether and under which conditions there is a trade-off for the choice of the sample size for estimating the SCMs. In a similar way, the impact of STOs on beamforming is discussed and analyzed.

The remainder of this thesis is structured as follows. In Chapter 2, the prerequisites needed to understand this thesis are briefly introduced. Chapter 3 provides an overview of the experimental setup that is used to analyze the effect of using a finite sample size for SCM estimation on beamforming if STOs and SROs are present. Chapter 4 starts with a derivation of a statistical model of SCMs which are estimated based on a finite sample size followed

by a description of the statistics of the MVDR beamformer coefficients that result from these SCM estimates. Based on this, a closed-form approximation of the performance of an MVDR beamformer as a function of the sample size used for SCM estimation is derived and analyzed in the second part of Chapter 4. The advantages of beamforming using distributed devices compared to beamforming using a compact microphone array are demonstrated in Sec. 5.1 at hand of meeting recognition as an example application. Moreover, the effects of STOs and SROs on SCM estimation as well as their influence on the performance of an MVDR beamformer are discussed in Sec. 5.2 and Sec. 5.3. In Chapter 6, first, the closed-form approximation of the performance of an MVDR beamformer is extended to be able to model the case with STOs and SROs being present. Employing the extended closed-form approximation of the performance of an MVDR beamformer, the interplay of the impact of SCM estimation from a finite set of samples on beamforming and the impact of STOs as well as the impact of SROs on beamforming is investigated. Chapter 7 presents an approach to STO estimation that enables to maintain the physical TDOF information after compensating for STOs. In addition to that, an SRO estimator for scenarios with time-varying SROs and speaker position changes is introduced in Chapter 7. Both, the STO estimator and the SRO estimator, are evaluated at the end of Chapter 7. Finally, the main contributions of this thesis are summarized and an outlook on future work is given in Chapter 8.

2 Prerequisites

In this chapter, the prerequisites of the remaining work are introduced. The general signal model, used throughout this work, is introduced in Sec. 2.1. Subsequently, the models for describing the effects of asynchronous sampling of the microphone signals, i.e., sampling rate offsets (SROs) and sampling time offsets (STOs), are presented in Sec. 2.2, before the fundamentals of minimum variance distortionless response (MVDR) beamforming are described in Sec. 2.3. Finally, the Wishart distribution and its properties are explained in Sec. 2.4.

2.1 Signal model

Throughout this work, a meeting-like scenario, i.e., a conversation between Q speakers, which may also contain speech pauses and partially overlapping speech, is considered. If not stated otherwise, the conversation is captured by M spatially distributed devices, which are equipped with a single microphone each. It is assumed that the positions of the speakers and microphones are fixed. Since human speakers emit sound in a limited region of space they can be modeled as point sources. Therefore, the corresponding propagation of the q-th speaker's continuous time signal $z_q(t)$ to the m-th microphone can be modeled via a convolution with the room impulse response (RIR) $h_{q,m}(t)$ [12], where t corresponds to the continuous-time index. This results in

$$y_m(t) = \sum_{q=0}^{Q-1} h_{q,m}(t) * z_q(t) + \nu_m(t) = \sum_{q=0}^{Q-1} x_{q,m}(t) + \nu_m(t) \tag{2.1}$$

for the m-th microphone signal, with $\nu_m(t)$ corresponding to additive noise, e.g., sensor noise or background noise, which is captured by the m-th microphone, $x_{q,m}(t)$ being the source image of the q-th speaker at the m-th microphone and $*$ denoting the convolution operator. Moreover, it is assumed that the noise signals of different microphones are uncorrelated.

The RIR $h_{q,m}(t)$ models the multi-path propagation of sound waves from the position of speaker q to the position of microphone m [12]. This means that the RIR $h_{q,m}(t)$ does not only model the direct path propagation but also reverberant components which are caused by reflections of sound at the room boundaries, i.e., walls, floor and ceiling, and other objects, like furniture. In this context, the reverberation consists of so-called early reflections, which correspond to the first reflection of the sound waves, and so-called late reverberation, which correspond to sound waves that are reflected multiple times [12]. The early reflections typically cause single peaks in the RIR [12], which can be well distinguished from each other.

These reflections contribute to the coherent part of the microphone signals [13], i.e., the portion that preserves consistent phase and amplitude relationships across microphones. In contrast, the late reverberation, which stem from sound waves that are reflected multiple times and generally cannot be distinguished from each other, results in an exponentially decaying reverberant tail in the RIR [12]. Often, these late reverberations are modeled by a diffuse sound field [13]. The exponent of the exponentially decaying envelope of the RIR's reverberant tail is inversely proportional to the sound decay time T_{60} [14]. Thus, a larger sound decay time comes with a longer RIR. The size of the sound decay time is dependent on the room dimension and the energy absorption coefficients of the surfaces at which the sound is reflected among other factors [15].

Assuming that the sampling processes of all different microphone signals are synchronized, i.e., all devices start recording at the same point in time and sample their microphone signals with the nominal sampling rate f_s, the discrete-time signal of the m-th microphone results from sampling the continuous-time signal $y_m(t)$ at the equidistant points in time n/f_s:

$$y_m(n) = y_m(t)|_{t=\frac{n}{f_\mathrm{s}}} \,. \tag{2.2}$$

Here, n denotes the discrete-time index.

2.1.1 Spectral model

Commonly, audio signals are processed in the short-time Fourier transform (STFT) domain. Using the so-called narrowband approximation [12], the convolution of the source signal $z_q(t)$ with the RIR $h_{q,m}(t)$ in (2.1) can be modeled by a multiplication of the STFT of the source signal $z_q(\ell, k)$ with the acoustic transfer function (ATF) $h_{q,m}(k)$ in the STFT domain. This results in

$$y_m(\ell, k) = \sum_{q=0}^{Q-1} h_{q,m}(k) \cdot z_q(\ell, k) + \nu_m(\ell, k) = \sum_{q=0}^{Q-1} x_{q,m}(\ell, k) + \nu_m(\ell, k) \tag{2.3}$$

for the representation of the m-th microphone signal in the STFT domain, with ℓ denoting the time frame index and k denoting the frequency bin index. The STFT of a signal $y_m(n)$ is calculated via

$$y_m(\ell, k) = \sum_{\widetilde{n}=0}^{N-1} y_m(\widetilde{n} - \ell \cdot B) \cdot \omega(\widetilde{n}) \cdot \exp\left(-\mathrm{j} \cdot \frac{2 \cdot \pi}{N} \cdot k \cdot \widetilde{n}\right), \tag{2.4}$$

where $\omega(n)$ denotes a window function of length N and B the frame advance of the STFT. It is to be mentioned that the narrowband approximation assumes that the length of the time frames N is sufficiently large compared to the length of the RIR [12].

For sake of a more compact notation, a vector notation for the observed microphone signals is introduced by stacking their STFTs. This leads to

$$\mathbf{y}(\ell, k) = \sum_{q=0}^{Q-1} \mathbf{x}_q(\ell, k) + \boldsymbol{\nu}(\ell, k) \tag{2.5}$$

for the observed microphone signals, with the vectors

$$\mathbf{y}(\ell,k)=\begin{bmatrix} y_0(\ell,k) \\ y_1(\ell,k) \\ \vdots \\ y_{M-1}(\ell,k) \end{bmatrix}, \mathbf{x}_q(\ell,k)=\begin{bmatrix} x_{q,0}(\ell,k) \\ x_{q,1}(\ell,k) \\ \vdots \\ x_{q,M-1}(\ell,k) \end{bmatrix} \text{ and } \boldsymbol{\nu}(\ell,k)=\begin{bmatrix} \nu_0(\ell,k) \\ \nu_1(\ell,k) \\ \vdots \\ \nu_{M-1}(\ell,k) \end{bmatrix}. \tag{2.6}$$

2.1.2 Statistical model

One way to overcome the limitations of the narrowband approximation, i.e., the need for sufficiently long time frames in the STFT domain relatively to the length of the RIRs, is to model the microphone signals statistically. As proposed in [12], [16], [17], the microphone signals $\mathbf{y}(\ell,k)$ can be modeled by the local Gaussian model (LGM) [12], [16], [17] based on the second-order moments of the single speaker's source images $\mathbf{x}_q(\ell,k)$ and the noise images $\boldsymbol{\nu}(\ell,k)$. The single time frequency bins of the q-th speaker's source images are modeled as independent and identically distributed (i.i.d.) samples from a zero-mean, circular Gaussian distribution with covariance matrix $\sigma_q^2(\ell,k)\cdot\mathbf{R}_q(k)$:

$$\mathbf{x}_q(\ell,k)\sim\mathcal{N}(\mathbf{0},\sigma_q^2(\ell,k)\cdot\mathbf{R}_q(k)). \tag{2.7}$$

Here, $\sigma_q^2(\ell,k)$ corresponds to the power of the q-th speaker's signal and $\mathbf{R}_q(k)$ to the spatial covariance matrix (SCM) of the q-th speaker's source images. In general, the SCMs $\mathbf{R}_q(k)$ have full rank, especially, if the size of the STFT analysis window is small relative to the sound decay time T_{60}, i.e., the length of the corresponding RIRs. If the narrowband approximation is sufficient, i.e., if the length of the STFT analysis window is sufficiently large compared to the length of the RIRs, the speaker's SCMs have rank one with $\mathbf{R}_q(k)=\mathbf{h}_q(k)\cdot\mathbf{h}_q^{\mathrm{H}}(k)$ and the vector of stacked ATFs $\mathbf{h}_q(k)=[h_{q,0}(k),h_{q,1}(k),\cdots,h_{q,M-1}(k)]^{\mathrm{T}}$, where $(\cdot)^{\mathrm{T}}$ denotes transposition.

The SCMs are hermitian so that $\mathbf{R}_q(k)=\mathbf{R}_q^{\mathrm{H}}(k)$ holds, where $(\cdot)^{\mathrm{H}}$ denotes the hermitian transpose of a vector or matrix. Further, the SCMs are positive semidefinite, i.e., all eigenvalues of the SCMs $\mathbf{R}_q(k)$ are non-negative. It follows that

$$\mathrm{tr}\{\mathbf{R}_q(k)\}\geq 0 \tag{2.8}$$

holds for the trace of $\mathbf{R}_q(k)$ and that

$$\widetilde{\mathbf{a}}^{\mathrm{H}}\cdot\mathbf{R}_q(k)\cdot\widetilde{\mathbf{a}}\geq 0 \tag{2.9}$$

hols for an arbitrary vector $\widetilde{\mathbf{a}}$. Both properties are important for the analysis of MVDR beamforming in Chapter 4.

To complete the statistical signal model, the noise is assumed to be stationary so that its power $\sigma_\nu^2(k)$ is time-invariant. Moreover, it is assumed that the noise signals of different microphones are uncorrelated. The stationary noise signals are modeled by a zero-mean, circular Gaussian distribution with

$$\boldsymbol{\nu}(\ell,k)\sim\mathcal{N}(\mathbf{0},\sigma_\nu^2(k)\cdot\mathbf{R}_\nu(k)) \tag{2.10}$$

and $\mathbf{R}_\nu(k) = \mathbf{I}$, where $\mathbf{I}$ corresponds to the identity matrix. With (2.7) and (2.10), it follows that the time frequency bins of the microphone signals' STFT $\mathbf{y}(\ell, k)$ correspond to i.i.d. samples from a zero-mean Gaussian distribution with

$$\mathbf{y}(\ell, k) \sim \mathcal{N}\left(\mathbf{0}, \sum_{q=0}^{Q-1} \sigma_q^2(\ell, k) \cdot \mathbf{R}_q(k) + \sigma_\nu^2(k) \cdot \mathbf{R}_\nu(k)\right). \tag{2.11}$$

2.2 Asynchronous sampling

The hardware of the different devices of a set of distributed recording devices usually is independent from each other so that the sampling processes of the different microphone signals are independent from each other, too. Therefore, the devices generally start recording the signals at different points in time, which leads to STOs between the signals. Moreover, the sampling rate of the different devices typically deviate slightly from the nominal sampling frequency f_s. This results in SROs between the signals. Although STOs and SROs occur at the same time, their effects will be first considered separately in the following.

First, the effect of STOs is considered. Therefore, it is assumed that the recording process of the m-th microphone signal does not start at the nominal point in time t=0 but rather at t=T_m, i.e., T_m seconds after the nominal start of the recording. Note that this deviation of the start of the recording can correspond to several seconds in real applications. In this case, the relationship between synchronously sampled microphone signal and the microphone signal which is sampled in the presence of an STO is given by

$$y_m^{\mathrm{STO}}(n) = y_m(t)|_{t=\frac{n}{f_\mathrm{s}}+T_m} = y_m(t)|_{t=\frac{n+T_m \cdot f_\mathrm{s}}{f_\mathrm{s}}}. \tag{2.12}$$

By comparing (2.12) and (2.2), it becomes obvious that an STO τ_m^{STO}=$T_m \cdot f_\mathrm{s}$ causes a time shift of τ_m^{STO} w.r.t. the synchronously sampled signal.

The time shift due to an STO is commonly modeled via a multiplication with a phase term in the frequency domain, as shown, e.g., in [1], [OC1]. Thus, it follows that the relationship between synchronously sampled microphone signal $y_m(\ell, k)$ and the microphone signal $y_m^{\mathrm{STO}}(\ell, k)$, which is sampled in the presence of an STO, can be modeled in the STFT domain via

$$y_m^{\mathrm{STO}}(\ell, k) = y_m(\ell, k) \cdot \exp\left(\mathrm{j} \cdot \frac{2 \cdot \pi}{N} \cdot k \cdot \tau_m^{\mathrm{STO}}\right). \tag{2.13}$$

It is to be mentioned that modeling STOs by a multiplication with a phase term in the STFT domain is only exact if the STO τ_m^{STO} is smaller than one sample. Otherwise, cyclic wrap around effects will occur. However, as long as the size of the STOs is much smaller than the size of the analysis window of the STFT, the cyclic wrap around effects have an negligibly small effect. Further, this model is also not able to reflect the fact that different parts of the continuous-time signal are extracted for the same time frame if the STO between two signals is large. Stacking the STFTs of the different channels results in

$$\mathbf{y}_{\mathrm{STO}}(\ell, k) = \mathbf{S}(k) \cdot \mathbf{y}(\ell, k), \tag{2.14}$$

with

$$\mathbf{S}(k) = \operatorname{diag}\left(\begin{bmatrix} \exp\left(\mathrm{j}\cdot\frac{2\cdot\pi}{N}\cdot k\cdot\tau_0^{\mathrm{STO}}\right) \\ \exp\left(\mathrm{j}\cdot\frac{2\cdot\pi}{N}\cdot k\cdot\tau_1^{\mathrm{STO}}\right) \\ \vdots \\ \exp\left(\mathrm{j}\cdot\frac{2\cdot\pi}{N}\cdot k\cdot\tau_{M-1}^{\mathrm{STO}}\right) \end{bmatrix}\right) \in \mathbb{C}^{M\times M} \tag{2.15}$$

summarizing the phase terms modeling the time shifts due to the STOs. The operator $\operatorname{diag}(\cdot)$ corresponds to the construction of a diagonal matrix from a given vector.

Next, the effect of SROs is considered. In the following, it is assumed that the m-th microphone signal is sampled with a time-varying sampling rate of

$$f_m(n) = (1 + \widetilde{\varepsilon}_m(n)) \cdot f_{\mathrm{s}}. \tag{2.16}$$

Here, $\widetilde{\varepsilon}_m(n)$ denotes a time-varying SRO in the order of a few parts per million(ppm) such that $\widetilde{\varepsilon}_m(n) \ll 1$ holds. In [18], SRO values in the range between $-40\,\mathrm{ppm}$ and $416\,\mathrm{ppm}$ were reported, where values exceeding the range $\pm 100\,\mathrm{ppm}$ are rarely observed. Although many works assume a time-invariant value for the SRO, the SRO typically changes over time [18], [19], with the time-varying behavior being dependent on many factors. The SRO is directly connected to the stability of the fundamental frequency of the crystal oscillator which controls the sampling process. This frequency is determined by constant factors like the shape of the crystal and the technique employed for cutting the crystal. The resulting deviations of the sampling frequency from the nominal sampling frequency are sufficiently modeled via time-invariant SROs. However, the fundamental frequency of the crystal oscillator additionally depends on environmental influences, like temperature and supply voltage, which can change over time, so that SROs are time-variant in practice. For instance, SROs may change by several ppm within a few minutes due to a changing temperature of the devices when they are switched on or change from sleeping to processing mode, as reported in [18]. SROs also slowly fluctuate by a few ppm after the start-up phase of the devices, e.g., due to changes of the supply voltage that result from a changing workload of the microprocessor [19].

With (2.16), it follows that the m-th microphone signal, which is sampled in the presence of an SRO, is given by

$$y_m^{\mathrm{SRO}}(n) = y_m(t)|_{t=\sum_{\widetilde{n}=0}^{n-1} 1/f_m(\widetilde{n})} = y_m(t)|_{t=\sum_{\widetilde{n}=0}^{n-1} 1/(f_{\mathrm{s}}\cdot(1+\widetilde{\varepsilon}_m(\widetilde{n})))} = y_m(t)|_{t=t_{\mathrm{SRO}}(n)}\,. \tag{2.17}$$

Utilizing that $\widetilde{\varepsilon}_m(n) \ll 1$ holds, the first-order Taylor-series approximation

$$\frac{1}{1+\widetilde{\varepsilon}_m(n)} \approx 1 - \widetilde{\varepsilon}_m(n) \tag{2.18}$$

can be used to simplify the formula for the sampling instant $t^{\mathrm{SRO}}(n)$ to

$$t_{\mathrm{SRO}}(n) = \sum_{\widetilde{n}=0}^{n-1} \frac{1}{f_{\mathrm{s}}\cdot(1+\widetilde{\varepsilon}(\widetilde{n}))} \approx \frac{n - \sum_{\widetilde{n}=0}^{n-1} \widetilde{\varepsilon}_m\left(\widetilde{n}\right)}{f_{\mathrm{s}}}. \tag{2.19}$$

Thus, an SRO leads to a time shift of $\sum_{\widetilde{n}=0}^{n-1} \widetilde{\varepsilon}_m\left(\widetilde{n}\right)$ w.r.t. the corresponding synchronously sampled microphone signal, which can be seen from comparing (2.17) and (2.2).

In the next step, the effect of SROs on the microphone signals is modeled in the STFT domain. As mentioned above, SROs generally change very slowly. Hence, it can be assumed that SROs are constant for short periods of time. By modeling the SRO $\widetilde{\varepsilon}_m(n)$ of the m-th device as $\varepsilon_m(\ell)$, i.e. as a constant value over the duration of a time frame, the SRO-induced shift can be modeled via a multiplication with a phase term in the STFT domain [2], [3], [OC1], [20], [21]. This results in

$$y_m^{\mathrm{SRO}}(\ell,k) = y_m(\ell,k) \cdot \exp\left(-\mathrm{j} \cdot \frac{2\cdot\pi}{N} \cdot k \cdot \left(\frac{N}{2} \cdot \varepsilon_m\left(0\right) + \sum_{\widetilde{\ell}=1}^{\ell} \varepsilon_m\left(\widetilde{\ell}\right) \cdot B\right)\right) \tag{2.20}$$

for the relationship between the m-th microphone signal $y_m^{\mathrm{SRO}}(\ell,k)$ which is sampled in the presence of an SRO and the synchronously sampled m-th microphone signal $y_m(\ell,k)$ [OC1]. Stacking the STFTs of the different channels, leads to

$$\mathbf{y}_{\mathrm{SRO}}(\ell,k) = \mathbf{E}(\ell,k) \cdot \mathbf{y}(\ell,k) \tag{2.21}$$

for the relationship between the observed microphone signals in the presence of SROs $\mathbf{y}_{\mathrm{SRO}}(\ell,k)$ and the synchronously sampled microphone signals $\mathbf{y}(\ell,k)$. The matrix

$$\mathbf{E}(\ell,k) = \mathrm{diag}\left(\begin{bmatrix} \exp\left(-\mathrm{j}\cdot\frac{2\cdot\pi}{N}\cdot k\cdot\left(\frac{N}{2}\cdot\varepsilon_0\left(0\right) + \sum_{\widetilde{\ell}=1}^{\ell}\varepsilon_0\left(\widetilde{\ell}\right)\cdot B\right)\right) \\ \exp\left(-\mathrm{j}\cdot\frac{2\cdot\pi}{N}\cdot k\cdot\left(\frac{N}{2}\cdot\varepsilon_1\left(0\right) + \sum_{\widetilde{\ell}=1}^{\ell}\varepsilon_1\left(\widetilde{\ell}\right)\cdot B\right)\right) \\ \vdots \\ \exp\left(-\mathrm{j}\cdot\frac{2\cdot\pi}{N}\cdot k\cdot\left(\frac{N}{2}\cdot\varepsilon_{M-1}\left(0\right) + \sum_{\widetilde{\ell}=1}^{\ell}\varepsilon_{M-1}\left(\widetilde{\ell}\right)\cdot B\right)\right) \end{bmatrix}\right) \in \mathbb{C}^{M\times M} \tag{2.22}$$

summarizes the phase terms used to model the SROs-induced time shifts.

Assuming a time-invariant SRO ε_m, (2.20) simplifies to

$$y_m^{\mathrm{SRO}}(\ell,k) = y_m(\ell,k) \cdot \exp\left(-\mathrm{j} \cdot \frac{2\cdot\pi}{N} \cdot k \cdot \left(\frac{N}{2} + \ell \cdot B\right) \cdot \varepsilon_m\right). \tag{2.23}$$

This corresponds to the model of the effects of an SRO that usually is used in literature [2], [3], [20], [21] and leads to a simplified matrix

$$\mathbf{E}(\ell,k) = \mathrm{diag}\left(\begin{bmatrix} \exp\left(-\mathrm{j}\cdot\frac{2\cdot\pi}{N}\cdot k\cdot\left(\frac{N}{2}+\ell\cdot B\right)\cdot\varepsilon_0\right) \\ \exp\left(-\mathrm{j}\cdot\frac{2\cdot\pi}{N}\cdot k\cdot\left(\frac{N}{2}+\ell\cdot B\right)\cdot\varepsilon_1\right) \\ \vdots \\ \exp\left(-\mathrm{j}\cdot\frac{2\cdot\pi}{N}\cdot k\cdot\left(\frac{N}{2}+\ell\cdot B\right)\cdot\varepsilon_{M-1}\right) \end{bmatrix}\right). \tag{2.24}$$

In [OC1] it was proposed to model the time-varying behavior of an SRO by an Ornstein-Uhlenbeck process [22]. An Ornstein-Uhlenbeck process is a stochastic process characterized

by a continuous tendency to return to a long-term mean, combined with random fluctuations. This property makes a Ornstein-Uhlenbeck process well suited to represent both transient changes from an initial value toward a steady state and the stationary fluctuations around the long-term mean in the steady state. Moreover, it enables a smooth transition from transient dynamics to steady-state behavior. Consequently, a Ornstein-Uhlenbeck process can model transient changes of an SRO, e.g., after switching on the device, as well as fluctuations of an SRO in the steady state after all transient changes are completed.

As proposed in [OC1], the Ornstein-Uhlenbeck process can be realized via the discrete-time Euler-Maruyama approximation [23]. This results in the auto-regressive process

$$\varepsilon_m(\ell) = \varepsilon_m(\ell-1) + \theta \cdot \left(\mu_m^{(\infty)} - \varepsilon_m(\ell-1)\right) + \iota_\varepsilon(\ell), \tag{2.25}$$

that describes the time-varying behavior of the SRO $\varepsilon_m(\ell)$ in the STFT domain. Here, $\theta \ll 1$ denotes a smoothing factor and $\iota_\varepsilon(n) \sim \mathcal{N}(0; \sigma_{\mathrm{OU}}^2)$ a sample which is drawn from a zero-mean Gaussian distribution with variance σ_{OU}^2. Further, $\mu_m^{(\infty)}$ corresponds to the long-term expected value of the SRO $\varepsilon_m(\ell)$ toward which the SRO converges.

In order to model a transient behavior of the SRO $\varepsilon_m(\ell)$, $\varepsilon_m(0) = \mu_m^{(\infty)} + \Delta_m^{\mathrm{start}}$ is chosen for the start value of the SRO trajectory. Typical values for $\Delta_m^{\mathrm{start}}$ lie in the range of ± 10 ppm [OC1]. It is to be mentioned that the stability of the sampling frequency and, therefore, the stability of the SRO are strongly dependent on the utilized hardware. For instance, if a temperature-compensated crystal oscillator (TXCO) is employed to drive the analog-digital converter (ADC), transient effects due to temperature changes are diminished and mainly the steady state fluctuation remain. For example, for a TXCO of type SiT5156, the SRO varies in a range of ± 2.5 ppm while the range of temperature-dependent variations of the SRO is limited to ± 0.5 ppm [24].

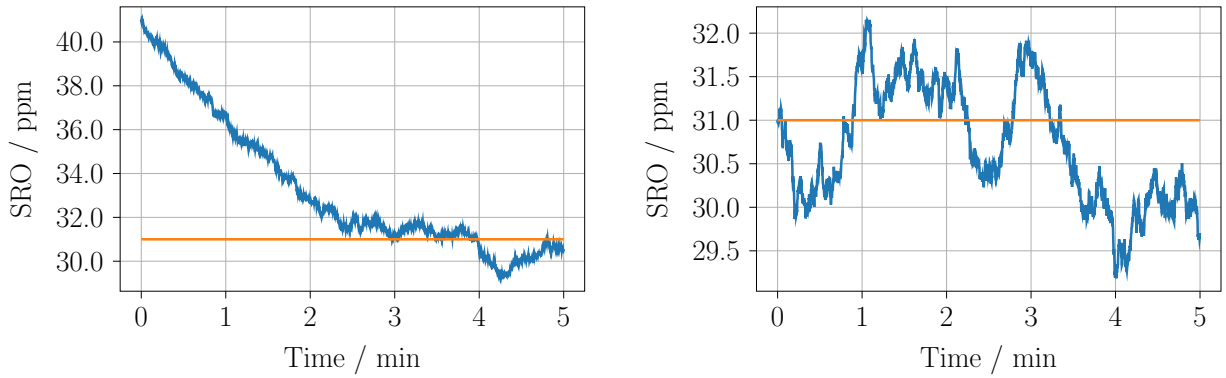


Figure 2.1: Representative SRO trajectories under (a) transient conditions ($\mu_m^{(\infty)}$=31 ppm and $\Delta_m^{\mathrm{start}}$=0 ppm) and (b) steady-state conditions ($\mu_m^{(\infty)}$=31 ppm and $\Delta_m^{\mathrm{start}}$=10 ppm) (*Adapted from* [OC1])

In Fig. 2.1 two representative SRO trajectories, one for the transient case and one for the steady-state case, are shown. Both SRO trajectories were generated using the autoregressive process in (2.25). While the SRO only fluctuates around its expected value in the steady-state

case (see Fig. 2.1(b)), it automatically transits from the start value to the steady-state case, where it remains, in the transient case (see Fig. 2.1(a)).

The relationship between the asynchronously sampled m-th microphone signal $y_m^{\mathrm{ASYNC}}(\ell, k)$, i.e., the microphone signal which is sampled in the presence of STOs and SROs, and the synchronously sampled m-th microphone signal $y_m(\ell, k)$ results from combining (2.13) and (2.20) in

$$y_m^{\mathrm{ASYNC}}(\ell, k) = y_m(\ell, k) \cdot \exp\left(-\mathrm{j} \cdot \frac{2 \cdot \pi \cdot k}{N} \cdot \left(-\tau_m^{\mathrm{STO}} + \frac{N}{2}\varepsilon_m(0) + \sum_{\widetilde{\ell}=1}^{\ell} \varepsilon_m\left(\widetilde{\ell}\right) \cdot B\right)\right). \tag{2.26}$$

Usually, STOs and SROs are not estimated w.r.t. the nominal start of the sampling process and the nominal sampling rate but they are rather estimated relative to the start of the sampling process and the sampling rate of a reference channel. Therefore, the STO and SRO of the j-th channel w.r.t. the i-th channel as reference are introduced which are defined as

$$\tau_{ij}^{\mathrm{STO}} = \tau_j^{\mathrm{STO}} - \tau_i^{\mathrm{STO}} \tag{2.27}$$

and

$$\varepsilon_{ij}(\ell) = \varepsilon_j(\ell) - \varepsilon_i(\ell). \tag{2.28}$$

The overall time difference of arrival (TDOA) $\tau_{ij}(\ell)$ between the i-th channel and the j-th channel, i.e., the time shift between the i-th channel and the j-th channel, corresponds to a superposition of the time shifts which are caused by an STO and an SRO between both channels and the time difference of flight (TDOF). Thereby, the TDOF is defined as the difference in time a signal needs to propagate from the corresponding source position to the different microphone positions w.r.t. the direct path component. Assuming that the q-th speaker is the only active speaker, so that there are no ambiguities, the TDOA between the i-th channel and the j-th channel is given by

$$\tau_{ij}(\ell) = \tau_{q,ij}^{\mathrm{TOF}} - \tau_{ij}^{\mathrm{STO}} + \frac{N}{2}\varepsilon_{ij}(0) + \sum_{\widetilde{\ell}=1}^{\ell} \varepsilon_{ij}\left(\widetilde{\ell}\right) \cdot B. \tag{2.29}$$

Here, $\tau_{q,ij}^{\mathrm{TOF}}$ denotes the TDOF of the q-th speaker's signal between the i-th channel and the j-th channel which is given by

$$\tau_{q,ij}^{\mathrm{TOF}} = \frac{r_{q,j} - r_{q,i}}{c}, \tag{2.30}$$

with c denoting the speed of sound and $r_{q,m}$ corresponding to the distance between the q-th speaker's position and the position of the m-th microphone.

2.3 Minimum variance distortionless response beamforming

In the following, the extraction of a target speaker's signal from the noisy and reverberant speech mixture, which is specified in (2.1), by employing MVDR beamforming is explained. Without loss of generality, speaker 0 is the considered as target speaker in the following. Based on this choice, the microphone signals

$$\mathbf{y}(\ell,k) = \underbrace{\mathbf{x}_0(\ell,k)}_{:=\mathbf{x}_{\text{tar}}(\ell,k)} + \underbrace{\sum_{q=1}^{Q-1} \mathbf{x}_q(\ell,k) + \boldsymbol{\nu}(\ell,k)}_{:=\mathbf{x}_{\text{int}}(\ell,k)} \tag{2.31}$$

can be decomposed into a target component $\mathbf{x}_{\text{tar}}(\ell,k)$ which correspond to the 0-th speaker's source images $\mathbf{x}_0(\ell,k)$ and a interference component $\mathbf{x}_{\text{int}}(\ell,k)$ which consists of the source images of all other speakers and the noise signals.

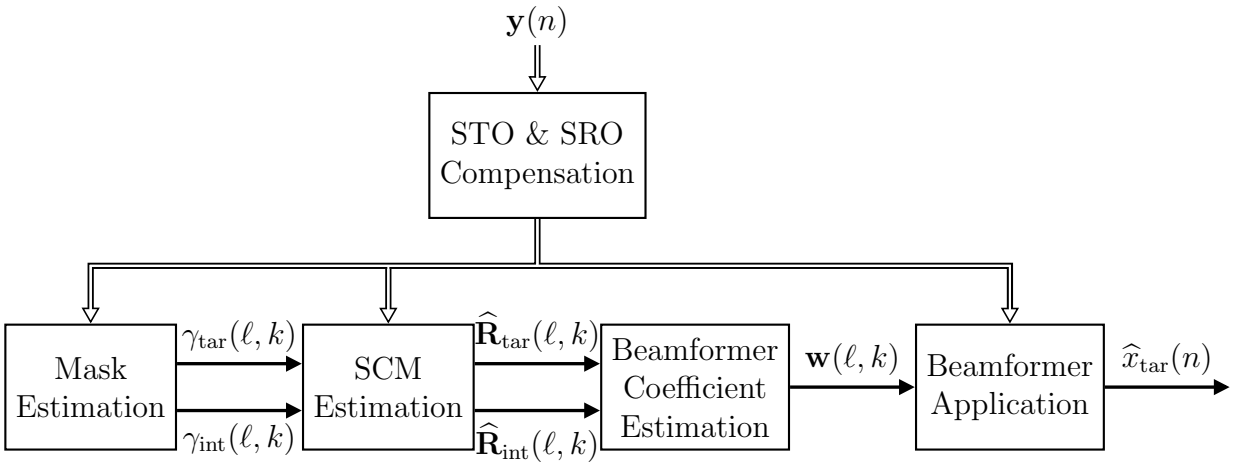


Figure 2.2: Overview of the beamforming system used to extract a target signal from a signal mixture. Multi-channel audio signals are indicated by double arrows.

The block diagram of the general structure of the beamforming system used to extract a target signal from a signal mixture throughout this work is shown in Fig. 2.2. First, the microphone signals are synchronized by compensating for STOs and SROs if spatially distributed devices are used for signal capture. If the speech mixture is recorded using a compact microphone array, this step can usually be omitted because the microphone signals are sampled using the same ADC. Refer to [25] for a hardware example. In addition, later chapters of this work analyze the impact of omitting the SRO and STO compensation also for a distributed setup. Afterwards, masks for the target speaker and the interference are estimated based on multi-channel or single-channel information. Based on these masks, SCMs are estimated for the target speaker and the interference from which the beamformer coefficients are derived. These beamformer coefficients are applied to the observed microphone signals to extract the target signal. In the following, the single steps are explained in more detail. The synchronization is assumed as already performed beforehand and is explained in

detail in Chapter 7. Furthermore, details on mask estimation are beyond the scope of this work and will not be discussed further.

Let $\mathbf{w}(\ell, k)$ denote the coefficients of an MVDR beamformer for the extraction of the target signal. The MVDR beamformer is designed to minimize the variance of the contribution of the interference at the beamformer output while a constraint ensures that the response to signals from a certain source position, which is defined by the steering vector $\mathbf{d}(k)$, is one. This results the following constrained optimization problem [26]:

$$\mathbf{w}(\ell, k) = \underset{\breve{\mathbf{w}}}{\operatorname{argmin}}\ \breve{\mathbf{w}}^{\mathrm{H}} \cdot \mathbf{R}_{\mathrm{int}}(\ell, k) \cdot \breve{\mathbf{w}} \text{ subject to } \breve{\mathbf{w}}^{\mathrm{H}} \cdot \mathbf{d}(k) = 1, \tag{2.32}$$

where $\mathbf{R}_{\mathrm{int}}(\ell, k)$ corresponds to the SCM of the interference. Note that the beamformer coefficients $\mathbf{w}(\ell, k)$ are generally treated to be time-dependent. However, the beamformer coefficients often are assumed to be fixed for blocks of time frames or whole utterances, based on the assumption of spatial stationarity of the sources, as for example in [27].

In order to extract the target signal, the beamformer coefficients are applied to the observed microphone signals for each frequency bin in the following way:

$$\widehat{x}_{\mathrm{tar}}(\ell, k) = \mathbf{w}^{\mathrm{H}}(\ell, k) \cdot \mathbf{y}(\ell, k)\,. \tag{2.33}$$

Beamforming is a linear operation. Therefore, the beamformer coefficients can be separately applied to the single source images and the output of the beamformer

$$\begin{aligned}\widehat{x}_{\mathrm{tar}}(\ell, k) &= \mathbf{w}^{\mathrm{H}}(\ell, k) \cdot \left(\sum_{q=0}^{Q-1} \mathbf{x}_q(\ell, k) + \boldsymbol{\nu}(\ell, k) \right) \\ &= \underbrace{\mathbf{w}^{\mathrm{H}}(\ell, k) \cdot \mathbf{x}_0(\ell, k)}_{=\mathbf{w}^{\mathrm{H}}(\ell,k)\cdot\mathbf{x}_{\mathrm{tar}}(\ell,k)} + \underbrace{\sum_{q=1}^{Q-1} \mathbf{w}^{\mathrm{H}}(\ell, k) \cdot \mathbf{x}_q(\ell, k) + \mathbf{w}^{\mathrm{H}}(\ell, k) \cdot \boldsymbol{\nu}(\ell, k)}_{=\mathbf{w}^{\mathrm{H}}(\ell,k)\cdot\mathbf{x}_{\mathrm{int}}(\ell,k)}\end{aligned} \tag{2.34}$$

can be decomposed into a contribution of the target speaker's signal $\mathbf{w}^{\mathrm{H}}(\ell, k) \cdot \mathbf{x}_{\mathrm{tar}}(\ell, k)$ and the contribution of the interference signals $\mathbf{w}^{\mathrm{H}}(\ell, k) \cdot \mathbf{x}_{\mathrm{int}}(\ell, k)$.

A key factor for the calculation of the MVDR beamformer coefficients are the SCMs of the target signal, to be extracted, and SCMs of the interference. The corresponding ground-truth SCMs follow from the LGM of a speech signal' STFT. From this, it follows that the SCMs of the target signal are given by

$$\mathbf{R}_{\mathrm{tar}}(\ell, k) = \mathbb{E}\big[\mathbf{x}_{\mathrm{tar}}(\ell, k) \cdot \mathbf{x}_{\mathrm{tar}}^{\mathrm{H}}(\ell, k)\big] = \sigma_0^2(\ell, k) \cdot \mathbf{R}_0(k), \tag{2.35}$$

with $\mathbb{E}[\cdot]$ corresponding to the expectation operator. The SCM of the interference is given by

$$\begin{aligned}\mathbf{R}_{\text{int}}(\ell,k) &= \mathbb{E}\big[\mathbf{x}_{\text{int}}(\ell,k)\cdot\mathbf{x}_{\text{int}}^{\text{H}}(\ell,k)\big]\\ &= \mathbb{E}\left[\left(\sum_{q=1}^{Q-1}\mathbf{x}_q(\ell,k)+\boldsymbol{\nu}(\ell,k)\right)\cdot\left(\sum_{q=1}^{Q-1}\mathbf{x}_q(\ell,k)+\boldsymbol{\nu}(\ell,k)\right)^{\text{H}}\right]\\ &= \sum_{q=1}^{Q-1}\mathbb{E}\big[\mathbf{x}_q(\ell,k)\cdot\mathbf{x}_q^{\text{H}}(\ell,k)\big]+\mathbb{E}\big[\boldsymbol{\nu}(\ell,k)\cdot\boldsymbol{\nu}^{\text{H}}(\ell,k)\big]\\ &= \sum_{q=1}^{Q-1}\sigma_q^2(\ell,k)\cdot\mathbf{R}_q(k)+\sigma_\nu^2(k)\cdot\mathbf{R}_\nu(k),\end{aligned} \tag{2.36}$$

where pairwise uncorrelated interference sources are assumed.

Usually, the steering vector and the interference SCM are unknown and, therefore, are replaced by estimates $\widehat{\mathbf{d}}(k)$ and $\widehat{\mathbf{R}}_{\text{int}}(k)$. Note that the dependence of the interference-SCM estimate on the time frame index ℓ is temporarily omitted since there are different ways to estimate the interference SCM based on different lengths of the estimation interval and fixed source positions are considered here. Solving the constrained optimization problem in (2.32) after replacing the steering vector and the interference SCM by their estimates, results in the MVDR beamformer coefficients

$$\mathbf{w}(k)=\frac{\widehat{\mathbf{R}}_{\text{int}}^{-1}(k)\cdot\widehat{\mathbf{d}}(k)}{\widehat{\mathbf{d}}^{\text{H}}(k)\cdot\widehat{\mathbf{R}}_{\text{int}}^{-1}(k)\cdot\widehat{\mathbf{d}}(k)}, \tag{2.37}$$

where $(\cdot)^{-1}$ corresponds to the inverse of a matrix.

In this work, the MVDR beamformer in the formulation that was proposed in [28] is utilized. This version of the MVDR beamformer is also called Souden-MVDR beamformer in the following. The Souden-MVDR beamformer follows from a slightly modified version of the optimization problem in (2.32):

$$\mathbf{w}(\ell,k)=\min_{\breve{\mathbf{w}}}\ \breve{\mathbf{w}}^{\text{H}}\cdot\mathbf{R}_{\text{int}}(\ell,k)\cdot\breve{\mathbf{w}}\ \text{subject to}\ \ \breve{\mathbf{w}}^{\text{H}}\cdot\mathbf{d}(k)=d_0(k). \tag{2.38}$$

Here, the response to signals from a certain source position, which is defined by the steering vector $\mathbf{d}(k)$, is not constrained to be one but to correspond to the element of the steering vector which belongs to the reference channel used for beamforming. Without loss of generality, the 0-th channel is chosen as reference channel. Solving the constrained optimization problem in (2.38) after replacing the steering vector and the interference SCM by their estimates, leads to

$$\mathbf{w}(k)=\widehat{d}_0^{*}(k)\cdot\frac{\widehat{\mathbf{R}}_{\text{int}}^{-1}(k)\cdot\widehat{\mathbf{d}}(k)}{\widehat{\mathbf{d}}^{\text{H}}(k)\cdot\widehat{\mathbf{R}}_{\text{int}}^{-1}(k)\cdot\widehat{\mathbf{d}}(k)}, \tag{2.39}$$

with $(\cdot)^*$ denoting complex conjugation. As demonstrated in [28], (2.39) can be reformulated as

$$\mathbf{w}(k)=\frac{\widehat{\mathbf{R}}_{\text{int}}^{-1}(k)\cdot\widehat{\mathbf{R}}_{\text{tar}}(k)}{\text{tr}\Big\{\widehat{\mathbf{R}}_{\text{int}}^{-1}(k)\cdot\widehat{\mathbf{R}}_{\text{tar}}(k)\Big\}}\cdot\mathbf{u} \tag{2.40}$$

by employing the rank-1 target-SCM estimate $\widehat{\mathbf{R}}_{\text{tar}}(k) = \widehat{\mathbf{d}}(k) \cdot \widehat{\mathbf{d}}^{\text{H}}(k)$, where $\mathbf{u}$ is a one-hot vector indicating the reference channel. As mentioned in [28], the dependence of the beamformer coefficients on the rank-1 assumption for the target-SCM estimate can be avoided via the formulation of the MVDR beamformer in (2.40). The Souden-MVDR beamformer and the classical steering vector based MVDR beamformer coincide if a rank-1 target-SCM estimate is utilized for the Souden-MVDR beamformer. In this case, both versions of the MVDR beamformer differ only in a multiplication with a complex-valued scalar $\widehat{d}_0^*(k)$.

Utilizing the approximate W-disjoint orthogonality of speech [29], i.e., the fact the STFTs of different speakers only show marginal overlap of activity, mask-based SCM estimation can be used to estimate the SCMs that are required to calculate the beamformer coefficients. The target and interference SCMs can be estimated based on target mask estimate $\gamma_{\text{tar}}(\ell, k)$, indicating the time frequency bins which are dominated by the target signal, and an interference mask estimate $\gamma_{\text{int}}(\ell, k)$, indicating the time frequency bins which are dominated by the interference [30]. Given an estimation interval with a length of L time frames, the target and interference SCMs are estimated via

$$\widehat{\mathbf{R}}_{\text{tar}}(k) = \frac{1}{\sum_{\ell=0}^{L-1} \gamma_{\text{tar}}(\ell, k)} \cdot \sum_{\ell=0}^{L-1} \gamma_{\text{tar}}(\ell, k) \cdot \mathbf{y}(\ell, k) \cdot \mathbf{y}^{\text{H}}(\ell, k) \tag{2.41}$$

and

$$\widehat{\mathbf{R}}_{\text{int}}(k) = \frac{1}{\sum_{\ell=0}^{L-1} \gamma_{\text{int}}(\ell, k)} \cdot \sum_{\ell=0}^{L-1} \gamma_{\text{int}}(\ell, k) \cdot \mathbf{y}(\ell, k) \cdot \mathbf{y}^{\text{H}}(\ell, k)\,, \tag{2.42}$$

as described in [30].

Based on the narrowband approximation in (2.3), it was proposed in [31] to approximate the target-SCM estimate by a rank-1 matrix. In this way, the beamforming performance was improved. As proposed in [31], the target SCM is given by the outer product of the steering vector estimate

$$\widehat{\mathbf{R}}_{\text{tar,r1}}(k) = \widehat{\mathbf{d}}(k) \cdot \widehat{\mathbf{d}}^{\text{H}}(k) \tag{2.43}$$

either with an eigenvalue decomposition (EVD) based estimate [31, Eq. (26)]

$$\widehat{\mathbf{d}}(k) = \widehat{\mathbf{d}}_{\text{EV}}(k) = \mathcal{P}\Big(\widehat{\mathbf{R}}_{\text{tar}}(k)\Big) \tag{2.44}$$

or with a generalized eigenvalue decomposition (GEVD) based estimate [31, Eq. (27)]

$$\widehat{\mathbf{d}}(k) = \widehat{\mathbf{d}}_{\text{GEV}}(k) = \widehat{\mathbf{R}}_{\text{int}}(k) \cdot \mathcal{P}\Big(\widehat{\mathbf{R}}_{\text{int}}^{-1}(k) \cdot \widehat{\mathbf{R}}_{\text{tar}}(k)\Big)\,. \tag{2.45}$$

Here, $\mathcal{P}(\cdot)$ denotes the principal eigenvector of a matrix, i.e., the eigenvector belonging to the largest eigenvalue. Note that in this work, $\widehat{\mathbf{d}}(k)$ is called steering vector since it takes the role of the classical steering vector in (2.39). Alternatively, external knowledge, like TDOA estimates, can be used to estimate the steering vector under the assumption of anechoic

signal propagation [OC2]. In this case, the elements of the steering vector correspond to complex-valued phase terms modeling the time shift of the channels relative to a reference channel.

For the analysis of MVDR beamforming in the presence of finite sample size SCM estimates in Chapter 4 and the interplay of SROs and SCM estimation using a finite sample size in Chapter 6, block-wise beamforming is considered to account for the increasing SRO-induced time shift between the different channels. Block-wise beamforming, as used here, shows a large similarity to block-online beamforming as used, e.g., in [32]. However, the recursive update of the SCM estimates is omitted here due to the SROs. For the b-th block which belongs to the set of time frames $\mathcal{B}(b)=\{b \cdot L, \ldots, (b+1) \cdot L\}$, the output of the beamformer is calculated via

$$\widehat{x}_{\text{tar}}(\ell, k) = \mathbf{w}^{\text{H}}(b, k) \cdot \mathbf{y}(\ell, k) \;\; \forall \ell \in \mathcal{B}(b). \tag{2.46}$$

The beamformer coefficients are estimated for each block via

$$\mathbf{w}(b, k) = \frac{\widehat{\mathbf{R}}_{\text{int}}^{-1}(b, k) \cdot \widehat{\mathbf{R}}_{\text{tar}}(b, k)}{\text{tr}\Big\{\widehat{\mathbf{R}}_{\text{int}}^{-1}(b, k) \cdot \widehat{\mathbf{R}}_{\text{tar}}(b, k)\Big\}} \cdot \mathbf{u}, \tag{2.47}$$

with the mask-based SCM estimates

$$\widehat{\mathbf{R}}_{\text{tar}}(b, k) = \frac{1}{\sum\limits_{\ell=b\cdot L}^{(b+1)\cdot L-1} \gamma_{\text{tar}}(\ell, k)} \cdot \sum_{\ell=b\cdot L}^{(b+1)\cdot L-1} \gamma_{\text{tar}}(\ell, k) \cdot \mathbf{y}(\ell, k) \cdot \mathbf{y}^{\text{H}}(\ell, k) \tag{2.48}$$

and

$$\widehat{\mathbf{R}}_{\text{int}}(b, k) = \frac{1}{\sum\limits_{\ell=b\cdot L}^{(b+1)\cdot L-1} \gamma_{\text{int}}(\ell, k)} \cdot \sum_{\ell=b\cdot L}^{(b+1)\cdot L-1} \gamma_{\text{int}}(\ell, k) \cdot \mathbf{y}(\ell, k) \cdot \mathbf{y}^{\text{H}}(\ell, k)\,. \tag{2.49}$$

2.4 Wishart distribution

The Wishart distribution plays a central role for the analysis of the MVDR beamformer in Chapter 4 and Chapter 6. Given A i.i.d. samples $\boldsymbol{\psi}(a)$ from an D-dimensional zero-mean, circular Gaussian distribution with $\boldsymbol{\psi}(a) \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, the matrix-valued random variable

$$\boldsymbol{\Psi} = \sum_{a=0}^{A-1} \boldsymbol{\psi}(a) \cdot \boldsymbol{\psi}^{\text{H}}(a) \tag{2.50}$$

is Wishart distributed [33], [34]:

$$\boldsymbol{\Psi} \sim \mathcal{W}_D(A, \boldsymbol{\Sigma})\,. \tag{2.51}$$

Here, $\mathcal{W}_D(A, \boldsymbol{\Sigma})$ denotes a Wishart distribution with A degrees of freedom and scale matrix $\boldsymbol{\Sigma}$ which belongs to a $D \times D$-dimensional Wishart matrix.

As shown in [34], the expected value of the Wishart matrix $\mathbf{\Psi}$ is given by

$$\mathbb{E}[\mathbf{\Psi}] = A \cdot \mathbf{\Sigma}. \tag{2.52}$$

The variance of the i-th row and j-th column element of the Wishart matrix $\mathbf{\Psi}$ is given by

$$\mathrm{var}\left(\Psi_{ij}\right) = A \cdot \Sigma_{ii} \cdot \Sigma_{jj}, \tag{2.53}$$

as derived in Appendix A.1.

Considering the ratio of the element-wise variance of the Wishart matrix to the squared absolute value of the expected value of the Wishart matrix

$$\frac{\mathrm{var}\left(\Psi_{ij}\right)}{\left|\mathbb{E}[\Psi_{ij}]\right|^2} = \frac{A \cdot \Sigma_{ii} \cdot \Sigma_{jj}}{\left|A \cdot \Sigma_{ij}\right|^2} = \frac{1}{A} \cdot \frac{\Sigma_{ii} \cdot \Sigma_{jj}}{\left|\Sigma_{ij}\right|^2}, \tag{2.54}$$

it becomes obvious that a Wishart matrix converges towards its expected value as the degrees of freedom A grow.

If $\mathbf{\Psi}$ is Wishart distributed with $\mathbf{\Psi} \sim \mathcal{W}_D(A, \mathbf{\Sigma})$ than its inverse $\mathbf{\Psi}^{-1}$ is inverse Wishart distributed [34] with

$$\mathbf{\Psi}^{-1} \sim \mathcal{W}_D^{-1}\left(A, \mathbf{\Sigma}^{-1}\right). \tag{2.55}$$

Here, $\mathcal{W}_D^{-1}\left(A, \mathbf{\Sigma}^{-1}\right)$ corresponds to an inverse Wishart matrix with A degrees of freedom and scale matrix $\mathbf{\Sigma}^{-1}$. For a detailed overview of the moments of a Wishart matrix and the moments of an inverse Wishart matrix, refer to [34].

3 Experimental setup for the analysis of MVDR beamforming

In the following, the experimental setup which is used for all experiments in Chapter 4 and in Chapter 6 is introduced. After the simulated dataset is presented in Sec. 3.1, the performance metrics are explained in Sec. 3.2. Subsequently, the parameterization used for source extraction via minimum variance distortionless response (MVDR) beamforming is given in Sec. 3.3 and the details on the statistical model for source extraction via MVDR beamforming used for the experiments are specified in Sec. 3.4.

3.1 Dataset

The dataset used for the analysis of MVDR beamforming consists of 100 simulated scenarios with Q=2 speakers. For each scenario, a rectangular conference room was modeled. The length and the width of the room are randomly drawn from the uniform distribution $\mathcal{U}(5\,\mathrm{m}, 7\,\mathrm{m})$. All rooms have a height of 3 m. In each room, a table with a height of 1 m and a length and a width drawn from the uniform distribution $\mathcal{U}(1.5\,\mathrm{m}, 3\,\mathrm{m})$ is placed with a random rotation and a minimum distance of 1 m to the closest wall. The two speakers are positioned around the table at a height of 1.6 m, with their distance to the closest edge of the table

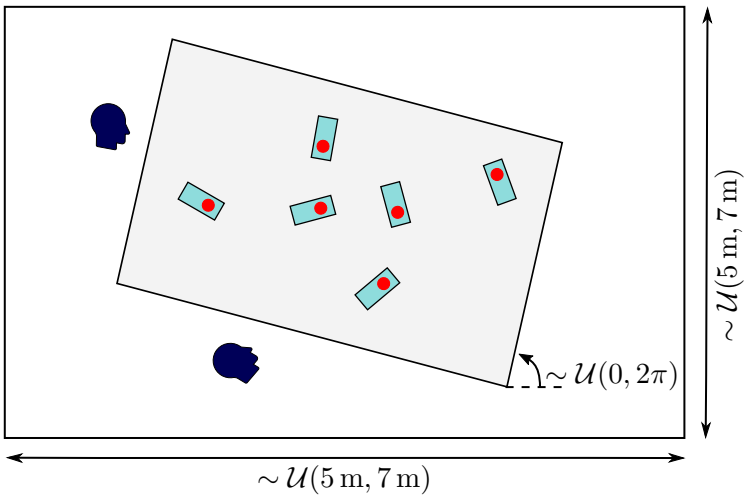


Figure 3.1: Illustration of the scenario utilized to simulate a noisy and reverberant speech mixture used in the analysis of MVDR beamforming. Red dots represent the microphones of the recording devices.

ranging from 0 m to 0.4 m. Furthermore, a minimum distance of 0.5 m between the speakers is guaranteed. The M microphones are randomly placed on the table. If not stated otherwise, M=6 microphones are assumed. Moreover, a minimum distance of 0.3 m of the microphones to the closest edge of the table and 0.5 m distance to the other microphones is guaranteed. If not otherwise specified, the sound decay time of the rooms is randomly drawn from the uniform distribution $\mathcal{U}(0.2\,\mathrm{s}, 0.8\,\mathrm{s})$. Figure 3.1 gives a pictographic illustration of the considered setup.

Fully overlapping speech signals are considered since overlapping speech typically is more challenging for source extraction and automatic speech recognition (ASR) as downstream task than single speaker regions. For the generation of the speech mixture, first, two speakers are drawn from the test-clean subset of the LibriSpeech dataset [35]. Then, utterances are randomly drawn and concatenated for each speaker until the resulting signal has a minimum length of 65 s. Note that the length of the mixtures is given by the longer of the two speakers' signals. The shorter of the two is padded with zeros.

Finally, the speakers' signals are convolved with simulated room impulse responses (RIRs) belonging to the conference room model. The RIRs have a length of 12,800 samples and are simulated via the image source method [36] employing the implementation of [37]. Additionally, additive white Gaussian noise (AWGN) is added with a signal-to-noise ratio (SNR) randomly drawn from the uniform distribution $\mathcal{U}(20\,\mathrm{dB}, 30\,\mathrm{dB})$ w.r.t. the average value of the speech mixtures' powers across the different microphone signals. All signals have a nominal sampling frequency of 16 kHz.

3.2 Performance metrics

As the main measure for the beamforming performance the invasive signal-to-distortion ratio (SDR) will be used in the following, where the invasive SDR quantifies the amount of distortion in the signal at the beamformer output relative to the target signal. This requires access to the target signal for the comparison. The definition, used here, is inspired from the definition proposed in [38] with

$$\mathrm{SDR} = \frac{\sum\limits_{n} \widetilde{x}_{\mathrm{tar}}^2(n)}{\sum\limits_{n} \left(\widehat{x}_{\mathrm{tar}}(n) - \widetilde{x}_{\mathrm{tar}}(n)\right)^2}. \tag{3.1}$$

Here, $\widehat{x}_{\mathrm{tar}}(n)$ corresponds to the output of the beamformer in the time domain and $\widetilde{x}_{\mathrm{tar}}(n)$ to the projection of $\widehat{x}_{\mathrm{tar}}(n)$ onto the subspace spanned by the target signal. For beamforming as a linear operation, the definition from [39] can be applied where $\widetilde{x}_{\mathrm{tar}}(n)$ corresponds to the output of the beamformer when applying it to the target speaker's source images $\mathbf{x}_{\mathrm{tar}}(\ell, k)$:

$$\widetilde{x}_{\mathrm{tar}}(n) = \mathrm{ISTFT}\left(\mathbf{w}^{\mathrm{H}}(\ell, k) \cdot \mathbf{x}_{\mathrm{tar}}(\ell, k)\right), \tag{3.2}$$

with $\mathrm{ISTFT}(\cdot)$ denoting the inverse short-time Fourier transform (ISTFT) over all time frames of the signal. Note that in block-wise beamforming $\mathbf{w}(\ell, k)$=$\mathbf{w}(b, k)$ holds for all time frames of the b-th block, i.e., for $\ell \in \mathcal{B}(b)$. Further, the difference $\widehat{x}_{\mathrm{tar}}(n) - \widetilde{x}_{\mathrm{tar}}(n)$

corresponds to the beamformer's output when applying it to the source images of the interference $\mathbf{x}_{\text{int}}(\ell, k)$:

$$\widehat{x}_{\text{tar}}(n) - \widetilde{x}_{\text{tar}}(n) = \text{ISTFT}\left(\mathbf{w}^{\text{H}}(\ell, k) \cdot \mathbf{x}_{\text{int}}(\ell, k)\right). \tag{3.3}$$

Considering block-wise beamforming with a block size of L time frames applied to a recording of a length of N_{B} blocks, the SDR as a function of the block size L can be expressed in the short-time Fourier transform (STFT) domain via

$$\begin{aligned}\text{SDR}(L) = \frac{P_{\text{tar}}(L)}{P_{\text{int}}(L)} &\approx \frac{\sum\limits_{b=0}^{N_{\text{B}}-1} \sum\limits_{\ell=b\cdot L}^{(b+1)\cdot L-1} \sum\limits_{k=0}^{K-1} \left|\mathbf{w}^{\text{H}}(b,k) \cdot \mathbf{x}_{\text{tar}}(\ell,k)\right|^2}{\sum\limits_{b=0}^{N_{\text{B}}-1} \sum\limits_{\ell=b\cdot L}^{(b+1)\cdot L-1} \sum\limits_{k=0}^{K-1} \left|\mathbf{w}^{\text{H}}(b,k) \cdot \mathbf{x}_1(\ell,k) + \mathbf{w}^{\text{H}}(b,k) \cdot \boldsymbol{\nu}(\ell,k)\right|^2} \\ &= \frac{\sum\limits_{b=0}^{N_{\text{B}}-1} \sum\limits_{\ell=b\cdot L}^{(b+1)\cdot L-1} \sum\limits_{k=0}^{K-1} \left|\mathbf{w}^{\text{H}}(b,k) \cdot \mathbf{x}_{\text{tar}}(\ell,k)\right|^2}{\sum\limits_{b=0}^{N_{\text{B}}-1} \sum\limits_{\ell=b\cdot L}^{(b+1)\cdot L-1} \sum\limits_{k=0}^{K-1} \left|\mathbf{w}^{\text{H}}(b,k) \cdot \mathbf{x}_{\text{int}}(\ell,k)\right|^2},\end{aligned} \tag{3.4}$$

where K denotes the number of frequency bins. This results from Parseval's theorem if a suitable analysis window function with a suitable block shift is used. The normalization factors, which account for the analysis window among others, cancel out and, therefore, are omitted. Note that for certain window functions the relation in (3.4) is identical to the value computed via (3.1). $P_{\text{tar}}(L)$ denotes the energy of the target speaker's signal at the beamformer output and $P_{\text{int}}(L)$ the energy of the interference at the beamformer output.

Similar to the results of previous works on antenna-based beamforming, e.g., [7], it is expected that the SDR increases with increasing block size L until it finally converges towards $\text{SDR}(\infty)$ for an infinitely large sample size used for spatial covariance matrix (SCM) estimation. Therefore, the SDR degradation $\Delta\text{SDR}(L)$ is introduced to simplify the representation of the degradation of the SDR, which is caused by the finite sample size used for SCM estimation. The SDR degradation $\Delta\text{SDR}(L)$, expressed in dB, is defined as

$$\Delta\text{SDR}(L) = 10 \cdot \log_{10}\left(\text{SDR}(\infty)\right) - 10 \cdot \log_{10}\left(\text{SDR}(L)\right). \tag{3.5}$$

Since $\text{SDR}(\infty)$ cannot be calculated it is approximated by $\text{SDR}(L_{\max})$ with $L_{\max}$ as highest considered block size. Hence, the SDR degradation is approximated as

$$\Delta\text{SDR}(L) \approx 10 \cdot \log_{10}\left(\text{SDR}(L_{\max})\right) - 10 \cdot \log_{10}\left(\text{SDR}(L)\right) \tag{3.6}$$

in the following. This can be justified by the fact that the SDR should already be converged to a value close to $\text{SDR}(\infty)$ if $L_{\max}$ is chosen large enough.

Since ASR is a typical downstream task, also the connection of the SDR to the word error rate (WER) is drawn at selected points. It is expected that the invasive SDR correlates well with the WER. However, in contrast to the invasive SDR, the WER, i.e., the ASR

performance, also depends on the strength of distortions and artifacts caused by beamforming. Although the MVDR beamformer has a distortionless response constraint, distortions can still occur, for example, due to an imperfect steering of the beamformer.

A pre-trained conformer-based ASR system [40] from the Espnet framework [41], which uses WAVLM features [42], is utilized to transcribe the extracted signals. The ASR system was trained based on Libri-Light [43], GigaSpeech [44] VoxPopuli [45], CHiME-4 [46], and WSJ0/1 [47] and achieves a WER of 1.9 % on the test-clean subset of LibriSpeech. In [27] it was mentioned that the pre-trained ASR system [48] for LibriSpeech from the ESPnet framework shows an undesired behavior when transcribing very long signals. The pre-trained conformer-based ASR system, used here, shows a similar behavior. Thus, all signals are segmented before being transcribed so that the maximum length of one segment is 18 s.

In the following chapters, a statistical model of source extraction using MVDR beamforming is considered. In order to be able to compare the second-order moments of two random variables, involved in this statistical model of beamforming, the correlation matrix distance, as proposed in [49], is employed. For example, the correlation matrix distance is used as metric to assess the quality of the approximation of a non-tractable probability distribution by a more simple probability distribution. The correlation matrix distance is defined as

$$d_{\text{corr}}(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) = 1 - \frac{\operatorname{tr}\left\{\widetilde{\mathbf{A}} \cdot \widetilde{\mathbf{B}}\right\}}{\left|\left|\widetilde{\mathbf{A}}\right|\right|_{\text{F}} \cdot \left|\left|\widetilde{\mathbf{B}}\right|\right|_{\text{F}}}, \tag{3.7}$$

with $||\cdot||_{\text{F}}$ denoting the Frobenius norm of a matrix. Note that smaller values are better, meaning that the second-order moments $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$ are closer to each other.

3.3 Source extraction

Throughout the experiments, the signals from both speakers are extracted. If not stated otherwise, a Blackman window with a size of 1024 samples (64 ms), a shift of 256 samples (16 ms) and a fast Fourier transform (FFT) size of 1024 samples are used for calculating the STFT of the signals. This window size and shift are typical STFT parameters used for beamforming. As shown in [50], this choice of the STFT window size maximizes the W-disjoint orthogonality of a speech mixture of two speakers. Here, the W-disjoint orthogonality quantifies the degree of non-overlap of the speaker's signals in the STFT domain. Thus, a large W-disjoint orthogonality is important for a good quality of mask-based SCM estimates.

Ideal ratio masks (IRMs) are used for mask-based SCM estimation. These are defined as the ratio between the absolute value of a source image $x_{i,m}(\ell, k)$, $i \in \{0, 1, \nu\}$, to the sum of absolute values of all source images [51]. Since asynchronous recordings will be considered later, here the IRMs are calculated for one reference channel. Thus, a synchronization of the microphone signals is not required for mask estimation. Note that the quality of the mask estimates and, therefore, the way how the masks are estimated indeed have an influence on the behavior of the beamformer with increasing block size. However, this will not be further investigated here to keep the scope of this work within reasonable limits.

3.4 Statistical model

As already mentioned, MVDR beamforming will be considered from a statistical point of view in the next chapters. This statistical point of view on MVDR beamforming is based on the local Gaussian model (LGM) of the source images, that was presented in Sec. 2.1.2. For all experiments involving the statistical model of the noisy and reverberant speech mixture which will be introduced in Sec. 4.1, the ground-truth SCMs of the LGM must be determined. The STFT window sizes which are typically used for beamforming are much smaller than the length of the RIRs for typical sound decay times. Thus, the rank-1 model of the SCMs based on the narrowband approximation of speech in the STFT domain would not be sufficient, as discussed in Sec. 2.1.1, so that full-rank SCMs are required. These are determined by employing the convergence behavior of a Wishart matrix, which was described in Sec. 2.4, and the LGM, which was discussed in Sec. 2.1.2.

Let $\widetilde{\mathbf{x}}_q(\ell,k)$ be the STFT of a stationary Gaussian noise signal convolved with the RIRs belonging to the full-rank SCM $\mathbf{R}_q(k)$ in the LGM of the speech images $\mathbf{x}_q(\ell,k)$. Assuming that the power of the stationary Gaussian noise signal is chosen such that $\widetilde{\mathbf{x}}_q(\ell,k)$ can be modeled as $\widetilde{\mathbf{x}}_q(\ell,k) \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_q(k))$, the sample SCM estimate

$$\widehat{\mathbf{R}}_q(k) = \frac{1}{L_{\mathrm{LGM}}} \cdot \sum_{\ell=0}^{L_{\mathrm{LGM}}-1} \widetilde{\mathbf{x}}_q(\ell,k) \cdot \widetilde{\mathbf{x}}_q^{\mathrm{H}}(\ell,k) \tag{3.8}$$

is Wishart distributed with $\widehat{\mathbf{R}}_q(k) \sim \mathcal{W}_M(L_{\mathrm{LGM}}, 1/L_{\mathrm{LGM}} \cdot \mathbf{R}_q(k))$. From the properties of the Wishart distribution, mentioned in Sec. 2.4, it follows that $\widehat{\mathbf{R}}_q(k)$ converges towards the ground-truth SCM $\mathbf{R}_q(k)$ as the sample size L_{LGM} grows. Thus, the full-rank SCMs $\mathbf{R}_q(k)$ of the LGM of the speakers' signals are approximated as the sample SCM estimates specified in (3.8) with $L_{\mathrm{LGM}} = 50{,}000$ time frames.

All expected values, which are needed in the experiments, are estimated using a Monte-Carlo simulation with 250 draws. This number of draws was chosen after a first convergence analysis which indicated that the estimated quantities, like the SDR, change only marginally when using a larger number of draws. Increasing the number of draws further would lead to a substantial rise in computational cost without a meaningful improvement in estimation accuracy.

4 Analysis of finite sample size effects on block-wise MVDR beamforming

In this chapter, the influence of a finite sample size used for estimating the spatial covariance matrices (SCMs), that are required to calculate the beamformer coefficients, on the performance of a minimum variance distortionless response (MVDR) beamformer is analyzed. It was shown in [7]–[11] that the performance of an MVDR beamformer improves with a growing sample size used for estimating the interference SCM. Note that in these works different names for the considered beamformers were used but the beamformers which were considered in these works only differ in minor details from the MVDR beamformer, e.g., in a different normalization of the beamformer coefficients.

For the analysis of the influence of SCM estimation from a finite sample size on MVDR beamforming, the expected value of the signal-to-interference-plus-noise ratio (SINR) after normalizing it by the optimal SINR was used in [8]–[10]. Note that the SINR coincides with the invasive signal-to-distortion ratio (SDR), which is used as main metric for the performance of a beamformer in this work. In [9], [11] deeper insights into the effect of SCM estimation based on a finite sample size were gained by first considering the power of the individual sources at the beamformer output separately before deriving a relationship between the output SINR and the sample size used for SCM estimation. Based on the probability distribution of the beamformer coefficients, which was derived in [52], [53], the first- and second-order moments of the MVDR coefficients were derived in [7]. On this basis a closed-form approximation of the expected value of the SINR was obtained.

Overall, the expressions for the expected SINR as a function of the sample size used for SCM estimation, the ground-truth SCMs and the steering vector, which were derived in these works, enable an assessment of the number of samples for SCM estimation that is needed to achieve a certain performance. However, the assumption underlying the derivations in [7]–[11] correspond to radio frequency (RF) transmissions and beamforming using antenna arrays so that these results cannot be directly transferred to the problem at hand with speech sources. On the one hand, narrowband beamforming is considered in RF transmissions while the extraction of speech signals via beamforming calls for broadband beamforming [54]. On the other hand, the signals considered in RF transmissions show a time-invariant power while the power of speech signals is strongly time-varying.

A central assumption of the derivations in [7]–[11] is that the SCM estimates are Wishart distributed, i.e., that the source signals correspond to independent and identically distributed (i.i.d.) samples from a zero-mean Gaussian random variables with fixed covariance matrix and constant power. For the problem at hand, the source signals can be modeled in the short-time Fourier transform (STFT) domain as i.i.d. samples from a zero-mean Gaussian

random variable via the local Gaussian model (LGM) but due to the time-varying power of speech their covariance matrices are time-variant. Hence, the SCM estimates are not Wishart distributed. To overcome this issue, an approach to approximate the probability distribution of the interference SCM by an equivalent Wishart distribution is introduced in this chapter. This approach builds upon the approximation of a weighted sum of Wishart matrices by an equivalent Wishart matrix that was proposed in [55]. By approximating the probability distribution of the interference-SCM estimates as a Wishart distribution, the first- and second-order moments of the MVDR beamformer coefficients, which were derived in [7], are used to derive a closed-form approximation of the SDR at the output of the beamformer.

In addition to that, previous works on the effects of SCMs which are estimated from a finite sample size on beamforming considered the case where the beamformer coefficients are estimated once and are applied to statistically independent signals afterwards. However, in block-wise beamforming, as considered here, the beamformer coefficients are estimated on the same part of the signals to which they are also applied. Therefore, statistical independence between the beamformer coefficients and the signals to which they are applied is not present here. In this chapter a closed-form approximation of the SDR at the beamformer output is derived which explicitly accounts for the statistical dependence between the beamformer coefficients and the signals to which they are applied.

The remainder of this chapter is organized as follows: First, the assumptions made for the statistical model of extracting a target signal via MVDR beamforming are introduced in Sec. 4.1. Afterwards, the approximation of the probability distribution of the interference-SCM estimates by a Wishart distribution and the first- and second-order moment of the beamformer coefficients, that follow from this approximation, are derived in Sec. 4.2. Subsequently, a closed-form approximation of the SDR at the beamformer output as function of the sample size used for SCM estimation is developed in Sec. 4.3 and analyzed in Sec. 4.4. In Sec. 4.5, the quality of the closed-form approximation of the SDR at the beamformer output is evaluated before the findings, made there, are further validated by an experimental investigation. Finally, the key findings about the effects of estimating SCMs from a finite sample size on MVDR beamforming are summarized in Sec. 4.6.

4.1 Modeling assumptions

For the analysis of the finite sample size effects on MVDR beamforming, a simplified version of the signal model which was introduced in Sec. 2.1 is considered in the following, with

$$\mathbf{y}(\ell,k) = \underbrace{\mathbf{x}_0(\ell,k)}_{:=\mathbf{x}_{\text{tar}}(\ell,k)} + \underbrace{\mathbf{x}_1(\ell,k) + \boldsymbol{\nu}(\ell,k)}_{:=\mathbf{x}_{\text{int}}(\ell,k)}. \tag{4.1}$$

Here, it is assumed that two speakers are simultaneously active. Moreover, the signal of speaker 0, belonging to the target source images $\mathbf{x}_{\text{tar}}(\ell,k) = \mathbf{x}_0(\ell,k)$, should be extracted while the interference signals $\mathbf{x}_{\text{int}}(\ell,k)$ corresponds to the superposition of the interfering speaker's source images $\mathbf{x}_1(\ell,k)$ and the noise images $\boldsymbol{\nu}(\ell,k)$.

As already mentioned before, block-wise beamforming is considered, where the block size L determines the sample size used for SCM estimation. The MVDR in the formulation specified in (2.47) is considered. In addition to that, a generalized eigenvalue decomposition (GEVD) based rank-1 approximation is applied to the target-SCM estimate which results in $\widehat{\mathbf{R}}_{\text{tar}}(b,k) = \widehat{\mathbf{d}}(b,k) \cdot \widehat{\mathbf{d}}^{\text{H}}(b,k)$. Here, $\widehat{\mathbf{d}}(b,k) = \widehat{\mathbf{d}}_{\text{GEV}}(b,k)$ is calculated from $\widehat{\mathbf{R}}_{\text{tar}}(b,k)$ and $\widehat{\mathbf{R}}_{\text{int}}(b,k)$ via (2.45). Without loss of generality, channel 0 is employed as reference channel from which $\mathbf{u} = \left[1, 0, \cdots, 0\right]^{\text{T}}$ follows. Thus, (2.47) becomes

$$\mathbf{w}(b,k) = \widehat{d}_0^*(b,k) \cdot \frac{\widehat{\mathbf{R}}_{\text{int}}^{-1}(b,k) \cdot \widehat{\mathbf{d}}(b,k)}{\widehat{\mathbf{d}}^{\text{H}}(b,k) \cdot \widehat{\mathbf{R}}_{\text{int}}^{-1}(b,k) \cdot \widehat{\mathbf{d}}(b,k)}. \tag{4.2}$$

In order to model the finite sample size effects on block-wise beamforming, first a statistical model for the invasive output SDR, as defined in (3.4), is developed. To this end, the statistical model of speech, which was introduced in Sec. 2.1.2, is utilized as model for the involved source images, with

$$\mathbf{x}_0(\ell,k) \sim \mathcal{N}(0, \sigma_0^2(\ell,k) \cdot \mathbf{R}_0(k)), \tag{4.3}$$

$$\mathbf{x}_1(\ell,k) \sim \mathcal{N}(0, \sigma_1^2(\ell,k) \cdot \mathbf{R}_1(k)), \tag{4.4}$$

$$\boldsymbol{\nu}(\ell,k) \sim \mathcal{N}(0, \sigma_\nu^2(k) \cdot \mathbf{R}_\nu(k)). \tag{4.5}$$

The power of the source signals $\sigma_0^2(\ell,k)$, $\sigma_1^2(\ell,k)$ and $\sigma_\nu^2(k)$ are modeled as deterministic quantities in the following. Additionally, the source images are assumed to be mutually statistically independent. Furthermore, the masks $\gamma_{\text{int}}(\ell,k)$ utilized for estimating the interference SCMs are modeled as deterministic quantities, too. Since the source images are modeled as random variables, the interference-SCM estimates $\widehat{\mathbf{R}}_{\text{int}}(b,k)$, the beamformer coefficients $\mathbf{w}(b,k)$ as well as the energy of the target signal at the beamformer output $P_{\text{tar}}(L)$ and interference signal at the beamformer output $P_{\text{int}}(L)$ also correspond to random variables. Hence, the output SDR, being defined in (3.4), is calculated via the expected values of the energies $P_{\text{tar}}(L)$ and $P_{\text{int}}(L)$ via

$$\text{SDR}(L) = \frac{\mathbb{E}\left[P_{\text{tar}}(L)\right]}{\mathbb{E}\left[P_{\text{int}}(L)\right]} = \frac{\sum_{b=0}^{N_{\text{B}}-1} \sum_{\ell=b\cdot L}^{(b+1)\cdot L-1} \sum_{k=0}^{K-1} \mathbb{E}\left[\left|\mathbf{w}^{\text{H}}(b,k) \cdot \mathbf{x}_{\text{tar}}(\ell,k)\right|^2\right]}{\sum_{b=0}^{N_{\text{B}}-1} \sum_{\ell=b\cdot L}^{(b+1)\cdot L-1} \sum_{k=0}^{K-1} \mathbb{E}\left[\left|\mathbf{w}^{\text{H}}(b,k) \cdot \mathbf{x}_{\text{int}}(\ell,k)\right|^2\right]}. \tag{4.6}$$

It is expected that a finite sample size used for interference-SCM estimation will generally have a larger impact on the beamforming performance than the steering of the beamformer based on the target-SCM estimate. On the one hand, this can be attributed to the rank-1 approximation of target-SCM estimate as outer product of steering vector estimate which mitigates the leakage of the interference signals into the target-SCM estimates. On the other hand, the suppression of all signal components of the interference is a more complex task than not distorting the main components of the target signal, which are represented by the

steering vector. Moreover, there are further ways to estimate the steering vector, e.g., by calculating the SCMs from time difference of arrival (TDOA) estimates assuming anechoic signal propagation, as for example in [OC2].

In consequence, the dependence of the beamforming performance on the block-size-dependent quality of the steering via the target-SCM estimate will be omitted in the following. The target-SCM estimate is modeled based on the outer product of the GEVD-based steering vector estimates $\widehat{\mathbf{d}}(b,k)$ which are derived from the ground-truth SCMs of the LGM via (2.45), with

$$\widehat{\mathbf{d}}(b,k) = (\mathbf{R}_1(k) + \mathbf{R}_\nu(k)) \cdot \mathcal{P}\big((\mathbf{R}_1(k) + \mathbf{R}_\nu(k))^{-1} \cdot \mathbf{R}_0(k)\big)\,. \tag{4.7}$$

Hence, the steering vector estimates $\widehat{\mathbf{d}}(b,k)$ correspond to a deterministic quantity which simplifies the derivation of the closed-form approximation of the invasive output SDR.

4.2 Statistical model of the beamformer coefficients

Next, a statistical model of the interference-SCM estimates and, based on this, a statistical model of the beamformer coefficients are derived. Following the LGM of the STFT of speech signals, which was presented in Sec. 2.1.2, the estimated interference SCMs in (2.49) correspond to a sum of dyadic products of samples from multivariate Gaussians whose covariance matrices vary over the time frame index. Thus, there is no closed-form expression for the probability distribution of the interference-SCM estimates. In addition to that, the non-linear transformations of the estimated SCMs involved in the calculation of the MVDR beamformer coefficients, e.g., calculating the matrix inverse, further complicate to determine a closed-form solution for the probability distribution of the MVDR beamformer coefficients.

4.2.1 Statistical model of the interference-SCM estimates

In order to derive a statistical model of the interference-SCM estimates, first, the properties of the interference-SCM estimates will be considered. For this, a single frequency bin k and block b=0 will be considered for sake of a simpler notation. Hence, the block index b as well as the frequency bin index k will be omitted in the following if not stated otherwise. After rewriting (2.49), it can be seen that the interference-SCM estimates

$$\widehat{\mathbf{R}}_{\text{int}} = \frac{1}{\sum_{\ell=0}^{L-1} \gamma_{\text{int}}(\ell)} \cdot \sum_{\ell=0}^{L-1} \gamma_{\text{int}}(\ell) \cdot \mathbf{y}(\ell) \cdot \mathbf{y}^{\text{H}}(\ell) = \sum_{\ell=0}^{L-1} \widetilde{\mathbf{y}}(\ell) \cdot \widetilde{\mathbf{y}}^{\text{H}}(\ell)\,, \tag{4.8}$$

$$\text{with } \widetilde{\mathbf{y}}(\ell) \sim \mathcal{N}\left(0, \widetilde{\sigma}_0^2(\ell) \cdot \mathbf{R}_0 + \widetilde{\sigma}_1^2(\ell) \cdot \mathbf{R}_1 + \widetilde{\sigma}_\nu^2 \cdot \mathbf{R}_\nu\right), \tag{4.9}$$

correspond to a sum of dyadic products of samples from zero-mean, circular multivariate Gaussians. The frame-wise weights involved in the calculation of the covariance matrix of these Gaussian are given by

$$\widetilde{\sigma}_i^2(\ell) = \frac{\gamma_{\text{int}}(\ell) \cdot \sigma_i^2(\ell)}{\sum_{\ell=0}^{L-1} \gamma_{\text{int}}(\ell)}, \; i \in \{0, 1, \nu\}. \tag{4.10}$$

Thus, the basic structure of the interference-SCM estimates is similar to a Wishart matrix but the interference-SCM estimates are not Wishart distributed due to the varying covariance matrices of the Gaussians following from the time-varying power of speech.

