Carl von Ossietzky
Universität Oldenburg

M.Sc. Engineering Physics

Master's Thesis

Blind Geometry Estimation of a Distributed Microphone Array Using Reverberant Speech

Presented by:
Klaus Brümann

First examiner
Prof. Simon Doclo

Second examiner
MSc. Daniel Fejgin

Oldenburg, April 6, 2021

Abstract

Knowing the positions of ad-hoc, distributed microphones is desirable for many multi-microphone speech enhancement applications. A practical way to estimate the microphone array geometry (MAG) is using acoustic signals. It was shown in [1] that the acoustic coherence could be used to estimate the MAG, only requiring a few seconds of reverberant speech. The framework relies on an iterative procedure, which is a variant of the well-known expectation maximization (EM) algorithm, namely expectation conditional-maximization (ECM), to estimate the coherence. From this coherence estimate, the pairwise distances (PDs) between microphones are estimated by finding a distance-dependent model which best fits estimated coherence. Lastly, the MAG is estimated using multi-dimensional scaling (MDS).

The main goals of this thesis are to analyse how generalizable this framework is and to find out how to best operate it in different scenarios, e.g., if the array is small or large and if the microphones are in free-field or if there is a head between the microphones of a pair of hearing aids. Therefore, various parameters crucial to the MAG-estimation are analysed with Monte-Carlo simulations in terms of PD and MAG estimation error. It is seen that ECM is very sensitive to the initialization of the estimated coherence and speech and reverberation power spectral densities (PSDs). Additionally, the PD estimation is sensitive to the the range of frequencies used. An alternative initial coherence estimate and filter for estimating the initial speech and reverberation PSDs are proposed which increase the robustness of the PD estimation for a wider range of frequencies and reduce the PD estimation error. Since the estimation of the relative-direct transfer function (RDTF) can suffer in highly reverberant environments, the influence of estimation errors in the RDTF are analysed in terms of MAG estimation error for a 3-dimensional microphone array.

Using the insight gained from the results of the Monte-Carlo simulations, the best-performing combination parameters is applied for estimating the geometry of a simulated, ad-hoc, 3-dimensional microphone array of different sizes as well as the distance between a pair of hearing aids. For the hearing-aid scenario, a psycho-acoustically motivated coherence model from [2] is suggested for estimating the inter-aural coherence to estimate the PD between hearing aids. The PD and MAG estimation errors are shown for a simulated, realistic scenario using measured head-related impulse responses (HRIRs) and the distributions of the errors are analysed, demonstrating the effect of outliers in the PD estimation on the MAG estimation. It is also shown how a-priori knowledge of known distances in a microphone array (e.g., PDs between microphones of a single hearing-aid) can be used to improve the MAG estimation error.

Nomenclature

Acronyms

AoA Angle of arrival

BTE Behind-the-ear

CM-step Conditional-maximization step

CL Cube-length

DFT Discrete Fourier transform

DTF Direct transfer function

DoA Direction of arrival

ECM Expectation conditional-maximization

EDM Euclidean distance matrix

EM Expectation maximization

E-step Expectation step

FIR Finite impulse response

GCC-PHAT Generalized cross-correlation with phase transform

IC Inter-aural coherence

ITD Inter-aural time difference

M-step Maximization step

ML Maximum likelihood

MAG Microphone array geometry

MPDR Minimum-power distortionless response

MPDR-S Minimum-power distortionless response (stationary)

MPDR-TV Minimum-power distortionless response (time-varying)

MVDR Minimum-variance distortionless response

MDS Multi-dimensional scaling

PD Pairwise distance

PDF Probability density function

PSD Power spectral density

RDTF Relative-direct transfer function

RDR Reverberant-to-direct ratio

STFT Short-time Fourier transform

SVD Singular-value decomposition

SNR Signal-to-noise ratio

TDoA Time-difference of arrival

ToA Time of arrival

Mathematical Notation

x or X Scalar

$\mathbf{x}$ Vector

$\mathbf{X}$ Matrix

$\mathbf{X}_{a,b}$ element in a-th row and b-th column of matrix $\mathbf{X}$

$\mathbf{X} = [X_{a,b}]$ Matrix constructor

$\mathbf{X}^{-1}$ Inverse of matrix $\mathbf{X}$

$|\mathbf{X}|$ Absolute value of scalar X

X^* Complex-conjugate of scalar X

Re{.} Real part operator

Im{.} Imaginary part operator

$\{.\}^{\mathrm{T}}$ Transpose operator

$\{.\}^{\mathrm{H}}$ Hermitian operator

Span{.} Span operator

diag(.) Diagonal value operator

Tr(.) Trace operator

rank . rank operator

$\|.\|_{\mathrm{F}}$ Frobenius norm

$\|.\|_2$ 2-norm

$|\mathbf{X}|_{\mathrm{mtx}}$ Absolute value of each entry of matrix $\mathbf{X}$

$\mathrm{sinc}(\mathbf{X})_{\mathrm{mtx}}$ sinc function ($\mathrm{sinc}(X) = \frac{\sin(X)}{X}$) applied to each entry of matrix $\mathbf{X}$

$\mathbb{E}\{.\}$ Expectation over realizations operator

$\mathbb{E}_t\{.\}$ Expectation over time operator

Fixed Symbols

$\mathbf{0}_{A,B}$ Matrix/vector of zeros with A rows and B columns

$\mathbf{1}$ Vector of ones

α Stretching parameter of modified sinc coherence

β Damping parameter of modified sinc coherence

c Speed of sound

$d_{a,b}$ Pairwise distance between microphones a and b

$\mathbf{d}_1$ vector consisting of 1^{st} column of EDM $\mathbf{D}_{\mathrm{EDM}}$

$\mathbf{D}_{\mathrm{EDM}}$ EDM matrix

$\boldsymbol{\delta}$ Delta vector (containing distances which are proportional to TDoAs)

$\mathbf{e}_n$ n-th basis vector (i.e., a 1 in the n-th entry and rest 0s)

ε Constant ($\varepsilon = \frac{2\pi f_s}{c}$)

ϵ_{MAG} MAG error

ϵ_{PD} PD error

eps Lower-bound

f Frequency

f_{lower} Lower frequency bound

f_{upper} Upper frequency bound

f_s Sampling frequency

F Loading factor

$\mathbf{g}[k]$ RDTF vector

$G_n[k]$ RDTF coefficient at n-th microphone

$\mathbf{G}$ Gram matrix

$\Gamma_{a,b}[k]$ Coherence between microphones a and b

$\mathbf{\Gamma}_{\mathrm{msinc}}(k,d)$ Inter-aural coherence model

$\mathbf{\Gamma}_{\mathrm{sinc}}(k,d)$ Diffuse coherence model

$\mathbf{\Gamma}[k]$ Coherence matrix

$\mathbf{h}[k]$ FIR filter

$\mathbf{h}_{\mathrm{MPDR}}[k]$ MPDR filter

$\mathbf{h}_{\mathrm{MPDR-S}}[k]$ MPDR-S filter

$\mathbf{h}_{\mathrm{MPDR-TV}}[k]$ MPDR-TV filter

$\mathbf{h}_{\mathrm{MVDR}}[k]$ MVDR filter

$\mathbf{H}$ Eigenvalue matrix with eigenvalues along diagonal, corresponding to Eigenvector matrix $\mathbf{W}$

i ECM iteration

I Maximal number of ECM iterations

$\mathbf{I}$ Identity matrix

k Frequency bin

K Number of frequency bins

$\mathcal{K}$ Set of frequencies used in PD estimation

l Time frame

L Number of time frames

$\mathbf{\Lambda}$ Eigenvalue matrix with eigenvalues along diagonal, corresponding to Eigenvector matrix $\mathbf{U}$

$\mathbf{m}_n$ Coordinates of n-th microphone

$\mathbf{M}$ MAG matrix

n microphone index

N Number of microphones

$\mathcal{N}^C(\mu, \sigma)$ Complex Gaussian with mean μ and standard-deviation σ

ν RDR lower threshold

P Array dimension

P' Room dimension

$\mathbf{P}$ PD matrix

$\phi_R[k, l]$ Reverberation PSD

$\overline{\phi}_R[k]$ Complete observed reverberation PSD

$\Phi_{\mathbf{r}}[k, l]$ Reverberation covariance matrix

$\phi_{S_{\text{ref}}}[k, l]$ Speech PSD

$\Phi_{\mathbf{s}_{\text{ref}}}[k, l]$ Speech covariance matrix

$\overline{\phi}_{S_{\text{ref}}}[k]$ Complete observed speech PSD

$\Phi_{\mathbf{x}}[k, l]$ Covariance matrix of recorded speech

ψ PDF

$\mathbf{q}$ Source coordinate vector

$\mathbb{R}$ Euclidean vector space

$R_n[k, l]$ Reverberation signal at n-th microphone

$\mathbf{r}[k, l]$ Multi-microphone vector of reverberation signal

$\overline{\mathbf{r}}[k]$ Complete observed multi-microphone reverberation signal

ρ Smoothing parameter

σ_τ RDTF error in samples

$S_{\text{ref}}[k, l]$ Speech signal at reference microphone

$\overline{S}_{\text{ref}}[k]$ Complete observed multi-microphone speech signal

t Sample

T_{60} Reverberation decay time

τ_{TDoA} Time-difference of arrival

τ_{ToA} Time of arrival

$\boldsymbol{\theta}$ Set of parameters estimated in ECM

$\mathbf{U}$ Eigenvector matrix corresponding to Eigenvalue matrix $\boldsymbol{\Lambda}$

$w[t]$ Impulse response

$\mathbf{W}$ Eigenvector matrix corresponding to Eigenvalue matrix $\mathbf{H}$

$X_n[k,l]$ n-th microphone signal

$\mathbf{x}[k,l]$ Multi-microphone vector of recorded speech signal

$\overline{\mathbf{x}}[k]$ Complete observed multi-microphone recorded speech signal

$\mathbf{z}$ Translation vector

Contents

List of Figures

List of Tables

1 Introduction

The research directly influencing this work and its applications are introduced in Chapter 1.1. Proposed changes to the state-of-the-art framework (introduced in Chapter 1.1), the goals of proposed analyses carried out in this thesis, and an outline of the other chapters of this thesis are included in Chapter 1.2.

1.1 Background

Recent years have seen ever increasing reliance on the digitalization of academic classes and meetings, requiring the use of webcams, but most importantly, microphones to present material. A common problem in this form of presentation is that the quality and intelligibility of the recorded speech is often adversely affected by additive background noise, such as a background fan, or reverberation, due to acoustically-reflective walls. Using an array of distributed microphones, there are various ways to increase the speech quality; many require knowledge or an estimate of the microphone array geometry (MAG) [5], i.e., the positions of the microphones in a relative coordinate system, or partial information of the geometry such as the microphone spacing relative to the source, in order to exploit spatial or temporal properties of the recorded acoustic signals.

Some methods which aim to improve speech quality increase the signal-to-noise ratio (SNR) or direct-to-reverberant ratio (DRR). To increase the SNR, a spatial beamformer can be employed, such as a delay-and-sum beamformer, to preserve acoustic information, i.e., the speech, coming from a certain direction while suppressing information (additive noise) from other directions [6]. In a reverberant room such as an office or a classroom, the recorded speech signal arriving at the microphones takes many paths from the source. In a free-field scenario, the component with the highest amplitude travels directly from the source to the microphones, i.e., the direct component. The early reflections acoustically reflect from or between the walls one or a few times before arriving at the microphones. The reverberation component reflects between the walls so many times that individual peaks corresponding to an acoustic reflection are no longer identifiable and has an adverse effect on speech quality and intelligibility. To increase the direct-to-reverberant ratio, dereverberation methods can be employed [1]. A simulated room-impulse-response (RIR) which shows the direct and early reflections, and reverberant tail, is shown in Fig. 1.1.

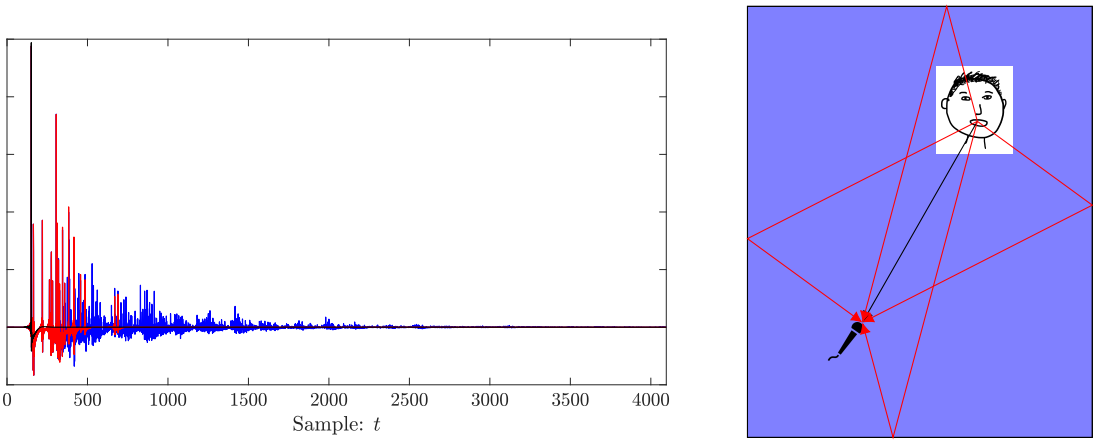


Fig. 1.1: Left: RIR of a simulated room, simulated using [3].
Right: depiction of how a localized speech signal propagates to a microphone within a room.
Black indicates the direct path, red the early reflections (only first-order reflections are shown), and blue the reverberation component.

A distributed array can take many shapes and sizes and has a clear advantage over a linear array, namely that there is no direction of arrival (DoA) of the source which is broadside. In the case that expensive, distributed multi-microphone setups with known geometries are not available, a cheap ad-hoc microphone array setup can be employed to take advantage of the aforementioned techniques, e.g., when giving an online or hybrid (both in-situ and streamed online) lecture, perhaps the lecturer has a laptop with 2 microphones and a phone available to form an ad-hoc microphone array. Precisely measuring the geometry becomes exponentially more tedious as the number of employed microphones increases. Conveniently, the MAG can be estimated in several ways only using sound input [7].

One method to estimate the MAG is using estimated time-differences of arrival (TDoAs) between microphones, relative to a source signal, e.g., using the generalized cross correlation with phase transform (GCC-PHAT) in [8]. However, it was shown in [9, 10] that estimating the MAG in this way requires the TDoAs to be estimated for sources at several unique source positions. This problem can be tackled by relying on a moving source or exploiting acoustic reflections as virtual sources as in [11]. A problem with these methods is that the presence of reverberation can adversely affect the accurate estimation of the TDoA [12].

Motivated by the acoustic coherence-based framework proposed in [13] for blindly estimating the MAG in diffuse noise, modifications were proposed for MAG estimation using reverberant speech in [14]. This enables MAG estimation in highly reverberant environments, where the TDoA-based methods suffer. The framework, in this work referred to as the state-of-the-art, consists of three main steps: esti-

mation of the coherence between microphones, estimation of the pairwise distance (PD) between microphones using the coherence, and finally, MAG estimation using the PDs. An example ad-hoc microphone array for a scenario with four hearing-aids and one mobile phone, and the corresponding PDs is shown in Fig., 1.2.

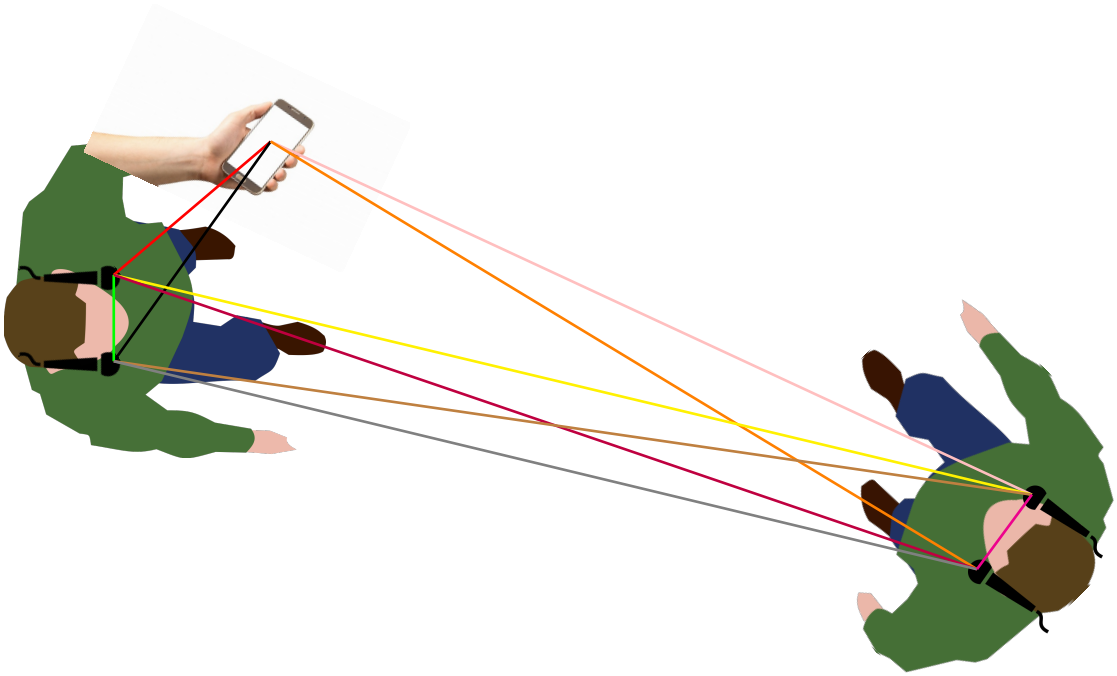


Fig. 1.2: An example ad-hoc microphone array consisting of hearing aids and a mobile phone. The PDs between each microphone of the array are indicated with different coloured lines. To estimate the coherence between a pair of microphones, an estimate of the coherence between them is required.

To estimate the coherence, a variant of the expectation maximization (EM) algorithm is used, i.e., expectation conditional-maximization [15], which iteratively estimates the coherence and the speech and reverberation PSDs given the recorded, reverberant speech signal and an estimate of the relative-direct transfer function (RDTF). The PDs between microphones can be estimated by finding the model coherence which best fits the estimated coherence from ECM and the MAG can be estimated from the squared PDs using multi-dimensional scaling (MDS) [16, 17]. Additionally in [14], it was proposed to use a cylindrically isotropic coherence model, instead of a spherically isotropic model, to more accurately model rooms with an absorbent floor and ceiling [18, 19].

1.2 Contributions and Outline

The goals of this thesis are to check how generalizable this framework is and to inform the reader on how to best operate the MAG estimation, i.e., which parameters need to be tuned or modified to estimate the MAG as accurately as possible in different scenarios such as conferencing with several distributed acoustic devices in the room or quickly calibrating the inter-aural distance of a pair of hearing-aids. In [13] and [14] the MAG estimation was only applied to 2-dimensional arrays with a maximal PD of 20 cm and 16 cm, respectively. Part of the generalization includes analysing the MAG estimation performance for larger PDs. Since the framework consists of three separate algorithms in series (coherence estimation, pairwise distance (PD) estimation, and MAG estimation), important parameters and their influence on the PD and MAG estimation accuracy are analysed with Monte-Carlo simulations to investigate error propagation through the framework.