The expected value of the interference-SCM estimates is given by

$$\mathbb{E}\left[\widehat{\mathbf{R}}_{\text{int}}\right] = \sum_{\ell=0}^{L-1} \widetilde{\sigma}_0^2(\ell) \cdot \mathbf{R}_0 + \widetilde{\sigma}_1^2(\ell) \cdot \mathbf{R}_1 + \widetilde{\sigma}_\nu^2(\ell) \cdot \mathbf{R}_\nu = \sum_{\ell=0}^{L-1} \widetilde{\mathbf{R}}(\ell), \tag{4.11}$$

with $\widetilde{\mathbf{R}}(\ell) = \widetilde{\sigma}_0^2(\ell) \cdot \mathbf{R}_0 + \widetilde{\sigma}_1^2(\ell) \cdot \mathbf{R}_1 + \widetilde{\sigma}_\nu^2(\ell) \cdot \mathbf{R}_\nu$. For the variance of the i-th row and j-th column element of the estimated interference SCMs

$$\text{var}\left(\left(\widehat{\mathbf{R}}_{\text{int}}\right)_{ij} \right) = \sum_{\ell=0}^{L-1} \widetilde{R}_{ii}(\ell) \cdot \widetilde{R}_{jj}(\ell) \tag{4.12}$$

holds, with $(\cdot)_{ij}$ corresponding to the i-th row and j-th column element of a matrix. Refer to Appendix A.2 for a detailed derivation of the element-wise variance.

Considering the ratio between the variance and the squared absolute value of the expected value of the main diagonal elements

$$\frac{\text{var}\left(\left(\widehat{\mathbf{R}}_{\text{int}}\right)_{ii} \right)}{\left| \mathbb{E}\left[\left(\widehat{\mathbf{R}}_{\text{int}}\right)_{ii} \right] \right|^2} = \frac{\sum_{\ell=0}^{L-1} \widetilde{R}_{ii}(\ell) \cdot \widetilde{R}_{ii}(\ell)}{\left| \sum_{\ell=0}^{L-1} \widetilde{R}_{ii}(\ell) \right|^2} = \frac{\sum_{\ell=0}^{L-1} \widetilde{R}_{ii}(\ell) \cdot \widetilde{R}_{ii}(\ell)}{\sum_{\ell=0}^{L-1} \widetilde{R}_{ii}(\ell) \cdot \widetilde{R}_{ii}(\ell) + \sum_{\ell=0}^{L-1} \sum_{\substack{\widetilde{\ell}=0 \\ \widetilde{\ell} \neq \ell}}^{L-1} \widetilde{R}_{ii}(\ell) \cdot \widetilde{R}_{ii}\left(\widetilde{\ell}\right)}, \tag{4.13}$$

it becomes apparent that this ratio decreases with growing block size L for a typical distribution of the time frequency bin wise powers of speech signals. Hence, the elements on the main diagonal of the interference-SCM estimates converge towards their expected value with growing block size. The off-diagonal elements also converge towards their expected value with growing block size (see Appendix A.3).

Thus, the basic structure and the convergence behavior of the interference-SCM estimates are similar to the ones of a Wishart-distributed matrix. Furthermore, the first- and second-order moments of the beamformer coefficients, which are relevant for the calculation of the power of the signals at the output of the beamformer later, were derived for Wishart-distributed SCM estimates in [7]. Moreover, it was shown in [55] that a weighted sum of Wishart-distributed matrices can be approximated by an equivalent Wishart matrix. Hence, the aim is to approximate the probability distribution of the mask-based SCM estimates as

an equivalent Wishart distribution in a similar way to the approximation in [55]. In this way, the derivation of a closed-form approximation of the output SDR of the beamformer can be kept tractable.

Firstly, the approximation of a weighted sum of Ξ Wishart matrices $\mathbf{W}_i$, with Ξ_i degrees of freedom,

$$\overline{\mathbf{W}} = \sum_{i=1}^{\Xi} \frac{a_i}{\Xi_i} \cdot \mathbf{W}_i \tag{4.14}$$

by an equivalent Wishart matrix $\widetilde{\mathbf{W}}$, presented in [55], is shortly recapitulated. Here, it is assumed that the matrices $\mathbf{W}_i$ are Wishart distributed, with $\mathbf{W}_i \sim \mathcal{W}_D(\Xi_i, \sigma^2 \cdot \mathbf{I})$, and the weights $a_i \in \mathbb{R}^+$ are positive and real-valued. As proposed in [55], $\overline{\mathbf{W}}$ can be approximated by a equivalent Wishart matrix $\widetilde{\mathbf{W}} \sim \mathcal{W}_D\big(\Xi_{\widetilde{\mathbf{W}}}, \mathbf{\Sigma}_{\widetilde{\mathbf{W}}}\big)$, with $\Xi_{\widetilde{\mathbf{W}}}$ corresponding to the equivalent degrees of freedom and $\mathbf{\Sigma}_{\widetilde{\mathbf{W}}}$ corresponding to the equivalent scale matrix. From matching the expected values of the matrix-valued random variables $\overline{\mathbf{W}}$ and $\widetilde{\mathbf{W}}$, i.e.,

$$\mathbb{E}\Big[\widetilde{\mathbf{W}}\Big] \overset{!}{=} \mathbb{E}\Big[\overline{\mathbf{W}}\Big] \tag{4.15}$$

$$\Leftrightarrow \Xi_{\widetilde{\mathbf{W}}} \cdot \mathbf{\Sigma}_{\widetilde{\mathbf{W}}} \overset{!}{=} \sum_{i=1}^{\Xi} \frac{a_i}{\Xi_i} \cdot \Xi_i \cdot \sigma^2 \cdot \mathbf{I}, \tag{4.16}$$

it follows that

$$\mathbf{\Sigma}_{\widetilde{\mathbf{W}}} = \frac{1}{\Xi_{\widetilde{\mathbf{W}}}} \cdot \sum_{i=1}^{\Xi} a_i \cdot \sigma^2 \cdot \mathbf{I}. \tag{4.17}$$

By matching the element-wise variances of the main-diagonal elements of the matrix-valued random variable $\overline{\mathbf{W}}$ and its approximate equivalent Wishart matrix $\widetilde{\mathbf{W}}$, i.e.,

$$\mathrm{var}\Big(\widetilde{W}_{jj}\Big) \overset{!}{=} \mathrm{var}\Big(\overline{W}_{jj}\Big) \;\forall j \in \{0, 1, \ldots, M-1\} \tag{4.18}$$

$$\Leftrightarrow \Xi_{\widetilde{\mathbf{W}}} \cdot \frac{1}{\big(\Xi_{\widetilde{\mathbf{W}}}\big)^2} \cdot \left(\sum_{i=1}^{\Xi} a_i \cdot \sigma^2\right)^2 \overset{!}{=} \sum_{i=1}^{\Xi} \Xi_i \cdot \left(\frac{a_i \cdot \sigma^2}{\Xi_i}\right)^2 \;\forall j \in \{0, 1, \ldots, M-1\}, \tag{4.19}$$

the equivalent degrees of freedom $\Xi_{\widetilde{\mathbf{W}}}$ are determined as

$$\Xi_{\widetilde{\mathbf{W}}} = \left\lfloor \frac{\left(\sum_{i=1}^{\Xi} a_i\right)^2}{\sum_{i=1}^{\Xi} \frac{a_i^2}{\Xi_i}} \right\rceil. \tag{4.20}$$

Here, $\lfloor \cdot \rceil$ denotes rounding to the next closest integer value.

The similarity of (4.14) to the interference-SCM estimates in (4.8) can be seen when the Wishart matrices $\mathbf{W}_i$ in (4.14) are expressed as sum over dyadic products which results in

$$\overline{\mathbf{W}} = \sum_{i=1}^{\Xi} \frac{a_i}{\Xi_i} \cdot \mathbf{W}_i = \sum_{i=1}^{\Xi} \sum_{j=1}^{\Xi_i} \mathbf{o}_i(j) \cdot \mathbf{o}_i^{\mathrm{H}}(j), \tag{4.21}$$

with $\mathbf{o}_i(j) \sim \mathcal{N}\left(0, \frac{a_i}{\Xi_i} \cdot \sigma^2 \cdot \mathbf{I}\right)$. Despite the similarity of (4.8) and (4.21), there are also non-negligible differences between the weighted sum of Wishart-distributed matrices and the interference-SCM estimates. While in (4.21) the covariance matrices of all Gaussians $\mathcal{N}\left(0, \frac{a_i}{\Xi_i} \cdot \sigma^2 \cdot \mathbf{I}\right)$ whose samples $\mathbf{o}_i(j)$ form the dyadic products have the same structure and are only scaled by different scalars, the structure of the covariance matrices in (4.8) can vary over the different summands since the ratio of $\widetilde{\sigma}_0(\ell)$, $\widetilde{\sigma}_1(\ell)$ and $\widetilde{\sigma}_\nu(\ell)$ might vary. However, in the frequency range in which speech has significant power, i.e., in the frequency range between 0.125 kHz and 3.5 kHz, the contribution of the interfering speaker usually dominates the covariance matrix of the Gaussians involved in (4.8). Thus, the rough structure of the covariance matrix will be maintained which should mitigate the effect of the change of the structure of the covariance matrices over the summands.

Although there are differences between the interference-SCM estimates $\widehat{\mathbf{R}}_{\text{int}}(b,k)$ and a weighted sum of Wishart matrices, the overall structure of both is very similar. Hence, the block-wise estimates of the interference-SCM estimates $\widehat{\mathbf{R}}_{\text{int}}(b,k)$ shall be approximated by equivalent Wishart matrices in a manner similar to the approximation from [55] in the following. In order to approximate the probability distribution of the interference-SCM estimates as an equivalent Wishart distribution, the first- and second-order moments of the estimated interference SCMs and their approximations are matched as described above. The starting point for the derivation of the equivalent Wishart matrix $\widetilde{\mathbf{R}}_{\text{int}}$ as approximation of the interference-SCM estimate $\widehat{\mathbf{R}}_{\text{int}}$ is chosen as

$$\widehat{\mathbf{R}}_{\text{int}} \approx \widetilde{\mathbf{R}}_{\text{int}}, \text{ with } \widetilde{\mathbf{R}}_{\text{int}} \sim \mathcal{W}_M\left(\widetilde{L}, \frac{1}{\widetilde{L}} \cdot \boldsymbol{\Sigma}_{\text{int}}\right). \tag{4.22}$$

By matching the expected values of the interference-SCM estimates and their equivalent Wishart matrices,

$$\mathbb{E}\left[\widetilde{\mathbf{R}}_{\text{int}}\right] \stackrel{!}{=} \mathbb{E}\left[\widehat{\mathbf{R}}_{\text{int}}\right] \tag{4.23}$$

$$\Leftrightarrow \qquad \boldsymbol{\Sigma}_{\text{int}} = \sum_{\ell=0}^{L-1} \widetilde{\sigma}_0^2(\ell) \cdot \mathbf{R}_0 + \widetilde{\sigma}_1^2(\ell) \cdot \mathbf{R}_1 + \widetilde{\sigma}_\nu^2(\ell) \cdot \mathbf{R}_\nu \tag{4.24}$$

follows for the structure of the equivalent scale matrix.

The equivalent degrees of freedom $\widetilde{L}$ are determined by matching the element-wise variances of interference-SCM estimates and their approximations. This results in

$$\operatorname{var}\left(\left(\widetilde{\mathbf{R}}_{\text{int}}\right)_{ij}\right) \stackrel{!}{=} \operatorname{var}\left(\left(\widehat{\mathbf{R}}_{\text{int}}\right)_{ij}\right) \tag{4.25}$$

$$\Leftrightarrow \frac{1}{\widetilde{L}} \cdot (\boldsymbol{\Sigma}_{\text{int}})_{ii} \cdot (\boldsymbol{\Sigma}_{\text{int}})_{jj} \stackrel{!}{=} \sum_{\ell=0}^{L-1} \left(\widetilde{\mathbf{R}}(\ell)\right)_{ii} \cdot \left(\widetilde{\mathbf{R}}(\ell)\right)_{jj}, \tag{4.26}$$

with $\widetilde{\mathbf{R}}(\ell) = \widetilde{\sigma}_0^2(\ell) \cdot \mathbf{R}_0 + \widetilde{\sigma}_1^2(\ell) \cdot \mathbf{R}_1 + \widetilde{\sigma}_\nu^2(\ell) \cdot \mathbf{R}_\nu$. As a consequence of the varying structure of the covariance matrices over the different summands in (4.8), which already was mentioned above, no closed-form solution for the equivalent degrees of freedom $\widetilde{L}$ exists.

To solve this issue, the system of equations, resulting from (4.26), is interpreted as an overdetermined least squares (LS) problem. Consider $\mathbf{V}$ as a matrix that summarizes the element-wise variances of the interference-SCM estimates, with $V_{ij} = \sum_{\ell=0}^{L-1} (\widetilde{\mathbf{R}}(\ell))_{ii} \cdot (\widetilde{\mathbf{R}}(\ell))_{jj}$, and $\widetilde{\mathbf{V}}$ as a matrix that summarizes the unscaled element-wise variances of the Wishart matrix approximations, with $\widetilde{V}_{ij} = (\mathbf{\Sigma}_{\text{int}})_{ii} \cdot (\mathbf{\Sigma}_{\text{int}})_{jj}$. With these definitions the following LS problem for the equivalent degrees of freedom is stated:

$$\widetilde{L} = \underset{\breve{L}}{\operatorname{argmin}} \left|\left| \breve{L} \cdot \mathbf{v} - \widetilde{\mathbf{v}} \right|\right|^2 . \tag{4.27}$$

Here, $\mathbf{v}$ and $\widetilde{\mathbf{v}}$ are defined as the vectorized elements of the upper-triangle part of the matrices $\mathbf{V}$ and $\widetilde{\mathbf{V}}$, respectively, and $||\cdot||^2$ corresponds to the squared value of the Euclidean norm. Finally, the equivalent degrees of freedom $\widetilde{L}$ result from solving the overdetermined LS problem:

$$\widetilde{L} = \left\lfloor \left(\mathbf{v}^{\mathrm{H}} \cdot \mathbf{v}\right)^{-1} \cdot \mathbf{v}^{\mathrm{H}} \cdot \widetilde{\mathbf{v}} \right\rceil = \left\lfloor \frac{\sum_{i=0}^{M-1} \sum_{j=i+1}^{M-1} \left(\sum_{\ell=0}^{L-1} (\widetilde{\mathbf{R}}(\ell))_{ii} \cdot (\widetilde{\mathbf{R}}(\ell))_{jj} \right) \cdot (\mathbf{\Sigma}_{\text{int}})_{ii} \cdot (\mathbf{\Sigma}_{\text{int}})_{jj}}{\sum_{i=0}^{M-1} \sum_{j=i+1}^{M-1} \left(\sum_{\ell=0}^{L-1} (\widetilde{\mathbf{R}}(\ell))_{ii} \cdot (\widetilde{\mathbf{R}}(\ell))_{jj} \right)^2} \right\rceil . \tag{4.28}$$

The scaling of the covariance matrices of all Gaussians might vary over all summands in (4.8). In combination with the high dynamic range of the weights $\widetilde{\sigma}_0(\ell)$ and $\widetilde{\sigma}_1(\ell)$ in (4.8) due to the high dynamic range of the power of speech, this can lead to individual summands dominating the sum rendering the effective number of terms in the sum small. Consequently, it might happen that the equivalent degrees of freedom, resulting from (4.28), are smaller than the number of microphones M and, therefore, also smaller than the dimension of the equivalent Wishart matrix. In this case, the equivalent Wishart matrix is singular and cannot be inverted which however is required for the calculation of the MVDR beamformer coefficients. Furthermore, first experiments have shown that the equivalent Wishart matrix might become ill-conditioned if the equivalent degrees of freedom are only marginally larger than the number of microphones M. As a result, the inverse of the equivalent Wishart matrix becomes unstable. To this end, the equivalent degrees of freedom are floored to the value $2 \cdot M$, which results in

$$\widetilde{L} = \text{maximum}\left(2 \cdot M, \left\lfloor \frac{\sum_{i=0}^{M-1} \sum_{j=i+1}^{M-1} \left(\sum_{\ell=0}^{L-1} (\widetilde{\mathbf{R}}(\ell))_{ii} \cdot (\widetilde{\mathbf{R}}(\ell))_{jj} \right) \cdot (\mathbf{\Sigma}_{\text{int}})_{ii} \cdot (\mathbf{\Sigma}_{\text{int}})_{jj}}{\sum_{i=0}^{M-1} \sum_{j=i+1}^{M-1} \left(\sum_{\ell=0}^{L-1} (\widetilde{\mathbf{R}}(\ell))_{ii} \cdot (\widetilde{\mathbf{R}}(\ell))_{jj} \right)^2} \right\rceil \right) . \tag{4.29}$$

The relation between the equivalent degrees of freedom $\widetilde{L}$ and the block size is investigated later (see Fig. 4.6 and Fig. 4.7).

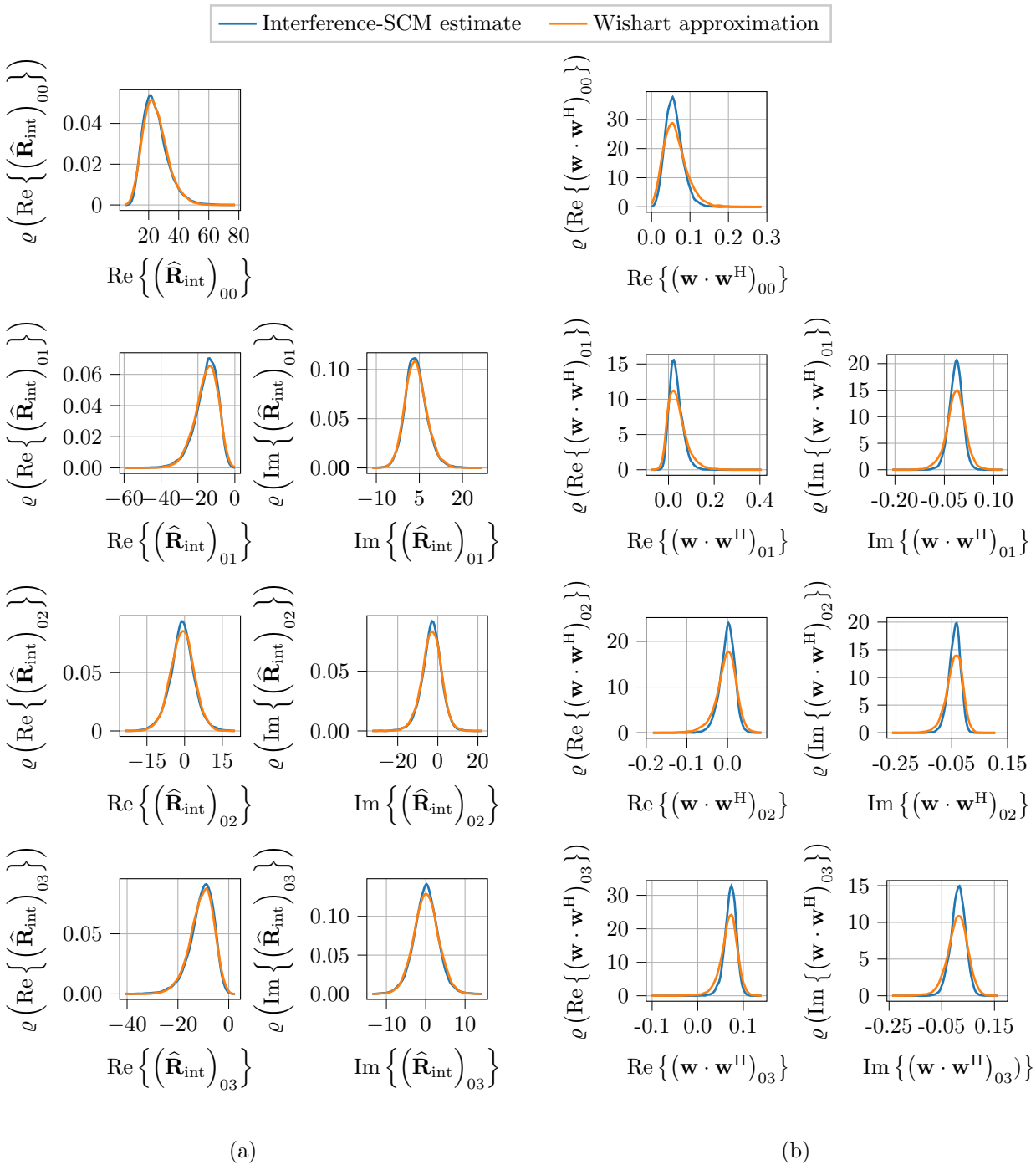


Figure 4.1: Probability density function (PDF) $\varrho(\cdot)$ of illustrative elements $(\widehat{\mathbf{R}}_{\text{int}})_{ij}$ of an interference-SCM estimate and its approximation based on an equivalent Wishart distribution (a). The PDFs of the elements $(\mathbf{w}\cdot\mathbf{w}^{\text{H}})_{ij}$ of the outer product of the resulting beamformer coefficients are shown in (b). The PDF of the real part $\text{Re}\{\cdot\}$ and the PDF of the imaginary part $\text{Im}\{\cdot\}$ of the matrix elements are separately shown. The probability distributions are estimated using a Gaussian kernel-density estimation method [56]–[58].

In Fig. 4.1(a) a comparison of the probability distribution of the first row elements of an interference-SCM estimate to the probability distribution of the corresponding elements of its equivalent Wishart matrix is shown. To be able to visualize the probability distribution of the generally complex-valued elements of the matrices, the probability distribution of their real and imaginary parts are depicted separately. Note that this form of representation does not capture all properties of the probability distribution of the matrices and should only give a first hint about the quality of the approximation of the interference-SCM estimate by an equivalent Wishart matrix. For example, the dependence between the real and imaginary part of the matrix elements as well as the correlation between the matrix elements is not represented. It can be seen in Fig. 4.1(a) that the probability distributions of the real and imaginary part of the elements of the estimated interference SCM and its approximation by an equivalent Wishart matrix match well. Hence, the approximation of the non-trivial probability distribution of the interference-SCM estimates by a Wishart distribution can be seen to be valid in general.

As explained later, the second-order moment of the beamformer coefficients is a fundamental component of the closed-form approximation of the output SDR of the beamformer. Hence, Fig. 4.1(b) illustrates a comparison between the probability distributions of the outer-product elements of the beamformer coefficients obtained from the estimated interference SCMs and those derived from the approximated interference SCMs modeled as a Wishart matrix. Although the probability distribution of the elements of the interference-SCM estimates and the probability distribution of their approximation as Wishart matrix seem to match well, the probability distributions of the outer products' elements of the resulting beamformer coefficients show a visible difference. This might be explained by the fact that the dependence between the real and imaginary part of the elements of the SCM as well as the correlation between the SCM elements are not represented in Fig. 4.1(a), as mentioned above. In addition to that, the non-linear transformations that are applied to the interference-SCM estimates when calculating the beamformer coefficients amplify the differences between the probability distributions of the interference-SCM estimates and their approximation as Wishart matrix even more. If the resulting deviations affect the expected value of the outer product of the beamformer coefficients, approximating the probability distribution of the interference-SCM estimates by a Wishart has a negative impact on the accuracy of the closed-form approximation of the beamformer's output SDR, that is derived later.

In order to be able to quantify the influence of approximating the interference-SCM estimates as Wishart matrices on the second-order moment of the beamformer coefficients, Fig. 4.2 depicts the correlation matrix distance $d_{\text{corr}}(\mathbb{E}[\mathbf{w} \cdot \mathbf{w}^{\text{H}}], \mathbb{E}[\widetilde{\mathbf{w}} \cdot \widetilde{\mathbf{w}}^{\text{H}}])$ between the second-order moment of the beamformer coefficients $\mathbf{w}$ and the second-order moment its approximation $\widetilde{\mathbf{w}}$, based on a Wishart distribution. To this end, the definition of the correlation matrix distance from Sec. 3.2 is utilized. Moreover, the dependence of correlation matrix distance $d_{\text{corr}}(\mathbb{E}[\mathbf{w} \cdot \mathbf{w}^{\text{H}}], \mathbb{E}[\widetilde{\mathbf{w}} \cdot \widetilde{\mathbf{w}}^{\text{H}}])$ on the block size and the frequency is considered. It becomes clear that the second-order moment of the beamformer coefficients and its approximation based on the Wishart distributed interference-SCM estimates show a deviation. In particular, this holds for small block sizes which are especially important for the investigation of the effects of a finite sample size used for SCM estimation on MVDR beamforming. With increasing block size, the approximation of the second-order moment of the beamformer coefficients based on Wishart distributed interference-SCM estimates becomes more accurate.

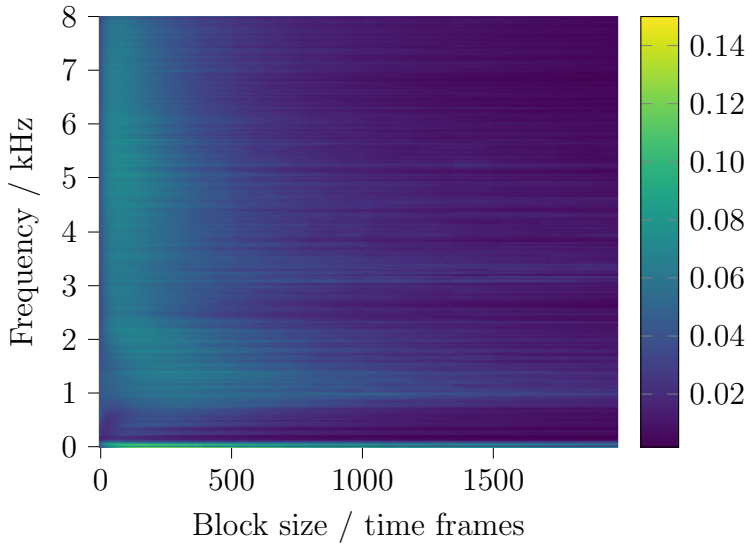


Figure 4.2: Average correlation matrix distance $d_{\text{corr}}(\mathbb{E}[\mathbf{w} \cdot \mathbf{w}^{\text{H}}], \mathbb{E}[\widetilde{\mathbf{w}} \cdot \widetilde{\mathbf{w}}^{\text{H}}])$ between the second-order moment of the beamformer coefficients $\mathbf{w}$ and its approximation $\widetilde{\mathbf{w}}$ using the Wishart approximation of the interference-SCM estimates

However, for the most relevant frequencies of speech, i.e., the frequency range from 0.125 kHz to 3.5 kHz, a deviation from the second-order moment of the beamformer coefficients without this approximation is still visible. First experiments have shown that the deviation in the second-order moment of the beamformer coefficients introduced by the approximation of the probability distribution of the interference-SCM estimates by a Wishart distribution negatively impacts the accuracy of the approximation of the power of the signal components at the beamformer output. Hence, this approximation has to be improved before using it to derive a closed-form approximation of the invasive SDR at the beamformer output.

As discussed before, the calculation of the beamformer coefficients involves non-linear transformations of the interference-SCM estimates. Hence, samples stemming from the tails of the probability distribution of the equivalent Wishart matrix might result in more pronounced tails of the probability distribution of the beamformer coefficients' outer product. In addition to that, only the second-order moment of the beamformer coefficients is of interest for the derivation of the closed-form approximation of the output SDR of the beamformer, as shown later. Thus, slight deviations between the probability distribution of the outer product of the beamformer coefficients caused by the approximation of the interference-SCM estimates by an equivalent Wishart matrix are acceptable as long as these deviations do not distort the expected value of the interference-SCM estimates. This insight can be used to narrow the gap between approximating the interference-SCM estimates by an equivalent Wishart matrix and approximating a weighted sum of Wishart matrices by a an equivalent Wishart matrix. To do so, the weights $\widetilde{\sigma}_i(\ell),\ \ i \in \{0, 1, \nu\}$, in (4.9) are smoothed based on calculating their

moving average via

$$\widetilde{\sigma}_i^2(\ell) = \text{movingavg}\left(\frac{\gamma_{\text{int}}(\ell)\cdot\sigma_i^2(\ell)}{\sum_{\ell=0}^{L-1}\gamma_{\text{int}}(\ell)};3\right),\ i\in\{0,1,\nu\}. \tag{4.30}$$

Here, movingavg$(\cdot;3)$ corresponds to a moving average which is computed using a centered window of size three, where each value is replaced by the average of itself and its two adjacent neighbors. After smoothing the weights $\widetilde{\sigma}_i(\ell)$, the structure of the covariance matrices of consecutive dyadic products in (4.8) becomes more similar so that (4.8) becomes more similar to (4.21). Note that the expected value of the interference-SCM estimates is not affected by smoothing the weights $\widetilde{\sigma}_i(\ell)$ via (4.30).

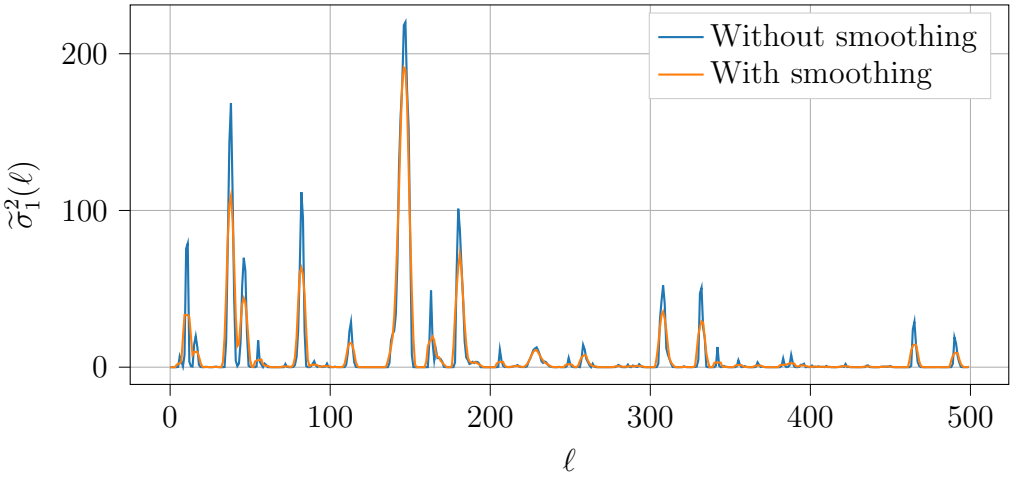


Figure 4.3: Visualization of weight smoothing via (4.30) used to approximate the probability distribution of the interference-SCM estimates by a Wishart distribution based on the weights $\widetilde{\sigma}_1^2(\ell)$ of $\mathbf{R}_1$ in (4.9)

The effect of smoothing the weights $\widetilde{\sigma}_i^2(\ell)$ of the sources' covariance matrices via (4.30) before calculating the covariance matrix of the Gaussian in (4.9) is visualized in Fig. 4.3. It can be seen that the sparsity of speech spectra and the high dynamic range of the weights are maintained. Furthermore, it becomes obvious that high peaks of the weights are mitigated but still are clearly dominant. Hence, the behavior of the effects of a finite sample size used for SCM estimation on MVDR beamforming, which will be investigated later, should not be affected too much by smoothing the weights $\widetilde{\sigma}_i^2(\ell)$ via (4.30).

Figure 4.4 shows how smoothing the weights $\widetilde{\sigma}_i^2(\ell)$ via (4.30) affects the probability distribution of the elements of the Wishart matrix as approximation of the interference-SCM estimates and the probability distribution of the outer product of the beamformer coefficients resulting from the Wishart approximation. It becomes obvious that smoothing the frame-wise weights involved in the statistical model of the SCM estimates before approximating the SCM estimates by an equivalent Wishart matrix reduces the element-wise variance of the equivalent Wishart matrix compared to the case without weight smoothing shown in Fig. 4.1(a) as depicted in Fig. 4.4(a). As shown in Fig. 4.4(b), the resulting reduced influence of the

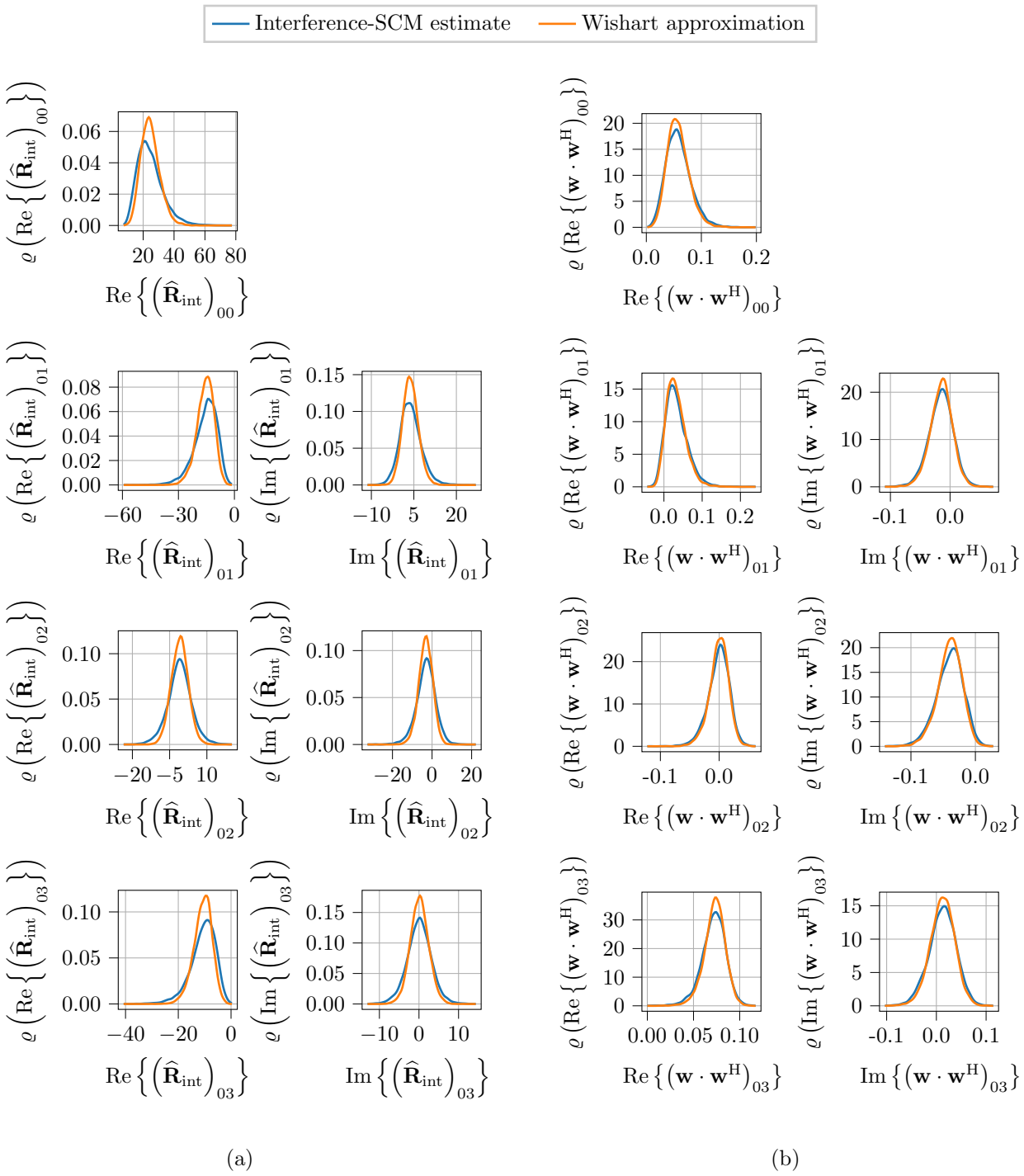


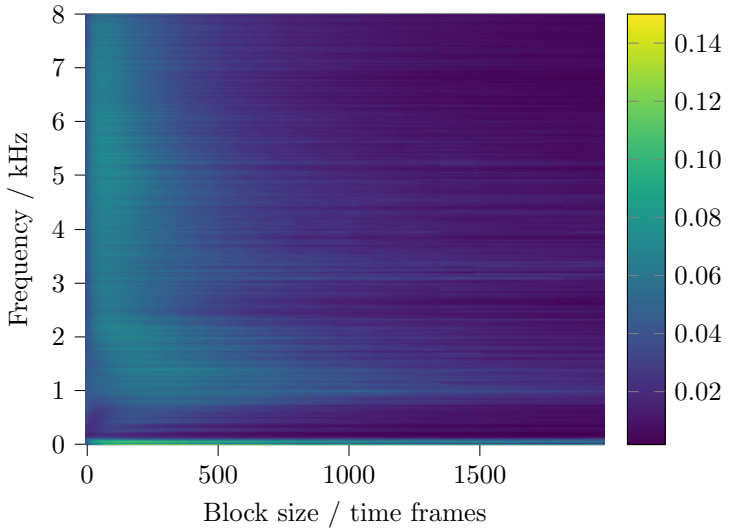
Figure 4.4: PDF $\varrho(\cdot)$ of illustrative elements $(\widehat{\mathbf{R}}_{\text{int}})_{ij}$ of an interference-SCM estimate and its approximation via an equivalent Wishart distribution using weight smoothing via (4.30) (a). The PDFs of the elements $(\mathbf{w} \cdot \mathbf{w}^{\text{H}})_{ij}$ of the outer product of the resulting beamformer coefficients are shown in (b). The PDF of the real part $\text{Re}\{\cdot\}$ and the PDF of the imaginary part $\text{Im}\{\cdot\}$ of the matrix elements are separately shown. The probability distributions are estimated using a Gaussian kernel-density estimation method [56]–[58].

tails of the probability distribution is beneficial to improve the accuracy of the probability distribution of the beamformer coefficients' outer product compared to the case without weight smoothing, being visualized in Fig. 4.1(b). Especially, the heavy tails, which in combination with the skewness of the probability distribution can lead to a distortion of the second-order moment of the beamformer coefficients, are nearly vanished. Thus, the Wishart approximation with an upstream smoothing of the weights $\widetilde{\sigma}_i^2(\ell)$ via (4.30) seems to be a suitable model of the probability distribution of the interference-SCM estimates around its expected value. However, the tails of the distribution are not accurately reflected by this approximation.

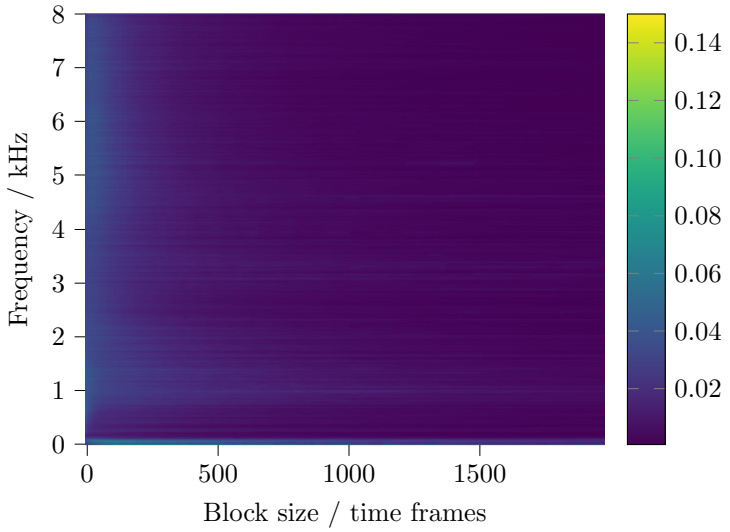
An alternative point of view on the variance reduction stems from the idea of variance reduction techniques to improve Monte-Carlo simulations [59, chapter 4]. The latter can also be used to calculate the expected value of a strongly non-linear transformation of a random variable, as in the problem at hand. Here, the aim is to reduce the variance of the samples in order to reduce the error of the estimate of the expected value. However, the weight smoothing in (4.30) changes the probability distribution of the random variable while variance reduction techniques do not affect the probability distribution but only the sampling process.

Figure 4.5 illustrates a comparison of the accuracy of the second-order moments of the beamformer coefficients resulting from the Wishart approximation of the interference-SCM estimates with weight smoothing via (4.30) to the accuracy of the second-order moments of the beamformer coefficients resulting from the Wishart approximation of the interference-SCM estimates without weight smoothing. As in Fig. 4.2, the correlation matrix distance $d_{\text{corr}}(\mathbf{w} \cdot \mathbf{w}^{\text{H}}, \mathbb{E}[\widetilde{\mathbf{w}} \cdot \widetilde{\mathbf{w}}^{\text{H}}])$ between second-order moment of the beamformer coefficients $\mathbf{w}$ and second-order moment of the beamformer coefficients $\widetilde{\mathbf{w}}$ following from the Wishart approximation of the interference-SCM estimates is considered. It becomes obvious that weight smoothing via (4.30) leads to a significant improvement of the approximation of the second-order moment of the beamformer coefficients. This improves the approximation of the beamformers' output SDR, which will be derived later, too. Weight smoothing especially improves the approximation for small block sizes. Good accuracy is particularly important for small block sizes since it is expected that the performance of an MVDR beamformer is especially affected by block size increases when the block size is small. But, there are also improvements for larger block sizes which become modeled almost perfectly when using weight smoothing. This especially holds for the frequency range between 0.5 kHz and 3.5 kHz, i.e., for a large part of the main frequency range of speech.

Note that the improvement of the second-order moment does not automatically guarantee a better model of the beamformer coefficients' probability distribution. For example, it could happen that the variance of the outer product of the beamformer coefficients becomes smaller than for its original probability distribution without affecting its second-order moment. However, this is immaterial because only an accurate model of the second-order moment of the beamformer coefficients is of interest for the derivation of the closed-form approximation of the SDR at the beamformer output, as shown later.



(a)



(b)

Figure 4.5: Comparison of the average correlation matrix distance $d_{\text{corr}}(\mathbf{w} \cdot \mathbf{w}^{\text{H}}, \mathbb{E}[\widetilde{\mathbf{w}} \cdot \widetilde{\mathbf{w}}^{\text{H}}])$ between the second-order moment of the beamformer coefficients $\mathbf{w}$ and its approximation $\widetilde{\mathbf{w}}$, which is based on approximating the probability distribution of the interference-SCM estimates by a Wishart distribution, for the case without weight smoothing via (4.30) (a) and the case with weight smoothing via (4.30) (b)

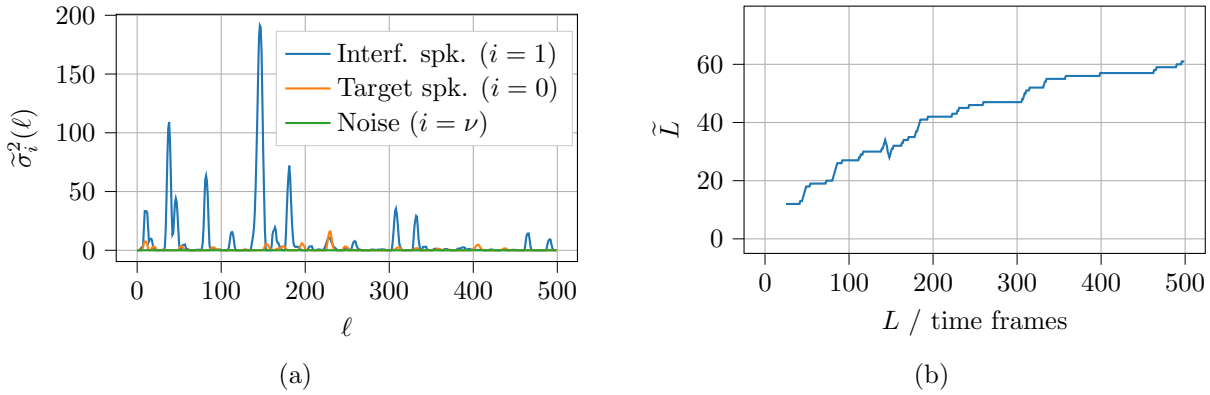


Figure 4.6: Relationship between the weights $\widetilde{\sigma}_i^2(\ell)$ of $\mathbf{R}_i$, $i \in \{0, 1, \nu\}$, in (4.9) (a) and the equivalent degrees of freedom $\widetilde{L}$ of the Wishart approximation of the SCM estimates (b) as a function of the block size L

Figure 4.6 shows the relationship between the block size L and the equivalent degrees of freedom $\widetilde{L}$ of the Wishart approximation of the interference-SCM estimates. It can be seen in Fig. 4.6(b) that the equivalent degrees of freedom grow with growing block size, as expected. However, the equivalent degrees of freedom are much smaller than the block size and also grow much slower than the block size. This behavior can be explained by the weights $\widetilde{\sigma}_i^2(\ell)$, $i \in \{0, 1, \nu\}$, of the sources' ground-truth SCMs in (4.9) which are depicted in Fig. 4.6(a). On the one hand, speech is sparse so that many time frames have only a negligible contribution to the interference-SCM estimate. On the other hand, the power of speech has a high dynamic range so that individual time frames might dominate the interference-SCM estimate. This results in an increased variance of the SCM estimates, which is equivalent to a smaller value of the degrees of freedom of a Wishart distribution. Overall, the structure of speech leads to a slower convergence of the SCM estimates towards their expected value compared to the case of sources that have a time-constant power. From this, it follows that more time frames are needed for SCM estimation in comparison to the case with sources that have time-constant power to achieve the same quality of the interference-SCM estimates.

The frequency dependence of the behavior of the average equivalent degrees of freedom with growing block size is depicted in Fig. 4.7. As already shown in Fig. 4.6, the equivalent degrees of freedom generally grow as the block size becomes larger. In addition to that, the equivalent degrees of freedom are much smaller than the block size for nearly all frequencies. Especially, in the frequency range between 0.5 kHz and 2.5 kHz, for which the interfering speaker has significant power, the equivalent degrees of freedom are significant smaller than the number of time frames used for SCM estimation.

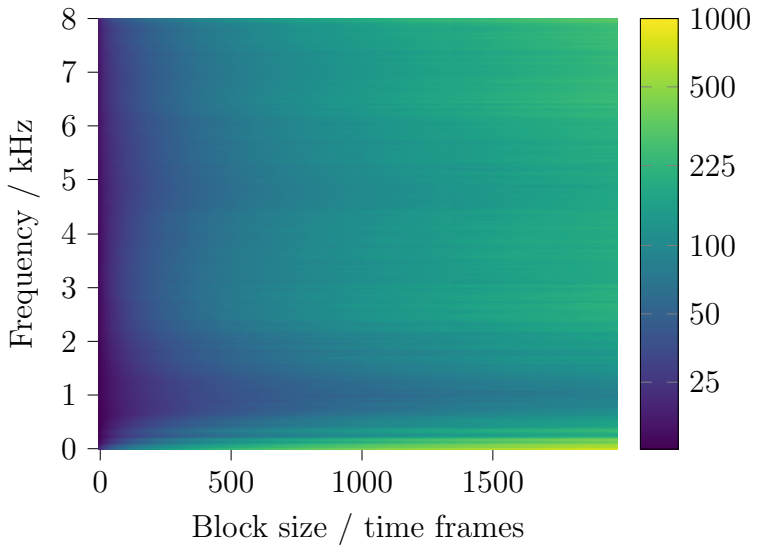


Figure 4.7: Average equivalent degrees of freedom $\widetilde{L}$ of the Wishart approximation of the interference-SCM estimates' probability distribution as a function of the block size L and the frequency

4.2.2 First- and second-order moment of the beamformer coefficients

Based on the approximation of the probability distribution of the interference-SCM estimates as a Wishart distribution, the interference-SCM estimates are assumed to be Wishart distributed in the following with $\widehat{\mathbf{R}}_{\text{int}} \sim \mathcal{W}_M\left(\widetilde{L}, \frac{1}{\widetilde{L}} \cdot \boldsymbol{\Sigma}_{\text{int}}\right)$. Note that the frequency bin index and the block index are omitted here and in the following for sake of a simpler notation. The first- and second-order moments of the beamformer coefficients were derived in [7] for the sample matrix inversion (SMI) beamformer, which corresponds an MVDR beamformer. With the results from [7], it follows that the expected value of the coefficients of a Souden-MVDR with rank-1 approximation of the target SCM estimate is given by

$$\mathbb{E}[\mathbf{w}] = \widehat{d}_0^* \cdot \frac{\boldsymbol{\Sigma}_{\text{int}}^{-1} \cdot \widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\text{H}} \cdot \boldsymbol{\Sigma}_{\text{int}}^{-1} \cdot \widehat{\mathbf{d}}}. \tag{4.31}$$

This means that the expected value of the beamformer coefficients corresponds to a beamformer that is calculated based on the expected value $\boldsymbol{\Sigma}_{\text{int}}$ of the Wishart approximation of the interference-SCM estimates. As specified in (4.24), $\boldsymbol{\Sigma}_{\text{int}}$ corresponds to the expected value of the interference-SCM estimates. For the problem at hand, this corresponds to calculating the beamformer coefficients from the ground-truth SCMs of the source images based on (4.24). Note that the multiplication by $\widehat{d}_0^*$ stems from the choice of the Souden-MVDR beamformer instead of the MVDR beamformer in its classic formulation.

As derived in [7], the second-order moment of the beamformer coefficients is given by

$$\mathbb{E}\left[\mathbf{w} \cdot \mathbf{w}^{\text{H}}\right] = \left|\widehat{d}_0\right|^2 \cdot \left(\frac{1}{\widetilde{L} - M + 1} \cdot \frac{\boldsymbol{\Sigma}_{\text{int}}^{-1}}{\widehat{\mathbf{d}}^{\text{H}} \cdot \boldsymbol{\Sigma}_{\text{int}}^{-1} \cdot \widehat{\mathbf{d}}} + \frac{\widetilde{L} - M}{\widetilde{L} - M + 1} \cdot \frac{\boldsymbol{\Sigma}_{\text{int}}^{-1} \cdot \widehat{\mathbf{d}} \cdot \widehat{\mathbf{d}}^{\text{H}} \cdot \boldsymbol{\Sigma}_{\text{int}}^{-1}}{\left(\widehat{\mathbf{d}}^{\text{H}} \cdot \boldsymbol{\Sigma}_{\text{int}}^{-1} \cdot \widehat{\mathbf{d}}\right)^2} \right). \tag{4.32}$$

For a growing value of the degrees of freedom $\widetilde{L}$ of the Wishart distribution, the second-order moment of the beamformer coefficients converges to the outer product of the expected value of the beamformer coefficients:

$$\lim_{\widetilde{L}\to\infty} \mathbb{E}\big[\mathbf{w}\cdot\mathbf{w}^{\mathrm{H}}\big] = \left|\widehat{d}_0\right|^2 \cdot \frac{\boldsymbol{\Sigma}_{\mathrm{int}}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int}}^{-1}}{\left(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int}}^{-1}\cdot\widehat{\mathbf{d}}\right)^2} = \mathbb{E}[\mathbf{w}]\cdot\left(\mathbb{E}[\mathbf{w}]\right)^{\mathrm{H}}. \tag{4.33}$$

This reflects the reduction of the variance of the SCM estimates with growing block size. Furthermore, there is an additive bias term which diminishes as the degrees of freedom $\widetilde{L}$ increase.

4.3 Derivation of a closed-form approximation of the output SDR

Based on the approximation of the probability distribution of the interference-SCM estimates via an equivalent Wishart distribution, a closed-form approximation of the SDR at the beamformer output will be derived in the following. For the calculation of the output SDR in (4.6), the expected values $\mathbb{E}\left[P_{\mathrm{tar}}(L)\right]$ and $\mathbb{E}\left[P_{\mathrm{int}}(L)\right]$ of the target and interference signals' energies at the beamformer output have to be calculated. Assuming that the samples used for interference-SCM estimation and the signals to which the beamformer is applied are statistically independent, as in [7], it holds that

$$\mathbb{E}\left[\left|\mathbf{w}^{\mathrm{H}}(b,k)\cdot\mathbf{x}_{\mathrm{tar}}(\ell,k)\right|^2\right] = \mathrm{tr}\big\{\mathbf{R}_{\mathrm{tar}}(\ell,k)\cdot\mathbb{E}\big[\mathbf{w}(b,k)\cdot\mathbf{w}^{\mathrm{H}}(b,k)\big]\big\} \tag{4.34}$$

and

$$\mathbb{E}\left[\left|\mathbf{w}^{\mathrm{H}}(b,k)\cdot\mathbf{x}_{\mathrm{int}}(\ell,k)\right|^2\right] = \mathrm{tr}\big\{\mathbf{R}_{\mathrm{int}}(\ell,k)\cdot\mathbb{E}\big[\mathbf{w}(b,k)\cdot\mathbf{w}^{\mathrm{H}}(b,k)\big]\big\}, \tag{4.35}$$

with the instantaneous SCMs

$$\mathbf{R}_{\mathrm{tar}}(\ell,k) = \sigma_0^2(\ell,k)\cdot\mathbf{R}_0(k) \tag{4.36}$$

and

$$\mathbf{R}_{\mathrm{int}}(\ell,k) = \sigma_1^2(\ell,k)\cdot\mathbf{R}_1(k) + \sigma_\nu^2(\ell,k)\cdot\mathbf{R}_\nu(k) \tag{4.37}$$

of the target and interference signals to which the beamformer is applied.

With (4.6), (4.34) and (4.35),

$$\mathrm{SDR}(L) = \frac{\mathbb{E}\left[P_{\mathrm{tar}}(L)\right]}{\mathbb{E}\left[P_{\mathrm{int}}(L)\right]} = \frac{\sum_{b=0}^{N_{\mathrm{B}}-1}\sum_{\ell=b\cdot L}^{(b+1)\cdot L-1}\sum_{k=0}^{K-1}\overbrace{\mathrm{tr}\big\{\mathbf{R}_{\mathrm{tar}}(\ell,k)\cdot\mathbb{E}\big[\mathbf{w}(b,k)\cdot\mathbf{w}^{\mathrm{H}}(b,k)\big]\big\}}^{:=p_{\mathrm{tar}}(b,\ell,k)}}{\sum_{b=0}^{N_{\mathrm{B}}-1}\sum_{\ell=b\cdot L}^{(b+1)\cdot L-1}\sum_{k=0}^{K-1}\underbrace{\mathrm{tr}\big\{\mathbf{R}_{\mathrm{int}}(\ell,k)\cdot\mathbb{E}\big[\mathbf{w}(b,k)\cdot\mathbf{w}^{\mathrm{H}}(b,k)\big]\big\}}_{:=p_{\mathrm{int}}(b,\ell,k)}} \tag{4.38}$$

follows for the SDR at the beamformer output. Note that the interference-SCM estimates

$$\widehat{\mathbf{R}}_{\text{int}}(b,k) \sim \mathcal{W}_M\left(\widetilde{L}(b,k), \frac{1}{\widetilde{L}(b,k)} \cdot \boldsymbol{\Sigma}_{\text{int}}(b,k)\right) \tag{4.39}$$

are approximated by an equivalent Wishart matrix for each block b and frequency bin k. Hence, the equivalent degrees of freedom $\widetilde{L}(b,k)$ and the equivalent scale matrix $\boldsymbol{\Sigma}_{\text{int}}(b,k)$ of the Wishart approximation of the interference-SCM estimates generally are block dependent. Utilizing the second-order moment of the beamformer coefficients in (4.32), it follows that the time frequency bin wise power of the target signal at the beamformer output is given by

$$\begin{aligned} &p_{\text{tar}}(b,\ell,k) \\ &= \left(\frac{1}{\widetilde{L}(b,k) - M + 1} \cdot \frac{|\widehat{d}_0(k)|^2 \cdot \text{tr}\left\{ \boldsymbol{\Sigma}_{\text{int}}^{-1}(b,k) \cdot \mathbf{R}_{\text{tar}}(\ell,k) \right\}}{\widehat{\mathbf{d}}^{\text{H}}(k) \cdot \boldsymbol{\Sigma}_{\text{int}}^{-1}(b,k) \cdot \widehat{\mathbf{d}}(k)} \right. \\ &\quad \left. + \frac{\widetilde{L}(b,k) - M}{\widetilde{L}(b,k) - M + 1} \cdot \frac{|\widehat{d}_0(k)|^2 \cdot \widehat{\mathbf{d}}^{\text{H}}(k) \cdot \boldsymbol{\Sigma}_{\text{int}}^{-1}(b,k) \cdot \mathbf{R}_{\text{tar}}(\ell,k) \cdot \boldsymbol{\Sigma}_{\text{int}}^{-1}(b,k) \cdot \widehat{\mathbf{d}}(k)}{\left(\widehat{\mathbf{d}}^{\text{H}}(k) \cdot \boldsymbol{\Sigma}_{\text{int}}^{-1}(b,k) \cdot \widehat{\mathbf{d}}(k) \right)^2} \right) \end{aligned} \tag{4.40}$$

and, analogously,

$$\begin{aligned} &p_{\text{int}}(b,\ell,k) \\ &= \left(\frac{1}{\widetilde{L}(b,k) - M + 1} \cdot \frac{|\widehat{d}_0(k)|^2 \cdot \text{tr}\left\{ \boldsymbol{\Sigma}_{\text{int}}^{-1}(b,k) \cdot \mathbf{R}_{\text{int}}(\ell,k) \right\}}{\widehat{\mathbf{d}}^{\text{H}}(k) \cdot \boldsymbol{\Sigma}_{\text{int}}^{-1}(b,k) \cdot \widehat{\mathbf{d}}(k)} \right. \\ &\quad \left. + \frac{\widetilde{L}(b,k) - M}{\widetilde{L}(b,k) - M + 1} \cdot \frac{|\widehat{d}_0(k)|^2 \cdot \widehat{\mathbf{d}}^{\text{H}}(k) \cdot \boldsymbol{\Sigma}_{\text{int}}^{-1}(b,k) \cdot \mathbf{R}_{\text{int}}(\ell,k) \cdot \boldsymbol{\Sigma}_{\text{int}}^{-1}(b,k) \cdot \widehat{\mathbf{d}}(k)}{\left(\widehat{\mathbf{d}}^{\text{H}}(k) \cdot \boldsymbol{\Sigma}_{\text{int}}^{-1}(b,k) \cdot \widehat{\mathbf{d}}(k) \right)^2} \right) \end{aligned} \tag{4.41}$$

holds for the time frequency bin wise power of the interference at the beamformer output.

In block-wise beamforming, which is considered here, the interference-SCM estimates and signals to which the beamformer coefficients are applied are not statistically independent, as mentioned above. This introduces effects which are not captured by (4.40) and (4.41), e.g., the effect that the beamformer can focus on the suppression of the specific characteristics of individual time frequency bins of the interference which dominate the interference-SCM estimate. In order to handle the statistical dependence between the beamformer coefficients and the signals to which they are applied, the aim is to rewrite the interference-SCM estimate so that results for applying a beamformer to statistically independent signals, described above, can be reused. In the following, a single frequency bin k and block b=0 will be considered for sake of a shorter notation. Hence, the frequency bin index k and block b will be neglected where possible.

First, the application of the beamformer to the ℓ-th time frame of the interfering speaker's signal $\mathbf{x}_1(\ell)$ is considered. To do so the Sherman-Morrison formula [60]

$$\left(\widetilde{\mathbf{A}} + \widetilde{\mathbf{b}} \cdot \widetilde{\mathbf{c}}^{\mathrm{H}}\right)^{-1} = \widetilde{\mathbf{A}}^{-1} - \frac{\widetilde{\mathbf{A}}^{-1} \cdot \widetilde{\mathbf{b}} \cdot \widetilde{\mathbf{c}}^{\mathrm{H}} \cdot \widetilde{\mathbf{A}}^{-1}}{1 + \widetilde{\mathbf{c}}^{\mathrm{H}} \cdot \widetilde{\mathbf{A}}^{-1} \cdot \widetilde{\mathbf{b}}} \tag{4.42}$$

is applied to the interference-SCM estimate in (2.49). Since the normalization of the interference-SCM estimate by $\sum_{\ell=0}^{L-1} \gamma_{\mathrm{int}}(\ell)$ cancels out when calculating the beamformer coefficients via (4.2), the normalization is omitted in the following to simplify the notation. By applying the Sherman-Morrison formula to the interference-SCM estimates

$$\widehat{\mathbf{R}}_{\mathrm{int}} = \underbrace{\sum_{\substack{\widetilde{\ell}=0 \\ \widetilde{\ell} \neq \ell}}^{L-1} \gamma_{\mathrm{int}}\left(\widetilde{\ell}\right) \cdot \mathbf{y}\left(\widetilde{\ell}\right) \cdot \mathbf{y}^{\mathrm{H}}\left(\widetilde{\ell}\right)}_{:=\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}} + \gamma_{\mathrm{int}}(\ell) \cdot \mathbf{y}\left(\ell\right) \cdot \mathbf{y}^{\mathrm{H}}\left(\ell\right), \tag{4.43}$$

its inverse can be written as

$$\widehat{\mathbf{R}}_{\mathrm{int}}^{-1} = \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} - \frac{\gamma_{\mathrm{int}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{y}\left(\ell\right) \cdot \mathbf{y}^{\mathrm{H}}\left(\ell\right) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}}{1 + \gamma_{\mathrm{int}}(\ell) \cdot \mathbf{y}^{\mathrm{H}}\left(\ell\right) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{y}\left(\ell\right)}. \tag{4.44}$$

Here, $\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}$ denotes the interference-SCM estimate that is calculated based on all time frames of the block under consideration except the ℓ-th time frame. With (4.44), the interfering speaker's contribution to the output of the beamformer for the ℓ-th time frame results in

$$\mathbf{w}^{\mathrm{H}} \cdot \mathbf{x}_1(\ell) = \widehat{d}_0 \cdot \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \left(\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} - \frac{\gamma_{\mathrm{int}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{y}(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}}{1 + \gamma_{\mathrm{int}}(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{y}(\ell)} \right)}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \left(\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} - \frac{\gamma_{\mathrm{int}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{y}(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}}{1 + \gamma_{\mathrm{int}}(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{y}(\ell)} \right) \cdot \widehat{\mathbf{d}}} \cdot \mathbf{x}_1(\ell). \tag{4.45}$$

Considering the W-disjoint orthogonality of speech, there typically are two possible cases. On the one hand, there are time frequency bins that are dominated by the interfering speaker's signal and, on the other hand, there are time frequency bins that are dominated by the signal of another source. If a time frequency bin is dominated by the interfering speaker's signal, it is expected that the beamformer focuses on the suppression of the interfering speaker's signal for this specific time frequency bin. Especially, this holds if the block size is small and the time frequency bin dominates the other time frequency bins contributing to the interference-SCM estimate. If a time frequency bin is not dominated by the interfering speaker's signal, it is not very likely that the beamformer focuses on the suppression of the interfering speaker's signal for that specific time frequency bin. This can be attributed to the fact that the time frequency bin corresponds either to a mix of multiple sources or even approximately corresponds to the signal of another source. Moreover, it is expected that in this case other time frequency bins will dominate the interference-SCM estimate so that the beamformer tends to focus rather on those time frequency bins. Hence, it is expected that a time frequency bin of the interfering speaker's signals tends to behave like signals that are

statistically independent from the beamformer coefficients if the time frequency bin is not dominated by the interfering speaker.

First, the case where the time frame of interest is dominated by the interfering speaker's signal is considered. In this case, $\mathbf{y}(\ell,k) \approx \mathbf{x}_1(\ell,k)$ follows from the approximate W-disjoint orthogonality of speech. Utilizing $\mathbf{y}(\ell,k) \approx \mathbf{x}_1(\ell,k)$, the contribution of the interfering speaker's signal to the output of the beamformer for the ℓ-th time frame can be reformulated as

$$\begin{aligned}
\mathbf{w}^{\mathrm{H}} \cdot \mathbf{x}_1(\ell) &\approx \widehat{d}_0 \cdot \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \left(\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} - \frac{\gamma_{\mathrm{int}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}}{1 + \gamma_{\mathrm{int}}(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell)} \right)}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \left(\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} - \frac{\gamma_{\mathrm{int}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}}{1 + \gamma_{\mathrm{int}}(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell)} \right) \cdot \widehat{\mathbf{d}}} \cdot \mathbf{x}_1(\ell) \\
&= \frac{\widehat{d}_0 \cdot \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}} \cdot \mathbf{x}_1(\ell)}{1 + \gamma_{\mathrm{int}}(\ell) \cdot \left(\mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) - \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}} \right)} \\
&= \frac{\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}_1(\ell)}{1 + \gamma_{\mathrm{int}}(\ell) \cdot \left(\mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) - \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}} \right)}.
\end{aligned} \tag{4.46}$$

Please refer to Appendix A.4 for a detailed derivation. Note that the numerator in the last row of (4.46) corresponds to the application of the beamformer coefficients

$$\mathbf{w}_{\setminus\ell} = \widehat{d}_0^{*} \cdot \frac{\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}, \tag{4.47}$$

which are calculated from all time frames of the block under consideration except the ℓ-th time frame, to the ℓ-th time frame of the interfering speaker's signals $\mathbf{x}_1(\ell)$. Consequently, the beamformer coefficients $\mathbf{w}_{\setminus\ell}$ and the ℓ-th time frame of the interfering speaker's signals $\mathbf{x}_1(\ell)$ are statistically independent.

Now, the case where the interfering speaker's signal $\mathbf{x}_1(\ell,k)$ does not dominate the corresponding time frequency bin is considered. As mentioned above, the beamformer coefficients and the interfering speaker's signal can be seen as approximately statistically independent in this case. Assuming that the block size is not impracticably small, the application of the beamformer coefficients to the interfering speaker's signal can be approximated via

$$\mathbf{w}^{\mathrm{H}} \cdot \mathbf{x}_1(\ell) \approx \mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}_1(\ell). \tag{4.48}$$

This means that the application of the beamformer coefficients to the ℓ-th time frame of the interfering speaker's source images is approximated by the application of the beamformer coefficients $\mathbf{w}_{\setminus\ell}$ which are statistically independent from the time frame $\mathbf{x}_1(\ell)$. Note that this approximation might introduce slight inaccuracies since the proportion of the contributions of the single sources to the interference-SCM estimate can slightly change when excluding the ℓ-th frame from SCM estimation. For example, it can happen that a non-negligible

contribution of the target speaker's signal to the interference-SCM estimate is not taken into account by the approximation in (4.48) if the target speaker's power $\sigma_1(\ell,k)$ is large and the interference mask $\gamma_{\text{int}}(\ell,k)$ is not zero for the time frame of interest. However, this way to achieve the statistical independence of the beamformer coefficients and ℓ-th time frame of the interfering speaker's signal has useful properties for the derivation of the power of the overall interference at the beamformer output, as shown later.

The decision whether $\mathbf{x}_1(\ell,k)$ dominates the ℓ-th time frame is based on the source signals' powers $\sigma_0^2(\ell,k)$, $\sigma_1^2(\ell,k)$ and $\sigma_\nu^2(k)$ by comparing the ratio of the power of the source of interest to the sum of all powers to a certain threshold th_{dom}:

$$\frac{\sigma_1^2(\ell,k)}{\sigma_0^2(\ell,k)+\sigma_1^2(\ell,k)+\sigma_\nu^2(k)} \geq \text{th}_{\text{dom}} \Rightarrow \mathbf{x}_1(\ell,k) \text{ is dominant}, \tag{4.49}$$

$$\frac{\sigma_1^2(\ell,k)}{\sigma_0^2(\ell,k)+\sigma_1^2(\ell,k)+\sigma_\nu^2(k)} < \text{th}_{\text{dom}} \Rightarrow \mathbf{x}_1(\ell,k) \text{ is not dominant}. \tag{4.50}$$

By generalizing the findings made for the contribution of the interfering speaker's signal to the beamformer output to the contribution of the signal of an arbitrary source, it follows that

$$\mathbf{w}^{\text{H}} \cdot \mathbf{x} \approx \frac{\mathbf{w}_{\setminus\ell}^{\text{H}} \cdot \mathbf{x}}{\delta(\mathbf{x})},\ \mathbf{x} \in \{\mathbf{x}_0(\ell), \mathbf{x}_1(\ell), \boldsymbol{\nu}(\ell)\}, \tag{4.51}$$

with

$$\delta(\mathbf{x}) = \begin{cases} 1+\gamma_{\text{int}}(\ell)\cdot\left(\mathbf{y}^{\text{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\text{int},\setminus\ell}^{-1}\cdot\mathbf{x} - \frac{\widehat{\mathbf{d}}^{\text{H}}\cdot\widehat{\mathbf{R}}_{\text{int},\setminus\ell}^{-1}\cdot\mathbf{x}\cdot\mathbf{y}^{\text{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\text{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\text{H}}\cdot\widehat{\mathbf{R}}_{\text{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}}\right) & ,\mathbf{x} \text{ is dominant} \\ 1 & ,\text{else} \end{cases}. \tag{4.52}$$

Based on (4.51) the power of the target signal at the beamformer output can be calculated for the ℓ-th time frame via

$$p_{\text{tar}}(\ell) = \mathbb{E}\Big[\big|\mathbf{w}^{\text{H}} \cdot \mathbf{x}_{\text{tar}}(\ell)\big|^2\Big] \approx \mathbb{E}\left[\frac{\big|\mathbf{w}_{\setminus\ell}^{\text{H}} \cdot \mathbf{x}_0(\ell)\big|^2}{|\delta(\mathbf{x}_0(\ell))|^2}\right]. \tag{4.53}$$

Analogously, the power of the interference at the beamformer output can be calculated for the ℓ-th time frame via

$$p_{\text{int}}(\ell) = \mathbb{E}\big[\mathbf{w}^{\text{H}} \cdot \mathbf{x}_{\text{int}}(\ell)\big] \approx \mathbb{E}\left[\left|\frac{\mathbf{w}_{\setminus\ell}^{\text{H}} \cdot \mathbf{x}_1(\ell)}{\delta(\mathbf{x}_1(\ell))} + \frac{\mathbf{w}_{\setminus\ell}^{\text{H}} \cdot \boldsymbol{\nu}(\ell)}{\delta(\boldsymbol{\nu}(\ell))}\right|^2\right]. \tag{4.54}$$

Since at most one source can be dominant for each time frequency bin, at least $\delta(\mathbf{x}_1(\ell))=1$ or $\delta(\boldsymbol{\nu}(\ell))=1$ holds. Additionally, the zero-mean source images are assumed to be statistically independent such that

$$p_{\text{int}}(\ell) \approx \mathbb{E}\left[\frac{\big|\mathbf{w}_{\setminus\ell}^{\text{H}} \cdot \mathbf{x}_1(\ell)\big|^2}{|\delta(\mathbf{x}_1(\ell))|^2}\right] + \mathbb{E}\left[\frac{\big|\mathbf{w}_{\setminus\ell}^{\text{H}} \cdot \boldsymbol{\nu}(\ell)\big|^2}{|\delta(\boldsymbol{\nu}(\ell))|^2}\right]. \tag{4.55}$$

follows for the power of the interference at the beamformer output for the ℓ-th time frame.

Both powers, $p_{\text{tar}}(\ell)$ as well as $p_{\text{int}}(\ell)$, require calculating terms of the form

$$p(\mathbf{x}) = \mathbb{E}\left[\frac{\left|\mathbf{w}_{\setminus\ell}^{\text{H}} \cdot \mathbf{x}\right|^2}{\left|\delta(\mathbf{x})\right|^2}\right],\ \mathbf{x} \in \{\mathbf{x}_0(\ell), \mathbf{x}_1(\ell), \boldsymbol{\nu}(\ell)\}\,. \tag{4.56}$$

This corresponds to a fraction of two functions of the same random variables.

In order to find a way to simplify the calculation of the expected value of the powers $p_{\text{tar}}(\ell)$ and $p_{\text{int}}(\ell)$, whose probability distributions are non-tractable, first the general case of calculating the expected value $\mathbb{E}[\varphi(u,v)]$ is considered, with $\varphi(u,v)$ denoting an arbitrary function of the two random variables u and v. As presented in [61], a Taylor series expansion is utilized to simplify the calculation of the expected value. The second-order Taylor series expansion of $\varphi(u,v)$ around the expected values of u and v, i.e., developing $\varphi(u,v)$ around (μ_u, μ_v), leads to

$$\begin{aligned}\varphi(u,v) \approx& \varphi(\mu_u,\mu_v) + \varphi'_u(\mu_u,\mu_v)\cdot(u-\mu_u) + \varphi'_v(\mu_u,\mu_v)\cdot(v-\mu_v) + \frac{1}{2}\cdot\varphi''_{uu}(\mu_u,\mu_v)\cdot(u-\mu_u)^2 \\ &+ \frac{1}{2}\cdot\varphi''_{uv}(\mu_u,\mu_v)\cdot(u-\mu_u)\cdot(v-\mu_v) + \frac{1}{2}\cdot\varphi''_{vv}(\mu_u,\mu_v)\cdot(v-\mu_v)^2,\end{aligned} \tag{4.57}$$

with the partial derivatives

$$\begin{aligned}&\varphi'_u(u,v) = \frac{\partial}{\partial u}\varphi(u,v),\ \varphi'_v(u,v) = \frac{\partial}{\partial v}\varphi(u,v),\ \varphi''_{uu}(u,v) = \frac{\partial^2}{\partial u^2}\varphi(u,v),\\ &\varphi''_{uv}(u,v) = \frac{\partial^2}{\partial u\,\partial v}\varphi(u,v),\ \varphi''_{vv}(u,v) = \frac{\partial^2}{\partial v^2}\varphi(u,v).\end{aligned}$$

Going back to the problem at hand, where the expected value of a fraction of two random variables has to be calculated, $\varphi(u,v)=\frac{u}{v}$ is considered. By applying the second-order Taylor series expansion which was introduced above, it follows, as derived in [61], that the expected value of the function $\varphi(u,v)=\frac{u}{v}$ can be approximated via

$$\mathbb{E}[\varphi(u,v)] \approx \frac{\mu_u}{\mu_v} - \frac{\operatorname{cov}(u,v)}{\mu_v^2} + \frac{\operatorname{var}(v)\cdot\mu_u}{\mu_v^3} = \frac{\mu_u}{\mu_v} + \underbrace{\frac{\mu_u}{\mu_v}\cdot\left(\frac{\operatorname{var}(v)}{\mu_v^2} - \frac{\operatorname{cov}(u,v)}{\mu_u\cdot\mu_v}\right)}_{:=\varkappa_1(u,v)}. \tag{4.58}$$

Here, $\operatorname{cov}(u,v)$ corresponds to the covariance between u and v. Accordingly, the power of the contribution of a source of interest, belonging to the source images $\mathbf{x}$, to the beamformer output can be approximated via

$$p(\mathbf{x}) = \mathbb{E}\left[\frac{\left|\mathbf{w}_{\setminus\ell}^{\text{H}} \cdot \mathbf{x}\right|^2}{\left|\delta(\mathbf{x})\right|^2}\right] \approx \frac{\mathbb{E}\left[\left|\mathbf{w}_{\setminus\ell}^{\text{H}} \cdot \mathbf{x}\right|^2\right]}{\mathbb{E}\left[\left|\delta(\mathbf{x})\right|^2\right]},\ \mathbf{x} \in \{\mathbf{x}_0(\ell), \mathbf{x}_1(\ell), \boldsymbol{\nu}(\ell)\} \tag{4.59}$$

if the additive bias $\widetilde{\varkappa}_1(\mathbf{x}) = \varkappa_1\left(\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}\right|^2, |\delta(\mathbf{x})|^2\right)$ in (4.58) is much smaller than the fraction of the expected values. This condition is fulfilled if it holds that

$$\frac{\operatorname{var}\left(|\delta(\mathbf{x})|^2\right)}{\left(\mathbb{E}\left[|\delta(\mathbf{x})|^2\right]\right)^2} \approx \frac{\operatorname{cov}\left(\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}\right|^2, \left|\delta(\mathbf{x})\right|^2\right)}{\mathbb{E}\left[\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}\right|^2\right] \cdot \mathbb{E}\left[\left|\delta(\mathbf{x})\right|^2\right]}, \; \mathbf{x} \in \{\mathbf{x}_0(\ell), \mathbf{x}_1(\ell), \boldsymbol{\nu}(\ell)\}\,. \tag{4.60}$$

Thus, the approximation of the power of the signals at the beamformer output as a fraction of the expected value of the numerator $\mathbb{E}\left[\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}\right|^2\right]$ and the expected value of the denominator $\mathbb{E}\left[|\delta(\mathbf{x})|^2\right]$ in (4.59) is valid if (4.60) is fulfilled. This means that the approximation is valid if the numerator $\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}\right|^2$ and denominator $|\delta(\mathbf{x})|^2$ are positively correlated and the relationship of the covariance between the numerator and the denominator and the variance of the denominator and the expected values of the numerator and the denominator in (4.60) is fulfilled. In addition to that, (4.60) holds if the numerator $\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}\right|^2$ and the denominator $|\delta(\mathbf{x})|^2$ are not or only very weakly correlated and the variance of the denominator $|\delta(\mathbf{x})|^2$ is small compared to its expected value, which matches intuition. For the problem at hand, it is to be expected that the numerator $\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}\right|^2$ and the denominator $|\delta(\mathbf{x})|^2$ are positively correlated since both grow if the power of the input signal increases. Hence, the additive bias $\widetilde{\varkappa}_1(\mathbf{x})$ is only negligible if (4.60) is fulfilled.

Fig. 4.8 illustrates the validity of the approximation of the expected value of the fraction of two functions of the same random variables as the fraction of the corresponding expected values introduced in (4.59). Therefore, all time frequency bins of the interfering speaker's signal at the beamformer output are considered for all examples from the simulated dataset, which was introduced in Sec. 3.1. Note that only the time frequency bins are considered for which the interfering speaker's signal is dominant since the approximation is not needed for all other time frequency bins. In Fig. 4.8(a) the cumulative distribution function (CDF) of the error

$$e_1(\mathbf{x}_1(\ell)) = \frac{\mathbb{E}\left[\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}_1(\ell)\right|^2\right]}{\mathbb{E}\left[|\delta(\mathbf{x}_1(\ell))|^2\right]} - \mathbb{E}\left[\frac{\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}_1(\ell)\right|^2}{|\delta(\mathbf{x}_1(\ell))|^2}\right], \tag{4.61}$$

is shown, which is introduced by the approximation in (4.59). Here, a sharper transition from zero to one around zero is better, which means that large errors occur less often. It becomes obvious, that the error $e_1(\mathbf{x}_1(\ell))$ is close to zero for the vast majority of time frequency bins which speaks for the suitability of the approximation specified in (4.59). In addition to that, the quality of the approximation in (4.59) becomes better with growing block size, i.e., when the variance of the SCM estimates decreases.

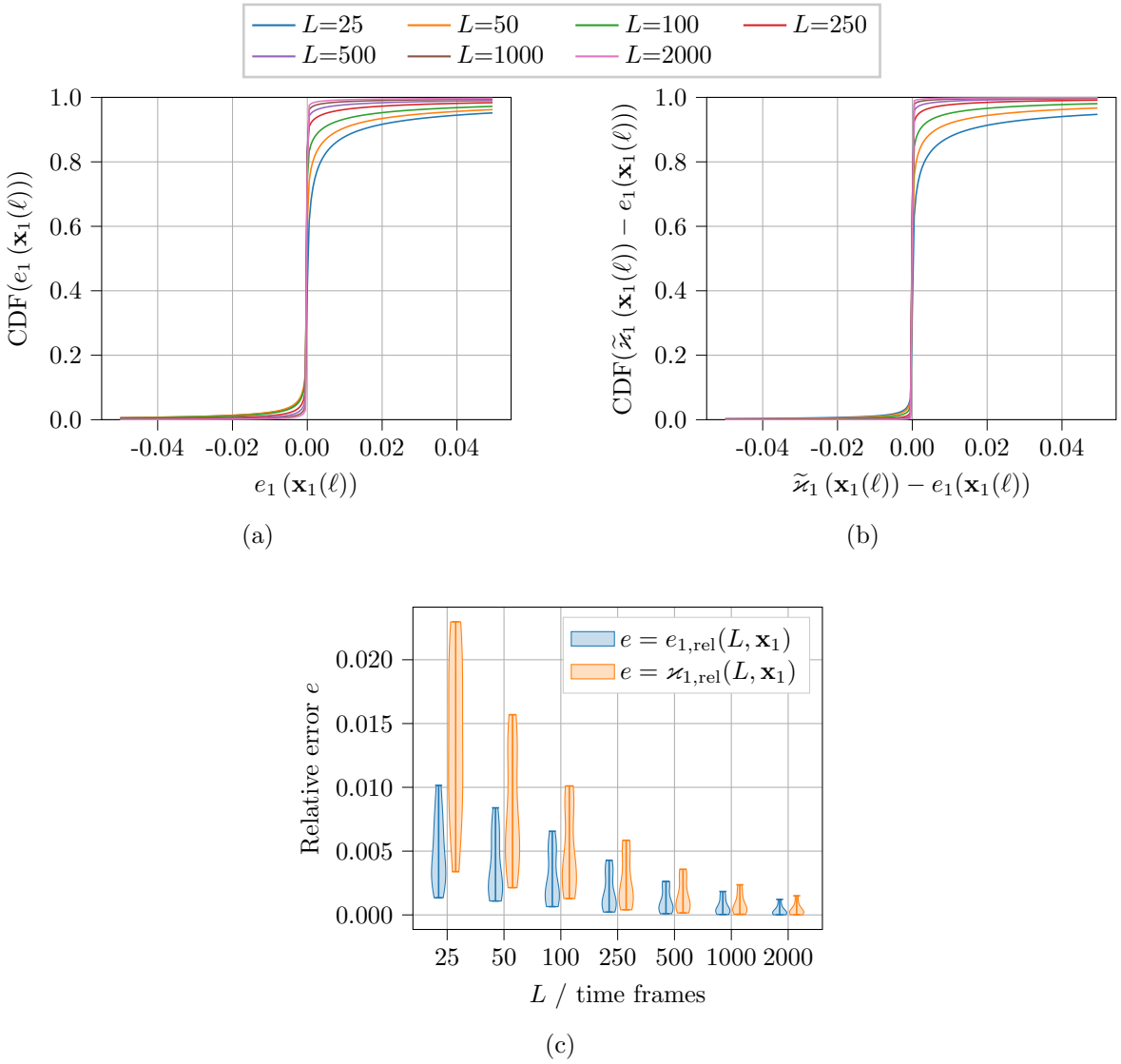


Figure 4.8: Accuracy of the approximation of the expected value of the fraction of two random variables as the fraction of expected values in (4.59) for the calculation of the power of the interfering speaker's signal at the beamformer output. (a) shows the cumulative distribution function (CDF) of the error $e_1(\mathbf{x}_1(\ell))$ due to the approximation in (4.59) and (b) the CDF of the its difference to its prediction $\widetilde{\varkappa}_1(\mathbf{x}_1(\ell))$ based on the second-order Taylor series expansion. (c) shows the distributions of the relative error of the overall energy of the interfering speaker's signal at the beamformer output $e_{1,\text{rel}}(L, \mathbf{x}_1)$ which follows from the approximation specified in (4.59) and $\varkappa_{1,\text{rel}}(L, \mathbf{x}_1)$ which follows from the additive bias $\widetilde{\varkappa}_1(\mathbf{x}_1(\ell))$ of the Taylor series approximation specified in (4.58).