In the coherence estimation step, the initialization of the ECM algorithm is investigated. Specifically, the initial coherence estimate and the filter which is used for initializing the direct speech and reverberation PSDs using a maximum-likelihood (ML) estimate. An alternative filter and different coherence initializations are proposed and compared. In addition to investigating the PD estimation accuracy for different ECM initializations, the effects of the range of frequencies used and the true (underlying) PD are analysed. In this thesis, unless otherwise stated, the RDTF vector is assumed to be known. Since in practice, the RDTF can contain estimation errors, especially when it is estimated using GCC-PHAT in very reverberant scenarios [12], one analysis is carried out to investigate the influence of estimation errors in the RDTF on the MAG estimation.

With the insight gained from the analysis of the Monte-Carlo simulations, the MAG estimation capabilities are applied in two simulated scenarios based on realistic situations. The first scenario involves estimating the MAG of randomly placed microphones in free-field. In the second scenario, the goal is to estimate the inter-aural distance between a pair of between behind-the-ear (BTE) hearing aids. An example of this scenario is depicted in Fig. 1.3. For this, it is proposed to use a psycho-acoustically motivated inter-aural coherence model [2]. It is also investigated whether incorporating prior knowledge into the MAG estimation can reduce the error of the estimated MAG.

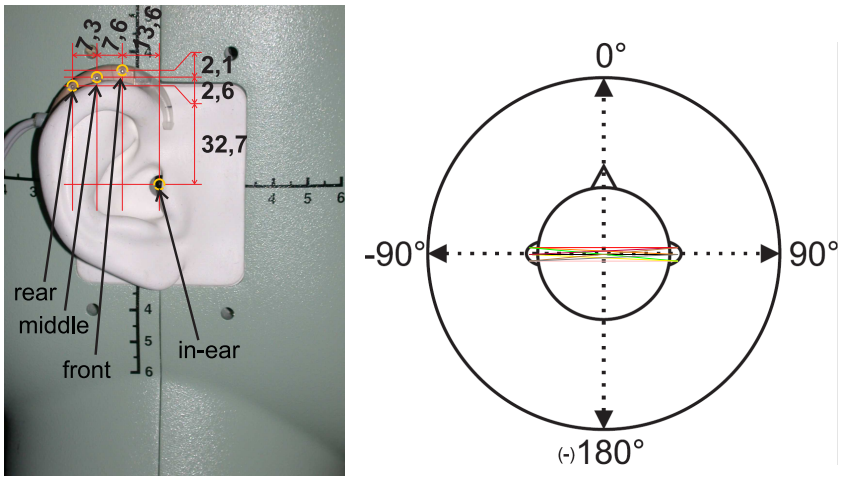


Fig. 1.3: Left: A hearing-aid located on an ear of a dummy [4].
Right: Birds-eye-view schematic of the inter-aural distances to be estimated, indicated with different colours.

In Chapter 2 the acoustic scene, the signal model, and underlying assumptions are introduced and each step of the state of the art MAG estimation is covered. Since the MAG estimate is a relative geometry (i.e., defined in a local coordinate system) and not one which is globally aligned, e.g., with the room or a source, the procedure to align the estimated geometry with the true geometry is included for the purpose of evaluating how accurate the MAG estimate is. The proposed analyses and changes to the MAG estimation are stated in Chapter 3. The experimental results of the analyses using Monte-Carlo simulations, where the state-of-the-art method is compared with implementations using proposed modifications, are covered in Chapter 4. Moreover, the influence of the PDs, the range of frequencies used, and estimation errors in the RDTF are analysed in terms of PD and/or MAG estimation errors. Applying the best performing combinations of parameters, the MAG estimation is applied for real-life-inspired scenarios, i.e., estimating the geometry of ad-hoc, distributed microphones and estimating the inter-aural distance between hearing-aids. This work is concluded in Chapter 5 with a summary and some ideas for future research.

2 Theory

In this Chapter, the state-of-the-art MAG estimation framework using reverberant speech is reviewed. The framework consists of a coherence estimation step, followed by a PD estimation step, from which the MAG is estimated. A framework overview schematic, with the corresponding chapters in which each individual method is described, is presented in Fig. 2.1.

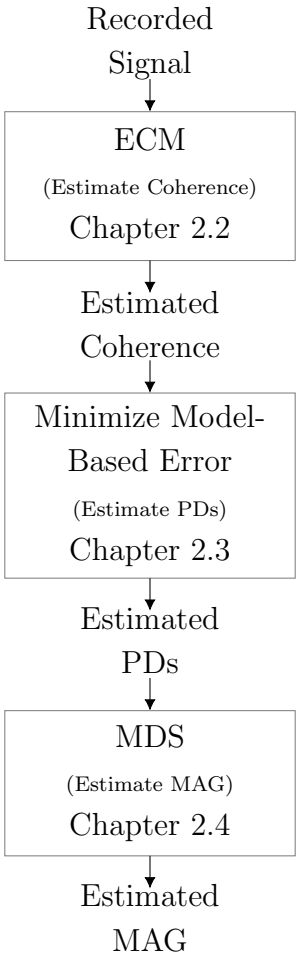


Fig. 2.1: MAG estimation Overview.

The acoustic scene, the signal model, and the signal statistics are introduced in Chapter 2.1. The ECM framework for estimating the coherence is covered in Chapter 2.2. The methods for PD estimation and MAG estimation are described in Chapters 2.3 and 2.4, respectively. An overview block-diagram with the mathematical notation introduced in this Chapter is presented in Fig. 2.5. The Procrustes analysis for aligning the estimated, relative coordinates with the true coordinates for evaluation, is described in Chapter 2.6.

2.1 Acoustic Scenario and Signal Model

The scenario geometry is described in the continuous spatial domain in P'-dimensional Euclidean space $\mathbb{R}^{P'}$ and the signal in the short-time Fourier transform domain. The acoustic scene in question comprises of one person speaking in a reverberant room, with ad-hoc, digital devices distributed within the room in the free-field, which are capable of recording audio, such as mobile phones, laptops, and/or hearing-aids. For a more concise presentation and simpler derivation it is assumed that the source and microphones are stationary, however, minor modifications could be made to the framework in order to be applicable for slowly moving scenarios.

An omnidirectional speech source is located at position $\mathbf{q} \in \mathbb{R}^{P'}$ within the reverberant room and emits an acoustic wave at the speed of sound c. Distributed within the room are N omnidirectional microphones, forming a P-dimensional array (with $P \leq P'$) and the position of the n-th microphone ($n \in \{1, 2, \ldots, N\}$) is denoted by the position vector $\mathbf{m}_n \in \mathbb{R}^P$. For convenience, the individual position vectors for each microphone of the microphone array are stacked into a MAG matrix $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_N]^{\mathrm{T}}$, with $\{.\}^{\mathrm{T}}$ the transpose operator.

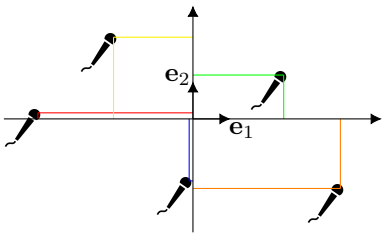


Fig. 2.2: Example MAG with $P = 2$ and $N = 5$. The coloured lines represent the coordinates of individual microphones in a relative coordinate system, with $\mathbf{e}_1 = [1, 0]^{\mathrm{T}}$, $\mathbf{e}_2 = [0, 1]^{\mathrm{T}} \in \mathbb{R}^2$.

The acoustic wave travelling from the source can take many paths to a given microphone. The direct path component of this emitted wave travels directly from the speech source to the distributed microphones. Another component of the wave is partially absorbed and reflected at the walls once or a few times before being captured by the microphones, these are called early reflections. The remaining component of the wave, captured by the microphones, has gone through so many reflections that they can not be distinguished from one-another. This component is the reverberation and it decreases exponentially in amplitude the longer it propagates throughout the room. A more detailed overview is described in [20]. In this work, the early reflections are considered part of the reverberation component.

If the STFT frames are long enough to capture the RIR, the impulse response, which would be applied to the source signal as a convolution in the time domain, can be

applied as a multiplicative transfer function [21]. Since is is difficult to model the acoustic transfer function from a source at an unknown location in the room to a microphone at an unknown location in the room, the relative-direct transfer function (RDTF) is used to describe the direct path between the reference microphone (arbitrarily chosen as $n = 1$) and other microphones.

The recorded microphone signal $X_n[k, l]$ of the n-th microphone at time frame $l \in \{1, 2, \ldots, L\}$ and frequency bin $k \in \{1, 2, \ldots, K\}$ in the STFT domain is stacked in the multichannel, recorded microphone signal vector $\mathbf{x}[k, l] = [X_1[k, l], X_2[k, l], \ldots, X_N[k, l]]^{\mathrm{T}}$ and can be decomposed into a direct speech component

$$\mathbf{s}[k, l] = \mathbf{g}[k] S_{\mathrm{ref}}[k, l] \tag{2.1}$$

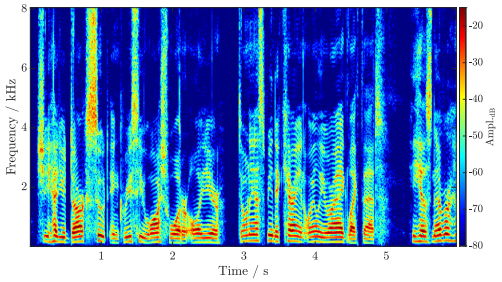
and a reverberant signal vector $\mathbf{r}[k, l] = [R_1[k, l], R_2[k, l], \ldots, R_N[k, l]]^{\mathrm{T}}$, defined similarly to $\mathbf{x}[k, l]$. In the free-field, the entries of the RDTF vector $\mathbf{g}[k] = [1, G_2[k], \ldots, G_N[k]]^{\mathrm{T}}$ are defined as

$$G_n[k] = \exp\left(\frac{-j2\pi\tau_n k}{K}\right), \tag{2.2}$$

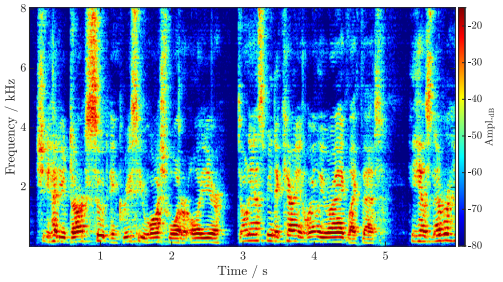
with $j^2 = -1$, and describe the multiplicative direct path transfer function between the speech signal at the reference microphone $S_{\mathrm{ref}}[k, l]$ and the speech signal $S_n[k, l]$ at the n-th microphone. The signal model is defined as

$$\boxed{\underbrace{\mathbf{x}[k, l]}_{\text{recorded}} = \underbrace{\mathbf{g}[k] S_{\mathrm{ref}}[k, l]}_{\text{direct}} + \underbrace{\mathbf{r}[k, l]}_{\text{reverberation}}}. \tag{2.3}$$

Example direct speech and speech reverberation spectrograms, $S_{\mathrm{ref}}[k, l]$ and $R_1[k, l]$ respectively, are depicted in Fig. 2.3. Visually, it is apparent that the late reflections from the reverberation smear the direct speech over time and *fill in some of the gaps*, i.e., low-amplitude time-frequency spectrogram coefficients.



(2.3.1) Direct speech



(2.3.2) Recorded speech

Fig. 2.3: Example spectrograms of direct speech and reverberant, recorded speech. The amplitude is plotted in dB.

The propagation of the reverberation signal component between a given pair of microphones $a \in [1, N]$ and $b \in [1, N]$, averaged over all realizations, is described by the coherence matrix $\mathbf{\Gamma}[k, l] = [\Gamma_{a,b}[k, l]]$, i.e.,

$$\begin{aligned}\Gamma_{a,b}[k,l] &= \frac{\mathbb{E}\left\{R_a[k,l]R_b^*[k,l]\right\}}{\sqrt{\mathbb{E}\left\{R_a[k,l]R_a^*[k,l]\right\}}\sqrt{\mathbb{E}\left\{R_b[k,l]R_b^*[k,l]\right\}}}\,, \\ &= \frac{\phi_{R_{a,b}}[k,l]}{\sqrt{\phi_{R_{a,a}}[k,l]\,\phi_{R_{b,b}}[k,l]}}\,,\end{aligned} \tag{2.4}$$

with $\{.\}^*$ denoting the complex conjugate operator and $\mathbb{E}\{.\}$ the expectation operator over realizations. Assuming ergodicity, the expectation over realizations in (2.4) can be replaced by the expectation over time $\mathbb{E}_{\mathrm{t}}\{.\}$ and since in this work it is assumed that the acoustic scenario is spatially stationary, the coherence coefficient $\Gamma_{a,b}[k, l]$ in (2.4) is only dependent on the frequency bin k and not time-frame l, i.e., $\Gamma_{a,b}[k]$. The signal reverberation PSD is defined as

$$\phi_{R_{a,a}}[k,l] = \mathbb{E}\left\{|R_a[k,l]|^2\right\} \tag{2.5}$$

and the corresponding cross-PSD as

$$\phi_{R_{a,b}}[k,l] = \mathbb{E}\left\{R_a[k,l]R_b^*[k,l]\right\}\,, \tag{2.6}$$

and they are are elements of the signal reverberation matrix $\mathbf{\Phi}_{\mathbf{r}}[k, l] = \left[\phi_{R_{a,b}}[k, l]\right]$. If the room's reverberation time is large enough, its reverberant sound field can be modeled as diffuse, homogeneous, and spherically isotropic [20], i.e., the coherence can be defined as a model function of the PD $d_{a,b}$ between microphones a and b, and frequency bin k, i.e.,

$$\Gamma_{\mathrm{sinc}}(k, d_{a,b}) = \mathrm{sinc}\left(\frac{\varepsilon k d_{a,b}}{K}\right), \tag{2.7}$$

with $\mathrm{sinc}\,(x) = \frac{\sin(x)}{x}$ and $\varepsilon = \frac{2\pi f_s}{c}$. The variation over $d_{a,b}$ and frequency is shown in Fig. 2.4.

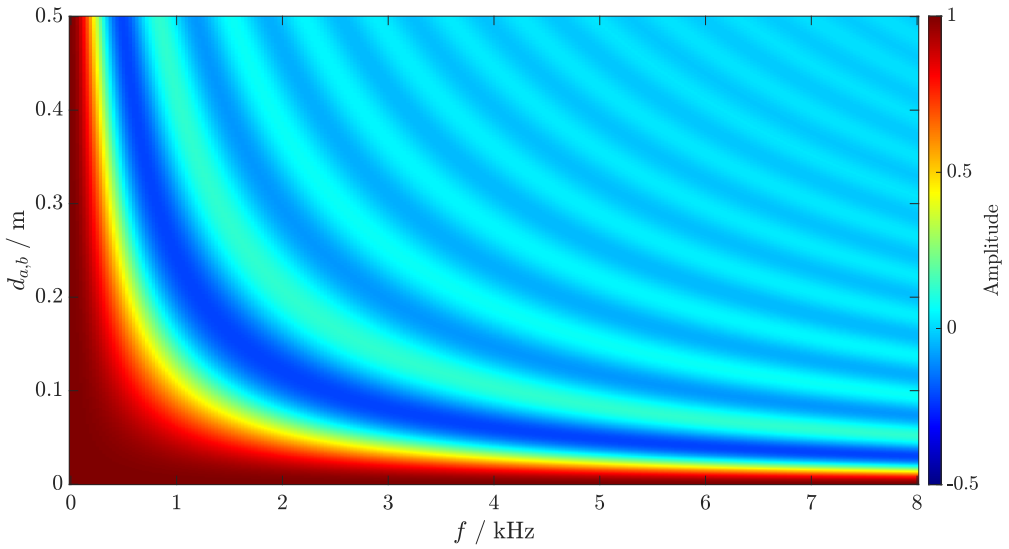


Fig. 2.4: Spherically isotropic coherence plotted over $d_{a,b}$ and f (with $f = \frac{k}{Kf_s}$)

By assuming that the direct and reverberation components in (2.3) are uncorrelated, the recorded signal covariance matrix $\mathbf{\Phi}_{\mathbf{x}}[k,l] = \mathbb{E}\left\{\mathbf{x}[k,l]\mathbf{x}^{\mathrm{H}}[k,l]\right\}$, with $\{.\}^{\mathrm{H}}$ the Hermitian operator, can be decomposed into the sum of the rank 1 direct covariance matrix $\mathbf{\Phi}_{\mathbf{s}_{\mathrm{ref}}}[k,l] = \phi_{S_{\mathrm{ref}}}[k,l]\mathbf{g}[k]\mathbf{g}^{\mathrm{H}}[k]$, with $\phi_{S_{\mathrm{ref}}}[k,l]$ the PSD of the direct speech signal at the reference microphone

$$\phi_{S_{\mathrm{ref}}}[k,l] = S_{\mathrm{ref}}[k,l]S^{*}_{\mathrm{ref}}[k,l]\ , \tag{2.8}$$

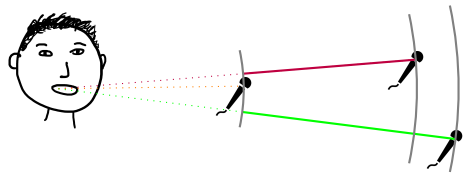
and the reverberant covariance matrix $\mathbf{\Phi}_{\mathbf{r}}[k,l] = \mathbb{E}\left\{\mathbf{r}[k,l]\mathbf{r}^{\mathrm{H}}[k,l]\right\}$.
Due to the homogeneous, reverberant sound field assumption, the reverberant covariance matrix can be described in terms of a time-varying reverberant PSD $\phi_R[k,l]$ and a stationary coherence matrix $\mathbf{\Gamma}[k]$, i.e.,

$$\mathbf{\Phi}_{\mathbf{r}}[k,l] = \phi_R[k,l]\mathbf{\Gamma}[k]\ . \tag{2.9}$$

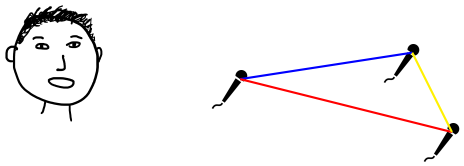
This gives the second-order statistics

$$\underbrace{\mathbf{\Phi}_{\mathbf{x}}[k,l]}_{\text{recorded covariance matrix}} = \underbrace{\phi_{S_{\mathrm{ref}}}[k,l]\mathbf{g}[k]\mathbf{g}^{\mathrm{H}}[k]}_{\text{direct covariance matrix: } \mathbf{\Phi}_{\mathbf{s}_{\mathrm{ref}}}[k,l]} + \underbrace{\phi_R[k,l]\mathbf{\Gamma}[k]}_{\text{reverberation covariance matrix: } \mathbf{\Phi}_{\mathbf{r}}[k,l]} \tag{2.10}$$

where both the direct and reverberation covariance matrices consist of a multiplication between a scalar and a time-invariant matrix.