In order to get deeper insights why the approximation in (4.59) works, Fig. 4.8(b)) shows the difference between the ground-truth error $e_1(\mathbf{x}_1(\ell))$, introduced by the approximation in (4.59), and the bias term $\widetilde{\varkappa}_1\left(\mathbf{x}_1(\ell)\right)$ following from the Taylor series expansion in (4.58). For most time frequency bins this difference is nearly zero which means that the additive bias $\widetilde{\varkappa}_1\left(\mathbf{x}_1(\ell)\right)$ fully describes the error following from the approximation of the expected value of a fraction of two random variables as fraction of the corresponding expected values. This implies that the validity of the approximation in (4.59) can be mainly explained by the fact that the condition stated in (4.60) is fulfilled. However, there are single time frequency bins for which the difference between the ground-truth error $e_1(\mathbf{x}_1(\ell))$ and the additive bias term $\widetilde{\varkappa}_1\left(\mathbf{x}_1(\ell)\right)$ is large. This especially occurs for small block sizes for which the variance of the involved random variables tends to be larger which may harm the quality of the second-order Taylor series expansion specified in (4.58).

Up to this point, the error introduced due to the approximation in (4.59), was only considered for single time frequency bins. However, the energies involved in the calculation of the SDR correspond to the sum over all time frequency bins. Hence, Fig. 4.8(c) visualizes the relative approximation error, as the sum over all time frequency bins,

$$e_{1,\mathrm{rel}}(L,\mathbf{x}_1) = \frac{\sum\limits_{b=0}^{N_\mathrm{B}-1}\sum\limits_{\ell=b\cdot L}^{(b+1)\cdot L-1}\sum\limits_{k=0}^{K-1}\left(\frac{\mathbb{E}\left[\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}(b,k)\cdot\mathbf{x}_1(\ell,k)\right|^2\right]}{\mathbb{E}\left[\left|\delta(\mathbf{x}_1(\ell,k))\right|^2\right]} - \mathbb{E}\left[\frac{\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}(b,k)\cdot\mathbf{x}_1(\ell,k)\right|^2}{\left|\delta(\mathbf{x}_1(\ell,k))\right|^2}\right]\right)}{\sum\limits_{b=0}^{N_\mathrm{B}-1}\sum\limits_{\ell=b\cdot L}^{(b+1)\cdot L-1}\sum\limits_{k=0}^{K-1}\mathbb{E}\left[\frac{\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}(b,k)\cdot\mathbf{x}_1(\ell,k)\right|^2}{\left|\delta(\mathbf{x}_1(\ell,k))\right|^2}\right]} \tag{4.62}$$

as well as the estimate of this relative error based on the second-order Taylor series expansion

$$\varkappa_{1,\mathrm{rel}}(L,\mathbf{x}_1) = \frac{\sum\limits_{b=0}^{N_\mathrm{B}-1}\sum\limits_{\ell=b\cdot L}^{(b+1)\cdot L-1}\sum\limits_{k=0}^{K-1}\widetilde{\varkappa}_1\left(\mathbf{x}_1(\ell,k)\right)}{\sum\limits_{b=0}^{N_\mathrm{B}-1}\sum\limits_{\ell=b\cdot L}^{(b+1)\cdot L-1}\sum\limits_{k=0}^{K-1}\mathbb{E}\left[\frac{\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}(b,k)\cdot\mathbf{x}_1(\ell,k)\right|^2}{\left|\delta(\mathbf{x}_1(\ell,k))\right|^2}\right]}. \tag{4.63}$$

It can be seen that the error tends to be smaller than one percentage point of the power of the interfering speaker's signal at the beamformer output. Since the SDR is typically expressed in dB, this tiny error will have a vanishingly small effect on the final result. Furthermore, it can be seen that the relative approximation error $e_{1,\mathrm{rel}}(L,\mathbf{x}_1)$ even tends to be smaller than its estimate $\varkappa_{1,\mathrm{rel}}(L,\mathbf{x}_1)$ based on the Taylor series expansion, with the gap between them becoming closer with growing block size. Nevertheless, the predicted size of the relative error which is calculated based on the Taylor series expansion also hints at the suitability of the approximation in (4.59).

Consider a general signal of interest $\mathbf{x} \in \{\mathbf{x}_0(\ell), \mathbf{x}_1(\ell), \boldsymbol{\nu}(\ell)\}$. Further, it is assumed that the interference-SCM estimates are Wishart distributed, with

$$\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell} \sim \mathcal{W}_M\left(L_{\setminus\ell}, \frac{1}{L_{\setminus\ell}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}\right). \tag{4.64}$$

Here, the matrix $\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}$ will be referred to as the unnormalized scale matrix in the following. As already mentioned before, the numerator of (4.59) corresponds to the power of a signal

at the beamformer output if the beamformer coefficients and the signal are statistically independent. Hence, the numerator follows the structure of (4.40) and (4.41), respectively. The numerator of (4.59) is given by

$$\begin{aligned}&\mathbb{E}\Big[\big|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}\cdot\mathbf{x}\big|^2\Big]\\ =&\frac{|\widehat{d}_0|^2}{L_{\setminus\ell}-M+1}\cdot\left(\frac{\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}}+(L_{\setminus\ell}-M)\cdot\frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}}{\left(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right)^2}\right),\end{aligned}\tag{4.65}$$

with the instantaneous SCM of the source of interest

$$\mathbf{R}_{\mathbf{x}}=\mathbb{E}\big[\mathbf{x}\cdot\mathbf{x}^{\mathrm{H}}\big]=\begin{cases}\sigma_0^2(\ell)\cdot\mathbf{R}_0, & \mathbf{x}=\mathbf{x}_0(\ell)\\ \sigma_1^2(\ell)\cdot\mathbf{R}_1, & \mathbf{x}=\mathbf{x}_1(\ell)\\ \sigma_\nu^2\cdot\mathbf{R}_\nu, & \mathbf{x}=\boldsymbol{\nu}(\ell)\end{cases}.\tag{4.66}$$

Note that the equivalent degrees of freedom $L_{\setminus\ell}$ and the equivalent $\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}$ depend on the time frame since the currently considered time frame is excluded from their calculation.

It holds for the expected value of the denominator in (4.59) that

$$\begin{aligned}&\mathbb{E}\big[|\delta(\mathbf{x})|^2\big]\\ =&\begin{cases}\mathbb{E}\left[\left|1+\gamma_{\mathrm{int}}(\ell)\cdot\left(\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{x}-\frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{x}\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}}\right)\right|^2\right], & \mathbf{x}\text{ is dominant}\\ 1, & \text{else}\end{cases}.\end{aligned}\tag{4.67}$$

If the source of interest is the dominant source for the considered time frame ℓ, the expected value of the denominator in (4.59) can be rewritten as

$$\begin{aligned}&\mathbb{E}\left[\left|1+\gamma_{\mathrm{int}}(\ell)\cdot\left(\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{x}-\frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{x}\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}}\right)\right|^2\right]\\ =&\mathbb{E}\left[\left|\frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\Big(\Big(1+\gamma_{\mathrm{int}}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{x}\Big)\cdot\mathbf{I}-\gamma_{\mathrm{int}}(\ell)\cdot\mathbf{x}\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\Big)\cdot\widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}}\right|^2\right]\\ =&\mathbb{E}\left[\left|\frac{\eta(\mathbf{x})}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}}\right|^2\right].\end{aligned}\tag{4.68}$$

By applying the Taylor series approximation from (4.58) to the denominator of (4.59), the power of the source of interest at the beamformer output $p(\mathbf{x})$ can be approximated as

$$
\begin{aligned}
p(\mathbf{x}) = \mathbb{E}\left[\frac{\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}\cdot\mathbf{x}\right|^2}{\left|\delta(\mathbf{x})\right|^2}\right] &\approx \frac{\mathbb{E}\left[\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}\cdot\mathbf{x}\right|^2\right]}{\mathbb{E}\left[\left|\delta(\mathbf{x})\right|^2\right]} = \frac{\mathbb{E}\left[\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}\cdot\mathbf{x}\right|^2\right]}{\mathbb{E}\left[\left|\frac{\eta(\mathbf{x})}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}}\right|^2\right]} \\
&\approx \frac{\mathbb{E}\left[\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}\cdot\mathbf{x}\right|^2\right]}{\frac{\mathbb{E}\left[\left|\eta(\mathbf{x})\right|^2\right]}{\mathbb{E}\left[\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right]}\cdot\left(1+\frac{\mathrm{var}\left(\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right)}{\left(\mathbb{E}\left[\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right]\right)^2}-\frac{\mathrm{cov}\left(\left|\eta(\mathbf{x})\right|^2,\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right)}{\mathbb{E}\left[\left|\eta(\mathbf{x})\right|^2\right]\cdot\mathbb{E}\left[\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right]}\right)} \\
&= \frac{\mathbb{E}\left[\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}\cdot\mathbf{x}\right|^2\right]}{\frac{\mathbb{E}\left[\left|\eta(\mathbf{x})\right|^2\right]}{\mathbb{E}\left[\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right]}} + \underbrace{\frac{\mathbb{E}\left[\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}\cdot\mathbf{x}\right|^2\right]}{\frac{\mathbb{E}\left[\left|\eta(\mathbf{x})\right|^2\right]}{\mathbb{E}\left[\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right]}}\cdot\frac{\frac{\mathrm{cov}\left(\left|\eta(\mathbf{x})\right|^2,\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right)}{\mathbb{E}\left[\left|\eta(\mathbf{x})\right|^2\right]\cdot\mathbb{E}\left[\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right]}-\frac{\mathrm{var}\left(\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right)}{\left(\mathbb{E}\left[\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right]\right)^2}}{1+\frac{\mathrm{var}\left(\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right)}{\left(\mathbb{E}\left[\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right]\right)^2}-\frac{\mathrm{cov}\left(\left|\eta(\mathbf{x})\right|^2,\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right)}{\mathbb{E}\left[\left|\eta(\mathbf{x})\right|^2\right]\cdot\mathbb{E}\left[\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right]}}}_{:=\varkappa_2(\mathbf{x})}.
\end{aligned}
\tag{4.69}
$$

Similar to (4.58), the Taylor series expansion results in a sum of a compound fraction of expected values and an additive bias $\varkappa_2(\mathbf{x})$. If the covariance between the components of the fraction $|\eta(\mathbf{x})|^2$ and $\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2$ in the denominator of the compound fraction and the variance of $\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2$ fulfill

$$
\frac{\mathrm{var}\left(\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right)}{\left(\mathbb{E}\left[\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right]\right)^2} \approx \frac{\mathrm{cov}\left(\left|\eta(\mathbf{x})\right|^2,\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right)}{\mathbb{E}\left[\left|\eta(\mathbf{x})\right|^2\right]\cdot\mathbb{E}\left[\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right]},
\tag{4.70}
$$

the additive bias $\varkappa_2(\mathbf{x})$ is much smaller than the compound fraction of expected values. In this case, the power of the source of interest at the beamformer output, which dominates the considered time frequency bin, can be approximated via

$$
p(\mathbf{x}) = \mathbb{E}\left[\frac{\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}\cdot\mathbf{x}\right|^2}{\left|\delta(\mathbf{x})\right|^2}\right] \approx \frac{\mathbb{E}\left[\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}\cdot\mathbf{x}\right|^2\right]}{\mathbb{E}\left[\left|\delta(\mathbf{x})\right|^2\right]} = \frac{\mathbb{E}\left[\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}\cdot\mathbf{x}\right|^2\right]}{\mathbb{E}\left[\left|\frac{\eta(\mathbf{x})}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}}\right|^2\right]} \approx \frac{\mathbb{E}\left[\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}\cdot\mathbf{x}\right|^2\right]}{\frac{\mathbb{E}\left[\left|\eta(\mathbf{x})\right|^2\right]}{\mathbb{E}\left[\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right]}}.
\tag{4.71}
$$

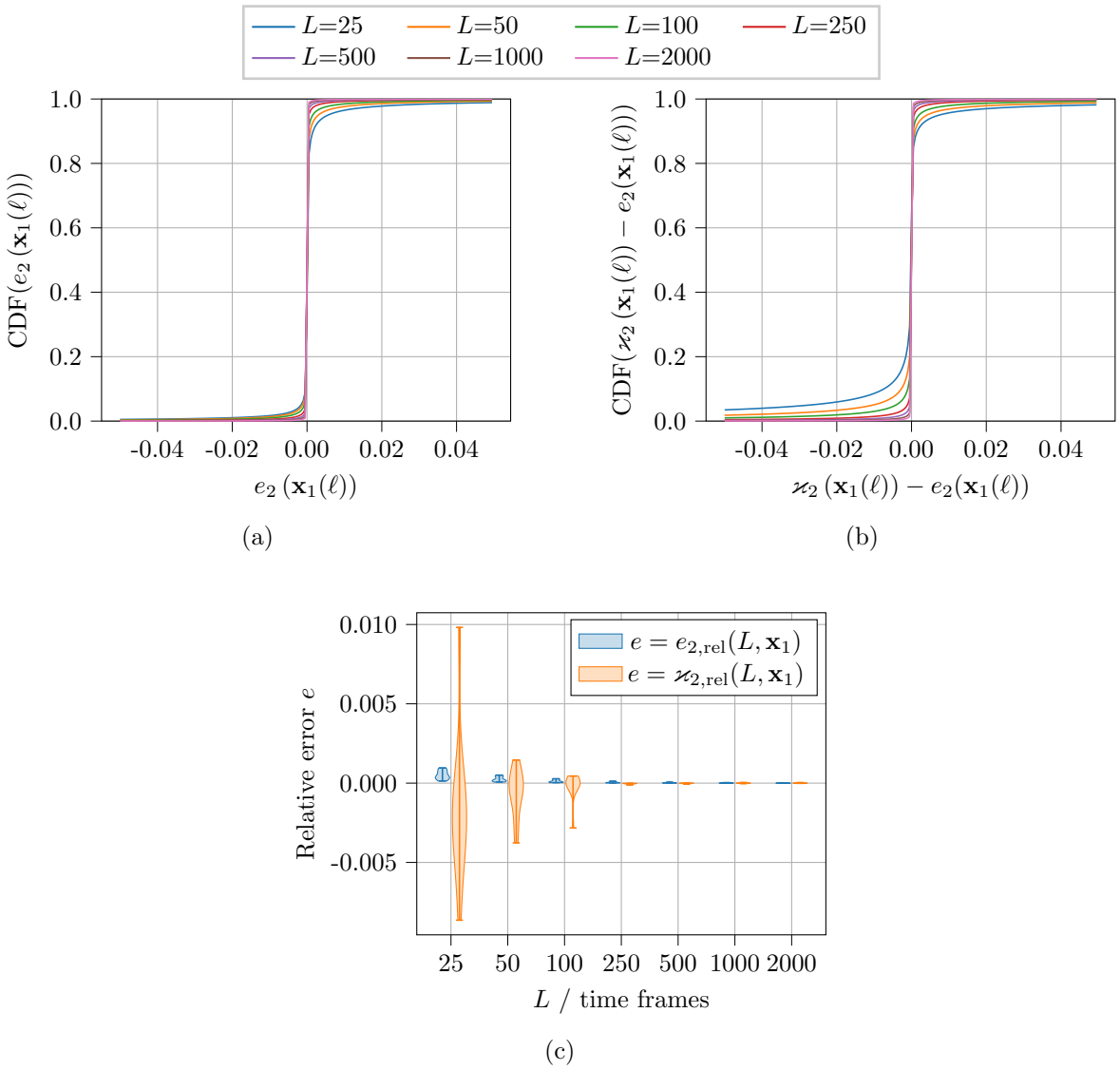


Figure 4.9: Accuracy of the approximation of the expected value of the compound fraction of functions of the same random variables as the compound fraction of expected values in (4.71) for the calculation of the power of the interfering speaker's signal at the beamformer output. (a) shows the CDF of the error $e_2(\mathbf{x}_1(\ell))$ due to the approximation in (4.71) and (b) the CDF of its difference to its prediction $\varkappa_2(\mathbf{x}_1(\ell))$ based on the second-order Taylor series expansion. (c) shows the distributions of the relative error of the energy of the interfering speaker's signal at the beamformer output $e_{2,\mathrm{rel}}(L, \mathbf{x}_1)$ which follows for the approximation in in (4.71) and $\varkappa_{2,\mathrm{rel}}(L, \mathbf{x}_1)$ which follows from the additive bias $\varkappa_2(\mathbf{x}_1(\ell))$ of the Taylor series approximation specified in (4.69).

Figure 4.9 visualizes the error caused the approximation of the expected value of a compound fraction as compound fraction of expected values in (4.71) in the same way as it was done in Fig. 4.8. The CDF of the error resulting from the approximation in (4.71)

$$e_2(\mathbf{x}_1(\ell)) = \frac{\mathbb{E}\left[\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}_1(\ell)\right|^2\right]}{\frac{\mathbb{E}\left[|\eta(\mathbf{x}_1(\ell))|^2\right]}{\mathbb{E}\left[\left|\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}\right|^2\right]}} - \frac{\mathbb{E}\left[\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}_1(\ell)\right|^2\right]}{\mathbb{E}\left[\left|\frac{\eta(\mathbf{x}_1(\ell))}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}\right|^2\right]}, \tag{4.72}$$

is depicted in Fig. 4.9(a). It becomes obvious that the error introduced by the approximation in (4.71) tends to be very close to zero so that this approximation can be seen to be a valid approximation of the power of signal at the beamformer output. By analyzing the difference between the error $e_2(\mathbf{x}_1(\ell))$ and the additive bias $\varkappa_2(\mathbf{x}_1(\ell))$ in (4.69), whose CDF is shown in Fig. 4.9(b), it can be seen that the error $e_2(\mathbf{x}_1(\ell))$ can be explained by the additive bias $\varkappa_2(\mathbf{x}_1(\ell))$ for the vast majority of time frequency bins. The rare larger differences between the error $e_2(\mathbf{x}_1(\ell))$ and the additive bias $\varkappa_2(\mathbf{x}_1(\ell))$ can be explained by the increasing error of the Taylor series expansion with growing variance of the involved random variables which was already discussed w.r.t. Fig. 4.8. Finally, Fig. 4.9(c) depicts the CDF of the relative error

$$e_{2,\mathrm{rel}}(L, \mathbf{x}_1) = \frac{\sum_{b=0}^{N_\mathrm{B}-1} \sum_{\ell=b\cdot L}^{(b+1)\cdot L-1} \sum_{k=0}^{K-1} \left(\frac{\mathbb{E}\left[\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}(b,k) \cdot \mathbf{x}_1(\ell,k)\right|^2\right]}{\frac{\mathbb{E}\left[|\eta(\mathbf{x}_1(\ell,k))|^2\right]}{\mathbb{E}\left[\left|\widehat{\mathbf{d}}^{\mathrm{H}}(k) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}(k) \cdot \widehat{\mathbf{d}}(k)\right|^2\right]}} - \frac{\mathbb{E}\left[\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}(b,k) \cdot \mathbf{x}_1(\ell,k)\right|^2\right]}{\mathbb{E}\left[\left|\frac{\eta(\mathbf{x}_1(\ell,k))}{\widehat{\mathbf{d}}^{\mathrm{H}}(k) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}(k) \cdot \widehat{\mathbf{d}}(k)}\right|^2\right]} \right)}{\sum_{b=0}^{N_\mathrm{B}-1} \sum_{\ell=b\cdot L}^{(b+1)\cdot L-1} \sum_{k=0}^{K-1} \frac{\mathbb{E}\left[\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}(b,k) \cdot \mathbf{x}_1(\ell,k)\right|^2\right]}{\mathbb{E}\left[\left|\frac{\eta(\mathbf{x}_1(\ell,k))}{\widehat{\mathbf{d}}^{\mathrm{H}}(k) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}(k) \cdot \widehat{\mathbf{d}}(k)}\right|^2\right]}}, \tag{4.73}$$

which reflects the error of the interfering speaker's total energy due to the approximation in (4.71), as well as the prediction of this relative error

$$\varkappa_{2,\mathrm{rel}}(L, \mathbf{x}_1) = \frac{\sum_{b=0}^{N_\mathrm{B}-1} \sum_{\ell=b\cdot L}^{(b+1)\cdot L-1} \sum_{k=0}^{K-1} \varkappa_2(\mathbf{x}_1(\ell,k))}{\sum_{b=0}^{N_\mathrm{B}-1} \sum_{\ell=b\cdot L}^{(b+1)\cdot L-1} \sum_{k=0}^{K-1} \frac{\mathbb{E}\left[\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}(b,k) \cdot \mathbf{x}_1(\ell,k)\right|^2\right]}{\mathbb{E}\left[\left|\frac{\eta(\mathbf{x}_1(\ell,k))}{\widehat{\mathbf{d}}^{\mathrm{H}}(k) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}(k) \cdot \widehat{\mathbf{d}}(k)}\right|^2\right]}}, \tag{4.74}$$

based on the second-order Taylor series expansion. It becomes obvious that the error is approximately zero in all cases and is even significantly smaller than its prediction based on the second-order Taylor series expansion. In conclusion, the approximation in (4.69) is suitable for the calculation of the energy of a source's signal at the beamformer output.

Calculating the expected values $\mathbb{E}\left[|\eta(\mathbf{x})|^2\right]$ and $\mathbb{E}[|\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}|^2]$ in (4.71) gives a closed-form approximation of the denominator $\mathbb{E}\left[|\delta(\mathbf{x})|^2\right]$ in (4.59) for the case where the source of interest is dominant:

$$
\begin{aligned}
\mathbb{E}\big[|\delta(\mathbf{x})|^2\big] \approx & \frac{\mathbb{E}\big[|\eta(\mathbf{x})|^2\big]}{\mathbb{E}\left[\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right]} \\
=& 1 + \frac{2\cdot\gamma_{\mathrm{int}}(\ell)\cdot L_{\backslash\ell}}{\big(L_{\backslash\ell}-M+1\big)}\cdot\left(\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\} - \frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}}\right) \\
&+ \frac{\gamma_{\mathrm{int}}^2(\ell)\cdot L_{\backslash\ell}^2}{\big(L_{\backslash\ell}-M+1\big)\cdot\big(L_{\backslash\ell}-M\big)^2}\cdot\left(\big(L_{\backslash\ell}-M\big)\cdot\left(\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\right)^2\right. \\
&+ 2\cdot\big(L_{\backslash\ell}-M\big)\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\} \\
&+ 3\cdot\big(L_{\backslash\ell}-M-1\big)\cdot\frac{\left(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\right)^2}{\left(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\right)^2} \\
&- \big(4\cdot\big(L_{\backslash\ell}-M\big)-2\big)\cdot\frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}} \\
&- \big(2\cdot\big(L_{\backslash\ell}-M\big)-1\big)\cdot\frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}} \\
&+ \big(L_{\backslash\ell}-M-1\big)\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\Big\} \\
&+ \big(L_{\backslash\ell}-M-1\big)\cdot\frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}}{\left(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\right)^2} \\
&- 2\cdot\big(L_{\backslash\ell}-M-1\big)\cdot\frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}} \\
&+ \mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\Big\} \\
&\left.- \frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}}\right),
\end{aligned}
\tag{4.75}
$$

with the ground-truth SCM of the signals of the sources that are not of interest

$$
\mathbf{R}_{\mathbf{r}} = \sigma_0^2(\ell)\cdot\mathbf{R}_0 + \sigma_1^2(\ell)\cdot\mathbf{R}_1 + \sigma_\nu^2\cdot\mathbf{R}_\nu - \mathbf{R}_{\mathbf{x}}. \tag{4.76}
$$

Refer to Appendix A.5 for a detailed derivation of this expected value.

4.4 Analysis of the closed-form approximation of the output SDR

In the following, the closed-form approximation of the output SDR is analyzed to get deeper insights into the effects of estimating the interference SCM from a finite sample size on MVDR beamforming. For explanation of the matrix algebra theorems used for this, refer to Appendix A.6.

4.4.1 Convergence behavior

First, the numerator of the signals' power at the beamformer output $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}|^2]$ in (4.59) is considered, which can be calculated via (4.65). It was shown in Sec. 4.2 that the interference-SCM estimates converge towards their expected value with growing block size and, therefore, also the beamformer coefficients converge towards their expected value. This is also reflected by (4.65), for which

$$\begin{aligned}
&\lim_{L_{\setminus\ell}\to\infty} \mathbb{E}\Big[\big|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}\big|^2\Big] \\
=&\lim_{L_{\setminus\ell}\to\infty} \left(\frac{|\widehat{d}_0|^2}{L_{\setminus\ell} - M + 1} \cdot \left(\frac{\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{R}_{\mathbf{x}}\Big\}}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}} + (L_{\setminus\ell} - M) \cdot \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{R}_{\mathbf{x}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}{\Big(\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}\Big)^2} \right) \right) \\
=&|\widehat{d}_0|^2 \cdot \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{R}_{\mathbf{x}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}{\Big(\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}\Big)^2}
\end{aligned} \tag{4.77}$$

holds. Utilizing the expected value of the beamformer coefficients, which is stated in (4.31), it follows that

$$\lim_{L_{\setminus\ell}\to\infty} \mathbb{E}\Big[\big|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}\big|^2\Big] = \mathrm{tr}\Big\{\big(\mathbb{E}\big[\mathbf{w}_{\setminus\ell}\big]\big)^{\mathrm{H}} \cdot \mathbf{R}_{\mathbf{x}} \cdot \mathbb{E}\big[\mathbf{w}_{\setminus\ell}\big]\Big\} \tag{4.78}$$

Consequently, the numerator of a source's signal at the beamformer output converges towards the power that results from applying the expected value of the beamformer coefficients to a signal as the block size L and, therefore, the equivalent degrees of freedom $L_{\setminus\ell}$ increase. Moreover, it holds that

$$\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-\frac{1}{2}} \cdot \mathbf{R}_{\mathbf{x}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-\frac{1}{2}}\Big\} \geq \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-\frac{1}{2}} \cdot \Big(\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-\frac{1}{2}} \cdot \mathbf{R}_{\mathbf{x}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-\frac{1}{2}}\Big) \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-\frac{1}{2}} \cdot \widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-\frac{1}{2}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-\frac{1}{2}} \cdot \widehat{\mathbf{d}}} \tag{4.79}$$

$$\Leftrightarrow \frac{\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{R}_{\mathbf{x}}\Big\}}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}} \geq \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{R}_{\mathbf{x}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}{\Big(\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}\Big)^2}, \tag{4.80}$$

where $(\cdot)^{-\frac{1}{2}}$ denotes the inverse of the matrix square root with $\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-\frac{1}{2}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-\frac{1}{2}} = \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}$. This results from the fact that the trace of a positive semidefinite matrix corresponds to the sum

of its eigenvalues and the Rayleigh quotient on the left hand side is upper-bounded by the largest eigenvalue (see Appendix A.6). From (4.80) it follows that the numerator of the signals' power at the beamformer output $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}|^2]$ in (4.59) always is larger than its limit value specified in (4.77) since

$$\begin{aligned}
&\mathbb{E}\Big[\big|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}\big|^2\Big] \\
=&\frac{|\widehat{d}_0|^2}{L_{\setminus\ell} - M + 1} \cdot \left(\frac{\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{R}_{\mathbf{x}}\Big\}}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}} + (L_{\setminus\ell} - M) \cdot \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{R}_{\mathbf{x}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}{\Big(\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}\Big)^2} \right) \\
\geq&\frac{|\widehat{d}_0|^2}{L_{\setminus\ell} - M + 1} \cdot \left((L_{\setminus\ell} - M + 1) \cdot \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{R}_{\mathbf{x}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}{\Big(\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}\Big)^2} \right) \\
=&|\widehat{d}_0|^2 \cdot \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{R}_{\mathbf{x}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}{\Big(\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}\Big)^2}.
\end{aligned} \tag{4.81}$$

In addition to that, it can be shown that the numerator of the power at the beamformer output is monotonically decreasing with increasing degrees of freedom $L_{\setminus\ell}$ of the Wishart approximation of the interference-SCM estimates. To this end, the derivative of (4.65) w.r.t. $L_{\setminus\ell}$ is considered which is given by

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}L_{\setminus\ell}} \mathbb{E}\Big[\big|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}\big|^2\Big] \\
=&\frac{\mathrm{d}}{\mathrm{d}L_{\setminus\ell}} \left(\frac{|\widehat{d}_0|^2}{L_{\setminus\ell} - M + 1} \cdot \left(\frac{\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{R}_{\mathbf{x}}\Big\}}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}} + (L_{\setminus\ell} - M) \cdot \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{R}_{\mathbf{x}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}{\Big(\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}\Big)^2} \right) \right) \\
=&\frac{|\widehat{d}_0|^2}{\big(L_{\setminus\ell} - M + 1\big)^2} \cdot \left(-\frac{\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{R}_{\mathbf{x}}\Big\}}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}} + \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{R}_{\mathbf{x}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}{\Big(\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}\Big)^2} \right).
\end{aligned} \tag{4.82}$$

By applying (4.80), it follows that

$$\frac{\mathrm{d}}{\mathrm{d}L_{\setminus\ell}} \mathbb{E}\Big[\big|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}\big|^2\Big] \leq 0. \tag{4.83}$$

From the monotonically decreasing behavior of the numerator of the signals' power at the beamformer output, it follows that the interference as well as the components of the target signal which are not represented by the steering vector become more suppressed as the block size and, therefore, the equivalent degrees of freedom $L_{\setminus\ell}$ of the Wishart approximation of the interference-SCM estimate grow.

It can be shown that the numerator of the power of the target speaker's signal at the beamformer output $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}_0(\ell)|^2]$ is not affected by the equivalent degrees of freedom $L_{\setminus\ell}$ due to the distortionless response constraint of the MVDR beamformer if the ground-truth

SCM of the target speaker's signals has rank one, i.e., $\mathbf{R}_0{=}\mathbf{d}{\cdot}\mathbf{d}^{\mathrm{H}}$, and the steering vector is perfect, i.e., $\widehat{\mathbf{d}} = \mathbf{d}$. To this end, $\mathbf{R}_{\mathbf{x}}{=}\mathbf{R}_0{=}\sigma_0(\ell){\cdot}\mathbf{d}{\cdot}\mathbf{d}^{\mathrm{H}}$, and $\widehat{\mathbf{d}}{=}\mathbf{d}$ are inserted into (4.65), which leads to

$$\mathbb{E}\Big[\big|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}_0(\ell)\big|^2\Big] = |\widehat{d}_0|^2 \cdot \sigma_0^2(\ell). \tag{4.84}$$

However, under reverberant conditions the SCM of the target speaker's signal typically is not of rank one but rather a full-rank matrix. In order to differentiate between the behavior of the components of the target speaker's signal, that are represented by the target-SCM estimate $\widehat{\mathbf{R}}_{\mathrm{tar}}{=}\widehat{\mathbf{d}} \cdot \widehat{\mathbf{d}}^{\mathrm{H}}$, and the behavior of the components, that are not represented by the target-SCM estimate, the ground-truth SCM of the target speaker's signals is decomposed in the following way:

$$\mathbf{R}_{0,\mathrm{steer}} = \frac{1}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \mathbf{R}_0^{-1} \cdot \widehat{\mathbf{d}}} \cdot \widehat{\mathbf{d}} \cdot \widehat{\mathbf{d}}^{\mathrm{H}}, \tag{4.85}$$

$$\mathbf{R}_{0,\mathrm{rest}} = \mathbf{R}_0 - \mathbf{R}_{0,\mathrm{steer}}. \tag{4.86}$$

Here, $\mathbf{R}_{0,\mathrm{steer}}$ corresponds to the signal components that are represented by the target-SCM estimate and $\mathbf{R}_{0,\mathrm{rest}}$ corresponds to the signal components that are not represented by the target-SCM estimate. This form of decomposition has the advantage that the resulting matrix $\mathbf{R}_{0,\mathrm{rest}}$ is ensured to be positive semidefinite which simplifies the interpretation of the behavior of the corresponding components. Refer to Appendix A.6 for a derivation of the decomposition specified in (4.85) and (4.86). Note that $\mathbf{R}_{0,\mathrm{steer}}{=}\lambda_{\max}(\mathbf{R}_0){\cdot}\mathbf{d}{\cdot}\mathbf{d}^{\mathrm{H}}$ holds if the steering vector estimate $\widehat{\mathbf{d}}$ perfectly matches the dominant eigenvector $\mathbf{d}$ of $\mathbf{R}_0$ with $\lambda_{\max}(\mathbf{R}_0)$ being the largest eigenvalue of $\mathbf{R}_0$. With (4.65) and the decomposition of the ground-truth SCM $\mathbf{R}_0$ in (4.85) and (4.86), it follows that

$$\begin{aligned}\mathbb{E}\Big[\big|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}_0(\ell)\big|^2\Big] =& \frac{\sigma_0^2(\ell) \cdot |\widehat{d}_0|^2}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \mathbf{R}_0^{-1} \cdot \widehat{\mathbf{d}}} + \frac{\sigma_0^2(\ell) \cdot |\widehat{d}_0|^2}{L_{\setminus\ell} - M + 1} \cdot \left(\frac{\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{R}_{0,\mathrm{rest}}\Big\}}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}} \right. \\ &\left. + (L_{\setminus\ell} - M) \cdot \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{R}_{0,\mathrm{rest}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}{\Big(\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}\Big)^2} \right)\end{aligned} \tag{4.87}$$

holds for the numerator in (4.59) if the target speaker's signal is considered. Following the discussion above, all components of the target signal which are not represented by the steering vector, i.e., the components belonging to $\mathbf{R}_{0,\mathrm{rest}}$, become more suppressed with growing block size while the power of the components that are represented by the steering vector is independent from the block size.

It is to be mentioned that the finite sample size effects on beamforming are non-negligible on the numerator of (4.59) although $(L_{\setminus\ell} - M)/(L_{\setminus\ell} - M + 1)$ and $1/(L_{\setminus\ell} - M + 1)$ become close to one and, respectively, quite small even for small values of the equivalent degrees of freedom $L_{\setminus\ell}$. This can be attributed to the fact that the equivalent degrees of freedom $L_{\setminus\ell}$ typically are much smaller than the block size. Furthermore, the $\geq$ relation in (4.80) usually

corresponds to $\gg$ when considering interfering signals. This can be seen from (4.79) for the interfering speaker, i.e., by inserting $\mathbf{R}_\mathrm{x}{=}\mathbf{R}_1$ into it. The left hand side of (4.79) contains the sum of all eigenvalues of the matrix $\widetilde{\mathbf{R}}_1 = \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-\frac{1}{2}} \cdot \mathbf{R}_1 \cdot \boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-\frac{1}{2}}$ via the trace-term and the right hand side of (4.79) contains the Rayleigh quotient belonging to the same matrix. If the dominant eigenvector of $\widetilde{\mathbf{R}}_1$ would be well aligned with the vector $\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-\frac{1}{2}} \cdot \widehat{\mathbf{d}}$ involved in the calculation of the Rayleigh quotient, the value of the Rayleigh quotient would be close to the largest eigenvalue of $\widetilde{\mathbf{R}}_1$. This results from the stationarity of the Rayleigh quotient at an eigenvector [62]. When the interfering speaker's signal at the beamformer output is considered and an accurate estimate of the steering vector is given, the dominant eigenvector of $\widetilde{\mathbf{R}}_1$ and the vector $\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-\frac{1}{2}} \cdot \widehat{\mathbf{d}}$ usually are not well aligned. In this case, the value of the Rayleigh quotient in (4.79) typically is much smaller than the value of the trace in (4.79).

Based on the previous discussion, the numerator of (4.59) can be decomposed into a transient component $p_\mathrm{t}(\mathbf{x})$, whose influence diminishes with growing degrees of freedom $L_{\setminus\ell}$, and a steady-state component $p_\infty(\mathbf{x})$, towards which the power at the beamformer output converges with growing degrees of freedom $L_{\setminus\ell}$, with

$$\begin{aligned}&\mathbb{E}\Big[\big|\mathbf{w}_{\setminus\ell}^\mathrm{H}{\cdot}\mathbf{x}\big|^2\Big]\\ =&\frac{1}{L_{\setminus\ell}{-}M{+}1}{\cdot}\underbrace{\frac{|\widehat{d}_0|^2{\cdot}\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}{\cdot}\mathbf{R}_\mathbf{x}\Big\}}{\widehat{\mathbf{d}}^\mathrm{H}{\cdot}\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}{\cdot}\widehat{\mathbf{d}}}}_{:=p_\mathrm{t}(\mathbf{x})}+\frac{L_{\setminus\ell}{-}M}{L_{\setminus\ell}{-}M{+}1}{\cdot}\underbrace{\frac{|\widehat{d}_0|^2{\cdot}\widehat{\mathbf{d}}^\mathrm{H}{\cdot}\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}{\cdot}\mathbf{R}_\mathbf{x}{\cdot}\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}{\cdot}\widehat{\mathbf{d}}}{\Big(\widehat{\mathbf{d}}^\mathrm{H}{\cdot}\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}{\cdot}\widehat{\mathbf{d}}\Big)^2}}_{:=p_\infty(\mathbf{x})}.\end{aligned}\tag{4.88}$$

In the following, the role of the denominator in (4.59) is investigated where at first the value range of the denominator is considered. For this purpose, only the case where the source of interest dominates the currently considered time frequency bin is considered since the denominator in (4.59) always has a value of one if the source of interest does not dominate the time frequency bin under consideration. Again, the interfering speaker's contribution to the output of the beamformer is discussed first.

If the interfering speaker dominates the currently considered time frequency bin, it follows that $\mathbf{y}(\ell){\approx}\mathbf{x}_1(\ell)$ and, therefore,

$$\begin{aligned}&\mathbb{E}\big[|\delta(\mathbf{x}_1(\ell))|^2\big]\\ \approx&\mathbb{E}\left[\left|1{+}\gamma_\mathrm{int}(\ell){\cdot}\left(\mathbf{x}_1^\mathrm{H}(\ell){\cdot}\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}{\cdot}\mathbf{x}_1(\ell){-}\frac{\widehat{\mathbf{d}}^\mathrm{H}{\cdot}\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}{\cdot}\mathbf{x}_1(\ell){\cdot}\mathbf{x}_1^\mathrm{H}(\ell){\cdot}\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}{\cdot}\widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^\mathrm{H}{\cdot}\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}{\cdot}\widehat{\mathbf{d}}}\right)\right|^2\right].\end{aligned}\tag{4.89}$$

Utilizing the Cauchy-Schwarz inequality

$$\widetilde{\mathbf{a}}^\mathrm{H}\cdot\widetilde{\mathbf{a}}\cdot\widetilde{\mathbf{b}}^\mathrm{H}\cdot\widetilde{\mathbf{b}}\geq\left|\widetilde{\mathbf{a}}^\mathrm{H}\cdot\widetilde{\mathbf{b}}\right|^2\tag{4.90}$$

with $\widetilde{\mathbf{a}}=\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-\frac{1}{2}}{\cdot}\mathbf{x}_1(\ell)$ and $\widetilde{\mathbf{b}}=\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-\frac{1}{2}}{\cdot}\widehat{\mathbf{d}}$, leads to

$$\mathbf{x}_1^\mathrm{H}(\ell){\cdot}\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}{\cdot}\mathbf{x}_1(\ell)\cdot\widehat{\mathbf{d}}^\mathrm{H}{\cdot}\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}{\cdot}\widehat{\mathbf{d}}\geq\widehat{\mathbf{d}}^\mathrm{H}{\cdot}\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}{\cdot}\mathbf{x}_1(\ell){\cdot}\mathbf{x}_1^\mathrm{H}(\ell){\cdot}\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}{\cdot}\widehat{\mathbf{d}}.\tag{4.91}$$

This results in

$$\mathbf{x}_1^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}_1(\ell) \geq \frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}_1(\ell)\cdot\mathbf{x}_1^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}}, \tag{4.92}$$

from which

$$\mathbb{E}\left[|\delta(\mathbf{x}_1(\ell))|^2\right] \geq 1 \tag{4.93}$$

follows. Thus, the denominator in (4.59) reflects an improved suppression of the interference when applying a beamformer to a statistically dependent interference signal.

A central component of the denominator in (4.59), which is given by (4.75), is the dot product $\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}$, with $\mathbf{R}_{\mathbf{x}}{=}\sigma_1^2(\ell)\cdot\mathbf{R}_1$ when considering the interfering speaker's signal at the output of the beamformer. Additionally, it is assumed that the unnormalized scale matrix $\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}$ of the approximate Wishart distribution of the interference-SCM estimates in (4.64) is given by

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell} &= \sum_{\substack{\widetilde{\ell}=0\\ \widetilde{\ell}\neq\ell}}^{L-1}\gamma_{\mathrm{int}}\left(\widetilde{\ell}\right)\cdot\sigma_0^2\left(\widetilde{\ell}\right)\cdot\mathbf{R}_0 + \sum_{\substack{\widetilde{\ell}=0\\ \widetilde{\ell}\neq\ell}}^{L-1}\gamma_{\mathrm{int}}\left(\widetilde{\ell}\right)\cdot\sigma_1^2\left(\widetilde{\ell}\right)\cdot\mathbf{R}_1 + \sum_{\substack{\widetilde{\ell}=0\\ \widetilde{\ell}\neq\ell}}^{L-1}\gamma_{\mathrm{int}}\left(\widetilde{\ell}\right)\cdot\sigma_\nu^2\left(\widetilde{\ell}\right)\cdot\mathbf{R}_\nu \\ &= \sum_{\substack{\widetilde{\ell}=0\\ \widetilde{\ell}\neq\ell}}^{L-1}\gamma_{\mathrm{int}}\left(\widetilde{\ell}\right)\cdot\sigma_1^2\left(\widetilde{\ell}\right)\cdot\check{\mathbf{R}}_1,\end{aligned} \tag{4.94}$$

with

$$\check{\mathbf{R}}_1 = \mathbf{R}_1 + \frac{\sum\limits_{\substack{\widetilde{\ell}=0\\ \widetilde{\ell}\neq\ell}}^{L-1}\gamma_{\mathrm{int}}\left(\widetilde{\ell}\right)\cdot\sigma_0^2\left(\widetilde{\ell}\right)}{\sum\limits_{\substack{\widetilde{\ell}=0\\ \widetilde{\ell}\neq\ell}}^{L-1}\gamma_{\mathrm{int}}\left(\widetilde{\ell}\right)\cdot\sigma_1^2\left(\widetilde{\ell}\right)}\cdot\mathbf{R}_0 + \frac{\sum\limits_{\substack{\widetilde{\ell}=0\\ \widetilde{\ell}\neq\ell}}^{L-1}\gamma_{\mathrm{int}}\left(\widetilde{\ell}\right)\cdot\sigma_\nu^2\left(\widetilde{\ell}\right)}{\sum\limits_{\substack{\widetilde{\ell}=0\\ \widetilde{\ell}\neq\ell}}^{L-1}\gamma_{\mathrm{int}}\left(\widetilde{\ell}\right)\cdot\sigma_1^2\left(\widetilde{\ell}\right)}\cdot\mathbf{R}_\nu. \tag{4.95}$$

This assumption is motivated by (4.24). Note that the normalization by the sum over the interference mask estimates is neglected for sake of a simpler notation. It follows from (4.94) that

$$\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\sigma_1^2(\ell)\cdot\mathbf{R}_1 = \frac{\sigma_1^2(\ell)}{\sum\limits_{\substack{\widetilde{\ell}=0\\ \widetilde{\ell}\neq\ell}}^{L-1}\gamma_{\mathrm{int}}\left(\widetilde{\ell}\right)\cdot\sigma_1^2\left(\widetilde{\ell}\right)}\cdot\check{\mathbf{R}}_1^{-1}\cdot\mathbf{R}_1. \tag{4.96}$$

The value of the fraction in (4.96) usually becomes large if the block size is small and the power of the interfering speaker's signal is dominant w.r.t. the contribution of the other time frames to the interference-SCM estimate. As the block size grows, the denominator of the fraction in (4.96) becomes larger for typically distributed powers of the speech source signal so that (4.96) converges towards a value of zero. Hence, it follows from (4.75) that the

denominator in (4.59) converges towards a value of one when the block size increases. Moreover, it follows from (4.75) that the denominator in (4.59) decreases as the equivalent degrees of freedom $L_{\setminus\ell}$ of the Wishart approximation of the interference-SCM estimates increase. Both effects reflect a diminishing behavior of the suppression of the signal components belonging to time frequency bins which dominate the interference-SCM estimates as the block size increases.

When considering the effect of the denominator in (4.59) on the power of the target speaker's signal at the beamformer output, it is expected that it is smaller compared to the effect on the power of the interference. This can be explained by the distortionless response constraint of the MVDR beamformer. To show this, first assume that the ground-truth target SCM $\mathbf{R}_0{=}\mathbf{d}\cdot\mathbf{d}^{\mathrm{H}}$ is of rank-1 and the steering vector estimate $\widehat{\mathbf{d}}{=}\mathbf{d}$ corresponds to the eigenvector $\mathbf{d}$ of the ground-truth target SCM $\mathbf{R}_0$. In this case, it follows from (4.75) that the denominator of (4.59) is given by

$$\begin{aligned}&\mathbb{E}\big[|\delta(\mathbf{x}_0(\ell))|^2\big]\\ \approx&1+\underbrace{\frac{\gamma_{\mathrm{int}}^2(\ell)\cdot L_{\setminus\ell}^2\cdot\Big(\mathbf{d}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{d}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\Big\}-\mathbf{d}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{d}\Big)}{\big(L_{\setminus\ell}-M+1\big)\cdot\big(L_{\setminus\ell}-M\big)^2}}_{:=\widetilde{\delta}_{\mathrm{steer}}(\mathbf{x}_0)}.\end{aligned}\tag{4.97}$$

In this case, most terms of (4.59) cancel out, which reflects the distortionless response constraint of the MVDR beamformer. However, there still is an additional suppression of the target speaker's signal since $\mathbb{E}\big[|\delta(\mathbf{x}_1(\ell))|^2\big]\geq 1$ which can be attributed to the cross terms between the target speaker's signal and the signals belonging to the other sources, which are represented by the SCM $\mathbf{R}_{\mathbf{r}}$ as defined in (4.76). This additional attenuation generally is much smaller than for the interfering signal since the other source signals tend to have rather small powers if the target speaker is dominant due to the approximate W-disjoint orthogonality of speech. Furthermore, the interference mask $\gamma_{\mathrm{int}}(\ell,k)$ should have a value that is close to zero if a time frequency bin is dominated by the target speaker's signal.

In general, the ground-truth SCMs belonging to the target speaker have full rank. Utilizing the composition of the target SCM $\mathbf{R}_0$ into a component $\mathbf{R}_{0,\mathrm{steer}}$ which is represented by the steering vector and its remainder $\mathbf{R}_{0,\mathrm{rest}}$ via (4.85) and (4.86), the denominator corresponds to the sum of (4.75) with $\mathbf{R}_{\mathbf{x}}{=}\mathbf{R}_{0,\mathrm{rest}}$ and $\widetilde{\delta}_{\mathrm{steer}}(\mathbf{x}_0)$, with $\widetilde{\delta}_{\mathrm{steer}}(\mathbf{x}_0)$ usually being much smaller than the other term. If the quality of the beamformer's steering is good, i.e., the steering vector is well aligned with the dominant eigenvector of the target speaker's SCM $\mathbf{R}_0$, the component $\mathbf{R}_{0,\mathrm{steer}}$ of the target SCM constitutes a significant proportion of $\mathbf{R}_0$ so that the denominator will generally be much smaller for the target speaker than for the interfering speaker.

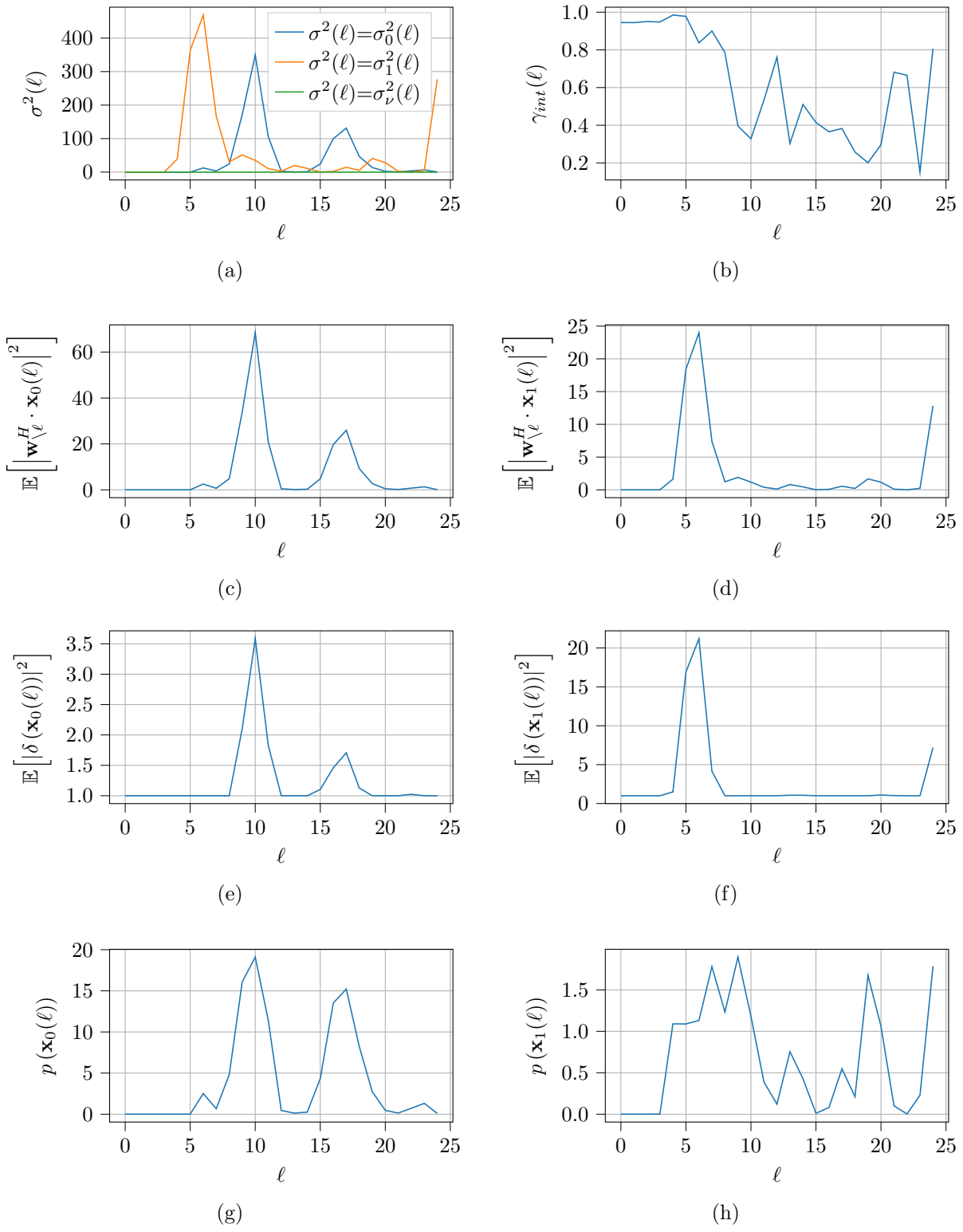


Figure 4.10: Interplay of the numerator and the denominator of the closed-form approximation of the power of the sources' signals at the beamformer output $p(\mathbf{x})$ in (4.59). The numerator $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}_0(\ell)|^2]$ (c), denominator $\mathbb{E}[|\delta(\mathbf{x}_0(\ell))|^2]$ (e) and the resulting power of the target signal at the beamformer output $p(\mathbf{x}_0(\ell))$ (g) are shown for each time frame within a block. The corresponding quantities for the interfering speaker's signal are shown in (d), (f) and (h). (a) depicts the powers $\sigma_i(\ell)$, $i \in \{0, 1, \nu\}$, of the individual source signals and (b) the interference mask $\gamma_{\mathrm{int}}(\ell)$.

Figure 4.10 visualizes the effect of the denominator on the beamformer output for the target speaker's and the interfering speaker's signal for all time frames within a block with a size of 25 time frames. Here, an atypically small block size is chosen to emphasize the effect of the denominator in (4.59). Figure 4.10(a) depicts the power of the source signals of the single sources and Fig. 4.10(b) the values of the interference mask $\gamma_{\text{int}}(\ell)$ used for SCM estimation. The behavior of the numerator $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\text{H}} \cdot \mathbf{x}_0(\ell)|^2]$, denominator $\mathbb{E}[|\delta(\mathbf{x}_0(\ell))|^2]$ and resulting power $p(\mathbf{x}_0(\ell))$ in (4.59) are depicted for the target speaker's signal at the beamformer output in Fig. 4.10(c), Fig. 4.10(e) and Fig. 4.10(g), respectively. Fig. 4.10(d), Fig. 4.10(f) and Fig. 4.10(h) show the corresponding quantities for the interfering speaker's signal at the beamformer output.

For both, the target speaker's signal and the interfering speaker's signal, the numerator in (4.59) is proportional to the frame-wise power of the source signals, i.e., the larger the power of the source signal the larger the value of the numerator. Considering the interfering speaker's signal at the beamformer output, it can be seen that the denominator $\mathbb{E}[|\delta(\mathbf{x}_1(\ell))|^2]$ shows a similar behavior so that the time frames of the interfering speaker's signal, whose power $\sigma_1^2(\ell)$ is very large, become additionally suppressed. This results in the power $p(\mathbf{x}_1(\ell))$ of the interfering speaker's signal at the beamformer output being similarly large for all time frames. A similar, but much less pronounced effect, can be seen for the target speaker. For the target speaker's signal the denominator $\mathbb{E}[|\delta(\mathbf{x}_0(\ell))|^2]$ is much smaller than the numerator $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\text{H}} \cdot \mathbf{x}_0(\ell)|^2]$ in (4.59) which might be explained by the low value of the interference mask $\gamma_{\text{int}}(\ell)$ for time frequency bins in which the target speaker's signal has large power. Further, a large part of the target speaker's signal is represented by the steering vector and, therefore, not affected by the denominator due to the distortionless response constraint.

4.4.2 SCM estimation from oracle-separated signals

For sake of a better interpretability, beamforming under optimal conditions will be considered next. Therefore, it is assumed that the interference only consists of the interfering speaker's signal and no noise. Moreover, it is assumed that oracle-separated signals are available so that the interference SCM can be estimated without contributions by the target speaker's signals and $\gamma_{\text{int}}(\ell)$=1 follows for the interference mask. From the latter assumption and (4.24), it follows that the unnormalized scale matrix $\mathbf{\Sigma}_{\text{int},\setminus\ell}$ of the Wishart approximation of the interference-SCM estimate in (4.64) is given by

$$\mathbf{\Sigma}_{\text{int},\setminus\ell} = \sum_{\substack{\widetilde{\ell}=0 \\ \widetilde{\ell}\neq\ell}}^{L-1} \sigma_1^2\left(\widetilde{\ell}\right) \cdot \mathbf{R}_1. \tag{4.98}$$

Further, it is assumed that the ground-truth target SCM has rank one with $\mathbf{R}_0$=$\mathbf{d}\cdot\mathbf{d}^{\text{H}}$ and the steering vector estimate $\widehat{\mathbf{d}}$ corresponds to the eigenvector $\mathbf{d}$ of the ground-truth target SCM $\mathbf{R}_0$.

Under the conditions specified above, the target speaker's signals are not involved in the estimation of the interference SCM such that the beamformer coefficients and the target speaker's

signals are statistically independent, i.e., $\mathbb{E}\big[|\delta(\mathbf{x}_0)|^2\big]=1$ holds for the denominator in (4.59). Hence, using (4.59) and (4.65) with $\mathbf{R}_{\mathbf{x}}=\sigma_0^2(\ell)\cdot\mathbf{R}_0=\sigma_0^2(\ell)\cdot\mathbf{d}\cdot\mathbf{d}^{\mathrm{H}}$ and $\widehat{\mathbf{d}}=\mathbf{d}$, the power of the target speaker's signal at the beamformer output can be expressed as

$$p_{\mathrm{tar}}(\ell) = p(\mathbf{x}_0(\ell)) = |d_0|^2 \cdot \sigma_0^2(\ell). \tag{4.99}$$

The independence from the sample size used for SCM estimation can be attributed to the perfect steering and the distortionless response constraint of the MVDR beamformer.

Next, the power of the interfering speaker's signal at the beamformer output is considered. For the optimal conditions described above, i.e., for $\mathbf{R}_0=\mathbf{d}\cdot\mathbf{d}^{\mathrm{H}}$, $\widehat{\mathbf{d}}=\mathbf{d}$ and $\mathbf{\Sigma}_{\mathrm{int},\setminus\ell}$ being defined as in (4.98) and $\mathbf{R}_{\mathrm{x}}=\sigma_1^2(\ell)\cdot\mathbf{R}_1$, the numerator of (4.59), being specified in (4.65), is given by

$$\mathbb{E}\Big[\big|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}\cdot\mathbf{x}_1(\ell)\big)\big|^2\Big] = \frac{L_{\setminus\ell}}{L_{\setminus\ell}-M+1}\cdot\frac{|d_0|^2\cdot\sigma_1^2(\ell)}{\mathbf{d}^{\mathrm{H}}\cdot\mathbf{R}_1^{-1}\cdot\mathbf{d}} \tag{4.100}$$

and the denominator of (4.59), being specified in (4.75), is given by

$$\begin{aligned}&\mathbb{E}\big[|\delta(\mathbf{x}_1(\ell))|^2\big]\\ =&1+\frac{2\cdot L_{\setminus\ell}\cdot(M-1)}{\big(L_{\setminus\ell}-M+1\big)}\cdot\frac{\sigma_1^2(\ell)}{\sum\limits_{\substack{\widetilde{\ell}=0\\ \widetilde{\ell}\neq\ell}}^{L-1}\sigma_1^2\left(\widetilde{\ell}\right)}+\frac{L_{\setminus\ell}^3\cdot(M^2-1)-L_{\setminus\ell}^2\cdot(M^3+1)}{\big(L_{\setminus\ell}-M+1\big)\cdot\big(L_{\setminus\ell}-M\big)^2}\cdot\frac{\sigma_1^4(\ell)}{\left(\sum\limits_{\substack{\widetilde{\ell}=0\\ \widetilde{\ell}\neq\ell}}^{L-1}\sigma_1^2\left(\widetilde{\ell}\right)\right)^2}.\end{aligned} \tag{4.101}$$

With (4.100) and (4.101) the power of the interfering speaker's signal at the beamformer output results in

$$\begin{aligned}p_{\mathrm{int}}(\ell) = p(\mathbf{x}_1(\ell)) &\approx \frac{\mathbb{E}\left[\left|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}\cdot\mathbf{x}_1(\ell)\big)\right|^2\right]}{\mathbb{E}\big[|\delta(\mathbf{x}_1(\ell))|^2\big]}\\ =&\frac{1}{\frac{L_{\setminus\ell}-M+1}{L_{\setminus\ell}}+2\cdot(M-1)\cdot\frac{\sigma_1^2(\ell)}{\sum\limits_{\substack{\widetilde{\ell}=0\\ \widetilde{\ell}\neq\ell}}^{L-1}\sigma_1^2\left(\widetilde{\ell}\right)}+\frac{L_{\setminus\ell}^2\cdot(M^2-1)-L_{\setminus\ell}\cdot(M^3+1)}{\big(L_{\setminus\ell}-M\big)^2}\cdot\frac{\sigma_1^4(\ell)}{\left(\sum\limits_{\substack{\widetilde{\ell}=0\\ \widetilde{\ell}\neq\ell}}^{L-1}\sigma_1^2\left(\widetilde{\ell}\right)\right)^2}}\cdot\frac{|d_0|^2\cdot\sigma_1^2(\ell)}{\mathbf{d}^{\mathrm{H}}\cdot\mathbf{R}_1^{-1}\cdot\mathbf{d}}.\end{aligned} \tag{4.102}$$

Note that the interfering speaker is the only source involved in estimating the interference SCM so that the interfering speaker's signal is the dominant source for all time frequency bins. Consequently, the denominator given in (4.101) is applied to all time frequency bins under the optimal conditions.

It can be observed from (4.102) that the power of the interfering speaker's signal at the beamformer output $p_{\mathrm{int}}(\ell)$ decreases when the equivalent degrees of freedom $L_{\setminus\ell}$ of the approximate Wishart distribution of interference-SCM estimate increase and when the ratio

$$\xi_{\sigma_1^2}(\ell) = \frac{\sigma_1^2(\ell)}{\sum\limits_{\substack{\widetilde{\ell}=0\\ \widetilde{\ell}\neq\ell}}^{L-1}\sigma_1^2\left(\widetilde{\ell}\right)} \tag{4.103}$$

decreases. Here, $\xi_{\sigma_1^2}(\ell)$ corresponds to the ratio between the power of the interfering speaker's signal in the ℓ-th time frame and its power over all other time frames within the considered block. The power of the interfering speaker's signal at the beamformer can become smaller than the steady-state power towards which it converges for very large block sizes, i.e.,

$$p_{\text{int}}(\ell) < \frac{|d_0|^2 \cdot \sigma_1^2(\ell)}{\widehat{\mathbf{d}}^{\text{H}} \cdot \mathbf{R}_1^{-1} \cdot \widehat{\mathbf{d}}} \tag{4.104}$$

if the ℓ-th time frame of the interfering speaker's signal dominates the interference-SCM estimate, i.e., $\xi_{\sigma_1^2}(\ell)$ is not close to zero, and the equivalent degrees of freedom of the interference-SCM estimate's approximate Wishart distribution $L_{\setminus\ell}$ are not much larger than the number of microphones M. Consequently, the power of the interference at the beamformer output increases, i.e., suppression of the interference diminishes, with increasing block size. This behavior results from a decreasing value of the power ratio $\xi_{\sigma_1^2}(\ell)$ and an increasing value of the equivalent degrees of freedom $L_{\setminus\ell}$. Note that the interference becomes more suppressed for small block sizes if more microphones are employed, i.e., if M is larger, which can be seen from (4.102).

In the following, the idealized conditions with interference-SCM estimation from oracle-separated signals is considered. However, the more general case with noise as additional interfering source, a full-rank target SCM and estimated steering vectors is considered.

Figure 4.11 visualizes the effects of SCM estimation from a finite sample size on the numerator of (4.59) that were theoretically discussed before. To this end, one time frequency bin of the interfering speaker's signal at the beamformer output is investigated. It becomes obvious that the numerator of the interfering speaker's power at the beamformer output $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\text{H}} \cdot \mathbf{x}_1(\ell)|^2]$ decreases with growing block size L (see Fig. 4.11(d)). The transient component $p_{\text{t}}(\mathbf{x}_1(\ell))$, which is shown in Fig. 4.11(a), and the steady-state component $p_\infty(\mathbf{x}_1(\ell))$, which is shown in Fig. 4.11(b), are quite independent from the block size and only change marginally due to slight changes of the portion of the contributions of interfering speaker's signal and noise to the interference-SCM estimate. In addition to that, it can be seen that the transient component $p_{\text{t}}(\mathbf{x}_1(\ell))$ is significantly larger than the steady-state component $p_\infty(\mathbf{x}_1(\ell))$, as discussed before. Thus, the decreasing behavior of the numerator $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\text{H}} \cdot \mathbf{x}_1(\ell)|^2]$ in (4.59) results from the convergence of the interference-SCM estimates towards their expected value. This is reflected by the growing degrees of freedom $L_{\setminus\ell}$ of the Wishart approximation of the interference-SCM estimates which is shown in Fig. 4.11(c). The growing degrees of freedom $L_{\setminus\ell}$ result in a decreasing weight of the transient component $p_{\text{t}}(\mathbf{x}_1(\ell))$ and a convergence of the weight of the steady-state component $p_\infty(\mathbf{x}_1(\ell))$ towards a value of one.

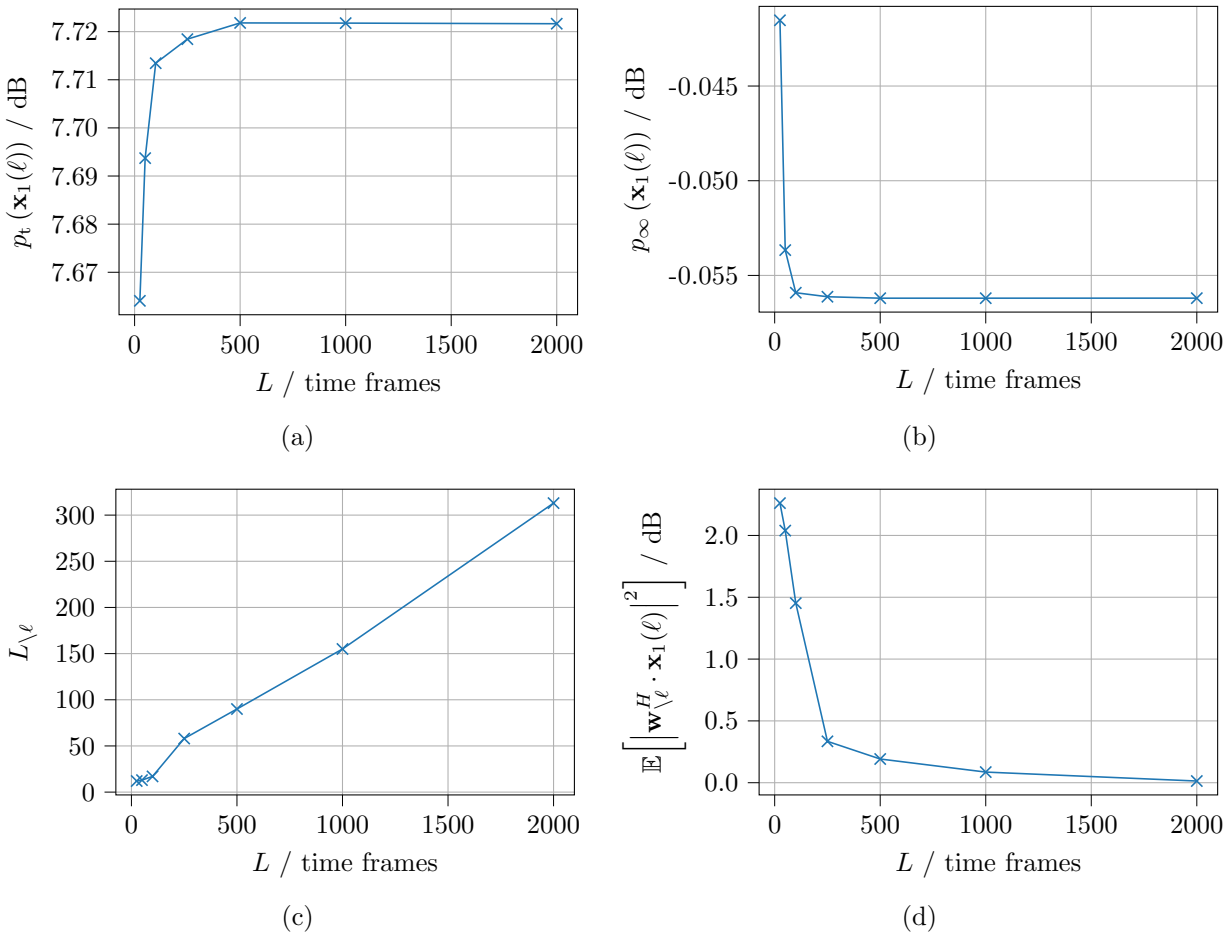


Figure 4.11: Visualization of the effects of using a finite sample size for SCM estimation on the numerator of the power of the interfering speaker's signal at the beamformer output in (4.59) as a function of the block size L. The same time frequency bin is considered but the block size is varied. The ideal case with interference SCM estimation without presence of the target speaker's signal is considered. (a) shows the transient component $p_t(\mathbf{x}_1(\ell))$, (b) the steady-state component $p_\infty(\mathbf{x}_1(\ell))$ and (d) the resulting value of the numerator $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^H \cdot \mathbf{x}_1(\ell)|^2]$ in (4.59). (c) shows the equivalent degrees of freedom $L_{\setminus\ell}$ of the Wishart approximation of the interference-SCM estimate.

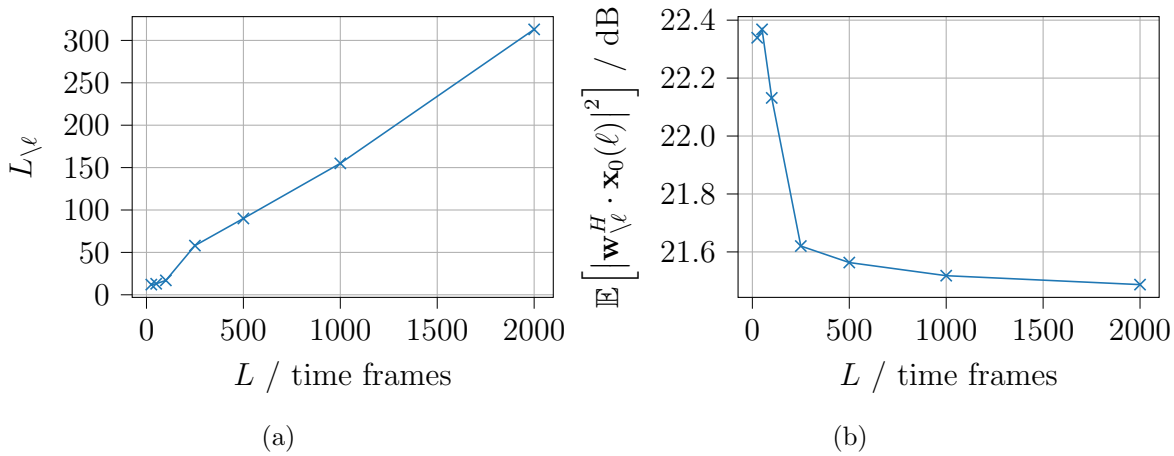


Figure 4.12: Visualization of the effects of using a finite sample size for SCM estimation on the numerator of the power of the target speaker's signal at the beamformer output in (4.59) as a function of the block size L. The same time frequency bin is considered but the block size is varied. The ideal case with interference SCM estimation without presence of the target speaker's signal is considered. (a) shows the equivalent degrees of freedom $L_{\setminus\ell}$ of the Wishart approximation of the interference-SCM estimates and (b) the value of the numerator $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}\cdot\mathbf{x}_0(\ell)|^2]$ in (4.59).

The effects of SCM estimation from a finite sample size on the numerator of (4.59) are illustrated for the target speaker's signal at the beamformer output in Fig. 4.12. Similar to the behavior of the numerator of (4.59) for the interfering speaker's signal, which was discussed w.r.t. Fig. 4.11, the value of numerator $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}\cdot\mathbf{x}_0(\ell)|^2]$ in (4.59) also decreases for the target speaker's signal at the beamformer output as the block size L increases. However, this effect is much smaller than for the interfering speaker's signal. This can be attributed to the distortionless response constraint of the MVDR beamformer and the mismatch between the interference-SCM estimate and the target speaker's SCM.

In Fig. 4.13 the numerator $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}\cdot\mathbf{x}_1(\ell)|^2]$ and the denominator $\mathbb{E}[|\delta(\mathbf{x}_1(\ell))|^2]$ in (4.59) as well as the resulting behavior of the power $p(\mathbf{x}_1(\ell))$ of a time frequency bin of the interfering speaker's signal at the beamformer output are shown as a function of the block size L. Note that the same time frequency bin as in Fig. 4.11 is investigated in the same way as in Fig. 4.11. The numerator and the denominator in (4.59) show a decreasing behavior with growing block size. However, the denominator $\mathbb{E}[|\delta(\mathbf{x}_1(\ell))|^2]$ can be much larger than the numerator $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}\cdot\mathbf{x}_1(\ell)|^2]$ if the considered time frequency bin dominates the interference-SCM estimate. Thus, the focus on special characteristics of the interference for small block sizes dominates the improvement of the general statistics captured by the interference-SCM estimate with a growing block size. This leads to a decreasing suppression of the interference with increasing block size as shown in Fig. 4.13(e). Moreover, it can be seen that the denominator $\mathbb{E}[|\delta(\mathbf{x}_1(\ell))|^2]$ in (4.59), whose behavior with increasing block size in visualized in Fig. 4.13(b), behaves very similar to the ratio $\xi_{\sigma_1^2}(\ell)$ between the power of the interfering speaker's signal in the ℓ-th time frame and the sum of its power over all other time frames within the considered block, which is depicted in Fig. 4.13(b). This reflects the fact that the additional suppression of the interference due to the denominator in (4.59) reflecting the statistical dependence between the beamformer coefficients and the signals to

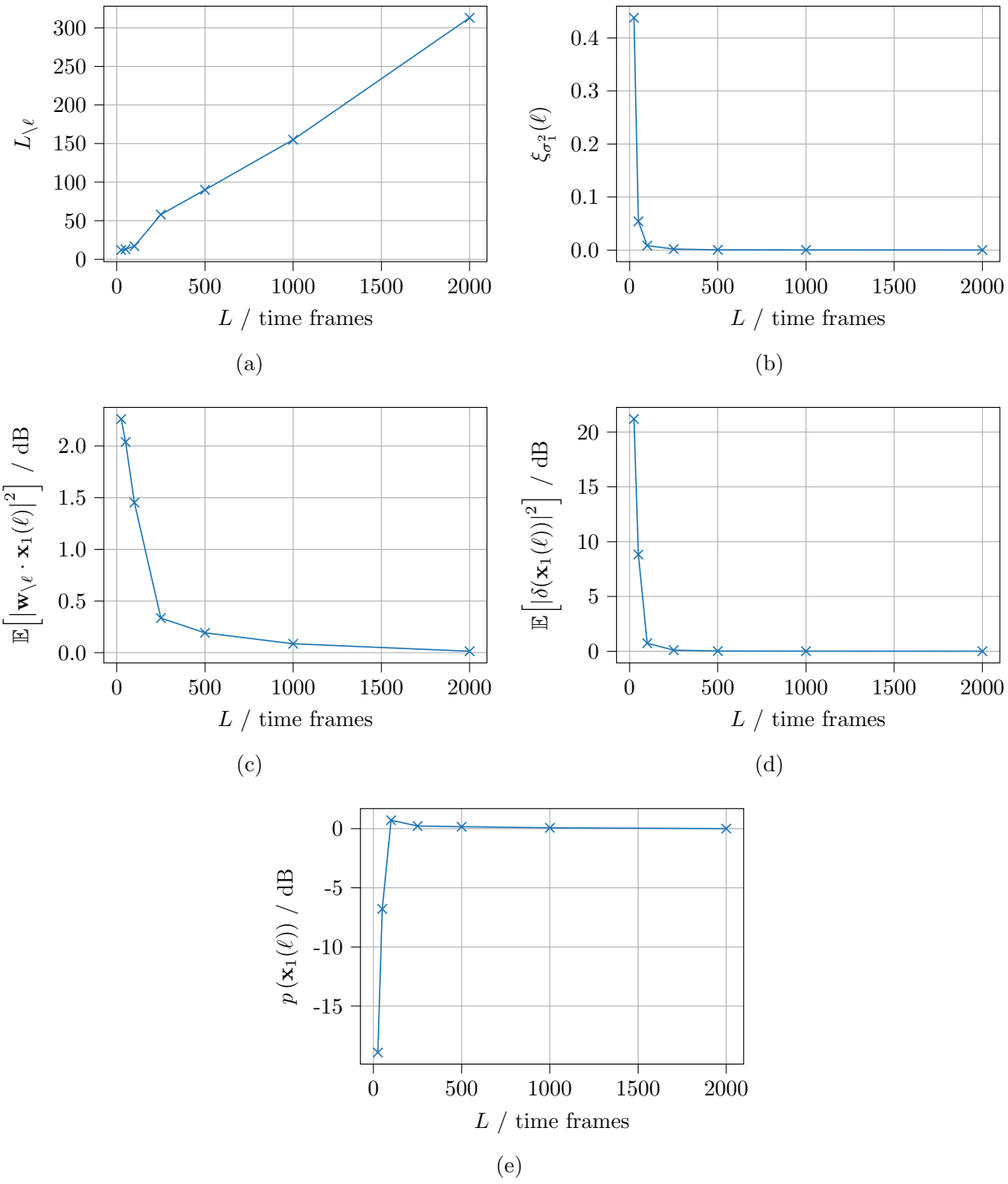


Figure 4.13: Visualization of the effects of using a finite sample size for SCM estimation on the power of the interfering speaker's contribution to the beamformer output as a function of the block size L. Thereby, the same time frequency bin is considered but the block size is varied. The ideal case with interference SCM estimation without presence of the target speaker's signal is considered. (a) shows the equivalent degrees of freedom $L_{\setminus\ell}$ of the Wishart approximation of the interference-SCM estimates, (b) the ratio between the power of the interfering speaker's signal in the ℓ-th time frame and its power over all other time frames within the considered block $\xi_{\sigma_1^2}(\ell)$, (c) the numerator $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}_1(\ell)|^2]$ in (4.59), (d) the denominator $\mathbb{E}[|\delta(\mathbf{x}_1(\ell))|^2]$ in (4.59) and (e) the resulting power $p(\mathbf{x}_1(\ell))$.

which they are applied diminishes if the contribution of a time frequency bin becomes less dominant.

4.4.3 Influence of the leakage of the target signal into the interference-SCM estimate

In addition to the convergence of the SCM estimates towards their expected value with growing block size, the portion of the unwanted contribution of the target speaker's signal to the interference-SCM estimates will also affect the performance of the beamformer. It is expected that the relative contribution of the target speaker's signal to the interference-SCM estimate decreases as the block size increases. This would result in an additional improvement of the quality of the interference-SCM estimates with increasing block size.

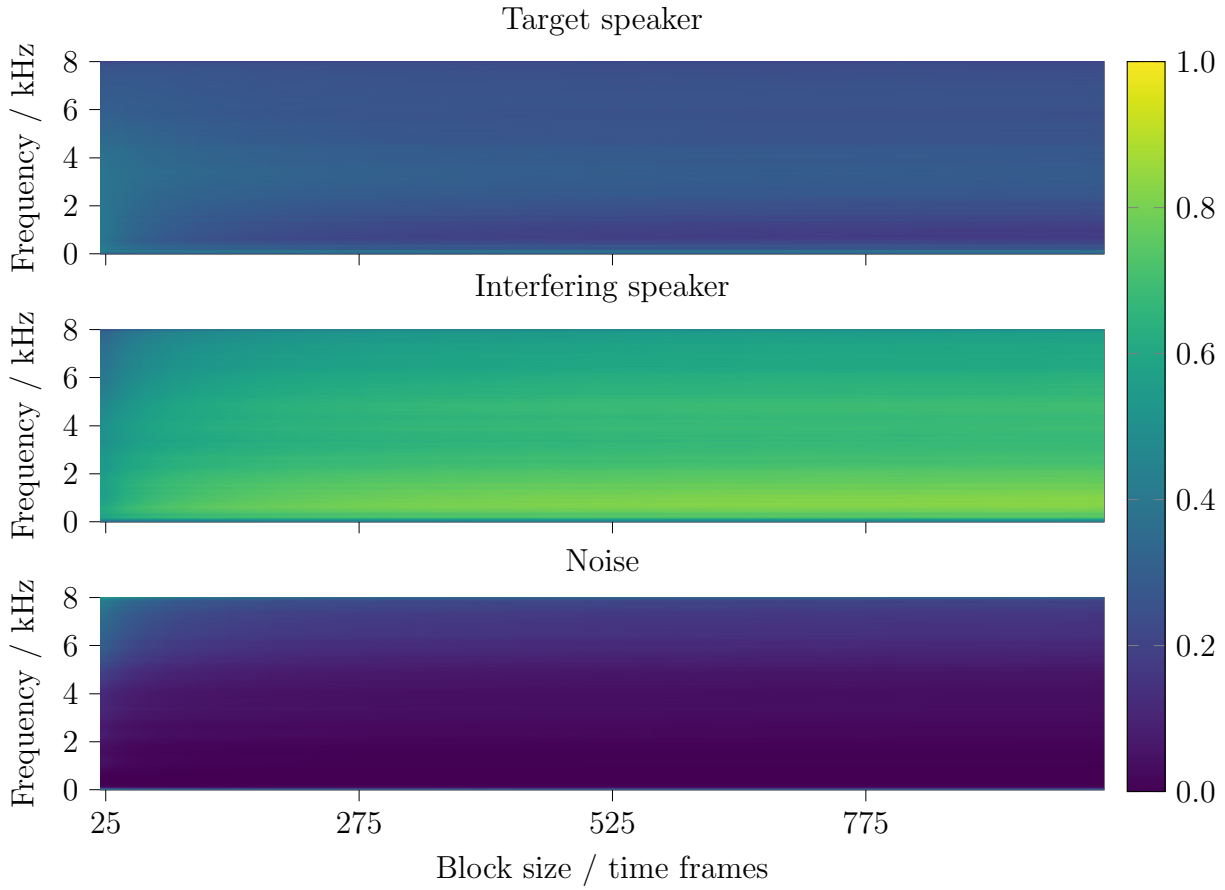


Figure 4.14: Average proportion of the single sources' contribution to the estimated interference SCMs $\Upsilon_i(L, b, k)$, with $i \in \{0, 1, \nu\}$.

The dependence of the average proportion of the single sources to the mask-based interference-SCM estimates on the block size is shown in Fig. 4.14. To this end, the interference estimates in (2.49) are decomposed into partial sums, to which only one source contributes. Note that additional cross terms, e.g., $\mathbf{x}_0(\ell, k) \cdot \mathbf{x}_1^{\mathrm{H}}(\ell, k)$, also contribute to the estimated interference SCMs. These terms are ignored in the investigation since their contribution should be

negligibly small if the block size is large enough assuming statistical independence of the source signals. For small block sizes, however, these cross terms might have a non-negligible contribution to the estimated interference SCMs which leads to an additional deterioration of the beamforming performance. Finally, the ratio of the trace of the source-wise submatrices to the sum of the traces of all source-wise submatrices is utilized as measure for the contribution of the single sources:

$$\Upsilon_i(L,b,k) = \frac{\mathrm{tr}\left\{\sum\limits_{\ell=b\cdot L}^{(b+1)\cdot L-1} \mathbf{x}_i(\ell,k)\cdot\mathbf{x}_i^{\mathrm{H}}(\ell,k)\right\}}{\mathrm{tr}\left\{\sum\limits_{\ell=b\cdot L}^{(b+1)\cdot L-1} (\mathbf{x}_0(\ell,k)\cdot\mathbf{x}_0^{\mathrm{H}}(\ell,k)+\mathbf{x}_1(\ell,k)\cdot\mathbf{x}_1^{\mathrm{H}}(\ell,k)+\mathbf{x}_\nu(\ell,k)\cdot\mathbf{x}_\nu^{\mathrm{H}}(\ell,k))\right\}}, \quad (4.105)$$

with $i \in \{0,1,\nu\}$.

It is noticeable that the target source has a non-negligible contribution to the estimated interference SCMs although the interference masks generally are close to zero for the time frequency bins for which the target source has a large power. Nevertheless, the contribution of the interfering speaker dominates the estimated interference SCMs, especially for those frequencies for which speech has the largest power. With growing block size the contribution of the target source decreases, but still cannot be completely neglected. Again, this particularly holds for the frequency range between 0.125 kHz and 2 kHz for which speech has significant power. It is expected that the leakage of the target speaker's signal into the interference-SCM estimate negatively influences the beamforming performance and that the strength of this effect is larger for smaller block sizes. Moreover, the contribution of the noise is mostly negligible compared to the overall contribution of both speech signals.

In the following, the effect of the leakage of the target speaker's signal into the interference-SCM estimates on the suppression of the interfering speaker's signal is considered. For sake of simplicity the noiseless case is considered. Note that the difference to the general case will typically be small since the influence of the noise signals on the interference-SCM estimates generally is comparably small, as visualized in Fig. 4.14.

First, the numerator of (4.59) is considered for the interfering speaker's signal, i.e., (4.88) is considered with $\mathbf{R}_\mathbf{x}$=$\mathbf{R}_1$. Without loss of generality, the scaling of the SCM by the time-varying power $\sigma_1^2(\ell)$ is omitted for sake of a simpler notation. By assuming that the unnormalized scale matrix of the approximate Wishart distribution of the interference-SCM estimates is given by $\mathbf{\Sigma}_{\mathrm{int},\setminus\ell} = \mathbf{R}_1 + \beta\cdot\mathbf{R}_0$ with real-valued $\beta \geq 0$, the influence of the degree of target signal leakage into the interference-SCM estimates on the transient component $p_\mathrm{t}(\mathbf{x}_1(\ell))$ in (4.88) is investigated, next. To this end, the derivative of the transient component $p_\mathrm{t}(\mathbf{x}_1(\ell))$ w.r.t. the scaling factor β of the contribution of the target speaker's signals to the interference-SCM estimates is considered, which is given by

$$\frac{\mathrm{d}}{\mathrm{d}\beta}p_{\mathrm{t}}(\mathbf{x}_1(\ell)) = \frac{\mathrm{d}}{\mathrm{d}\beta}\frac{\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_1\Big\}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}} = \frac{\mathrm{d}}{\mathrm{d}\beta}\frac{\mathrm{tr}\big\{(\mathbf{R}_1+\beta\cdot\mathbf{R}_0)^{-1}\cdot\mathbf{R}_1\big\}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot(\mathbf{R}_1+\beta\cdot\mathbf{R}_0)^{-1}\cdot\widehat{\mathbf{d}}}$$
$$=\frac{1}{\Big(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\Big)^2}\cdot\Big(-\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_0\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_1\Big\}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}$$
$$+\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_1\Big\}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_0\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\Big)\,. \tag{4.106}$$

It is expected that the term at hand grows with increasing value of β, i.e., that the suppression of the interfering speaker's signal degrades with a growing portion of contribution of the target signal to the inference-SCM estimate. In order to show this, it has to be proven that the derivative given in (4.106) is positive, i.e., it has to be investigated whether the following relation holds [1]:

$$\frac{\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_0\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_1\Big\}}{\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_1\Big\}} \overset{?}{\leq} \frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_0\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}}. \tag{4.107}$$

Here, it is utilized that the denominator $(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}})^2$ in (4.106) always is positive. First, the left hand side of (4.107) is considered, which can be rewritten as

$$\frac{\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_0\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_1\Big\}}{\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_1\Big\}} = \frac{\mathrm{tr}\Big\{\overbrace{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-\frac{1}{2}}\cdot\mathbf{R}_0\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-\frac{1}{2}}}^{:=\widetilde{\mathbf{A}}}\cdot\overbrace{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-\frac{1}{2}}\cdot\mathbf{R}_1\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-\frac{1}{2}}}^{:=\widetilde{\mathbf{B}}}\Big\}}{\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-\frac{1}{2}}\cdot\mathbf{R}_1\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-\frac{1}{2}}\Big\}}. \tag{4.108}$$

With the eigenvalue decomposition (EVD) of the auxiliary matrix $\widetilde{\mathbf{B}}=\mathbf{O}\cdot\boldsymbol{\Lambda}\cdot\mathbf{O}^{\mathrm{H}}$, it follows that

$$\frac{\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_0\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_1\Big\}}{\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_1\Big\}} = \frac{\mathrm{tr}\Big\{\widetilde{\mathbf{A}}\cdot\mathbf{O}\cdot\boldsymbol{\Lambda}\cdot\mathbf{O}^{\mathrm{H}}\Big\}}{\mathrm{tr}\Big\{\widetilde{\mathbf{B}}\Big\}} = \frac{\sum\limits_{i=0}^{M-1}\lambda_i\cdot\Big(\mathbf{O}^{\mathrm{H}}\cdot\widetilde{\mathbf{A}}\cdot\mathbf{O}\Big)_{ii}}{\sum\limits_{i=0}^{M-1}\lambda_i}. \tag{4.109}$$

Here, the matrix $\mathbf{O}$ corresponds to the matrix of eigenvectors of the auxiliary matrix $\widetilde{\mathbf{B}}$ and $\boldsymbol{\Lambda}$ to a diagonal matrix containing the eigenvalues λ_i of the matrix $\widetilde{\mathbf{B}}$. Since the unitary transformation $\mathbf{O}^{\mathrm{H}}\cdot\widetilde{\mathbf{A}}\cdot\mathbf{O}$ of the auxiliary matrix $\widetilde{\mathbf{A}}$ preserves the eigenvalues of $\widetilde{\mathbf{A}}$, the left hand side of (4.107) is upper bounded by the largest eigenvalue of $\widetilde{\mathbf{A}}$. This can be explained by he fact that, usually, the diagonal elements of the positive semidefinite matrix $\widetilde{\mathbf{A}}$ are smaller than the largest eigenvalue of $\widetilde{\mathbf{A}}$ [63] so that the weighted average of the diagonal

[1] A large language model (LLM) was used in the initial research to investigate how this inequality could be proven. This gave rise to the idea of proving the inequality based on the eigenvalues of $\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-\frac{1}{2}}\cdot\mathbf{R}_0\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-\frac{1}{2}}$.

elements of the matrix $\widetilde{\mathbf{A}}$ in (4.109) typically is smaller than the largest eigenvalue of the matrix $\widetilde{\mathbf{A}} = \boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-\frac{1}{2}} \cdot \mathbf{R}_0 \cdot \boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-\frac{1}{2}}$.

Next, the right hand side of (4.107) is investigated. By rewriting the right hand side, it follows that

$$\frac{\widehat{\mathbf{d}}^{\text{H}} \cdot \boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-1} \cdot \mathbf{R}_0 \cdot \boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\text{H}} \cdot \boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}} = \frac{\widetilde{\mathbf{d}}^{\text{H}} \cdot \boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-\frac{1}{2}} \cdot \mathbf{R}_0 \cdot \boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-\frac{1}{2}} \cdot \widetilde{\mathbf{d}}}{\widetilde{\mathbf{d}}^{\text{H}} \cdot \widetilde{\mathbf{d}}}, \tag{4.110}$$

with $\widetilde{\mathbf{d}} = \boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-\frac{1}{2}} \cdot \widehat{\mathbf{d}}$. Thus, the right hand side of (4.107) corresponds to the Rayleigh quotient of the matrix $\boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-\frac{1}{2}} \cdot \mathbf{R}_0 \cdot \boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-\frac{1}{2}}$ at vector $\widetilde{\mathbf{d}}$ and, therefore, is lower bounded by the smallest eigenvalue of the matrix $\boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-\frac{1}{2}} \cdot \mathbf{R}_0 \cdot \boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-\frac{1}{2}}$ and upper-bounded by the largest eigenvalue of the matrix $\boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-\frac{1}{2}} \cdot \mathbf{R}_0 \cdot \boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-\frac{1}{2}}$. Assuming that the steering vector estimate $\widehat{\mathbf{d}}$ corresponds to the GEVD-based steering vector estimate from (2.45), which is calculated based on the SCMs $\mathbf{R}_0$ and $\boldsymbol{\Sigma}_{\text{int},\ell}$,

$$\widetilde{\mathbf{d}} = \boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-\frac{1}{2}} \cdot \widehat{\mathbf{d}} = \boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-\frac{1}{2}} \cdot \boldsymbol{\Sigma}_{\text{int}} \cdot \bar{\mathbf{d}} = \boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{\frac{1}{2}} \cdot \bar{\mathbf{d}} \tag{4.111}$$

follows, with $\bar{\mathbf{d}} = \mathcal{P}\left(\boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-1} \cdot \mathbf{R}_0\right)$ denoting the dominant generalized eigenvector of the matrix pair $(\mathbf{R}_0, \boldsymbol{\Sigma}_{\text{int},\setminus\ell})$. Inserting (4.111) into (4.110), leads to

$$\frac{\widehat{\mathbf{d}}^{\text{H}} \cdot \boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-1} \cdot \mathbf{R}_0 \cdot \boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\text{H}} \cdot \boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}} = \frac{\bar{\mathbf{d}}^{\text{H}} \cdot \mathbf{R}_0 \cdot \bar{\mathbf{d}}}{\bar{\mathbf{d}}^{\text{H}} \cdot \boldsymbol{\Sigma}_{\text{int},\setminus\ell} \cdot \bar{\mathbf{d}}}. \tag{4.112}$$

Since $\bar{\mathbf{d}}$ corresponds to the dominant generalized eigenvector of the matrix pair $(\mathbf{R}_0, \boldsymbol{\Sigma}_{\text{int},\setminus\ell})$ in this case, the generalized Rayleigh quotient [60] in (4.112) becomes the corresponding largest generalized eigenvalue of the matrix pair $(\mathbf{R}_0, \boldsymbol{\Sigma}_{\text{int},\setminus\ell})$. It can be shown that the generalized eigenvalues of the matrix pair $(\mathbf{R}_0, \boldsymbol{\Sigma}_{\text{int},\setminus\ell})$ corresponds to the eigenvalues of the matrix $\boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-\frac{1}{2}} \cdot \mathbf{R}_0 \cdot \boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-\frac{1}{2}}$ such that (4.110) corresponds to the largest eigenvalue of the matrix $\boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-\frac{1}{2}} \cdot \mathbf{R}_0 \cdot \boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-\frac{1}{2}}$. If the steering vector is only close to the GEVD-based steering vector estimate which is calculated based on the SCMs $\mathbf{R}_0$ and $\boldsymbol{\Sigma}_{\text{int}}(\ell)$ as in (4.111), the Rayleigh quotient in (4.110) still is close to the largest eigenvalue of the matrix $\boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-\frac{1}{2}} \cdot \mathbf{R}_0 \cdot \boldsymbol{\Sigma}_{\text{int},\setminus\ell}^{-\frac{1}{2}}$ due to the stationarity of the Rayleigh quotient at an eigenvector [62].

By comparing the left hand side and the right hand side of (4.107), it can be seen that the derivative of the transient component $p_{\text{t}}(\mathbf{x}_1(\ell))$ in (4.65) w.r.t. β typically is positive, where β is the scaling factor of the contribution of the target speaker's signals to the interference-SCM estimate. This follows from the fact that the weighted average of eigenvalues on the left hand side of (4.107) usually is smaller than the value of the right hand side of (4.107) which usually is close the corresponding largest eigenvalue. Thus, the transient component $p_{\text{t}}(\mathbf{x}_1(\ell))$ in (4.65) increases as the contribution of the target speaker's signals to the interference-SCM estimate increases, i.e, the suppression of the interfering speaker's signal degrades. Note that

there can be rare cases where the transient component $p_{\mathrm{t}}(\mathbf{x}_1(\ell))$ in (4.65) is not monotonically increasing with β, e.g., if the accuracy of the steering vector is bad.

Next, the steady-state component $p_\infty(\mathbf{x}_1(\ell))$ in (4.88) is investigated in a similar way. For this purpose, $\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell} = \beta \cdot \mathbf{R}_1 + \mathbf{R}_0$ with real-valued $\beta \geq 0$ is considered for the unnormalized scale matrix of the approximate Wishart distribution of the interference-SCM estimate. From this it follows that the portion of the contribution of the target speaker's signal to the interference-SCM estimates decreases with growing β. The derivative of the steady-state component $p_\infty(\mathbf{x}_1(\ell))$ in (4.88) w.r.t. the scaling factor β of the contribution of the interfering speaker's signals to the interference-SCM estimates is given by

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\beta} p_\infty(\mathbf{x}_1(\ell)) = \frac{\mathrm{d}}{\mathrm{d}\beta} \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{R}_{\mathbf{x}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \widehat{\mathbf{d}}}{\left(\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \widehat{\mathbf{d}}\right)^2} \\
=&\frac{\mathrm{d}}{\mathrm{d}\beta} \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot (\beta \cdot \mathbf{R}_1 + \mathbf{R}_0)^{-1} \cdot \mathbf{R}_{\mathbf{x}} \cdot (\beta \cdot \mathbf{R}_1 + \mathbf{R}_0)^{-1} \cdot \widehat{\mathbf{d}}}{\left(\widehat{\mathbf{d}}^{\mathrm{H}} \cdot (\beta \cdot \mathbf{R}_1 + \mathbf{R}_0)^{-1} \cdot \widehat{\mathbf{d}}\right)^2} \\
=&\frac{2}{\left(\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \widehat{\mathbf{d}}\right)^3} \cdot \Bigg(-\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{R}_1 \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{R}_1 \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \widehat{\mathbf{d}} \cdot \widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \widehat{\mathbf{d}} \\
&+ \left(\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{R}_1 \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \widehat{\mathbf{d}}\right)^2 \Bigg).
\end{aligned} \tag{4.113}$$

It is expected that that the suppression of the interfering speaker's signal improves as the portion of the contribution of the target speaker's signals to the interference-SCM estimates decreases, i.e., the term at hand decreases with increasing value of β. In order to show this, it has to be considered if

$$\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{R}_1 \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{R}_1 \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \widehat{\mathbf{d}} \cdot \widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \widehat{\mathbf{d}} \overset{?}{\geq} \left(\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{R}_1 \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \widehat{\mathbf{d}}\right)^2 \tag{4.114}$$

holds to show that the considered derivative is negative, since $\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \widehat{\mathbf{d}}$ is larger than zero. By defining the auxiliary vectors

$$\widetilde{\mathbf{a}} = \boldsymbol{\Sigma}^{-\frac{1}{2}} \cdot \mathbf{R}_1 \cdot \boldsymbol{\Sigma}^{-1}\mathbf{d} \text{ and } \widetilde{\mathbf{b}} = \boldsymbol{\Sigma}^{-\frac{1}{2}} \cdot \mathbf{d}, \tag{4.115}$$

it can be seen that (4.114) corresponds to a Cauchy-Schwarz inequality (see (4.90)) and, therefore, is fulfilled. Overall, both terms in (4.88) grow with growing portion of the contribution of the target speaker's signals to the interference-SCM estimates. In consequence, the suppression of interference becomes better when the sample size used for SCM estimation grows so that the contribution of the target signals to the interference-SCM estimates diminishes.

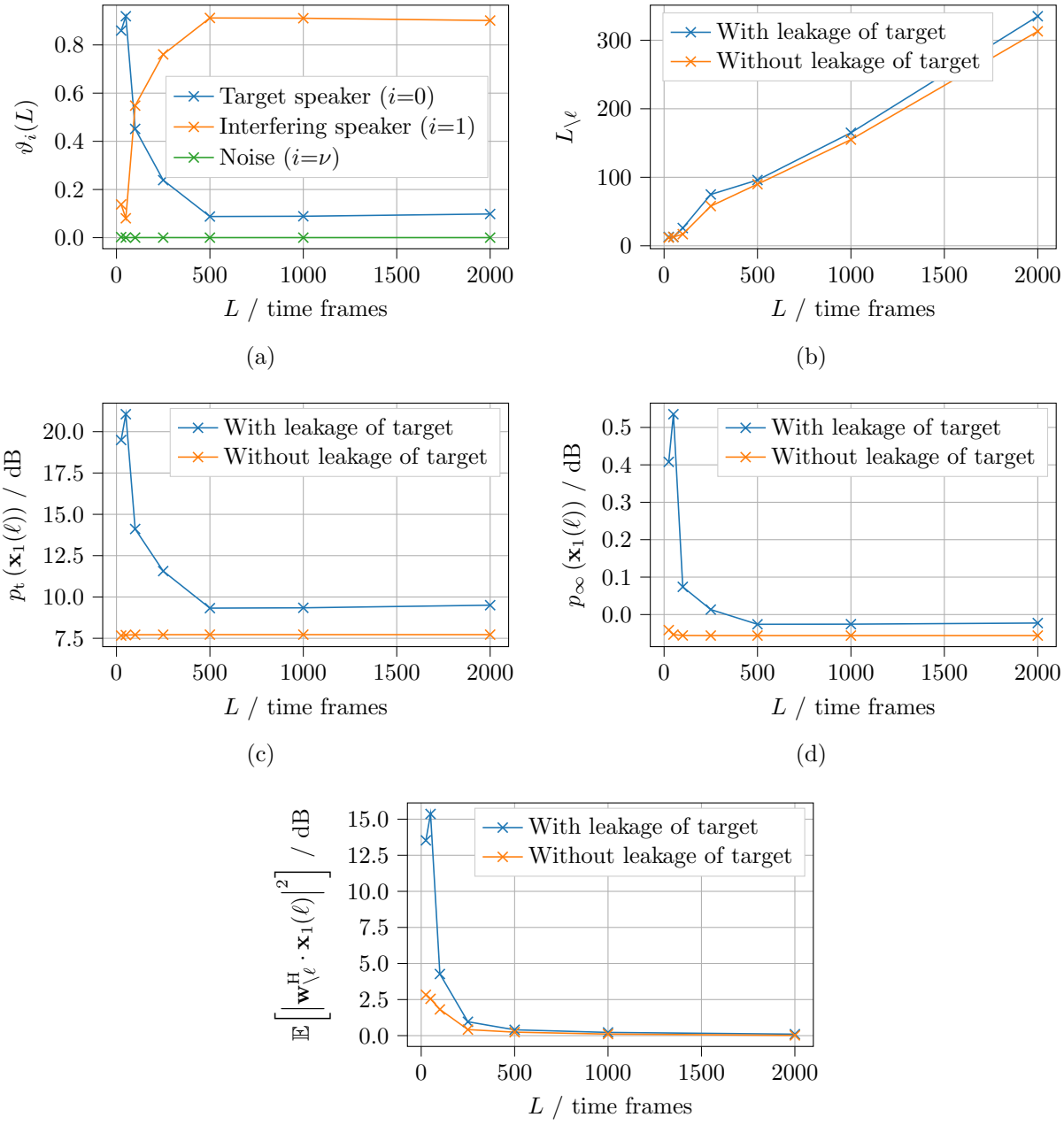


Figure 4.15: Visualization of the effects of the leakage of the target speaker's signal into the interference-SCM estimate on the numerator of the interfering speaker's power in (4.59) as a function of the block size L. The same time frequency bin is considered but the block size is varied. The case with leakage of the target signals into the interference-SCM estimates corresponds to beamforming using mask-based SCM estimates while the case without leakage corresponds to beamforming based on SCMs that are estimated from oracle-separated signals. (a) shows the portion of contribution $\vartheta_i(L)$ of the individual sources to the interference SCM estimate, (b) shows the equivalent degrees of freedom $L_{\setminus\ell}$ of the Wishart approximation of the interference-SCM estimates, (c) the transient component $p_t(\mathbf{x}_1(\ell))$, (d) the steady-state component $p_\infty(\mathbf{x}_1(\ell))$ and (e) the value of the numerator $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}_1(\ell)|^2]$ in (4.59).

Figure 4.15 visualizes the effects of the leakage of the target speaker's signal into the interference-SCM estimate on the numerator $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}_1(\ell)|^2]$ in (4.59). The same setup as in Fig. 4.11, where the case without leakage of the target speaker's signals into the interference-SCM estimate was discussed, is considered. In order to measure the portion of contribution of the individual sources to the interference-SCM estimate,

$$\vartheta_i(L) = \frac{\sum_{\ell=0}^{L-1} \gamma_{\mathrm{int}}(\ell) \cdot \sigma_i^2(\ell) \cdot \mathrm{tr}\{\mathbf{R}_i\}}{\sum_{\ell=0}^{L-1} \left(\gamma_{\mathrm{int}}(\ell) \cdot \sigma_0^2(\ell) \cdot \mathrm{tr}\{\mathbf{R}_0\} + \gamma_{\mathrm{int}}(\ell) \cdot \sigma_1^2(\ell) \cdot \mathrm{tr}\{\mathbf{R}_1\} + \gamma_{\mathrm{int}}(\ell) \cdot \sigma_\nu^2(\ell) \cdot \mathrm{tr}\{\mathbf{R}_\nu\}\right)} \tag{4.116}$$

is introduced, with $i \in \{0, 1, \nu\}$. This ratio reflects the contribution of a source to the expected value of the interference-SCM estimate, which is calculated using (4.11). Thus, $\vartheta_i(L)$ also reflects the contribution of a source to the equivalent scale matrix $\mathbf{\Sigma}_{\mathrm{int},\setminus\ell}$ of the approximation of the interference SCM estimate by a Wishart matrix, as derived in (4.24).

With leakage of the target speaker's signals into the interference-SCM estimate, the transient component $p_{\mathrm{t}}(\mathbf{x}_1(\ell))$ and the steady-state component $p_\infty(\mathbf{x}_1(\ell))$ of the power of the interfering speaker's signal at the beamformer output become significantly larger than for the case without this leakage. Further, it becomes obvious that the transient component $p_{\mathrm{t}}(\mathbf{x}_1(\ell))$ and the steady-state component $p_\infty(\mathbf{x}_1(\ell))$ decrease with growing block size. This can be attributed to the fact that the portion of the contribution of the target speaker's signals to the interference-SCM estimates $\vartheta_0(L)$ decreases as the block size grows. As shown in Fig. 4.15(e), this phenomenon can significantly emphasize the improvement of the suppression of the interfering speaker's signal at the beamformer output with growing block size compared to the case without leakage of the target speaker's signals into the interference-SCM estimate. Note that the considered case corresponds to a quite extreme case and the contribution of the target speaker's signal for small block sizes typically is much smaller than for the case shown in Fig. 4.15.

In Fig. 4.16 the numerator $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}_1(\ell)|^2]$ and denominator $\mathbb{E}[|\delta(\mathbf{x}_1(\ell))|^2]$ in (4.59) as well as the resulting behavior of the power $p(\mathbf{x}_1(\ell))$ of a time frequency bin of the interfering speaker's signal at the beamformer output are shown as a function of the block size L. Note that the same time frequency bin as in Fig. 4.13 is investigated in the same way as in Fig. 4.13. Additionally, the effect of the leakage of the target speaker's signals into the interference-SCM estimate and its dependence on the block size are visualized. As discussed w.r.t. Fig. 4.13, the suppression of the interference degrades with growing block size if the interference SCMs are estimated from oracle-separated interference signals so that there is no leakage of the target speaker's signal into the interference-SCM estimates. The overall behavior changes for mask-based beamforming when there is leakage of the target speaker's signal into the interference-SCM estimates. In this case, the tendency can be seen that the suppression of the interfering speaker's signal improves with growing block size. This effect can be attributed to the usually decreasing relative contribution $\vartheta_0(L)$ of the target speaker's signal to the interference-SCM estimate which is shown in Fig. 4.16(a) for the time frequency bin that is considered here. It is to be mentioned that the dominance of the portion of the contribution of the target speaker's signal to the interference-SCM estimate corresponds to a quite extreme case and usually is much less pronounced for the vast majority of time frequency bins.

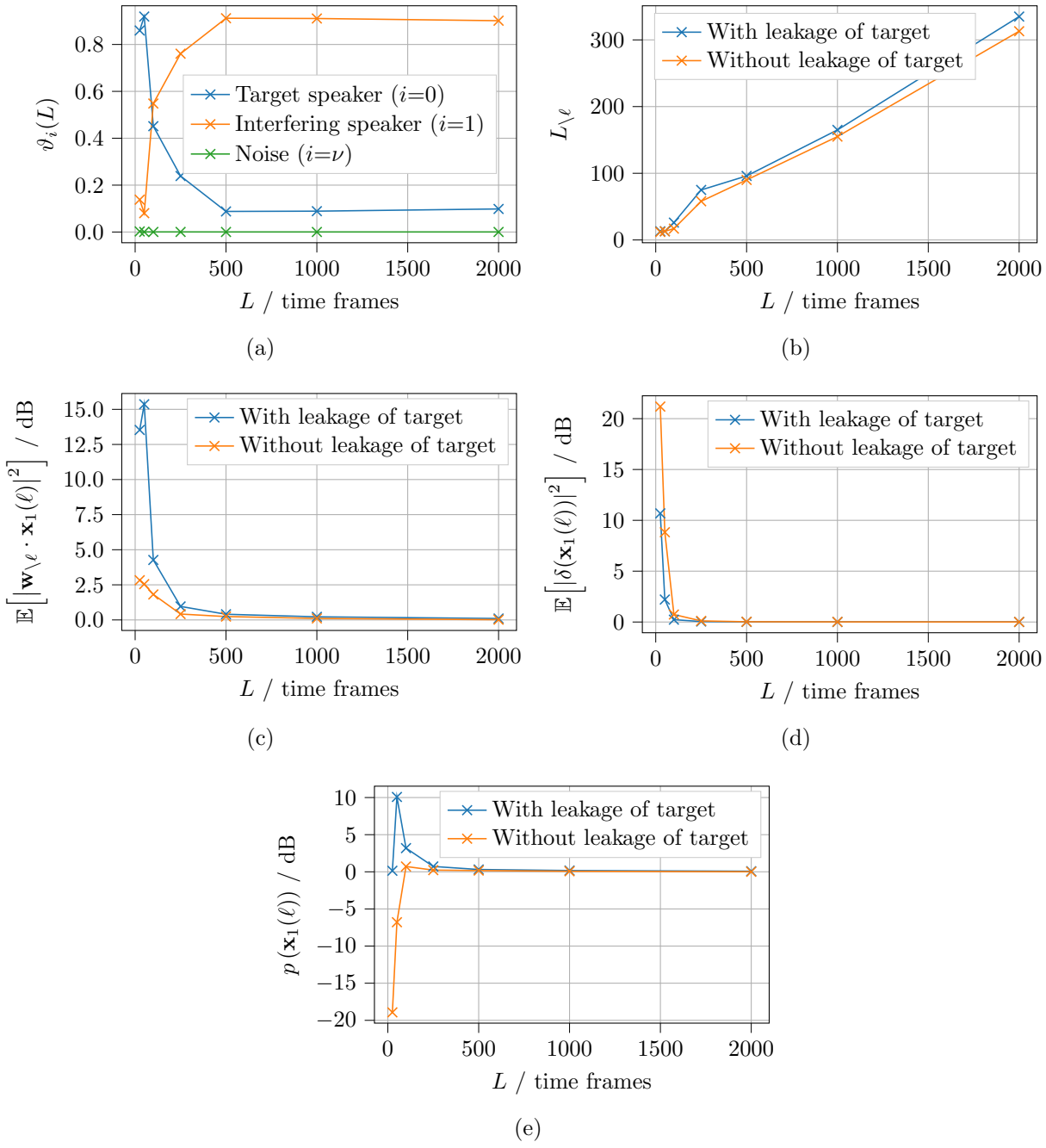


Figure 4.16: Visualization of the effects of the leakage of the target speaker's signal into the interference-SCM estimate on the power of the interfering speaker's components at the beamformer output as a function of the block size L. The same time frequency bin is considered but the block size is varied. The case with leakage of the target signals into the interference-SCM estimates corresponds to beamforming using mask-based SCM estimates while the case without leakage corresponds to beamforming based on SCMs that are estimated from oracle-separated signals. (a) shows the portion of contribution $\vartheta_i(L)$ of the single sources to the interference SCM estimate, (b) shows the equivalent degrees of freedom $L_{\setminus\ell}$ of the Wishart approximation of the interference-SCM estimates, (c) the numerator $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}_1(\ell)|^2]$ in (4.59), (d) the denominator $\mathbb{E}[|\delta(\mathbf{x}_1(\ell))|^2]$ in (4.59) and (e) the resulting power $p(\mathbf{x}_1(\ell))$.

The contribution of the target speaker's signal to the estimate of the interference SCM results in a larger value for the numerator $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}_1(\ell)|^2]$ in (4.59) and a smaller value for the denominator $\mathbb{E}[|\delta(\mathbf{x}_1(\ell))|^2]$ in (4.59) especially if the denominator exhibits a large value for small block sizes. Here, the effect on the numerator is much larger than the effect on the denominator. This corresponds to the intuition since the numerator reflects the effects of the quality of the general statistics of the interference-SCM estimates on the performance of the MVDR beamformer. In contrast, the denominator reflects the focus on the special characteristics of single time frequency bins which is less dependent on the quality of the general statistics of the interference-SCM estimates. Since the numerator is much more affected by the contribution of the target speaker's signals to the interference-SCM estimates, its behavior with growing block size dominates the overall behavior of the power of the interfering speaker's signal at the beamformer output.

Fig. 4.17 depicts the behavior of the numerator $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}_0(\ell)|^2]$ and denominator $\mathbb{E}[|\delta(\mathbf{x}_0(\ell))|^2]$ in (4.59) as well as the resulting behavior of the power $p(\mathbf{x}_0(\ell))$ of a time frequency bin of the target speaker's signal at the beamformer output in the same way as it was done for the interfering speaker's signal in Fig. 4.16. Without leakage of the target speaker's signal into the interference-SCM estimate, the beamformer coefficients and the target speaker's signal are statistically independent. Hence, the denominator becomes one and the behavior of the power of the target speaker's signal at the beamformer output corresponds to the behavior of the numerator which already was discussed w.r.t. Fig. 4.12.

The contribution of the target speaker's signal to the interference-SCM estimate has only a very small effect on the behavior of the numerator $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}_0(\ell)|^2]$ (see Fig. 4.17(c)). With leakage of the target speaker's signals into the interference-SCM estimate, the denominator $\mathbb{E}[|\delta(\mathbf{x}_0(\ell))|^2]$ in (4.59) becomes larger than one, i.e., the target speaker's signal becomes additionally suppressed due to the statistical dependence between the beamformer coefficients and the target speaker's signal. For small block sizes this additional suppression can become quite large since single time frequency bins of the target speaker's signals can dominate the interference-SCM estimate if the interference shows low activity within a block. As for the interfering speaker's case, this leads to a growing power of the target speaker's signal at the beamformer output when the block size grows since the relative contribution of the target speaker's signals to the interference-SCM estimate decreases.

Note that the strong dominance of a single time frame of the target speaker's signal in interference-SCM estimate corresponds to an extreme case and, typically, is less pronounced as in Fig. 4.17, as already discussed before. However, the target speaker's signal becomes strongly suppressed for individual time frequency bins. This can introduce distortions of the target speaker's signal which can harm the listening experience and the performance of downstream tasks, like automatic speech recognition (ASR), if the block size becomes too small and if this case occurs for too many time frequency bins.

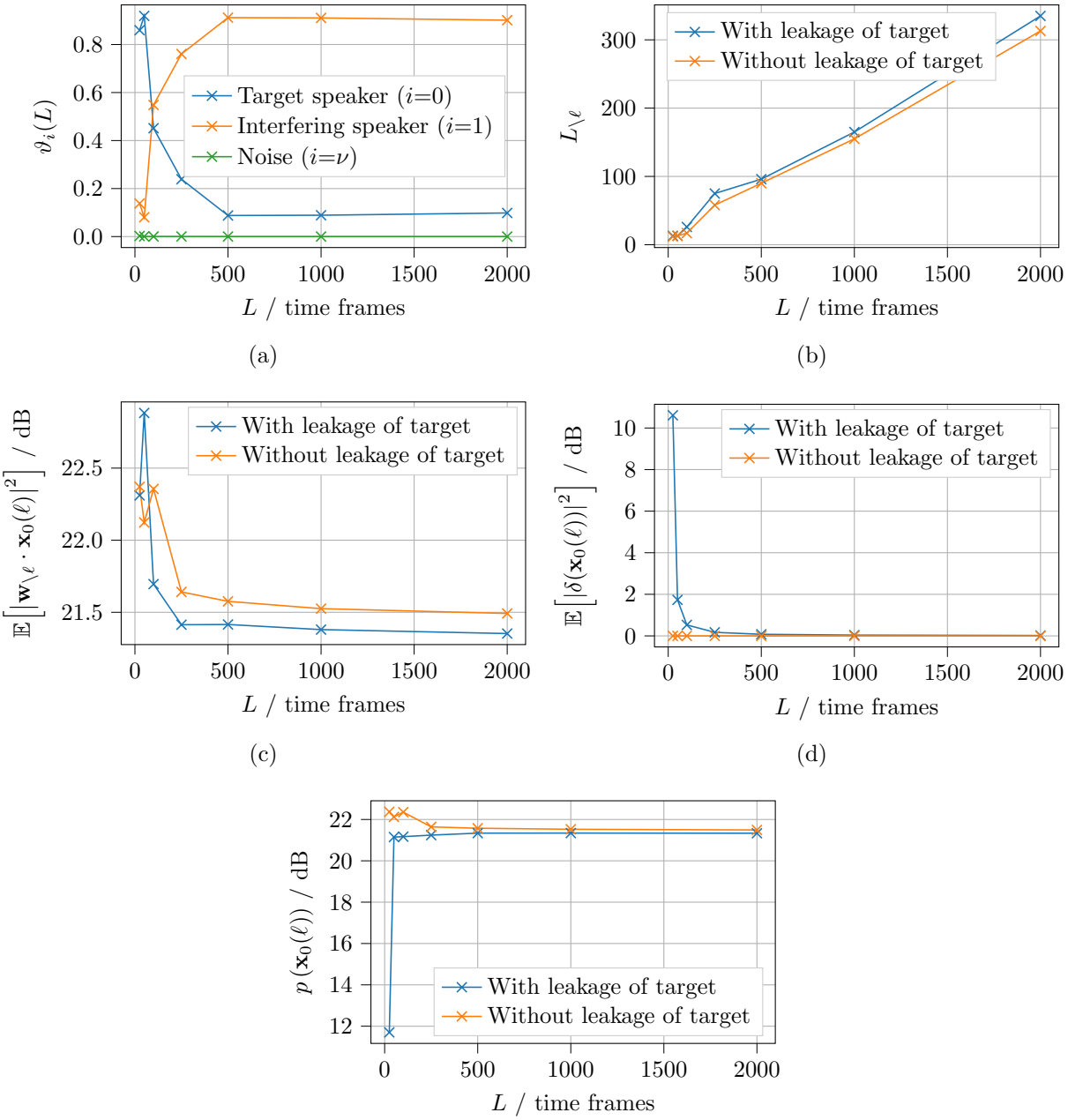


Figure 4.17: Visualization of the effects of the leakage of the target speaker's signal into the interference-SCM estimate on the power of the target speaker's components at the beamformer output as a function of the block size L. The same time frequency bin is considered but the block size is varied. The case with leakage of the target signals into the interference-SCM estimates corresponds to beamforming using mask-based SCM estimates while the case without leakage corresponds to beamforming based on SCMs that are estimated from oracle-separated signals. (a) shows the portion of contribution $\vartheta_i(L)$ of the single sources to the interference SCM estimate, (b) shows the equivalent degrees of freedom $L_{\setminus\ell}$ of the Wishart approximation of the interference-SCM estimates, (c) the numerator $\mathbb{E}[|\mathbf{w}_{\setminus\ell}^{\mathrm{H}} \cdot \mathbf{x}_0(\ell)|^2]$ in (4.59), (d) the denominator $\mathbb{E}[|\delta(\mathbf{x}_0(\ell))|^2]$ in (4.59) and (e) the resulting power $p(\mathbf{x}_0(\ell))$.

4.4.4 Summary

In the following, the analysis of the closed-form approximation of the power of the sources' signals at the beamformer output will be summarized and set into context with the performance of the MVDR beamformer measured by the output SDR. Assuming that the quality of the steering vector and, therefore, the quality of the target SCM estimate are good, the suppression of the interference increases more with growing block size than the suppression of the components of the target signal which are not represented by the target-SCM estimate. This means that the behavior of the output SDR with growing block size is dominated by the behavior of the interference's power at the beamformer output. Consequently, the focus of the discussion of the dependence of the output SDR on the block size particularly lies on the effects of a finite block size on the suppression of the interference.

The behavior of the performance of an MVDR beamformer with an increasing block size is quite complex since it results from a superposition of multiple effects which partially show opposite behavior. As discussed w.r.t. Fig. 4.16, the behavior of the denominator $\mathbb{E}[|\delta(\mathbf{x}_1(\ell))|^2]$ in (4.59) dominates the overall behavior of the power $p(\mathbf{x}_1(\ell))$ of the interference at the beamformer output if the interference SCM can be estimated without the target speaker's signal being present. Consequently, the SDR typically decreases with growing block size in this case, so that a small block size leads to an improved performance of the beamformer. In practice, however, the interference SCM is always estimated from a mix of interference and target signal. As also shown in Fig. 4.16, the behavior reverses, i.e., the suppression of the interference and also the output SDR improve with growing block size. This effect can be attributed to the leakage of the target signal into the interference-SCM estimate. In general, the contribution of the target signal to the interference-SCM estimates decreases with growing block size. Thus, the effect of the declining contribution of the target speaker's signal to the interference-SCM estimates with growing block size typically dominates the overall behavior of the beamformer's performance.

4.5 Evaluation

In this section, the accuracy of the closed-form approximation of the SDR at the beamformer output that was derived in the previous section is first investigated in Sec. 4.5.1. Afterwards, the findings which were made based on the closed-form approximation of the output SDR are further validated by additional experiments in Sec. 4.5.2. In both subsections the experimental setup which was introduced in Chapter 3 is employed.

4.5.1 Evaluation of the closed-form approximation of the output SDR

In the following, the suitability of the statistical model for beamforming and the single steps of the derivation of the closed-form approximation of the output SDR are investigated. This should give insights into the accuracy of the single steps that are used to derive the closed-form approximation:

- The approximation to handle the statistical dependence between the beamformer coefficients and the signals to which they are applied
- The two-stage approximation of the expected value of a compound fraction by a compound fraction of expected values
- The approximation of the probability distribution of the interference-SCM estimates by an equivalent Wishart distribution

If not stated otherwise, the statistical model of MVDR beamforming which was presented in Sec. 4.1 is considered rather than applying beamforming to deterministic speech mixtures.

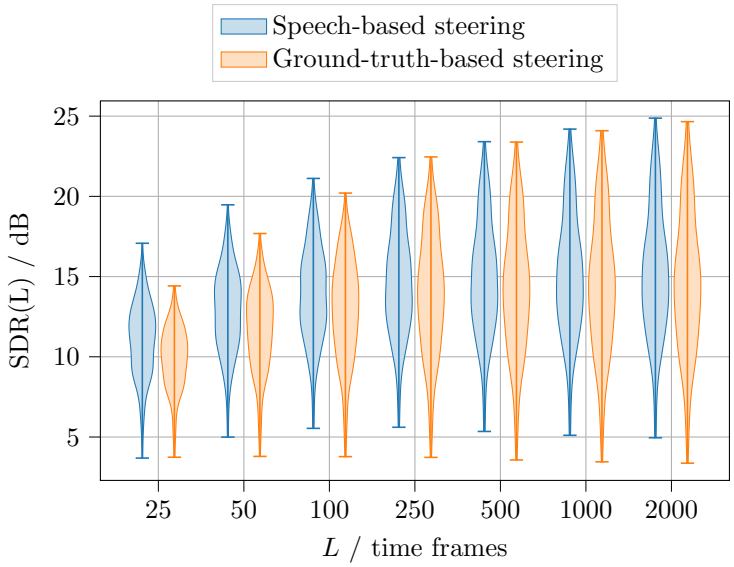
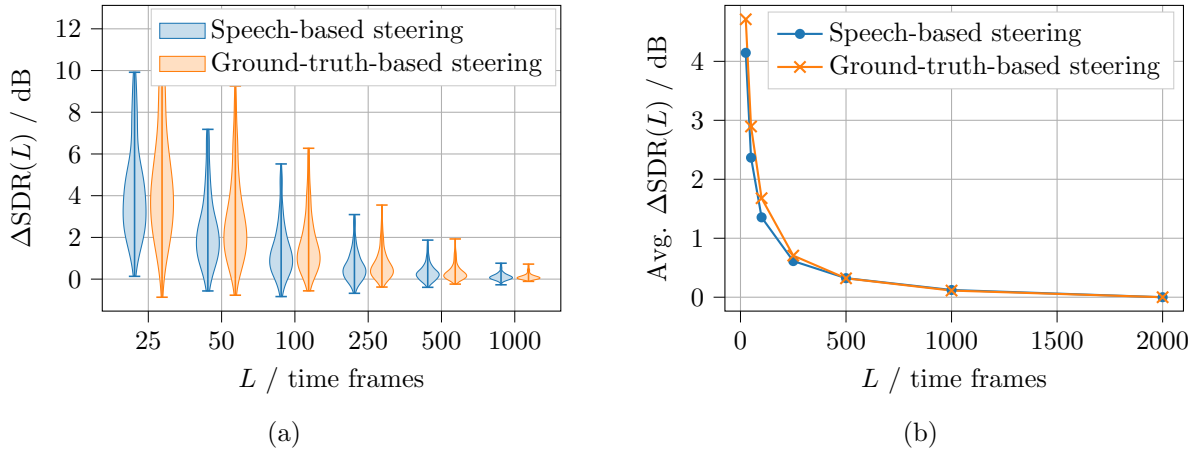


Figure 4.18: Comparison of the output SDR as a function of the block size L between MVDR beamformers using a rank-1 target SCM that is estimated from the speech signals to which the beamformer is applied and MVDR beamformers using a rank-1 target SCM which is estimated via (4.7) based on ground-truth SCMs

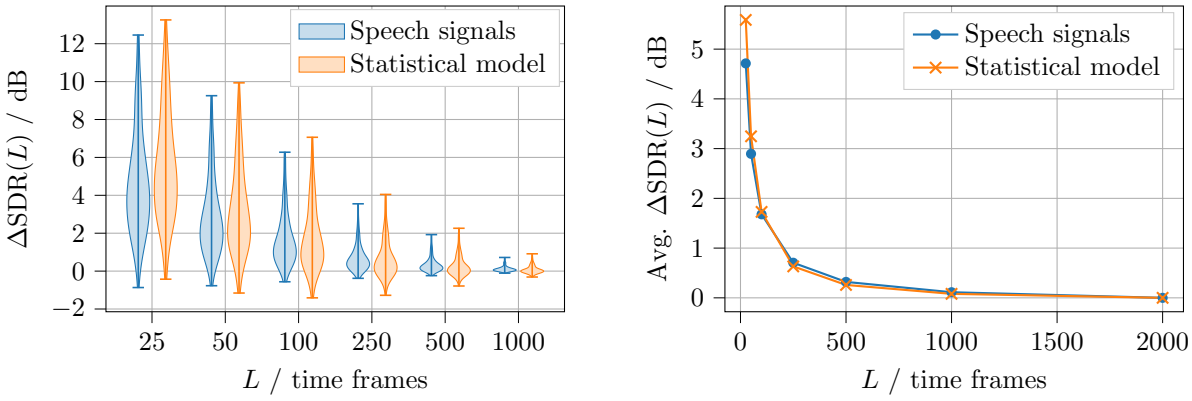
Figure 4.18 shows a comparison of beamforming using GEVD-based steering vectors, which are calculated from SCMs estimated from speech signals, and beamforming using steering vectors, which are estimated via a GEVD of the ground-truth SCM via (4.7). Both types of beamformers are compared based on the distribution of the output SDR as a function of the block size L. Further, beamforming is applied to deterministic speech mixtures rather than considering the statistical model which was introduced in Sec. 4.1. It can be seen that the behavior of the SDR with growing block size is very similar for both variants. However, the beamformer which uses steering vectors estimated from the speech mixture to which the beamformer coefficients are applied later systematically outperforms the beamformer which uses the ground-truth-based steering vectors. This can be explained by the better adaptation to the underlying signals of the current block when estimating the steering vectors form the speech mixture.



(a) (b)

Figure 4.19: Comparison of the output SDR degradation between MVDR beamformers using a rank-1 target SCM that is estimated from the speech signals to which the beamformer is applied and MVDR beamformers using a rank-1 target SCM which is estimated via (4.7) based on ground-truth SCMs. (a) shows the distribution of the SDR degradation $\Delta\mathrm{SDR}(L)$ and (b) the corresponding average value as a function of the block size L.

Since the focus of this analysis of MVDR beamforming lies on the finite sample effects of SCM estimation on beamforming, i.e., the change of the output SDR with growing block sizes, it is more important that the usage of the oracle-based steering vector does not change the behavior of the SDR degradation $\Delta\mathrm{SDR}(L)$, which was introduced in Sec. 3.2. A comparison between the SDR degradation $\Delta\mathrm{SDR}(L)$ for both variants of the steering vector is depicted in Fig. 4.19. It becomes obvious that the SDR degradation $\Delta\mathrm{SDR}(L)$ shows the same behavior for the usage of both variants of the steering vector, which verifies the suitability of the oracle-based steering vector for the previous derivations.



(a) (b)

Figure 4.20: Comparison of the SDR degradation $\Delta\mathrm{SDR}(L)$ calculated via the statistical model of MVDR beamforming and the SDR degradation $\Delta\mathrm{SDR}(L)$ for MVDR beamforming applied to deterministic speech mixtures. (a) shows the distribution of the SDR degradation $\Delta\mathrm{SDR}(L)$ and (b) the corresponding average value as a function of the block size L.

What remains is the question how good the statistical model reflects the behavior of applying MVDR beamforming to deterministic speech mixtures. To this end, Fig. 4.20 compares the behavior of the distribution and the average value of the SDR degradation $\Delta\mathrm{SDR}(L)$ for applying beamforming to deterministic speech mixtures to the behavior of the SDR degradation for the statistical model of beamforming which was presented in Sec. 4.1. It becomes apparent that the statistical model for beamforming well reflects the behavior of the SDR degradation $\Delta\mathrm{SDR}(L)$ for applying beamforming to deterministic speech mixtures.

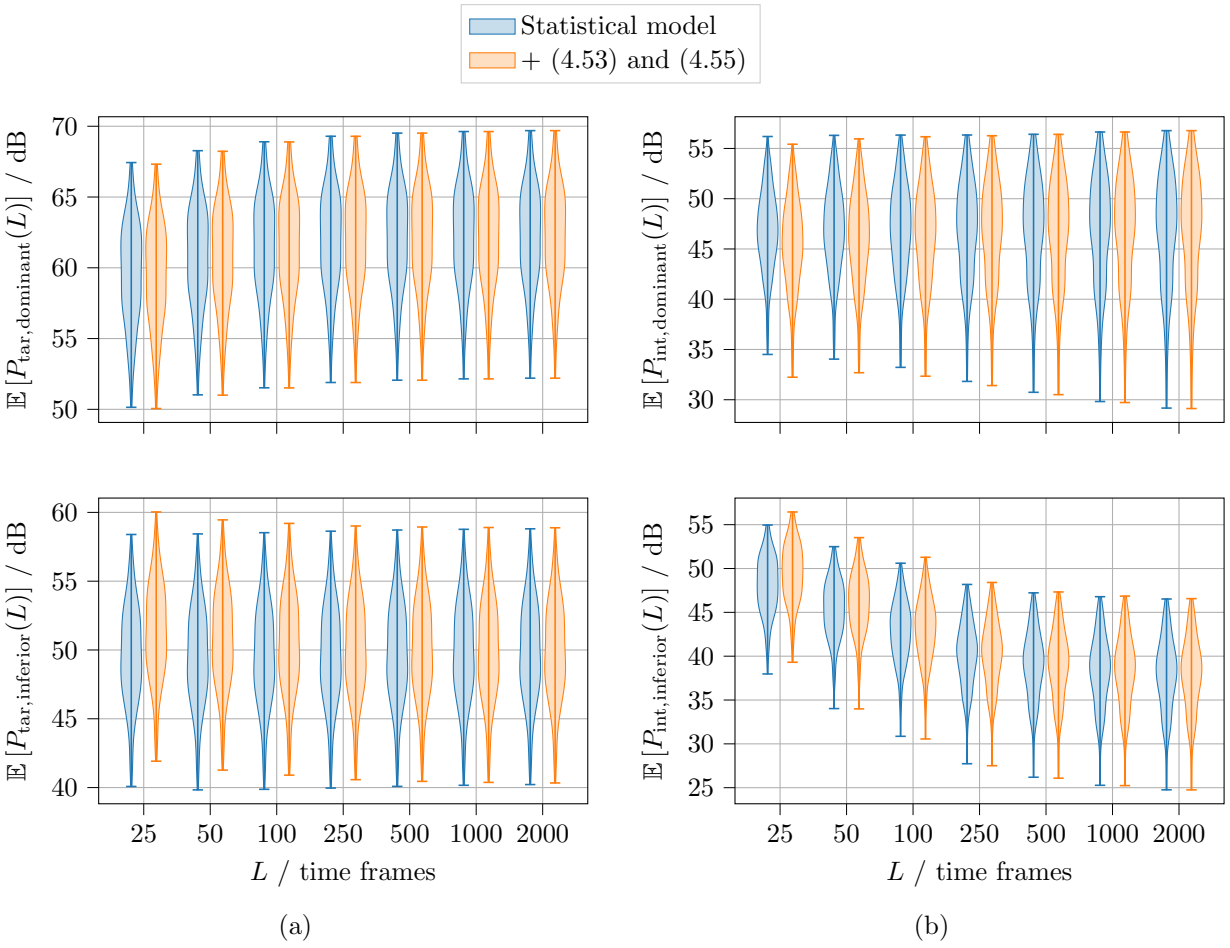


Figure 4.21: Validation of modeling the energy of the sources' signals at the beamformer output by decomposing it into a set of time frequency bins that are dominated by the source of interest (top row: $\mathbb{E}[P_{\mathrm{tar,dominant}}(L)]$ and $\mathbb{E}[P_{\mathrm{int,dominant}}(L)]$) and a set of time frequency bins that are dominated by the source of interest (bottom row: $\mathbb{E}[P_{\mathrm{tar,inferior}}(L)]$ and $\mathbb{E}[P_{\mathrm{int,inferior}}(L)]$). (a) shows the corresponding energies for the target signal at the beamformer output as a function of the block size L and (b) the corresponding energies for the interference signal at the beamformer output as a function of the block size L. The energies are compared to their approximation based on (4.53) and (4.55).

Figure 4.21 shows a comparison of the distributions of the energies of the target speaker's signal and the interference signal at the beamformer output to their approximation via (4.53) and (4.55). It is distinguished between the expected values of the energies $\mathbb{E}[P_{\mathrm{tar,dominant}}(L)]$ and $\mathbb{E}[P_{\mathrm{int,dominant}}(L)]$ for the time frequency bins which are dominated by the source of

interest (see top row of Fig. 4.21) and the expected values of the energies $\mathbb{E}\left[P_{\text{tar,inferior}}(L)\right]$ and $\mathbb{E}\left[P_{\text{int,inferior}}(L)\right]$ for the time frequency bins which are not dominated by the source of interest (see bottom row of Fig. 4.21). These result from restricting the sum over all time frequency bins in (4.38) to the sum over the set of time frequency bins which are dominated by the source of interest and to the sum over the set of time frequency bins which are not dominated by the source of interest, respectively, using the definitions in (4.49) and (4.50).

It becomes obvious that the error of the energy of the sources' signals at the beamformer output which is introduced by the approximations in (4.53) and (4.55) are acceptably small. Moreover, these errors even vanish with growing block size. In addition to that, it can be seen that the general behavior of the energy of the sources' signals at the beamformer output, which already has been discussed for the power of single time frequency bins in the previous section, is well represented.

Thereby, the energy of the target signal at the beamformer output is less influenced by the choice of the block size than the energy of the interference signal. Furthermore, the energy $\mathbb{E}\left[P_{\text{tar,dominant}}(L)\right]$ of the target signal's components that dominate their corresponding time frequency bins grows as the block size increases which reflects the tendency of an unwanted additional suppression of these time frequency bins due to statistical dependence to the beamformer coefficients. However, this effect diminishes with growing block size, as discussed before.

Additionally, it can be observed that the energy $\mathbb{E}\left[P_{\text{int,inferior}}(L)\right]$ of the components of the interference signal which are not dominant w.r.t. their corresponding time frequency bins decreases with growing block size L until it converges for large block sizes. This means that the suppression of these components improves with growing block size. The energy $\mathbb{E}\left[P_{\text{int,dominant}}(L)\right]$ of the interference signal which are dominant w.r.t. their corresponding time frequency bins show an increasing behavior with growing block size in some cases and a decreasing behavior with growing block size for other cases. This reflects the interplay of the diminishing focus on the special characteristics of individual time frequency bins and the improvement of the accuracy of the general statistics captured by the interference-SCM estimates as the block size increases.

The accuracy of the approximations given by (4.53) and (4.55) w.r.t. the resulting distribution of the SDR degradation $\Delta\text{SDR}(L)$ (see Fig. 4.22(a)) and the resulting average SDR degradation (see Fig. 4.22(b)) as a function of the block size L is visualized in Fig. 4.22. As for the single energies which are involved in the calculation of the SDR and which were considered in Fig. 4.21, the approximations that were introduced to handle the statistical dependence between the beamformer coefficients and the signal to which they are applied maintain the generally decreasing behavior of the SDR degradation with growing block size. There are only minimal differences in the distribution of the SDR degradation and, therefore, also in the average SDR degradation. This especially holds for small block sizes. It is to be mentioned that the SDR degradation becomes negative for some examples. For those examples the improved suppression of the interference due to the statistical dependence of the beamformer coefficients and the interference signals dominates the overall behavior. Moreover, the generally decreasing behavior of the SDR degradation $\Delta\text{SDR}(L)$ suggests that the impact of the leakage of the target speaker's signal into the interference-SCM estimates,

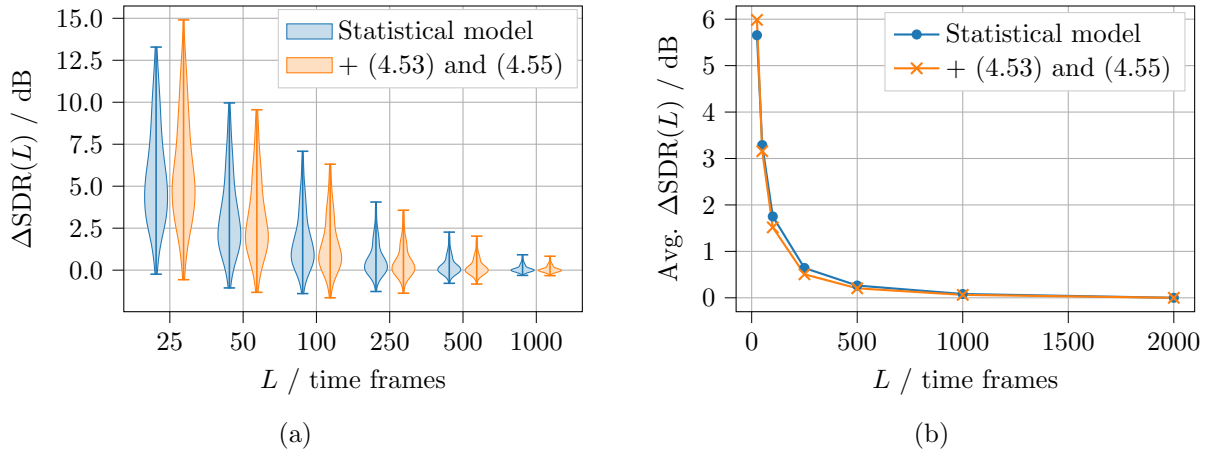


Figure 4.22: Validation of the approximation of the SDR degradation via (4.53) and (4.55). (a) shows the distribution of the SDR degradation $\Delta\mathrm{SDR}(L)$ and (b) the average SDR degradation as a function of the block size L.

which diminishes with growing block size, dominates the behavior of the MVDR beamformer's performance.

Figure 4.23 visualizes the accuracy of the stepwise approximation of the expected value of the compound fraction by a compound fraction of expected values as specified in (4.59) and (4.71), respectively. For this purpose, the distribution of the SDR degradation $\Delta\mathrm{SDR}(L)$ with these approximations and the distribution of the SDR degradation without the approximations of the expected value of a compound fraction are compared. The first stage of this approximation,

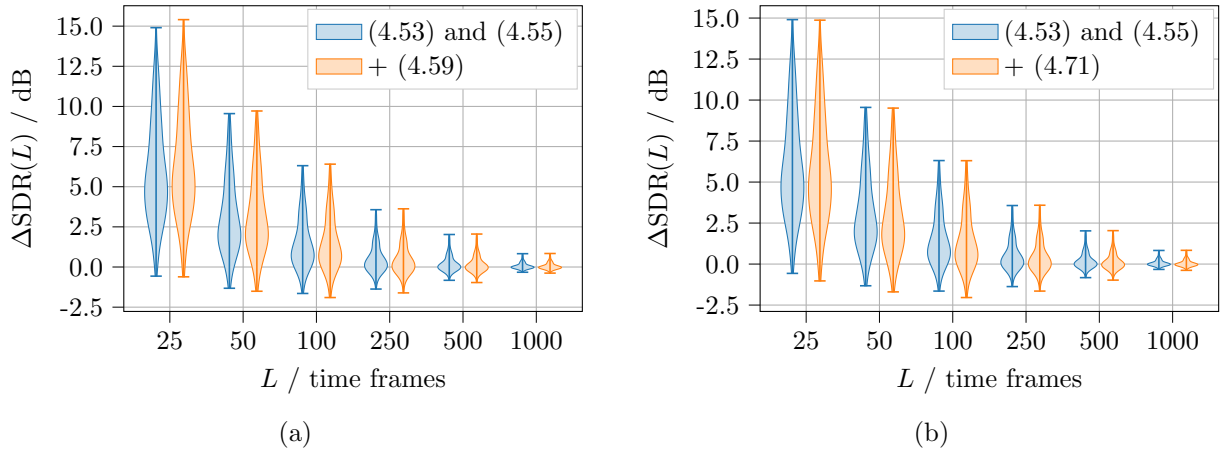


Figure 4.23: Validation of the two stages involved in the approximation of the expected value of a compound fraction as compound fraction of expected values used for the closed-form approximation of the power of the sources' signals at the beamformer output. The distribution of the SDR degradation $\Delta\mathrm{SDR}(L)$ as a function of the block size L is considered to measure the accuracy of the two-stage approximation. (a) corresponds to stage one specified in (4.59) and (b) corresponds to stage two specified in (4.71).

which is specified in (4.59), is considered in Fig. 4.23(a) and the second stage, which is specified in (4.71), is considered in Fig. 4.23(b). As already indicated by considering the effect of the stagewise approximation of the expected value of the compound fraction by a compound fraction of expected values on the interfering speaker's signal at the beamformer output in Fig. 4.8 and Fig. 4.9, it becomes obvious that both stages lead to a valid approximation of the SDR degradation.

Finally, Fig. 4.24 visualizes the accuracy of the closed-form approximation of the SDR at the beamformer output that was derived in Sec. 4.3. Figure 4.24(a) shows the distribution of the SDR at the beamformer output, the distribution of the SDR degradation $\Delta\mathrm{SDR}(L)$ and the average SDR degradation. The quantities are calculated based a Monte-Carlo simulation that applies all approximations that were already validated in this section and then compared to the corresponding quantities which result from the closed-form approximation. In this way, the errors introduced by the approximation of the probability distribution of the interference-SCM estimates by an equivalent Wishart distribution can be isolated from the errors introduced by the other approximations.

It can be seen that the closed-form approximation is able to fundamentally model the behavior of the output SDR of an MVDR beamformer as a function of the block size L. However, the deviations which are introduced by the Wishart approximation of the interference-SCM estimates are larger than the errors that are introduced by the previously discussed approximations. This especially holds for small block sizes for which the approximation of the interference-SCM estimates' probability distribution by a Wishart distribution has weaknesses. These weaknesses were shown, for example, in Fig. 4.5(b) for the resulting second-order moment of the beamformer coefficients which are decisive for the quality of the numerator of the power of the sources' signals at beamformer output in (4.59). But, these errors still are acceptable when considering the relation to the distribution towards which the SDR converges for large block sizes.

A similar error pattern due to the Wishart approximation of the probability distribution of the interference-SCM estimates arises for the SDR degradation $\Delta\mathrm{SDR}(L)$. However, the final closed-form approximation of the SDR at the beamformer output can be seen as a suitable model for investigating the relationship between the performance of an MVDR beamformer and the sample size used for SCM estimation. This follows from the fact that the overall behavior of the SDR degradation as a function of the block size nevertheless is well modeled and the resulting relative errors still are acceptably small.

Finally, the closed-form approximation of the output SDR is compared to the output SDR for applying the beamformer to the deterministic, simulated speech mixtures that correspond to the scenarios considered for the closed-form approximation of the output SDR. This comparison is shown in Fig. 4.24(b). Again, the comparison is based on the distribution of the output SDR, the distribution of the SDR degradation $\Delta\mathrm{SDR}(L)$ and the average SDR degradation. It can be observed that the deviations of the closed-form approximation of the SDR at the beamformer output to the output SDR resulting from applying the MVDR beamformer to deterministic speech mixtures are acceptably small for the SDR as well as for the SDR degradation $\Delta\mathrm{SDR}(L)$ w.r.t. the value range of these quantities. Hence, the closed-form approximation of the output SDR is not only able to model the rather

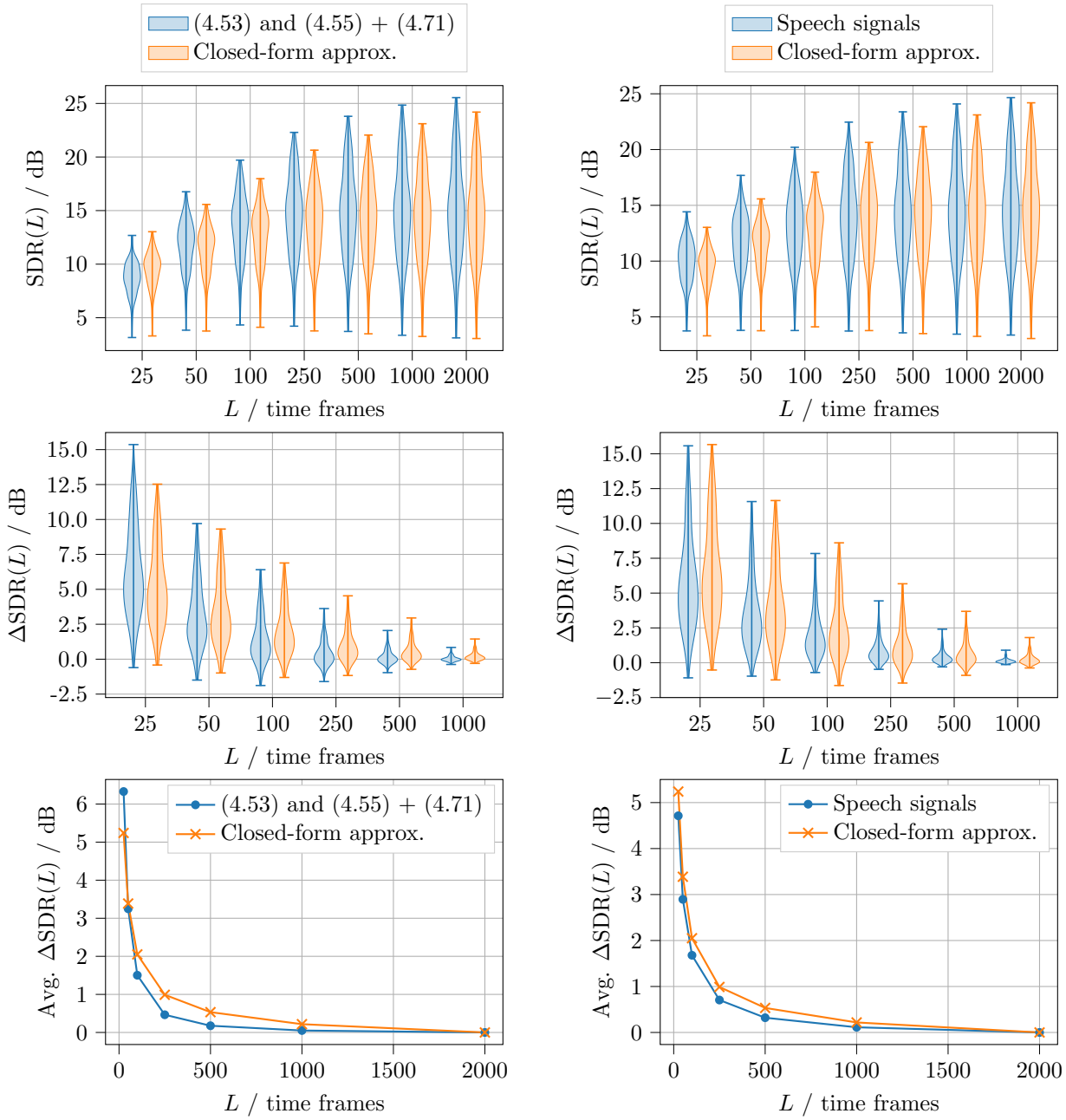


Figure 4.24: Validation of the closed-form approximation of the SDR at the beamformer output. To this end, the closed-form approximation of the SDR at the beamformer output is compared to the SDR following from the statistical model for beamforming (a), including all approximation which were already validated in this section, and the SDR for applying the beamformer to deterministic speech mixtures (b). The distribution of the SDR as a function of the block size L is shown in the top row, the distribution of the SDR degradation $\Delta\text{SDR}(L)$ as a function of the block size L is shown in the middle row and the average value of the SDR degradation $\Delta\text{SDR}(L)$ as a function of the block size L is shown in the bottom row.

theoretical statistical model of MVDR beamforming but also reflects the practical application of beamforming to speech signals.

4.5.2 Experimental validation

This subsection aims at getting deeper insights into the effect of a finite sample size used for SCM estimation on the performance of an MVDR beamformer by investigating the influence of the STFT window size, the influence of the sound decay time and the influence of the number of microphones on this effect. For this purpose, the application of MVDR beamforming to deterministic speech mixtures is considered instead of using the statistical model of MVDR beamforming which was employed for the derivation of the closed-form approximation of the output SDR in Sec. 4.3. In order to extract the target speaker's signal, the formulation of the Souden-MVDR beamformer with a GEVD-based rank-1 approximation of the target-SCM estimate is utilized. Again, block-wise beamforming is considered.

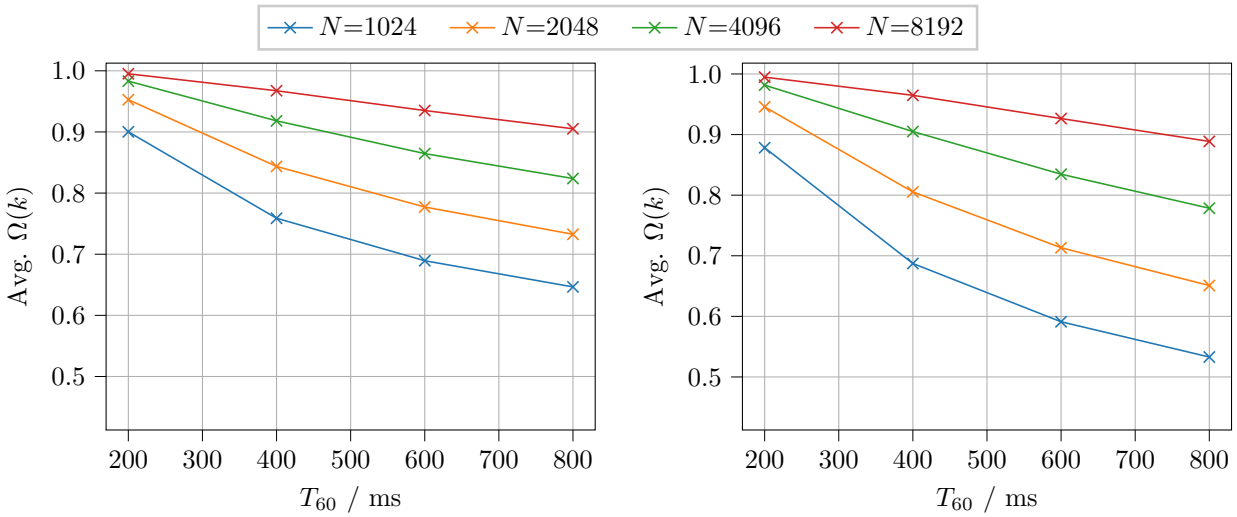


Figure 4.25: Influence of the sound decay time T_{60} and the size of the STFT analysis window N on the dominance of the largest eigenvalue of the SCMs for M=3 microphones (a) and M=6 microphones (b). The dominance of the largest eigenvalue is measured by the ratio $\Omega(k)$ between the largest eigenvalue and the sum of all eigenvalues.

Figure 4.25 shows the relationship between the size of the STFT analysis window N and the dominance of the largest eigenvalue. To this end, the full-rank SCMs $\mathbf{R}_q(k)$ in the LGM of the STFT of speech signals are estimated as described in Sec. 3.4 and an EVD is performed. As a measure of the degree of dominance, the eigenvalue ratio

$$\Omega(k) = \frac{\lambda_0(k)}{\sum_{i=0}^{M-1} \lambda_i(k)} \tag{4.117}$$

is utilized, with $\lambda_i(k)$ denoting the eigenvalues of the SCM $\mathbf{R}_q(k)$ and $\lambda_0(k)$ being the largest eigenvalue. This ratio is averaged over all frequency bins and the SCMs of all speakers in the dataset. The value of the eigenvalue ratio becomes closer to one if the eigenvector belonging to the largest eigenvalue becomes more dominant.

As described in [12], the narrowband approximation of the spectral model in (2.3) becomes more accurate if the size of the STFT window N is sufficiently large w.r.t. to the length of the room impulse response (RIR) and, respectively, the sound decay time T_{60}. This implies that larger window sizes are needed to model acoustic transfer functions (ATFs) belonging to larger sound decay times. Hence, the largest eigenvalue and the corresponding eigenvector become more dominant with increasing size of the STFT analysis window. Note that this effects also shows a slight dependence of the number of microphones M and, thus, the size of the SCMs. Here, the tendency is that the largest eigenvalue and the corresponding eigenvector are more dominant for the case with fewer microphones than for the case with more microphones.

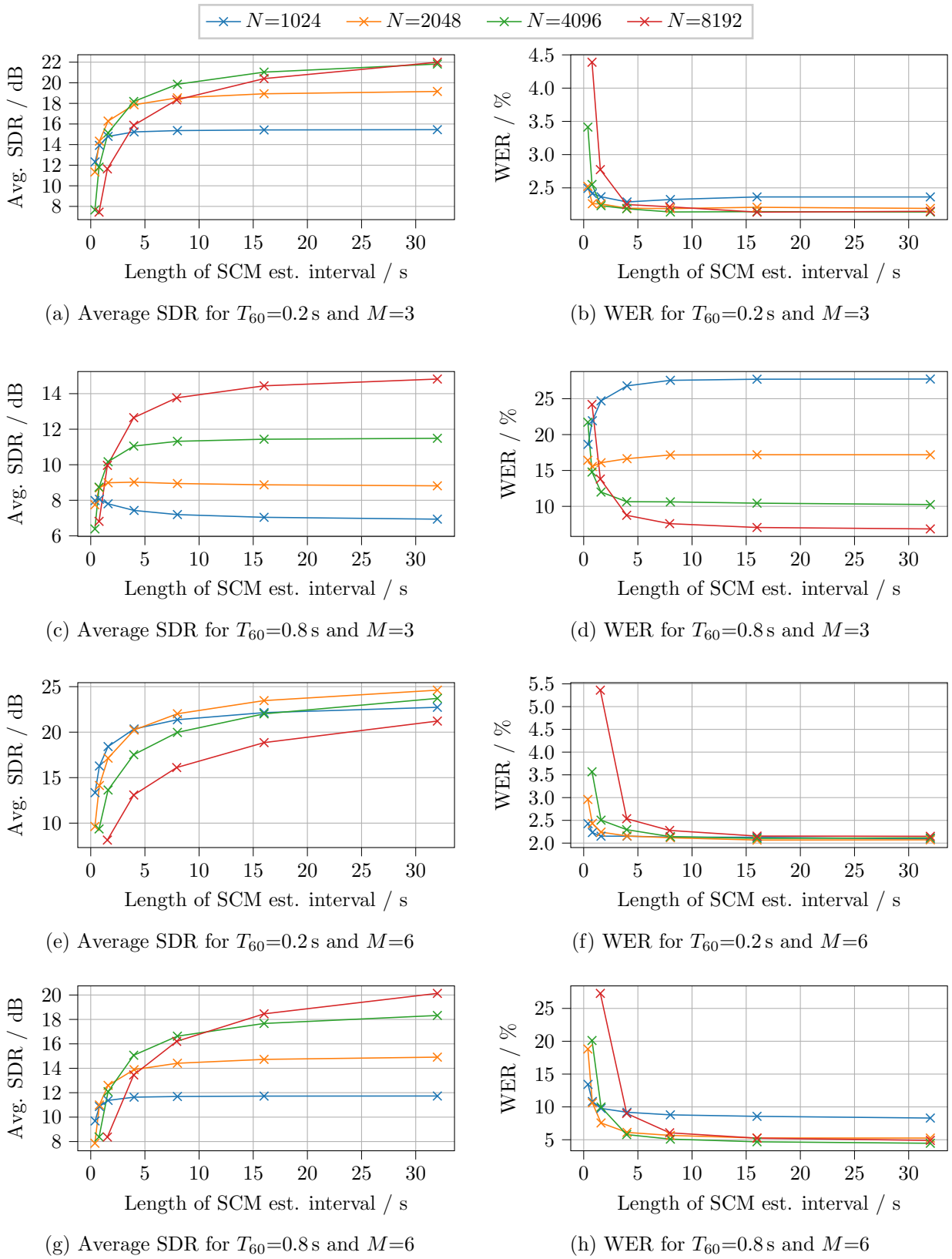
Figure 4.26 visualizes the effect of the size of the STFT analysis window N on the beamforming performance with special focus on the block size L, i.e., the sample size used for SCM estimation. First of all, it can be seen that the invasive SDR and the word error rate (WER) correlate well, i.e., a higher average SDR usually comes with a lower WER. Therefore, the results of the previous investigation of the finite sample size effects on the output SDR should allow good conclusions to be drawn about the performance of ASR as a downstream task. However, for high SDR values the correlation might be worse since the SDR might be improving with an increasing sample size used for SCM estimation but the interference is already suppressed such good that it has no influence of the ASR performance. For example, this can be observed in Fig. 4.26(e) and Fig. 4.26(f).

Moreover, some general tendencies can be seen in Fig. 4.26. On the one hand, the beamforming performance is generally better if more microphones are used since more microphones typically increase the degrees of freedom of the beamformer. On the other hand, the beamforming performance typically degrades for higher sound decay times T_{60} such that the eigenvector of the SCMs belonging to the largest eigenvalue is less dominant and the structure of the SCMs generally is more complex.

In addition to that, the beamforming performance, i.e., the output SDR and the WER, seem to converge more slowly for larger STFT window sizes N. This can be explained by the fact that the frame advance of the STFT is chosen as 25 % of the window size. Thus, for a given length of the SCM estimation interval in seconds the resulting number of time frames used for SCM estimation is smaller for larger window sizes.

Furthermore, it can be seen that a better beamforming performance can be achieved by using larger STFT window sizes N if the length of the signal segments from which the SCMs are estimated is large enough. This might be explained by the fact that the eigenvector belonging to the largest eigenvalue becomes more dominant. Consequently, the beamformer can better focus on the suppression of the interference.

It is to be mentioned that constellations exist for which a smaller sample size leads to better beamforming performance (see Fig. 4.26(c)) although it is to be expected that a larger sample size leads to better beamforming performance, as discussed before. For example, considering



(a) Average SDR for T_{60}=0.2 s and M=3

(b) WER for T_{60}=0.2 s and M=3

(c) Average SDR for T_{60}=0.8 s and M=3

(d) WER for T_{60}=0.8 s and M=3

(e) Average SDR for T_{60}=0.2 s and M=6

(f) WER for T_{60}=0.2 s and M=6

(g) Average SDR for T_{60}=0.8 s and M=6

(h) WER for T_{60}=0.8 s and M=6

Figure 4.26: Impact of estimating SCMs from a finite sample size on the beamforming performance, across different STFT window sizes N, sound decay times T_{60} and numbers of microphones M. The left column shows the average SDR at the output of the beamformer as a function of the length of the interval used for SCM estimation and the right column the word error rate (WER) as a function of the length of the interval used for SCM estimation for ASR as downstream task.

beamforming with M=3 microphones and large sound decay times T_{60}, it is better to use fewer time frames for SCM estimation if the STFT window size N is small. This effect becomes even more visible in the significant increase of the WER (see Fig. 4.26(d)).