(2.5.1) RDTF $\mathbf{g}(k)$ encodes TDoA (depends on source angle & PD)



(2.5.2) Spatial coherence $\mathbf{\Gamma}(k)$ encodes PD

Fig. 2.5: Spatial signal parameters

2.2 Coherence Estimation Using ECM

In this work, the estimated PD matrix, which is used to estimate the MAG, requires an estimate of the coherence matrix $\hat{\mathbf{\Gamma}}[k]$ for different frequency bins k. In order to estimate $\mathbf{\Gamma}[k]$, the expectation conditional-maximization (ECM) algorithm [15] which is a generalized form of the EM algorithm, can be used. These are iterative algorithms which, in each iteration, estimate a set of parameters in an expectation step and then maximize (or increase) the likelihood that the estimated parameters describe the observed and estimated data. In the M-step, ECM maximizes the likelihood conditionally on some functions of the parameters which are being estimated (so-called conditional M-step (CM-step)), often making the likelihood maximization simpler than in EM, with the drawback that it does not maximize the likelihood in each iteration but only increases it.

A few assumptions are made for the derivation. The complete observed recorded signal is available for each frequency bin k, i.e., $\overline{\mathbf{x}}[k] = [\mathbf{x}[k,1], \mathbf{x}[k,2], \ldots, \mathbf{x}[k,L]]$. The complete speech $\overline{\mathbf{S}}_{\text{ref}}[k]$ and reverberation $\overline{\mathbf{r}}[k]$ are defined similarly. The complete speech and reverberation PSDs are defined as $\overline{\boldsymbol{\phi}}_{S_{\text{ref}}}[k] = [\phi_{S_{\text{ref}}}[k,1], \phi_{S_{\text{ref}}}[k,2], \ldots, \phi_{S_{\text{ref}}}[k,L]]$ and $\overline{\boldsymbol{\phi}}_{R}[k] = [\phi_{R}[k,1], \phi_{R}[k,2], \ldots, \phi_{R}[k,L]]$, respectively. In this work, unless otherwise stated, it is assumed that the true RDTF vector $\mathbf{g}[k]$ is known. In practice, it can be computed using (2.2), with TDoAs τ_n estimated using e.g., GCC-PHAT as in [8].

In order to fully describe the second-order statistics of the recorded speech in (2.10), it is necessary to estimate the speech PSDs $\overline{\boldsymbol{\phi}}_{S_{\text{ref}}}[k]$, reverberant PSDs $\overline{\boldsymbol{\phi}}_{R}[k]$, and the acoustic coherence matrix $\mathbf{\Gamma}[k]$, which encodes the PD information, depicted in Fig. 2.5. The set of parameters to be estimated is encapsulated by $\boldsymbol{\theta}[k]$, i.e.,

$$\boldsymbol{\theta}[k] = \left\{ \overline{\boldsymbol{\phi}}_{S_{\text{ref}}}[k], \overline{\boldsymbol{\phi}}_{R}[k], \mathbf{\Gamma}[k] \right\}. \tag{2.11}$$

In [22], it was shown that the super-Gaussian distribution was a better model than

a Gaussian distribution for speech, however, in [1] it was argued that using the Gaussian distribution led to simpler derivations. It can be seen in Figs. 2.6 and 2.7 that using simulated RIRs, a zero-mean Gaussian distribution is a reasonable fit for both the real and imaginary components of speech $S_1[k,l]$ and reverberation $R_1[k,l]$ at the reference microphone. Fig. 2.8 shows that the diffuse coherence from (2.7) fits well with the reverberation data.

Thus, it is assumed that in any given microphone, the speech $S_{\text{ref}}[k,l]$ and reverberation $\mathbf{r}[k,l]$ are distributed as independent complex Gaussians $\mathcal{N}^C(0,\sigma)$, with standard deviation σ and zero-mean $\mu = 0$. Of course, the speech and reverberation covariance matrices are not just defined by this time-varying PSD, but also the stationary components $\mathbf{g}[k]$ and $\mathbf{\Gamma}[k]$, respectively. Within the ECM framework, to estimate the parameters in $\boldsymbol{\theta}[k]$ (in the i-th iteration), their likelihoods are iteratively maximized given the recorded microphone signal $\mathbf{x}[k,l]$ and the parameters in the previous $((i-1)$-th) iteration (or using initial estimates in the first iteration), i.e., the expectation of the log-likelihood function $\log\psi$ is maximized.

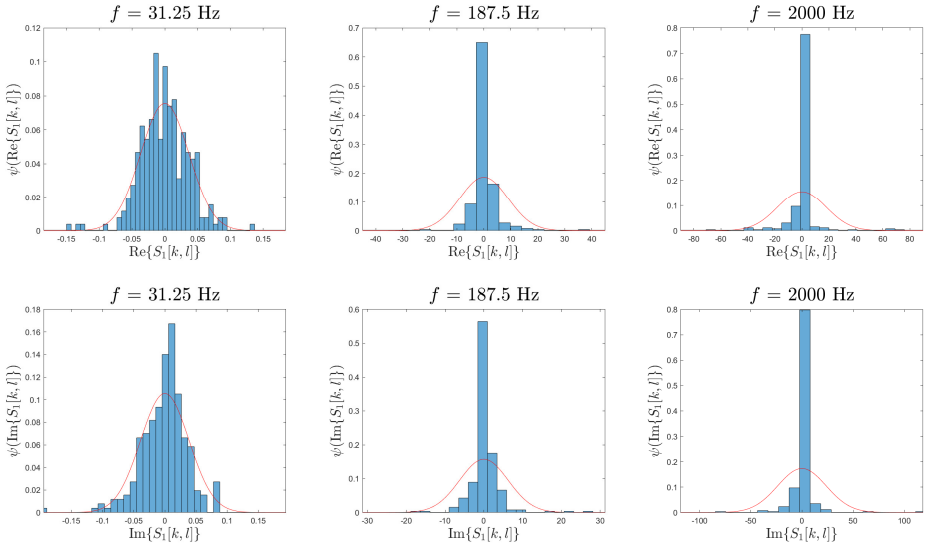


Fig. 2.6: Probability density functions (PDF) ψ of recorded, direct speech amplitudes (real- and imaginary-parts) at the frequency bins with centre-frequencies $\{31.25, 187.25, 2000\}$ Hz. Re{.} and Im{.} denote the real- and imaginary-part operators. The blue histograms show the discrete probability densities for 64 s of recorded, direct speech data and the red line depicts the Gaussian probability density function with the mean and standard deviation which best represent the recorded, direct speech.

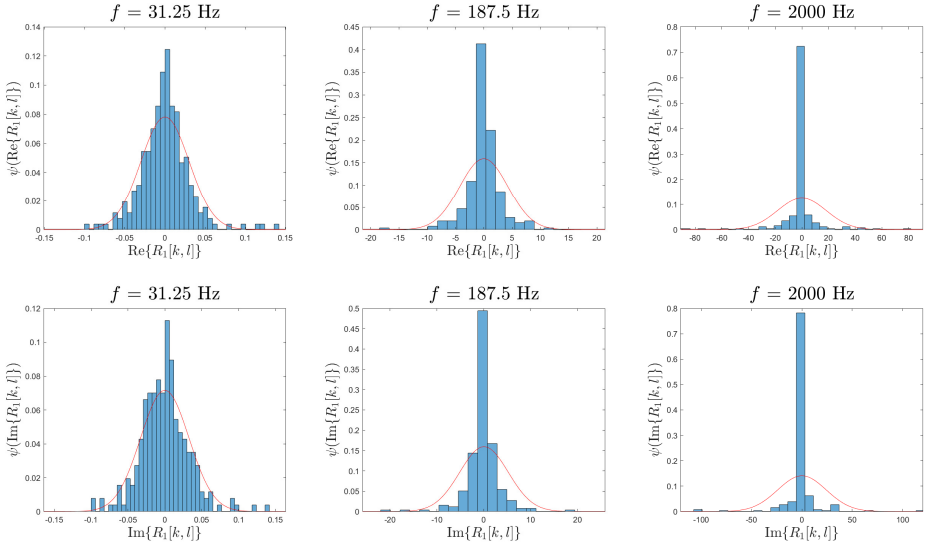


Fig. 2.7: Probability density functions (PDF) of recorded, speech reverberation amplitudes (real- and imaginary-parts) at the frequency bins with centre-frequencies $\{31.25, 187.25, 2000\}$ Hz. The blue histograms show the discrete probability densities for 64 s of recorded, speech reverberation data (recorded speech, omitting the direct path) and the red line depicts the Gaussian probability density function with the mean and standard deviation which best represent the speech reverberation.

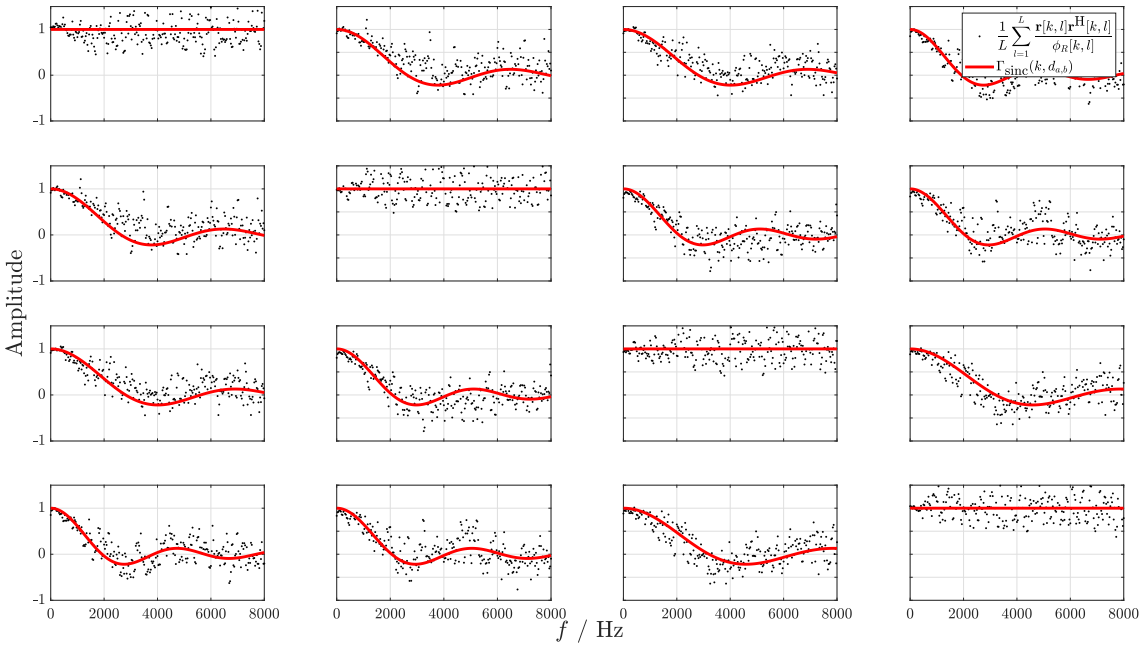


Fig. 2.8: Simulated array with $N = 4$ microphones at random positions.
In black: coherence plotted over frequency f of speech reverberation data (recorded speech, omitting the direct path), simulated using [3] with 64 s of recorded, speech reverberation data.
In red: diffuse noise model coherence for a given PD $d_{a,b}$, plotted over frequency f.

$$Q\left(\boldsymbol{\theta}[k];\hat{\boldsymbol{\theta}}_{(i-1)}[k]\right) = \mathbb{E}\left\{\log\psi(\overline{S}_{\text{ref}}[k],\overline{\mathbf{r}}[k]) \mid \overline{\mathbf{x}}[k];\hat{\boldsymbol{\theta}}_{(i-1)}[k]\right\}. \tag{2.12}$$

The PDF of the complete data $\psi(\overline{S}_{\text{ref}}[k,l],\overline{\mathbf{r}}[k,l];\boldsymbol{\theta})$ is defined as the product of the joint probabilities of the speech and reverberation

$$\psi(\overline{S}_{\text{ref}}[k,l],\overline{\mathbf{r}}[k,l];\boldsymbol{\theta}[k]) \;=\; \prod_{l=1}^{L}\psi\left(S_{\text{ref}}[k,l],\mathbf{r}[k,l];\boldsymbol{\theta}[k]\right), \tag{2.13}$$

with the joint probability defined as

$$\psi(S_{\text{ref}}[k,l],\mathbf{r}[k,l];\boldsymbol{\theta}[k]) \;=\; \mathcal{N}^{C}_{S_{\text{ref}}[k,l]}\left(0,\phi_{S_{\text{ref}}}[k,l]\right)\;\mathcal{N}^{C}_{\mathbf{r}[k,l]}\left(\mathbf{0},\Phi_{\mathbf{r}}[k,l]\right). \tag{2.14}$$

2.2.1 Initialization

To begin the ECM procedure, the estimated parameter set $\hat{\boldsymbol{\theta}}[k]$ is initialized with the maximum likelihood (ML) estimates [23, 24] for the initial speech and reverberation PSDs $\phi_{(0),S_{\text{ref}}}[k,l]$ and $\phi_{(0),R}[k,l]$, respectively, i.e.,

$$\hat{\phi}_{(0),S_{\text{ref}}}[k,l] = \mathbf{h}^{\text{H}}_{\text{MVDR}}[k]\left[\mathbf{x}[k,l]\mathbf{x}^{\text{H}}[k,l] - \hat{\phi}_{(0),R}[k,l]\hat{\boldsymbol{\Gamma}}_{(0)}[k]\right]\mathbf{h}_{\text{MVDR}}[k], \tag{2.15}$$

and

$$\hat{\phi}_{(0),R}[k,l] = \frac{1}{N-1}\mathbf{x}^{\mathrm{H}}[k,l]\ \hat{\mathbf{\Gamma}}_{(0)}^{-1}[k]\ \left[\mathbf{I} - \mathbf{g}[k]\mathbf{h}_{\mathrm{MVDR}}^{\mathrm{H}}[k]\right]\mathbf{x}[k,l]\ , \tag{2.16}$$

using the multi-channel, minimum-variance, distortionless response (MVDR) filter $\mathbf{h}_{\mathrm{MVDR}}[k]$. $\mathbf{I}$ denotes the identity matrix and $\hat{\mathbf{\Gamma}}_{(0)}[k]$ denotes the initial, estimated coherence matrix, which can be estimated a few different ways, discussed in Chapter 3.1.2.

The MVDR filter $\mathbf{h}_{\mathrm{MVDR}}[k]$ is derived by finding the filter coefficients $\mathbf{h}[k]$ which minimize the reverberation output power while preserving the direct path $\mathbf{g}[k]$ [25], i.e.,

$$\min_{\mathbf{h}[k]}\ \mathbf{h}^{\mathrm{H}}[k]\ \mathbf{\Gamma}[k]\ \mathbf{h}[k]\ \ ,\quad \text{s.t.}\quad \mathbf{h}^{\mathrm{H}}[k]\mathbf{g}[k]\ =\ 1\ . \tag{2.17}$$

Because of the decomposition in (2.9), the reverberation covariance matrix $\mathbf{\Phi}_{\mathbf{r}}[k,l]$ can be replaced with the coherence matrix $\mathbf{\Gamma}[k]$, making the MVDR a stationary filter. The solution to (2.17) is

$$\mathbf{h}_{\mathrm{MVDR}}[k] = \frac{\mathbf{\Gamma}^{-1}[k]\mathbf{g}[k]}{\mathbf{g}^{\mathrm{H}}[k]\mathbf{\Gamma}^{-1}[k]\mathbf{g}[k]}\ . \tag{2.18}$$

For the ECM initialization, $\mathbf{\Gamma}[k]$ is not known so it is replaced with the initial estimate $\hat{\mathbf{\Gamma}}_{(0)}[k]$.

2.2.2 Iteration

After the initialization, the ECM algorithm enters an iterative process which iterates I times and consists of an expectation step (E-step) and a conditional maximization step (CM-step). At the beginning of each iteration, the estimated recorded covariance matrix is assembled using available estimates of each component of the signal model in (2.3), i.e.,

$$\hat{\mathbf{\Phi}}_{(i),x}[k,l] = \hat{\phi}_{(i-1),S_{\mathrm{ref}}}[k,l]\ \mathbf{g}[k]\mathbf{g}^{\mathrm{H}}[k]\ +\ \hat{\phi}_{(i-1),R}[k,l]\ \hat{\mathbf{\Gamma}}_{(i-1)}[k]\ . \tag{2.19}$$

In the E-step, the expected speech coefficient and reverberation vector are estimated using the observed data and estimated parameters from either the initialization or the previous ($i-1$-th) iteration. For this, the multi-channel Wiener filter is used. Using these estimates, the expected speech PSD and reverberation covariance matrix are estimated. The expected speech coefficient is estimated as

$$\hat{S}_{(i),\mathrm{ref}}[k,l]\ =\ \hat{\phi}_{(i-1),S_{\mathrm{ref}}}[k,l]\ \mathbf{g}^{\mathrm{H}}[k]\ \hat{\mathbf{\Phi}}_{(i),\mathbf{x}}^{-1}[k,l]\ \mathbf{x}[k,l], \tag{2.20}$$

thus, the expected speech PSD is estimated as

$$\widehat{|S_{\text{ref}}|^2}_{(i)}[k,l] \;=\; |\hat{S}_{(i),\text{ref}}[k,l]|^2 + \hat{\phi}_{(i-1),S_{\text{ref}}}[k,l]\;\left[\mathbf{I} - \hat{\phi}_{(i-1),S_{\text{ref}}}[k,l]\;\mathbf{g}^{\text{H}}[k]\;\hat{\mathbf{\Gamma}}_{(i-1)}[k]\;\mathbf{g}[k]\right]. \quad (2.21)$$

The expected reverberation vector is estimated as

$$\hat{\mathbf{r}}_{(i)}[k,l] = \hat{\phi}_{(i-1),R}[k,l]\;\hat{\mathbf{\Gamma}}_{(i-1)}[k]\;\hat{\mathbf{\Phi}}^{-1}_{(i),\mathbf{x}}[k,l]\;\mathbf{x}[k,l], \quad (2.22)$$

thus, the reverberation covariance matrix is estimated as

$$\widehat{\mathbf{r}\mathbf{r}^{\text{H}}}_{(i)}[k,l] = \hat{\mathbf{r}}_{(i)}[k,l]\hat{\mathbf{r}}^{\text{H}}_{(i)}[k,l] + \hat{\phi}_{(i-1),R}[k,l]\;\hat{\mathbf{\Gamma}}_{(i-1)}[k]\;\left[\mathbf{I} - \hat{\mathbf{\Phi}}^{-1}_{(i),\mathbf{x}}[k,l]\;\hat{\mathbf{\Gamma}}_{(i-1)}[k]\;\hat{\phi}_{(i-1),R}[k,l]\right]. \quad (2.23)$$

In the M-step, the parameters $\boldsymbol{\theta}_{(i)}[k]$ are estimated using the observed data and expected estimated parameters. The speech PSD is estimated as

$$\hat{\phi}_{(i),S_{\text{ref}}}[k,l] = \widehat{|S_{\text{ref}}|^2}_{(i)}[k,l]\;. \quad (2.24)$$

In the CM-step, the estimation of the coherence matrix and reverberation PSD is interlaced (the estimation of $\hat{\phi}_{(i),R}[k,l]$ requires the estimate $\hat{\mathbf{\Gamma}}_{(i)}[k]$ from the same iteration i), i.e.,

$$\hat{\mathbf{\Gamma}}_{(i)}[k] \;=\; \frac{1}{L}\sum_{l=1}^{L}\frac{\widehat{\mathbf{r}\mathbf{r}^{\text{H}}}_{(i)}[k,l]}{\hat{\phi}_{(i),R}[k,l]} \quad (2.25)$$

and

$$\hat{\phi}_{(i),R}[k,l] \;=\; \widehat{\mathbf{r}\mathbf{r}^{\text{H}}}_{(i)}[k,l]\;\hat{\mathbf{\Gamma}}^{-1}_{(i)}[k]\;. \quad (2.26)$$

After iterating I times, the coherence $\mathbf{\Gamma}_{(I)}[k]$ estimated in the I-th iteration is used to estimate the PDs in Chapter 2.3.