Despite the better beamforming performance which can be achieved with larger STFT window sizes N, larger window sizes might be impractical in distributed setups with sampling rate offsets (SROs). Here, larger signal lengths, which are required to have enough time frames for SCM estimation, come with larger time drifts during the SCM estimation interval due to the SROs. However, this is detrimental for beamforming which will be discussed in Sec. 5.3. Additionally, the length of the signal segment in which the interference can be observed is often limited to a few seconds in practical applications, like meeting recognition. Thus, in the remainder of this thesis a typically used STFT window size of N=1024 samples will be considered to keep the effects of SROs in a manageable range.

The dependence of the finite sample size effects on the sound decay time T_{60} and the number of microphones M is depicted in Fig. 4.27. It can be seen that for smaller sound decay times T_{60}, i.e., less reverberation, the possible improvement of the output SDR with an increasing block size is larger than for large sound decay times. This might be explained by the fact that the SCMs to be estimated are closer to a rank-1 matrix for small sound decay times and, thus, the beamformer can benefit to a greater extent from more accurate estimates of the spatial statistics of the interference due to a larger sample size. However, this improvement is not of great importance for downstream tasks, like ASR, since the interference is already suppressed such well that the ASR system does not benefit from further improvement.

Under more reverberant conditions the possible improvement due to a larger sample size used for SCM estimation is much smaller than for a low sound decay time T_{60}. However, these small possible improvements are of larger importance for ASR as downstream task. This can be explained by the fact that especially examples with low SDR values dominate the WER. The number of examples with low output SDR is reduced due to the better estimates of the spatial statistics following from an increased sample size used for SCM estimation.

In addition to that, it becomes obvious that the SDR improves more with increasing sample size for SCM estimation if the number of microphones M, used for beamforming, is large. This can be explained by the additional degrees of freedom due to the additional microphones as a result of which more samples are needed for a robust estimate of the more complex structure of the SCMs.

Again, it can be seen that the finite sample size effects can also lead to an improved beamforming performance for SCMs estimated from a small sample size compared to SCMs estimated from a large sample size. For example, this is the case for large sound decay times T_{60} and a small number of microphones M.

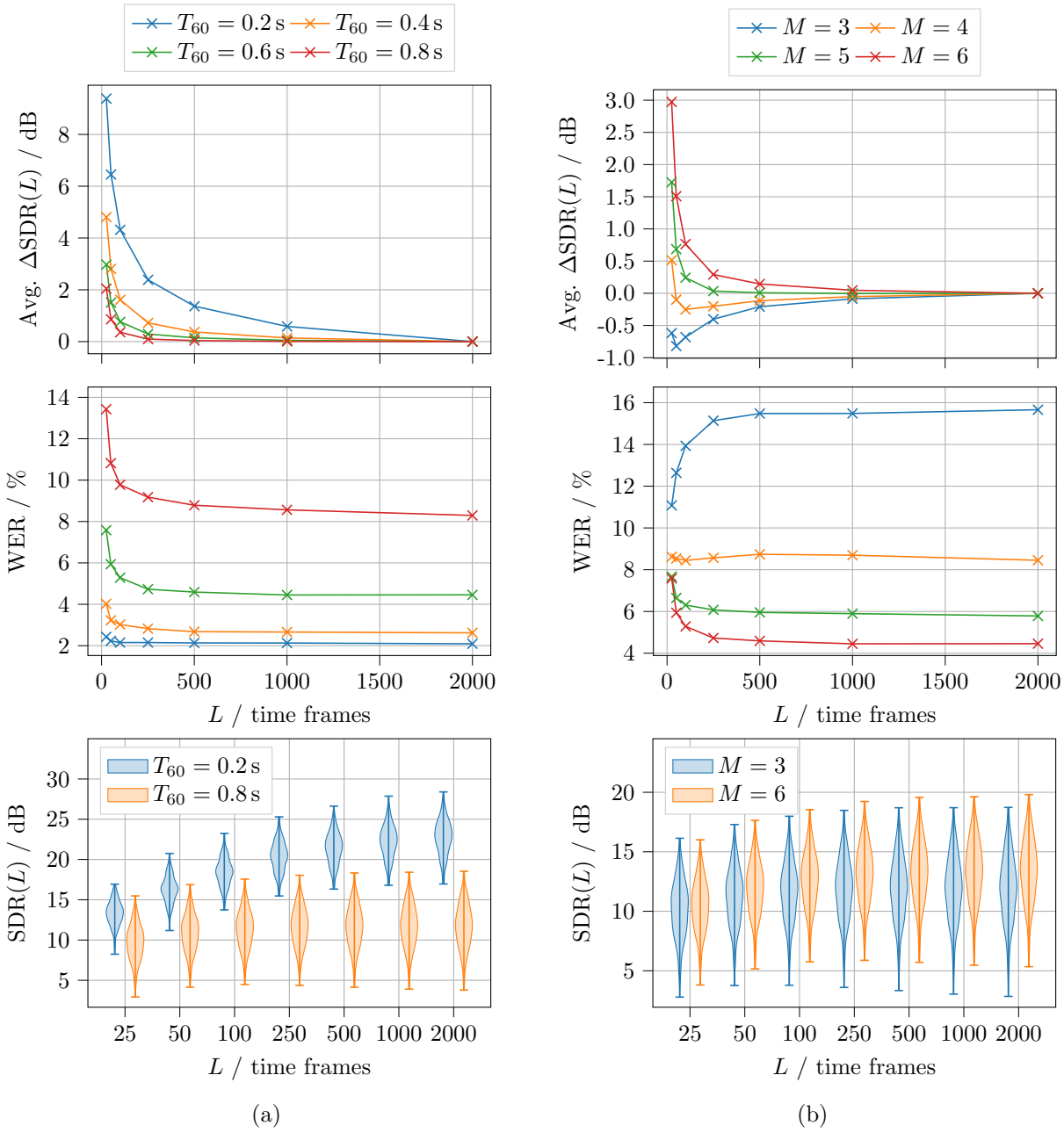


Figure 4.27: Impact of SCMs estimated from a finite sample size on the beamforming performance, across different sound decay times T_{60} for M=6 microphones (a) and different numbers of microphones M for a sound decay time of T_{60}=0.6 s (b). The average value of the SDR degradation $\Delta\mathrm{SDR}(L)$ as a function of the block size L is shown in the top row, the WER as a function of the block size L for ASR as downstream task is shown in the middle row and the distribution of the SDR as a function of the block size L is shown in the bottom row.

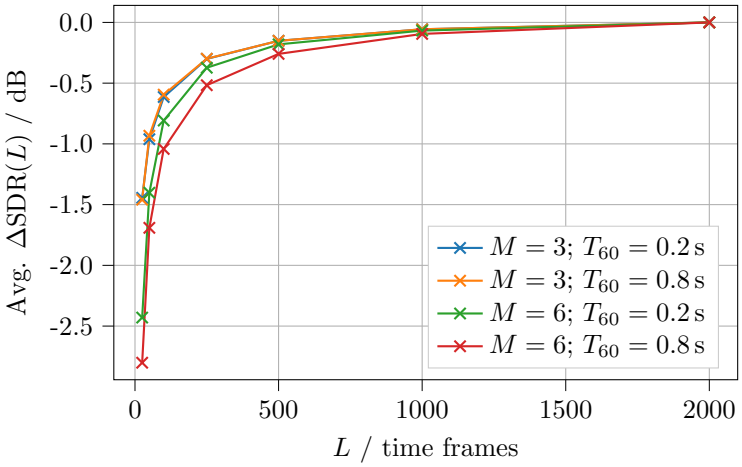


Figure 4.28: Visualization of the effect of SCM estimation from a finite sample size on the performance of an MVDR beamformer for SCM estimation from oracle-separated target and interference signals. The average value of the SDR degradation $\Delta\mathrm{SDR}(L)$ as a function of the block size L is shown for different numbers of microphones M and sound decay times T_{60}.

In order to assess the influence of the leakage of the target speaker's signals into the interference-SCM estimate and the interference signals into the target-SCM estimate, Fig. 4.28 depicts the average SDR degradation $\Delta\mathrm{SDR}(L)$ as a function of the block size L for the case that the SCMs are estimated from oracle-separated target and interference signals. In this case, there is no leakage of unwanted signal components into the SCM estimates. It becomes apparent that the SDR degradation is negative for all considered cases, i.e., the SDR generally is better when using a small block size for SCM estimation. This can be attributed to the improved suppression of the interference due to the statistical dependence between the beamformer coefficients and the signals to which they are applied. This was already discussed in Sec. 4.4. By comparing these results to the results for mask-based beamforming in Fig. 4.26 where the SDR generally improves as the block size increases, it can be concluded that in mask-based beamforming the diminishing leakage of unwanted signal components into the SCM estimates with growing block size dominates the overall behavior of the MVDR beamformer with growing sample size used for SCM estimation in most cases.

4.6 Summary

In this chapter, a statistical model was developed to describe the dependence between the performance of an MVDR beamformer and the sample size used for estimating the interference SCM that is needed to calculate the beamformer coefficients. Based on this statistical model an approximation of the probability distribution of the interference-SCM estimates as well as a closed-form approximation of the SDR at the beamformer output were derived. It was shown that the closed-form approximation of the SDR at the beamformer output provides a good model of the relationship between the performance of an MVDR beamformer and the size of the SCM estimation interval. From the closed-form approximation of the output SDR, it can be seen that the behavior of the beamformer's performance with growing block size is a result of three partially counteracting effects:

- The accuracy of the estimated statistics of the interference in form of the interference-SCM estimates improves with growing sample size used for SCM estimation. This effect is reflected by the numerator in (4.59) and results in an improved suppression of the interference with growing block size among other effects.
- Due to the statistical dependence between the beamformer coefficients and the signals to which they are applied in block-wise beamforming, time frames that dominate the interference-SCM estimate get additionally suppressed. This effect arises from the denominator in (4.59). It occurs especially for small block sizes and diminishes with growing block size.
- The decreasing relative contribution of the target speaker's signals to the interference-SCM estimate with growing block size has a large effect on the performance of the beamformer. The leakage generally results in a decreasing suppression of target speaker's signal components which are not represented by the steering vector and an improved suppression of the interference with growing block size.

Typically, the contribution of the target speaker's signals to the interference-SCM estimates dominates the performance of the MVDR beamformer as a function of the block size so that the SDR at the beamformer output improves with growing block size. Therefore, it is recommended to choose the block size, i.e., the SCM estimation interval, as large as possible in practice if typical masks, which tend to make errors and do not set target dominated time frequency bins explicitly to zero, are used for SCM estimation. However, if the masks tend to approximate SCM estimation from oracle-separated signals, e.g., by forcing time frequency bins that are very likely not dominated by the interference to zero, short SCM estimation intervals are better in block-wise beamforming. This choice, however, comes with a high risk and, therefore, is not recommended in practice where the mask estimates exhibit errors.

5 MVDR beamforming using asynchronous, distributed devices

The spatial distribution of recording devices offers the possibility of an improved signal capture and signal enhancement, e.g., via beamforming, as indicated, for example, in [OC3], [OC4]. However, distributed recording devices typically come with the drawback of independent sampling clocks, which results in sampling time offsets (STOs) and sampling rate offsets (SROs) between the recorded signals. In previous works [1]–[6], it was shown that STOs and SROs can have a significant negative impact on the performance of a beamformer if they are not compensated for.

This chapter shows that a spatially distributed recording setup can lead to an improved performance of source extraction via beamforming, if the STOs and SROs are compensated for, compared to the case of beamforming using a compact microphone array. While the spacing between the microphones of a compact microphone array typically is a few centimeters, the spacing between the microphones of spatially distributed recording setup is much larger and can even exhibit values of a few meters. Moreover, the effects of STOs and SROs on spatial covariance matrix (SCM) estimation as well as their impact on the performance of an minimum variance distortionless response (MVDR) beamformer are discussed in this chapter.

A typical application of acoustic beamforming is meeting recognition, i.e., recording a meeting with a set of microphones in order to transcribe who said what. Many works on meeting recognition either only consider the case of a compact microphone array [27], [64]–[75] or the case of distributed recording devices [6], [76]–[78]. A direct comparison of both cases only is performed in few works. The results presented in [OC3], [OC4] indicate that an overall better performance can be achieved when using a setup with distributed recording devices rather than a compact microphone array. However, both studies do not allow to draw conclusions about the advantages of distributed recording devices for beamforming. Despite the pre-processing stages, like mask estimation, being the same in [OC3], the comparison of beamforming utilizing a compact microphone array and beamforming utilizing distributed recording devices in [OC3] was not fair. This is due to the different numbers of microphones used in the two setups. While the number of microphones was the same for both recording conditions in [OC4], only results for the overall meeting recognition performance were reported. Thus, it is not possible to differentiate between the advantage due to a distributed recording setup which is made in the early stages of the system, e.g., in the mask estimation stage, and the advantage which is made in the beamforming stage.

This chapter aims at providing a systematic comparison of beamforming using a compact microphone array and beamforming using a distributed recording setup by taking meeting

recognition as example application and using MVDR beamforming for source extraction. Here, the focus lies on source extraction via beamforming since earlier stages of the processing pipeline and the system used for automatic speech recognition (ASR) as downstream task can be arbitrarily exchanged by other algorithms and systems.

In addition to the advantages of distributed recording devices, the occurring STOs and SROs might have detrimental impacts on the performance of an MVDR beamformer. It was demonstrated in [1] that even STOs of only a few samples significantly degrade the performance of an MVDR beamformer when the steering vector is not estimated from the signals to which the beamformer is applied. This was shown for steering vectors which are based on relative acoustic transfer functions (ATFs) being estimated from the ground-truth room impulse responses (RIRs). Furthermore, it was proven in [1] that the detrimental impact of STOs on MVDR beamforming can be counteracted by employing a generalized eigenvalue decomposition (GEVD) based steering vector that is estimated from the observed signals. In this thesis, it is demonstrated that the invariance of a GEVD-based MVDR beamformer from STOs, which was proven in [1], only holds if the STOs are much smaller than the size of the analysis window of the short-time Fourier transform (STFT).

Furthermore, a significant drop in performance of different types of beamformers due to SROs was reported before [2]–[6]. It was shown in [4] that this drop in performance can be attributed to a distortion of the beampattern. Further insights into the reasons for the SRO-induced degradation of the beamforming performance were gained in [5], where it was shown that SROs induce a distortion of the amplitude of the SCM estimates used to calculate the beamformer coefficients. The strength of this distortion increases as the sample size used for SCM estimation grows. As demonstrated in this chapter, the degradation of the performance of an MVDR beamformer due to SROs additionally results from a mismatch between the average phase of the signals captured by the SCM estimates and the drifting phase of the signals to which the beamformer is applied.

The remainder of this chapter is organized as follows: First, the advantages of distributed recording devices are motivated using meeting recognition as example application for MVDR beamforming in Sec. 5.1. Next, the effects of STOs on MVDR beamforming are discussed in Sec. 5.2. Finally, the effects of SROs on SCM estimation and MVDR beamforming are addressed in Sec. 5.3.

5.1 Compact microphone array vs. distributed recording devices for meeting recognition

In the following, beamforming using a compact microphone array and beamforming using distributed recording devices are compared.

5.1.1 System overview

The system used for the comparison of beamforming using a compact microphone array and beamforming using distributed recording devices is based on the reference system of the

LibriWASN dataset [OC3] which exhibits the structure of the beamforming system shown in Fig. 2.2. First, the SROs are compensated for if distributed recording devices are employed. Afterwards, masks for the target speaker and the interference are computed, from which the target and the interference SCMs are estimated. Based on the SCM estimates the beamformer coefficients are derived which are subsequently applied to the microphone signals to extract the signal of the target speaker. Finally, an ASR system is applied to transcribe the extracted signals.

To compensate for the SROs, the dynamic weighted average coherence drift (DWACD) method [OC1], which is presented in detail in Sec. 7.2.2, is utilized to estimate the SROs w.r.t. first channel. The SROs are compensated for by using the STFT-based resampling method from [79].

The masks used to estimate the SCMs, as described in Sec. 2.3, are estimated using a complex angular central Gaussian mixture model (cACGMM) [80] which is applied to the signals of the compact microphone array. These masks are used for the case of beamforming using a compact microphone array and the case of beamforming using distributed devices. Therefore, differences in the performance can directly be attributed to the usage of a compact or a distributed recording setup for beamforming. Since mask estimation is not the focus of this work, further investigation of mask estimation from the microphone signals of the distributed devices is waived.

The cACGMM for mask estimation corresponds to a spatial mixture model. This means that the cACGMM utilizes spatial information to estimate the likelihood that the individual sources are active in a time frequency bin. Based on this, a mask for each of the speakers and an additional mask for noise are obtained. The cACGMM is used in the same way as in the reference system of the LibriWASN dataset in [OC3]. As mask estimation is beyond the scope of this work, it will not be discussed in a more detailed way here. For more details, refer to [OC3].

In order to extract the q-th speaker's signal, the MVDR beamformer in the formulation specified in (2.40) is used. For this, the microphone signals are divided into segments of continuous activity of the target speaker which are determined from the cACGMM in the same way as in [64]. A visualization of the segmentation is shown in Fig. 5.1, where $\mathcal{L}_s$ denotes the set of time frames belonging to the s-th segment. The beamformer is applied to each resulting segment in the same manner as it is applied to a single block in block-wise beamforming. To this end, the target SCM is estimated for the s-th segment via (2.41) employing all time frames in $\mathcal{L}_s$ and the target mask estimate

$$\gamma_{\mathrm{tar},s}(\ell,k) = \begin{cases} \gamma_q(\ell,k), & \ell \in \mathcal{L}_s \\ 0, & \text{else} \end{cases}. \tag{5.1}$$

Here, $\gamma_q(\ell,k)$ is the mask belonging to the q-th speaker.

As proposed in [OC2], [OC14], each segment is further split into subsegments to account for changes in the activity of the interfering sources within the segments of continuous activity of the target speaker. In this way, the beamformer is able to better focus on the suppression of the currently active interfering sources. The boundaries of the subsegments are given by the change points of the interfering sources' activities, as shown in Fig. 5.1. Hence, the set

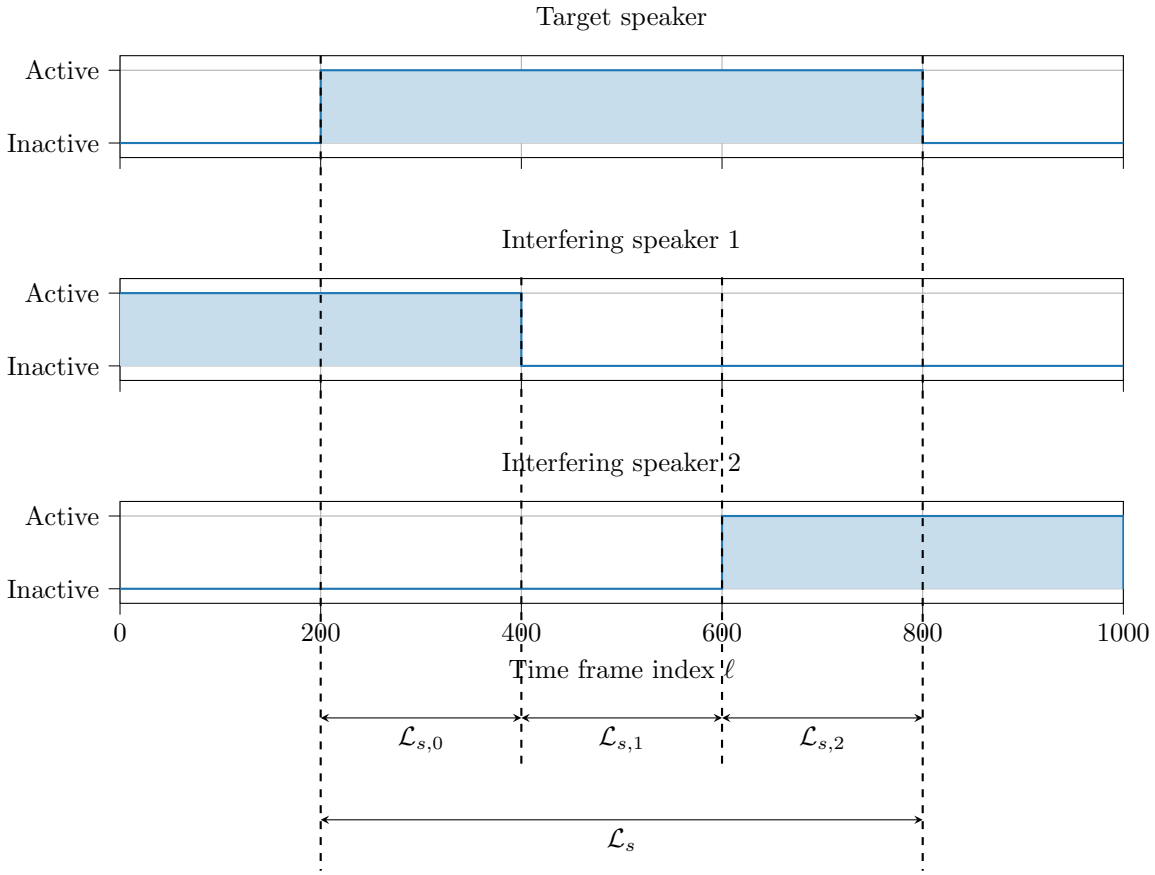


Figure 5.1: Visualization of the segments used for source extraction via beamforming in meeting recognition

of time frames $\mathcal{L}_s$ is divided into subsets $\mathcal{L}_{s,s'}$, with s' corresponding to the index of the subsegment and $\mathcal{L}_{s,s'}$ being the set of time frames belonging to the s'-th subsegment. For each subsegment the interference SCM is estimated via (2.42) using the interference mask estimate

$$\gamma_{\text{int},s,s'}(\ell,k) = \begin{cases} 1-\gamma_q(\ell,k), & \ell \in \mathcal{L}_{s,s'} \\ 0, & \text{else} \end{cases} . \tag{5.2}$$

The beamformer coefficients are calculated via (2.40) using the SCM estimates which result from $\gamma_{\text{tar},s}(\ell,k)$ as target mask and $\gamma_{\text{int},s,s'}(\ell,k)$ as interference mask for each subsegment. Rather than choosing an arbitrary reference channel for beamforming, the reference channel is chosen such that the expected signal-to-distortion ratio (SDR) is maximized [81]. In order to avoid changes in the MVDR beamformer's response to signals originating from the target speaker which might harm the ASR performance, the same reference channel is used for the complete segment of continuous activity of the target speaker. For this purpose, the reference channel is chosen such that the expected SDR of the subsegment, which exhibits the lowest expected SDR, is maximized [OC14]. The resulting beamformer coefficients are then used to extract the target speaker's signal for all time frames $\mathcal{L}_{s,s'}$ belonging to the s'-th subsegment as in block-wise beamforming which was introduced in Sec. 2.3.

Finally, the extracted signals are transcribed using the ASR system also used for the reference system of the LibriWASN dataset. The used ASR system [48] is a pretrained model from the ESPnet framework [41] that is based on a transformer architecture and was trained on LibriSpeech [35]. The ASR system achieves a word error rate (WER) of 2.7 % on the clean test set of LibriSpeech.

5.1.2 Experimental setup

To compare beamforming with a compact and distributed microphone devices, the LibriWASN dataset is used. The LibriWASN dataset is a re-recording of the widely used LibriCSS dataset [82] which consists of recordings of simulated meetings that were played back by loudspeakers and recorded by a compact microphone array. These simulated meetings exhibit different overlap ratios, namely no overlap with long silence during speaker changes (0L), no overlap with short silence during speaker changes (0S) as well as 10 % (OV10), 20 % (OV20), 30 % (OV30) and 40 % (OV40) overlap. In the LibriWASN dataset the same simulated meetings were played back and were captured by multiple distributed, asynchronous devices. In total, recordings from four compact microphone arrays and five smartphones with one channel each are available. The SROs between the signals of different devices range from -23 ppm to 110 ppm.

The LibriWASN dataset was recorded in two different acoustic environments resulting in the subsets LibriWASN^{200} and LibriWASN^{800}. Each of these subsets has its own arrangement of loudspeakers and recording devices. While LibriWASN^{200} was recorded in an acoustic laboratory with a sound decay time of $T_{60}\approx$200 ms, LibriWASN^{800} was recorded in a computer laboratory with a sound decay time of $T_{60}\approx$800 ms and several computers running in the background. In this work, either the compact microphone array *asnupb4* with six microphones or a distributed recording setup are used. The distributed recording setup consists of one channel from the devices *asnupb4*, *asnupb2* and *asnupb7* each as well as the channels from the smartphones *Pixel6a*, *Pixel6b* and *Pixel7*. Thus, the distributed recording setup has the same number of microphones as the compact microphone array which enables a fair comparison.

The source extraction system via beamforming uses the same parametrization as in [OC3]. Furthermore, the DWACD method for SRO estimation mainly utilizes the same parameters as in [OC1]. However, α was increased to 0.99 to mitigate the effect of frequent speaker changes which can happen in parts of a meeting. The role of speaker changes will be discussed in a more detailed way in Sec. 7.2.2.

As performance metric, the concatenated minimum-permutation word error rate (cpWER) [83] is utilized, which corresponds to a speaker-attributed version of the WER. For the calculation of this metric, the reference utterances and the hypothesis segments are concatenated per speaker. Afterwards, the conventional WER is calculated for all permutations of reference and hypothesis. The minimum value of the resulting set of WERs corresponds to the cpWER.

5.1.3 Results

Table 5.1: Comparison of MVDR beamforming using a compact microphone array and MVDR beamforming using distributed recording devices on the LibriWASN dataset

		cpWER / %						
Dataset	Microphone setup	0L	0S	OV10	OV20	OV30	OV40	Avg.
LibriWASN^{200}	Compact	3.20	3.00	4.34	4.39	3.97	3.16	3.70
LibriWASN^{200}	Distributed	3.08	2.81	4.25	4.09	3.38	2.95	3.43
LibriWASN^{800}	Compact	3.91	3.59	4.34	5.58	5.77	5.30	4.85
LibriWASN^{800}	Distributed	3.70	3.33	4.01	5.23	4.87	4.43	4.33

A comparison of the performance which is achieved for MVDR beamforming using a compact microphone array and MVDR beamforming using distributed recording devices is shown in Tab. 5.1 . For both subsets of the LibriWASN dataset, beamforming using a set of distributed recording devices consistently outperforms beamforming using a compact microphone array. While the advantages of a distributed recording setup for beamforming are small for the LibriWASN^{200} subset, which is characterized by a high signal-to-noise ratio (SNR) and low reverberation, they are more prominent under the more difficult acoustic conditions, i.e., a lower SNR and higher reverberation, of the LibriWASN^{800} dataset. This is particularly evident for the cases with overlapping speech for which the gain in performance due to the distributed recording setup compared to the compact microphone array is larger than for the case without overlap.

The improvements w.r.t. the results presented in [OC3] can be mainly explained by the division of the segments of continuous activity of the target speaker into subsegments at the change points of the activity of the interfering sources. This matches the theoretical investigations in the previous chapter where it is shown that the optimal suppression of an interfering source is achieved if the interference SCM only represents the currently active interfering source. This was discussed at hand of the leakage of the target speaker's signals into the interference-SCM estimate in Sec. 4.4.

Although beamforming using distributed recording devices introduces new challenges, like the need for synchronization algorithms, it promises an improved beamforming performance. Note that the gap can become even larger when also pre-processing stages that utilize spatial information, such as mask-estimation, take advantage of the generally larger time differences of arrival (TDOAs) coming with the distributed recording setup. This can be seen for example in [OC4]. However, this aspect will not be discussed further in this work.

5.2 Impact of STOs on MVDR beamforming

Although STOs and SROs typically are both present in signals being captured by multiple distributed devices, the effect of both are separately considered in the following. While there are multiple works considering the effect of SROs on the performance of a beamformer, the effects of STOs is less investigated in literature. The same holds for the compensation for STOs which is typically done by a coarse alignment of the signals based on a long-term correlation estimate as it will be discussed in Sec. 7.1.1. However, in this way, the STOs are generally only minimized rather than completely compensated for.

In [1] the effect of STOs on low-rank multichannel Wiener filters, with the MVDR beamformer corresponding to a special case of this class of beamformers, is examined. It is shown in [1] that the performance of an MVDR beamformer degrades significantly even for STOs with a value in the subsample range when the steering vector is estimated from external information. This was demonstrated for steering vectors which were derived from the ground-truth RIRs. Furthermore, it is proven in [1] that an MVDR beamformer that uses a GEVD-based steering vector is not affected by STOs. This can be attributed to the fact that the SCMs which are used for the estimation of the GEVD-based steering vector and the interference-SCM estimates are directly estimated from the signals and, therefore, take into account the STOs. In the following, the results from [1] are transferred to the mask-based Souden-MVDR beamformer and the shortcomings of the STO model used in [1] for large STOs are discussed. The effects of STOs on the Souden-MVDR beamformer are analyzed based on the STO model in (2.14) which also served as basis for the analysis in [1]. Note that there is a large similarity between these derivations and the analysis of the effects of STOs on an MVDR beamformer with GEVD-based steering vector in [1].

It is assumed that the microphone signals show an STO τ_m^{STO}, where, w.l.o.g., the STO of channel 0 is set to zero, i.e., $\tau_0^{\mathrm{STO}}{=}0\,$samples. With (2.14), the relationship between the target-SCM estimate in the presence of STOs $\widehat{\mathbf{R}}_{\mathrm{tar}}^{\mathrm{STO}}(k)$ and the target SCM $\widehat{\mathbf{R}}_{\mathrm{tar}}(k)$, which is estimated from synchronously sampled signals, is given by

$$\begin{aligned}
\widehat{\mathbf{R}}_{\mathrm{tar}}^{\mathrm{STO}}(k) &= \frac{1}{\sum_{\ell=0}^{L-1} \gamma_{\mathrm{tar}}(\ell,k)} \cdot \sum_{\ell=0}^{L-1} \gamma_{\mathrm{tar}}(\ell,k) \cdot \mathbf{y}_{\mathrm{STO}}(\ell,k) \cdot \mathbf{y}_{\mathrm{STO}}^{\mathrm{H}}(\ell,k) \\
&= \frac{1}{\sum_{\ell=0}^{L-1} \gamma_{\mathrm{tar}}(\ell,k)} \cdot \sum_{\ell=0}^{L-1} \gamma_{\mathrm{tar}}(\ell,k) \cdot \mathbf{S}(k) \cdot \mathbf{y}(\ell,k) \cdot \mathbf{y}^{\mathrm{H}}(\ell,k) \cdot \mathbf{S}^{\mathrm{H}}(k) \\
&= \mathbf{S}(k) \cdot \widehat{\mathbf{R}}_{\mathrm{tar}}(k) \cdot \mathbf{S}^{\mathrm{H}}(k)\,. \qquad (5.3)
\end{aligned}$$

Here, $\mathbf{S}(k)$ is a diagonal matrix which consists of the phase terms modeling the STOs as specified in (2.15). Analogously, it follows for the interference-SCM estimate that

$$\begin{aligned}
\widehat{\mathbf{R}}_{\mathrm{int}}^{\mathrm{STO}}(k) &= \frac{1}{\sum_{\ell=0}^{L-1} \gamma_{\mathrm{tar}}(\ell,k)} \cdot \sum_{\ell=0}^{L-1} \gamma_{\mathrm{tar}}(\ell,k) \cdot \mathbf{y}_{\mathrm{STO}}(\ell,k) \cdot \mathbf{y}_{\mathrm{STO}}^{\mathrm{H}}(\ell,k) \\
&= \mathbf{S}(k) \cdot \widehat{\mathbf{R}}_{\mathrm{int}}(k) \cdot \mathbf{S}^{\mathrm{H}}(k)\,. \qquad (5.4)
\end{aligned}$$

Inserting $\widehat{\mathbf{R}}_{\text{tar}}^{\text{STO}}(k)$ and $\widehat{\mathbf{R}}_{\text{int}}^{\text{STO}}(k)$ into (2.40), leads to

$$\begin{aligned}\mathbf{w}_{\text{STO}}(k) &= \frac{\left(\widehat{\mathbf{R}}_{\text{int}}^{\text{STO}}(k)\right)^{-1} \cdot \widehat{\mathbf{R}}_{\text{tar}}^{\text{STO}}(k)}{\text{tr}\left\{\left(\widehat{\mathbf{R}}_{\text{int}}^{\text{STO}}(k)\right)^{-1} \cdot \widehat{\mathbf{R}}_{\text{tar}}^{\text{STO}}(k)\right\}} \cdot \mathbf{u} \\ &= \frac{\left(\mathbf{S}(k) \cdot \widehat{\mathbf{R}}_{\text{int}}(k) \cdot \mathbf{S}^{\text{H}}(k)\right)^{-1} \cdot \mathbf{S}(k) \cdot \widehat{\mathbf{R}}_{\text{tar}}(k) \cdot \mathbf{S}^{\text{H}}(k)}{\text{tr}\left\{\left(\mathbf{S}(k) \cdot \widehat{\mathbf{R}}_{\text{int}}(k) \cdot \mathbf{S}^{\text{H}}(k)\right)^{-1} \cdot \mathbf{S}(k) \cdot \widehat{\mathbf{R}}_{\text{tar}}(k) \cdot \mathbf{S}^{\text{H}}(k)\right\}} \cdot \mathbf{u}\end{aligned} \tag{5.5}$$

for the beamformer coefficients $\mathbf{w}_{\text{STO}}(k)$ estimated in the presence of STOs. Utilizing $\mathbf{S}^{\text{H}}(k) = \mathbf{S}^{-1}(k)$, which arises from the fact that $\mathbf{S}$ corresponds to a diagonal matrix of phase terms, and assuming that the first channel is utilized as reference channel for beamforming, i.e, $\mathbf{u} = \begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix}^{\text{T}}$ and, therefore, $\mathbf{S}^{\text{H}}(k) \cdot \mathbf{u} = \mathbf{u}$, leads to

$$\mathbf{w}_{\text{STO}}(k) = \frac{\mathbf{S}(k) \cdot \widehat{\mathbf{R}}_{\text{int}}^{-1}(k) \cdot \widehat{\mathbf{R}}_{\text{tar}}(k)}{\text{tr}\left\{\widehat{\mathbf{R}}_{\text{int}}^{-1}(k) \cdot \widehat{\mathbf{R}}_{\text{tar}}(k)\right\}} \cdot \mathbf{u} = \mathbf{S}(k) \cdot \mathbf{w}(k) \tag{5.6}$$

for the relationship between the beamformer coefficients $\mathbf{w}_{\text{STO}}(k)$ and the beamformer coefficients $\mathbf{w}(k)$ being estimated from synchronously sampled signals. Thus, the beamformer output in the presence of STOs

$$\widehat{\mathbf{x}}_{\text{tar}}^{\text{STO}}(\ell, k) = \mathbf{w}_{\text{STO}}^{\text{H}}(k) \cdot \mathbf{y}_{\text{STO}}(\ell, k) = \mathbf{w}^{\text{H}}(k) \cdot \mathbf{S}^{\text{H}}(k) \cdot \mathbf{S}(k) \cdot \mathbf{y}(\ell, k) = \mathbf{w}^{\text{H}}(k) \cdot \mathbf{y}(\ell, k) \tag{5.7}$$

is equal to the beamformer output for the case without STOs. Note that utilizing another reference channel with a non-zero STO results in an additional multiplication with a scalar phase term in (5.6) which models the effect of the STO of the reference channel. Thus, the choice of a reference channel with non-zero STO only affects the absolute phase of the extracted signal. Consequently, STOs do not affect the performance of the Souden-MVDR as long as the STO model in (2.13) is suitable to model the STOs.

However, the accuracy of the STO model in (2.13) decreases with growing STO as already discussed in Sec. 2.2. For example, the reduction of the correlation between the microphone signals with growing STO is not represented by a simple multiplication with a phase term. This reduced correlation between the microphone signals can negatively influence the beamforming performance. Thus, the STOs have to be much smaller than the size of the analysis window of the STFT. For typical sizes of the STFT analysis window used for beamforming this can be already achieved by a coarse compensation for STOs which will be described in Sec. 7.1.1.

As discussed in Sec. 2.3, the target-SCM estimate can be approximated as a rank-1 matrix based on an eigenvalue decomposition (EVD) or a GEVD. First, the EVD-based rank-1 approximation of the target SCM estimated in the presence of STOs considered for which

$$\widehat{\mathbf{R}}_{\text{tar,EV}}^{\text{STO}}(k) = \widehat{\mathbf{d}}_{\text{EV}}^{\text{STO}}(k) \cdot \left(\widehat{\mathbf{d}}_{\text{EV}}^{\text{STO}}(k)\right)^{\text{H}} \tag{5.8}$$

holds, with the principal eigenvector being calculated via

$$\widehat{\mathbf{d}}_{\mathrm{EV}}^{\mathrm{STO}}(k) = \mathcal{P}\Big(\widehat{\mathbf{R}}_{\mathrm{tar}}^{\mathrm{STO}}(k)\Big) = \mathcal{P}\Big(\mathbf{S}(k) \cdot \widehat{\mathbf{R}}_{\mathrm{tar}}(k) \cdot \mathbf{S}^{\mathrm{H}}(k)\Big)\,. \tag{5.9}$$

In order to find the relationship to the case without STOs being present, the corresponding eigenvalue problem is considered:

$$\widehat{\mathbf{R}}_{\mathrm{tar}}^{\mathrm{STO}}(k) \cdot \mathbf{o}_{\mathrm{STO}} = \lambda \cdot \mathbf{o}_{\mathrm{STO}} \tag{5.10}$$

$$\Leftrightarrow \mathbf{S}(k) \cdot \widehat{\mathbf{R}}_{\mathrm{tar}}(k) \cdot \mathbf{S}^{\mathrm{H}}(k) \cdot \mathbf{o}_{\mathrm{STO}} = \lambda \cdot \mathbf{o}_{\mathrm{STO}}. \tag{5.11}$$

Assuming $\mathbf{o}_{\mathrm{STO}}$=$\mathbf{S}(k) \cdot \mathbf{o}$, leads to

$$\mathbf{S}(k) \cdot \widehat{\mathbf{R}}_{\mathrm{tar}}(k) \cdot \mathbf{S}^{\mathrm{H}}(k) \cdot \mathbf{S}(k) \cdot \mathbf{o} = \lambda \cdot \mathbf{S}(k) \cdot \mathbf{o} \tag{5.12}$$

$$\Leftrightarrow \widehat{\mathbf{R}}_{\mathrm{tar}}(k) \cdot \mathbf{o} = \lambda \cdot \mathbf{o}, \tag{5.13}$$

where $\mathbf{S}^{\mathrm{H}}(k)$ =$\mathbf{S}^{-1}(k)$ is utilized. This is the eigenvalue problem for the case that STOs are not present. Hence, it follows that the relationship between the principal eigenvector of $\widehat{\mathbf{R}}_{\mathrm{tar}}^{\mathrm{STO}}(k)$ for the case of STOs being present to the principal eigenvector $\widehat{\mathbf{d}}_{\mathrm{EV}}(k)$ of $\widehat{\mathbf{R}}_{\mathrm{tar}}(k)$ is given by

$$\widehat{\mathbf{d}}_{\mathrm{EV}}^{\mathrm{STO}}(k) = \mathbf{S}(k) \cdot \widehat{\mathbf{d}}_{\mathrm{EV}}(k)\,. \tag{5.14}$$

Inserting (5.14) into (5.8), leads to

$$\begin{aligned}
\widehat{\mathbf{R}}_{\mathrm{tar,EV}}^{\mathrm{STO}}(k) &= \widehat{\mathbf{d}}_{\mathrm{EV}}^{\mathrm{STO}}(k) \cdot \Big(\widehat{\mathbf{d}}_{\mathrm{EV}}^{\mathrm{STO}}(k)\Big)^{\mathrm{H}} \\
&= \mathbf{S}(k) \cdot \widehat{\mathbf{d}}_{\mathrm{EV}}(k) \cdot \mathbf{S}^{\mathrm{H}}(k) \cdot \mathbf{S}(k) \cdot \widehat{\mathbf{d}}_{\mathrm{EV}}^{\mathrm{H}}(k) \cdot \mathbf{S}^{\mathrm{H}}(k) \\
&= \mathbf{S}(k) \cdot \widehat{\mathbf{R}}_{\mathrm{tar,EV}}(k) \cdot \mathbf{S}^{\mathrm{H}}(k)
\end{aligned} \tag{5.15}$$

for the relationship of the EVD-based rank-1 approximation of the target-SCM estimate $\widehat{\mathbf{R}}_{\mathrm{tar,EV}}^{\mathrm{STO}}(k)$ and the corresponding quantity $\widehat{\mathbf{R}}_{\mathrm{tar,EV}}(k)$ in absence of STOs. Hence, it can be shown in the same way as for the Souden-MVDR beamformer with a full-rank target-SCM estimate that also the Souden-MVDR beamformer with EVD-based rank-1 estimate of the target SCM is unaffected by STOs if their value is not too large.

By considering the GEVD-based steering vector $\widehat{\mathbf{d}}_{\mathrm{GEV}}^{\mathrm{STO}}(k)$ which is calculated from the SCMs $\widehat{\mathbf{R}}_{\mathrm{tar}}^{\mathrm{STO}}(k)$ and $\widehat{\mathbf{R}}_{\mathrm{int}}^{\mathrm{STO}}(k)$ via (2.45), it can be shown that

$$\begin{aligned}
\widehat{\mathbf{d}}_{\mathrm{GEV}}^{\mathrm{STO}}(k) &= \widehat{\mathbf{R}}_{\mathrm{int}}^{\mathrm{STO}}(k) \cdot \mathcal{P}\Bigg(\Big(\widehat{\mathbf{R}}_{\mathrm{int}}^{\mathrm{STO}}(k)\Big)^{-1} \cdot \widehat{\mathbf{R}}_{\mathrm{tar}}^{\mathrm{STO}}(k)\Bigg) \\
&= \mathbf{S}(k) \cdot \widehat{\mathbf{R}}_{\mathrm{int}}(k) \cdot \mathbf{S}^{\mathrm{H}}(k) \cdot \mathcal{P}\Bigg(\Big(\mathbf{S}(k) \cdot \widehat{\mathbf{R}}_{\mathrm{int}}(k) \cdot \mathbf{S}^{\mathrm{H}}(k)\Big)^{-1} \cdot \mathbf{S}(k) \cdot \widehat{\mathbf{R}}_{\mathrm{tar}}(k) \cdot \mathbf{S}^{\mathrm{H}}(k)\Bigg) \\
&= \mathbf{S}(k) \cdot \widehat{\mathbf{R}}_{\mathrm{int}}(k) \cdot \mathbf{S}^{\mathrm{H}}(k) \cdot \mathcal{P}\Big(\mathbf{S}(k) \cdot \widehat{\mathbf{R}}_{\mathrm{int}}^{-1}(k) \cdot \widehat{\mathbf{R}}_{\mathrm{tar}}(k) \cdot \mathbf{S}^{\mathrm{H}}(k)\Big) \\
&= \mathbf{S}(k) \cdot \widehat{\mathbf{R}}_{\mathrm{int}}(k) \cdot \mathbf{S}^{\mathrm{H}}(k) \cdot \mathbf{S}(k) \cdot \mathcal{P}\Big(\widehat{\mathbf{R}}_{\mathrm{int}}^{-1}(k) \cdot \widehat{\mathbf{R}}_{\mathrm{tar}}(k)\Big) \\
&= \mathbf{S}(k) \cdot \widehat{\mathbf{R}}_{\mathrm{int}}(k) \cdot \mathcal{P}\Big(\widehat{\mathbf{R}}_{\mathrm{int}}^{-1}(k) \cdot \widehat{\mathbf{R}}_{\mathrm{tar}}(k)\Big) \\
&= \mathbf{S}(k) \cdot \widehat{\mathbf{d}}_{\mathrm{GEV}}(k)\,.
\end{aligned} \tag{5.16}$$

Thus, the GEVD-based steering vector $\widehat{\mathbf{d}}_{\mathrm{GEV}}^{\mathrm{STO}}(k)$ with STOs being present relates to the GEVD-based steering vector $\widehat{\mathbf{d}}_{\mathrm{GEV}}(k)$ for the case without STOs being present in the same way as the corresponding EVD-based estimates in (5.14). Consequently, a Souden-MVDR beamformer with a GEVD-based rank-1 approximation of the target-SCM estimate behaves like a Souden-MVDR beamformer with an EVD-based rank-1 approximation of the target-SCM estimate. Hence, a Souden-MVDR beamformer with a GEVD-based rank-1 approximation of the target-SCM estimate is not affected by STOs as long as the STOs are sufficiently small.

This is consistent with the results from [1]. For a rank-1 target-SCM estimate the Souden-MVDR coincides, up to a multiplication with a complex scalar, with an MVDR beamformer in its classical steering-vector-based formulation as shown in [28]. The classical MVDR beamformer corresponds to a special case of the low-rank multichannel Wiener filter which was considered in [1]. As demonstrated in [1], this type of beamformer is unaffected by STOs if a GEVD-based steering vector is employed which is estimated from the signals to which the beamformer is applied and the STOs are such small that they can be modeled via (2.13).

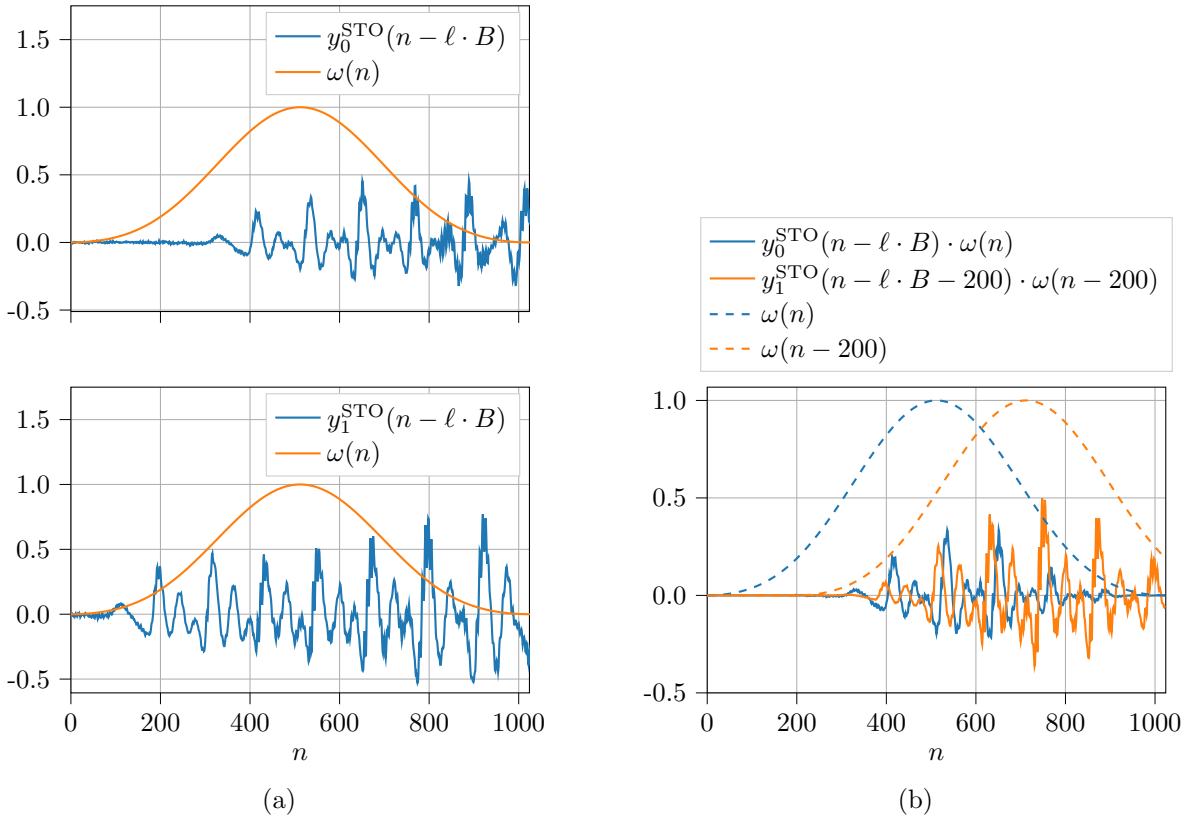


Figure 5.2: Visualization of the reduction of the correlation between two microphone signals $y_0^{\mathrm{STO}}(n)$ and $y_1^{\mathrm{STO}}(n)$ due to large STOs. (a) visualizes the extraction of the ℓ-th time frame of the STFT for an STO of 200 samples between the signals and (b) visualizes the reduced overlap of the analysis windows $\omega(n)$ for the extracted time frames in (a) when aligning the extracted signals

The reasons for the reduced correlation between two channels due to large STOs are illustrated in Fig. 5.2 by considering the extraction of a single time frame of the STFTs of the two microphone signals $y_0^{\mathrm{STO}}(n)$ and $y_1^{\mathrm{STO}}(n)$ that exhibit an STO between them. It can be seen in Fig. 5.2(a) that a large part of the underlying source signal is only present in one of the two time frames. Additionally, Fig. 5.2(b) shows that this effect is even intensified by the analysis window of the STFT $\omega(n)$ since the analysis window for both signals also do not overlap completely when compensating for the STO.

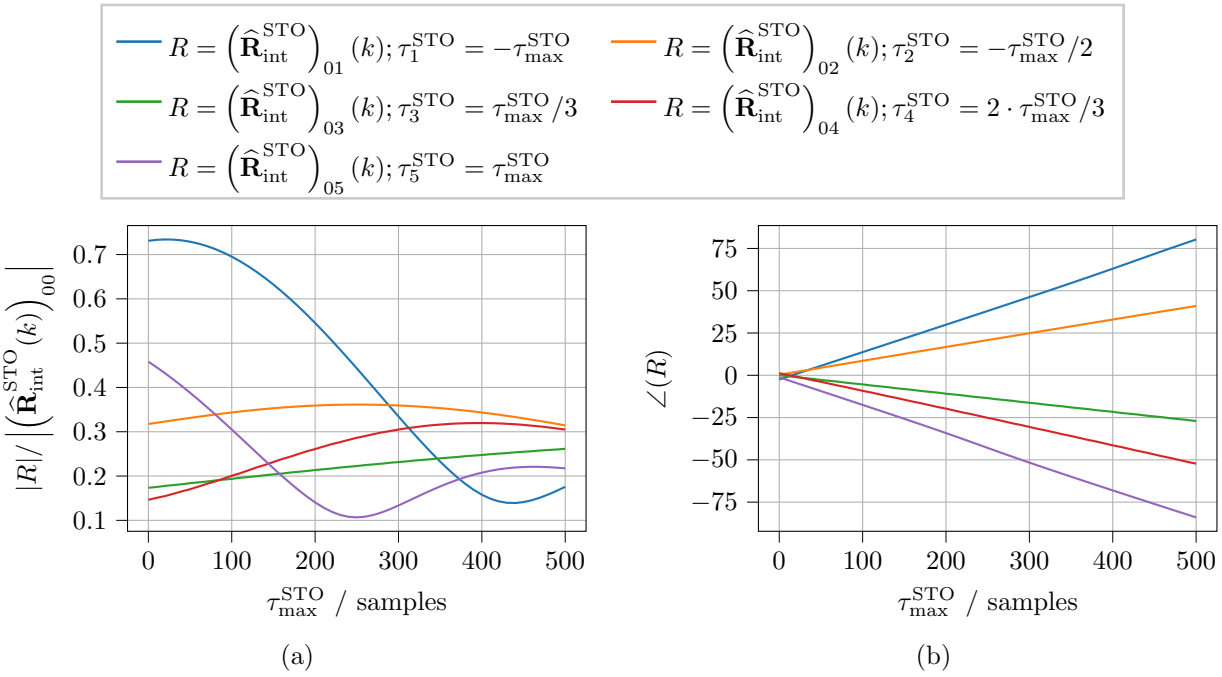


Figure 5.3: Example of the STO-induced distortion of the amplitude $|(\widehat{\mathbf{R}}_{\mathrm{int}}^{\mathrm{STO}})_{ij}(k)|$ (a) and phase $\angle((\widehat{\mathbf{R}}_{\mathrm{int}}^{\mathrm{STO}})_{ij}(k))$ (b) of the elements of the SCM estimates for k=26. The STO of the individual microphone signals is specified w.r.t. the maximum STO $\tau_{\max}^{\mathrm{STO}}$ which is varied to investigate the effect of the strength of the STOs. The amplitude of the considered elements is normalized by the variance of the 0-th channel, i.e., $(\widehat{\mathbf{R}}_{\mathrm{int}}^{\mathrm{STO}})_{00}(k)$, to isolate the STO-induced amplitude distortion from other changes of the amplitude.

Figure 5.3 shows the effects of STOs on the SCM estimates. STOs distort the phase $\angle(\cdot)$ of the elements of the SCM estimates which can be observed as linear phase drift with growing STO in Fig. 5.3(b). This effect already was discussed before in literature, e.g., in [1]. In addition to that, STOs distort the amplitude of the elements of the SCM estimates, as shown in Fig. 5.3(a). This means that the performance of the beamformer is only unaffected by STOs if their value is small. Although it is expected that the correlation between the signals and, therefore, the absolute value of the elements of the SCM estimates decreases with increasing STOs, the absolute value of the elements of the SCM estimates might increase for individual channel combinations. This phenomenon may be attributed to fact that the signal components corresponding to the early reflections of sound are time-aligned as a result of the STO.

5.3 Impact of SROs on MVDR beamforming

The effects of SROs on beamforming were already investigated in literature before, e.g., in [2]–[6]. All these publications report a deterioration of the performance of a beamformer due to SROs if the SROs are not compensated for. However, most of these works purely consider the effects of SROs in an experimental way. Only [4], [5] obtained deeper theoretical insights into reasons for the SRO-induced degradation of the beamforming performance.

In [4], the degradation of the beamforming performance due to SROs was attributed to a distortion of the beampattern. Moreover, it was shown in [5] that SROs cause a distortion of the amplitude of the SCM estimates' elements. This amplitude distortion was modeled as a function of the SRO and the sample size, i.e., the number of time frames used for SCM estimation. However, a connection to the influence on the performance of the beamformer was not established. Additionally, a noise source with time-constant power was considered as interfering source in [5]. Thus, the results cannot be directly transferred to the problem at hand with an interfering speech signal which has a time-varying power.

In the following, the model of the SRO-induced distortion of the amplitude of the SCM estimates' elements is transferred to the case of speech as interfering signal. Here, it is shown that SROs do not only result in a distortion of the SCM estimates' amplitude but also introduce a phase mismatch between the SCM estimates and the signal to which the beamformer is applied. It is experimentally demonstrated that this phase mismatch may lead to a distortion of the target signal at the beamformer output and to a degradation of the suppression of the interfering signals.

For sake of a simpler notation, time-invariant SROs are considered in the following. However, the findings can also be transferred to the case of time-varying SROs. Without loss of generality, the SRO of channel 0 is set to zero, i.e., ε_0=0 ppm. From (2.21) and (2.41),

$$\begin{aligned}
\widehat{\mathbf{R}}_{\text{tar}}^{\text{SRO}}(k) &= \frac{1}{\sum_{\ell=0}^{L-1} \gamma_{\text{tar}}(\ell,k)} \cdot \sum_{\ell=0}^{L-1} \gamma_{\text{tar}}(\ell,k) \cdot \mathbf{y}_{\text{SRO}}(\ell,k) \cdot \mathbf{y}_{\text{SRO}}^{\text{H}}(\ell,k) \\
&= \frac{1}{\sum_{\ell=0}^{L-1} \gamma_{\text{tar}}(\ell,k)} \cdot \sum_{\ell=0}^{L-1} \gamma_{\text{tar}}(\ell,k) \cdot (\mathbf{E}(\ell,k) \cdot \mathbf{y}(\ell,k)) \cdot (\mathbf{E}(\ell,k) \cdot \mathbf{y}(\ell,k))^{\text{H}} \\
&= \frac{1}{\sum_{\ell=0}^{L-1} \gamma_{\text{tar}}(\ell,k)} \cdot \sum_{\ell=0}^{L-1} \gamma_{\text{tar}}(\ell,k) \cdot \mathbf{E}(\ell,k) \cdot \mathbf{y}(\ell,k) \cdot \mathbf{y}^{\text{H}}(\ell,k) \cdot \mathbf{E}^{\text{H}}(\ell,k)
\end{aligned} \tag{5.17}$$

follows for the target-SCM estimate in the presence of SROs. Here, $\mathbf{E}(\ell,k)$ is a diagonal matrix which consists of the phase terms modeling the SRO-induced time shifts between the signals as specified in (2.24). Analogously, the interference SCM estimated in the presence of SROs is given by

$$\begin{aligned}\widehat{\mathbf{R}}_{\text{int}}^{\text{SRO}}(k) &= \frac{1}{\sum_{\ell=0}^{L-1}\gamma_{\text{int}}(\ell,k)} \cdot \sum_{\ell=0}^{L-1}\gamma_{\text{int}}(\ell,k)\cdot\mathbf{y}_{\text{SRO}}(\ell,k)\cdot\mathbf{y}_{\text{SRO}}^{\text{H}}(\ell,k) \\ &= \frac{1}{\sum_{\ell=0}^{L-1}\gamma_{\text{int}}(\ell,k)} \cdot \sum_{\ell=0}^{L-1}\gamma_{\text{int}}(\ell,k)\cdot\mathbf{E}(\ell,k)\cdot\mathbf{y}(\ell,k)\cdot\mathbf{y}^{\text{H}}(\ell,k)\cdot\mathbf{E}^{\text{H}}(\ell,k)\,. \end{aligned} \tag{5.18}$$

In the following, the insights into the effect of SROs on the amplitude of the SCM estimates which were obtained in [5] are recapitulated. In [5], the source images of the interference were modeled as samples from a zero-mean Gaussian. Furthermore, it was assumed that the interfering source can be observed without the target source being present. In this case, no masking was needed so that $\gamma_{\text{int}}(\ell,k)=1$ was assumed for the estimation of the interference SCM. Using the truncated geometric series, it was shown in [5] that the expected value of elements of the interference-SCM estimates $\widehat{\mathbf{R}}_{\text{int}}^{\text{SRO}}(k)$ estimated in the presence of SROs is given by

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{\mathbf{R}}_{\text{int}}^{\text{SRO}}(k)\right)_{ij}\right] &= (\mathbf{R}_{\text{int}}(k))_{ij}\cdot\frac{1}{L}\cdot\sum_{\ell=0}^{L-1}\exp\left(-\text{j}\cdot\frac{2\cdot\pi\cdot k}{N}\cdot\left(\frac{N}{2}+\ell\cdot B\right)\cdot\varepsilon_{ji}\right) \\ &= (\mathbf{R}_{\text{int}}(k))_{ij}\cdot\exp(-\text{j}\cdot\pi\cdot\varepsilon_{ji})\cdot\frac{1}{L}\cdot\frac{1-\exp\left(-\text{j}\cdot\frac{2\cdot\pi}{N}B\cdot\varepsilon_{ji}\cdot k\cdot L\right)}{1-\exp\left(-\text{j}\cdot\frac{2\cdot\pi}{N}\cdot B\cdot\varepsilon_{ji}\cdot k\right)}. \end{aligned} \tag{5.19}$$

The SRO-induced distortion follows a sinc-like function of the sample size L used for SCM estimation and the SRO ε_{ji}. Hence, the SRO can even drive the value of individual entries of the estimated SCMs to zero for specific values of L.

For the problem at hand, with mask-based SCM estimation and speech signals with time-varying power that are statistically modeled by the local Gaussian model (LGM),

$$\begin{aligned}\mathbb{E}\left[\widehat{\mathbf{R}}_{\text{int}}^{\text{SRO}}(k)\right] &= \mathbb{E}\left[\frac{1}{\sum_{\ell=0}^{L-1}\gamma_{\text{int}}(\ell,k)}\cdot\sum_{\ell=0}^{L-1}\gamma_{\text{int}}(\ell,k)\cdot\mathbf{y}_{\text{SRO}}(\ell,k)\cdot\mathbf{y}_{\text{SRO}}^{\text{H}}(\ell,k)\right] \\ &= \frac{1}{\sum_{\ell=0}^{L-1}\gamma_{\text{int}}(\ell,k)}\cdot\sum_{\ell=0}^{L-1}\gamma_{\text{int}}(\ell,k)\cdot\mathbf{E}(\ell,k)\cdot\mathbb{E}\left[\mathbf{y}(\ell,k)\cdot\mathbf{y}^{\text{H}}(\ell,k)\right]\cdot\mathbf{E}^{\text{H}}(\ell,k) \end{aligned} \tag{5.20}$$

follows for the expected value of the interference-SCM estimate in the presence of SROs. With the second-order moment of the microphone signals

$$\mathbb{E}\left[\mathbf{y}(\ell,k)\cdot\mathbf{y}^{\text{H}}(\ell,k)\right] = \sigma_0^2(\ell,k)\cdot\mathbf{R}_0(k)+\sigma_1^2(\ell,k)\cdot\mathbf{R}_1(k)+\sigma_\nu^2(k)\cdot\mathbf{R}_\nu(k)\,, \tag{5.21}$$

which results from (2.11), it follows that

$$\mathbb{E}\left[\widehat{\mathbf{R}}_{\text{int}}^{\text{SRO}}(k)\right] = \frac{1}{\sum_{\ell=0}^{L-1}\gamma_{\text{int}}(\ell,k)} \cdot \left(\sum_{\ell=0}^{L-1} \gamma_{\text{int}}(\ell,k) \cdot \sigma_0^2(\ell,k) \cdot \mathbf{E}(\ell,k) \cdot \mathbf{R}_0(k) \cdot \mathbf{E}^{\text{H}}(\ell,k) + \sum_{\ell=0}^{L-1} \gamma_{\text{int}}(\ell,k) \cdot \sigma_1^2(\ell,k) \cdot \mathbf{E}(\ell,k) \cdot \mathbf{R}_1(k) \cdot \mathbf{E}^{\text{H}}(\ell,k) + \sum_{\ell=0}^{L-1} \gamma_{\text{int}}(\ell,k) \cdot \sigma_\nu^2(k) \cdot \mathbf{E}(\ell,k) \cdot \mathbf{R}_\nu(k) \cdot \mathbf{E}^{\text{H}}(\ell,k) \right) \tag{5.22}$$

holds for the expected value of the interference SCM estimated in the presence of SROs. Thus, the expected value of the interference-SCM estimate in the presence of SROs corresponds to a weighted sum of the sources' ground-truth SCMs $\mathbf{R}_0(k)$, $\mathbf{R}_1(k)$ and $\mathbf{R}_\nu(k)$ where the elements of each ground-truth SCM is distorted by a weighted sum of phase terms modeled by $\mathbf{E}(\ell,k)$. This can be better seen when considering a single element of $\widehat{\mathbf{R}}_{\text{int}}^{\text{SRO}}(k)$ with

$$\begin{aligned} &\mathbb{E}\left[\left(\widehat{\mathbf{R}}_{\text{int}}^{\text{SRO}}(k)\right)_{ij}\right] \\ &= \frac{1}{\sum_{\ell=0}^{L-1}\gamma_{\text{int}}(\ell,k)} \cdot \left((\mathbf{R}_0(k))_{ij} \cdot \sum_{\ell=0}^{L-1} \gamma_{\text{int}}(\ell,k) \cdot \sigma_0^2(\ell,k) \cdot \exp\left(-\text{j}\cdot\frac{2\cdot\pi\cdot k}{N}\cdot\left(\frac{N}{2}+\ell\cdot B\right)\cdot\varepsilon_{ji}\right) \right. \\ &\quad + (\mathbf{R}_1(k))_{ij} \cdot \sum_{\ell=0}^{L-1} \gamma_{\text{int}}(\ell,k) \cdot \sigma_1^2(\ell,k) \cdot \exp\left(-\text{j}\cdot\frac{2\cdot\pi\cdot k}{N}\cdot\left(\frac{N}{2}+\ell\cdot B\right)\cdot\varepsilon_{ji}\right) \\ &\quad \left. + (\mathbf{R}_\nu(k))_{ij} \cdot \sum_{\ell=0}^{L-1} \gamma_{\text{int}}(\ell,k) \cdot \sigma_\nu^2(k) \cdot \exp\left(-\text{j}\cdot\frac{2\cdot\pi\cdot k}{N}\cdot\left(\frac{N}{2}+\ell\cdot B\right)\cdot\varepsilon_{ji}\right) \right). \end{aligned} \tag{5.23}$$

Similar to (5.19) the elements of the ground-truth SCMs are multiplied by a weighted sum of phase terms, which drift over time. It is expected that the multiplication with the weighted sum of phase terms reduces the amplitude of the SCM estimates' elements as the sample size increases. This can be be attributed to the fact that phase terms with a phase shift of approximately π and a similar weight mutually cancel. However, a simplification to a closed-form expression as in (5.19) is not possible due to the time-varying weights $\sigma_0^2(\ell,k)$, $\sigma_1^2(\ell,k)$ and $\sigma_\nu^2(\ell,k)$. In general, the SRO-induced distortion of the elements of the ground-truth SCMs as a function of the sample size used for SCM estimation will be more irregular than the sinc-like function in [5] due to the high dynamic range of the time-varying weights $\sigma_0^2(\ell,k)$ and $\sigma_1^2(\ell,k)$.

Figure 5.4 visualizes the SRO-induced distortion of the SCM estimates' amplitude as a function of the sample size L used for SCM estimation. Here, the amplitude of representative elements of the interference-SCM estimates, which correspond to the covariance between the 0-th channel and another channel, are considered. The amplitude of the considered elements is normalized by the variance of the 0-th channel, i.e., $(\widehat{\mathbf{R}}_{\text{int}}^{\text{SRO}})_{00}(k)$ or $\mathbb{E}[(\widehat{\mathbf{R}}_{\text{int}}^{\text{SRO}})_{00}(k)]$ depending on whether beamforming of deterministic speech mixtures or the statistical model

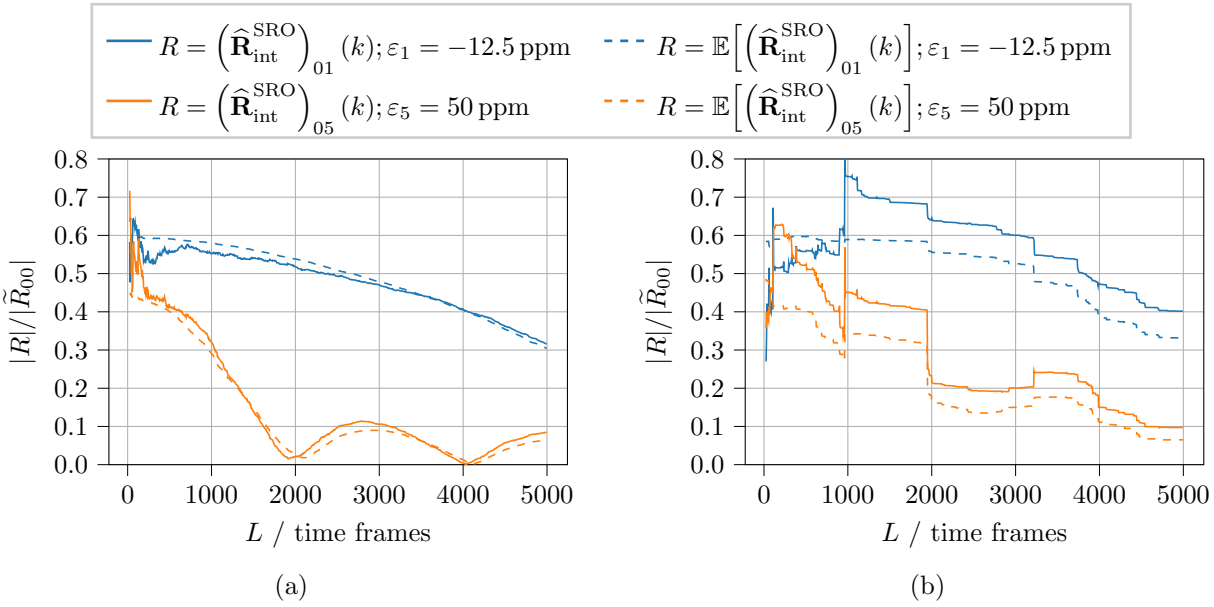


Figure 5.4: SRO-induced distortion of the amplitude of the elements of an interference-SCM estimate $|(\widehat{\mathbf{R}}_{\text{int}}^{\text{SRO}})_{ij}(k)|$ for k=38 as a function of the sample size L used for SCM estimation for the case of stationary noise as interference (a), as considered in [5], and the case of mask-based estimation from a noisy speech mixture (b). The SROs ε_i specified in the legend correspond to the relative SRO between channel 0 and the other channel corresponding to the element of the SCM. $(\widehat{\mathbf{R}}_{\text{int}}^{\text{SRO}})_{ij}(k)$ is estimated from a mixture of deterministic source images. $\mathbb{E}[(\widehat{\mathbf{R}}_{\text{int}}^{\text{SRO}})_{ij}(k)]$ is calculated via a Monte-Carlo simulation with 250 draws based on the statistical model of stationary noise or the LGM for speech sources. $\widetilde{R}_{00}$ corresponds to $(\widehat{\mathbf{R}}_{\text{int}}^{\text{SRO}})_{00}(k)$ or $\mathbb{E}[(\widehat{\mathbf{R}}_{\text{int}}^{\text{SRO}})_{00}(k)]$ depending on whether beamforming is applied to deterministic speech mixtures or the statistical model of beamforming is considered.

of beamforming is considered, to isolate the SRO-induced amplitude distortion from other changes of the amplitude.

It can be seen in Fig. 5.4(a) that the distortion of the amplitude of the SCM estimates follows a sinc-like function of the sample size L which is used for SCM estimation if the interfering source corresponds to stationary noise as in [5]. In contrast to this, the amplitude distortion for mask-based beamforming with speech sources is a irregular function of the sample size L and even shows discontinuities as depicted in Fig. 5.4(b). These discontinuities can be attributed to the existence of individual time frames with exceptionally high signal power which dominate the SCM estimate. However, the tendency of the amplitudes of the off-diagonal elements of the SCM estimates to drift towards a value of zero for a growing sample size L can be seen for both types of interfering sources. Furthermore, Fig. 5.4 indicates that the statistical perspective by modeling the samples used for SCM estimation as realization of a random variable and considering the expected value of the SCM estimates in (5.23) is suitable to reflect the behavior of estimating SCMs from deterministic signals. Thus, a statistical point of view on beamforming in the presence of SROs, which is introduced in the next chapter, is justified.

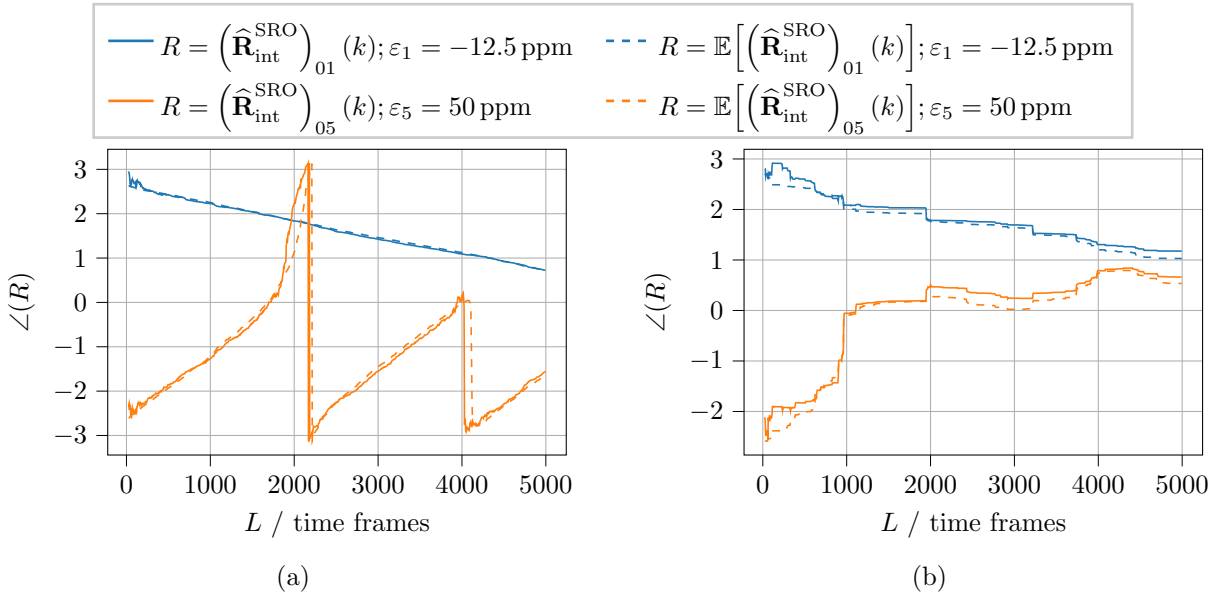


Figure 5.5: SRO-induced distortion of the phase of the elements of an interference-SCM estimate $\angle((\widehat{\mathbf{R}}_{\text{int}}^{\text{SRO}})_{ij}(k))$ for k=38 as a function of the sample size L used for SCM estimation for the case of stationary noise as interference (a), as considered in [5], and the case of mask-based estimation from a noisy speech mixture (b). The SROs ε_i specified in the legend correspond to the relative SRO between channel 0 and the other channel corresponding to the element of the SCM. $(\widehat{\mathbf{R}}_{\text{int}}^{\text{SRO}})_{ij}(k)$ is estimated from a mixture of deterministic source images. $\mathbb{E}[(\widehat{\mathbf{R}}_{\text{int}}^{\text{SRO}})_{ij}(k)]$ is calculated via a Monte-Carlo simulation with 250 draws based on the statistical model of stationary noise or the LGM for speech sources.

In addition to the effect of SROs on the amplitude of the SCM estimates, the distortion of their phase as a function of the sample size L used for SCM estimation is depicted in Fig. 5.5. It becomes apparent that SROs induce a drifting phase of the elements of the SCM estimates with growing sample size. This can be explained by the fact that the SCM estimates reflect the average phase of the signals within the SCM estimation interval. For an interfering source with time-constant power this drift mostly is an approximately linear function of the sample size except when the amplitude of the SCM estimates approaches a value of zero (see Fig. 5.4) as shown in Fig. 5.5(a). In contrast to this, the phase drift corresponds to an irregular function of the sample size with discontinuities for mask-based SCM estimation from speech signals as demonstrated in Fig. 5.5(b). As for the amplitude distortion, the statistical perspective via expected value of the SCM estimate is able to represent the case where the SCM is estimated from the corresponding deterministic signals.

Figure 5.6 compares the phase of expected value of elements of the interference-SCM estimate $\mathbb{E}\left[(\widehat{\mathbf{R}}_{\text{int}}^{\text{SRO}}(k))_{0i}\right]$ and the instantaneous phase of the corresponding elements of the expected value of interfering speaker's SCMs $\mathbb{E}\left[x_{1,0}(\ell,k)\cdot x_{1,i}^{*}(\ell,k)\right]$ for the time frames within the SCM estimation interval. It becomes obvious that SROs induce a linear phase drift within the SCM estimation interval. Due to the different SROs of the channels, the strength of the phase drift is different for each channel combination. For block-wise beamforming this causes a phase mismatch between the SCM estimates, which represent a weighted average of the

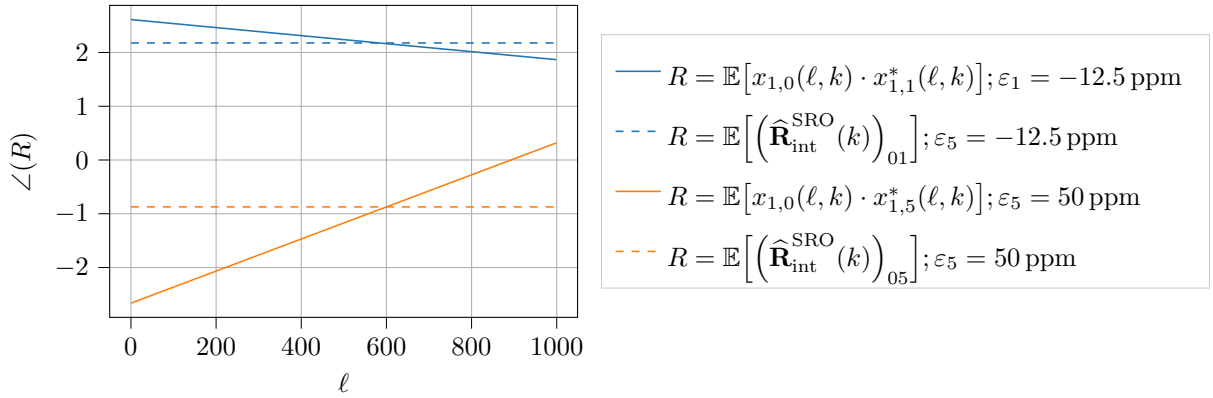


Figure 5.6: Comparison of the phase of the expected value of the elements of the interference SCM estimate $\mathbb{E}[\widehat{\mathbf{R}}_{\text{int}}^{\text{SRO}}(k)]$ and the value of the drifting phase of the interfering speaker's instantaneous second-order moment $\mathbb{E}[\mathbf{x}_1(\ell,k)\cdot\mathbf{x}_1^{\text{H}}(\ell,k)]$ within the SCM estimation interval for $k = 38$. The SROs ε_i specified in the legend correspond to the relative SRO between channel 0 and the other channel corresponding the element of the SCM. The expected values are estimated using a Monte-Carlo simulation with 250 draws.

signals' drifting phase, and the signals to which the beamformer is applied. As the sample size used for SCM estimation decreases, the strength of the drift within the SCM estimation interval also decreases, leading to a smaller phase mismatch. Note that the instantaneous phase of the signals to which the beamformer is applied changes more rapidly for higher frequency bin indices k or larger SROs ε_i. This might cause the phase difference between the beamformer's input signals and the phase captured by the SCM estimates to vary periodically, alternating between alignment and misalignment.

For the target speaker's signal at the beamformer output this phase mismatch has a similar effect as an imperfect steering vector estimate in steering-vector-based beamforming. In previous works [10], [84], [85] it has been shown that an imperfect steering of the beamformer results in a degradation of its performance. An intuitive explanation of this effect can be given by the fact that the distortionless response is requested for signals stemming from another position. Moreover, the phase mismatch deteriorates the attenuation of the interference signal at the beamformer output. In a strongly simplified way this can be attributed to the fact that the beamformer tries to minimize the power of an interfering source, where the position of this source was estimated wrongly.

Figure 5.7 depicts the effect of the SRO-induced phase drift within the SCM estimation interval on the target signal (see Fig. 5.7(a)) and the interference signal (see Fig. 5.7(b)) at the beamformer output. Here, expected values based on a speech mixture are simulated using the LGM rather than considering true speech to obtain a less noisy figure. In order to avoid a switching behavior between time frames which are dominated by the interfering speaker and time frames that are dominated by noise, a single interfering speaker and absence of noise is considered as interference. The SROs are given by $\varepsilon_0 = 0\,\text{ppm}$, $\varepsilon_1 = -12.5\,\text{ppm}$, $\varepsilon_2 = -6.26\,\text{ppm}$, $\varepsilon_3 = 16.67\,\text{ppm}$, $\varepsilon_4 = 33.33\,\text{ppm}$ and $\varepsilon_5 = 50\,\text{ppm}$. For consistency with the next chapter a GEVD-based rank-1 estimate, which is calculated based on the ground-truth SCMs of the LGM, is used for the target SCM. Thus, the target-SCM estimate $\widehat{\mathbf{R}}_{\text{tar}}(k)$ is assumed to be a deterministic quantity from here on.