2.2.3 Recap

An algorithm overview is summarized with pseudo-code in Algorithm 1.

Algorithm 1: ECM

Input: $\mathbf{x}[k,l]\ \forall\ k,l$, $\mathbf{g}[k]\ \forall\ k$;

% initialize:
$\hat{\phi}_{(0),S_{\text{ref}}}[k,l]\ \forall\ k,l$ using (2.15);
$\hat{\phi}_{(0),R}[k,l]\ \forall\ k,l$ using (2.16);
$\mathbf{\Gamma}_{(0)}\ \forall\ k$ using a method described in Chapter 3.1;

% iterate:
for *i = 1:I* **do**
  Estimate $\hat{\Phi}_{(i),x}[k,l]\ \forall\ k,l$ using (2.19);

  % E-step:
  for *k = 1:K* **do**
    for *l = 1:L* **do**
      $\hat{S}_{(i),\text{ref}}[k,l]$ using (2.20);
      $\widehat{|S_{\text{ref}}|^2}_{(i)}[k,l]$ using (2.21);
      $\hat{\mathbf{r}}_{(i)}[k,l]$ using (2.22);
      $\widehat{\mathbf{r}\mathbf{r}^{\text{H}}}_{(i)}[k,l]$ using (2.23);
    end
  end

  % CM-step:
  for *k = 1:K* **do**
    for *l = 1:L* **do**
      $\hat{\phi}_{(i),S_{\text{ref}}}[k,l]$ using (2.24);
      $\hat{\mathbf{\Gamma}}_{(i)}[k]$ using (2.25);
      $\hat{\phi}_{(i),R}[k,l]$ using (2.26);
    end
  end
end
return $\hat{\mathbf{\Gamma}}_{(I)}[k]$;

2.2.4 Practical Considerations

A few measures are important when implementing the ECM algorithm. Some are based on practical considerations mentioned in [1] and others are to avoid numerical or other unwanted errors. They are briefly listed in the following:

Numerical corrections

1. Enforce Hermitian structure for any matrix which should be Hermitian, e.g., matrix $\mathbf{A}$
$$\mathbf{A} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^{\mathrm{H}}) \tag{2.27}$$

2. Force any variables, which should be real-valued, to be real-valued, e.g. $\hat{\phi}_{(i),S,\mathrm{ref}}$, $\hat{\phi}_{(i),R}$, the entries of $\hat{\mathbf{\Gamma}}_{(i)}[k]$, and even the denominator of the MVDR filter in (2.18), to prevent the imaginary-valued parts from growing and introducing more estimation errors as the code iterates.

Heuristics

1. Before computing the inverse of any matrix $\mathbf{A}$, apply regularization via diagonal loading [26, 27], i.e.,
$$\mathbf{A} = \mathbf{A} + \mathbf{I}\,\frac{1}{N}\mathrm{Tr}(\mathbf{A}F + \mathrm{eps}\mathbf{1}\mathbf{1}^{\mathrm{T}}) \;, \tag{2.28}$$
with Tr(.) denoting the Trace operator, F a loading factor (typically in the region of $L = 0.05$), $\mathbf{1}$ denoting a column-vector of ones, and a lower-bound $\mathrm{eps} = 10^{-6}$ when dealing with .wav files.

2. Set an upper-bound for the reverberation PSD estimate $\hat{\phi}_{(i),R}[k,l]$
$$\hat{\phi}_{(i),R}[k,l] = \min(\; \frac{1}{N}\mathbf{x}^{\mathrm{H}}[k,l]\mathbf{x}[k,l] \;,\;\; \hat{\phi}_{(i),R}[k,l] \;) \;, \tag{2.29}$$
based on the average recorded signal PSD, which includes both the direct speech and reverberation PSDs.

3. In a given frequency bin k, only use data from relevant time-frames l to estimate stationary parameters (i.e., $\hat{\mathbf{\Gamma}}_{(i)}[k]$). The relevant time-frames for the estimation of $\mathbf{\Gamma}[k]$ in (2.25) are those where the reverberant-to-direct ratio (RDR) at a given time frame and frequency bin $\mathrm{RDR}[k,l]$ is above a defined threshold ν_{RDR}, i.e.,
$$\widehat{\mathrm{RDR}}[k,l] = \frac{\hat{\phi}_{(i),R,\mathrm{ref}}[k,l]}{\hat{\phi}_{(i),S,\mathrm{ref}}[k,l]} > \nu_{\mathrm{RDR}} \tag{2.30}$$

4. Set a lower bound for all estimated PSDs

$$\hat{\phi}_{(i),S,\text{ref}} = \max(\ \hat{\phi}_{(i),S,\text{ref}}\ ,\ \ \text{eps}\ ) \tag{2.31a}$$

$$\hat{\phi}_{(i),R} = \max(\ \hat{\phi}_{(i),R}\ \ \ ,\ \ \text{eps}\ ) \tag{2.31b}$$

   in order to avoid dividing by 0, e.g., in (2.25).

5. Normalize coherence estimate $\hat{\mathbf{\Gamma}}_{(i)}[k]$ by its trace

$$\hat{\mathbf{\Gamma}}_{(i)}[k] = \frac{\hat{\mathbf{\Gamma}}_{(i)}[k]}{\frac{1}{N}\text{Tr}(\hat{\mathbf{\Gamma}}_{(i)}[k])} \tag{2.32}$$

   Informal tests indicated that omitting the division by the denominator in (2.25) before this practical consideration resulted in a more robust coherence estimation (due to reducing the number of divisions by small numbers).

2.3 PD Estimation

Once the coherence has been estimated using the ECM algorithm in Chapter 2.2, the pairwise distances can be found by minimizing the difference between the estimated coherence $\hat{\Gamma}_{(I),a,b}[k]$ between microphones a and b, and the model coherence function $\Gamma(k,d)$ (which, importantly, is a function of distance). In [13, 14] it was shown that the PD $\hat{d}_{a,b}$ could be estimated as the distance d for which the model coherence best matches the estimated coherence over a set of selected frequency bins $k \in \mathcal{K}$, i.e.,

$$\hat{d}_{a,b} = \underset{d}{\text{argmin}} \left\{ \sum_{k\in\mathcal{K}} |\hat{\Gamma}_{(I),a,b}[k] - \Gamma(k,d))|^2 \right\} . \tag{2.33}$$

The set of frequency bins is discussed further in Chapter 3.3.

2.4 MAG Estimation Using MDS

In Chapters 2.2 and 2.3 it was shown how to estimate the coherence and the pairwise distances between microphones, respectively. MDS computes relative coordinates in P-dimensional space (i.e., coordinates arbitrarily rotated, translated, or reflected in relation to the true coordinates) from squared pairwise distances. So in this chapter is is shown how the estimated distances can be used to estimate the MAG. The reason why only the MAG (i.e., the relative coordinates) is estimated and not the true coordinates is because the information describing the rotation, translation, and reflection, relative to the source or the room is lost when describing the problem in terms of PDs.

Rather than using the distances $d_{a,b}$ in the PD matrix $\mathbf{P} = [d_{a,b}]$, the squared distances $d_{a,b}^2$ are used in the Euclidean distance matrix (EDM) $\mathbf{D}_{\text{EDM}} = [d_{a,b}^2]$ in order to be able to exploit a rank property in the eigenvalue decomposition [16, 28]. The EDM can be constructed from the individual squared PDs $d_{a,b}^2$ or in matrix form, i.e.,

$$\mathbf{D}_{\text{EDM}} = [d_{a,b}^2] = \underbrace{\mathbf{1}\text{diag}(\mathbf{M}\mathbf{M}^{\text{T}})^{\text{T}}}_{\text{rank}=1} - \underbrace{2\mathbf{M}\mathbf{M}^{\text{T}}}_{\text{rank}\leq P} + \underbrace{\text{diag}(\mathbf{M}\mathbf{M}^{\text{T}})\mathbf{1}^{\text{T}}}_{\text{rank}=1}, \tag{2.34}$$

with diag(.) denoting the diagonal-element operator, which returns the diagonal elements of a square matrix as a column-vector. (2.34) clearly shows that the rank of an EDM can be at most $P + 2$ as stated in [28], i.e.,

$$\text{Rank}(\mathbf{D}_{\text{EDM}}) \leq P + 2 \; . \tag{2.35}$$

This coincides with the expectation because the underlying coordinate matrix $\mathbf{M}$ is at most rank P, regardless of the number of microphones. The matrix terms $\mathbf{1}\text{diag}(\mathbf{M}\mathbf{M}^{\text{T}})^{\text{T}}$ and $\text{diag}(\mathbf{M}\mathbf{M}^{\text{T}})\mathbf{1}^{\text{T}}$ are translation operations and can not be directly estimated without knowing $\mathbf{M}$, however, other transformations can be applied which preserve the geometry of the remaining term with a dependence on $\mathbf{M}$. Reformulating (2.34) as follows

$$\begin{aligned}
\mathbf{M}\mathbf{M}^{\text{T}} &= -\frac{1}{2}(\mathbf{D}_{\text{EDM}} - \mathbf{1}\text{diag}(\mathbf{M}\mathbf{M}^{\text{T}})^{\text{T}} - \text{diag}(\mathbf{M}\mathbf{M}^{\text{T}})\mathbf{1}^{\text{T}}) && \text{(2.36a)} \\
&= -\frac{1}{2}(\mathbf{D}_{\text{EDM}} - \mathbf{1}\mathbf{d}_1 - \mathbf{d}_1\mathbf{1}^{\text{T}}) && \text{(2.36b)} \\
&= \mathbf{G} \; , && \text{(2.36c)}
\end{aligned}$$

with each element of

$$\mathbf{1}\text{diag}(\mathbf{M}\mathbf{M}^{\text{T}})^{\text{T}} = \mathbf{1}\mathbf{d}_1^{\text{T}} \; , \tag{2.37}$$

leaves the Gram matrix $\mathbf{G}$ with $\text{rank}\,\mathbf{G} \leq P$. It was shown in [29] that any rank 2 matrix subtraction from $\mathbf{D}_{\text{EDM}}$, structured like the one in (2.36), that leaves $\mathbf{G}$ positive semi-definite also ensures that $\mathbf{G}$ is a Gram matrix and can be replaced by a multiplication with $\mathbf{D}_{\text{EDM}}$ from both sides, i.e.,

Theorem 2.1 *$\mathbf{D}_{EDM}$ is an EDM iff*

$$-\frac{1}{2}(\mathbf{I} - \mathbf{1}\mathbf{z}^T)\mathbf{D}_{EDM}(\mathbf{I} - \mathbf{z}\mathbf{1}^T)$$

is positive semi-definite for any $\mathbf{z}$ such that $\mathbf{z}^T\mathbf{1} = 1$ and $\mathbf{z}^T\mathbf{D}_{EDM} \neq 0$.

Applying Theorem 2.1 yields a Gram matrix $\mathbf{G}_{\text{rel}} = \mathbf{M}_{\text{rel}}\mathbf{M}_{\text{rel}}^{\text{T}}$ where the underlying co-ordinates $\mathbf{M}_{\text{rel}}$ are translated by $\mathbf{z}$ in addition to the arbitrary rotation or

reflection from expressing the problem in terms of distances, i.e.,

$$\mathbf{G}_{\text{rel}} = -\frac{1}{2}(\mathbf{I} - \mathbf{1}\mathbf{z}^{\text{T}})\mathbf{D}_{\text{EDM}}(\mathbf{I} - \mathbf{z}\mathbf{1}^{\text{T}}) \ . \tag{2.38}$$

As shown in [28], $\mathbf{z} = \mathbf{e}_1$ in (2.38) is equivalent to using $\mathbf{d}_1$ in (2.36), centering the co-ordinates $\mathbf{M}_{\text{rel}}$ or $\mathbf{M}$, respectively, at the origin. A commonly used alternative to $\mathbf{z}$ is the geometric centering matrix $\mathbf{z}_{\text{c}}$, i.e.,

$$\mathbf{z}_{\text{c}} = \frac{1}{N}\mathbf{1}^{\text{T}} \ , \tag{2.39}$$

which translates the underlying coordinates $\mathbf{M}_{\text{rel}}$ such that their centroid is at the origin (instead of the reference microphone). With $\mathbf{G}_{\text{rel}}$, the coordinates can be estimated with the help of an eigenvalue decomposition

$$\mathbf{G}_{\text{rel}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\text{T}}. \tag{2.40}$$

After sorting the eigenvalues in order of decreasing magnitude, $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_N|$, the P-dimensional coordinates are obtained by truncating the eigenvalues to the P largest eigenvalues (if the PDs are not corrupted by estimation errors then there should be at most $\min(P, N)$ non-zero eigenvalues)

$$\hat{\mathbf{M}}_{\text{rel}} = \mathbf{U}[\text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_P}), \mathbf{0}_{P\times(N-P)}]^{\text{T}} \ , \tag{2.41}$$

with $\mathbf{0}_{P\times(N-P)}$ a $P \times (N-P)$-dimensional matrix of zeros. Each step of the algorithm is summarized with pseudo-code in Algorithm 2.

Algorithm 2: MDS

Input: $\hat{\mathbf{D}}_{\text{EDM}}$, P;
$\hat{\mathbf{G}}_{\text{rel}} \leftarrow -\frac{1}{2}(\mathbf{I} - \mathbf{1}\mathbf{z}_{\text{c}}^{\text{T}})\hat{\mathbf{D}}_{\text{EDM}}(\mathbf{I} - \mathbf{z}_{\text{c}}\mathbf{1}^{\text{T}})$;
$\mathbf{U}\Lambda\mathbf{U}^{\text{T}} \leftarrow \hat{\mathbf{G}}_{\text{rel}}$;
$\hat{\mathbf{M}}_{\text{rel}} \leftarrow \mathbf{U}[\text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_P}), \mathbf{0}_{P\times(N-P)}]^{\text{T}}$;
return $\hat{\mathbf{M}}_{\text{rel}}$;

In order to analyse the similarity of the EDM $\mathbf{D}_{\text{EDM}}$ with the low-rank approximation based on the estimated, relative MAG $\hat{\mathbf{M}}_{\text{rel}}$, the EDM can be reconstructed using (2.34), i.e.,

$$\hat{\mathbf{D}}_{\text{EDM,MDS}} = \mathbf{1}\text{diag}(\hat{\mathbf{M}}_{\text{rel}}\hat{\mathbf{M}}_{\text{rel}}^{\text{T}})^{\text{T}} - 2\hat{\mathbf{M}}_{\text{rel}}\hat{\mathbf{M}}_{\text{rel}}^{\text{T}} + \text{diag}(\hat{\mathbf{M}}_{\text{rel}}\hat{\mathbf{M}}_{\text{rel}}^{\text{T}})\mathbf{1}^{\text{T}} \ . \tag{2.42}$$

The difference between (2.34) and (2.42) is that if $\mathbf{D}_{\text{EDM}}$ has any estimation errors,

they can increase its rank and the rank of the underlying Gram matrix $\mathbf{G}$, while (2.42) is constructed based on an approximation of $\mathbf{G}_{\text{rel}}$ which is at most rank P.

2.5 Overview of Complete MAG Estimation Framework

A summary of the whole MAG estimation algorithm is presented in this section. The most important parameter estimation steps are presented in Fig. 2.9. First, the recorded reverberant speech in the STFT domain and the initial coherence estimate are passed into the ECM algorithm. As well as these inputs, the RDTF is also required, which in this thesis is assumed to be known (unless otherwise stated), but can be estimated using GCC-PHAT as in [8]. After I iterations of the ECM algorithm, ECM is terminated and the estimated coherence is passed to the PD estimation step. The PD is determined by finding the model coherence which best matches the estimated coherence. Using the squared estimated PDs, the MAG can be estimated using MDS.

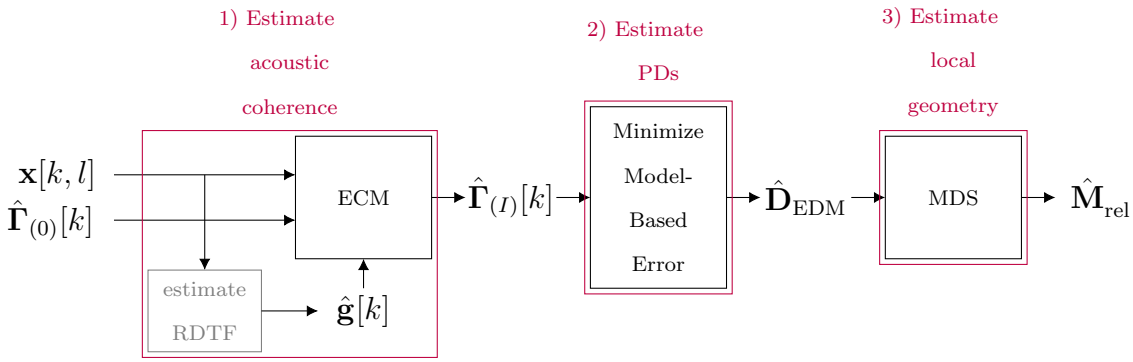


Fig. 2.9: MAG estimation block-diagram. Since in this work (unless otherwise stated) it is assumed that the RDTF $\mathbf{g}[k]$ is available, the "estimate RDTF" block is greyed-out.

2.6 Realigning the Coordinates

For the purpose of evaluating the accuracy of the estimated MAG $\hat{\mathbf{M}}_{\text{rel}}$, which is arbitrarily rotated, translated, and reflected in reference to the true coordinates $\mathbf{M}$, the whole geometry must be aligned with the true coordinates to find the error of each microphone. This problem, where a matrix must be mapped to another as closely as possible is called the orthogonal Procrustes problem, the solution to which is presented in [30]. For related applications of this solution, see [7, 28].

The alignment of the relative estimated microphone positions $\hat{\mathbf{M}}_{\text{rel}}$ with the true positions $\mathbf{M}$ is simplified by manually setting the origin as a common point-of-reference as it is a point which can be defined in both coordinate systems, e.g., the centroid or a specific microphone. For this, the centroids (mean coordinate) $\mathbf{m}_{\text{c}}$ and

$\mathbf{m}_{\text{rel,c}}$ are subtracted from (each column of) the matrices $\mathbf{M}$ and $\hat{\mathbf{M}}_{\text{rel}}$, respectively, such that their centroids are at the origin i.e.,

$$\begin{aligned} \mathbf{M}_{\text{c}} &= \mathbf{M} - \mathbf{1}\mathbf{m}_{\text{c}}^{\text{T}} \ , \\ \hat{\mathbf{M}}_{\text{rel,c}} &= \hat{\mathbf{M}}_{\text{rel}} - \mathbf{1}\hat{\mathbf{m}}_{\text{rel,c}}^{\text{T}} \ , \end{aligned} \tag{2.43}$$

placing the centroids of both matrices at the origin. Thus the alignment problem is no longer about translation, but instead, just about rotating/reflecting $\hat{\mathbf{M}}_{\text{rel,c}}$ to $\mathbf{M}_{\text{c}}$, which can be applied as a matrix operation $\mathbf{Q}$, i.e.,

$$\begin{aligned} \mathbf{Q} \quad = \quad & \underset{\mathbf{Q}}{\text{argmin}} \quad ||\hat{\mathbf{M}}_{\text{rel,c}}\mathbf{Q} - \mathbf{M}_{\text{c}}||_{\text{F}}^2 \\ & \text{s.t. } \mathbf{Q}\mathbf{Q}^{\text{T}} = \mathbf{I} \end{aligned} \tag{2.44}$$

with $||.||_{\text{F}}$ the Frobenius norm. The condition $\mathbf{Q}\mathbf{Q}^{\text{T}} = \mathbf{I}$ must be fulfilled for $\mathbf{Q}$ to not apply any translation or scaling, i.e., to be a rotation and/or reflection operation. The solution to (2.44) can be found by using the singular-value decomposition (SVD) of $\hat{\mathbf{M}}_{\text{rel,c}}^{\text{T}}\mathbf{M}_{\text{c}}$, i.e.,

$$\hat{\mathbf{M}}_{\text{rel,c}}^{\text{T}}\mathbf{M}_{\text{c}} = \mathbf{U}_{\text{SVD}}\mathbf{\Sigma}\mathbf{V}_{\text{SVD}}^{\text{T}} \ . \tag{2.45}$$

Rewriting (2.44) using $||\hat{\mathbf{M}}_{\text{rel,c}}\mathbf{Q} - \mathbf{M}_{\text{c}}||_{\text{F}}^2 = \text{Tr}\left((\hat{\mathbf{M}}_{\text{rel,c}}\mathbf{Q} - \mathbf{M}_{\text{c}})^{\text{T}}(\hat{\mathbf{M}}_{\text{rel,c}}\mathbf{Q} - \mathbf{M}_{\text{c}})\right)$ reduces the number of terms which influence the minimization, i.e.,

$$\begin{aligned} \mathbf{Q} \quad = \quad & \underset{\mathbf{Q}}{\text{argmin}} \quad \left\{-\text{Tr}\left(\mathbf{Q}^{\text{T}}\hat{\mathbf{M}}_{\text{rel,c}}^{\text{T}}\mathbf{M}_{\text{c}}\right)\right\} && \text{(2.46a)} \\ & \text{s.t. } \mathbf{Q}\mathbf{Q}^{\text{T}} = \mathbf{I} \\ = \quad & \underset{\mathbf{Q}}{\text{argmax}} \quad \left\{\text{Tr}\left(\mathbf{Q}^{\text{T}}\mathbf{U}_{\text{SVD}}\mathbf{\Sigma}\mathbf{V}_{\text{SVD}}^{\text{T}}\right)\right\} && \text{(2.46b)} \\ & \text{s.t. } \mathbf{Q}\mathbf{Q}^{\text{T}} = \mathbf{I} \\ = \quad & \underset{\mathbf{Q}}{\text{argmax}} \quad \left\{\text{Tr}\left(\mathbf{V}_{\text{SVD}}^{\text{T}}\mathbf{Q}^{\text{T}}\mathbf{U}_{\text{SVD}}\mathbf{\Sigma}\right)\right\} \ . && \text{(2.46c)} \\ & \text{s.t. } \mathbf{Q}\mathbf{Q}^{\text{T}} = \mathbf{I} \end{aligned}$$

Finding the solution involves exploiting the diagonal elements of the orthogonal matrix $\mathbf{V}_{\text{SVD}}^{\text{T}}\mathbf{Q}^{\text{T}}\mathbf{U}_{\text{SVD}}$, which can not exceed 1 due to the constraint. Thus, a solution which maximizes $\text{Tr}\left(\mathbf{V}_{\text{SVD}}^{\text{T}}\mathbf{Q}^{\text{T}}\mathbf{U}_{\text{SVD}}\mathbf{\Sigma}\right)$ in (2.46) involves setting $\mathbf{V}_{\text{SVD}}^{\text{T}}\mathbf{Q}^{\text{T}}\mathbf{U}_{\text{SVD}} = \mathbf{I}$, i.e.,

$$\mathbf{Q}^{\text{T}} = \mathbf{V}_{\text{SVD}}\,\mathbf{U}_{\text{SVD}}^{\text{T}} \ . \tag{2.47}$$

To evaluate how accurate the estimated MAG $\hat{\mathbf{M}}_{\text{rel}}$ is, compared to the true coordinates $\mathbf{M}$, the estimated MAG is centered at the origin as in (2.43), the optimal

$\mathbf{Q}$ is found with (2.47), and then the rotated estimated MAG $\hat{\mathbf{M}}_{\mathrm{rel,c}}\mathbf{Q}$ is compared with $\mathbf{M}_{\mathrm{c}}$. Error measures are discussed in Chapter 4.

3 Proposed Analyses and Changes to the MAG Estimation

Building on the theory discussed in Chapter 2, in this Chapter, some modifications are proposed to the state-of-the-art MAG estimation as well as parameters to be analysed. Changes to the initialization of the ECM algorithm from Chapter 2.2 are proposed in Chapter 3.1. A proposed analysis of the effect of estimation errors of the RDTF $\mathbf{g}[k]$, used in ECM, is described in Chapter 3.2. A proposed analysis of the frequency range used in the PD estimation in Chapter 2.3 is discussed in Chapter 3.3 and an alternative coherence model is suggested for estimating the distance between hearing-aids. In Chapter 3.4, a method is proposed to incorporate prior knowledge in MDS with the aim of improving the MAG estimation accuracy.

3.1 ECM Initialization

Although the ECM algorithm maximizes the likelihood in each iteration, it is not guaranteed to converge to the global-optimum solution after several iterations. This is because it could arrive in a local optimum. This means that the initialization of the ECM algorithm is important, because by initializing the parameters more closely to the global optimum, they are more likely to converge to the global optimum than a local optimum. Two changes are proposed to the initialization. In Chapter 3.1.1, it is proposed to use a different filter for estimating the initial speech and reverberation PSDs and in Chapter 3.1.2 alternative initial coherence estimates are proposed.

3.1.1 Filter for PSD Initialization

When initializing the ECM algorithm, the initial speech and reverberation PSDs are estimated using an MVDR filter in (2.18). In this work it is assumed that the direct and reverberant speech components are uncorrelated, therefore the MVDR filter is equivalent to the minimum-power distortionless response (MPDR) filter $\mathbf{h}_{\text{MPDR}}[k]$ [25], which aims to minimize the total output power while preserving the direct path, i.e.,

$$\begin{aligned} &\min_{\mathbf{h}[k]} \ \mathbf{h}^{\text{H}}[k]\ \mathbf{\Phi}_{\mathbf{x}}[k,l]\ \mathbf{h}[k]\ , \\ &\text{s.t.} \quad \mathbf{h}^{\text{H}}[k]\mathbf{g}[k]\ =\ 1\ . \end{aligned} \tag{3.1}$$

The solution to (3.1) is

$$\mathbf{h}_{\text{MPDR}}[k,l] = \frac{\mathbf{\Phi}_{\mathbf{x}}^{-1}[k,l]\mathbf{g}[k]}{\mathbf{g}^{\text{H}}[k]\mathbf{\Phi}_{\mathbf{x}}^{-1}[k,l]\mathbf{g}[k]}\ . \tag{3.2}$$

This equivalence of course only holds if there are no estimation errors in $\mathbf{g}[k]$, $\mathbf{\Phi}_{\mathbf{x}}^{-1}[k,l]$, or $\mathbf{\Gamma}[k]$. In this instance, the MPDR filter is arguably a better choice,

because the coherence $\mathbf{\Gamma}[k]$ or an estimate thereof is not available in the ECM initialization whereas $\mathbf{\Phi}_{\mathbf{x}}[k,l]$ can be estimated directly from the reverberant speech data $\mathbf{x}[k,l]$ by averaging over time. The MAG estimation capabilities of both filters are compared in Chapter 4.1.

Two Implementations are considered, the first is a stationary implementation of the filter $\widehat{\mathbf{h}_{\text{MPDR-S}}}[k]$ which relies on a stationary covariance matrix estimate

$$\widehat{\mathbf{h}_{\text{MPDR-S}}}[k] = \frac{\left(\frac{1}{L}\sum_{l=1}^{L}\mathbf{x}[k,l]\mathbf{x}^{\text{H}}[k,l]\right)^{-1}\mathbf{g}[k]}{\mathbf{g}^{\text{H}}[k]\left(\frac{1}{L}\sum_{l=1}^{L}\mathbf{x}[k,l]\mathbf{x}^{\text{H}}[k,l]\right)^{-1}\mathbf{g}[k]} \tag{3.3}$$

and the second is a time-varying implementation $\widehat{\mathbf{h}_{\text{MPDR-TV}}}[k,l]$ which relies on a recursively-smoothed covariance matrix estimate

$$\hat{\mathbf{\Phi}}_{\mathbf{x}}[k,l] = \rho\hat{\mathbf{\Phi}}_{\mathbf{x}}[k,l-1] + (1-\rho)\mathbf{x}[k,l]\mathbf{x}^{\text{H}}[k,l] \tag{3.4a}$$

$$\widehat{\mathbf{h}_{\text{MPDR-TV}}}[k,l] = \frac{\hat{\mathbf{\Phi}}_{\mathbf{x}}^{-1}[k,l]\mathbf{g}[k]}{\mathbf{g}^{\text{H}}[k]\hat{\mathbf{\Phi}}_{\mathbf{x}}^{-1}[k,l]\mathbf{g}[k]}\,, \tag{3.4b}$$

with smoothing parameter $\rho \in [0,1]$. The reason why it is implemented as a time-varying covariance matrix is because although $\mathbf{\Gamma}[k]$ is stationary, $\mathbf{\Phi}[k,l]$ is not, because of (2.9). This means that to take the variation of the speech and reverberation PSDs into account, the filter (3.4) should also be time-varying. To apply the time-varying MPDR-TV filter, the stationary filter $\mathbf{h}_{\text{MVDR}}[k]$ in (2.15) and (2.16) is replaced with $\widehat{\mathbf{h}_{\text{MPDR-TV}}}[k,l]$.

Another important consideration is the initialization of the coherence estimate $\hat{\mathbf{\Gamma}}_{(0)}[k]$. Without a-priori knowledge about the PDs, the choice of initializations is quite limited. In the following, a few blind coherence initializations are presented as well as some coherence initializations which rely on oracle knowledge (i.e., access to latent signal components which are not available in practice).

3.1.2 Coherence Initializations

Blind Coherence Initializations:

- Identity matrix (state-of-the-art in [14]):

$$\hat{\mathbf{\Gamma}}_{(0),\mathbf{I}}[k] = \mathbf{I}\,. \tag{3.5}$$

  Initializing using the identity matrix is a *safe bet*, however, when referring to (2.7), it is seen that if the coherence matrix is an identity matrix it means that

the PDs are inifinitely large.

- Deltas

$$\boldsymbol{\Delta} = [0, \delta_2, \ldots, \delta_N]^{\mathrm{T}} = c\boldsymbol{\tau} \ , \tag{3.6}$$

with TDoAs $\boldsymbol{\tau} = [0, \tau_2, \ldots, \tau_N]^{\mathrm{T}}$. The initial coherence estimate using $\boldsymbol{\Delta}$ is defined as

$$\hat{\boldsymbol{\Gamma}}_{(0),\boldsymbol{\Delta}}[k] = \mathrm{sinc}(\frac{\varepsilon k}{K}|\boldsymbol{\Delta}\mathbf{1}^{\mathrm{T}} - \mathbf{1}\boldsymbol{\Delta}^{\mathrm{T}}|_{\mathrm{mtx}})_{\mathrm{mtx}} \tag{3.7}$$

where $|.|_{\mathrm{mtx}}$ denotes the absolute value operator applied to each entry of the matrix and similarly, $\mathrm{sinc}(.)_{\mathrm{mtx}}$ is the sinc function applied to each entry of the matrix.

The motivation behind this initialization is to initialize the coherence closer to the true coherence, compared to (3.5) where the coherence represents infinitely large PDs. For this, a distance, which is a lower-bound of the true PD is used, directly proportional to the TDoA. An example array with labeled ds and $\boldsymbol{\Delta}$s is presented in Fig. 3.1. Although $\boldsymbol{\Delta}$s represents a lower-bound for each PD, it is not known exactly how close it is to the true PD because the angle between the line connecting a pair of microphones and the axis of propagation is not known. This means that relative to the model coherence, the coherence will be *stretched* over frequency.

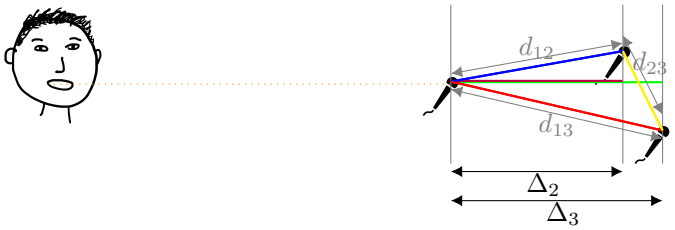


Fig. 3.1: Visual representation of the entries of $\boldsymbol{\Delta}$ for a 2-dimensional array with $N = 3$ microphones

Since the RDTF $\mathbf{g}[k]$ is assumed to be known in this work, it is assumed that $\boldsymbol{\tau}$ is available, however, in practice, $\boldsymbol{\tau}$ can be blindly estimated using GCC-PHAT as in [8].

- Recorded Signal covariance matrix:

$$\hat{\boldsymbol{\Gamma}}_{(0),\Phi_{\mathbf{x}}}[k] = \frac{1}{L}\sum_{l=1}^{L} \frac{\mathbf{x}[k,l]\mathbf{x}^{\mathrm{H}}[k,l]}{\frac{1}{N}\mathrm{Tr}(\mathbf{x}[k,l]\mathbf{x}^{\mathrm{H}}[k,l])} \tag{3.8}$$

Although this estimate obviously includes the direct path, the motivation behind this proposed initialization is that it could be closer to the true coherence

matrix than $\mathbf{I}$.

Coherence Initializations Using Oracle Knowledge:

- sinc

$$\hat{\mathbf{\Gamma}}_{(0),\text{sinc}}[k] = \text{sinc}(\frac{\varepsilon k}{K}\mathbf{P})_{\text{mtx}} \tag{3.9}$$

with $\mathbf{P} = [d_{a,b}]$ the pairwise distance matrix containing the pairwise distances $d_{a,b}$ between each microphone combination a and b.

- Signal reverberation

$$\hat{\mathbf{\Gamma}}_{(0),\text{rev}}[k] = \frac{1}{L}\sum_{l=1}^{L}\frac{\mathbf{r}[k,l]\mathbf{r}^{\text{H}}[k,l]}{\frac{1}{N}\text{Tr}(\mathbf{r}[k,l]\mathbf{r}^{\text{H}}[k,l])} \tag{3.10}$$

In this initialization $\mathbf{r}[k,l]$ is generated using the STFT of the speech signal convolved with the complete RIR excluding the direct path.

The different coherence initializations are compared in Chapter 4.1.

3.2 Proposed Analysis of Erroneous RDTF Estimation

Since in practice, the RDTF must be blindly estimated, e.g., with TDoAs, estimated using GCC-PHAT [8], the TDoAs could contain estimation errors, affecting the estimation of parameters such as the coherence in ECM. A controlled analysis of the influence of TDoA errors is proposed to show how well the MAG can still be estimated with erroneously estimated RDTFs.

The TDoA, in samples, between the reference microphone and the n-th microphone, relative to the source, τ_n, can be determined directly from the geometry with

$$\tau_n = \frac{f_s}{c}(||\mathbf{q} - \mathbf{e}_n^{\text{T}}\mathbf{M}||_2 - ||\mathbf{q} - \mathbf{e}_1^{\text{T}}\mathbf{M}||_2)\ . \tag{3.11}$$

using the difference in 2-norms $||.||_2$. With these TDoAs, the RDTF is estimated using (2.2). To simulate erroneous RDTF estimation due to errors in the TDoAs, it is proposed to add Gaussian-distributed error with a zero-mean and standard-deviation of σ_τ samples to each entry τ_n (except for $\tau_1 = 0$).

3.3 Proposed Analysis and Changes to the Coherence-Based PD Estimation

This Chapter builds on Chapter 2.3, where the PD is estimated by finding the model-based coherence which best fits the estimated coherence for a set of frequencies. The frequency selection is an important aspect of the PD estimation, since

the coherence estimation in (2.25) in each frequency bin depends on how well the estimated coherence matches the model coherence.

The set of frequency bins used in the state-of-the-art in [14] contained frequency bins which corresponded to frequencies between $[1, 3]$ kHz. Also, in Fig. 2.8 it is seen that even using simulated reverberant data, the data-based coherence deviates from the model coherence which describes a diffuse sound field. This raises the need for an analysis of the frequency range, in order to investigate how to determine the best frequency range and whether this is dependent on the distance between two microphones.

A modification to the model-based coherence function for hearing-aid distance estimation is suggested in Chapter 3.3.1. The results of this analysis are presented in Chapter 4.3.