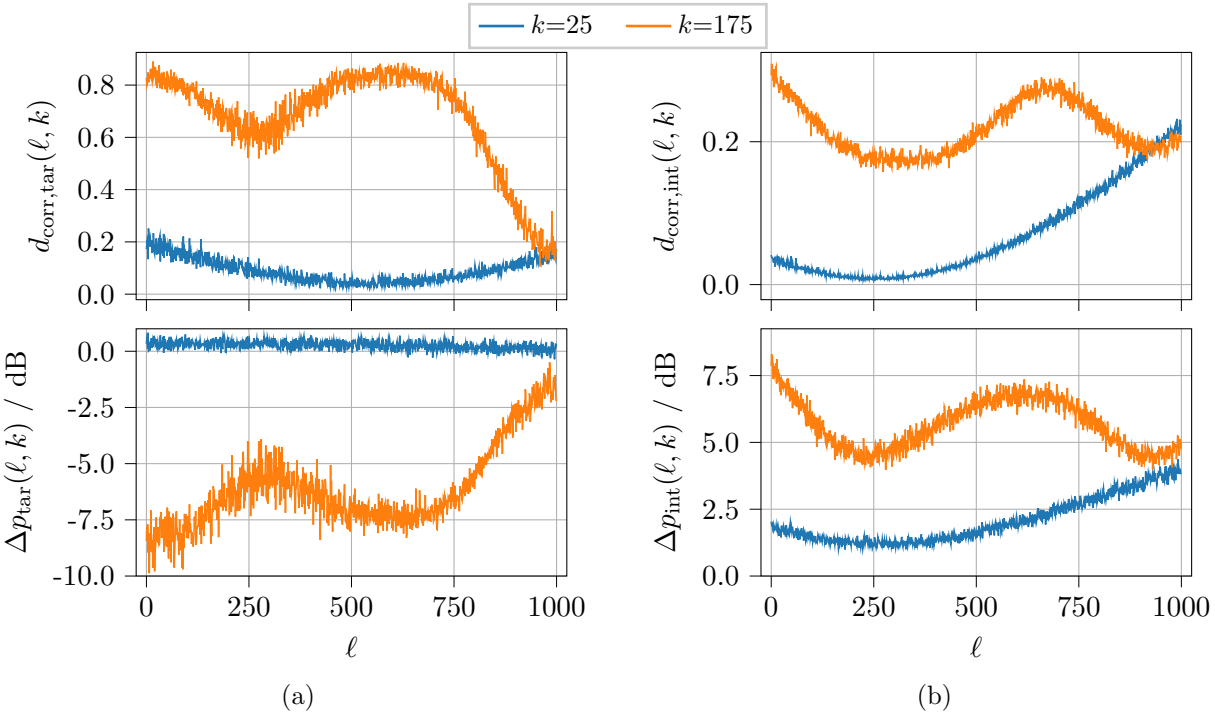


Figure 5.7: Influence of the SRO-induced phase mismatch between the SCM estimates and the signals to which the beamformer is applied on MVDR beamforming. $d_{\text{corr,tar}}(\ell, k)$ corresponds to the correlation matrix distance between the second-order moment of $\mathbf{x}_{\text{tar}}(\ell, k)$ and the expected value of the target-SCM estimate, as defined in (5.24). $\Delta p_{\text{tar}}(\ell, k)$ corresponds to the expected value of the difference in the power of the target signal at the beamformer output between the cases with and without SROs for time frame ℓ, as defined in (5.26). $d_{\text{corr,int}}(\ell, k)$ corresponds to the correlation matrix distance between the second-order moment of $\mathbf{x}_{\text{int}}(\ell, k)$ and the expected value of the interference-SCM estimate, as defined in (5.25). $\Delta p_{\text{int}}(\ell, k)$ corresponds to the expected value of the difference in the power of the interference signal at the beamformer output between the cases with and without SROs for time frame ℓ, as defined in (5.27). The expected values are estimated using a Monte-Carlo simulation with 500 draws.

In order to visualize the SRO-induced phase mismatch between the SCM estimates and the signals to which the beamformer is applied, the correlation matrix distance $d_{\text{corr}}(\cdot, \cdot)$, as introduced in Sec. 3.2, is utilized, with

$$d_{\text{corr,tar}}(\ell, k) = d_{\text{corr}}\left(\mathbb{E}\left[\mathbf{x}_0(\ell, k) \cdot \mathbf{x}_0^{\text{H}}(\ell, k)\right], \mathbb{E}\left[\widehat{\mathbf{R}}_{\text{tar}}(k)\right]\right) \tag{5.24}$$

measuring the distance between the target-SCM estimate and the instantaneous second-order moment of the target signal and

$$d_{\text{corr,int}}(\ell, k) = d_{\text{corr}}\left(\mathbb{E}\left[\mathbf{x}_1(\ell, k) \cdot \mathbf{x}_1^{\text{H}}(\ell, k)\right], \mathbb{E}\left[\widehat{\mathbf{R}}_{\text{int}}(k)\right]\right) \tag{5.25}$$

measuring the distance between the interference-SCM estimate and the instantaneous second-order moment of the interfering speaker's signal. Lower values for these distances signify a

smaller phase mismatch between the SCM estimates and the signals to which the beamformer is applied. Hence, lower values are better.

The impact of these phase mismatches on the beamformer's performance is measured by the differences $\Delta p_{\text{tar}}(\ell, k)$ and $\Delta p_{\text{int}}(\ell, k)$ between the power of the individual source signals at the beamformer output $\mathbb{E}\Big[\big|\mathbf{w}_{\text{SRO}}^{\text{H}}(k) \cdot \mathbf{x}_i^{\text{SRO}}(\ell, k)\big|^2\Big]$, $i \in \{0, 1\}$, with SROs being present and the corresponding quantities $\mathbb{E}\Big[\big|\mathbf{w}^{\text{H}}(k) \cdot \mathbf{x}_i(\ell, k)\big|^2\Big]$, $i \in \{0, 1\}$, without SROs being present, with

$$\Delta p_{\text{tar}}(\ell, k) = 10 \cdot \log_{10}\left(\frac{\mathbb{E}\Big[\big|\mathbf{w}_{\text{SRO}}^{\text{H}}(k) \cdot \mathbf{x}_0^{\text{SRO}}(\ell, k)\big|^2\Big]}{\mathbb{E}\big[|\mathbf{w}^{\text{H}}(k) \cdot \mathbf{x}_0(\ell, k)|^2\big]}\right) \tag{5.26}$$

and

$$\Delta p_{\text{int}}(\ell, k) = 10 \cdot \log_{10}\left(\frac{\mathbb{E}\Big[\big|\mathbf{w}_{\text{SRO}}^{\text{H}}(k) \cdot \mathbf{x}_1^{\text{SRO}}(\ell, k)\big|^2\Big]}{\mathbb{E}\big[|\mathbf{w}^{\text{H}}(k) \cdot \mathbf{x}_1(\ell, k)|^2\big]}\right). \tag{5.27}$$

The closer these values are to zero the smaller the impact on the beamforming performance, with positive values corresponding to a larger output power compared to the case without SROs and negative values corresponding to a smaller output power compared to the case without SROs. As mentioned above, it is expected that SROs result in a smaller power of the target component at the beamformer output belonging to a negative value of $\Delta p_{\text{tar}}(\ell, k)$ and in a deterioration of the interference suppression corresponding to a positive value of $\Delta p_{\text{int}}(\ell, k)$.

The correlation matrix distance $d_{\text{corr,tar}}(\ell, k)$ and the SRO-induced changes in the power of the target signal at the beamformer output $\Delta p_{\text{tar}}(\ell, k)$ are strongly correlated as shown in Fig. 5.7. The same holds for the correlation matrix distance $d_{\text{corr,int}}(\ell, k)$ and the SRO-induced changes in the power of the interference signal at the beamformer output $\Delta p_{\text{int}}(\ell, k)$ Thus, a larger SRO-induced distance between the second-order moment of the target signal and the target-SCM estimate comes with a larger reduction of the power of the target signal at the beamformer output compared to the case without SROs being present. This outcome is consistent with intuition since the target signal is suppressed more strongly when the difference of its statistics to the target-SCM estimate, which is responsible for the distortionless response, is larger. Similarly, a larger SRO-induced distance between the second-order moment of the interference signal and the interference-SCM estimate comes with a larger power of the interference signal at the beamformer output compared to the case without SROs being present. Again, this outcome is consistent with intuition since the interference is less suppressed when the difference of its statistics to the interference-SCM estimate, which defines the statistics of the signal whose variance should be minimized, is larger.

For low frequencies ($k = 25 \triangleq 390.625\,\text{Hz}$ in Fig. 5.7) the phase drift is slow so that the correlation matrix distances $d_{\text{corr,tar}}(\ell, k)$ and $d_{\text{corr,int}}(\ell, k)$ slowly increase with increasing temporal distance to their minima. These minima correspond to the time frames at which the phase of the signals to which the beamformer is applied and the phase of the SCM estimates

match best. The target signal slowly becomes more suppressed and the interference signal slowly becomes less suppressed. For higher frequencies ($k = 175 \triangleq 2734.375\,\text{Hz}$ in Fig. 5.7), a more complex behavior of the correlation matrix distances $d_{\text{corr,tar}}(\ell, k)$ and $d_{\text{corr,int}}(\ell, k)$ as well as a more complex behavior of the changes of the power of the signals at the beamformer output $\Delta p_{\text{tar}}(\ell, k)$ and $\Delta p_{\text{int}}(\ell, k)$ can be observed. The more complex behavior results from the faster drifting phase and the fact that the value of the phase drift differs for each microphone signal due to the different SROs, as shown in Fig. 5.6.

The changes of the correlation matrix distances $d_{\text{corr,tar}}(\ell, k)$ and $d_{\text{corr,int}}(\ell, k)$ during the SCM estimation interval can be fully attributed to the SRO-induced phase mismatch between the SCM estimates and the signals to which the beamformer is applied. The SRO-induced distortion of the amplitude, which was discussed w.r.t. Fig. 5.4, has the same influence on the signal's power across all time frames. Hence, it can be concluded that the SRO-induced phase mismatch between the SCM estimates and the signals to which the beamformer is applied can have a significant negative impact on the beamforming performance alongside the negative effects of the SRO-induced distortion of the SCM estimates' amplitude.

Note that the similarity of the second-order moment of the target speaker's signal $\mathbf{x}_1(\ell, k)$ to the interference-SCM estimate $\widehat{\mathbf{R}}_{\text{int}}(k)$ also has a small impact on the SRO-induced change of the power of the target signal at the beamformer output $\Delta p_{\text{tar}}(\ell, k)$. This might explain the negligibly small positive value of $\Delta p_{\text{tar}}(\ell, k)$ for k=25 in Fig. 5.7(a). Similarly, the SRO-induced change of the power of the interference signal at the beamformer output $\Delta p_{\text{int}}(\ell, k)$ is also slightly affected by the similarity between the second-order moment of the interference signal and the target-SCM estimate.

5.4 Summary

In this chapter, the advantages and disadvantages of spatially distributed recording devices compared to compact microphone arrays were discussed. Using meeting recognition as an example application, it was shown that a spatial distribution of the recording devices can lead to an improved performance of an MVDR beamformer compared to beamforming using a compact microphone array if a compensation for STOs and SROs is performed beforehand.

STOs do not affect an MVDR in the formulation given in (2.40) if the SCMs that are required to calculate the beamformer coefficients are estimated from the signals to which the beamformer is applied and the STOs do not become too large. However, large STOs lead to a degradation of the correlations between the signals and, therefore, might result in a deterioration of the performance of the beamformer. Thus, at least a coarse time alignment of the signals is required to minimize the STOs. For example, simple correlation methods can be used for this, as discussed in Sec. 7.1.1.

In contrast to this, SROs can heavily affect SCM estimation and, accordingly, the beamformer calculated from these SCM estimates. SROs do not only distort the amplitude of the SCM estimates as already reported in literature before but also introduce a phase mismatch between the SCM estimates and the signals to which the beamformer is applied. In this chapter, it

was shown that both effects negatively impact the performance of an MVDR beamformer, with their impact becoming more pronounced with increasing sample size used for SCM estimation. Hence, either a compensation for SROs is necessary or a small SCM estimation interval has to be chosen in order to mitigate the negative impact of SROs on the performance of the beamformer.

6 Interplay of SCM estimation from a finite sample size and asynchronous sampling

In this chapter, the interplay of spatial covariance matrix (SCM) estimation from a finite sample size and sampling time offsets (STOs) as well as sampling rate offsets (SROs) are considered. Here, the focus lies on the investigation of the resulting impact on the performance of an minimum variance distortionless response (MVDR) beamformer as a function of the sample size used for SCM estimation. In most previous works either the effects of SCMs that are estimated from a finite sample size or the effects of SROs were investigated in isolation. Only in [5] the influence of the strength of the SRO-induced distortions of the interference-SCM estimates on the sample size used for SCM estimation was examined. However, the impact of the SRO-induced distortions on the performance of the beamformer was not further investigated. In this chapter, the closed-form approximation of the signal-to-distortion ratio (SDR) at the ouput of an MVDR beamformer which was derived in Sec. 4.3 is extended to model STOs and SROs. Finally, the resulting closed-form approximation of the output SDR in the presence of SROs is employed to investigate the trade-off between utilizing a small sample size for SCM estimation to mitigate the effects of SROs and utilizing a large sample size for SCM estimation to diminish the finite sample size effects.

6.1 Derivation of a closed-form approximation of the output SDR in the presence of STOs

In the following, the effect of STOs will be integrated into the closed-form approximation of the output SDR of an MVDR beamformer which was derived in Sec. 4.3. Note that the block index, the frame index and frequency bin index are omitted where possible and only block b=0 is considered for these derivations to keep the notation as simple as possible.

By introducing STOs as in (2.14),

$$\mathbf{x}_0^{\mathrm{STO}}(\ell,k) \sim \mathcal{N}(\mathbf{0}, \sigma_0^2(\ell,k) \cdot \mathbf{S}(k) \cdot \mathbf{R}_0(k) \cdot \mathbf{S}^{\mathrm{H}}(k)), \tag{6.1}$$

$$\mathbf{x}_1^{\mathrm{STO}}(\ell,k) \sim \mathcal{N}(\mathbf{0}, \sigma_1^2(\ell,k) \cdot \mathbf{S}(k) \cdot \mathbf{R}_1(k) \cdot \mathbf{S}^{\mathrm{H}}(k)), \tag{6.2}$$

$$\boldsymbol{\nu}^{\mathrm{STO}}(\ell,k) \sim \mathcal{N}(\mathbf{0}, \sigma_\nu^2(k) \cdot \mathbf{S}(k) \cdot \mathbf{R}_\nu(k) \cdot \mathbf{S}^{\mathrm{H}}(k)), \tag{6.3}$$

follow for the local Gaussian model (LGM) of the short-time Fourier transform (STFT) of the source images, where $\mathbf{S}(k)$ is a diagonal matrix which consists of the phase terms modeling the STOs as specified in (2.15). As discussed in Sec. 5.2, the steering vector for the rank-1 target SCM estimate becomes

$$\widehat{\mathbf{d}}^{\mathrm{STO}} = \mathbf{S} \cdot \widehat{\mathbf{d}} \tag{6.4}$$

when STOs are present, with $\widehat{\mathbf{d}}$ corresponding to the steering vector estimated without STOs being present.

Next, the modifications to the approximation of the interference-SCM estimates by an equivalent Wishart matrix, as introduced in Sec. 4.2, are presented to account for STOs. To this end, it is assumed that

$$\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{\mathrm{STO}} \sim \mathcal{W}_M\left(L_{\backslash\ell}^{\mathrm{STO}}, \frac{1}{L_{\backslash\ell}^{\mathrm{STO}}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{STO}}\right) \tag{6.5}$$

approximately holds for the interference-SCM estimate in the presence of STOs. By applying the expected value operator to (5.4), it becomes obvious that

$$\mathbb{E}\left[\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{\mathrm{STO}}\right] = \mathbb{E}\left[\mathbf{S} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell} \cdot \mathbf{S}^{\mathrm{H}}\right] = \mathbf{S} \cdot \mathbb{E}\left[\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}\right] \cdot \mathbf{S}^{\mathrm{H}} \tag{6.6}$$

follows for the expected value of the interference-SCM estimate in the presence of STOs. From (4.23), $\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{STO}}=\mathbb{E}\left[\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{\mathrm{STO}}(k)\right]$ follows for the unnormalized equivalent scale matrix of the Wishart approximation in the presence of STOs. Hence, its relation to the unnormalized equivalent scale matrix $\boldsymbol{\Sigma}_{\mathrm{int}}=\mathbb{E}\left[\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}\right]$ for the case that there are no STOs is given by

$$\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{STO}} = \mathbf{S} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell} \cdot \mathbf{S}^{\mathrm{H}}. \tag{6.7}$$

Moreover, the variance of the elements of the interference-SCM estimate and the variance of the elements of the equivalent Wishart matrix are not affected by the phase shifts resulting from the STOs since both are calculated from the main-diagonal elements of the ground-truth SCMs and the equivalent scale matrix, respectively, as shown in (4.12) and (2.53). These are not affected by STOs. Hence, the equivalent degrees of freedom of the Wishart distribution as approximation of the distribution of the interference-SCM estimates are still calculated via (4.28) even if STOs are present, i.e., $L_{\backslash\ell}^{\mathrm{STO}}=L_{\backslash\ell}$ holds.

Note that $\mathbf{R}_{\mathbf{x}}^{\mathrm{STO}}=\mathbf{S}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{S}^{\mathrm{H}}$ and $\mathbf{R}_{\mathbf{r}}^{\mathrm{STO}}=\mathbf{S}\cdot\mathbf{R}_{\mathbf{r}}\cdot\mathbf{S}^{\mathrm{H}}$ results from (4.66) and (4.76) by comparing the LGM of the source images in the presence of STOs (see (6.1), (6.2), (6.3)) and the LGM of the source images if there are no STOs (see (4.3), (4.4), (4.5)). Utilizing $\mathbf{S}(k) \cdot \mathbf{S}^{\mathrm{H}}(k) = \mathbf{I}$, the numerator of the fraction in (4.59), which is calculated via (4.65), becomes

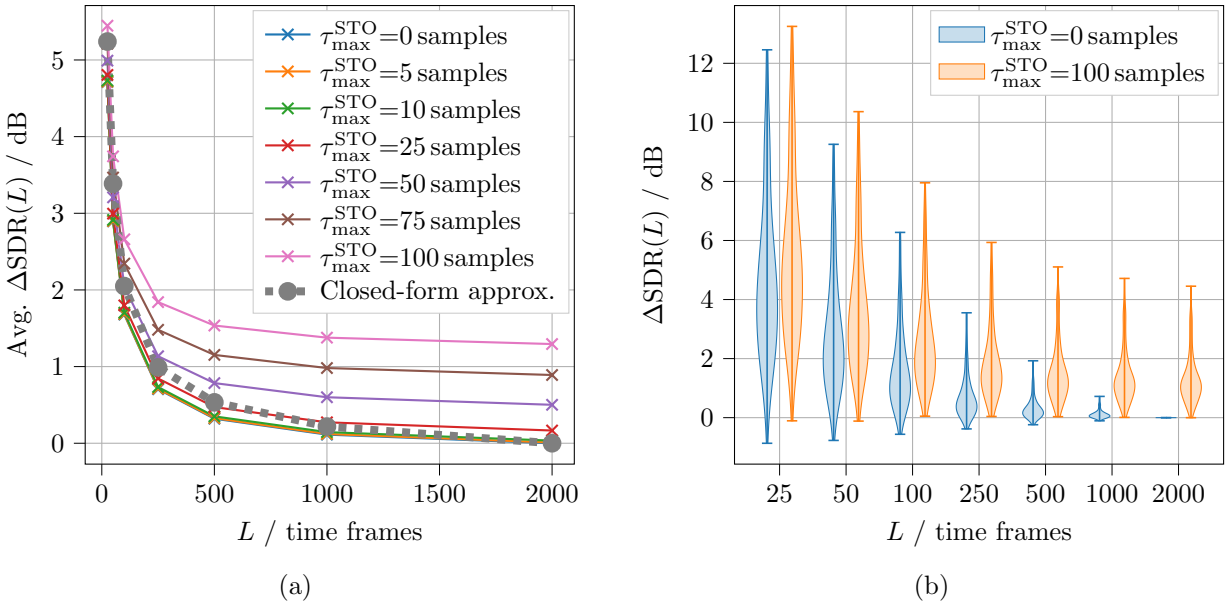
$$
\begin{aligned}
&\mathbb{E}\Big[\big|\mathbf{w}_{\mathrm{STO},\setminus\ell}^{\mathrm{H}}\cdot\mathbf{x}_{\mathrm{STO}}\big|^2\Big]\\
=&\frac{|\widehat{d}_0^{\mathrm{STO}}|^2}{L_{\setminus\ell}^{\mathrm{STO}}-M+1}\cdot\left(\frac{\mathrm{tr}\Big\{\big(\mathbf{\Sigma}_{\mathrm{int},\setminus\ell}^{\mathrm{STO}}\big)^{-1}\cdot\mathbf{R}_{\mathbf{x}}^{\mathrm{STO}}\Big\}}{\Big(\widehat{\mathbf{d}}^{\mathrm{STO}}\Big)^{\mathrm{H}}\cdot\big(\mathbf{\Sigma}_{\mathrm{int},\setminus\ell}^{\mathrm{STO}}\big)^{-1}\cdot\widehat{\mathbf{d}}^{\mathrm{STO}}}\right.\\
&\left.+\,(L_{\setminus\ell}^{\mathrm{STO}}-M)\cdot\frac{\Big(\widehat{\mathbf{d}}^{\mathrm{STO}}\Big)^{\mathrm{H}}\cdot\big(\mathbf{\Sigma}_{\mathrm{int},\setminus\ell}^{\mathrm{STO}}\big)^{-1}\cdot\mathbf{R}_{\mathbf{x}}^{\mathrm{STO}}\cdot\big(\mathbf{\Sigma}_{\mathrm{int},\setminus\ell}^{\mathrm{STO}}\big)^{-1}\cdot\widehat{\mathbf{d}}^{\mathrm{STO}}}{\left(\Big(\widehat{\mathbf{d}}^{\mathrm{STO}}\Big)^{\mathrm{H}}\cdot\big(\mathbf{\Sigma}_{\mathrm{int},\setminus\ell}^{\mathrm{STO}}\big)^{-1}\cdot\widehat{\mathbf{d}}^{\mathrm{STO}}\right)^2}\right)\\
=&\frac{|\widehat{d}_0\cdot S_{00}|^2}{L_{\setminus\ell}-M+1}\cdot\left(\frac{\mathrm{tr}\Big\{\mathbf{S}\cdot\mathbf{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{S}^{\mathrm{H}}\Big\}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\mathbf{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\widehat{\mathbf{d}}}\right.\\
&\left.+\,(L_{\setminus\ell}-M)\cdot\frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\mathbf{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\mathbf{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\widehat{\mathbf{d}}}{\Big(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\mathbf{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\widehat{\mathbf{d}}\Big)^2}\right)\\
=&\frac{|\widehat{d}_0|^2}{L_{\setminus\ell}-M+1}\cdot\left(\frac{\mathrm{tr}\Big\{\mathbf{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\mathbf{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}}+(L_{\setminus\ell}-M)\cdot\frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\mathbf{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}}{\Big(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\mathbf{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\Big)^2}\right)\\
=&\mathbb{E}\Big[\big|\mathbf{w}_{\setminus\ell}^{\mathrm{H}}\cdot\mathbf{x}\big|^2\Big],
\end{aligned}
\tag{6.8}
$$

when STOs are present, where $\mathbf{w}_{\mathrm{STO},\setminus\ell}^{\mathrm{H}}\cdot\mathbf{x}_{\mathrm{STO}}$ corresponds the output of the beamformer when applying the beamformer to the source images of the source of interest in the presence of STOs. This means that the numerator of (4.59) is not affected by STOs. In a similar way, it can can be shown that the denominator of (4.59) is not affected by STOs since all STO matrices $\mathbf{S}(k)$ cancel out when applying $\mathbf{S}(k)\cdot\mathbf{S}^{\mathrm{H}}(k)=\mathbf{I}$ (see Appendix A.7). In consequence, STOs do not affect the closed-from approximation of the output SDR as long as the STOs are so small that the model in (2.14) is valid. However, the effect of the reduced correlation between the channels due to large STOs is not reflected by the closed-form approximation of the SDR at the beamformer output.

6.2 Evaluation of the interplay of SCM estimation from a finite sample size and STOs

In the following, the interplay of the effects of SCM estimation from a finite sample size and the effects of STOs on MVDR beamforming is examined. Furthermore, the accuracy of the closed-form approximation of the SDR at the beamformer output in the presence of STOs, that was derived in the previous section, is investigated. For the most part, the same simulation framework as described in Chapter 3 is utilized. However, now the STOs $\tau_0^{\mathrm{STO}}{=}0\,$samples, $\tau_1^{\mathrm{STO}}{=}-\tau_{\max}^{\mathrm{STO}}$, $\tau_2^{\mathrm{STO}}{=}-\frac{\tau_{\max}^{\mathrm{STO}}}{2}$, $\tau_3^{\mathrm{STO}}{=}\frac{\tau_{\max}^{\mathrm{STO}}}{3}$, $\tau_4^{\mathrm{STO}}{=}\frac{2\cdot\tau_{\max}^{\mathrm{STO}}}{3}$ and $\tau_5^{\mathrm{STO}}{=}\tau_{\max}^{\mathrm{STO}}$ are

introduced where $\tau_{\max}^{\mathrm{STO}}$ is varied to assess the influence of the strength of the STOs on the effect of STOs on beamforming. The STOs are chosen in this way in order to guarantee that there always is an STO for all channel combinations which generally is not guaranteed for random sampling.



(a) (b)

Figure 6.1: Interplay of the effect of SCM estimation from a finite sample size and the effect of STOs on the performance of an MVDR beamformer. (a) shows the average SDR degradation $\Delta\mathrm{SDR}(L)$ as function of the block size L for different strengths of the STOs defined by the maximum STO $\tau_{\max}^{\mathrm{STO}}$. Here, solid lines correspond to beamforming applied to deterministic speech mixtures and the dashed line to the closed-form approximation. (b) shows a comparison of the distribution of the SDR degradation with STOs being present to the distribution of the SDR degradation without STOs being present.

Figure 6.1 visualizes how the beamforming performance is affected by the interplay of STOs and a finite sample size used for SCM estimation based on the average SDR degradation $\Delta\mathrm{SDR}(L)$ as a function of the block size L. For this purpose, the average SDR degradation for applying the beamformer to deterministic, simulated speech mixtures as well as the corresponding values resulting from closed-form approximations which was derived in Sec. 4.3 are shown. Since the closed-form approximation is not affected by STOs, as mentioned above, it is only depicted for the case without STOs being present. For small block sizes, the effect of estimating SCMs from a finite set of samples dominates the behavior of the beamformer such that the effect of STOs becomes negligible. While small STOs do not affect the MVDR beamformer, as reflected by the closed-form approximation of the output SDR, large STOs negatively influence the SDR at the beamformer output for large block sizes. However, the drop in the SDR due to STOs is less than 1 dB so that the effect on downstream tasks like automatic speech recognition (ASR) is expected to be small.

In Fig. 6.1(b) the distribution of the SDR degradation $\Delta\mathrm{SDR}(L)$ as a function of the block size L is shown for the case without STOs and the case of a maximum STO $\tau_{\max}^{\mathrm{STO}}$=100 samples. By comparing the SDR degradation with STOs being present to the case without STOs being

present, it can be seen that STOs do not only cause an increased SDR degradation in individual cases but lead to a degraded beamforming performance in all cases.

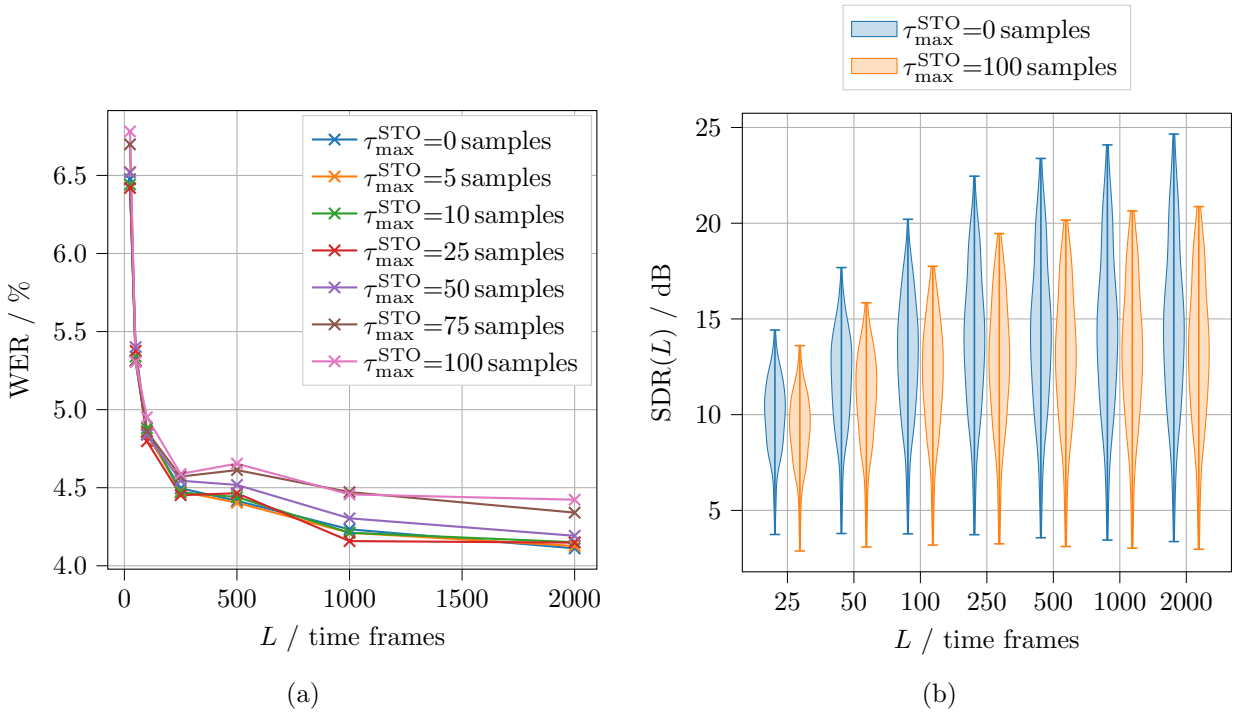


Figure 6.2: Influence of the interplay of the effect of SCM estimation from a finite sample size and the effect of STOs on the performance of ASR as a downstream task. (a) shows the word error rate (WER) as a function of the block size L and the strength of the STOs defined by the maximum STO $\tau_{\max}^{\mathrm{STO}}$. (b) shows a comparison of the distribution of the SDR at the beamformer output with STOs being present to the distribution of the SDR without STOs being present.

Up to this point only the SDR degradation due to finite sample size effects and STOs was considered which however does not reflect if these affect the SDR so much that downstream tasks, like ASR, are affected or the SDR still is high enough for these tasks despite these effects. To this end, Fig. 6.2(a) shows the word error rate (WER), that belongs to the SDR degradation shown in Fig. 6.1, for ASR as a downstream task. It can be observed that the ASR performance is mainly affected by the effects of a finite sample size used for estimating the beamformer coefficients for small block sizes. With growing block size the effects of STOs become more visible. However, the resulting increase of the WER is quite small and, therefore, acceptable. This holds even for STOs of a value of up to 100 samples. Note that such high STOs in the range of up to 100 samples typically occur rarely after a coarse synchronization of the signals but they can occur in practice. For example, consider a correlation-based coarse synchronization at the beginning of a meeting where only one speaker is active. In this case, the coarse synchronization drives the time shift between the signals towards zero and, therefore, the position-based time differences of flight (TDOFs), which can be in this range, correspond to the remaining STOs.

Figure 6.2(b) gives insights into the reasons why the SDR degradation does not heavily impact the ASR performance by comparing the distribution of the SDR as a function of the

block size L for the case without STOs being present to the case of $\tau_{\max}^{\mathrm{STO}}$=100 samples. It becomes obvious that STOs lead to a degradation of performance especially in cases for which high SDRs can be achieved at the beamformer output if the block size becomes so large that the effects of SCM estimation from a finite sample size become negligible. In contrast to that, the SDR degradation due to STOs is comparably small for examples for which only small SDRs can be achieved. However, for most downstream tasks the slight degradation of the SDR of up to 2 dB for the cases with high SDRs at the beamformer output has a vanishingly small effect and the cases with small SDRs are more fundamental for the performance. Thus, for most downstream tasks a coarse correlation-based synchronization should be good enough to compensate for STOs before beamforming.

6.3 Derivation of a closed-form approximation of the output SDR in the presence of SROs

Next, the closed-form approximation of the SDR at the output of an MVDR beamformer from Sec. 4.3 is extended to be able to model SROs. Note that the block index, the frame index and frequency bin index are omitted where possible and only block b=0 is considered for these derivations to keep the notation as simple as possible. First, the approximation of the probability distribution of the interference-SCM estimates by an equivalent Wishart distribution is adapted. To this end, it is assumed that

$$\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}} \sim \mathcal{W}_M\left(L_{\backslash\ell}^{\mathrm{SRO}}, \frac{1}{L_{\backslash\ell}^{\mathrm{SRO}}} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right) \tag{6.9}$$

approximately holds for the interference-SCM estimate in the presence of SROs. Using the SRO model in (2.21),

$$\mathbf{x}_0^{\mathrm{SRO}}(\ell,k) \sim \mathcal{N}(\mathbf{0}, \sigma_0^2(\ell,k) \cdot \mathbf{E}(\ell,k) \cdot \mathbf{R}_0(k) \cdot \mathbf{E}^{\mathrm{H}}(\ell,k)), \tag{6.10}$$

$$\mathbf{x}_1^{\mathrm{SRO}}(\ell,k) \sim \mathcal{N}(\mathbf{0}, \sigma_1^2(\ell,k) \cdot \mathbf{E}(\ell,k) \cdot \mathbf{R}_1(k) \cdot \mathbf{E}^{\mathrm{H}}(\ell,k)), \tag{6.11}$$

$$\boldsymbol{\nu}^{\mathrm{SRO}}(\ell,k) \sim \mathcal{N}(\mathbf{0}, \sigma_\nu^2(k) \cdot \mathbf{E}(\ell,k) \cdot \mathbf{R}_\nu(k) \cdot \mathbf{E}^{\mathrm{H}}(\ell,k)) \tag{6.12}$$

follows for the LGM of the source images when SROs are present, where $\mathbf{E}(\ell,k)$ is a diagonal matrix which consists of the phase terms modeling the SRO-induced time shifts between the signals as specified in (2.24). With $\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}$=$\mathbb{E}\left[\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right]$, which follows from (4.23), and (5.18), it follows that the unnormalized equivalent scale matrix of the approximate Wishart distribution of the interference-SCM estimates is given by

$$\begin{aligned}\boldsymbol{\Sigma}_{\text{int},\backslash\ell}^{\text{SRO}} =& \mathbb{E}\left[\widehat{\mathbf{R}}_{\text{int},\backslash\ell}^{\text{SRO}}\right] \\ =& \frac{1}{\sum_{\ell=0}^{L-1}\gamma_{\text{int}}(\ell)} \cdot \left(\sum_{\ell=0}^{L-1}\gamma_{\text{int}}(\ell)\cdot\sigma_0^2(\ell)\cdot\mathbf{E}(\ell)\cdot\mathbf{R}_0\cdot\mathbf{E}^{\text{H}}(\ell)\right. \\ &+ \sum_{\ell=0}^{L-1}\gamma_{\text{int}}(\ell)\cdot\sigma_1^2(\ell)\cdot\mathbf{E}(\ell)\cdot\mathbf{R}_1\cdot\mathbf{E}^{\text{H}}(\ell) \\ &\left.+ \sum_{\ell=0}^{L-1}\gamma_{\text{int}}(\ell)\cdot\sigma_\nu^2(\ell)\cdot\mathbf{E}(\ell)\cdot\mathbf{R}_\nu\cdot\mathbf{E}^{\text{H}}(\ell)\right) \end{aligned} \tag{6.13}$$

when SROs are present. Hence, the unnormalized scale matrix of the Wishart approximation of the interference-SCM estimates reflects the SRO-induced amplitude and phase distortions which were discussed in Sec. 5.3. Following the argumentation in Sec. 6.1 for the case of STOs being present, the element-wise variances of the SCM estimates and their approximation by a Wishart matrix are not affected by the SRO-induced distortion of the phase. Hence, the equivalent degrees of freedom of the equivalent Wishart matrix are still calculated via (4.28) when SROs are present, i.e., $L_{\backslash\ell}^{\text{SRO}}{=}L_{\backslash\ell}$. Thus, SROs distort only the phase and amplitude of the scale matrix of the Wishart matrix as approximation of the interference-SCM estimates but not the corresponding degrees of freedom and, therefore, the speed of convergence towards its expected value.

Next, the closed-form approximation of the power of the sources' signals at the beamformer output in (4.59) is extended in order to enable modeling SROs. Comparing the LGM of the source images in the presence of SROs (see (6.10), (6.11), (6.12)) and the LGM of the source images if there are no SROs (see (4.3), (4.4), (4.5)), $\mathbf{R}_{\mathbf{x}}^{\text{SRO}}{=}\mathbf{E}(\ell){\cdot}\mathbf{R}_{\mathbf{x}}{\cdot}\mathbf{E}^{\text{H}}(\ell)$ and $\mathbf{R}_{\mathbf{r}}^{\text{SRO}}{=}\mathbf{E}(\ell){\cdot}\mathbf{R}_{\mathbf{r}}{\cdot}\mathbf{E}^{\text{H}}(\ell)$ follows from (4.66) and (4.76). Accounting for the effects of SROs, the numerator of the power of the sources' signals at the beamformer output in (4.59), which is can be calculated via (4.65), becomes

$$\begin{aligned}&\mathbb{E}\left[\left|\mathbf{w}_{\text{SRO},\backslash\ell}^{\text{H}}{\cdot}\mathbf{x}_{\text{SRO}}\right|^2\right] \\ =& \frac{|\widehat{d}_0|^2}{L_{\backslash\ell}^{\text{SRO}}{-}M{+}1}\cdot\left(\frac{\text{tr}\left\{\left(\boldsymbol{\Sigma}_{\text{int},\backslash\ell}^{\text{SRO}}\right)^{-1}\cdot\mathbf{R}_{\mathbf{x}}^{\text{SRO}}\right\}}{\widehat{\mathbf{d}}^{\text{H}}\cdot\left(\boldsymbol{\Sigma}_{\text{int},\backslash\ell}^{\text{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}} + (L_{\backslash\ell}^{\text{SRO}}{-}M)\cdot\frac{\widehat{\mathbf{d}}^{\text{H}}\cdot\left(\boldsymbol{\Sigma}_{\text{int},\backslash\ell}^{\text{SRO}}\right)^{-1}\cdot\mathbf{R}_{\mathbf{x}}^{\text{SRO}}\cdot\left(\boldsymbol{\Sigma}_{\text{int},\backslash\ell}^{\text{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}}{\left(\widehat{\mathbf{d}}^{\text{H}}\cdot\left(\boldsymbol{\Sigma}_{\text{int},\backslash\ell}^{\text{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}\right)^2}\right) \\ =& \frac{|\widehat{d}_0|^2}{L_{\backslash\ell}{-}M{+}1}\cdot\left(\frac{\text{tr}\left\{\left(\boldsymbol{\Sigma}_{\text{int},\backslash\ell}^{\text{SRO}}\right)^{-1}\cdot\mathbf{E}(\ell)\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{E}^{\text{H}}(\ell)\right\}}{\widehat{\mathbf{d}}^{\text{H}}\cdot\left(\boldsymbol{\Sigma}_{\text{int},\backslash\ell}^{\text{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}}\right. \\ &\left.+ (L_{\backslash\ell}{-}M)\cdot\frac{\widehat{\mathbf{d}}^{\text{H}}\cdot\left(\boldsymbol{\Sigma}_{\text{int},\backslash\ell}^{\text{SRO}}\right)^{-1}\cdot\mathbf{E}(\ell)\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{E}^{\text{H}}(\ell)\cdot\left(\boldsymbol{\Sigma}_{\text{int},\backslash\ell}^{\text{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}}{\left(\widehat{\mathbf{d}}^{\text{H}}\cdot\left(\boldsymbol{\Sigma}_{\text{int},\backslash\ell}^{\text{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}\right)^2}\right).\end{aligned} \tag{6.14}$$

In the following, it is assumed that the MVDR beamformer is applied to the interfering speaker's signals. In this case, the second term in (6.14) corresponds to the power of the

interfering speaker's signal at the beamformer output when calculating the beamformer coefficients based on the ground-truth SCMs. Note that the unnormalized scale matrix $\mathbf{\Sigma}_{\text{int},\backslash\ell}^{\text{SRO}}$ of the Wishart approximation of the interference-SCM estimates takes the role of the interference-SCM estimates when calculating the beamformer coefficient in (6.14). Thus, it is expected that the phase mismatch between the unnormalized scale matrix $\mathbf{\Sigma}_{\text{int},\backslash\ell}^{\text{SRO}}$ of the Wishart approximation of the interference-SCM estimates and the SCM of the interfering speaker's signal $\mathbf{E}(\ell)\cdot\mathbf{R}_1\cdot\mathbf{E}^{\text{H}}(\ell)$ degrades the suppression of the interfering speaker's signal. Therefore, the SRO-induced amplitude and phase distortions of the interference-SCM estimates as well as the drifting phase of the interfering speaker's signal within one block typically lead to a degradation of the suppression of the interference. However, a decreased similarity of the interference-SCM estimate to the target SCM estimate, i.e., the outer product of the steering vector estimates $\widehat{\mathbf{d}}$, can also result in a tiny improvement of the suppression of the interference if the phase mismatch to the interference signals is small. For the target speaker's signal at the beamformer output, the second term in (6.14) can be seen as generalized Rayleigh quotient of the target SCM $\mathbf{E}(\ell)\cdot\mathbf{R}_0\cdot\mathbf{E}^{\text{H}}(\ell)$ at the steering vector $\widehat{\mathbf{d}}$ (see Appendix A.6). Thus, it follows that a larger phase mismatch between the target SCM and the steering vector results in a stronger suppression of the target speaker's signal.

SROs can be introduced for the denominator of the power of the sources' signals at the beamformer output in (4.59) in a similar way as it was done for the numerator in (4.59) above. However, the effect of SROs on the denominator in (4.59) is expected to be much more negligible since the denominator models the focus on the special characteristics of the currently considered time frame or has a value of one if the source of interest does not dominate the currently considered time frame. For sake of completeness the denominator of the power of the sources' signals at the beamformer output in (4.59) in the presence of SROs is stated in Appendix A.8.

6.4 Evaluation of the interplay of SCM estimation from a finite sample size and SROs

In the following, the interplay of the effect of SCM estimation from a finite sample size and the effect of SROs on the performance of an MVDR beamforming is investigated. Additionally, the accuracy of the closed-form approximation of the SDR at the beamformer output in the presence of SROs, that was derived in the previous section, is examined. The simulation framework employed throughout this section corresponds to the one that was described in Chapter 3. Additionally, SROs are introduced with $\varepsilon_0{=}0\,\text{ppm}$, $\varepsilon_1{=}-\frac{\varepsilon_{\max}}{4}$, $\varepsilon_2{=}-\frac{\varepsilon_{\max}}{8}$, $\varepsilon_3{=}\frac{\varepsilon_{\max}}{3}$, $\varepsilon_4{=}\frac{2\cdot\varepsilon_{\max}}{3}$ and $\varepsilon_5{=}\varepsilon_{\max}$. In order to investigate the influence of the strength of the SROs on beamforming the maximum SRO $\varepsilon_{\max}$ is varied in the following. As for the choice of the STOs, this definition should guarantee that there always is an SRO between two channels for all channel combinations. When dealing with deterministic speech signal, the SROs are simulated using the STFT-based resampling method from [79].

Moreover, the SRO-induced drifting phase of the signals also influences the steering vector estimates. However, this will not be further discussed since the quality of the steering vector

is not the focus of this consideration. Instead, a rather simple model for the effect of the SRO-induced drifting phase is employed. Therefore, the phase of the steering vector is aligned with the phase of the target speaker's signals at the center of each block.

Figure 6.3 visualizes how good the closed-form approximation of the SDR at the beamformer output, which was derived in Sec. 6.3, reflects the effects of the SRO-induced phase drift within one block. Note that the same example as for the experimental investigation of these effects in 5.7 is considered here. The left column (Fig. 6.3(a), Fig. 6.3(c), Fig. 6.3(e), Fig. 6.3(g)) corresponds to the signal of the target speaker at the beamformer output and the right column (Fig. 6.3(b), Fig. 6.3(d), Fig. 6.3(f), Fig. 6.3(h)) to the signal of the interfering speaker at the beamformer output. For both, the correlation matrix distance, which was introduced in Sec. 3.2, is used to measure the similarity of the sources' instantaneous second-order moments $\mathbf{E}(\ell) \cdot \mathbf{R}_0 \cdot \mathbf{E}^{\mathrm{H}}(\ell)$ and $\mathbf{E}(\ell) \cdot \mathbf{R}_1 \cdot \mathbf{E}^{\mathrm{H}}(\ell)$ to the target SCM estimate, i.e, the outer product of the steering vector $\widehat{\mathbf{d}}$, and the scale matrix $\boldsymbol{\Sigma}^{\mathrm{SRO}}_{\mathrm{int},\setminus\ell}$ of the approximate Wishart distribution of the interference-SCM estimates:

$$d^{\mathrm{steer,SRO}}_{\mathrm{corr,tar}}(\ell,k) = d_{\mathrm{corr}}\left(\mathbf{E}(\ell,k)\cdot\mathbf{R}_0(k)\cdot\mathbf{E}^{\mathrm{H}}(\ell,k), \widehat{\mathbf{d}}(k)\cdot\widehat{\mathbf{d}}^{\mathrm{H}}(k)\right), \tag{6.15}$$

$$d^{\mathrm{SCM,SRO}}_{\mathrm{corr,tar}}(\ell,k) = d_{\mathrm{corr}}\left(\mathbf{E}(\ell,k)\cdot\mathbf{R}_0(k)\cdot\mathbf{E}^{\mathrm{H}}(\ell,k), \boldsymbol{\Sigma}^{\mathrm{SRO}}_{\mathrm{int},\setminus\ell}(k)\right), \tag{6.16}$$

$$d^{\mathrm{steer,SRO}}_{\mathrm{corr,int}}(\ell,k) = d_{\mathrm{corr}}\left(\mathbf{E}(\ell,k)\cdot\mathbf{R}_1(k)\cdot\mathbf{E}^{\mathrm{H}}(\ell,k), \widehat{\mathbf{d}}(k)\cdot\widehat{\mathbf{d}}^{\mathrm{H}}(k)\right), \tag{6.17}$$

$$d^{\mathrm{SCM,SRO}}_{\mathrm{corr,int}}(\ell,k) = d_{\mathrm{corr}}\left(\mathbf{E}(\ell,k)\cdot\mathbf{R}_1(k)\cdot\mathbf{E}^{\mathrm{H}}(\ell,k), \boldsymbol{\Sigma}^{\mathrm{SRO}}_{\mathrm{int},\setminus\ell}(k)\right). \tag{6.18}$$

Further, the SRO-induced changes of the numerator and the denominator of the closed-form approximation of the power of the sources' signals at the beamformer output in (4.59) w.r.t. the resulting quantities without SROs being present are depicted:

$$\Delta^{\mathrm{SRO}}_{\mathrm{num,tar}}(\ell,k) = 10\cdot\log_{10}\left(\frac{\mathbb{E}\left[\left|\mathbf{w}^{\mathrm{H}}_{\mathrm{SRO},\setminus\ell}(k)\cdot\mathbf{x}^{\mathrm{SRO}}_0(\ell,k)\right|^2\right]}{\mathbb{E}\left[\left|\mathbf{w}^{\mathrm{H}}_{\setminus\ell}(k)\cdot\mathbf{x}_0(\ell,k)\right|^2\right]}\right), \tag{6.19}$$

$$\Delta^{\mathrm{SRO}}_{\mathrm{denom,tar}}(\ell,k) = 10\cdot\log_{10}\left(\frac{\mathbb{E}\left[\left|\delta\left(\mathbf{x}^{\mathrm{SRO}}_0(\ell,k)\right)\right|^2\right]}{\mathbb{E}\left[\left|\delta\left(\mathbf{x}_0(\ell,k)\right)\right|^2\right]}\right), \tag{6.20}$$

$$\Delta^{\mathrm{SRO}}_{\mathrm{num,int}}(\ell,k) = 10\cdot\log_{10}\left(\frac{\mathbb{E}\left[\left|\mathbf{w}^{\mathrm{H}}_{\mathrm{SRO},\setminus\ell}(k)\cdot\mathbf{x}^{\mathrm{SRO}}_1(\ell,k)\right|^2\right]}{\mathbb{E}\left[\left|\mathbf{w}^{\mathrm{H}}_{\setminus\ell}(k)\cdot\mathbf{x}_1(\ell,k)\right|^2\right]}\right), \tag{6.21}$$

$$\Delta^{\mathrm{SRO}}_{\mathrm{denom,int}}(\ell,k) = 10\cdot\log_{10}\left(\frac{\mathbb{E}\left[\left|\delta\left(\mathbf{x}^{\mathrm{SRO}}_1(\ell,k)\right)\right|^2\right]}{\mathbb{E}\left[\left|\delta\left(\mathbf{x}_1(\ell,k)\right)\right|^2\right]}\right) \tag{6.22}$$

where $\mathbf{w}_{\mathrm{SRO},\setminus\ell}(k)$ denotes the beamforming coefficients that are estimated in the presence of SROs.

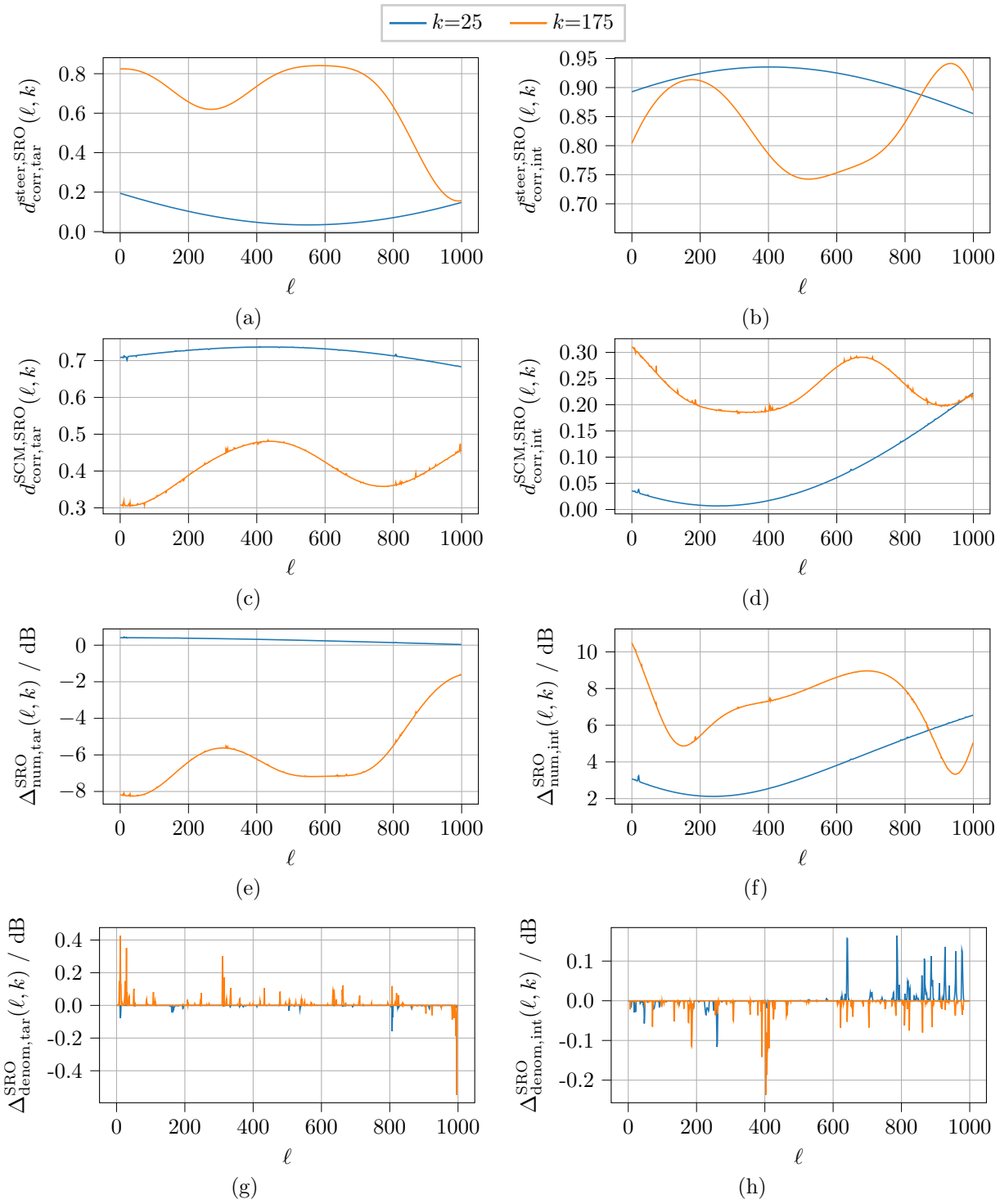


Figure 6.3: Influence of the SRO-induced phase mismatch between the SCM estimates and the signals to which the beamformer is applied on the closed-form approximation of the SDR at the beamformer output. $d^{\text{steer,SRO}}_{\text{corr,tar}}(\ell,k)$, $d^{\text{SCM,SRO}}_{\text{corr,tar}}(\ell,k)$, $d^{\text{steer,SRO}}_{\text{corr,int}}(\ell,k)$ and $d^{\text{SCM,SRO}}_{\text{corr,int}}(\ell,k)$ denote the correlation matrix distance of the sources' instantaneous second-order moment to the target SCM estimate and the scale matrix of the approximate Wishart distribution of the interference-SCM estimates. $\Delta^{\text{SRO}}_{\text{num,tar}}(\ell,k)$, $\Delta^{\text{SRO}}_{\text{denom,tar}}(\ell,k)$, $\Delta^{\text{SRO}}_{\text{num,int}}(\ell,k)$ and $\Delta^{\text{SRO}}_{\text{denom,int}}(\ell,k)$ correspond to the SRO-induced changes of the numerator and the denominator of the closed-form approximation of the power of the sources' signals at the beamformer output in (4.59) w.r.t. the resulting quantities without SROs being present.

It can be seen that the extension of the closed-form approximation of the power of the sources' signals at the beamformer output is able to reflect the SRO-induced effects on the beamformer output which were discussed in Sec. 5.3. The SRO-induced phase drift within one block mainly influences the numerator of the power of a sources' signal. This reflects the findings from Fig. 5.7, namely an increasing suppression of the target, i.e., $\Delta^{\mathrm{SRO}}_{\mathrm{num,tar}}(\ell, k)$ is negative, as well as a decreasing suppression of the interference, i.e., $\Delta^{\mathrm{SRO}}_{\mathrm{num,int}}(\ell, k)$ is positive, with growing dissimilarity to the corresponding SCM estimate, measured by $d^{\mathrm{steer,SRO}}_{\mathrm{corr,tar}}(\ell, k)$ and $d^{\mathrm{SCM,SRO}}_{\mathrm{corr,int}}(\ell, k)$. However, the SRO-induced changes of the similarity between the instantaneous target SCM and the interference-SCM estimate, measured by $d^{\mathrm{SCM,SRO}}_{\mathrm{corr,tar}}(\ell, k)$, and the SRO-induced changes of the similarity between the instantaneous interference SCM to the rank-1 target-SCM estimate, measured by $d^{\mathrm{steer,SRO}}_{\mathrm{corr,int}}(\ell, k)$, influence the numerator of a signal's power at the beamformer output to a certain extent, too. A signal becomes less suppressed if the corresponding instantaneous second-order moment is more similar to the target-SCM estimate and a signal becomes more suppressed if the corresponding second-order moment is more similar to the interference-SCM estimate. For example, this can be observed for the time frames 800 to 1000 of the interfering speaker's signal.

The effect of the drifting phase of the signal within one block on the denominator is comparably small, as reflected by $\Delta^{\mathrm{SRO}}_{\mathrm{denom,tar}}(\ell, k)$ and $\Delta^{\mathrm{SRO}}_{\mathrm{denom,int}}(\ell, k)$. This supports the role of the denominator to model the specialization of the MVDR beamformer to the special characteristics of individual time frequency bins due to the statistical dependence between the beamformer coefficients and the signals to which they are applied. As for the numerator these small changes of the denominator can be explained by the SRO-induced changes of similarity of a sources' instantaneous SCM to the target-SCM estimate and of its similarity to the interference-SCM estimate.

Figure 6.4 looks into the effects of a finite sample size used for SCM estimation and SROs on the performance of an MVDR beamformer. As shown in Fig. 6.4(a), the closed-form approximation is able to reflect the average SDR degradation $\Delta\mathrm{SDR}(L)$ due to SCM estimation from a finite sample size and SROs. Since the closed-form approximation is not able to reflect the reduced correlation between the channels due to large STOs, it is also not able to reflect the reduction of the correlation between the channels if the average SRO-induced time shift within one block becomes large. Hence, the reduced correlation between the channels due to the growing average SRO-induced time shift between the channels per block has a negligible effect compared to the other effects of SROs on the beamforming performance. Note that this only holds as long as the overlap between the time frames, which are extracted from different channels, is sufficiently large.

Moreover, it can be seen that the impact of a finite sample size used for SCM estimation dominates the beamforming performance for small block sizes, while for large block sizes, the negative effects of the SROs dominate. Even moderate SROs, i.e., SROs up to $\varepsilon_{\max}$=25 ppm, can cause a large degradation of the beamforming performance if the block size is too large. Without compensating for SROs, there is a trade-off between reducing the effects of finite sample size used for SCM estimation by increasing size of the estimation interval and minimizing the interval size to reduce the negative effect of the SROs. For small to moderate SROs, the negative effects of SROs can be mitigated by choosing a block size in the range of 100 time frames to 250 time frames, which corresponds to a 1.6 s to 4 s long

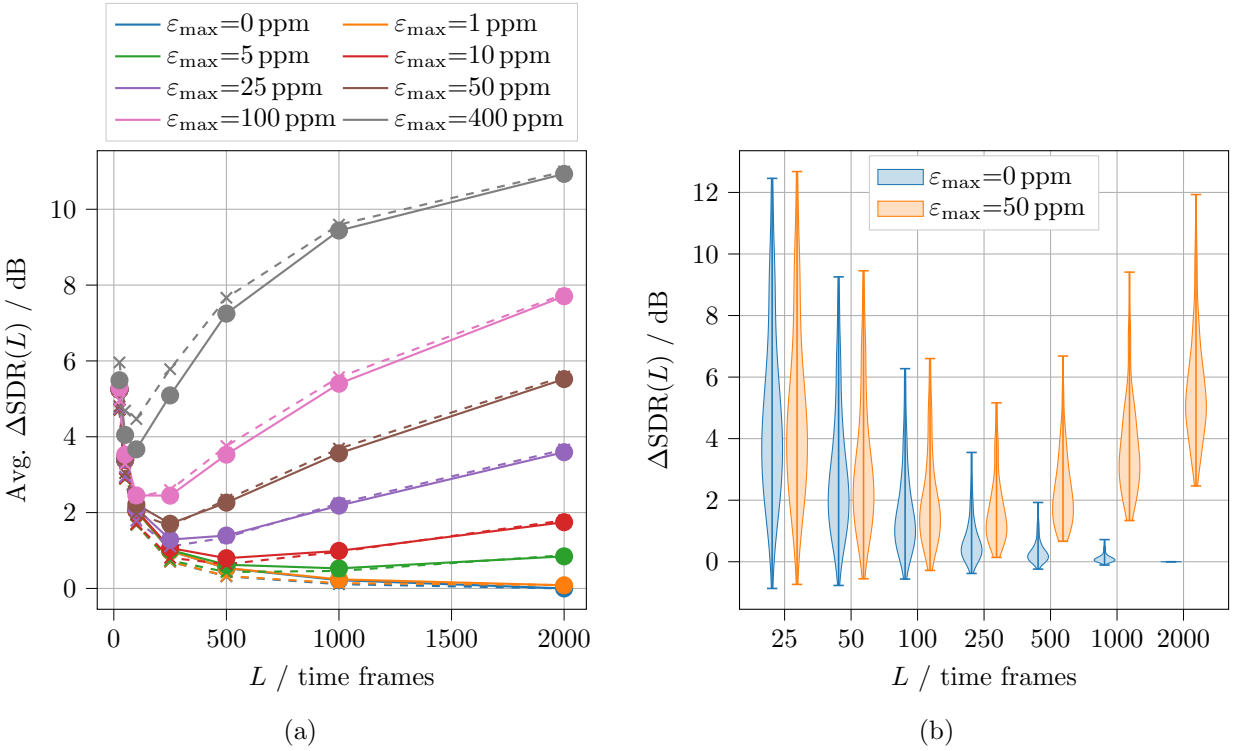


Figure 6.4: Interplay of the effect of SCM estimation from a finite sample size and the effect of SROs on the performance of an MVDR beamformer. (a) shows the average SDR degradation $\Delta\text{SDR}(L)$ as a function of the block size L for different strengths of the SROs defined by the maximum STO $\varepsilon_{\max}$. Here, solid lines correspond to beamforming applied to deterministic speech mixtures and dashed lines to the corresponding closed-form approximation. (b) shows a comparison of the distribution of the SDR degradation with SROs being present to the distribution of the SDR degradation without SROs being present.

part of the observed signals, at the cost of an acceptable loss of performance w.r.t. the optimal performance that can be achieved without SROs. Larger SROs always lead to a significant degradation of the beamforming performance w.r.t. the performance which is possible without SROs, even for the optimal choice of the block size.

The effect of a maximum SRO $\varepsilon_{\max}$ of 50 ppm on the distribution of the SDR degradation $\Delta\text{SDR}(L)$ as a function of the block size L is depicted in Fig. 6.4(b). As mentioned for the average SDR degradation, SROs do not influence the distribution of the SDR degradation if the block size is small. For large block sizes, SROs always cause a degradation of the beamforming performance measured by the SDR at the beamformer output, i.e., the increasing average value of the SDR degradation $\Delta\text{SDR}(L)$ reflects the general trend and does not results from single outliers.

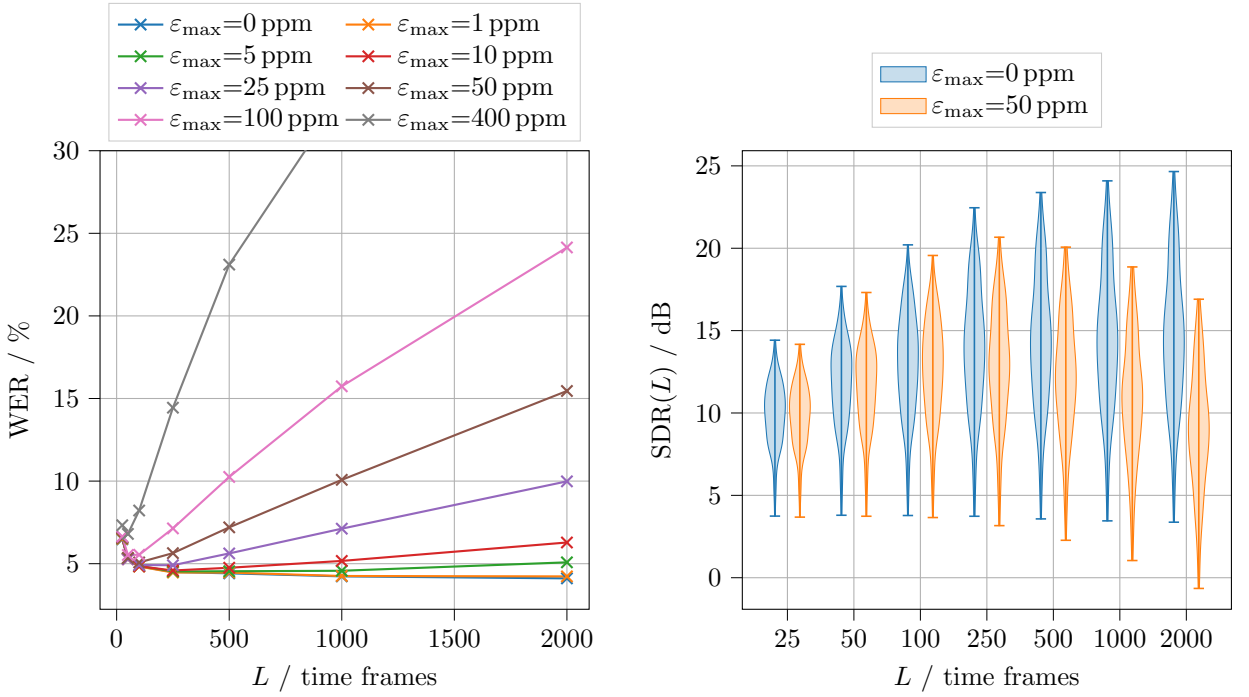


Figure 6.5: Influence of the interplay of the effect of SCM estimation from a finite sample size and the effect of SROs on the performance of ASR as a downstream task. (a) shows the WER as a function of the block size L and the strength of the SROs defined by the maximum STO $\varepsilon_{\max}$. (b) shows a comparison of the distribution of the SDR at the beamformer output with SROs being present to the distribution of the SDR without SROs being present.

In order to be able to assess the consequences of the SDR degradation, stemming from the interplay of a finite sample size used for SCM estimation and SROs, for downstream tasks, Fig. 6.5(a) shows the ASR performance measured by the WER as a function of the block size and the size of the SROs. It can be seen that the WER reflects the behavior of the corresponding SDR degradation in Fig. 6.4. The WER improves for small block sizes, where the finite sample size effects dominate, until a certain block size is reached. For larger block sizes the effects of the SROs dominate the beamforming performance so that the WER significantly degrades. The significantly higher WER compared to the case with STOs, even in cases with a similar average SDR degradation, might be attributed to the varying character of the strength of the negative impacts of SROs within one block, as shown for example in Fig. 6.3.

Figure 6.5(b) gives insights into the distribution of the SDR as a function of the block size L for the case without SROs being present and the changes of this distribution for a maximum SRO of 50 ppm. For small block sizes, the distribution of the SDR is not affected by SROs. As the block size grows, especially examples which exhibit a high SDR are affected. Thus, a small effect on the WER is expected similar to the case with STOs being present. However, the WER degradation due to SROs is much larger than for the case of STOs being present. This might be attributed to local distortions of the target signal at the beamformer output, e.g., due to an imperfect steering of the beamformer. For large SROs, the SDR tends to be

much smaller than for the case without SROs which leads to a significant increase of the WER.

6.5 Summary

The impact of SCM estimation based on a finite sample size and the effect of asynchronous signal sampling, i.e., STOs and SROs, on MVDR beamforming as well as their interplay were discussed on a theoretical basis and experimentally investigated in this chapter. To this end, the closed-form approximation of the SDR at the beamformer output was extended to model the effects of STOs and SROs. If the STOs are in the range of ± 25 samples, their effect on the beamforming performance is negligible. This degree of precision can be achieved by simple correlation methods, rendering sophisticated compensation for STOs unnecessary. Matters are different for SROs. There is a trade-off between an increased variance of the SCM estimates for small SCM-estimation intervals and the negative effects of SROs for large SCM-estimation intervals. Although the drop in performance of the beamformer can be kept at an acceptable level by carefully selecting the sample size for SCM estimation if the SROs are small, large SROs always result in a significant degradation of the performance. In consequence, a compensation for SROs is inevitable in practice to ensure the best-possible performance of an MVDR beamformer.

7 Signal synchronization

As mentioned in the previous chapter, synchronizing signals captured by different devices is crucial to ensure that the performance of algorithms, such as beamforming, does not degrade. In this context, the synchronization of the signals refers to compensating for the sampling time offsets (STOs) and sampling rate offsets (SROs) between the signals. Note that some algorithms, like beamforming as discussed in the previous chapter, do not require an accurate compensation for STOs but merely a minimization of the STOs. Compensating for SRO can be achieved either by resampling in software [3], [21], [79], [86], [87] or by adjusting the sampling frequency in hardware [OC3]. While many algorithms only require long-term stability of the time alignment between the signals, some algorithms also profit from an accurate alignment of the signals' sampling rates, as shown for beamforming in the previous chapter.

In most cases, STOs are estimated in form of the time shift between the signals via a correlation [21]. Thus, no distinction is made between the physical time differences of flight (TDOFs), i.e., the difference in time a signal needs to propagate from the position of the emitting source to the microphones, and the STOs so that both are compensated for at once. In this case, the resulting time differences of arrival (TDOAs) between the signal cannot be mapped to the sources' positions anymore. Only a few methods exists that differentiate between the physical TDOFs and the STOs by considering position estimation and STO estimation at once. In this chapter an alternative approach to STOs estimation is discussed that utilizes estimates for the distance between the sources and the microphones to enable a synchronization that maintains the physical TDOFs. For example, this enables the usage of source localization techniques to support the steering of the beamformer.

Typically, SROs are blindly estimated from the recorded signals. Previous studies on blind SRO estimation often rely on simplifying yet unrealistic assumptions, such as SROs that remain constant over time or the presence of a single acoustic source at a fixed position that is active throughout the complete estimation interval. In practice the conditions are more dynamic. This includes time-varying SROs that arise from the properties of the hardware, as discussed in Sec. 2.2. Further, in natural conversations, like meetings, multiple speakers are present and speaker changes occur, which imply changes of the position of the active speaker. In this chapter, an SRO estimator is developed which is able to blindly estimate the SRO from recordings of natural conversations under a hardware-realistic model for time-varying SROs.

The remainder of this chapter is organized as follows: First, existing methods for STO estimation are reviewed in Sec. 7.1.1. Subsequently, the proposed approach to STO estimation is introduced in Sec. 7.1.2. After reviewing existing SRO estimators in Sec. 7.2.1, an SRO estimator for scenarios with time-varying SROs and speaker position changes is proposed in

Sec. 7.2.1. Finally, the proposed approaches to STO and SRO estimation are evaluated in Sec. 7.3 before conclusions are drawn in Sec. 7.4.

7.1 Compensation for STOs

In the following, the estimation of STOs is considered, where STOs are estimated w.r.t. a reference channel. The estimated STOs are compensated for by time-shifting the signals by the estimated STOs [21].

7.1.1 Existing approaches to STO estimation

While there exist many works on SRO estimation, STO estimation is comparatively rarely considered. Moreover, many algorithms only require a minimization of the STOs rather than an accurate compensation for them. For example, the minimum variance distortionless response (MVDR) beamformer, which was analyzed in the previous chapter, or the meeting recognition pipeline which was proposed in [6] only require a coarse time alignment between the signals, meaning that the STO is minimized.

A coarse time alignment between the signals is often achieved via time-shifting the signals based on time-shift estimates which are obtained by searching for the maximum of the corresponding long-term time-domain correlations [6], [21]. Thus, the time shifts, which are compensated for, correspond to a superposition of the STO and the physical TDOF, i.e., the difference in time a signal requires for propagation from the acoustic source to the different microphones, as described in Sec. 2.2. Alternatively to calculating correlations in the time domain, the estimation of the time shifts between the signals can also be integrated into the SRO estimation framework. In [88] the online SRO estimates are used to obtain a SRO-compensated version of the generalized cross power spectral density (GCPSD) which in its SRO-uncompensated version is also the basis for SRO estimation. Finally, the time shift is estimated by aggregating the SRO-compensated version of the GCPSD and searching for the maximum of the corresponding generalized cross-correlation (GCC). As mentioned in [88], the STOs can be derived from the time shift estimates when external knowledge of the physical TDOFs is available.

In addition to that, there exist more sophisticated approaches to STO estimation which explicitly distinguish between STOs and TDOFs by considering the estimation of the microphone positions, the source positions and STOs as one problem [89]–[92]. In order to estimate the STOs and position information, time of arrival (TOA) or TDOA estimates are utilized. The non-linear optimization problem, which follows in this case, is solved in an iterative manner. A TOA-based approach that solves the optimization problem by alternatingly optimizing timing information, including the STOs, and position information is presented in [89]. More commonly, TDOA information is used for the joint estimation of STOs and the positions of the microphones and sources [90]–[92]. For STO estimation based on TDOA information, the non-linear optimization often is solved in two steps by either first estimating the position of the microphones and sources and with this knowledge the STOs or by solving these two steps in reverse order. For instance, first an unscaled version of the relative positions of the

microphones and sources was estimated using direction of arrival (DOA) information in [91]. Subsequently, the scaling of the relative position estimates was jointly estimated together with the STOs based on TDOAs. The authors in [92] proposed to create a pseudo TOA matrix from TDOA estimates by introducing STOs among other timing information. Next, the STOs were estimated using the low-rank property of the pseudo TOA matrix. In [90] an auxiliary function for the non-linear optimization problem was proposed which can be iteratively solved based on closed-form updates of the involved parameters.

7.1.2 Physically correct synchronization

The joint consideration of position and STO estimation, which enables distinguishing between STOs and TDOF information, requires an iterative solution of a non-linear optimization problem. It is to be mentioned that iterative methods require a good initialization to guarantee that they do not diverge or converge towards a local optimum. Further, the additional unknowns which are introduced in form of the positions of the microphones and sources increase the number of observations that is needed to be able to solve the optimization problem. This means that sources at a certain number of positions have to be observed before the STOs can be estimated.

To overcome these issues, an alternative approach to physically correct STO estimation, which was first proposed in [OC1], is presented in the following. It is based on estimates of the distance between the sources' positions and the microphones' position. The estimation of these distances has received a lot of interest in recent years. Here, deep neural network (DNN) based estimators have prevailed which either make use of single-channel information [93] or multi-channel information [OC5], [94], [95].

In the following, the estimation of the STO τ_{ij}^{STO} between channel i and channel j is considered. To this end, it is assumed that the SRO have already been compensated for so that the time shift between the i-th and the j-th microphone signal in (2.29) corresponds to the sum of the STO and the TDOF. Further, it is assumed that the STO is minimized using a correlation-based coarse time-alignment in order to enable a coherent processing of the signals. For sake of simplicity, this effect is not regarded in the following equations.

Given L_{STO} segment-wise estimates of the time shift $\widehat{\tau}_{ij}(l_{\mathrm{STO}})$ between the i-th and the j-th microphone signal, with l_{STO} denoting the index of the segments used for STO estimation, and the corresponding L_{STO} segment-wise estimates of the distance between the sources and the microphones $\widehat{r}_i(l_{\mathrm{STO}})$ and $\widehat{r}_j(l_{\mathrm{STO}})$, STO estimation can be formulated as an least squares (LS) problem. First, the distance estimates are used to estimate the TDOF via

$$\widehat{\tau}_{ij}^{\mathrm{TOF}}(l_{\mathrm{STO}}) = \frac{\widehat{r}_j(l_{\mathrm{STO}}) - \widehat{r}_i(l_{\mathrm{STO}})}{c} \cdot f_{\mathrm{s}}, \tag{7.1}$$

with c denoting the speed of sound and f_{s} the nominal sampling frequency. Based on these estimates and (2.29), the STO estimate is obtained as the solution to the following LS

problem:

$$\widehat{\tau}_{ij}^{\mathrm{STO}} = \underset{\breve{\tau}_{ij}^{\mathrm{STO}}}{\operatorname{argmin}} \sum_{l_{\mathrm{STO}}=0}^{L_{\mathrm{STO}}-1} \left(\breve{\tau}_{ij}^{\mathrm{STO}} - \frac{\widehat{r}_j(l_{\mathrm{STO}}) - \widehat{r}_i(l_{\mathrm{STO}})}{c} \cdot f_{\mathrm{s}} + \widehat{\tau}_{ij}(l_{\mathrm{STO}}) \right)^2 . \tag{7.2}$$

Solving the LS problem, leads to

$$\widehat{\tau}_{ij}^{\mathrm{STO}} = \frac{1}{L_{\mathrm{STO}}} \cdot \sum_{l_{\mathrm{STO}}=0}^{L_{\mathrm{STO}}-1} \left(\frac{\widehat{r}_j(l_{\mathrm{STO}}) - \widehat{r}_i(l_{\mathrm{STO}})}{c} \cdot f_{\mathrm{s}} - \widehat{\tau}_{ij}(l_{\mathrm{STO}}) \right) . \tag{7.3}$$

Note that segments without source activity are excluded from the LS problem based on an energy-based source activity detector.

It is to be mentioned that the described method leads to a closed-form solution for physically correct STO estimation in contrast to the STO estimators which were described before. Moreover, a source at a single position would be sufficient to estimate the STO if the distance and time shift estimates would be perfect. However, the proposed STO estimator benefits from additional observations so that it can compensate for errors of the distance and time shift estimates. Since the distance and time shift estimates may have very large errors, which can dominate the LS problem, e.g., when multiple sources are active within a segment, the LS solver is embedded in a random sample consensus (RANSAC) method [96] in order to reject outliers.

7.2 Compensation for SROs

SROs can be compensated for either via resampling in software or by adjusting the sampling frequency of the hardware in online applications during the recording if possible. Thereby, resampling in software, which can be applied to the already recorded signals offline and online, is the much more common way. The simplest form of resampling is the multiplication with a complex-valued phase term [20] in the short-time Fourier transform (STFT)-domain which cancels the SRO-induced phase term of the linear drift model which was introduced in Sec. 2.2. However, this way of resampling becomes inaccurate if the SRO-induced time shift between the signals becomes much larger than one sample due to the cyclic wrap-around effects as mentioned in [79]. In order to handle this effect, it was proposed in [79] to compensate for the integer-value component of the SRO-induced time shifts by shifting the analysis window of the STFT and only compensate for the remainder of the SRO-induced time shifts via a multiplication with a complex-valued phase term. In this work the focus lies on on SRO estimation rather than on resampling. Therefore, refer to the literature [3], [21], [79], [86], [87] for a more detailed view of resampling methods.

7.2.1 Existing SRO estimators

One way for SRO estimation is to utilize time stamps which are transmitted among the recording devices. Here, a two-way message exchange protocol is utilized to estimate the SRO for a

pair of devices [97], [98]. In [98] an additional Kalman filter was introduced to track the generally time-varying SRO values. However, the estimation of SROs based on time stamps comes with a huge communication overhead. Further, the accuracy of the SRO estimates can be shown to grow proportional to the number of time stamp exchanges [99].

The more common way is to estimate SROs from the recorded signals. In most cases, only the assumption is made that the recorded signals are coherent, i.e. they show consistent phase and amplitude relationships across microphones. This approach to SRO estimation is referred to as blind method. However, there are also active approaches to SRO estimation. For instance, in [100] specially designed sinusoidal signals were proposed that are actively played back for SRO estimation.

A common way to estimate SROs is to maximize the correlation or the coherence between the signals, corresponding to a normalized version of the cross power spectral density (CPSD) between the signals. In [101] a wideband correlation processor is proposed which utilizes the cross correlation between the signals, after resampling them based on a pre-defined set of SROs values, as ambiguity function. The recursive band-limited interpolation (RBI) method, which was presented in [102], circumvents the exhaustive search for the maximum of this ambiguity function by approximating it via a truncated sinc-interpolation which enables an optimization via a gradient-descent method. Instead of maximizing the correlation between the signals in time domain, the estimation of SROs by maximizing the coherence between the signals in the STFT domain after compensating for the SRO-induced phase drifts via a multiplication with a phase term was proposed in [4]. In [20] a correlation maximization (CM) algorithm was proposed which performs SRO estimation via an exhaustive search of the SRO value from a set of predefined values that maximizes the correlation coefficient between the signals after compensating for this SRO.

A maximum likelihood method for SRO estimation was proposed in [103]. This approach builds upon the local Gaussian model (LGM) as statistical model for the STFT of the signals where the corresponding likelihood is estimated based on the STFT coefficients after compensating for a predefined set of SROs. It utilizes the fact that the likelihood of the STFT coefficients is maximized if the remaining SRO between both signals approaches a value of zero.

A large group of SRO estimators is based on the linear phase-drift model which was presented in Sec. 2.2 where either the phase-drift of the coherence or the phase-drift of the correlation is employed for SRO estimation. These methods generally combine consecutive coherence or correlation estimates to map the SRO-induced phase-drift to a constant phase term which is proportional to the SRO. The average coherence drift (ACD) method [3] employs the drift between consecutive coherence functions where the SRO is estimated by averaging the phase of the quotient of consecutive coherence functions over the frequency range that shows no wrap-around effects. This method was improved in [2] by replacing the quotient of consecutive coherence functions by the complex-conjugated product of consecutive coherence functions which results in the weighted average coherence drift (WACD) method. In this way an signal-to-noise ratio (SNR)-related weighting of the time-frequency bins is introduced which leads to more accurate SRO estimates. In [104], the least-squares coherence drift (LCD) method was proposed that estimates the SRO via a LS problem which is based on the phase of the quotient of consecutive coherence functions. As an alternative to the usage of the phase

drift of the coherence, it was proposed in [105] to estimate the SRO based the drift of the maximum of a time-domain correlation. To this end, the proposed double-cross-correlation processor (DXCP) utilizes that the positions of the maximum of a secondary cross correlation between two primary cross correlation functions is proportional to the SROs. This method was further improved via a frequency-domain implementation named double-cross-correlation processor with phase transform (DXCP-PhaT) [88] which builds upon the phase-drift of the GCPSD.