3.3.1 Modifications for Hearing Aid Distance Estimation

Till now, it has been assumed that the microphones are in free-field. This means that in a highly reverberant room, a simple coherence model $\Gamma(k, d)$ could be used in (2.33) such as the spherical isotropic model in (2.7). In [14], a cylindrical isotropic coherence model was proposed for (shoebox) rooms where the floor and ceiling were more absorbing/less reflective (c.f. [31]) than the other walls. To analytically model the changes that a head between hearing-aids applies to the sinc-coherence, a modification was proposed in [2] to the sinc coherence to analytically describe the changes to the , the so-called inter-aural coherence (IC)

$$\Gamma_{\text{msinc}}(k, d_{a,b}) = \text{sinc}\left(\alpha \frac{\varepsilon k d_{a,b}}{K}\right) \frac{1}{\sqrt{1 + \left(\beta \frac{\varepsilon k d_{a,b}}{K}\right)^4}} \,. \tag{3.12}$$

The parameter α compresses the sinc along the frequency axis and the parameter β the sinc increasingly at high frequencies.

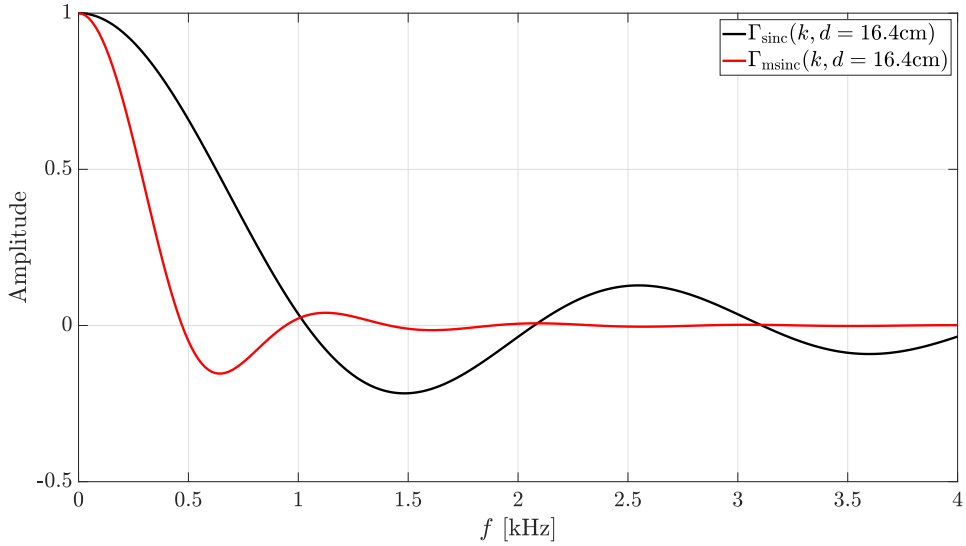


Fig. 3.2: Comparison of coherence models. The free-field sinc is plotted in black and the modified sinc is plotted in red, using $\alpha = 2.2$ and $\beta = 0.5$.

3.4 Exploiting Prior Geometry Knowledge in MAG Estimation with MDS

Modern hearing-aids often have more than one microphone per device. Since the distances between microphones of a single hearing-aid are already known and do not change, it should be possible to exploit this information in the MAG estimation. Another situation where it could be useful to incorporate other available knowledge than the estimated PDs is if there is a reliable TDoA estimate relative to a source in the far-field. The far-field TDoAs are directly proportional to coordinates in one dimension of $\mathbb{R}^P$ so if the TDoA data is more reliable than the data from the reverberation (e.g. if the DRR of the recorded signal is high due to acoustically absorbing walls, floor, and ceiling), then incorporating the more reliable spatial information would lead to a more accurate MAG estimate. Incorporating the TDoAs also has the added benefit that it orients one of the coordinate axes, which describe the relative MAG, relative to the source, i.e., acting as a DoA estimate.

To incorporate the prior-knowledge in a hearing-aid scenario, it is assumed that the hearing-aids are parallel to each-other, this means that the entire MAG only has one degree-of-freedom, i.e. only the distance between hearing-aids is variable. The idea is that the known information is subtracted (from the Gram matrix) before applying MDS, and then *added in again* afterwards. If using the TDoAs, this same procedure can be applied by using the TDoAs to estimate the spacing between microphones, relative to the source. These spacings *are* coordinates when the source is in the far-field. Hence, similarly to using prior knowledge in the hearing-aid scenario, the

coordinates estimated using the TDoAs are subtracted before applying MDS, and then replaced afterwards.

In (2.40), it was seen that the Gram matrix $\mathbf{G}$ (related to $\mathbf{D}_{\text{EDM}}$ via (2.38)) can be decomposed via an eigenvalue decomposition. In (2.41), each root-eigenvalue multiplied together with its corresponding eigenvector represents the coordinates in one dimension of $\mathbb{R}^P$. The way the prior knowledge about the hearing-aid geometry can be exploited is by substituting the Gram matrices of the known coordinates in dimensions, which are orthogonal to the axis of freedom (defined here as Span $\{\mathbf{e}_1\}$). This is equivalent to truncating eigenvalues from the eigenvalue decomposition (2.40), however, in this instance, the eigenvectors are defined based on the prior knowledge and are not simply an outcome of computing an eigenvalue decomposition.

Starting with the relative Gram matrix $\mathbf{G}_{\text{rel,c}}$ in (2.38), obtained from $\mathbf{D}_{\text{EDM}}$ using $\mathbf{z}_{\text{c}}$ in (2.39), this Gram matrix is a function of the relative microphone geometry $\mathbf{M}_{\text{rel,c}}$ which is centered at the origin and is at most rank P. Assuming that the coordinates are known in ξ dimensions, i.e., they lie in Span $\{\mathbf{e}_{P-\xi+1}, \ldots, \mathbf{e}_P\}$ and can as such be selected from the known columns of the MAG $\mathbf{M}_{\text{rel,c}}$ (e.g., $\mathbf{M}_{\text{rel,c}}\mathbf{e}_p$ for the p-th column). The MAG estimation can be performed similarly to (2.40) and (2.41), except the rank ξ subtraction of the known coordinates from $\mathbf{G}_{\text{rel},c}$ is applied before the eigenvalue decomposition by subtracting ξ orthogonal, rank 1 Gram matrices from $\mathbf{G}_{\text{rel},c}$, i.e.,

$$\underbrace{\mathbf{G}_{\text{rel},c} - \sum_{p=P-\xi+1}^{P} (\mathbf{M}_{\text{rel,c}}\mathbf{e}_p)(\mathbf{M}_{\text{rel,c}}\mathbf{e}_p)^{\text{T}}}_{\text{rank}\leq P-\xi} = \mathbf{W}\mathbf{H}\mathbf{W}^{\text{T}} \,. \tag{3.13}$$

The remaining $(P-\xi)$ coordinates to be determined in Span $\{\mathbf{e}_1, \ldots, \mathbf{e}_{P-\xi}\}$ can be reconstructed as follows

$$\widehat{\mathbf{M}_{\text{rel,c,partial}}} = \mathbf{W}[\text{diag}(\sqrt{\eta_1}, \ldots, \sqrt{\eta_{P-\xi}}), \mathbf{0}_{(P-\xi,N-(P-\xi))}]^{\text{T}} \,, \tag{3.14}$$

with $\mathbf{H}$ the eigenvalue matrix containing eigenvalues $|\eta_1| \geq \cdots \geq |\eta_{P-\xi}|$ along the diagonal and the rest zeros, $\mathbf{W}$ is a matrix containing the corresponding eigenvectors, and . If there are estimation errors in $\mathbf{G}_{\text{rel},c}$ due to estimation errors in $\mathbf{D}_{\text{EDM}}$, then $\mathbf{H}$ should be truncated similar to $\mathbf{\Lambda}$ in (2.40), keeping only the largest $P-\xi$ eigenvalues. In order to reconstruct the whole geometry, the remaining coordinates in Span $\{\mathbf{e}_{P-\xi}, \ldots, \mathbf{e}_P\}$ which were subtracted in (3.13) must be re-included, i.e.

$$\hat{\mathbf{M}}_{\text{rel,c}} = [\widehat{\mathbf{M}_{\text{rel,c,partial}}}, \left[\mathbf{M}_{\text{rel,c}}\mathbf{e}_{P-\xi}, \ldots, \mathbf{M}_{\text{rel,c}}\mathbf{e}_P\right]] \,. \tag{3.15}$$

An example 3-dimensional scenario, incorporating prior knowledge of known PDs, is shown in Fig. 3.3, where after subtracting the coordinates in Span $\{\mathbf{e}_2, \mathbf{e}_3\}$ from $\mathbf{G}_{\text{rel,c}}$, like in (3.13), the only coordinates left to estimate in (3.14) are those in Span $\{\mathbf{e}_1\}$.

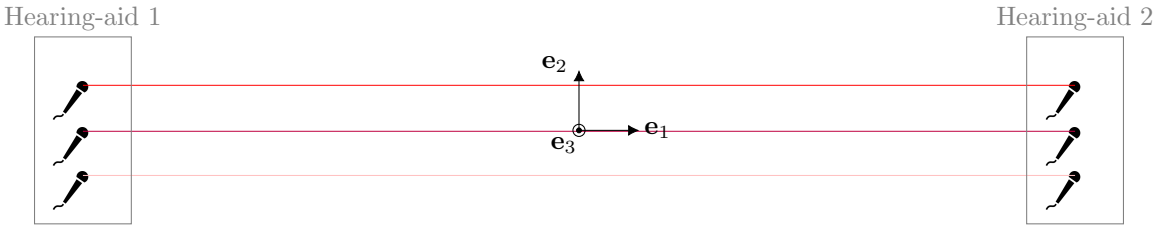


Fig. 3.3: Example hearing-aid geometry where each hearing-aid has 3 microphones. The coloured lines depict PDs which are estimated in (3.14). The coordinates in Span $\{\mathbf{e}_2, \mathbf{e}_3\}$ (with $\mathbf{e}_3$ coming *out of the page*) are known and $\mathbf{e}_1$ is the axis of freedom along which the position of the hearing-aids can change.

In order to use the prior knowledge from TDoAs, the matrix $\mathbf{M}_{\text{rel,c}}\mathbf{e}_p$ in (3.13) can be replaced with the distances shown in Fig. 3.1, i.e.

$$(\mathbf{M}_{\text{rel,c}}\mathbf{e}_P)(\mathbf{M}_{\text{rel,c}}\mathbf{e}_P)^{\text{T}} = \boldsymbol{\Delta}\boldsymbol{\Delta}^{\text{T}}, \tag{3.16}$$

which reduces the rank of the remaining Gram matrix by one. After applying MDS to estimate the partial geometry in (3.14), the coordinates are added in again to reconstruct the complete MAG matrix

$$\hat{\mathbf{M}}_{\text{rel,c}} = [\widehat{\mathbf{M}_{\text{rel,c,partial}}}, \boldsymbol{\Delta}] \, . \tag{3.17}$$

4 Experimental Simulations

In order to answer the question of whether the MAG estimation is generalizable, first, the reliability of the coherence estimation and PD estimation was analysed in terms of PD estimation error. In Chapter 4.1, the state-of-the-art ECM initialization (using an MVDR filter and the initial coherence estimate $\hat{\mathbf{\Gamma}}_{(0),\mathbf{I}}[k]$) was compared with proposed alternatives suggested in Chapter 3, i.e., using an MPDR filter and/or proposed blind and oracle coherence estimate initializations, for different frequency ranges. To see how well the PD can be estimated for different PDs, the PD estimation error was analysed for different frequency ranges and PDs in Chapter 4.2. Since, in practice, the TDoA is not available and must be blindly estimated, the influence of erroneous TDoA estimation is analysed in Chapter 4.3.

After finding the optimal parameter combination for PD estimation (e.g., ECM initialization parameters and frequency range), these parameters were then used for MAG estimation in simulated scenarios inspired by real life applications. The first scenario in Chapter 4.4.1 was a free-field scenario inspired by a real-life conference-room scenario with an ad-hoc constellation of $N = 6$ microphones (e.g., representing electronic devices such as laptops or phones). The second scenario in Chapter 4.4.2 involved estimating the distance between a pair of hearing-aids, as well as the effect of incorporating prior knowledge, to see whether it is possible to extend the MAG estimation framework to non-free-field scenarios.

4.1 ECM Initialization Parameter Influence

The acoustic scenarios were simulated using anechoic speech [32] originally sampled at 48 kHz, downsampled to 16 kHz, and convolved with simulated RIRs, 4096 samples long, using a free-field RIR generator [3] based on the image method [31]. The speed of sound in the simulations was set to $c = 340$ m/s. The centroid (average position) of the microphone array with $N = 2$ microphones was randomly defined within the room with dimensions $6 \times 6 \times 2.4$ m but not located too close to the walls nor the speech source. The reflection coefficients of each wall were equal and determined by the T_{60} which resulted in a DRR of 0 dB (± 1 dB) at the reference microphone $n = 1$, based on [33], i.e.,

$$\mathrm{DRR} = 10 \log_{10} \frac{\sum_{t=0}^{f_s(\tau_{\mathrm{ToA},1}+\tau_{\mathrm{cut}})} w_1[t]}{\sum_{f_s(\tau_{\mathrm{ToA},1}+\tau_{\mathrm{cut}}+1)}^{4096} w_1[t]}\,, \tag{4.1}$$

where the selection of the direct component was determined by the maximum peak of the impulse response $w_1[t]$ at sample t plus a few samples (i.e., $\tau_{\mathrm{cut}} = 8$ ms) to

capture the whole direct path.

Regarding the STFT framework, a frame length of $K = 512$ with a frame shift of $K/2$ was used and a square-root-Hann window was used as the analysis window.

The RDTF $\mathbf{g}[k]$ was computed with oracle anechoic RIRs using the method described in [34], i.e., as the principal eigenvector of the covariance matrix of the anechoic RTF normalized by the first entry (the reference microphone).

Regarding regularization of inverse matrices in 2.28, the loading factor was set to $F = 0.05$ and the lower-bound eps $= 10^{-6}$.

The influence of the initial coherence estimate and the filter used to estimate the initial speech and reverberation PSDs (discussed in Chapter 3.1) on the PD estimation error ϵ_{PD}, i.e.,

$$\epsilon_{\text{PD}} = |\hat{d}_{a,b} - d_{a,b}|, \tag{4.2}$$

was investigated (MAG estimation is trivial for $N = 2$ so it was omitted). The analysis was carried out for frequency bins corresponding to frequencies between $[0, 4]$ kHz, i.e., the lower frequency $f_{\text{lower}} \geq 0$ kHz, the upper frequency $f_{\text{upper}} \leq 4$ kHz, and $f_{\text{lower}} < f_{\text{upper}}$. The median error of 50 scenarios was reported, where one scenario was defined as a unique combination of an $N = 2$ microphone array with a pre-determined PD but random centroid and orientation and a randomly selected 5 s speech signal convolved with the RIR corresponding to its randomly generated source position.

In Chapter 3.1, some modifications were proposed to the ECM algorithm, namely to the filter in (3.2) which was used to estimate the initial speech and reverberation PSDs, as well as proposed alternative coherence initializations. In this Chapter, these proposed modifications are compared with state-of-the-art method, i.e., using $\mathbf{\Gamma}_{(0),\mathbf{I}}[k]$ and the MVDR filter to estimate the initial speech and reverberation PSDs, in terms of PD estimation error, for different frequency ranges.

The PD estimation error was analysed using different initial coherence estimates for a fixed PD $d_{a,b} = 20$ cm and using the MVDR filter in (2.18) or the MPDR filter in (3.2) (implemented as the MPDR-S filter in (3.3) and the MPDR-TV filter in (3.4)). The corresponding Figs. are listed in Tab. 4.1.

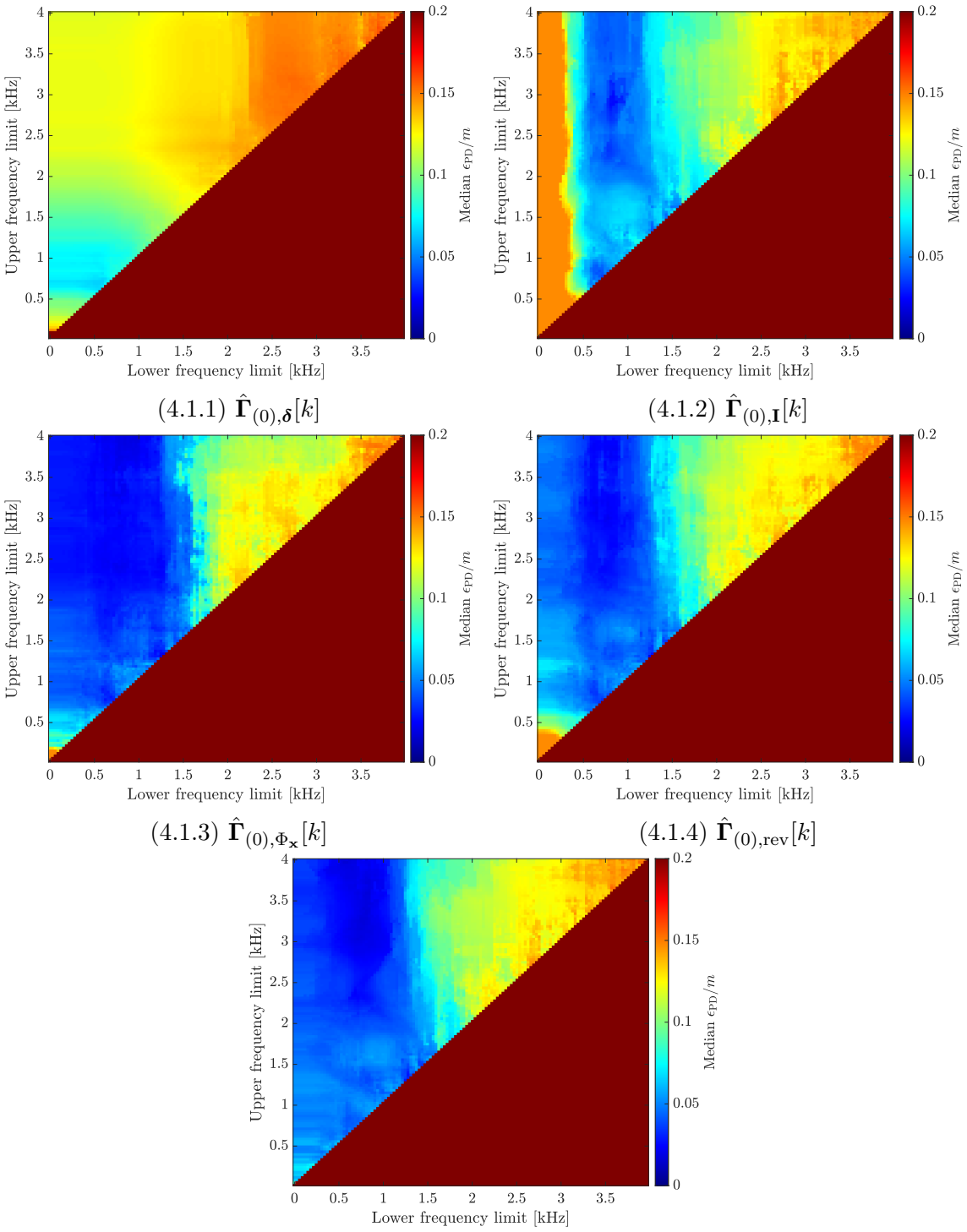
The results in Fig. 4.1 show that when using the MVDR filter, for $d_{a,b} = 20$ cm, both the choice of the initial coherence estimate and the frequency range are important to obtain a low estimation error. The proposed coherence initialization $\hat{\mathbf{\Gamma}}_{(0),\boldsymbol{\delta}}[k]$ which

Tab. 4.1: Figure guide for the analysis of PD estimation error using different initial coherence estimates and different frequency ranges.

MVDR filter	MPDR-S filter	MPDR-TV filter
Fig. 4.1	Fig. 4.2	Fig. 4.3

is based on the TDoAs is unreliable for all frequency ranges. The implementation using the initial coherence estimate $\hat{\mathbf{\Gamma}}_{(0),\mathbf{I}}[k]$ (together with the MVDR) represents the state-of-the-art implementation and gives an estimation error under 5 cm using a lower frequency $f_{\text{lower}} = 1$ kHz and upper frequency $f_{\text{upper}} = 3$ kHz. When setting the lower frequency to between 0.5 and 1.5 kHz, the estimation error is around 5 cm. The reason for the high estimation error when using lower frequencies smaller than 0.5 kHz is probably because of the poor initial estimation of the speech and reverberation PSDs in (2.15) and (2.16) using the MVDR filter. All other initializations (the proposed blind initialization $\hat{\mathbf{\Gamma}}_{(0),\Phi_{\mathbf{x}}}[k]$ and oracle initializations $\hat{\mathbf{\Gamma}}_{(0),\text{rev}}[k]$ and $\hat{\mathbf{\Gamma}}_{(0),\text{sinc}}[k]$) are far less sensitive to the selection of the lower frequency. They mostly give an estimation error of 5 cm or lower for lower frequencies $f_{\text{lower}} \leq 1.5$ kHz and appear insensitive to the upper frequency f_{upper}. Using the coherence initialization $\hat{\mathbf{\Gamma}}_{(0),\text{sinc}}[k]$ shows the lowest error when using a lower frequency bound and a frequency bound $f_{\text{upper}} < 0.5$ kHz.

When using the MVDR, at $d_{a,b} = 20$ cm, $\hat{\mathbf{\Gamma}}_{(0),\Phi_{\mathbf{x}}}[k]$ is the blind coherence estimate with the lowest PD estimation error initialization and has the benefit that it is the most robust for different frequency ranges.



(4.1.1) $\hat{\mathbf{\Gamma}}_{(0),\boldsymbol{\delta}}[k]$

(4.1.2) $\hat{\mathbf{\Gamma}}_{(0),\mathbf{I}}[k]$

(4.1.3) $\hat{\mathbf{\Gamma}}_{(0),\Phi_{\mathbf{x}}}[k]$

(4.1.4) $\hat{\mathbf{\Gamma}}_{(0),\mathrm{rev}}[k]$

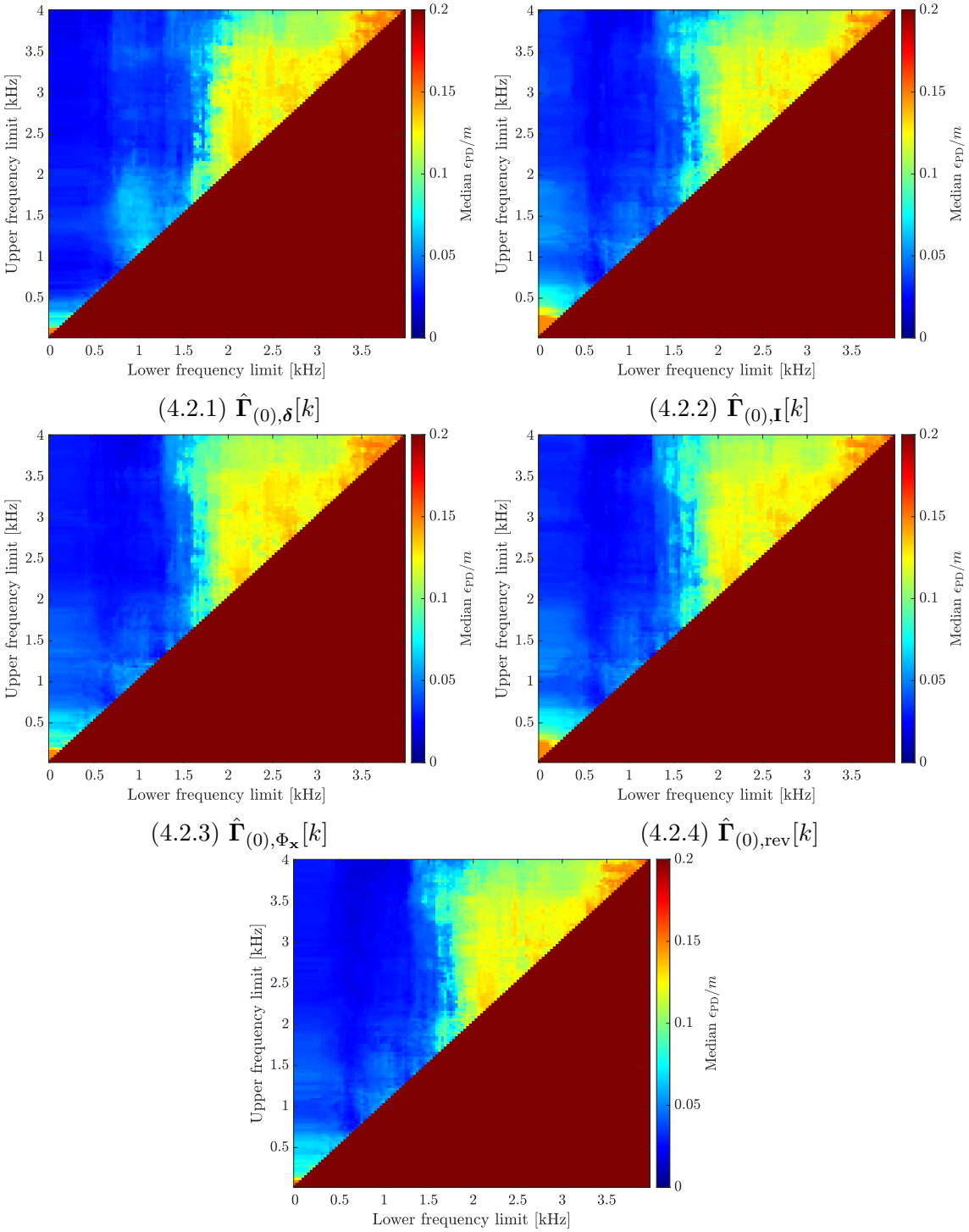
(4.1.5) $\hat{\mathbf{\Gamma}}_{(0),\mathrm{sinc}}[k]$

Fig. 4.1: PD estimation error for different frequency ranges and coherence initializations. Using an MVDR filter in the ECM initialization and for a fixed PD $d_{a,b} = 20$ cm

The results in Fig. 4.2 show that when using the MPDR-S filter to estimate the initial PSDs for $d_{a,b} = 20$ cm, the PD estimation is quite insensitive to the initial-

ization of the coherence estimate. For each initial coherence estimate, there is little variation between the PD estimation error at different frequency ranges. Using a lower frequency $f_{\text{lower}} \leq 1.5$ kHz and an upper frequency $f_{\text{upper}} > 2$ kHz mostly gives an estimation error of 5 cm or lower for each initial coherence estimate.

Comparing the results using the MVDR filter in Fig. 4.1 with those of the MPDR-S filter in Fig. 4.2, the MPDR-S filter appears to generally either produce the same or a smaller PD estimation error at different frequency ranges and using any initial coherence matrix. Since poorly initialized coherence matrices are not used in the MPDR-S filter, but instead a quantity (the covariance matrix of the recorded signal) based on the available recorded signal is used, it makes sense that smaller/fewer estimation errors are introduced in ECM and thus the result is a smaller PD error. Using the MPDR-S filter, the initial speech and reverberation PSDs in (2.15) and (2.15), respectively, rely less on the initial coherence estimate, so the improved performance suggests that estimating the PSDs is crucial for ECM.



(4.2.1) $\hat{\mathbf{\Gamma}}_{(0),\boldsymbol{\delta}}[k]$

(4.2.2) $\hat{\mathbf{\Gamma}}_{(0),\mathbf{I}}[k]$

(4.2.3) $\hat{\mathbf{\Gamma}}_{(0),\Phi_{\mathbf{x}}}[k]$

(4.2.4) $\hat{\mathbf{\Gamma}}_{(0),\mathrm{rev}}[k]$

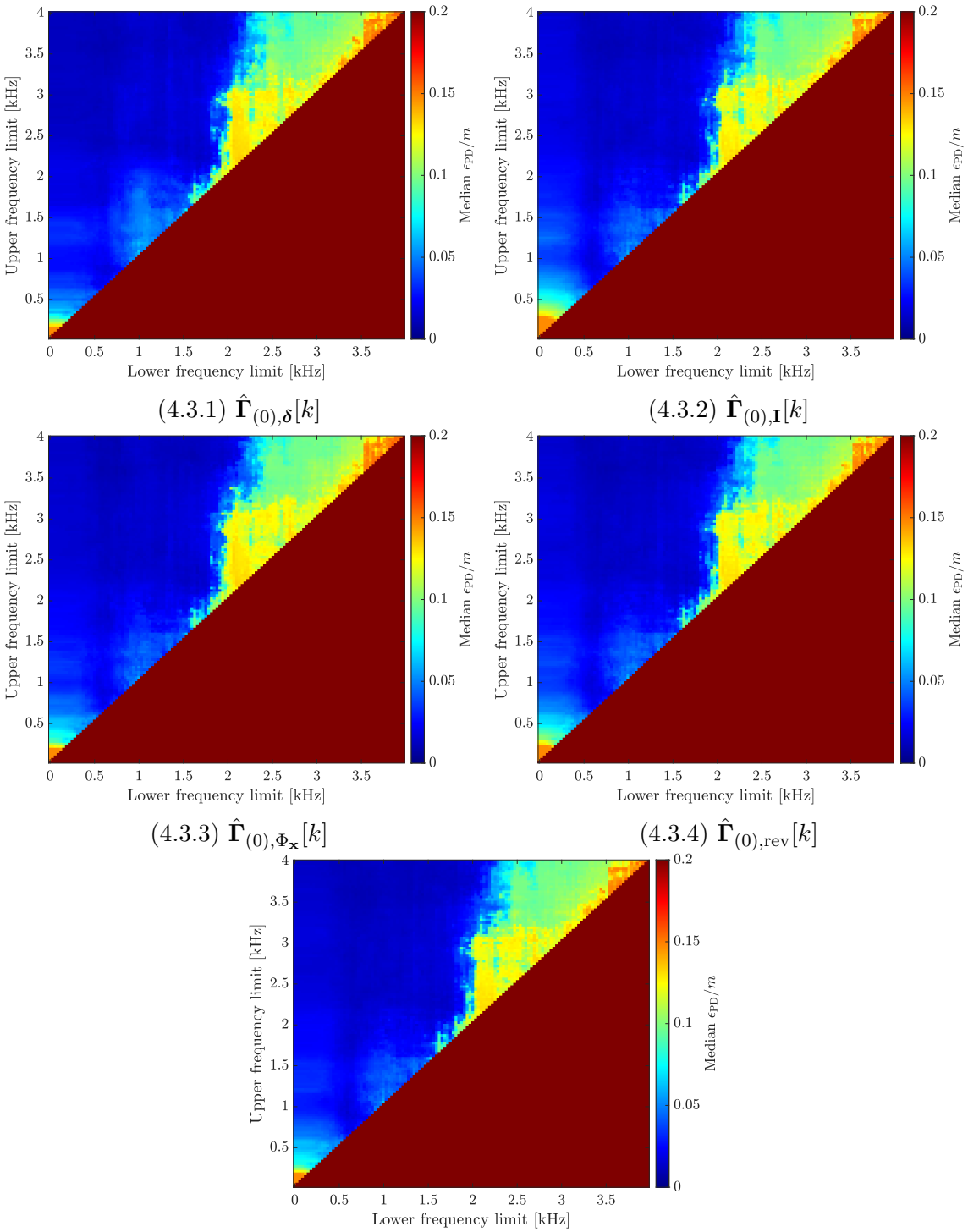
(4.2.5) $\hat{\mathbf{\Gamma}}_{(0),\mathrm{sinc}}[k]$

Fig. 4.2: PD estimation error for different frequency ranges and coherence initializations. Using the stationary MPDR filter in the ECM initialization and for a fixed PD $d_{a,b} = 20$ cm

Looking at Fig. 4.3, it is apparent that independently of the initial coherence estimate, the PD estimation is mostly lower than 3 cm for a lower frequency $f_{\mathrm{lower}} \leq 2$

kHz and an upper frequency $f_{\text{upper}} \geq 2$ kHz (if the frequency range is large enough, i.e., $f_{\text{upper}} - f_{\text{lower}} \geq 500$ Hz).

Comparing the results of the MPDR-TV filter in Fig. 4.3 with those of the MPDR-S filter in Fig. 4.2 and MVDR filter in Fig 4.1, it is evident that the MPDR-TV filter results in the lowest PD estimation error regardless of the frequency range and coherence estimate initialization. In order to reflect the highly time-varying nature of the recorded speech signal, the covariance matrix of the recorded speech should also be time-varying in the MPDR filter. This is because $\mathbf{\Phi}_{\mathbf{x}}[k, l]$ is time-varying as it is a *mixture of matrices* $\mathbf{g}[k]\mathbf{g}^{\text{H}}[k]$ and $\mathbf{\Gamma}[k]$, where each matrix is weighted by different time-varying PSDs in each time-frame l.



(4.3.1) $\hat{\mathbf{\Gamma}}_{(0),\boldsymbol{\delta}}[k]$

(4.3.2) $\hat{\mathbf{\Gamma}}_{(0),\mathbf{I}}[k]$

(4.3.3) $\hat{\mathbf{\Gamma}}_{(0),\Phi_{\mathbf{x}}}[k]$

(4.3.4) $\hat{\mathbf{\Gamma}}_{(0),\mathrm{rev}}[k]$

(4.3.5) $\hat{\mathbf{\Gamma}}_{(0),\mathrm{sinc}}[k]$

Fig. 4.3: PD estimation error for different frequency ranges and coherence initializations. Using the time-varying MPDR filter in the ECM initialization and for a fixed PD $d_{a,b} = 20$ cm

Tab. 4.2: Figure guide for the analysis of PD estimation error at different PDs and different frequency ranges.

	MVDR filter	MPDR-S filter	MPDR-TV filter
$\hat{\mathbf{\Gamma}}_{(0),\mathbf{I}}[k]$	Fig. 4.5	Fig. 4.7	Fig. 4.9
$\hat{\mathbf{\Gamma}}_{(0),\Phi_{\mathbf{x}}}[k]$	Fig. 4.6	Fig. 4.8	Fig. 4.10

4.2 Influence of Pairwise Distance

In order to see how accurately the coherence estimate from ECM can estimate the PDs for different different sized arrays, in this Chapter the PD estimation accuracy is analysed at different PDs $d_{1,2} = \{5, 10, 20, 30, 40, 50\}$ cm for an array with $N = 2$ microphones. The PD estimation error is analysed for different frequency ranges at each PD to see how dependent the frequency range is to changes in the PD. The two best blind initial coherence estimates with the lowest PD estimation error in Chapter 4.1 together with either the MVDR, MPDR-S, or MPDR-TV filter were considered. Tab. 4.2 shows an overview of compared parameters.

For the purpose of comparing the PD estimation accuracy at different frequencies with the Figs. in Tab. 4.2 with the spherical-isotropic coherence model from (2.7), the coherence model is plotted in Fig. 4.4 for the tested PDs.

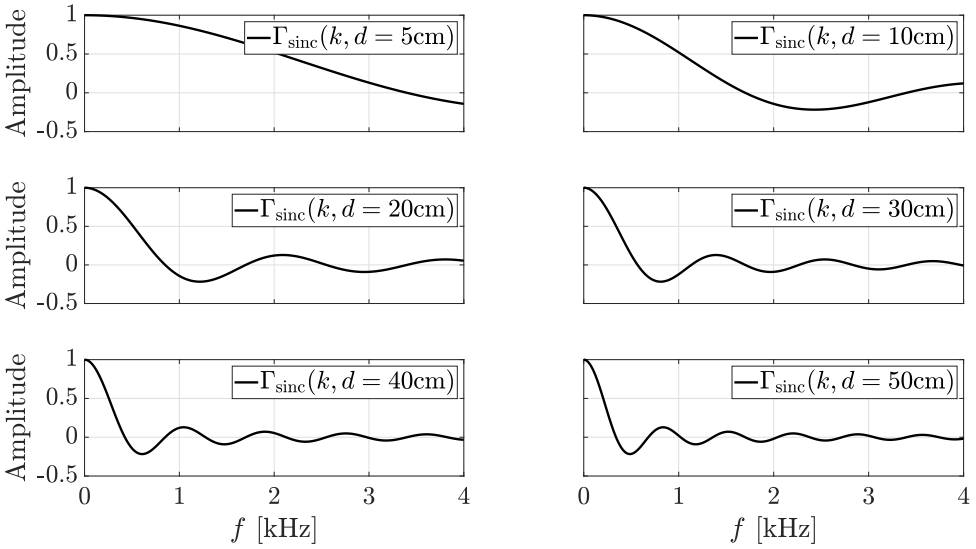
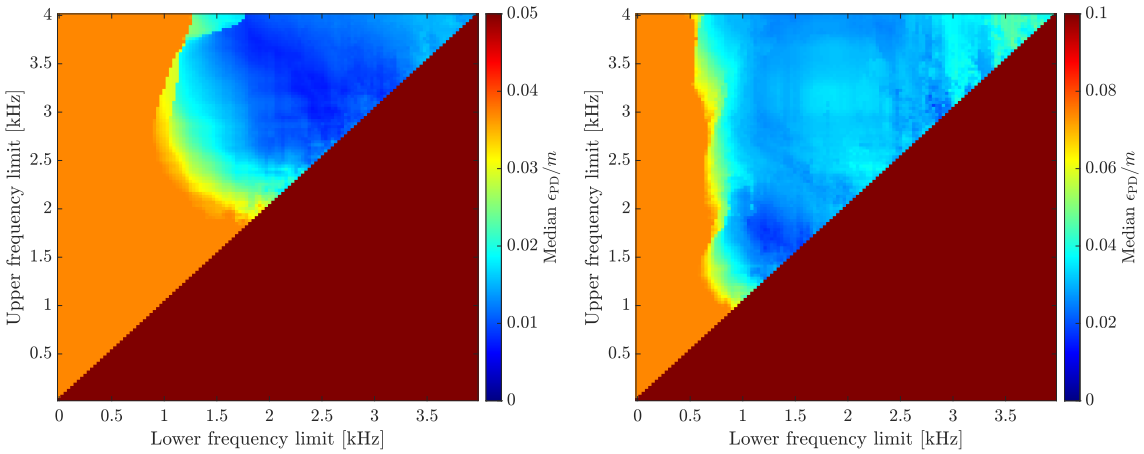


Fig. 4.4: Model coherence over frequency for different PDs.