Another aspect to be mentioned is that the performance of SRO estimators typically degrade for large SROs or long recordings [20]. As mentioned in [20], the performance of SRO estimators degrades with decreasing coherence between the signals which results from the SRO-induced time-shift which increases over time. In addition to that, modeling the effect of SROs using a multiplication with a phase term in the STFT domain, which is the basis for a large class of SRO estimators, becomes less accurate for long recordings and large SROs, as discussed in [20]. The authors of [106] proposed a coarse synchronization scheme before applying the maximum likelihood method for SRO estimation from [103] to overcome the issues due to long recordings. Therefore, a coarse estimate of the SRO and STO is obtained based on finding two pairs of short intervals that belong to the same part of the underlying continuous-time signals. This method was further improved in [21] by alternating SRO and STO estimation to account for the mutual dependence of the SRO and STO estimates. Moreover, a two stage approach to SRO estimation was proposed in [20], where the SRO estimate from the first stage is compensated for to obtain a more accurate SRO estimate in the second stage. Similarly, the WACD was extended in [2] to a multi-stage SRO estimate procedure with intermediate resampling steps which utilizes more than two stages.

With the upcoming transition from offline to online SRO estimation new methods to improve SRO estimation for large SROs or long recordings emerged. For instance, the online WACD method from [OC6] accounts for the SRO-induced increasing time shift between the signals by shifting the analysis window used for coherence estimation based on the estimated SROs. In this way, the integer-value part of SRO-induced time shift between the signals is compensated for without need for an explicit resampling. In [107] a control architecture for the feedback of the SRO into an upstream resampling before SRO estimation was introduced. Similarly, the SRO estimates are used for an upstream resampling in [OC7] however without the need of a control architecture. In addition to that, online SRO estimation enables a compensation for SROs in streaming applications, such as video conferencing.

In early works on SRO estimation very static conditions, including time-constant SROs and a single acoustic source at a fixed position which is permanently active, are considered. However, in recent years the trend goes towards more realistic conditions like time-varying SROs, multi-source scenarios and source movements. In order to account for negative effects resulting from multiple sources being active, the LCD method was extended to a weighted least-squares estimator with an additional outlier rejection in [108]. Moreover, time-varying SROs were considered in online SRO estimation in [102], [107]. In these works however SROs with unrealistic jumping discontinuities were used in the experiments. In [109] SRO estimation for conversations with natural speaker changes, like meetings, in low SNR environments was considered. In order to handle the challenges emerging from this scenario an online SRO estimator was proposed, which maximizes the real part of a recursively smoothed version of a

secondary cross correlation in the frequency domain. The problem of source movements was addressed in [110] by searching for periods in time without source movements and estimate the SROs only during these periods in time.

7.2.2 Development of an SRO estimator for time-varying SROs in the presence of speaker position changes

In the following, the dynamic weighted average coherence drift (DWACD) method for SRO estimation is presented, which was first introduced in [OC1]. This method is an extension of the online WACD method from [OC6] in order to enable the handling of dynamic scenarios with time-varying SROs and speaker position changes.

First, the SRO-induced phase drift of the coherence, which forms the basis of the DWACD method, and its use for SRO estimation are recapitulated. For sake of a simpler notation, it is assumed that only a single speaker is active although later the general signal model from Sec. 2.1 with more than one speaker is considered. Moreover, the speaker index of the acoustic transfer function (ATF) will be omitted in the following for sake of a shorter notation. Further, the narrow-band approximation of the STFT coefficients of the microphone signals, which is presented in Sec. 2.1.1, is used. It is to be mentioned that this approximation might not be accurate for typical choices of the STFT parameters used for SRO estimation. However, this choice can be justified due to the fact that the diffuse components of the late reverberation can be absorbed into the noise signal since their coherence is comparably small compared to the coherence of the direct path component and the early reflection components of the speech signals. This can be motivated by the definition of the coherent-to-diffuse power ratio (CDR) [13].

In the DWACD method, the coherence $\Gamma_{ij}(\ell, k)$ between microphone signal i and microphone signal j builds the basis for estimating the SRO between these signals. The coherence is defined as

$$\Gamma_{ij}(\ell, k) = \frac{\Phi_{y_i y_j}(\ell, k)}{\sqrt{\Phi_{y_i y_i}(\ell, k) \cdot \Phi_{y_j y_j}(\ell, k)}} \tag{7.4}$$

with the power spectral densitys (PSDs)

$$\Phi_{y_i y_j}(\ell, k) = \frac{1}{L_\mathrm{w}} \cdot \sum_{\widetilde{\ell}=\ell}^{\ell+L_\mathrm{w}-1} y_i\Big(\widetilde{\ell}, k\Big) \cdot y_j^*\Big(\widetilde{\ell}, k\Big) \tag{7.5}$$

being estimated via a Welch method using an estimation interval of L_w time frames.

As shown in [2], the coherence can be decomposed into 3 factors:

$$\Gamma_{ij}(\ell, k) = \widetilde{h}_{ij}(k) \cdot \rho_{ij}(\ell, k) \cdot \exp\left(\mathrm{j} \cdot \frac{2 \cdot \pi \cdot k}{N} \cdot \left(\left(\frac{N}{2} \cdot \varepsilon_{ij}(0) + \sum_{\widetilde{\ell}=1}^{\ell} \varepsilon_{ij}\left(\widetilde{\ell}\right)\right) \cdot B - \tau_{ij}^{\mathrm{STO}}\right)\right). \tag{7.6}$$

The first factor in (7.6)

$$\widetilde{h}_{ij}(k) = \frac{h_i(k) \cdot h_j^*(k)}{\sqrt{|h_i(k)|^2 \cdot |h_j(k)|^2}} \tag{7.7}$$

summarizes the dependence of the coherence on the ATFs and the second factor in (7.6)

$$\rho_{ij}(\ell, k) \approx \frac{\sum_{\widetilde{\ell}=\ell}^{\ell+L_\mathrm{w}-1} \left| z\left(\widetilde{\ell}, k\right) \right|^2 \cdot \exp\left(\mathrm{j} \cdot \frac{2 \cdot \pi \cdot k}{N} \cdot \varepsilon_{ij}(\ell) \cdot \widetilde{\ell} \cdot B \right)}{\sqrt{\left(\sum_{\widetilde{\ell}=\ell}^{\ell+L_\mathrm{w}-1} \left| z\left(\widetilde{\ell}, k\right) \right|^2 + \left| \frac{\nu_i\left(\widetilde{\ell}, k\right)}{h_i(k)} \right|^2 \right) \cdot \left(\sum_{\widetilde{\ell}=\ell}^{\ell+L_\mathrm{w}-1} \left| z\left(\widetilde{\ell}, k\right) \right|^2 + \left| \frac{\nu_j\left(\widetilde{\ell}, k\right)}{h_j(k)} \right|^2 \right)}} \tag{7.8}$$

is an SNR-related weight. Note that in the numerator of (7.8) the contribution of the noise and the cross terms between the noise and the speech signal were omitted in [2] utilizing the uncorrelatedness of these signals and the fact that all signal have an expected value of zero. Further, the cross terms between noise and speech signal in the denominator of (7.8) were neglected in [2] due to the same reason. For sake of simplicity, these additional terms in the numerator and denominator of (7.8) are neglected here, too, because their contribution generally is much smaller than the contribution of the other terms. The last factor in (7.6) corresponds to an SRO-related phase term which is employed for SRO estimation. Note that the SRO $\varepsilon_{ij}(\ell)$ is assumed to be constant during the interval used for estimating the coherence which is motivated by its very slow drift.

The SRO-related phase term in (7.6) does not only contain a drifting term which is linked to the SRO but also a constant component which arises from the STO. Thus, it was proposed in [3] to employ the ratio of two consecutive coherence functions to get rid of the influence of the STO and to remove the effect of accumulating the phase drift before estimating the SRO from this phase term. However, this also means that the information whether the signals are coherent and, therefore, lead to reliable SRO estimates is lost. Instead, signal segments with a coherent source signal and signal segments without a coherent source signal are weighted equally when estimating the SRO.

In order to maintain the information if a coherent source signal is present, the WACD method for SRO estimation was introduced in [2]. This method utilizes the complex-conjugated product of consecutive coherence functions to maintain the information how coherent a signal segment is as weighting of the individual signal segments used for SRO estimation. Assuming a slowly changing SRO, the SRO can be approximated to be constant during the whole estimation interval for both consecutive coherence functions. Thus, the complex-conjugated product of two consecutive coherence functions with a temporal distance of ℓ_d time frames is given by

$$\begin{aligned} g_\Gamma(\ell, k) &= \Gamma_{ij}(\ell, k) \cdot \Gamma_{ij}^*(\ell - \ell_\mathrm{d}, k) \\ &= \widetilde{h}_{ij}(k) \cdot \widetilde{h}_{ij}^*(k) \cdot \rho_{ij}(\ell, k) \cdot \rho_{ij}^*(\ell - \ell_\mathrm{d}, k) \cdot \exp\left(\mathrm{j} \cdot \frac{2 \cdot \pi \cdot k}{N} \cdot \ell_\mathrm{d} \cdot \varepsilon_{ij}(\ell) \cdot B \right). \end{aligned} \tag{7.9}$$

The resulting product consists of a phase term which is proportional to the SRO $\varepsilon_{ij}(\ell)$ and a prefactor. As already discussed in [2], the phase of the prefactor optimally has to be zero for SRO estimation.

In order to achieve that the prefactor in (7.9) has a phase of zero, multiple conditions have to be fulfilled. An important requirement is that the position of the coherent source is the same for the signal segments used to estimate the two consecutive coherence functions. In this case, $\widetilde{h}_{ij}(k) \cdot \widetilde{h}_{ij}^{*}(k) = 1$ holds so that the influence of the ATFs in (7.9) is completely removed. For the problem at hand with speaker changes or even multiple speaker's being active at the same time, the requirement that the two consecutive coherence functions belong to the same source position is not always fulfilled. This problem was already discussed in [108] and taken into account by extending a previously proposed LS method for SRO estimation to a weighted LS method. In the DWACD method this issue is taken into account by using shorter segments for coherence estimation and reducing the temporal distance between the two coherence functions that are used to calculate the complex-conjugated product of consecutive coherence functions $g_{\Gamma}(\ell, k)$ to decrease the probability of speaker changes within the SRO estimation interval.

Moreover, the SRO has to be zero or at least very close to zero such that the SNR-related term $\rho_{ij}(\ell, k)$, which is specified in (7.8), has a phase that is close to zero. This requirement can be achieved either via a multi-stage approach [2] with alternating SROs estimation and resampling based on the SROs estimates or in an online fashion where the SRO estimates are fed back into an upstream resampling before SRO estimation [107]. The DWACD method accounts for this requirements by a lightweight resampling without the need of a controller, as described later.

Next, the DWACD method is presented, The corresponding block diagram is shown in Fig. 7.1. In the following, the different steps shown in this block diagram will be successively explained. A core component of the online WACD method, that was presented in [OC6] and is extended to the DWACD method here, is the compensation for the integer part of the SRO-induced time shift between the signals which generally grows over time. In order to compensate for a time shift of ζ, the auxiliary signal

$$y_m^{(\zeta)}(n) = y_m(n - \zeta) \tag{7.10}$$

is defined as shifted version of the m-th microphone signal. Note that the calculation of the STFT $y_m^{(\zeta)}(\ell, k)$ can also be interpreted as calculating the STFT of $y_i(n)$ using an analysis window shifted by ζ samples. With the definition in (7.10), the coherence function with a compensation for a time shift of ζ is given by

$$\Gamma_{ij}^{(\zeta)}(\ell, k) = \frac{\Phi^{\mathrm{cmp}}_{y_i^{(\zeta)} y_j}(\ell, k)}{\sqrt{\Phi_{y_i^{(\zeta)} y_i^{(\zeta)}}(\ell, k) \cdot \Phi_{y_j y_j}(\ell, k)}}. \tag{7.11}$$

Additionally, the SRO-induced drift within the coherence estimation interval is compensated for via a lightweight resampling based on a multiplication with a complex-valued phase term:

$$\Phi^{\mathrm{cmp}}_{y_i^{(\zeta)} y_j}(\ell, k) = \frac{1}{L_{\mathrm{w}}} \cdot \sum_{\widetilde{\ell}=\ell}^{\ell+L_{\mathrm{w}}-1} y_i^{(\zeta)}\left(\widetilde{\ell}, k\right) \cdot y_j^{*}\left(\widetilde{\ell}, k\right) \cdot \exp\left(-\mathrm{j} \cdot \frac{2 \cdot \pi \cdot k}{N} \cdot \widehat{\varepsilon}_{ij}(\ell - 1) \cdot \widetilde{\ell} \cdot B\right), \tag{7.12}$$

where $\widehat{\varepsilon}_{ij}(\ell - 1)$ denotes the SRO estimate for the previous time frame. This lightweight resampling is introduced to drive the phase of the SNR-related weight $\rho_{ij}(\ell, k)$, which is

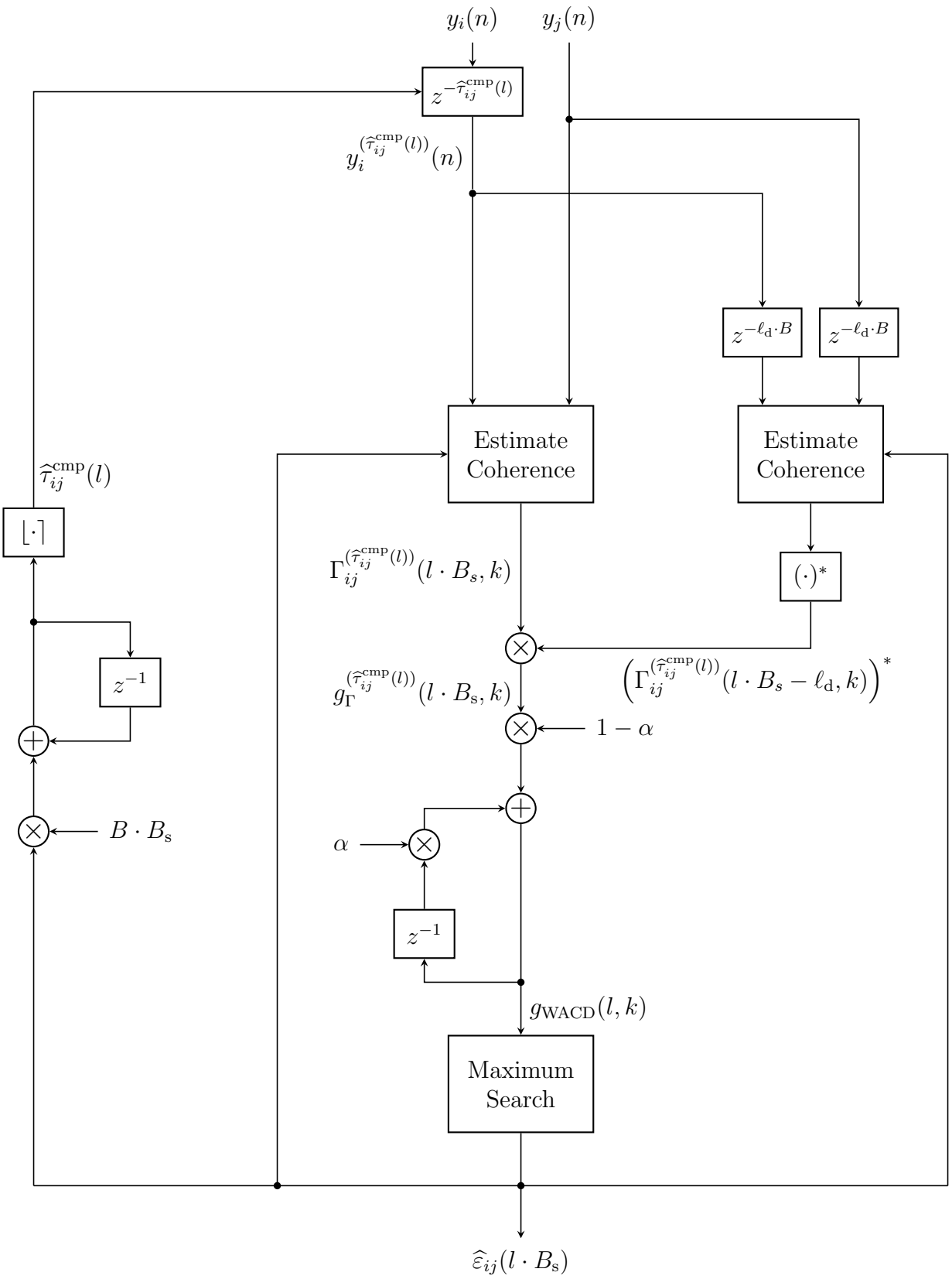


Figure 7.1: Block diagram of the DWACD method

specified in (7.8), towards zero, as discussed before. Note that the online WACD method does not specifically address this issue. By combining the compensation for the integer part of the SRO-induced time shift between the signals via shifting the analysis window of the STFT by a value of ζ and the lightweight resampling via a multiplication with a phase term during PSD estimation, a computationally lightweight online version of the multi-stage WACD method from [2] can be realized.

Since the SRO varies slowly, it is not estimated on a frame basis but rather every B_s time frames and is assumed to be constant during interval between two SRO estimates. Therefore, the complex-conjugated product of SRO-compensated coherence functions is also only is calculated every B_s time frames via

$$g_\Gamma^{(\tau_{ij}^{\mathrm{cmp}}(l))}(l \cdot B_\mathrm{s}, k) = \Gamma_{ij}^{(\tau_{ij}^{\mathrm{cmp}}(l))}(l \cdot B_\mathrm{s}, k) \cdot \left(\Gamma_{ij}^{(\tau_{ij}^{\mathrm{cmp}}(l))}(l \cdot B_\mathrm{s} - \ell_\mathrm{d}, k)\right)^*. \tag{7.13}$$

The integer part of the SRO-induced time shift between the signals, which is compensated for, is estimated based on the previous SRO estimates with

$$\tau_{ij}^{\mathrm{cmp}}(l) = \left\lfloor \sum_{\widetilde{l}=1}^{l-1} \widehat{\varepsilon}_{ij}\left(\widetilde{l} \cdot B_\mathrm{s}\right) \cdot B_\mathrm{s} \cdot B \right\rfloor. \tag{7.14}$$

In order to improve the accuracy of the SRO estimates, the complex-conjugated product of SRO-compensated coherence functions is averaged over time. While the offline and online WACD method assign the same weight to all time frames, the DWACD method utilizes an auto-regressive smoothing to introduce a temporal weighting. In this way, a higher weight is assigned to more recent time frames to enable a faster adaptation to large or fast changes of the SROs. The averaged complex-conjugated product of SRO-compensated coherence functions is updated very B_s time frames via

$$g_{\mathrm{WACD}}(l, k) = \alpha \cdot g_{\mathrm{WACD}}(l-1, k) + (1-\alpha) \cdot g_\Gamma^{(\tau_{ij}^{\mathrm{cmp}}(l))}(l \cdot B_\mathrm{s}, k), \tag{7.15}$$

with α being a smoothing factor with a value that it close to one and l denoting the SRO estimation index. In addition to that, a sound activity detection (SAD) is used to update the averaged complex-conjugated product of consecutive SRO-compensated coherence functions only if a coherent source is active within the estimation interval of both consecutive coherence functions.

Finally, the SRO is estimated based on the complex-conjugated product of consecutive SRO-compensated coherence functions $g_{\mathrm{WACD}}(l, k)$. The offline and online WACD method estimate the SRO based on the aggregated amplitude and phase information of the average complex-conjugated product of consecutive coherence functions via

$$\widehat{\epsilon}_{ij}^{\mathrm{WACD}}(l) = \frac{\epsilon_{\max}}{\pi} \cdot \angle\left(\sum_{k=K_{\min}}^{K_{\max}} |g_{\mathrm{WACD}}(l, k)| \cdot \exp\left(\frac{\mathrm{j} \cdot N \cdot \angle\left(g_{\mathrm{WACD}}(l, k)\right)}{2 \cdot \ell_\mathrm{d} \cdot B \cdot k \cdot \epsilon_{\max}}\right)\right). \tag{7.16}$$

Here, $K_{\max}$ denotes the largest frequency bin index that does not show a phase ambiguity for an assumed maximum SRO $\epsilon_{\max}$ [2], [3]. This way of SRO estimation does only make use of a fraction of the available frequency range. Hence, it is less robust to time frequency

bins with exceptionally large phase errors. However, such time frequency bins occur more frequently in the multi-speaker scenario with speaker changes where also shorter averaging periods, shorter segments and shorter temporal distances between the consecutive coherence function have to be utilized.

In the DWACD method a more robust approach, which also does not require to take into account the 2π-periodicity of the phase of the complex-conjugated product of SRO-compensated coherence functions $g_{\text{WACD}}(l,k)$, is employed for SRO estimation. To this end, the weighted average coherence drift $g_{\text{WACD}}(l,k)$ is interpreted as a GCPSD [111] with an SNR-related weighting of the frequencies. In this case, the SRO can be estimated based on the time lag

$$\kappa_{\max}(l) = \underset{\kappa}{\operatorname{argmax}} \left|\chi(l,\kappa)\right|, \tag{7.17}$$

that maximizes the GCC

$$\chi(l,\kappa) = \text{IFFT}\left(g_{\text{WACD}}(l,k)\right). \tag{7.18}$$

Here, IFFT$(\cdot)$ denotes the inverse fast Fourier transform (IFFT). To improve the accuracy of the SRO estimate, a golden section search is performed to find the non-integer time lag $\widetilde{\kappa}_{\max}(l)$ that maximizes $|\chi(l,\kappa)|$ in the interval $[\kappa_{\max}(l)-0.5, \kappa_{\max}(l)+0.5]$. From this,

$$\widehat{\varepsilon}_{ij}(l \cdot B_s) = -\frac{1}{\ell_{\text{d}} \cdot B} \cdot \widetilde{\kappa}_{\max}(l) \tag{7.19}$$

follows for the SRO estimate. In order to prevent an unwanted behavior due to a compensation for erroneous SRO estimates, a settling time is introduced, i.e., the SRO is only estimated and compensated for if $l \geq 40$ holds.

7.3 Evaluation

In the following, the proposed approaches to STO and SRO estimation are evaluated. Note that the presented results are taken from [OC1].

7.3.1 Dataset

In order to be able to control the STO and SRO such that ground-truth values are available, the simulated dataset that was presented in [OC1] is utilized for the evaluation of the proposed STO and SRO estimators. Recording a similar dataset instead of using simulations comes with an immense hardware effort.

The dataset consist of four different scenarios, whose properties are given in Tab. 7.1, with 100 examples each. This dataset was also used in a modified way for geometry calibration in [OC8]. Scenario-1 corresponds to the simplest scenario with time-constant SROs and a single speaker being always active at the same position. This scenario is extended by

Table 7.1: Overview over the scenarios employed in the evaluation of STO and SRO estimators

	Time-varying SROs	Speaker position changes	Speech pauses
Scenario-1			
Scenario-2	✓		
Scenario-3	✓	✓	✓
Scenario-4	✓	✓	

time-varying SROs in Scenario-2. In Scenario-3 and Scenario-4 speaker changes and, therefore, speaker positions changes are added, where in Scenario-3 there are speech pauses when the speaker changes.

For each example a room with a random width drawn from the uniform distribution $\mathcal{U}(6\,\mathrm{m}, 7\,\mathrm{m})$, a random length drawn from the uniform distribution $\mathcal{U}(5\,\mathrm{m}, 6\,\mathrm{m})$ and a fixed height of 3 m is simulated. The sound decay time of the rooms is T_{60}=300 ms. Each room is divided into four quadrants, each originating from a corner. One recording device is randomly placed in each quadrant, with a minimum spacing of 1 m being guaranteed between the recording devices. The speaker positions are randomly placed in the room. Both, the recording devices and the speakers, are placed at a height of 1.4 m with a minimum distance of 0.1 m to the closest wall.

Signals with a length of 5 min are created utilizing utterances from the TIMIT dataset [112]. For scenarios with a single fixed speaker position, utterances are concatenated until the target signal length is achieved. In scenarios with speaker changes up to four utterances are concatenated per speaker position. In order to achieve the target signal length of 5 min, new speaker positions are drawn until the target signal length is reached. Generally, this results in an unusually large amount of speaker positions for a meeting-like scenario under consideration in the other parts of this work. Note that the suitability of the DWACD method for SRO estimation in meeting scenarios was proven in Sec. 5.1 at hand of meeting recognition using distributed recording devices. The resulting large number of speaker positions can be justified via the fact that in realistic scenarios speakers might move over time. Therefore, the number of observed speaker positions grows over time. Optionally, speech pauses with a length between 0.5 s and 2 s are added when the speaker position changes in Scenario-3. The clean speech signals are convolved with room impulse responses (RIRs) that are simulated via the image source method using the implementation of [37]. Additionally, white Gaussian sensor noise is added. The SNR has an average value of 30 dB for a speaker-microphone distance of 3.2 m. The nominal sampling frequency of the microphone signals is f_s=16 kHz.

The simulated STOs are in the range of ±16,000 samples. Furthermore, SROs in the range of ±100 ppm are simulated. For the scenarios with time-varying SROs θ=0.001 and σ_OU=0.05 ppm are used to generate SRO trajectories via (2.25) which results in fluctuations with a standard deviation of the SRO of 1.25 ppm around $\mu_m^{(\infty)}$, which is drawn from the uniform distribution $\mathcal{U}(-100\,\mathrm{ppm}, 100\,\mathrm{ppm})$, in the steady state. The offset of the start value Δ_start of the SRO trajectories is randomly drawn from the uniform distribution

$\mathcal{U}(0\,\text{ppm}, 10\,\text{ppm})$. In order to generate signals with an SRO, the STFT-based resampling method from [79] is employed.

7.3.2 Parametrization

Without loss of generality, the microphone signal of the 0-th device is used as reference signal for SRO and STO estimation. In order to enable a coherent signal processing of the signals stemming from different devices, the STOs are minimized before SRO and STO estimation. For example, an initial STO minimization is required to achieve accurate SRO estimates, as discussed in [OC6]. In order to minimize the STOs, the integer time shift between the microphone signals is estimated during the first 20 s of speech activity based on the time-domain cross-correlation between the microphone signals. Afterwards, the STOs are compensated for by time-shifting the microphone signals.

The TDOA and distance estimates that are utilized for STO estimation are estimated from overlapping signal segments with length of 16,384 samples and a shift of 2048 samples. For TDOA estimation the generalized cross-correlation with phase transform (GCC-PhaT) algorithm [111] is employed which utilizes an STFT with a blackman window with a size of 16,384 samples and a shift 2048 samples. The distance estimates are gathered by the DNN-based distance estimator that was proposed in [OC5]. Here, the version with CDR and STFT as input features is employed.

For the DWACD method the parametrization that was proposed in [OC1] it utilized: L_w=16, ℓ_d=4, B_s=4 and α=0.95. Further, an STFT with a Blackman window of size N=4096 and a shift of B=512 is employed. In order to decide if speech is active, an energy-based voice activity detection (VAD) with a fixed threshold is used.

The online WACD and the DXCP-PhaT method are used as reference methods for SRO estimation. For both SRO estimators the parametrization that was proposed in the corresponding publications is utilized whereby slight modification were necessary to enable a fair comparison for the scenarios at hand. Since the online WACD method was applied to constant SROs in [OC6], it averages the complex-conjugated coherence products over all past time frames. Here, the complex-conjugated product of consecutive coherence functions is only averaged over the last 160 time frames in order to adapt to time-varying SROs. Moreover, an alternative parametrization is introduced for the DXCP-PhaT method, denoted as DXCP-PhaT$_8$ in the following, in order to enable a fair comparison in scenarios with speaker changes. The need for this new parametrization arises from the fact that the original DXCP-PhaT method uses a large temporal distance of approximately 5 s between the two signal segments used to calculate the secondary GCPSD. In the parametrization of DXCP-PhaT$_8$ this value was reduced to 8 time frames, which results in an approximate temporal distance of 1 s. This new parametrization can be seen as trade-off between robust SRO estimates in static scenarios and an improved ability to handle speaker changes.

7.3.3 Metrics

For the evaluation of the proposed STO and SRO estimators, the metrics that were defined in [OC1] are utilized. The accuracy of the STO estimates is measured by their absolute error (AE)

$$\left|\widehat{\tau}_{0m}^{\mathrm{STO}} - \tau_{0m}^{\mathrm{STO}}\right|. \tag{7.20}$$

In order to evaluate the quality of the SRO estimates, the root mean square error (RMSE) of the SRO estimates

$$\mathrm{RMSE}\left(\varepsilon_{0m}\right) = \sqrt{\frac{1}{L_{\mathrm{SRO}}} \cdot \sum_{l=0}^{L_{\mathrm{SRO}}-1} \left(\varepsilon_{0m}(l \cdot B_{\mathrm{s}}) - \widehat{\varepsilon}_{0m}(l \cdot B_{\mathrm{s}})\right)^2} \tag{7.21}$$

is employed. This can be seen as a natural extension of the RMSE of the SRO estimates, which is often reported when dealing with time-constant SROs, to SRO trajectories.

However, the remaining SRO-induced time shift after compensating for the estimated SROs is more important than the RMSE of the SRO. Thus, two SRO estimators might have a similar RMSE of the SRO but one of them shows a bias and the other is bias-free but shows larger fluctuations. In this case, the latter is preferred w.r.t. the accuracy of the resulting compensation for SROs. Even a small bias in the SRO estimates can cause an SRO-induced time shift between the signals that grows over time after compensation for the estimated SRO. In contrast, bias-free but fluctuating SRO estimates lead to a residual SRO-induced time shift between the signals that merely oscillates around zero. To account for this fact the RMSE of the corresponding SRO-induced time shift

$$\mathrm{RMSE}\left(\tau_{0m}^{\mathrm{SRO}}\right) = \sqrt{\frac{1}{L_{\mathrm{SRO}}} \cdot \sum_{l=0}^{L_{\mathrm{SRO}}-1} \left(\tau_{0m}^{\mathrm{SRO}}(l) - \widehat{\tau}_{0m}^{\mathrm{SRO}}(l)\right)^2} \tag{7.22}$$

is introduced with

$$\tau_{0m}^{\mathrm{SRO}}(l) = \sum_{\widetilde{l}=0}^{l} \varepsilon_{0m}\left(\widetilde{l} \cdot B_{\mathrm{s}}\right) \cdot B \cdot B_{\mathrm{s}} \text{ and } \widehat{\tau}_{0m}^{\mathrm{SRO}}(l) = \sum_{\widetilde{l}=0}^{l} \widehat{\varepsilon}_{0m}\left(\widetilde{l} \cdot B_{\mathrm{s}}\right) \cdot B \cdot B_{\mathrm{s}}. \tag{7.23}$$

For both, the SROs and the SRO-induced time shifts, the average RMSE over all examples and devices is reported. Further, the maximum RMSE over all examples and devices is reported to get insights into the worst-case qualitiy of the synchronization resulting from the SRO estimates.

7.3.4 STO estimation

Figure 7.2 shows the distribution of the AE $\left|\widehat{\tau}_{0m}^{\mathrm{STO}} - \tau_{0m}^{\mathrm{STO}}\right|$ of the STO estimates for Scenario-3 as a function of the signal length used for STO estimation. Here, larger signal lengths used for STO estimation come with the advantage that speakers at more positions are

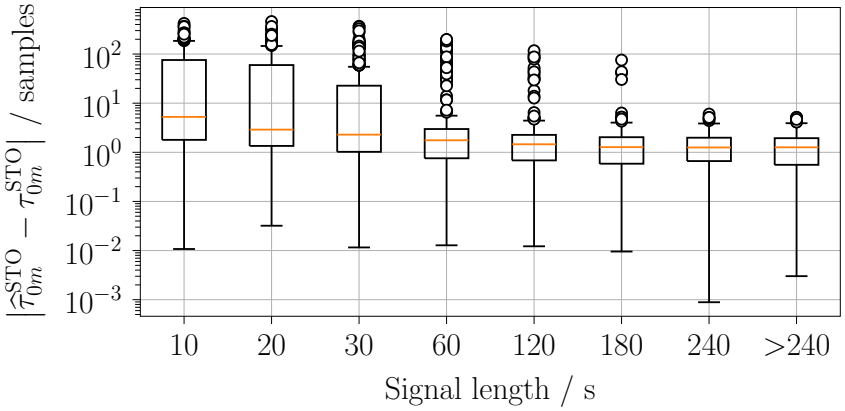


Figure 7.2: STO estimation performance as a function of the length of the signals used for STO estimation (*Adapted from* [OC1])

observed which usually should improve the quality of the solution of the LS problem for STO estimation that is embedded into a RANSAC method. At the same time, a larger signal length used for STO estimation means a larger latency when applying the STO in online applications.

It can be seen that the STO errors get smaller with growing signal lengths, as expected. Despite some outliers, an STO error smaller than 10 samples can be achieved in most cases if a signal lasting 1 min or longer is used for STO estimation. Converting the STO error to a TDOA-based localization error, an error of 10 samples would result in a localization error of approximately 20 cm which is sufficient for most downstream tasks. Using signals of 4 min duration or longer for STO estimation, the outliers disappear completely.

It can be concluded that the proposed STO estimator in (7.3) that maintains the physical TDOFs between the signals is accurate enough for typical downstream tasks, like TDOA-based localization. However, quite long signals are needed so that the required accuracy of the STO estimates will only be guaranteed in offline applications where latency does not matter. Note that much shorter signal segments are sufficient when the physical TDOFs do not need to be mainlined. In this case, a coarse minimization of the STOs based on the time lag which maximizes the correlation between signal segments of a few seconds can be used.

7.3.5 SRO estimation

A comparison of the proposed DWACD method to the online WACD and the DXCP-PhaT method for the four considered scenarios which were introduced in Tab. 7.1 is shown in Tab. 7.2. It becomes obvious that all SRO estimators show a decent performance for Scenario-1 with time-constant SROs and a single speaker being always active at the same position. Furthermore, it can be seen that the performance of the DXCP-PhaT$_8$ method significantly degrades compared to the DXCP-PhaT method which reflects the tuning of many SRO estimators for unrealistic scenarios, e.g., with time-constant SROs. Note that the DWACD method shows a quite high RMSE of the SRO RMSE (ε_{0m}) compared to the

Table 7.2: Comparison of SRO estimators for scenarios with different degrees of dynamism (*Adapted from* [OC1])

	Method	Average RMSE (ε_{0m}) / ppm	Average RMSE $\left(\tau_{0m}^{\mathrm{SRO}}\right)$ / samples	Maximum RMSE $\left(\tau_{0m}^{\mathrm{SRO}}\right)$ / samples
Scenario-1	Online WACD	0.21	0.14	0.50
	DXCP-PhaT	0.15	0.36	0.68
	DXCP-PhaT$_8$	0.69	1.73	3.65
	DWACD	0.40	0.15	0.50
Scenario-2	Online WACD	0.63	0.73	2.09
	DXCP-PhaT	0.66	0.97	2.73
	DXCP-PhaT$_8$	0.95	1.83	4.66
	DWACD	0.51	0.27	1.04
Scenario-3	Online WACD	2.98	6.04	21.00
	DXCP-PhaT	28.96	21.84	161.54
	DXCP-PhaT$_8$	1.31	2.70	7.76
	DWACD	0.57	0.32	1.20
Scenario-4	Online WACD	2.80	3.25	10.96
	DXCP-PhaT	22.42	16.61	160.49
	DXCP-PhaT$_8$	1.28	2.81	6.93
	DWACD	0.64	0.32	1.10

online WACD and the DXCP-PhaT method but the RMSE of the SRO-induced time shift RMSE $\left(\tau_{0m}^{\mathrm{SRO}}\right)$ as more direct measure of the resulting synchronization accuracy is on par with the online WACD method, which performs best. This shows that the RMSE of the SRO RMSE (ε_{0m}) is only conditionally meaningful since it does not differentiate between bias-free, fluctuating SRO estimates and biased SRO estimates which are significantly more detrimental for the resulting synchronization.

Introducing time-varying SROs in Scenario-2, leads to a degradation in performance for all estimators where the performance of the DWACD method deteriorates significantly less compared to the performance of the other estimators. However, the RMSE of the SRO-induced time shift RMSE $\left(\tau_{0m}^{\mathrm{SRO}}\right)$ indicates that all estimators enable an acceptable accuracy for the synchronization. Matters are different in Scenario-3 and Scenario-4 with additional speaker changes. In these scenarios the error of the SRO estimates from the online WACD method and both considered versions of the DXCP-PhaT method grow significantly which is particularly evident in the RMSE of the SRO-induced time shifts RMSE $\left(\tau_{0m}^{\mathrm{SRO}}\right)$. In contrast to that, the performance of the DWACD method is nearly the same as for a fixed speaker position. The large errors reported for the original parametrization of the DXCP-PhaT method might be mainly attributed to cases in which the secondary GCPSD function is often calculated from signal segments corresponding to two different speaker positions. While the DWACD method can use the SAD to skip segments, when the speaker position changes, in Scenario-3 with speech pauses, there are no speech pauses when the speaker position changes

in Scenario-4. Despite that, the resulting degradation in performance of the DWACD method is negligibly small.

Table 7.3: Dependence of the average RMSE of the SRO estimates RMSE (ε_{0m}) / ppm on the standard deviation σ_ε of the ground-truth SRO trajectories (*Adapted from* [OC1]). For this purpose, Scenario-2 is considered.

σ_ε / ppm	Online WACD	DXCP-PhaT	DWACD
0 - 1	0.55	0.54	0.49
1 - 2	0.60	0.61	0.51
2 - 3	0.63	0.69	0.51
3 - 4	0.71	0.80	0.54

Table 7.3 presents the dependence of the different SRO estimators' performance on the standard deviation σ_ε of the ground-truth SRO trajectories based on Scenario-2. To this end, the examples are divided into subsets w.r.t. the standard deviation of the ground-truth SRO trajectories. Afterwards, the average RMSE of the SROs RMSE (ε_{0m}) is calculated for each subset separately. A small standard deviation σ_ε of a ground-truth SRO trajectory corresponds to the steady-state case. In contrast to this, a large standard deviation σ_ε indicates a transient behavior of the SRO.

It can be seen that the SRO estimation error increases for the online WACD method and the DXCP-PhaT method as the standard deviation σ_ε of the SRO trajectories grows. This might be attributed to the fact that both methods were initially designed for applications with time-constant SROs and cannot immediately follow fast and large changes of the SRO. In contrast, the SRO error is nearly constant for the DWACD method.

7.4 Summary

In this chapter, an approach to STO estimation was presented that exploits DNN-based distance estimates to separate STOs from the physical TDOFs. This allows to compensate for STOs while preserving the spatial information about the positions of the microphones and the speakers contained in the TDOFs. However, many algorithms, such as MVDR beamforming, require only a coarse minimization of the STOs and the proposed STO estimation approach entails a considerably higher computational effort compared to simpler synchronization methods. Therefore, a correlation-based coarse alignment of the signals is preferable for downstream tasks that do not rely on mapping TDOAs to physical positions.

Furthermore, the DWACD method for SRO estimation in the presence of time-varying SROs and changing speaker positions was presented. While the performance of previously proposed online SRO estimators degrades under such dynamic conditions, especially in scenarios involving speaker position changes, the DWACD method can effectively handle both effects. This is achieved by extending the previously proposed online WACD method by utilizing shorter signal segments to estimate the phase-drift of the coherence, which is combined with a more robust GCC-based approach to estimating the SRO based on this phase-drift. In

static scenarios with time-invariant SROs and a single coherent source at a fixed position, all considered SRO estimators achieve comparable and reliable results.

8 Summary and future work

This thesis answered the question whether a trade-off exists between choosing a large sample size for spatial covariance matrix (SCM) estimation in order to mitigate the negative effects of SCM estimation from a finite sample size on beamforming and choosing a small sample size to diminish the negative effects of sampling rate offsets (SROs) on beamforming. In order to answer this question, the extraction of a target speaker's signal from a noisy and reverberant speech mixture via mask-based, block-wise minimum variance distortionless response (MVDR) beamforming was considered.

First, the effect of SCMs that are estimated from a finite sample size was considered in isolation from SROs. To this end, a statistical model for the extraction of a target speaker's signal from a noisy and reverberant speech mixture was introduced in Sec. 4.1. Further, it was shown in Sec. 4.2 that the non-tractable probability distribution of the mask-based SCM estimates can be approximated by a Wishart distribution. Based on this approximation of the probability distribution of the SCM estimates, a closed-form approximation of the signal-to-distortion ratio (SDR) at the beamformer output as a function of the sample size used for SCM estimation was derived in Sec. 4.3. The evaluation of the closed-form approximation of the enhancement quality in terms of the SDR at the beamformer output in Sec. 4.5 revealed that the largest possible sample size should be used for SCM estimation in practice if the sampling processes of the microphone signals are synchronized.

Additionally, the analysis of the closed-form approximation in Sec. 4.4 has shown that the behavior of the beamforming performance with growing sample size used for SCM estimation results from the superposition of three effects. In general, the variance of the SCM estimates becomes smaller as the sample size used for SCM estimation increases, resulting in an improvement of the beamformer's performance. Moreover, the statistical dependence between the beamformer coefficients and the signals to which they are applied in block-wise beamforming leads to an improved suppression of the interfering sources. This effect diminishes with growing sample size used for SCM estimation. Considering the combination of the two effects, mentioned above, the latter effect usually dominates the behavior of the beamforming performance. This results in a degradation of the beamforming performance as the sample size used for SCM estimation grows. However, the inevitable leakage of the target speaker's signals into the mask-based interference-SCM estimates degrades the suppression of the interference and results in a suppression of the target speaker. In most cases, this effect dominates the overall behavior of the beamforming performance. Due to the fact that the leakage of the target speaker's signals into the interference-SCM estimates generally diminishes with growing sample size for SCM estimation, all in all, the performance of an MVDR beamformer improves as the sample size used for SCM estimation increases.

In contrast to that, the detrimental effect of SROs on beamforming becomes stronger with growing sample size used for SCM estimation, as discussed in Sec. 5.3. It was demonstrated that this degradation of the performance of a beamformer is not only caused by a distortion of the SCM estimates' amplitude which already has been reported in literature before but also results from a mismatch between the phase of the SCM estimates and the instantaneous phase of the signals to which the beamformer is applied.

Finally, the interplay of the effect of SCM estimation from a finite sample size and the effect of SROs on the beamforming performance was considered by extending the closed-form approximation of the SDR at the beamformer output in Sec. 6.3 to the case with SROs being present. The evaluation of the extended closed-form approximation shows that a trade-off for the choice of the sample size used for SCM estimation exists for small and moderate SROs up to $\pm 25\,\text{ppm}$ at cost of a marginal loss in the performance of the beamformer. This renders an accurate compensation for SROs unnecessary in cases with small SROs, e.g., when using multiple devices of the same type. However, large SROs always lead to an unacceptably high drop in the beamformer's performance even when choosing the optimal sample size for SCM estimation. Hence, an accurate compensation for SROs is inevitable to guarantee the best possible beamforming performance especially when different types of devices are employed so that the occurrence of large SROs cannot be excluded.

Due to the fact that generally a compensation for SROs is necessary to guarantee the best possible performance of a beamformer, SRO estimation was considered in Sec. 7.2.2. While previous works often assumed a static scenario with a single source being always active at the same positions and time-constant SROs, here an SRO estimator for more realistic scenarios with time-varying SROs and speaker position changes was considered. The experiments in Sec. 7.3 show that the performance of online SRO estimators which were developed for scenarios with time-constant SROs and a single source at a fixed position in previous works significantly degrades in dynamic scenarios with time-varying SROs and speaker position changes. In contrast, the proposed SRO estimator exhibits similar performance for dynamic scenarios as for the static scenario.

Moreover, the effect of sampling time offsets (STOs) on MVDR beamforming was discussed. If the STOs are much smaller than the size of the analysis window of the short-time Fourier transform (STFT), the effect of STOs on the microphone signals can be modeled via a multiplication with a complex-valued phase term. By extending the closed-form approximation of the SDR at the beamformer output to the case with STOs based on this model, it was proven in Sec. 6.1 that STOs do not affect the beamforming performance at all in this case. However, it was also discussed that modeling the effect of STOs via a multiplication with a complex-valued phase term is only sufficient if the STOs are much smaller than the size of the STFT analysis window. The experimental analysis of MVDR beamforming in the presence of STOs in Sec. 6.2 revealed that the effect of STOs on the beamforming performance is only negligibly small for STOs in the range $\pm 25\,\text{samples}$. This precision can be achieved by compensating for the coarse time shift between the signals which can be estimated via simple correlation methods.

Future work

The approximation of the probability distribution of the SCM estimates by an equivalent Wishart distribution corresponds to the step of the derivation of the closed-form approximation of the SDR at the beamformer output which exhibits the largest error. Thus, the improvement of the approximation of the SCM estimates' probability distribution would be a next step to improve the closed-form approximation of the SDR at the beamformer output. However, this extension should be performed while maintaining the Wishart distribution in this case so that the effort for the derivation of the closed-form approximation of the SDR at the beamformer output can be kept at an acceptable level. Alternatively, the resulting inverse of the SCM estimates involved in the calculation of the beamformer coefficients could be directly approximated by an equivalent inverse Wishart matrix.

Further, new ways to improve the performance of block-wise beamforming can be derived from the insights into the reasons for the degradation of the beamforming performance due to a finite sample size used for SCM estimation. One possible direction of research would be the improvement of the masks that are used for SCM estimation to reduce the detrimental leakage of the target speaker's signals into the interference-SCM estimates. For instance, this could be done by setting the interference masks to zero if their likelihood to dominate a time frequency bin is low. By reducing the leakage of the target speaker's signals into the interference-SCM estimates, the improved suppression of the interference for small block sizes, from which an improved performance of the beamformer follows, can be fully exploited. In this way the negative effects of SROs on beamforming can be mitigated using a small sample size for SCM estimation.

Moreover, improving the robustness of SRO estimation under real-world conditions is of larger importance than further improvement of SRO estimation in artificial scenarios with a single static source position and time-constant SROs where SRO estimation can be seen as solved. For example, many open questions remain like how to deal with small movements of the sources, like head movements, or even larger movements of the sources.

A Appendix

A.1 Derivation of the variance of the elements of a Wishart matrix

Let $\mathbf{\Psi}$ be Wishart distributed with $\mathbf{\Psi} \sim \mathcal{W}_D(A, \mathbf{\Sigma})$. The variance of i-th row and j-th column element of the Wishart matrix $\mathbf{\Psi}$ is given by

$$\operatorname{var}(\Psi_{ij}) = \mathbb{E}\big[|\Psi_{ij}|^2\big] - |\mathbb{E}[\Psi_{ij}]|^2 . \tag{A.1}$$

With (2.52), it follows that

$$|\mathbb{E}[\Psi_{ij}]|^2 = A^2 \cdot |\Sigma_{ij}|^2 . \tag{A.2}$$

Further, it follows from (2.50) and the independent and identically distributed (i.i.d.) assumption for the samples $\boldsymbol{\psi}(a)$ that

$$\begin{aligned}
\mathbb{E}\big[|\Psi_{ij}|^2\big] &= \mathbb{E}\left[\left|\sum_{a=0}^{A-1} \psi_i(a) \cdot \psi_j^*(a)\right|^2\right] \\
&= \mathbb{E}\left[\left(\sum_{a=0}^{A-1} \psi_i(a) \cdot \psi_j^*(a)\right) \cdot \left(\sum_{\widetilde{a}=0}^{A-1} \psi_i(\widetilde{a}) \cdot \psi_j^*(\widetilde{a})\right)^*\right] \\
&= \mathbb{E}\left[\sum_{a=0}^{A-1} \psi_i(a) \cdot \psi_j^*(a) \cdot \psi_i^*(a) \cdot \psi_j(a)\right] + \mathbb{E}\left[\sum_{a=0}^{A-1} \sum_{\substack{\widetilde{a}=0 \\ \widetilde{a} \neq a}}^{A-1} \psi_i(a) \cdot \psi_j^*(a) \cdot \psi_i^*(\widetilde{a}) \cdot \psi_j(\widetilde{a})\right] \\
&= \mathbb{E}\left[\sum_{a=0}^{A-1} \psi_i(a) \cdot \psi_j^*(a) \cdot \psi_i^*(a) \cdot \psi_j(a)\right] + \sum_{a=0}^{A-1} \sum_{\substack{\widetilde{a}=0 \\ \widetilde{a} \neq a}}^{A-1} \Sigma_{ij} \cdot \Sigma_{ij}^* \\
&= \sum_{a=0}^{A-1} \mathbb{E}\big[\psi_i(a) \cdot \psi_j^*(a) \cdot \psi_i^*(a) \cdot \psi_j(a)\big] + A \cdot (A-1) \cdot |\Sigma_{ij}|^2 .
\end{aligned} \tag{A.3}$$

Utilizing Isserlis' theorem [113] and the fact that the Gaussians involved in the calculation of the Wishart matrix are circular, gives

$$\begin{aligned}
\mathbb{E}\big[\psi_i(a) \cdot \psi_j^*(a) \cdot \psi_i^*(a) \cdot \psi_j(a)\big] &= \mathbb{E}\big[|\psi_i(a)|^2\big] \cdot \mathbb{E}\big[|\psi_j(a)|^2\big] + \big|\mathbb{E}\big[\psi_i(a) \cdot \psi_j^*(a)\big]\big|^2 \\
&= \Sigma_{ii} \cdot \Sigma_{jj} + |\Sigma_{ij}|^2 .
\end{aligned} \tag{A.4}$$

Inserting (A.4) into (A.3), leads to

$$\begin{aligned}\mathbb{E}\left[|\Psi_{ij}|^2\right] &= \sum_{a=0}^{A-1}\left(\Sigma_{ii}\cdot\Sigma_{jj}+|\Sigma_{ij}|^2\right)+A\cdot(A-1)\cdot|\Sigma_{ij}|^2\\ &= A\cdot\Sigma_{ii}\cdot\Sigma_{jj}+A^2\cdot|\Sigma_{ij}|^2\,.\end{aligned}\tag{A.5}$$

From this it follows that

$$\mathrm{var}(\Psi_{ij}) = A\cdot\Sigma_{ii}\cdot\Sigma_{jj}+A^2\cdot|\Sigma_{ij}|^2-A^2\cdot|\Sigma_{ij}|^2 = A\cdot\Sigma_{ii}\cdot\Sigma_{jj}.\tag{A.6}$$

A.2 Derivation of the variance of the elements of an SCM estimate

Given (4.8), the estimate of the interference spatial covariance matrix (SCM) is assumed to be

$$\widehat{\mathbf{R}}_{\mathrm{int}} = \sum_{\ell=0}^{L-1}\widetilde{\mathbf{y}}(\ell)\cdot\widetilde{\mathbf{y}}^{\mathrm{H}}(\ell)\,,\tag{A.7}$$

$$\text{with } \widetilde{\mathbf{y}}(\ell)\sim\mathcal{N}\left(0,\widetilde{\sigma}_0^2(\ell)\cdot\mathbf{R}_0+\widetilde{\sigma}_1^2(\ell)\cdot\mathbf{R}_1+\widetilde{\sigma}_\nu^2(\ell)\cdot\mathbf{R}_\nu\right).\tag{A.8}$$

The variance of i-th row and j-th column element of the SCM estimate $\widehat{\mathbf{R}}_{\mathrm{int}}$ is given by

$$\mathrm{var}\left(\left(\widehat{\mathbf{R}}_{\mathrm{int}}\right)_{ij}\right) = \mathbb{E}\left[\left|\left(\widehat{\mathbf{R}}_{\mathrm{int}}\right)_{ij}\right|^2\right]-\left|\mathbb{E}\left[\left(\widehat{\mathbf{R}}_{\mathrm{int}}\right)_{ij}\right]\right|^2.\tag{A.9}$$

With (4.11) it follows that

$$\left|\mathbb{E}\left[\left(\widehat{\mathbf{R}}_{\mathrm{int}}\right)_{ij}\right]\right|^2 = \left|\sum_{\ell=0}^{L-1}\widetilde{R}_{ij}(\ell)\right|^2,\tag{A.10}$$

with $\widetilde{\mathbf{R}}(\ell)=\widetilde{\sigma}_0^2(\ell)\cdot\mathbf{R}_0+\widetilde{\sigma}_1^2(\ell)\cdot\mathbf{R}_1+\widetilde{\sigma}_\nu^2(\ell)\cdot\mathbf{R}_\nu$.

Further, it follows with (A.7) that

$$\begin{aligned}
&\mathbb{E}\left[\left|\left(\widehat{\mathbf{R}}_{\text{int}}\right)_{ij}\right|^2\right] \\
=&\mathbb{E}\left[\left|\sum_{\ell=0}^{L-1}\widetilde{y}_i(\ell)\cdot\widetilde{y}_j(\ell)\right|^2\right] \\
=&\mathbb{E}\left[\left(\sum_{\ell=0}^{L-1}\widetilde{y}_i(\ell)\cdot\widetilde{y}_j^*(\ell)\right)\cdot\left(\sum_{\widetilde{\ell}=0}^{L-1}\widetilde{y}_i\left(\widetilde{\ell}\right)\cdot\widetilde{y}_j^*\left(\widetilde{\ell}\right)\right)^*\right] \\
=&\mathbb{E}\left[\sum_{\ell=0}^{L-1}\widetilde{y}_i(\ell)\cdot\widetilde{y}_j^*(\ell)\cdot\widetilde{y}_i^*(\ell)\cdot\widetilde{y}_j(\ell)\right]+\mathbb{E}\left[\sum_{\ell=0}^{L-1}\sum_{\substack{\widetilde{\ell}=0\\ \widetilde{\ell}\neq\ell}}^{L-1}\widetilde{y}_i(\ell)\cdot\widetilde{y}_j^*(\ell)\cdot\widetilde{y}_i^*\left(\widetilde{\ell}\right)\cdot\widetilde{y}_j\left(\widetilde{\ell}\right)\right] \\
=&\sum_{\ell=0}^{L-1}\mathbb{E}\left[\widetilde{y}_i(\ell)\cdot\widetilde{y}_j^*(\ell)\cdot\widetilde{y}_i^*(\ell)\cdot\widetilde{y}_j(\ell)\right]+\sum_{\ell=0}^{L-1}\sum_{\substack{\widetilde{\ell}=0\\ \widetilde{\ell}\neq\ell}}^{L-1}\widetilde{R}_{ij}(\ell)\cdot\widetilde{R}_{ij}^*\left(\widetilde{\ell}\right). \qquad \text{(A.11)}
\end{aligned}$$

Utilizing Isserlis' theorem [113] and the fact that the Gaussians involved in the calculation of the interference-SCM estimate are circular, gives

$$\begin{aligned}
\mathbb{E}\left[\widetilde{y}_i(\ell)\cdot\widetilde{y}_j^*(\ell)\cdot\widetilde{y}_i^*(\ell)\cdot\widetilde{y}_j(\ell)\right] &= \mathbb{E}\left[|\widetilde{y}_i(\ell)|^2\right]\cdot\mathbb{E}\left[|\widetilde{y}_j(\ell)|^2\right]+\left|\mathbb{E}\left[\widetilde{y}_i(\ell)\cdot\widetilde{y}_j^*(\ell)\right]\right|^2 \\
&= \widetilde{R}_{ii}(\ell)\cdot\widetilde{R}_{jj}(\ell)+\left|\widetilde{R}_{ij}(\ell)\right|^2. \qquad \text{(A.12)}
\end{aligned}$$

Inserting (A.12) into (A.11), leads to

$$\begin{aligned}
\mathbb{E}\left[\left|\left(\widehat{\mathbf{R}}_{\text{int}}\right)_{ij}\right|^2\right] &= \sum_{\ell=0}^{L-1}\left(\widetilde{R}_{ii}(\ell)\cdot\widetilde{R}_{jj}(\ell)+\left|\widetilde{R}_{ij}(\ell)\right|^2\right)+\sum_{\ell=0}^{L-1}\sum_{\substack{\widetilde{\ell}=0\\ \widetilde{\ell}\neq\ell}}^{L-1}\widetilde{R}_{ij}(\ell)\cdot\widetilde{R}_{ij}^*\left(\widetilde{\ell}\right) \\
&= \sum_{\ell=0}^{L-1}\widetilde{R}_{ii}(\ell)\cdot\widetilde{R}_{jj}(\ell)+\left|\sum_{\ell=0}^{L-1}\widetilde{R}_{ij}(\ell)\right|^2. \qquad \text{(A.13)}
\end{aligned}$$

From this it follows that

$$\begin{aligned}
\operatorname{var}\left(\left(\widehat{\mathbf{R}}_{\text{int}}\right)_{ij}\right) &= \sum_{\ell=0}^{L-1}\widetilde{R}_{ii}(\ell)\cdot\widetilde{R}_{jj}(\ell)+\left|\sum_{\ell=0}^{L-1}\widetilde{R}_{ij}(\ell)\right|^2-\left|\sum_{\ell=0}^{L-1}\widetilde{R}_{ij}(\ell)\right|^2 \\
&= \sum_{\ell=0}^{L-1}\widetilde{R}_{ii}(\ell)\cdot\widetilde{R}_{jj}(\ell). \qquad \text{(A.14)}
\end{aligned}$$

A.3 Convergence of the off-diagonal elements of the SCM estimates

In the following, the ratio of the variance and the squared absolute value of the expected value of the off-diagonal elements of the interference-SCM estimates

$$\frac{\operatorname{var}\left(\left(\widehat{\mathbf{R}}_{\text{int}}\right)_{ij}\right)}{\left|\mathbb{E}\left[\left(\widehat{\mathbf{R}}_{\text{int}}\right)_{ij}\right]\right|^2} = \frac{\sum_{\ell=0}^{L-1}\widetilde{R}_{ii}(\ell)\cdot\widetilde{R}_{jj}(\ell)}{\left|\sum_{\ell=0}^{L-1}\widetilde{R}_{ij}(\ell)\right|^2} = \frac{\sum_{\ell=0}^{L-1}\widetilde{R}_{ii}(\ell)\cdot\widetilde{R}_{jj}(\ell)}{\sum_{\ell=0}^{L-1}\sum_{\tilde{\ell}=0}^{L-1}\widetilde{R}_{ij}(\ell)\cdot\widetilde{R}_{ij}^{*}\left(\tilde{\ell}\right)} \tag{A.15}$$

is considered, with $\widetilde{R}_{ij}(\ell) = \widetilde{\sigma}_0^2(\ell)\cdot(\mathbf{R}_0)_{ij} + \widetilde{\sigma}_1^2(\ell)\cdot(\mathbf{R}_1)_{ij} + \widetilde{\sigma}_\nu^2(\ell)\cdot(\mathbf{R}_\nu)_{ij}$. The numerator of (A.15) is a sum of non-negative, bounded terms and, thus, grows at most linearly with L, which means it is $\mathcal{O}(L)$ as $L\to\infty$ in Landau notation. Here, $\mathcal{O}(\cdot)$ denotes an upper bound on the asymptotic growth of the given function Assuming that the interfering speaker typically dominates $\widetilde{R}_{ij}(\ell)$ so that $\widetilde{\sigma}_1^2(\ell) \gg \widetilde{\sigma}_0^2(\ell)$ and $\widetilde{\sigma}_\nu^2(\ell) \gg \widetilde{\sigma}_0^2(\ell)$ holds for the vast majority of time frames, the pathological case that the weighted sum of all complex-valued covariance terms in the denominator of (A.15) converges to a value of zero is very unlikely. The denominator of (A.15) rather scales at least with L^2 as $L\to\infty$ in this case. Hence, (A.15) scales with $\mathcal{O}(\frac{1}{L})$ as $L\to\infty$. This means that the off-diagonal elements of the interference-SCM estimates converge towards their expected value with growing sample size.

A.4 Reformulation of the beamformer output for time frames that are dominated by a single source

The starting point for the reformulation of the interfering speaker's signal at the beamformer output for time frames that are dominated by the interfering speaker's signal is (4.45) after applying the Shermann-Morrison formula to the interference-SCM estimate:

$$\mathbf{w}^{\mathrm{H}}\cdot\mathbf{x}_1(\ell) = \widehat{d}_0\cdot\frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\left(\widehat{\mathbf{R}}_{\text{int},\setminus\ell}^{-1} - \frac{\gamma_{\text{int}}(\ell)\cdot\widehat{\mathbf{R}}_{\text{int},\setminus\ell}^{-1}\cdot\mathbf{y}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\text{int},\setminus\ell}^{-1}}{1+\gamma_{\text{int}}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\text{int},\setminus\ell}^{-1}\cdot\mathbf{y}(\ell)}\right)}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\left(\widehat{\mathbf{R}}_{\text{int},\setminus\ell}^{-1} - \frac{\gamma_{\text{int}}(\ell)\cdot\widehat{\mathbf{R}}_{\text{int},\setminus\ell}^{-1}\cdot\mathbf{y}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\text{int},\setminus\ell}^{-1}}{1+\gamma_{\text{int}}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\text{int},\setminus\ell}^{-1}\cdot\mathbf{y}(\ell)}\right)\cdot\widehat{\mathbf{d}}}\cdot\mathbf{x}_1(\ell). \tag{A.16}$$

Next, it is assumed that $\mathbf{x}_1(\ell)$ dominates the currently considered time frequency bin, i.e., the observed microphone signals can be approximated via $\mathbf{y}(\ell)\approx\mathbf{x}_1(\ell)$. Note that only selected incidents of $\mathbf{y}(\ell)$ in (A.16) are approximated as $\mathbf{y}(\ell)\approx\mathbf{x}_1(\ell)$ to account as much as possible for the contribution of the other sources to the interference-SCM estimate:

$$
\begin{aligned}
&\mathbf{w}^{\mathrm{H}} \cdot \mathbf{x}_1(\ell) \\
\approx &\widehat{d}_0 \cdot \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \left(\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} - \frac{\gamma_{\mathrm{int}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}}{1+\gamma_{\mathrm{int}}(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell)} \right)}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \left(\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} - \frac{\gamma_{\mathrm{int}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}}{1+\gamma_{\mathrm{int}}(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell)} \right) \cdot \widehat{\mathbf{d}}} \cdot \mathbf{x}_1(\ell) \\
= &\widehat{d}_0 \cdot \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) \cdot \left(1 - \frac{\gamma_{\mathrm{int}}(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell)}{1+\gamma_{\mathrm{int}}(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell)} \right)}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \left(\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} - \frac{\gamma_{\mathrm{int}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}}{1+\gamma_{\mathrm{int}}(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell)} \right) \cdot \widehat{\mathbf{d}}} \\
= &\widehat{d}_0 \cdot \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) \cdot \frac{1}{1+\gamma_{\mathrm{int}}(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell)}}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \left(\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} - \frac{\gamma_{\mathrm{int}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}}{1+\gamma_{\mathrm{int}}(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell)} \right) \cdot \widehat{\mathbf{d}}} \\
= &\frac{\widehat{d}_0 \cdot \widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell)}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}} \cdot \left(1 + \gamma_{\mathrm{int}}(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) \right) - \gamma_{\mathrm{int}}(\ell) \cdot \widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}} \\
= &\frac{\widehat{d}_0 \cdot \widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) \cdot \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}} \cdot \left(1 + \gamma_{\mathrm{int}}(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) \right) - \gamma_{\mathrm{int}}(\ell) \cdot \widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}} \\
= &\frac{\widehat{d}_0 \cdot \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}} \cdot \mathbf{x}_1(\ell)}{1 + \gamma_{\mathrm{int}}(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) - \frac{\gamma_{\mathrm{int}}(\ell) \cdot \widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}} \\
= &\frac{\widehat{d}_0 \cdot \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1}}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}} \cdot \mathbf{x}_1(\ell)}{1 + \gamma_{\mathrm{int}}(\ell) \cdot \left(\mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) - \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \mathbf{x}_1(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}} \right)}.
\end{aligned}
\tag{A.17}
$$

A.5 Expected value of the denominator accounting for the statistical dependence of the beamformer coefficients and the signals

In the following, the expected value $\mathbb{E}\left[|\delta(\mathbf{x})|^2\right]$ of the denominator in (4.59) will be derived for the case that $\mathbf{x}$ dominates the considered time frame with

$$
\mathbb{E}\left[|\delta(\mathbf{x})|^2\right] = \mathbb{E}\left[\left| \frac{\eta(\mathbf{x})}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}}} \right|^2 \right] \approx \frac{\mathbb{E}\left[|\eta(\mathbf{x})|^2\right]}{\mathbb{E}\left[\left| \widehat{\mathbf{d}}^{\mathrm{H}} \cdot \widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}^{-1} \cdot \widehat{\mathbf{d}} \right|^2 \right]}. \tag{A.18}
$$

The term in the numerator can be factored as follows:

$$
\begin{aligned}
|\eta(\mathbf{x})|^2 &= \left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\left(\left(1+\gamma_{\mathrm{int}}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\right)\cdot\mathbf{I}-\gamma_{\mathrm{int}}(\ell)\cdot\mathbf{x}\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\right)\cdot\widehat{\mathbf{d}}\right|^2 \\
&= \left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\left(1+\gamma_{\mathrm{int}}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\right)-\gamma_{\mathrm{int}}(\ell)\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2 \\
&= \left(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\right)^2+2\cdot\gamma_{\mathrm{int}}(\ell)\cdot\mathrm{Re}\left\{\left(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\right)^2\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\right\} \\
&\quad-2\cdot\gamma_{\mathrm{int}}(\ell)\cdot\mathrm{Re}\left\{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\right\} \\
&\quad-2\cdot\gamma_{\mathrm{int}}^2(\ell)\cdot\mathrm{Re}\left\{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\right\} \\
&\quad+\gamma_{\mathrm{int}}^2(\ell)\cdot\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\right|^2 \\
&\quad+\gamma_{\mathrm{int}}^2(\ell)\cdot\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\right|^2. \qquad \text{(A.19)}
\end{aligned}
$$

In the following, it is assumed that

$$\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}\sim\mathcal{W}_M\left(L_{\backslash\ell},\frac{1}{L_{\backslash\ell}}\cdot\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}\right), \qquad \text{(A.20)}$$

$$\mathbf{x}\sim\mathcal{N}\left(\mathbf{0},\mathbf{R}_{\mathbf{x}}\right), \qquad \text{(A.21)}$$

$$\mathbf{r}(\ell)=\mathbf{y}(\ell)-\mathbf{x}, \qquad \text{(A.22)}$$

$$\mathbf{r}(\ell)\sim\mathcal{N}\left(\mathbf{0},\mathbf{R}_{\mathbf{r}}\right) \qquad \text{(A.23)}$$

hold with

$$\mathbf{R}_{\mathbf{r}}=\sigma_0^2(\ell)\cdot\mathbf{R}_0+\sigma_1^2(\ell)\cdot\mathbf{R}_1+\sigma_\nu^2\cdot\mathbf{R}_\nu-\mathbf{R}_{\mathbf{x}}. \qquad \text{(A.24)}$$

Moreover, $\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}$, $\mathbf{x}$ and $\mathbf{r}(\ell)$ are mutually statistically independent. Further, let $\widetilde{\mathbf{A}}$, $\widetilde{\mathbf{B}}$, $\widetilde{\mathbf{C}}$ and $\widetilde{\mathbf{D}}$ be deterministic matrices.

In [34] the following higher order moments of an inverse Wishart matrix

$$\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\sim\mathcal{W}_M^{-1}\left(L_{\backslash\ell},L_{\backslash\ell}\cdot\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\right) \qquad \text{(A.25)}$$

are derived, where $\mathcal{O}(\cdot)$ denotes an upper bound on the asymptotic growth of the given function:

$$
\begin{aligned}
\mathbb{E}\left[\mathrm{tr}\left\{\widetilde{\mathbf{A}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{B}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\right\}\right] =& \frac{L_{\backslash\ell}^2}{\left(L_{\backslash\ell}-M\right)^2}\cdot\mathrm{tr}\left\{\widetilde{\mathbf{A}}\cdot\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{B}}\cdot\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\right\} \\
&+\frac{L_{\backslash\ell}^2}{\left(L_{\backslash\ell}-M\right)^3}\cdot\mathrm{tr}\left\{\widetilde{\mathbf{A}}\cdot\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\right\}\cdot\mathrm{tr}\left\{\widetilde{\mathbf{B}}\cdot\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\right\} \\
&+\mathcal{O}\left(\frac{L_{\backslash\ell}^2}{\left(L_{\backslash\ell}-M\right)^4}\right), \qquad \text{(A.26)}
\end{aligned}
$$

$$\begin{aligned}\mathbb{E}\Big[\mathrm{tr}\Big\{\widetilde{\mathbf{A}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\cdot\mathrm{tr}\Big\{\widetilde{\mathbf{B}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\Big] =& \frac{L_{\backslash\ell}^2}{\left(L_{\backslash\ell}-M\right)^2}\cdot\mathrm{tr}\Big\{\widetilde{\mathbf{A}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\cdot\mathrm{tr}\Big\{\widetilde{\mathbf{B}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\\ &+\frac{L_{\backslash\ell}^2}{\left(L_{\backslash\ell}-M\right)^3}\cdot\mathrm{tr}\Big\{\widetilde{\mathbf{A}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{B}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\\ &+\mathcal{O}\left(\frac{L_{\backslash\ell}^2}{\left(L_{\backslash\ell}-M\right)^4}\right),\end{aligned} \tag{A.27}$$

$$\begin{aligned}&\mathbb{E}\Big[\mathrm{tr}\Big\{\widetilde{\mathbf{A}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{B}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\cdot\mathrm{tr}\Big\{\widetilde{\mathbf{C}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\Big]\\ =&\frac{L_{\backslash\ell}^3}{\left(L_{\backslash\ell}-M\right)^3}\cdot\mathrm{tr}\Big\{\widetilde{\mathbf{A}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{B}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\cdot\mathrm{tr}\Big\{\widetilde{\mathbf{C}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\\ &+\frac{L_{\backslash\ell}^3}{\left(L_{\backslash\ell}-M\right)^4}\cdot\Big(\mathrm{tr}\Big\{\widetilde{\mathbf{A}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\cdot\mathrm{tr}\Big\{\widetilde{\mathbf{B}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\cdot\mathrm{tr}\Big\{\widetilde{\mathbf{C}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\\ &+\mathrm{tr}\Big\{\widetilde{\mathbf{A}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{C}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{B}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\\ &+\mathrm{tr}\Big\{\widetilde{\mathbf{A}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{B}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{C}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\Big)\\ &+\mathcal{O}\left(\frac{L_{\backslash\ell}^3}{\left(L_{\backslash\ell}-M\right)^5}\right),\end{aligned} \tag{A.28}$$

$$\begin{aligned}&\mathbb{E}\Big[\mathrm{tr}\Big\{\widetilde{\mathbf{A}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{B}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\cdot\mathrm{tr}\Big\{\widetilde{\mathbf{C}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{D}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\Big]\\ =&\frac{L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^4}\cdot\mathrm{tr}\Big\{\widetilde{\mathbf{A}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{B}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\cdot\mathrm{tr}\Big\{\widetilde{\mathbf{C}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{D}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\\ &+\frac{L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^5}\cdot\Big(\mathrm{tr}\Big\{\widetilde{\mathbf{A}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{B}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{C}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{D}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\\ &+\mathrm{tr}\Big\{\widetilde{\mathbf{A}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{B}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{D}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{C}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\\ &+\mathrm{tr}\Big\{\widetilde{\mathbf{A}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{C}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{D}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{B}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\\ &+\mathrm{tr}\Big\{\widetilde{\mathbf{A}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{D}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{C}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{B}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\\ &+\mathrm{tr}\Big\{\widetilde{\mathbf{A}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{B}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\cdot\mathrm{tr}\Big\{\widetilde{\mathbf{C}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\cdot\mathrm{tr}\Big\{\widetilde{\mathbf{D}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\\ &+\mathrm{tr}\Big\{\widetilde{\mathbf{A}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\cdot\mathrm{tr}\Big\{\widetilde{\mathbf{B}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\cdot\mathrm{tr}\Big\{\widetilde{\mathbf{C}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widetilde{\mathbf{D}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\Big)\\ &+\mathcal{O}\left(\frac{L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^6}\right).\end{aligned} \tag{A.29}$$

Moreover, it was shown in [114] for hermitian matrices $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$ that

$$\mathbb{E}\Big[\mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{A}} \cdot \mathbf{x} \cdot \mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{B}} \cdot \mathbf{x}\Big] = \mathrm{tr}\Big\{\widetilde{\mathbf{A}} \cdot \mathbf{R}_{\mathbf{x}}\Big\} \cdot \mathrm{tr}\Big\{\widetilde{\mathbf{B}} \cdot \mathbf{R}_{\mathbf{x}}\Big\} + 2 \cdot \mathrm{tr}\Big\{\widetilde{\mathbf{A}} \cdot \mathbf{R}_{\mathbf{x}} \cdot \widetilde{\mathbf{B}} \cdot \mathbf{R}_{\mathbf{x}}\Big\}. \tag{A.30}$$

Utilizing that $\mathbf{x}$ and $\mathbf{r}(\ell)$ are statistically independent and their expected value is zero, results in

$$\begin{aligned}
\mathbb{E}\Big[\mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{A}} \cdot \mathbf{y}(\ell) \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widetilde{\mathbf{B}} \cdot \mathbf{x}\Big] &= \mathbb{E}\Big[\mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{A}} \cdot (\mathbf{x} + \mathbf{r}(\ell)) \cdot (\mathbf{x} + \mathbf{r}(\ell))^{\mathrm{H}} \cdot \widetilde{\mathbf{B}} \cdot \mathbf{x}\Big] \\
&= \mathbb{E}\Big[\mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{A}} \cdot \mathbf{x} \cdot \mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{B}} \cdot \mathbf{x} + \mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{A}} \cdot \mathbf{r}(\ell) \cdot \mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{B}} \cdot \mathbf{x} \\
&\qquad + \mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{A}} \cdot \mathbf{x} \cdot \mathbf{r}^{\mathrm{H}}(\ell) \cdot \widetilde{\mathbf{B}} \cdot \mathbf{x} + \mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{A}} \cdot \mathbf{r}(\ell) \cdot \mathbf{r}^{\mathrm{H}}(\ell) \cdot \widetilde{\mathbf{B}} \cdot \mathbf{x}\Big] \\
&= \mathbb{E}\Big[\mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{A}} \cdot \mathbf{x} \cdot \mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{B}} \cdot \mathbf{x} + \mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{A}} \cdot \mathbf{r}(\ell) \cdot \mathbf{r}^{\mathrm{H}}(\ell) \cdot \widetilde{\mathbf{B}} \cdot \mathbf{x}\Big] \\
&= \mathrm{tr}\Big\{\widetilde{\mathbf{A}} \cdot \mathbf{R}_{\mathbf{x}}\Big\} \cdot \mathrm{tr}\Big\{\widetilde{\mathbf{B}} \cdot \mathbf{R}_{\mathbf{x}}\Big\} + 2 \cdot \mathrm{tr}\Big\{\widetilde{\mathbf{A}} \cdot \mathbf{R}_{\mathbf{x}} \cdot \widetilde{\mathbf{B}} \cdot \mathbf{R}_{\mathbf{x}}\Big\} \\
&\quad + \mathrm{tr}\Big\{\widetilde{\mathbf{A}} \cdot \mathbf{R}_{\mathbf{r}} \cdot \widetilde{\mathbf{B}} \cdot \mathbf{R}_{\mathbf{x}}\Big\}
\end{aligned} \tag{A.31}$$

and

$$\begin{aligned}
\mathbb{E}\Big[\mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{A}} \cdot \mathbf{x} \cdot \mathbf{y}^{\mathrm{H}}(\ell) \cdot \widetilde{\mathbf{B}} \cdot \mathbf{y}(\ell)\Big] &= \mathbb{E}\Big[\mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{A}} \cdot \mathbf{x} \cdot (\mathbf{x} + \mathbf{r}(\ell))^{\mathrm{H}} \cdot \widetilde{\mathbf{B}} \cdot (\mathbf{x} + \mathbf{r}(\ell))\Big] \\
&= \mathbb{E}\Big[\mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{A}} \cdot \mathbf{x} \cdot \mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{B}} \cdot \mathbf{x} + \mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{A}} \cdot \mathbf{x} \cdot \mathbf{r}^{\mathrm{H}}(\ell) \cdot \widetilde{\mathbf{B}} \cdot \mathbf{x} \\
&\qquad + \mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{A}} \cdot \mathbf{x} \cdot \mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{B}} \cdot \mathbf{r}(\ell) + \mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{A}} \cdot \mathbf{x} \cdot \mathbf{r}^{\mathrm{H}}(\ell) \cdot \widetilde{\mathbf{B}} \cdot \mathbf{r}(\ell)\Big] \\
&= \mathbb{E}\Big[\mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{A}} \cdot \mathbf{x} \cdot \mathbf{x}^{\mathrm{H}} \cdot \widetilde{\mathbf{B}} \cdot \mathbf{x} \\
&\qquad + \mathrm{tr}\Big\{\widetilde{\mathbf{A}} \cdot \mathbf{x} \cdot \mathbf{x}^{\mathrm{H}}\Big\} \cdot \mathrm{tr}\Big\{\widetilde{\mathbf{B}} \cdot \mathbf{r}(\ell) \cdot \mathbf{r}^{\mathrm{H}}(\ell)\Big\}\Big] \\
&= \mathrm{tr}\Big\{\widetilde{\mathbf{A}} \cdot \mathbf{R}_{\mathbf{x}}\Big\} \cdot \mathrm{tr}\Big\{\widetilde{\mathbf{B}} \cdot \mathbf{R}_{\mathbf{x}}\Big\} + 2 \cdot \mathrm{tr}\Big\{\widetilde{\mathbf{A}} \cdot \mathbf{R}_{\mathbf{x}} \cdot \widetilde{\mathbf{B}} \cdot \mathbf{R}_{\mathbf{x}}\Big\} \\
&\quad + \mathrm{tr}\Big\{\widetilde{\mathbf{A}} \cdot \mathbf{R}_{\mathbf{x}}\Big\} \cdot \mathrm{tr}\Big\{\widetilde{\mathbf{B}} \cdot \mathbf{R}_{\mathbf{r}}\Big\}.
\end{aligned} \tag{A.32}$$

Moreover, it is used that the expected value of a function of the random variables $\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}$, $\mathbf{x}$ and $\mathbf{y}(\ell)$ can be calculated using conditional expected values:

$$\mathbb{E}[\cdot] = \mathbb{E}_{\mathbf{x},\mathbf{y}(\ell)}\left[\mathbb{E}_{\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}|\mathbf{x},\mathbf{y}(\ell)}[\cdot]\right] = \mathbb{E}_{\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}}\left[\mathbb{E}_{\mathbf{x},\mathbf{y}(\ell)|\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}}[\cdot]\right]. \tag{A.33}$$

Here, $\mathbb{E}_{\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}}[\cdot]$ denotes the expected value over $\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}$, $\mathbb{E}_{\mathbf{x},\mathbf{y}(\ell)}[\cdot]$ the expected value over $\mathbf{x}$ and $\mathbf{y}(\ell)$, $\mathbb{E}_{\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}|\mathbf{x},\mathbf{y}(\ell)}[\cdot]$ the expected value over $\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}$ for a given value of $\mathbf{x}$ and $\mathbf{y}(\ell)$, and $\mathbb{E}_{\mathbf{x},\mathbf{y}(\ell)|\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}}[\cdot]$ the the expected value over $\mathbf{x}$ and $\mathbf{y}(\ell)$ for a given value of $\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}$.

Employing the linearity of the expected value operator, (4.75) follows from the following expected values:

$$\begin{aligned}\mathbb{E}\Big[\Big(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\Big)^2\Big] &= \mathbb{E}\Big[\mathrm{tr}\Big\{\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\Big]\\ &= \frac{\big(L_{\backslash\ell}-M+1\big)\cdot L_{\backslash\ell}^2}{\big(L_{\backslash\ell}-M\big)^3}\cdot\Big(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\Big)^2,\end{aligned}\tag{A.34}$$

$$\begin{aligned}&\mathbb{E}\Big[\mathrm{Re}\Big\{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\Big\}\Big]\\ &=\mathrm{Re}\Big\{\mathbb{E}\Big[\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\Big]\Big\}\\ &=\mathrm{Re}\Big\{\mathbb{E}\Big[\mathrm{tr}\Big\{\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\cdot\mathrm{tr}\Big\{\mathbf{x}\cdot\mathbf{y}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\Big]\Big\}\\ &=\mathrm{Re}\Big\{\mathbb{E}\Big[\mathrm{tr}\Big\{\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\cdot\mathrm{tr}\Big\{\mathbf{R}_{\mathbf{x}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\Big]\Big\}\\ &\approx\frac{\big(L_{\backslash\ell}-M+1\big)\cdot L_{\backslash\ell}^3}{\big(L_{\backslash\ell}-M\big)^4}\cdot\Big(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\Big)^2\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\\ &\quad+\frac{2\cdot L_{\backslash\ell}^3}{\big(L_{\backslash\ell}-M\big)^4}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\Big\},\end{aligned}\tag{A.35}$$

$$\begin{aligned}&\mathbb{E}\Big[\mathrm{Re}\Big\{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\Big\}\Big]\\ &=\mathrm{Re}\Big\{\mathbb{E}\Big[\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\Big]\Big\}\\ &=\mathrm{Re}\Big\{\mathbb{E}\Big[\mathrm{tr}\Big\{\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\cdot\mathbf{y}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\cdot\mathrm{tr}\Big\{\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\Big]\Big\}\\ &=\mathrm{Re}\Big\{\mathbb{E}\Big[\mathrm{tr}\Big\{\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\cdot\mathrm{tr}\Big\{\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\Big\}\Big]\Big\}\\ &\approx\frac{\big(L_{\backslash\ell}-M+2\big)\cdot L_{\backslash\ell}^3}{\big(L_{\backslash\ell}-M\big)^4}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\\ &\quad+\frac{L_{\backslash\ell}^3}{\big(L_{\backslash\ell}-M\big)^4}\cdot\Big(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\Big)^2\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\},\end{aligned}\tag{A.36}$$

$$
\begin{aligned}
&\mathbb{E}\Big[\mathrm{Re}\left\{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{y}(\ell)\cdot\mathbf{x}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\right\}\Big]\\
&=\mathrm{Re}\left\{\mathbb{E}\Big[\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{y}(\ell)\cdot\mathbf{x}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\Big]\right\}\\
&=\mathrm{Re}\left\{\mathbb{E}\Big[\mathbf{x}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{y}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\Big]\right\}\\
&=\mathrm{Re}\left\{\mathbb{E}\Big[\mathrm{tr}\Big\{\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\cdot\mathrm{tr}\Big\{\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\right.\\
&\quad+2\cdot\mathrm{tr}\Big\{\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\\
&\quad\left.+\ \mathrm{tr}\Big\{\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\Big]\right\}\\
&\approx\frac{\big(L_{\backslash\ell}-M+2\big)\cdot L_{\backslash\ell}^{4}}{\big(L_{\backslash\ell}-M\big)^{5}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\Big\}\\
&\quad+\frac{\big(L_{\backslash\ell}-M+4\big)\cdot L_{\backslash\ell}^{4}}{\big(L_{\backslash\ell}-M\big)^{5}}\cdot2\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\\
&\quad+\frac{\big(L_{\backslash\ell}-M+4\big)\cdot L_{\backslash\ell}^{4}}{\big(L_{\backslash\ell}-M\big)^{5}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\\
&\quad+\frac{L_{\backslash\ell}^{4}}{\big(L_{\backslash\ell}-M\big)^{5}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\Big\}\\
&\quad+\frac{L_{\backslash\ell}^{4}}{\big(L_{\backslash\ell}-M\big)^{5}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\Big\}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\\
&\quad+\frac{L_{\backslash\ell}^{4}}{\big(L_{\backslash\ell}-M\big)^{5}}\cdot\Big(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\Big)^{2}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\Big\}\\
&\quad+\frac{L_{\backslash\ell}^{4}}{\big(L_{\backslash\ell}-M\big)^{5}}\cdot2\cdot\Big(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\Big)^{2}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\\
&\quad+\frac{L_{\backslash\ell}^{4}}{\big(L_{\backslash\ell}-M\big)^{5}}\cdot\Big(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\Big)^{2}\cdot\Big(\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\Big)^{2}\\
&\quad+\frac{L_{\backslash\ell}^{4}}{\big(L_{\backslash\ell}-M\big)^{5}}\cdot2\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\\
&\quad+\frac{L_{\backslash\ell}^{4}}{\big(L_{\backslash\ell}-M\big)^{5}}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\Big\}\\
&\quad+\frac{L_{\backslash\ell}^{4}}{\big(L_{\backslash\ell}-M\big)^{5}}\cdot\Big(\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\Big)^{2},
\end{aligned}
\tag{A.37}
$$

$$
\begin{aligned}
&\mathbb{E}\left[\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\right|^2\right]\\
=&\mathbb{E}\left[\mathrm{tr}\left\{\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\right\}\cdot\mathrm{tr}\left\{\mathbf{x}\cdot\mathbf{x}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{y}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\right\}\right]\\
\approx&\mathbb{E}\left[\frac{L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^4}\cdot\mathrm{tr}\left\{\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\right\}\cdot\mathrm{tr}\left\{\mathbf{x}\cdot\mathbf{x}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{y}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\right\}\right.\\
&+\frac{L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^5}\cdot\left(\mathrm{tr}\left\{\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\cdot\mathbf{x}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{y}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\right\}\right.\\
&+\mathrm{tr}\left\{\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{y}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\cdot\mathbf{x}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\right\}\\
&+\mathrm{tr}\left\{\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\cdot\mathbf{x}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{y}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\right\}\\
&+\mathrm{tr}\left\{\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{y}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\cdot\mathbf{x}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\right\}\\
&+\mathrm{tr}\left\{\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\right\}\cdot\mathrm{tr}\left\{\mathbf{x}\cdot\mathbf{x}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\right\}\cdot\mathrm{tr}\left\{\mathbf{y}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\right\}\\
&\left.\left.+\mathrm{tr}\left\{\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\right\}\cdot\mathrm{tr}\left\{\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\right\}\cdot\mathrm{tr}\left\{\mathbf{x}\cdot\mathbf{x}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{y}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\right\}\right)\right]\\
=&\mathbb{E}\left[\frac{L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^4}\cdot\left(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\right)^2\cdot\mathbf{x}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{y}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\right.\\
&+\frac{L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^5}\cdot\left(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathbf{x}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{y}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\right.\\
&+\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathbf{x}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{y}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\\
&+\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathbf{x}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{y}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\\
&+\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathbf{x}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{y}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\\
&+\left(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\right)^2\cdot\mathbf{x}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{y}(\ell)\\
&\left.\left.+\left(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\right)^2\cdot\mathbf{x}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{y}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\right)\right]\\
=&\frac{\left(L_{\backslash\ell}-M+2\right)\cdot L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^5}\cdot2\cdot\left(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\right)^2\cdot\mathrm{tr}\left\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\right\}\\
&+\frac{\left(L_{\backslash\ell}-M+1\right)\cdot L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^5}\cdot\left(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\right)^2\cdot\mathrm{tr}\left\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\right\}\\
&+\frac{\left(L_{\backslash\ell}-M+2\right)\cdot L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^5}\cdot\left(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\right)^2\cdot\left(\mathrm{tr}\left\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\right\}\right)^2\\
&+\frac{L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^5}\cdot4\cdot\mathrm{tr}\left\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\right\}\cdot\mathrm{tr}\left\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\right\}
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^5}\cdot 8\cdot \mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\\
&+\frac{L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^5}\cdot 2\cdot \mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\Big\}\\
&+\frac{L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^5}\cdot 2\cdot \mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\\
&+\frac{L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^5}\cdot\Big(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\Big)^2\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\Big\},
\end{aligned}
\tag{A.38}
$$

$$
\begin{aligned}
&\mathbb{E}\Big[\Big|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\Big|^2\Big]\\
=&\mathbb{E}\Big[\mathbf{x}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{y}(\ell)\cdot\mathbf{y}^{\mathrm{H}}(\ell)\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{x}\Big]\\
=&\mathbb{E}\Big[\mathrm{tr}\Big\{\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\cdot\mathrm{tr}\Big\{\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\\
&+2\cdot\mathrm{tr}\Big\{\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\\
&+\mathrm{tr}\Big\{\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\Big)\Big]\\
=&\mathbb{E}\Big[\mathrm{tr}\Big\{\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\cdot\mathrm{tr}\Big\{\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\\
&+2\cdot\mathrm{tr}\Big\{\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\cdot\mathrm{tr}\Big\{\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\\
&+\mathrm{tr}\Big\{\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\cdot\mathrm{tr}\Big\{\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\Big\}\Big]\\
\approx&\frac{\left(L_{\backslash\ell}-M\right)\cdot L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^5}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\Big\}\\
&+\frac{\left(L_{\backslash\ell}-M+1\right)\cdot L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^5}\cdot 2\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\\
&+\frac{\left(L_{\backslash\ell}-M+1\right)\cdot L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^5}\cdot\Big(\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\Big)^2\\
&+\frac{L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^5}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\Big\}\\
&+\frac{L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^5}\cdot 2\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\Big\}\\
&+\frac{L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^5}\cdot 10\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\\
&+\frac{L_{\backslash\ell}^4}{\left(L_{\backslash\ell}-M\right)^5}\cdot 5\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}
\end{aligned}
$$

$$+\frac{L_{\setminus\ell}^4}{\left(L_{\setminus\ell}-M\right)^5}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\Big\}$$
$$+\frac{L_{\setminus\ell}^4}{\left(L_{\setminus\ell}-M\right)^5}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\Big\}$$
$$+\frac{L_{\setminus\ell}^4}{\left(L_{\setminus\ell}-M\right)^5}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathrm{tr}\Big\{\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{R}_{\mathbf{x}}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{R}_{\mathbf{r}}\Big\}. \tag{A.39}$$

A.6 Matrix algebra

In the following, the matrix algebra fundamentals which are used for the analysis of the closed-form approximation of the signal-to-distortion ratio (SDR) at the beamformer output in Sec. 4.4 are explained. To this end, the $D\times D$-dimensional, positive definite, hermitian matrices $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$ are considered, with the eigenvalues $\lambda_0(\widetilde{\mathbf{A}})\geq\cdots\geq\lambda_{D-1}(\widetilde{\mathbf{A}})>0$ and $\lambda_0(\widetilde{\mathbf{B}})\geq\cdots\geq\lambda_{D-1}(\widetilde{\mathbf{B}})>0$, respectively. The corresponding eigenvectors are denoted as $\mathbf{o}_0(\widetilde{\mathbf{A}}),\ldots,\mathbf{o}_{D-1}(\widetilde{\mathbf{A}})$ and $\mathbf{o}_0(\widetilde{\mathbf{B}}),\ldots,\mathbf{o}_{D-1}(\widetilde{\mathbf{B}})$, respectively. Further, let $\widetilde{\mathbf{a}}$ and $\widetilde{\mathbf{c}}$ be complex-valued vectors.

Trace operator

The trace of a matrix is given by

$$\mathrm{tr}\Big\{\widetilde{\mathbf{A}}\Big\}=\sum_{i=0}^{D-1}\widetilde{A}_{ii}=\sum_{i=0}^{D-1}\lambda_i(\widetilde{\mathbf{A}}). \tag{A.40}$$

Since $\widetilde{\mathbf{A}}$ is positive definite, it holds that

$$\mathrm{tr}\Big\{\widetilde{\mathbf{A}}\Big\}>0. \tag{A.41}$$

Rayleigh quotient

As mentioned in [63], the Rayleigh quotient is defined as

$$\frac{\widetilde{\mathbf{c}}^{\mathrm{H}}\cdot\widetilde{\mathbf{A}}\cdot\widetilde{\mathbf{c}}}{\widetilde{\mathbf{c}}^{\mathrm{H}}\cdot\widetilde{\mathbf{c}}}. \tag{A.42}$$

The Rayleigh quotient is bounded by the smallest and largest eigenvalue of the matrix $\widetilde{\mathbf{A}}$:

$$\lambda_{D-1}(\widetilde{\mathbf{A}})\leq\frac{\widetilde{\mathbf{c}}^{\mathrm{H}}\cdot\widetilde{\mathbf{A}}\cdot\widetilde{\mathbf{c}}}{\widetilde{\mathbf{c}}^{\mathrm{H}}\cdot\widetilde{\mathbf{c}}}\leq\lambda_0(\widetilde{\mathbf{A}}). \tag{A.43}$$

If the vector $\widetilde{\mathbf{c}}$ corresponds to the eigenvector $\mathbf{o}_i(\widetilde{\mathbf{A}})$, the Rayleigh quotient equals the corresponding eigenvalue:

$$\frac{\left(\mathbf{o}_i(\widetilde{\mathbf{A}})\right)^{\mathrm{H}} \cdot \widetilde{\mathbf{A}} \cdot \mathbf{o}_i(\widetilde{\mathbf{A}})}{\left(\mathbf{o}_i(\widetilde{\mathbf{A}})\right)^{\mathrm{H}} \cdot \mathbf{o}_i(\widetilde{\mathbf{A}})} = \lambda_i(\widetilde{\mathbf{A}}). \tag{A.44}$$

Generalized Rayleigh quotient

As mentioned in [60], the generalized Rayleigh quotient is defined as

$$\frac{\widetilde{\mathbf{c}}^{\mathrm{H}} \cdot \widetilde{\mathbf{A}} \cdot \widetilde{\mathbf{c}}}{\widetilde{\mathbf{c}}^{\mathrm{H}} \cdot \widetilde{\mathbf{B}} \cdot \widetilde{\mathbf{c}}}. \tag{A.45}$$

The generalized Rayleigh quotient is linked to the following generalized eigenvalue problem:

$$\widetilde{\mathbf{A}} \cdot \widetilde{\mathbf{o}} = \widetilde{\lambda} \cdot \widetilde{\mathbf{B}} \cdot \widetilde{\mathbf{o}}, \tag{A.46}$$

with the generalized eigenvalue $\widetilde{\lambda}$ and the generalized eigenvector $\widetilde{\mathbf{o}}$. Similar as the Rayleigh quotient, the generalized Rayleigh quotient is bounded by the smallest generalized eigenvalue $\widetilde{\lambda}_{D-1}\left(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}\right)$ and the largest generalized eigenvalue $\widetilde{\lambda}_0\left(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}\right)$:

$$\widetilde{\lambda}_{D-1} \leq \frac{\widetilde{\mathbf{c}}^{\mathrm{H}} \cdot \widetilde{\mathbf{A}} \cdot \widetilde{\mathbf{c}}}{\widetilde{\mathbf{c}}^{\mathrm{H}} \cdot \widetilde{\mathbf{B}} \cdot \widetilde{\mathbf{c}}} \leq \widetilde{\lambda}_0. \tag{A.47}$$

If the vector $\widetilde{\mathbf{c}}$ corresponds to the generalized eigenvector $\mathbf{o}_i\left(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}\right)$, the generalized Rayleigh quotient equals the corresponding generalized eigenvalue $\lambda_i\left(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}\right)$:

$$\frac{\left(\mathbf{o}_i\left(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}\right)\right)^{\mathrm{H}} \cdot \widetilde{\mathbf{A}} \cdot \mathbf{o}_i\left(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}\right)}{\left(\mathbf{o}_i\left(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}\right)\right)^{\mathrm{H}} \cdot \widetilde{\mathbf{B}} \cdot \mathbf{o}_i\left(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}\right)} = \lambda_i\left(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}\right). \tag{A.48}$$

Matrix decomposition

In the following, the matrix $\mathbf{R}_0$ should be decomposed into $\mathbf{R}_{0,\mathrm{steer}}$ which corresponds to the contribution along $\widehat{\mathbf{d}} \cdot \widehat{\mathbf{d}}^{\mathrm{H}}$ and the remainder $\mathbf{R}_{0,\mathrm{rest}}$, which should be positive semidefinite. The starting point is:

$$\mathbf{R}_{0,\mathrm{steer}} = a \cdot \widehat{\mathbf{d}} \cdot \widehat{\mathbf{d}}^{\mathrm{H}}, \tag{A.49}$$

$$\mathbf{R}_{0,\mathrm{rest}} = \mathbf{R}_0 - \mathbf{R}_{0,\mathrm{steer}}. \tag{A.50}$$

Next, a is chosen such that $\mathbf{R}_{0,\text{rest}}$ is positive semidefinite while the maximal rank-1 component of $\mathbf{R}_0$ along $\widehat{\mathbf{d}} \cdot \widehat{\mathbf{d}}^{\mathrm{H}}$ is subtracted in (A.50). Since $\mathbf{R}_{0,\text{rest}}$ should be positive semidefinite,

$$\det\left(\mathbf{R}_{0,\text{rest}}\right) = \det\left(\mathbf{R}_0 - a \cdot \widehat{\mathbf{d}} \cdot \widehat{\mathbf{d}}^{\mathrm{H}}\right) \geq 0 \tag{A.51}$$

must hold for its determinant. Applying Cauchy's formula for the determinant of a rank-1 perturbation [63]

$$\det\left(\widetilde{A} + \widetilde{b} \cdot \widetilde{c}^{\mathrm{H}}\right) = \det\left(\widetilde{A}\right) \cdot \left(1 + \widetilde{c}^{\mathrm{H}} \cdot \widetilde{A}^{-1} \cdot \widetilde{b}\right), \tag{A.52}$$

leads to

$$\det\left(\mathbf{R}_{0,\text{rest}}\right) = \det\left(\mathbf{R}_0\right) \cdot \left(1 - a \cdot \widehat{\mathbf{d}}^{\mathrm{H}} \cdot \mathbf{R}_0^{-1} \cdot \widehat{\mathbf{d}}\right) \geq 0. \tag{A.53}$$

Employing that $\mathbf{R}_0$ is positive definite, results in the following condition for a in order to force $\mathbf{R}_{0,\text{rest}}$ to be positive semidefinite:

$$1 - a \cdot \widehat{\mathbf{d}}^{\mathrm{H}} \cdot \mathbf{R}_0^{-1} \cdot \widehat{\mathbf{d}} \geq 0 \Leftrightarrow a \leq \frac{1}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \mathbf{R}_0^{-1} \cdot \widehat{\mathbf{d}}}. \tag{A.54}$$

In order to subtract the maximal rank-1 component of $\mathbf{R}_0$ along $\widehat{\mathbf{d}} \cdot \widehat{\mathbf{d}}^{\mathrm{H}}$, the largest possible value for a has to be chosen. This leads to

$$\mathbf{R}_{0,\text{steer}} = \frac{1}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \mathbf{R}_0^{-1} \cdot \widehat{\mathbf{d}}} \cdot \widehat{\mathbf{d}} \cdot \widehat{\mathbf{d}}^{\mathrm{H}}. \tag{A.55}$$

Matrix derivatives

As stated in [115], it holds that

$$\frac{\mathrm{d}}{\mathrm{d}a}\widetilde{\mathbf{B}}^{-1} = -\widetilde{\mathbf{B}}^{-1} \cdot \frac{\mathrm{d}\widetilde{\mathbf{B}}}{\mathrm{d}a} \cdot \widetilde{\mathbf{B}}^{-1}. \tag{A.56}$$

Further, the following derivatives are stated in [116]:

$$\frac{\mathrm{d}}{\mathrm{d}\widetilde{\mathbf{B}}}\widetilde{\mathbf{a}}^{\mathrm{H}} \cdot \widetilde{\mathbf{B}}^{-1} \cdot \widetilde{\mathbf{c}} = -\left(\widetilde{\mathbf{B}}^{-1} \cdot \widetilde{\mathbf{c}} \cdot \widetilde{\mathbf{a}}^{\mathrm{H}} \cdot \widetilde{\mathbf{B}}^{-1}\right)^{\mathrm{T}}, \tag{A.57}$$

$$\frac{\mathrm{d}}{\mathrm{d}\widetilde{\mathbf{B}}}\mathrm{tr}\left\{\widetilde{\mathbf{A}} \cdot \widetilde{\mathbf{B}}^{-1} \cdot \widetilde{\mathbf{C}}\right\} = -\left(\widetilde{\mathbf{B}}^{-1} \cdot \widetilde{\mathbf{C}} \cdot \widetilde{\mathbf{A}} \cdot \widetilde{\mathbf{B}}^{-1}\right)^{\mathrm{T}}. \tag{A.58}$$

The chain rule is given by

$$\frac{\mathrm{d}}{\mathrm{d}a}\varphi(a) = \left(\frac{\mathrm{d}}{\mathrm{d}\widetilde{\mathbf{B}}}\varphi(a)\right)^{\mathrm{T}} \cdot \frac{\mathrm{d}}{\mathrm{d}a}\widetilde{\mathbf{B}}. \tag{A.59}$$

A.7 Denominator of the signal power at the beamformer output in the presence of STOs

In the following, the denominator of the signal power at the beamformer output in the presence of sampling time offsets (STOs) is considered. To this end, one time frequency bin is considered, as in Sec. 6.1, and the frequency bin index k is neglected. As described in Sec. 6.1, the following relationships hold between the quantities in the presence of STOs and the corresponding quantities if STOs are not present:

$$\widehat{\mathbf{d}}_{\mathrm{STO}} = \mathbf{S} \cdot \widehat{\mathbf{d}}, \tag{A.60}$$

$$\boldsymbol{\Sigma}_{\mathrm{int}}^{\mathrm{STO}} = \mathbf{S} \cdot \boldsymbol{\Sigma}_{\mathrm{int}} \cdot \mathbf{S}^{\mathrm{H}}, \tag{A.61}$$

$$\mathbf{R}_{\mathbf{x}}^{\mathrm{STO}} = \mathbf{S} \cdot \mathbf{R}_{\mathbf{x}} \cdot \mathbf{S}^{\mathrm{H}}, \tag{A.62}$$

$$\mathbf{R}_{\mathbf{r}}^{\mathrm{STO}} = \mathbf{S} \cdot \mathbf{R}_{\mathbf{r}} \cdot \mathbf{S}^{\mathrm{H}}. \tag{A.63}$$

The denominator of the signal power at the beamformer output in (4.59) in the presence of STOs follows from replacing $\widehat{\mathbf{d}}$, $\boldsymbol{\Sigma}_{\mathrm{int}}$, $\mathbf{R}_{\mathbf{x}}$ and $\mathbf{R}_{\mathbf{r}}$ in (4.75) by $\widehat{\mathbf{d}}_{\mathrm{STO}}$, $\boldsymbol{\Sigma}_{\mathrm{int}}^{\mathrm{STO}}$, $\mathbf{R}_{\mathbf{x}}^{\mathrm{STO}}$ and $\mathbf{R}_{\mathbf{r}}^{\mathrm{STO}}$:

$$\begin{aligned}
&\mathbb{E}\Big[|\delta_{\mathrm{STO}}(\mathbf{x})|^2\Big] \approx \frac{\mathbb{E}\Big[|\eta_{\mathrm{STO}}(\mathbf{x})|^2\Big]}{\mathbb{E}\left[\left|\widehat{\mathbf{d}}_{\mathrm{STO}}^{\mathrm{H}} \cdot \left(\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{\mathrm{STO}}\right)^{-1} \cdot \widehat{\mathbf{d}}_{\mathrm{STO}}\right|^2\right]} \\
&= 1 + \frac{2 \cdot \gamma_{\mathrm{int}}(\ell) \cdot L_{\backslash\ell}}{\left(L_{\backslash\ell} - M + 1\right)} \cdot \Bigg(\mathrm{tr}\Big\{\mathbf{S} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \mathbf{R}_{\mathbf{x}} \cdot \mathbf{S}^{\mathrm{H}}\Big\} \\
&\quad - \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \mathbf{R}_{\mathbf{x}} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \widehat{\mathbf{d}}} \Bigg) \\
&\quad + \frac{\gamma_{\mathrm{int}}^2(\ell) \cdot L_{\backslash\ell}^2}{\left(L_{\backslash\ell} - M + 1\right) \cdot \left(L_{\backslash\ell} - M\right)^2} \cdot \Bigg(\left(L_{\backslash\ell} - M\right) \cdot \Big(\mathrm{tr}\Big\{\mathbf{S} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \mathbf{R}_{\mathbf{x}} \cdot \mathbf{S}^{\mathrm{H}}\Big\}\Big)^2 \\
&\quad + 2 \cdot \left(L_{\backslash\ell} - M\right) \cdot \mathrm{tr}\Big\{\mathbf{S} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \mathbf{R}_{\mathbf{x}} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \mathbf{R}_{\mathbf{x}} \cdot \mathbf{S}^{\mathrm{H}}\Big\} \\
&\quad + 3 \cdot \left(L_{\backslash\ell} - M - 1\right) \cdot \frac{\left(\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \mathbf{R}_{\mathbf{x}} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \widehat{\mathbf{d}}\right)^2}{\left(\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \widehat{\mathbf{d}}\right)^2} \\
&\quad - \left(4 \cdot \left(L_{\backslash\ell} - M\right) - 2\right) \cdot \frac{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \mathbf{R}_{\mathbf{x}} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \mathbf{R}_{\mathbf{x}} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \boldsymbol{\Sigma}_{\mathrm{int},\backslash\ell}^{-1} \cdot \mathbf{S}^{\mathrm{H}} \cdot \mathbf{S} \cdot \widehat{\mathbf{d}}}
\end{aligned}$$

$$
\begin{aligned}
&- \left(2\cdot\left(L_{\setminus\ell}-M\right)-1\right)\cdot\frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\widehat{\mathbf{d}}\cdot\mathrm{tr}\left\{\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{S}^{\mathrm{H}}\right\}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\widehat{\mathbf{d}}}\\
&+ \left(L_{\setminus\ell}-M-1\right)\cdot\mathrm{tr}\left\{\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\mathbf{R}_{\mathbf{r}}\cdot\mathbf{S}^{\mathrm{H}}\right\}\\
&+ \left(L_{\setminus\ell}-M-1\right)\cdot\Bigg(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\widehat{\mathbf{d}}\\
&\cdot\frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\mathbf{R}_{\mathbf{r}}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\widehat{\mathbf{d}}}{\left(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\widehat{\mathbf{d}}\right)^{2}}\Bigg)\\
&- 2\cdot\left(L_{\setminus\ell}-M-1\right)\cdot\frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\mathbf{R}_{\mathbf{r}}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\widehat{\mathbf{d}}}\\
&+ \mathrm{tr}\left\{\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{S}^{\mathrm{H}}\right\}\cdot\mathrm{tr}\left\{\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\mathbf{R}_{\mathbf{r}}\cdot\mathbf{S}^{\mathrm{H}}\right\}\\
&- \frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\mathbf{R}_{\mathbf{r}}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\widehat{\mathbf{d}}\cdot\mathrm{tr}\left\{\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{S}^{\mathrm{H}}\right\}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}^{-1}\cdot\mathbf{S}^{\mathrm{H}}\cdot\mathbf{S}\cdot\widehat{\mathbf{d}}}\Bigg).
\end{aligned}
\tag{A.64}
$$

Utilizing that $\mathbf{S}^{\mathrm{H}}{=}\mathbf{S}^{-1}$ holds, it can be shown that $\mathbb{E}\left[\left|\delta_{\mathrm{STO}}(\mathbf{x})\right|^2\right]$ corresponds to the denominator of the signal power $\mathbb{E}\left[\left|\delta(\mathbf{x})\right|^2\right]$ for the case without STOs.

A.8 Denominator of the signal power at the beamformer output in the presence of SROs

In the following, the denominator of the signal power at the beamformer output is derived for the case that sampling rate offsets (SROs) are present. To this end, one time frequency bin is considered, as in Sec. 6.3, and the frequency bin index k is neglected. Let the unnormalized scale matrix $\boldsymbol{\Sigma}_{\mathrm{int}}^{\mathrm{SRO}}$ of the equivalent Wishart distribution of the interference-SCM estimates in the presence of SROs be defined as in (6.13). As described in Sec. 6.3, the following relationships hold between the quantities in the presence of SROs and the corresponding quantities if SROs are not present: $\mathbf{R}_{\mathbf{x}}^{\mathrm{SRO}} = \mathbf{E}(\ell)\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{E}^{\mathrm{H}}(\ell)$ and $\mathbf{R}_{\mathbf{r}}^{\mathrm{SRO}} = \mathbf{E}(\ell)\cdot\mathbf{R}_{\mathbf{r}}\cdot\mathbf{E}^{\mathrm{H}}(\ell)$. For sake of a simpler notation, the time frame index ℓ of the matrix $\mathbf{E}(\ell)$ which represents the SRO-induced time shifts will be omitted in the following.

The denominator of the signal power at the beamformer output in the presence of SROs follows from replacing $\boldsymbol{\Sigma}_{\mathrm{int}}$, $\mathbf{R}_{\mathbf{x}}$ and $\mathbf{R}_{\mathbf{r}}$ in (4.75) by $\boldsymbol{\Sigma}_{\mathrm{int}}^{\mathrm{SRO}}$, $\mathbf{R}_{\mathbf{x}}^{\mathrm{SRO}}$ and $\mathbf{R}_{\mathbf{r}}^{\mathrm{SRO}}$:

$$
\begin{aligned}
\mathbb{E}\Big[|\delta_{\mathrm{SRO}}(\mathbf{x})|^2\Big] &\approx \frac{\mathbb{E}\Big[|\eta_{\mathrm{SRO}}(\mathbf{x})|^2\Big]}{\mathbb{E}\left[\left|\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\left(\widehat{\mathbf{R}}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}\right|^2\right]} \\
=1+\frac{2\cdot\gamma_{\mathrm{int}}(\ell)\cdot L_{\backslash\ell}}{\left(L_{\backslash\ell}-M+1\right)}\cdot&\left(\mathrm{tr}\left\{\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\mathbf{E}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{E}^{\mathrm{H}}\right\}\right. \\
&\left.-\frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\mathbf{E}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{E}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}}\right) \\
&+\frac{\gamma_{\mathrm{int}}^2(\ell)\cdot L_{\backslash\ell}^2}{\left(L_{\backslash\ell}-M+1\right)\cdot\left(L_{\backslash\ell}-M\right)^2}\cdot\left(\left(L_{\backslash\ell}-M\right)\cdot\left(\mathrm{tr}\left\{\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\mathbf{E}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{E}^{\mathrm{H}}\right\}\right)^2\right. \\
&+2\cdot\left(L_{\backslash\ell}-M\right)\cdot\mathrm{tr}\left\{\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\mathbf{E}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{E}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\mathbf{E}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{E}^{\mathrm{H}}\right\} \\
&+3\cdot\left(L_{\backslash\ell}-M-1\right)\cdot\frac{\left(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\mathbf{E}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{E}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}\right)^2}{\left(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}\right)^2} \\
&-\left(4\cdot\left(L_{\backslash\ell}-M\right)-2\right)\cdot\frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\mathbf{E}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{E}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\mathbf{E}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{E}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}} \\
&-\left(2\cdot\left(L_{\backslash\ell}-M\right)-1\right)\cdot\frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\mathbf{E}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{E}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathrm{tr}\left\{\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\mathbf{E}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{E}^{\mathrm{H}}\right\}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}} \\
&+\left(L_{\backslash\ell}-M-1\right)\cdot\mathrm{tr}\left\{\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\mathbf{E}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{E}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\mathbf{E}\cdot\mathbf{R}_{\mathbf{r}}\cdot\mathbf{E}^{\mathrm{H}}\right\} \\
&+\left(L_{\backslash\ell}-M-1\right)\cdot\frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\mathbf{E}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{E}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}\cdot\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\mathbf{E}\cdot\mathbf{R}_{\mathbf{r}}\cdot\mathbf{E}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}}{\left(\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}\right)^2} \\
&-2\cdot\left(L_{\backslash\ell}-M-1\right)\cdot\frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\mathbf{E}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{E}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\mathbf{E}\cdot\mathbf{R}_{\mathbf{r}}\cdot\mathbf{E}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}} \\
&+\mathrm{tr}\left\{\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\mathbf{E}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{E}^{\mathrm{H}}\right\}\cdot\mathrm{tr}\left\{\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\mathbf{E}\cdot\mathbf{R}_{\mathbf{r}}\cdot\mathbf{E}^{\mathrm{H}}\right\} \\
&\left.-\frac{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\mathbf{E}\cdot\mathbf{R}_{\mathbf{r}}\cdot\mathbf{E}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}\cdot\mathrm{tr}\left\{\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\mathbf{E}\cdot\mathbf{R}_{\mathbf{x}}\cdot\mathbf{E}^{\mathrm{H}}\right\}}{\widehat{\mathbf{d}}^{\mathrm{H}}\cdot\left(\mathbf{\Sigma}_{\mathrm{int},\backslash\ell}^{\mathrm{SRO}}\right)^{-1}\cdot\widehat{\mathbf{d}}}\right). \qquad \text{(A.65)}
\end{aligned}
$$

Symbols and notation

In the following, the symbols and notation used throughout this thesis are explained. Note that only symbols which are used throughout the thesis or are especially important are listed below. Symbols which are only used in a limited local context are introduced at their first appearance.

For vectors bold lower-case characters are employed while bold upper-case characters are utilized for matrices. a_i corresponds to the i-th element of the vector $\mathbf{a}$. A_{ij} and $(\mathbf{A})_{ij}$ correspond to the i-th row and j-th column element of the matrix $\mathbf{A}$. In order to resolve ambiguities between ground-truth values and the corresponding estimates, symbols with a hat indicate estimates, i.e., $\widehat{a}$ corresponds to the estimate of a . If only the estimated value and not the corresponding ground-truth value appear in this work, the hat is omitted. Whenever there is no ambiguity, indices, like the time frame index or the frequency bin index, are omitted. However, this is also mentioned in the relevant section of this thesis.

Symbols

b	Block index in block-wise beamforming
B	Advance of the analysis window used for STFT calculation
B_s	The SRO is estimated every B_s time frames in the DWACD method
c	Speed of sound
$\widehat{\mathbf{d}}(b,k)$	Block-wise steering vector estimate
$\widehat{\mathbf{d}}(k)$	Steering vector estimate
$\mathbf{E}(\ell,k)$	Diagonal matrix containing the SRO-induced phase terms used to model the SRO-induced time shift in the STFT domain
f_s	Nominal sampling frequency
$g_\mathrm{WACD}(l,k)$	Temporally averaged version of the complex-conjugated product of consecutive coherence functions $g_\Gamma^{(\zeta)}(\ell,k)$
$g_\Gamma(\ell,k)$	Complex-conjugated product of consecutive coherence functions
$g_\Gamma^{(\zeta)}(\ell,k)$	Complex-conjugated product of consecutive coherence functions after compensating for a shift of ζ between the signals and compensating for SRO

$h_{q,m}(k)$	ATF belonging to the RIR which models the propagation of sound waves from the position of speaker q to the position of microphone m
$\mathbf{h}_q(k)$	Vector of stacked ATFs belonging to the q-th speaker
$h_{q,m}(n)$	RIR modeling the propagation of sound waves from the position of speaker q to the position of microphone m
$\mathbf{I}$	Identity matrix
j	Imaginary unit
k	Frequency bin index
K	Number of frequency bins
ℓ	Time frame index
ℓ_d	Temporal distance between the two consecutive coherence function when calculating the complex-conjugated product of consecutive coherence functions
L	Block size in block-wise beamforming / sample size used for SCM estimation
L_W	Length of the estimation interval used for PSD estimation in the DWACD method
$L_{\setminus\ell}$	Equivalent degrees of freedom of the approximate Wishart distribution of the interference-SCM estimate $\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}$
$\widetilde{L}(b,k)$	Equivalent degrees of freedom of the approximate Wishart distribution of the interference-SCM estimate $\widehat{\mathbf{R}}_\mathrm{int}(b,k)$
m	Microphone index
M	Number of microphones
n	Discrete-time index
N	Size of the analysis window used for STFT calculation
N_B	Number of blocks a signal consist of for block-wise beamforming
$p(\mathbf{x})$	Power of the source of interest $\mathbf{x}$ at the beamformer output
$p_\mathrm{t}(\mathbf{x})$	Transient component of the power of the source of interest $\mathbf{x}$ at the beamformer output
$p_\infty(\mathbf{x})$	Steady-state component of the power of the source of interest $\mathbf{x}$ at the beamformer output
$p_\mathrm{int}(b,\ell,k)$	Instantaneous power of the interference component at the beamformer output for one time frequency bin in block-wise beamforming

$P_{\text{int}}(L)$	Energy of the interference signal component at the beamformer output as a function of the block size L
$p_{\text{tar}}(b, \ell, k)$	Instantaneous power of the target component at the beamformer output for one time frequency bin in block-wise beamforming
$P_{\text{tar}}(L)$	Energy of the target signal component at the beamformer output as a function of the block size L
q	Speaker index
Q	Number of speakers
$r_{q,m}$	Distance between the position of the q-th speaker and the m-th microphone
$\mathbf{R}_{\text{int}}(\ell, k)$	Instantaneous interference SCM
$\widehat{\mathbf{R}}_{\text{int}}(b, k)$	Block-wise estimate of the interference SCM
$\widehat{\mathbf{R}}_{\text{int},\setminus\ell}$	Interference-SCM estimate excluding the ℓ-th time frame from the block-wise estimate
$\widehat{\mathbf{R}}_{\text{int}}^{\text{SRO}}(k)$	Block-wise interference-SCM estimate in the presence of SROs
$\widehat{\mathbf{R}}_{\text{int}}^{\text{STO}}(k)$	Block-wise interference-SCM estimate in the presence of STOs
$\mathbf{R}_q(k)$	SCM of the q-th speaker for frequency bin k
$\mathbf{R}_{\mathbf{r}}$	Sum of the instantaneous second-order moments of the signals of all sources except the source of interest $\mathbf{x}$
$\mathbf{R}_{\mathbf{r}}^{\text{SRO}}$	Sum of the instantaneous second-order moments of the signals of all sources except the source of interest $\mathbf{x}$ in the presence of SROs
$\mathbf{R}_{\mathbf{r}}^{\text{STO}}$	Sum of the instantaneous second-order moments of the signals of all sources except the source of interest $\mathbf{x}$ in the presence of STOs
$\mathbf{R}_{\text{tar}}(\ell, k)$	Instantaneous target SCM
$\widehat{\mathbf{R}}_{\text{tar}}(b, k)$	Block-wise estimate of the target SCM
$\widehat{\mathbf{R}}_{\text{tar}}^{\text{SRO}}(k)$	Block-wise target-SCM estimate in the presence of SROs
$\widehat{\mathbf{R}}_{\text{tar}}^{\text{STO}}(k)$	Block-wise target-SCM estimate in the presence of STOs
$\mathbf{R}_{\mathbf{x}}$	Instantaneous second-order moment of the signals of the source of interest $\mathbf{x}$
$\mathbf{R}_{\mathbf{x}}^{\text{SRO}}$	Instantaneous second-order moment of the signals of the source of interest $\mathbf{x}$ in the presence of SROs
$\mathbf{R}_{\mathbf{x}}^{\text{STO}}$	Instantaneous second-order moment of the signals of the source of interest $\mathbf{x}$ in the presence of STOs
$\mathbf{R}_{\nu}(k)$	SCM of the stationary noise signal for frequency bin k

$\mathbf{R}_{0,\text{steer}}$	Component of the target speaker's SCM that is aligned with the outer product of the steering vector
$\mathbf{R}_{0,\text{rest}}$	Remainder of the target speaker's SCM that is not aligned with the outer product of the steering vector
$\mathbf{S}(k)$	Diagonal matrix containing the STO-induced phase terms used to model the corresponding time shift in the frequency domain
$\text{SDR}(L)$	Invasive SDR at the beamformer output as a function of the block size L
t	Continuous-time index
T_{60}	Sound decay time
$\mathbf{u}$	One-hot vector indicating the reference channel used for MVDR beamforming
$\mathbf{w}(b,k)$	Block-wise estimate of the coefficients of an MVDR beamformer
$\mathbf{w}(k)$	Coefficients of an MVDR beamformer
$\mathbf{w}_{\text{SRO}}(k)$	Block-wise estimate of the MVDR beamformer coefficients in the presence of SROs
$\mathbf{w}_{\text{STO}}(k)$	Block-wise estimate of the MVDR beamformer coefficients in the presence of STOs
$\mathbf{w}_{\setminus\ell}$	Block-wise estimate of the MVDR beamformer coefficients excluding the ℓ-th time frame from estimation
$\mathbf{w}_{\setminus\ell}^{\text{SRO}}$	Block-wise estimate of the MVDR beamformer coefficients in the presence of SROs excluding the ℓ-th time frame from estimation
$\mathbf{w}_{\setminus\ell}^{\text{STO}}$	Block-wise estimate of the MVDR beamformer coefficients in the presence of STOs excluding the ℓ-th time frame from estimation
$\mathbf{x}_{\text{int}}(\ell,k)$	Vector of stacked STFTs of the interfering signals
$x_{q,m}(\ell,k)$	STFT of the source image of the q-th speaker at the m-th microphone
$\mathbf{x}_q(\ell,k)$	Vector of stacked STFTs of all source images belonging to the q-th speaker
$x_{q,m}(n)$	Source image of the q-th speaker at the m-th microphone
$\mathbf{x}_{\text{tar}}(\ell,k)$	Vector of stacked STFTs of the target speaker's source images
$\widehat{x}_{\text{tar}}(\ell,k)$	Signal extracted via beamforming / output of the beamformer

$\mathbf{y}(\ell, k)$	Vector of stacked STFTs of all microphone signals
$y_m(\ell, k)$	STFT of the m-th microphone signal
$y_m^{\mathrm{ASYNC}}(\ell, k)$	Signal of the m-th microphone in the presence of an STO and an SRO
$\mathbf{y}_{\mathrm{SRO}}(\ell, k)$	Vector of stacked STFTs of all microphone signals in the presence of SROs
$\mathbf{y}_{\mathrm{STO}}(\ell, k)$	Vector of stacked STFTs of all microphone signals in the presence of STOs
$y_m(n)$	Discrete-time signal captured by the m-th microphone
$y_m^{(\zeta)}(n)$	m-th microphone signal after compensating for a shift of ζ
$y_m^{\mathrm{SRO}}(n)$	Signal of the m-th microphone in the presence of an SRO
$y_m^{\mathrm{STO}}(n)$	Signal of the m-th microphone in the presence of an STO
$y_m(t)$	Continuous-time signal captured by the m-th microphone
$z_q(n)$	Source signal of the q-th speaker
α	Smoothing factor when calculating the temporally averaged version of the complex-conjugated product of consecutive coherence functions in the DWACD method
$\gamma_{\mathrm{int}}(\ell, k)$	Interference mask
$\gamma_{\mathrm{tar}}(\ell, k)$	Target mask
$\Gamma_{ij}(\ell, k)$	Coherence between the i-th microphone signal and the j-th microphone signal
$\Gamma_{ij}^{(\zeta)}(\ell, k)$	Coherence between the i-th microphone signal and the j-th microphone signal after compensating for a shift of ζ between the signals and compensating for SRO
$\delta(\mathbf{x})$	Denominator of the instantaneous power of the source of interest $\mathbf{x}$ at the beamformer output
Δ_{start}	Start offset of the SRO trajectories created by the Ornstein-Uhlenbeck process w.r.t. the steady-state expected value
$\Delta\mathrm{SDR}(L)$	SDR degradation due to block-wise SCM w.r.t. using a infinitely large sample size for SCM estimation as a function of the block size L
$\varepsilon_{ij}(\ell)$	Time-varying SRO between the i-th microphone signal and the j-th microphone signal
$\widehat{\varepsilon}_{ij}(\ell)$	Estimate of the SRO between the i-th microphone signal and the j-th microphone signal

ε_m	Time-constant SRO of the m-th microphone signal w.r.t. the nominal sampling frequency
$\varepsilon_m(\ell)$	Time-varying SRO of the m-th microphone signal w.r.t. the nominal sampling frequency
θ	Smoothing factor used in the auto-regressive realization of the Ornstein-Uhlenbeck process for modeling time-varying SROs
$\mu_m^{(\infty)}$	Steady-state expected value of the SRO trajectories created by the Ornstein-Uhlenbeck process
$\nu_m(\ell, k)$	STFT of the noise signal captured by the m-th microphone
$\boldsymbol{\nu}(\ell, k)$	Vector of stacked STFTs of all noise signals
$\nu_m(n)$	Noise signal captured by the m-th microphone
$\sigma_q^2(\ell, k)$	Power of the q-th speaker's signal for time frequency bin (ℓ, k)
σ_{OU}	Standard deviation of the Gaussian distribution involved in the auto-regressive realization of the Ornstein-Uhlenbeck process used to model time-varying SROs
$\sigma_\nu^2(k)$	Time-constant power of the stationary noise signal for frequency bin k
$\boldsymbol{\Sigma}_{\mathrm{int}}(b, k)$	Unnormalized equivalent scale matrix of the approximate Wishart distribution of the interference-SCM estimate $\widehat{\mathbf{R}}_{\mathrm{int}}(b, k)$
$\boldsymbol{\Sigma}_{\mathrm{int},\setminus\ell}$	Unnormalized equivalent scale matrix of the approximate Wishart distribution of the interference-SCM estimate $\widehat{\mathbf{R}}_{\mathrm{int},\setminus\ell}$
τ_{ij}	TDOA between the i-th microphone signal and the j-th microphone signal
$\tau_{ij}^{\mathrm{cmp}}(l)$	Integer part of the SRO-induced time shift between the signals to be compensated for before coherence estimation
τ_{ij}^{STO}	STO between the i-th microphone signal and the j-th microphone signal
τ_m^{STO}	STO of the m-th microphone signal w.r.t. the nominal start of the recording
$\tau_{q,ij}^{\mathrm{TOF}}$	TDOF between the i-th microphone signal and the j-th microphone signal for a signal emitted by the q-th speaker
$\Phi_{y_i y_j(\ell,k)}$	PSD of the i-th microphone signal and the j-th microphone signal

$\omega(n)$	Analysis window used for STFT calculation

Operators

$\sim$	Draw a sample from a given probability distribution / a random variable is distributed according to a given probability distribution
$*$	Linear convolution
$(\cdot)^{-1}$	Inverse of a matrix
$(\cdot)^{\frac{1}{2}}$	Matrix square root
$(\cdot)^{\mathrm{T}}$	Transpose of a vector or matrix
$(\cdot)^{\mathrm{H}}$	Hermitian transpose of a vector or matrix
$(\cdot)^{*}$	Complex conjugation of a complex-valued scalar
$\lvert\cdot\rvert$	Absolute value of a complex-valued scalar
$\angle(\cdot)$	Phase of a complex-valued scalar
$\mathrm{Re}\{\cdot\}$	Real part of a complex-valued scalar
$\mathrm{Im}\{\cdot\}$	Imaginary part of a complex-valued scalar
$\mathbb{E}\{\cdot\}$	Expectation operator
$\mathrm{var}(\cdot)$	Variance of a random variable
$y_m(t)\vert_{t=\frac{n}{f_\mathrm{s}}}$	Sampling of the continuous-time signal $y_m(t)$ at $t = \frac{n}{f_\mathrm{s}}$
$\mathrm{tr}\{\cdot\}$	Trace operator
$\mathrm{diag}(\cdot)$	Operator creating a diagonal matrix from a vector
$\exp(\cdot)$	Exponential function
$\mathcal{P}(\cdot)$	Pricincipal eigenvector of a matrix
$d_\mathrm{corr}(\cdot,\cdot)$	Correlation matrix distance
$\lvert\lvert\cdot\rvert\rvert_\mathrm{F}$	Frobenius norm
$\lvert\lvert\cdot\rvert\rvert^2$	Squared Euclidean nrom
$\lfloor\cdot\rceil$	Rounding to the next integer value

Probability distributions

$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution with expected value $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathcal{U}(i, j)$	Uniform distribution in the interval $[i, j]$
$\mathcal{W}_D(A, \boldsymbol{\Sigma})$	Wishart distribution with A degrees of freedom and scale matrix $\boldsymbol{\Sigma}$ which belongs to a $D \times D$-dimensional Wishart matrix
$\mathcal{W}_D^{-1}\left(A, \boldsymbol{\Sigma}^{-1}\right)$	Inverse Wishart matrix with A degrees of freedom and scale matrix $\boldsymbol{\Sigma}^{-1}$ which belongs to the inverse of a $D \times D$-dimensional Wishart matrix

List of figures

List of tables

Acronyms

ACD average coherence drift.

ADC analog-digital converter.

AE absolute error.

ASR automatic speech recognition.

ATF acoustic transfer function.

AWGN additive white Gaussian noise.

cACGMM complex angular central Gaussian mixture model.

CDF cumulative distribution function.

CDR coherent-to-diffuse power ratio.

CM correlation maximization.

CPSD cross power spectral density.

cpWER concatenated minimum-permutation word error rate.

DNN deep neural network.

DOA direction of arrival.

DWACD dynamic weighted average coherence drift.

DXCP double-cross-correlation processor.

DXCP-PhaT double-cross-correlation processor with phase transform.

EVD eigenvalue decomposition.

FFT fast Fourier transform.

GCC generalized cross-correlation.

GCC-PhaT generalized cross-correlation with phase transform.

GCPSD generalized cross power spectral density.

GEVD generalized eigenvalue decomposition.

i.i.d. independent and identically distributed.

IFFT inverse fast Fourier transform.

IRM ideal ratio mask.

ISTFT inverse short-time Fourier transform.

LCD least-squares coherence drift.

LGM local Gaussian model.

LS least squares.

MVDR minimum variance distortionless response.

PDF probability density function.

ppm parts per million.

PSD power spectral density.

RANSAC random sample consensus.

RBI recursive band-limited interpolation.

RF radio frequency.

RIR room impulse response.

RMSE root mean square error.

SAD sound activity detection.

SCM spatial covariance matrix.

SDR signal-to-distortion ratio.

SINR signal-to-interference-plus-noise ratio.

SMI sample matrix inversion.

SNR signal-to-noise ratio.

SRO sampling rate offset.

STFT short-time Fourier transform.

STO sampling time offset.

TDOA time difference of arrival.

TDOF time difference of flight.

TOA time of arrival.

TXCO temperature-compensated crystal oscillator.

VAD voice activity detection.

WACD weighted average coherence drift.

WER word error rate.

Bibliography

List of peer-reviewed publications with own contributions (OC)

[OC1] **T. Gburrek**, J. Schmalenstroeer, and R. Haeb-Umbach, "On synchronization of wireless acoustic sensor networks in the presence of time-varying sampling rate offsets and speaker changes," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 916–920. DOI: 10.1109/ICASSP43922.2022.9746284 (cited on pp. 7, 9, 10, 94, 96, 129, 133, 138, 140–144).

[OC2] **T. Gburrek**, J. Schmalenstroeer, J. Heitkaemper, and R. Haeb-Umbach, "Informed vs. blind beamforming in ad-hoc acoustic sensor networks for meeting transcription," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022. DOI: 10.1109/IWAENC53105.2022.9914772 (cited on pp. 16, 26, 94).

[OC3] J. Schmalenstroeer, **T. Gburrek**, and R. Haeb-Umbach, "LibriWASN: A data set for meeting separation, diarization, and recognition with asynchronous recording devices," in *Proc. ITG Conference on Speech Communication*, 2023, pp. 86–90. DOI: 10.30420/456164016 (cited on pp. 92, 94, 96, 97, 127).

[OC4] T. Cord-Landwehr, **T. Gburrek**, M. Deegen, and R. Haeb-Umbach, *Spatio-spectral diarization of meetings by combining TDOA-based segmentation and speaker embedding-based clustering*, Co–first author (equal contribution)., 2025. DOI: 10.21437/Interspeech.2025-1663 (cited on pp. 92, 97).

[OC5] **T. Gburrek**, J. Schmalenstroeer, and R. Haeb-Umbach, "On source-microphone distance estimation using convolutional recurrent neural networks," in *Proc. ITG Conference on Speech Communication*, 2021, pp. 1–5 (cited on pp. 129, 140).

[OC6] A. Chinaev, G. Enzner, **T. Gburrek**, and J. Schmalenstroeer, "Online estimation of sampling rate offsets in wireless acoustic sensor networks with packet loss," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1110–1114. DOI: 10.23919/EUSIPCO54536.2021.9616037 (cited on pp. 132, 133, 135, 140).

[OC7] **T. Gburrek**, J. Schmalenstroeer, and R. Haeb-Umbach, "On the integration of sampling rate synchronization and acoustic beamforming," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2023, pp. 11–15. DOI: 10.23919/EUSIPCO58844.2023.10289831 (cited on p. 132).

[OC8] **T. Gburrek**, J. Schmalenstroeer, and R. Haeb-Umbach, "Geometry calibration in wireless acoustic sensor networks utilizing DOA and distance information," *EURASIP Journal on Audio, Speech, and Music Processing*, 2021, ISSN: 1687-4722. DOI: 10.1186/s13636-021-00210-x (cited on p. 138).

[OC9] **T. Gburrek**, T. Glarner, J. Ebbers, R. Haeb-Umbach, and P. Wagner, "Unsupervised learning of a disentangled speech representation for voice conversion," in *Proc. ISCA Workshop on Speech Synthesis (SSW)*, 2019, pp. 81–86. DOI: 10.21437/SSW.2019-15.

[OC10] **T. Gburrek**, J. Schmalenstroeer, A. Brendel, W. Kellermann, and R. Haeb-Umbach, "Deep neural network based distance estimation for geometry calibration in acoustic sensor networks," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2021, pp. 196–200. DOI: 10.23919/Eusipco47968.2020.9287583.

[OC11] **T. Gburrek**, J. Schmalenstroeer, and R. Haeb-Umbach, "Iterative geometry calibration from distance estimates for wireless acoustic sensor networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 741–745. DOI: 10.1109/ICASSP39728.2021.9413831.

[OC12] H. Afifi, H. Karl, **T. Gburrek**, and J. Schmalenstroeer, "Data-driven time synchronization in wireless multimedia networks," in *Proc. International Wireless Communications and Mobile Computing (IWCMC)*, 2022, pp. 161–166. DOI: 10.1109/IWCMC55113.2022.9824980.

[OC13] **T. Gburrek**, A. Meise, J. Schmalenstroeer, and R. Haeb-Umbach, "Diminishing domain mismatch for DNN-based acoustic distance estimation via stochastic room reverberation models," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2024, pp. 279–283. DOI: 10.1109/IWAENC61483.2024.10694103.

List of other publications with own contributions (OC)

[OC14] **T. Gburrek**, J. Schmalenstroeer, and R. Haeb-Umbach, *Spatial diarization for meeting transcription with ad-hoc acoustic sensor networks*, Only an extended summary has been peer-reviewed., 2023. DOI: 10.1109/IEEECONF59524.2023.10476794 (cited on pp. 94, 95).

[OC15] **T. Gburrek**, C. Boeddeker, T. von Neumann, T. Cord-Landwehr, J. Schmalenstroeer, and R. Haeb-Umbach, "A meeting transcription system for an ad-hoc acoustic sensor network," *arXiv preprint arxiv.2205.00944*, 2022.

Other References

[1] S.-E. Kotti, R. Heusdens, and R. C. Hendriks, "Clock-offset and microphone gain mismatch invariant beamforming," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2021, pp. 176–180. DOI: 10.23919/Eusipco47968.2020.9287852 (cited on pp. 2, 7, 92, 93, 98, 101, 102).

[2] J. Schmalenstroeer, J. Heymann, L. Drude, C. Boeddeker, and R. Haeb-Umbach, "Multi-stage coherence drift based sampling rate synchronization for acoustic beamforming," in *Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2017 (cited on pp. 2, 9, 92, 93, 103, 131–135, 137).

[3] S. Markovich-Golan, S. Gannot, and I. Cohen, "Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012, pp. 1–4 (cited on pp. 2, 9, 92, 93, 103, 127, 130, 131, 134, 137).

[4] D. Cherkassky, S. Markovich-Golan, and S. Gannot, "Performance analysis of MVDR beamformer in WASN with sampling rate offsets and blind synchronization," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2015, pp. 245–249. DOI: 10.1109/EUSIPCO.2015.7362382 (cited on pp. 2, 92, 93, 103, 131).

[5] J. Schmalenstroeer and R. Haeb-Umbach, "Insights into the interplay of sampling rate offsets and MVDR beamforming," in *Proc. ITG Conference on Speech Communication*, 2018, pp. 1–5 (cited on pp. 2, 92, 93, 103–107, 113).

[6] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Meeting recognition with asynchronous distributed microphone array using block-wise refinement of mask-based MVDR beamformer," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5694–5698. DOI: 10.1109/ICASSP.2018.8462458 (cited on pp. 2, 92, 93, 103, 128).

[7] L. Yu, W. Liu, and R. Langley, "SINR analysis of the subtraction-based SMI beamformer," *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5926–5932, 2010. DOI: 10.1109/TSP.2010.2058801 (cited on pp. 2, 20, 23, 24, 27, 39, 40).

[8] I. Reed, J. Mallett, and L. Brennan, "Rapid convergence rate in adaptive arrays," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-10, no. 6, pp. 853–863, 1974. DOI: 10.1109/TAES.1974.307893 (cited on pp. 2, 23).

[9] L. Chang and C.-C. Yeh, "Performance of DMI and eigenspace-based beamformers," *IEEE Transactions on Antennas and Propagation*, vol. 40, no. 11, pp. 1336–1347, 1992. DOI: 10.1109/8.202711 (cited on pp. 2, 23).

[10] D. Feldman and L. Griffiths, "A projection approach for robust adaptive beamforming," *IEEE Transactions on Signal Processing*, vol. 42, no. 4, pp. 867–876, 1994. DOI: 10.1109/78.285650 (cited on pp. 2, 23, 108).

[11] M. Wax and Y. Anu, "Performance analysis of the minimum variance beamformer," *IEEE Transactions on Signal Processing*, vol. 44, no. 4, pp. 928–937, 1996. DOI: 10.1109/78.492545 (cited on pp. 2, 23).

[12] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, Language Processing*, vol. 25, no. 4, pp. 692–730, 2017. DOI: 10.1109/TASLP.2016.2647702 (cited on pp. 4–6, 86).

[13] A. Brendel and W. Kellermann, "Distributed source localization in acoustic sensor networks using the coherent-to-diffuse power ratio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 61–75, 2019. DOI: 10.1109/JSTSP.2019.2900911 (cited on pp. 5, 133).

[14] K. Lebart, J.-M. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, pp. 359–366, May 2001 (cited on p. 5).

[15] H. Kuttruff, *Room acoustics*. CRC Press, 2016 (cited on p. 5).

[16] C. Fevotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian source models," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2005, pp. 78–81. DOI: 10.1109/ASPAA.2005.1540173 (cited on p. 6).

[17] E. Vincent, S. Arberet, and R. Gribonval, "Underdetermined instantaneous audio source separation via local Gaussian modeling," in *Independent Component Analysis and Signal Separation*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 775–782, ISBN: 978-3-642-00599-2. DOI: doi.org/10.1007/978-3-642-00599-2_97 (cited on p. 6).

[18] M. Guggenberger, M. Lux, and L. Böszörmenyi, "An analysis of time drift in hand-held recording devices," in *Proc. MultiMedia Modeling*, Springer International Publishing, 2015, pp. 203–213, ISBN: 978-3-319-14445-0. DOI: doi.org/10.1007/978-3-319-14445-0_18 (cited on p. 8).

[19] J. Schmalenstroeer, P. Jebramcik, and R. Haeb-Umbach, "A combined hardware–software approach for acoustic sensor network synchronization," *Signal Processing*, vol. 107, pp. 171–184, 2015, ISSN: 0165-1684. DOI: doi.org/10.1016/j.sigpro.2014.06.030 (cited on p. 8).

[20] L. Wang and S. Doclo, "Correlation maximization-based sampling rate offset estimation for distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech, Language Processing*, vol. 24, no. 3, pp. 571–582, 2016. DOI: 10.1109/TASLP.2016.2517326 (cited on pp. 9, 130–132).

[21] S. Miyabe, N. Ono, and S. Makino, "Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation," *Signal Processing*, vol. 107, pp. 185–196, 2015, ISSN: 0165-1684. DOI: https://doi.org/10.1016/j.sigpro.2014.09.015 (cited on pp. 9, 127, 128, 130, 132).

[22] G. E. Uhlenbeck and L. S. Ornstein, "On the theory of the Brownian motion," *Physical Review*, vol. 36, pp. 823–841, 5 Sep. 1930. DOI: 10.1103/PhysRev.36.823 (cited on p. 9).

[23] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations* (Stochastic Modelling and Applied Probability), 1st ed. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 1992, vol. 23. DOI: 10.1007/978-3-662-12616-5 (cited on p. 10).

[24] *TCXO frequency stability and frequency accuracy budget*, SiTime, SiT-AN10039 Rev 1.1, Jul. 2014 (cited on p. 10).

[25] H. Afifi, J. Schmalenstroeer, J. Ullmann, R. Haeb-Umbach, and H. Karl, "MARVELO - a framework for signal processing in wireless acoustic sensor networks," in *Proc. ITG Conference on Speech Communication*, 2018, pp. 1–5 (cited on p. 12).

[26] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969. DOI: 10.1109/PROC.1969.7278 (cited on p. 13).

[27] C. Boeddeker, A. S. Subramanian, G. Wichern, R. Haeb-Umbach, and J. Le Roux, "TS-SEP: Joint diarization and separation conditioned on estimated speaker embeddings," *IEEE/ACM Transactions on Audio, Speech, Language Processing*, vol. 32, pp. 1185–1197, 2024. DOI: 10.1109/TASLP.2024.3350887 (cited on pp. 13, 21, 92).

[28] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on Audio, Speech, Language Processing*, vol. 18, no. 2, pp. 260–276, 2010 (cited on pp. 14, 15, 101).

[29] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2002, pp. 529–532. DOI: 10.1109/ICASSP.2002.5743771 (cited on p. 15).

[30] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Transactions on Audio, Speech, Language Processing*, vol. 21, no. 9, pp. 1913–1928, 2013. DOI: 10.1109/TASL.2013.2263137 (cited on p. 15).

[31] Z. Wang, E. Vincent, R. Serizel, and Y. Yan, "Rank-1 constrained multichannel Wiener filter for speech recognition in noisy environments," *Computer Speech & Language*, vol. 49, pp. 37–51, 2018, ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2017.11.003 (cited on p. 15).

[32] J. M. Martín-Doñas, J. Heitkaemper, R. Haeb-Umbach, A. M. Gomez, and A. M. Peinado, "Multi-channel block-online source extraction based on utterance adaptation," in *Proc. ISCA Interspeech*, 2019, pp. 96–100. DOI: 10.21437/Interspeech.2019-2244 (cited on p. 16).

[33] N. R. Goodman, "Statistical analysis based on a certain multivariate complex Gaussian distribution," *Proceedings of the IEEE*, 1963 (cited on p. 16).

[34] D. Maiwald and D. Kraus, "On moments of complex Wishart and complex inverse Wishart distributed matrices," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1997, pp. 3817–3820. DOI: 10.1109/ICASSP.1997.604712 (cited on pp. 16, 17, 154).

[35] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964 (cited on pp. 19, 96).

[36] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, Apr. 1979. DOI: 10.1121/1.382599 (cited on p. 19).

[37] E. A. Habets, "Room impulse response generator," *Eindhoven University of Technology, Technical Report*, vol. 2, no. 2.4, 2006 (cited on pp. 19, 139).

[38] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006. DOI: 10.1109/TSA.2005.858005 (cited on p. 19).

[39] L. Drude, J. Heitkaemper, C. Boeddeker, and R. Haeb-Umbach, "SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition," *arXiv preprint arXiv:1910.13934*, 2019 (cited on p. 19).

[40] ESPnet, *ESPnet2 pretrained automatic speech recognition model*, Jan. 2022. [Online]. Available: https://huggingface.co/espnet/simpleoier_librispeech_asr_train_asr_conformer7_wavlm_large_raw_en_bpe5000_sp1 (cited on p. 21).

[41] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Yalta, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proc. ISCA Interspeech*, 2018, pp. 2207–2211. DOI: 10.21437/Interspeech.2018-1456 (cited on pp. 21, 96).

[42] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022. DOI: 10.1109/JSTSP.2022.3188113 (cited on p. 21).

[43] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-Light: A benchmark for ASR with limited or no supervision," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7669–7673. DOI: 10.1109/ICASSP40776.2020.9052942 (cited on p. 21).

[44] G. Chen, S. Chai, G.-B. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Z. You, and Z. Yan, "GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio," in *Proc. ISCA Interspeech*, 2021, pp. 3670–3674. DOI: 10.21437/Interspeech.2021-1965 (cited on p. 21).

[45] C. Wang, M. Rivière, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proc. Annual Meeting of the Association for Computational Linguistics*, Aug. 2021, pp. 993–1003. DOI: 10.18653/v1/2021.acl-long.80 (cited on p. 21).

[46] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017, ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2016.11.005 (cited on p. 21).

[47] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 1992, pp. 899–902. DOI: 10.21437/ICSLP.1992-277 (cited on p. 21).

[48] S. Watanabe, *ESPnet2 pretrained automatic speech recognition model*, Jul. 2020. DOI: 10.5281/zenodo.3966501. [Online]. Available: https://doi.org/10.5281/zenodo.3966501 (cited on pp. 21, 96).

[49] M. Herdin, N. Czink, H. Ozcelik, and E. Bonek, "Correlation matrix distance, a meaningful measure for evaluation of non-stationary MIMO channels," in *Proc. IEEE Vehicular Technology Conference*, vol. 1, 2005, pp. 136–140. DOI: 10.1109/VETECS.2005.1543265 (cited on p. 21).

[50] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004. DOI: 10.1109/TSP.2004.828896 (cited on p. 21).

[51] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 708–712. DOI: 10.1109/ICASSP.2015.7178061 (cited on p. 21).

[52] A. Steinhardt, "The PDF of adaptive beamforming weights," *IEEE Transactions on Signal Processing*, vol. 39, no. 5, pp. 1232–1235, 1991. DOI: 10.1109/78.80979 (cited on p. 23).

[53] C. Richmond, "PDF's, confidence regions, and relevant statistics for a class of sample covariance-based array processors," *IEEE Transactions on Signal Processing*, vol. 44, no. 7, pp. 1779–1793, 1996. DOI: 10.1109/78.510624 (cited on p. 23).

[54] B. Van Veen and K. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988. DOI: 10.1109/53.665 (cited on p. 23).

[55] G. F. Pivaro, S. Kumar, G. Fraidenraich, and C. F. Dias, "On the exact and approximate eigenvalue distribution for sum of Wishart matrices," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp. 10 537–10 541, 2017. DOI: 10.1109/TVT.2017.2727259 (cited on pp. 24, 27–29).

[56] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *The Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 832–837, 1956. DOI: 10.1214/aoms/1177728190 (cited on pp. 31, 35).

[57] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962. DOI: 10.1214/aoms/1177704472 (cited on pp. 31, 35).

[58] B. W. Silverman, *Density estimation for statistics and data analysis*. Routledge, 1998 (cited on pp. 31, 35).

[59] P. Glasserman, *Monte Carlo Methods in Financial Engineering*. Springer New York, NY, 2010 (cited on p. 36).

[60] G. H. Golub and C. F. Van Loan, *Matrix computations (3rd ed.)* USA: Johns Hopkins University Press, 1996, ISBN: 0801854148 (cited on pp. 42, 70, 162).

[61] R. C. Elandt-Johnson and N. L. Johnson, *Survival Models and Data Analysis*. 1980, ISBN: 9780471031741 (cited on p. 45).

[62] B. N. Parlett, "The Rayleigh quotient iteration and some generalizations for non-normal matrices," *Mathematics of Computation*, vol. 28, pp. 679–693, 1974 (cited on pp. 57, 70).

[63] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd. Cambridge University Press, 2013, ISBN: 9780521839402 (cited on pp. 69, 161, 163).

[64] C. Boeddeker, T. Cord-Landwehr, T. von Neumann, and R. Haeb-Umbach, "An initialization scheme for meeting separation with spatial mixture models," in *Proc. ISCA Interspeech*, 2022, pp. 271–275. DOI: 10.21437/Interspeech.2022-10929 (cited on pp. 92, 94).

[65] F. Asano, K. Yamamoto, J. Ogata, M. Yamada, and M. Nakamura, "Detection and separation of speech events in meeting recordings using a microphone array," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, no. 1, 2007. DOI: doi.org/10.1155/2007/27616 (cited on p. 92).

[66] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, and J. Yamato, "Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, no. 2, pp. 499–513, 2012. DOI: 10.1109/TASL.2011.2164527 (cited on p. 92).

[67] S. Araki, M. Okada, T. Higuchi, A. Ogawa, and T. Nakatani, "Spatial correlation model based observation vector clustering and MVDR beamforming for meeting recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 385–389. DOI: 10.1109/ICASSP.2016.7471702 (cited on p. 92).

[68] S. Araki, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, T. Higuchi, T. Yoshioka, D. Tran, S. Karita, and T. Nakatani, "Online meeting recognition in noisy environments with time-frequency mask based MVDR beamforming," in *Proc. Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 16–20. DOI: 10.1109/HSCMA.2017.7895553 (cited on p. 92).

[69] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, and T. Nakatani, "Speaker activity driven neural speech extraction," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6099–6103. DOI: 10.1109/ICASSP39728.2021.9414998 (cited on p. 92).

[70] Z.-Q. Wang, P. Wang, and D. Wang, "Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2001–2014, 2021. DOI: 10.1109/TASLP.2021.3083405 (cited on p. 92).

[71] D. Raj, P. Denisov, Z. Chen, H. Erdogan, Z. Huang, M. He, S. Watanabe, J. Du, T. Yoshioka, Y. Luo, N. Kanda, J. Li, S. Wisdom, and J. R. Hershey, "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 897–904. DOI: 10.1109/SLT48900.2021.9383556 (cited on p. 92).

[72] D. Raj, D. Povey, and S. Khudanpur, "GPU-accelerated guided source separation for meeting transcription," in *Proc. ISCA Interspeech*, 2023, pp. 3507–3511. DOI: 10.21437/Interspeech.2023-42 (cited on p. 92).

[73] H. Taherian and D. Wang, "Multi-channel conversational speaker separation via neural diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2467–2476, 2024. DOI: 10.1109/TASLP.2024.3393726 (cited on p. 92).

[74] S. Niu, R. Wang, J. Du, G. Yang, Y. Tu, S. Wu, S. Qian, H. Wu, H. Xu, X. Zhang, G. Zhong, X. Yu, J. Chen, M. Wang, D. Cai, T. Gao, G. Wan, F. Ma, J. Pan, and J. Gao, "The USTC-NERCSLIP systems for the CHiME-8 NOTSOFAR-1 Challenge," in *Proc. International Workshop on Speech Processing in Everyday Environments (CHiME)*, 2024, pp. 31–36. DOI: 10.21437/CHiME.2024-7 (cited on p. 92).

[75] N. Kamo, N. Tawara, A. Ando, T. Kano, H. Sato, R. Ikeshita, T. Moriya, S. Horiguchi, K. Matsuura, A. Ogawa, A. Plaquet, T. Ashihara, T. Ochiai, M. Mimura, M. Delcroix, T. Nakatani, T. Asami, and S. Araki, "Microphone array geometry-independent multi-talker distant ASR: NTT system for DASR task of the CHiME-8 challenge," *Computer Speech & Language*, vol. 95, 2025, ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2025.101820 (cited on p. 92).

[76] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Meeting recognition with asynchronous distributed microphone array," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 32–39. DOI: 10.1109/ASRU.2017.8268913 (cited on p. 92).

[77] T. Yoshioka, D. Dimitriadis, A. Stolcke, W. Hinthorn, Z. Chen, M. Zeng, and X. Huang, "Meeting transcription using asynchronous distant microphones," in *Proc. ISCA Interspeech*, 2019, pp. 2968–2972. DOI: 10.21437/Interspeech.2019-3088 (cited on p. 92).

[78] K. Ochi, N. Ono, S. Miyabe, and S. Makino, "Multi-talker speech recognition based on blind source separation with ad hoc microphone array using smartphones and cloud storage," in *Proc. ISCA Interspeech*, 2016, pp. 3369–3373. DOI: 10.21437/Interspeech.2016-758 (cited on p. 92).

[79] J. Schmalenstroeer and R. Haeb-Umbach, "Efficient sampling rate offset compensation - An overlap-save based approach," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2018, pp. 499–503. DOI: 10.23919/EUSIPCO.2018.8553379 (cited on pp. 94, 120, 127, 130, 140).

[80] N. Ito, S. Araki, and T. Nakatani, "Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1153–1157. DOI: 10.1109/EUSIPCO.2016.7760429 (cited on p. 94).

[81] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks.," in *Proc. ISCA Interspeech*, 2016, pp. 1981–1985. DOI: 10.21437/Interspeech.2016-552 (cited on p. 95).

[82] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous speech separation: Dataset and analysis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7284–7288. DOI: 10.1109/ICASSP40776.2020.9053426 (cited on p. 96).

[83] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. International Workshop on Speech Processing in Everyday Environments (CHiME)*, 2020, pp. 1–7. DOI: 10.21437/CHiME.2020-1 (cited on p. 96).

[84] L. Chang and C.-C. Yeh, "Effect of pointing errors on the performance of the projection beamformer," *IEEE Transactions on Antennas and Propagation*, vol. 41, no. 8, pp. 1045–1056, 1993. DOI: 10.1109/8.244645 (cited on p. 108).

[85] M. Wax and Y. Anu, "Performance analysis of the minimum variance beamformer in the presence of steering vector errors," *IEEE Transactions on Signal Processing*, vol. 44, no. 4, pp. 938–947, 1996. DOI: 10.1109/78.492546 (cited on p. 108).

[86] A. Chinaev, P. Thüne, and G. Enzner, "Low-rate Farrow structure with discrete-lowpass and polynomial support for audio resampling," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2018, pp. 475–479. DOI: 10.23919/EUSIPCO.2018.8553469 (cited on pp. 127, 130).

[87] A. Chinaev, G. Enzner, and J. Schmalenstroeer, "Fast and accurate audio resampling for acoustic sensor networks by polyphase-Farrow filters with FFT realization," in *Proc. ITG Conference on Speech Communication*, 2018, pp. 1–5 (cited on pp. 127, 130).

[88] A. Chinaev, P. Thüne, and G. Enzner, "Double-cross-correlation processing for blind sampling-rate and time-offset estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1881–1896, 2021. DOI: 10.1109/TASLP.2021.3071967 (cited on pp. 128, 132).

[89] N. D. Gaubitch, W. B. Kleijn, and R. Heusdens, "Auto-localization in ad-hoc microphone arrays," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 106–110. DOI: 10.1109/ICASSP.2013.6637618 (cited on p. 128).

[90] N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind alignment of asynchronously recorded signals for distributed microphone array," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 161–164. DOI: 10.1109/ASPAA.2009.5346505 (cited on pp. 128, 129).

[91] S. Woźniak and K. Kowalczyk, "Passive joint localization and synchronization of distributed microphone arrays," *IEEE Signal Processing Letters*, vol. 26, no. 2, pp. 292–296, 2019. DOI: 10.1109/LSP.2018.2889438 (cited on pp. 128, 129).

[92] L. Wang, T.-K. Hon, J. D. Reiss, and A. Cavallaro, "Self-localization of ad-hoc arrays using time difference of arrivals," *IEEE Transactions on Signal Processing*, vol. 64, no. 4, pp. 1018–1033, 2016. DOI: 10.1109/TSP.2015.2498130 (cited on pp. 128, 129).

[93] M. Neri, A. Politis, D. Krause, M. Carli, and T. Virtanen, "Single-channel speaker distance estimation in reverberant environments," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023, pp. 1–5. DOI: 10.1109/WASPAA58266.2023.10248087 (cited on p. 129).

[94] M. Zohourian, J. Stinner, and R. Martin, "Speaker Distance Estimation using Binaural Hearing Aids and Deep Neural Networks," in *Proc. of the 23rd International Congress on Acoustics*, Berlin, Germany: Deutsche Gesellschaft für Akustik, Sep. 9, 2019, pp. 3297–3304. DOI: 10.18154/RWTH-CONV-239773 (cited on p. 129).

[95] D. Neudek, B. Stodt, S. Getzmann, and R. Martin, "Investigation of binaural distance estimation with artificial neural networks trained on simulated data models," in *Proc. DAS —DAGA*, 2025. DOI: doi.org/10.71568/dasdaga2025.355 (cited on p. 129).

[96] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981, ISSN: 0001-0782. DOI: 10.1145/358669.358692 (cited on p. 130).

[97] Q. M. Chaudhari, "A simple and robust clock synchronization scheme," *IEEE Transactions on Communications*, vol. 60, no. 2, pp. 328–332, 2012. DOI: 10.1109/TCOMM.2011.110711.100136 (cited on p. 131).

[98] J. Schmalenstroeer and R. Haeb-Umbach, "Sampling rate synchronisation in acoustic sensor networks with a pre-trained clock skew error model," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2013, pp. 1–5 (cited on p. 131).

[99] Y. Zeng, R. C. Hendriks, and N. D. Gaubitch, "On clock synchronization for multi-microphone speech processing in wireless acoustic sensor networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 231–235. DOI: 10.1109/ICASSP.2015.7177966 (cited on p. 131).

[100] R. Wang, Z. Chen, and F. Yin, "Active sampling rate calibration method for acoustic sensor networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 3095–3107, 2020. DOI: 10.1109/TASLP.2020.3037514 (cited on p. 131).

[101] D. Cherkassky and S. Gannot, "Blind synchronization in wireless sensor networks with application to speech enhancement," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 183–187. DOI: 10.1109/IWAENC.2014.6954003 (cited on p. 131).

[102] D. Cherkassky and S. Gannot, "Blind synchronization in wireless acoustic sensor networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 651–661, 2017. DOI: 10.1109/TASLP.2017.2655259 (cited on pp. 131, 132).

[103] S. Miyabe, N. Ono, and S. Makino, "Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in STFT domain," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 674–678. DOI: 10.1109/ICASSP.2013.6637733 (cited on pp. 131, 132).

[104] M. H. Bahari, A. Bertrand, and M. Moonen, "Blind sampling rate offset estimation based on coherence drift in wireless acoustic sensor networks," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2015, pp. 2281–2285. DOI: 10.1109/EUSIPCO.2015.7362791 (cited on p. 131).

[105] A. Chinaev, P. Thüne, and G. Enzner, "A double-cross-correlation processor for blind sampling rate offset estimation in acoustic sensor networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 641–645. DOI: 10.1109/ICASSP.2019.8683605 (cited on p. 132).

[106] S. Miyabe, N. Ono, and S. Makino, "Optimizing frame analysis with non-integrer shift for sampling mismatch compensation of long recording," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, pp. 1–4. DOI: 10.1109/WASPAA.2013.6701833 (cited on p. 132).

[107] A. Chinaev, S. Wienand, and G. Enzner, "Control architecture of the double-cross-correlation processor for sampling-rate-offset estimation in acoustic sensor networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 801–805. DOI: 10.1109/ICASSP39728.2021.9413768 (cited on pp. 132, 135).

[108] M. H. Bahari, A. Bertrand, and M. Moonen, "Blind sampling rate offset estimation for wireless acoustic sensor networks through weighted least-squares coherence drift estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 674–686, 2017. DOI: 10.1109/TASLP.2016.2647713 (cited on pp. 132, 135).

[109] S. Guan, J. Wang, M. Wang, J. Chen, and J. Benesty, "Online sampling rate offset estimation via real part maximization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 3623–3637, 2025. DOI: 10.1109/TASLPRO.2025.3599776 (cited on p. 132).

[110] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Estimation of sampling frequency mismatch between distributed asynchronous microphones under existence of source movements with stationary time periods detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 785–789. DOI: 10.1109/ICASSP.2019.8683192 (cited on p. 133).

[111] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976, ISSN: 00963518. DOI: 10.1109/TASSP.1976.1162830 (cited on pp. 138, 140).

[112] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, *TIMIT acoustic-phonetic continuous speech corpus*, Linguistic Data Consortium (LDC), 1993 (cited on p. 139).

[113] L. Isserlis, "On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables," *Biometrika*, vol. 12, no. 1/2, pp. 134–139, 1918, ISSN: 00063444, 14643510 (cited on pp. 149, 151).

[114] D. S. Tracy and S. A. Sultan, "Higher order moments of multivariate normal distribution using matrix derivatives," *Stochastic Analysis and Applications*, vol. 11, no. 3, pp. 337–348, 1993. DOI: 10.1080/07362999308809320 (cited on p. 156).

[115] S. Selby, *Standard Mathematical Tables*. CRC Press, 1974 (cited on p. 163).

[116] K. B. Petersen and M. S. Pedersen, *The matrix cookbook*, 2012 (cited on p. 163).