The results in Fig. 4.5 show the PD estimation error at different PDs using $\hat{\mathbf{\Gamma}}_{(0),\mathbf{I}}[k]$ as the coherence initialization together with the MVDR filter. Overall, using this combination for PD estimation requires very careful tuning of the frequency range to even be able to estimate the PD within an error of 25% of the PD. While the upper frequency f_{upper} has little influence on the PD estimation, the lower frequency must

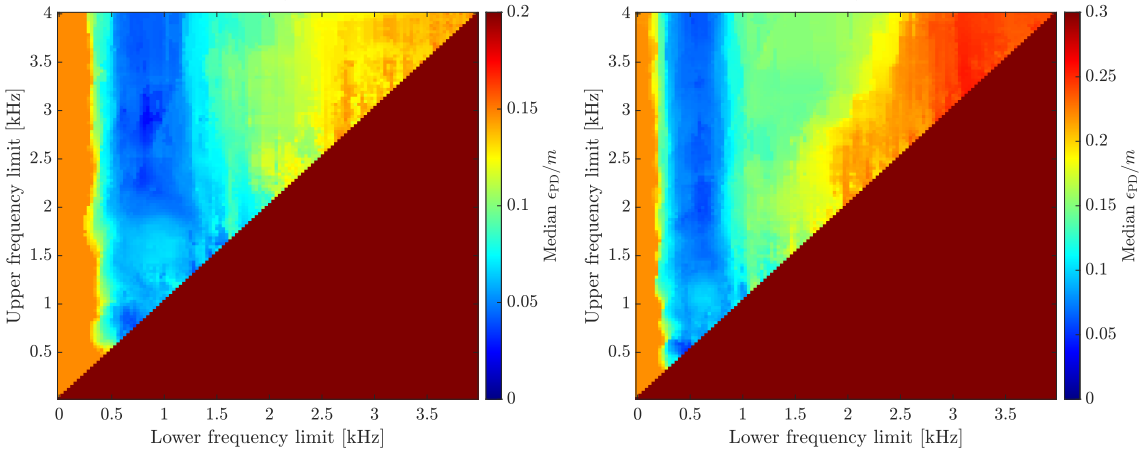
be within a certain band of frequencies in order to reliably estimate the PD. Comparing with Figs. 4.5 and 4.4, it seems as though setting the lower frequency f_{lower} to the frequency corresponding to the point at which the sinc coherence $\Gamma_{\text{sinc}}(k, d)$ has an amplitude of larger than 0.5 leads to erroneous PD estimation. Since the band of frequencies appears to be dependent on the PD itself, which makes this combination of parameters non-generalizable for MAG estimation, because in order to estimate the PD, the PD must be known in order to tune the frequency range!

The overall trend of the PD estimation errors indicates that as the PD increases, the lower frequency bound f_{lower} should be reduced.



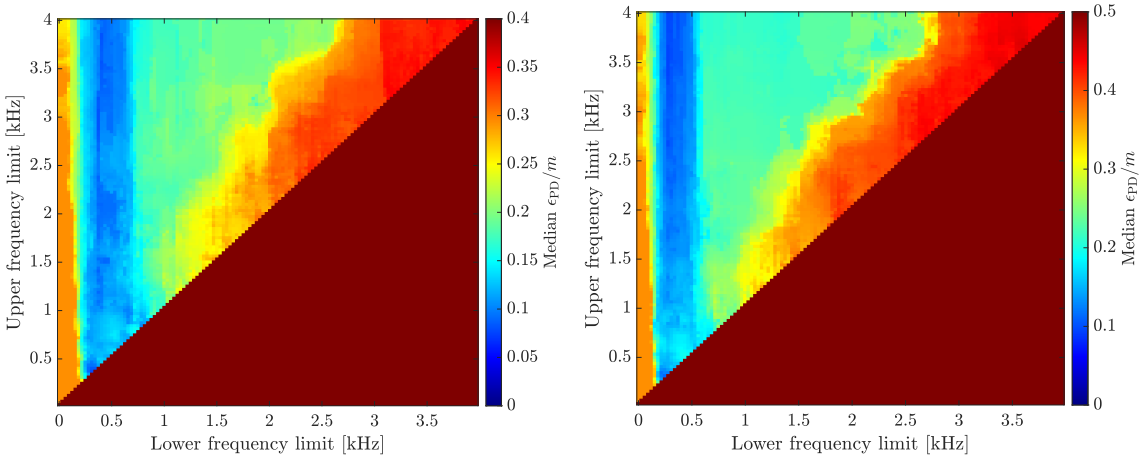
(4.5.1) $d_{a,b} = 5$ cm

(4.5.2) $d_{a,b} = 10$ cm



(4.5.3) $d_{a,b} = 20$ cm

(4.5.4) $d_{a,b} = 30$ cm



(4.5.5) $d_{a,b} = 40$ cm

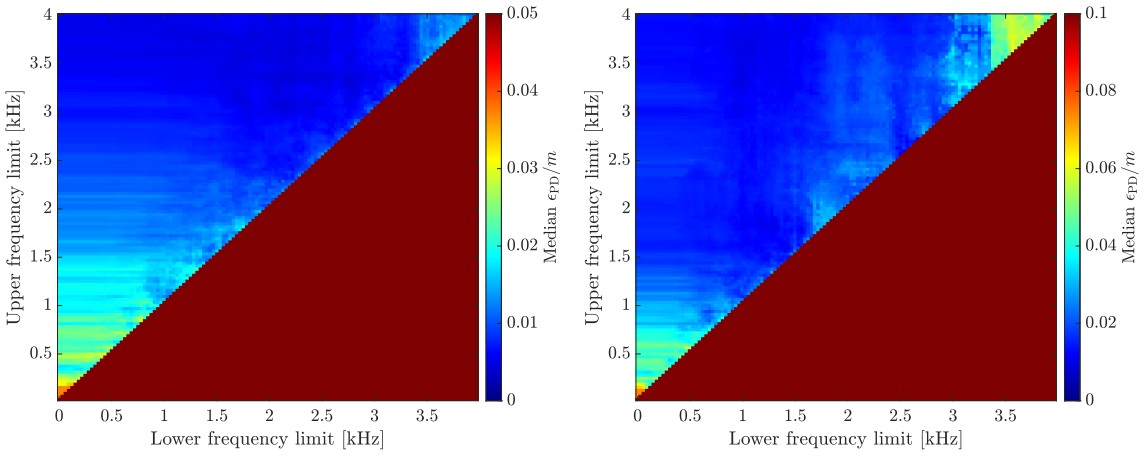
(4.5.6) $d_{a,b} = 50$ cm

Fig. 4.5: PD estimation error for different frequency ranges and different PDs. Using an MVDR filter in the ECM initialization and initial coherence estimate $\hat{\mathbf{\Gamma}}_{(0),\mathbf{I}}[k]$.

The results in Fig. 4.6 show the PD estimation error at different PDs using $\hat{\mathbf{\Gamma}}_{(0),\mathbf{\Phi_x}}[k]$ as the coherence initialization together with the MVDR filter. Compared to using

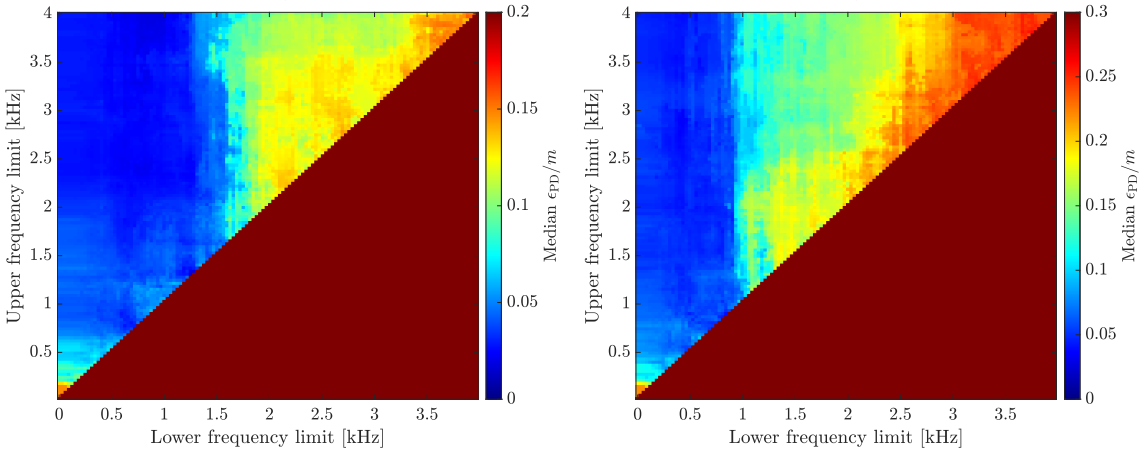
the state-of-the-art initial coherence estimate $\hat{\mathbf{\Gamma}}_{(0),\mathbf{I}}[k]$, using $\hat{\mathbf{\Gamma}}_{(0),\Phi_{\mathbf{x}}}[k]$ gives a lower PD estimation error for a larger range of lower frequencies f_{lower} at each tested PD (independently of the upper frequency bound f_{upper}). Using $\hat{\mathbf{\Gamma}}_{(0),\Phi_{\mathbf{x}}}[k]$, the lower bound frequency bound becomes less sensitive to the array size than with $\hat{\mathbf{\Gamma}}_{(0),\mathbf{I}}[k]$, which means that the PD can be estimated with a lower error using a fixed low frequency bound (e.g., 300 Hz) at all tested distances. Using the MVDR filter together with $\hat{\mathbf{\Gamma}}_{(0),\Phi_{\mathbf{x}}}[k]$ and using an adequate frequency range, the median PD estimation error ϵ_{PD} can be estimated in the region of 20% or lower of the PD itself for all tested distances.

When using the MVDR filter, it is important to initialize the estimated coherence matrix as well as possible to reduce the PD estimation error, for this, $\hat{\mathbf{\Gamma}}_{(0),\Phi_{\mathbf{x}}}[k]$ seems more suitable than $\hat{\mathbf{\Gamma}}_{(0),\mathbf{I}}[k]$.



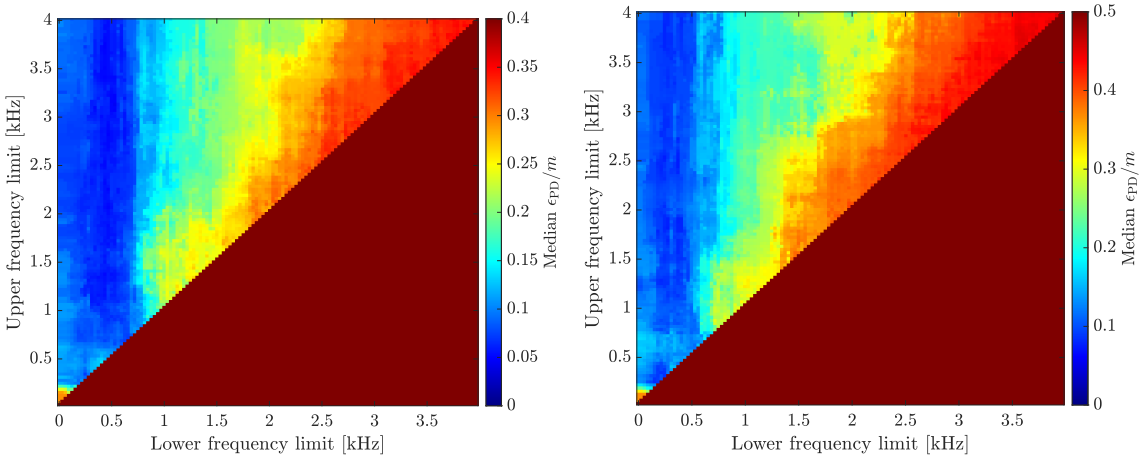
(4.6.1) $d_{a,b} = 5$ cm

(4.6.2) $d_{a,b} = 10$ cm



(4.6.3) $d_{a,b} = 20$ cm

(4.6.4) $d_{a,b} = 30$ cm



(4.6.5) $d_{a,b} = 40$ cm

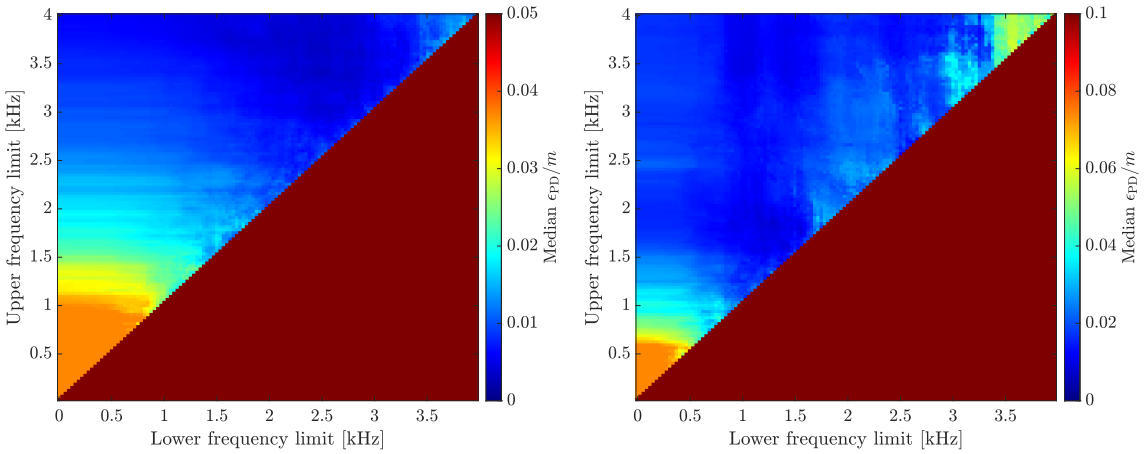
(4.6.6) $d_{a,b} = 50$ cm

Fig. 4.6: PD estimation error for different frequency ranges and different PDs. Using an MVDR filter in the ECM initialization and initial coherence estimate $\hat{\mathbf{\Gamma}}_{(0),\mathbf{\Phi_x}}[k]$.

The results in Fig. 4.7 show the PD estimation error at different PDs using $\hat{\mathbf{\Gamma}}_{(0),\mathbf{I}}[k]$ as the coherence initialization together with the proposed MPDR-S filter. The re-

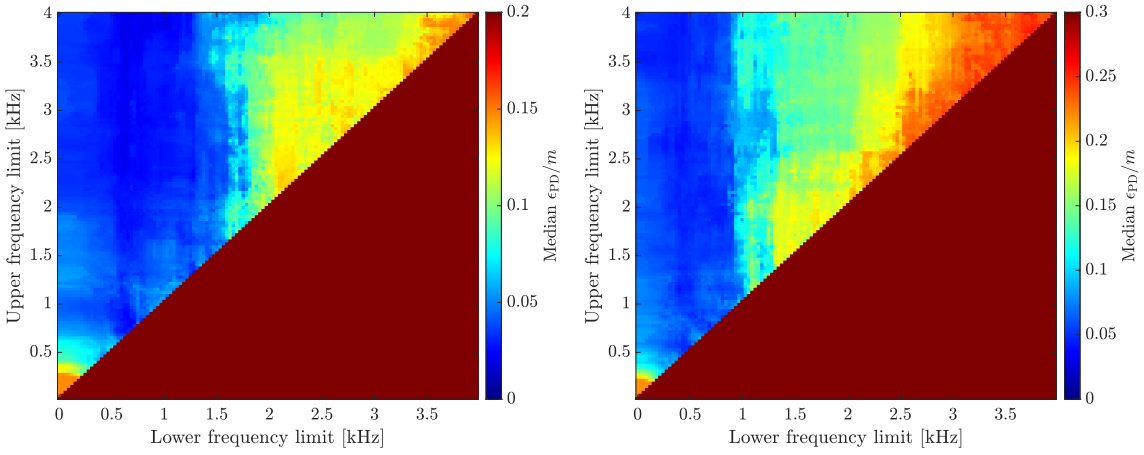
sults are similar to Fig. 4.6, with only minor differences. At large distances $d_{a,b} \geq 40$ cm, the lower frequency bound f_{lower} should be set above 250 Hz to estimate the PD error more accurately, whereas when using the MVDR together with $\hat{\mathbf{\Gamma}}_{(0),\Phi_{\mathbf{x}}}[k]$, there are smaller variations in PD estimation error at these low frequencies.

The dissimilarity of the results to Fig. 4.5 suggests that if the initial coherence estimate is poor, using the proposed MPDR-S filter instead of an MVDR filter can improve the PD estimation accuracy.



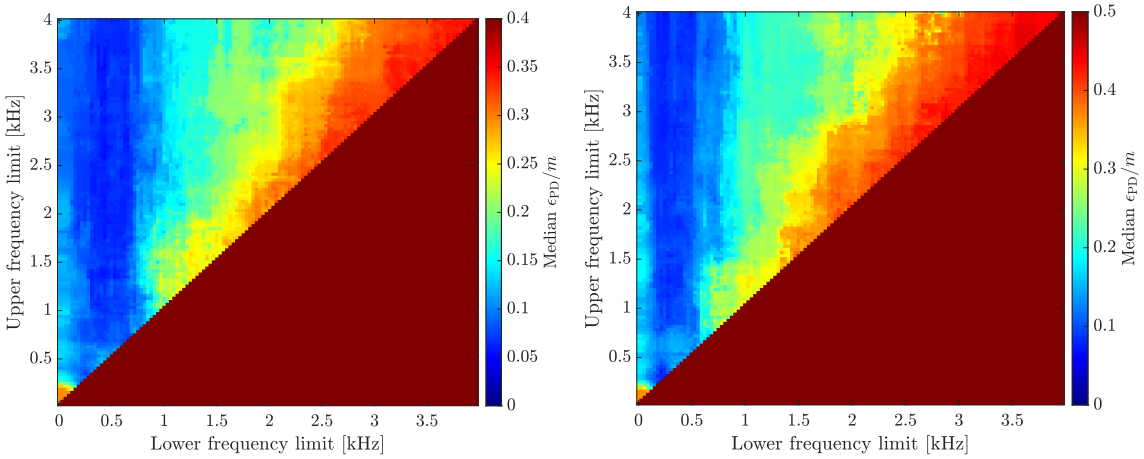
(4.7.1) $d_{a,b} = 5$ cm

(4.7.2) $d_{a,b} = 10$ cm



(4.7.3) $d_{a,b} = 20$ cm

(4.7.4) $d_{a,b} = 30$ cm



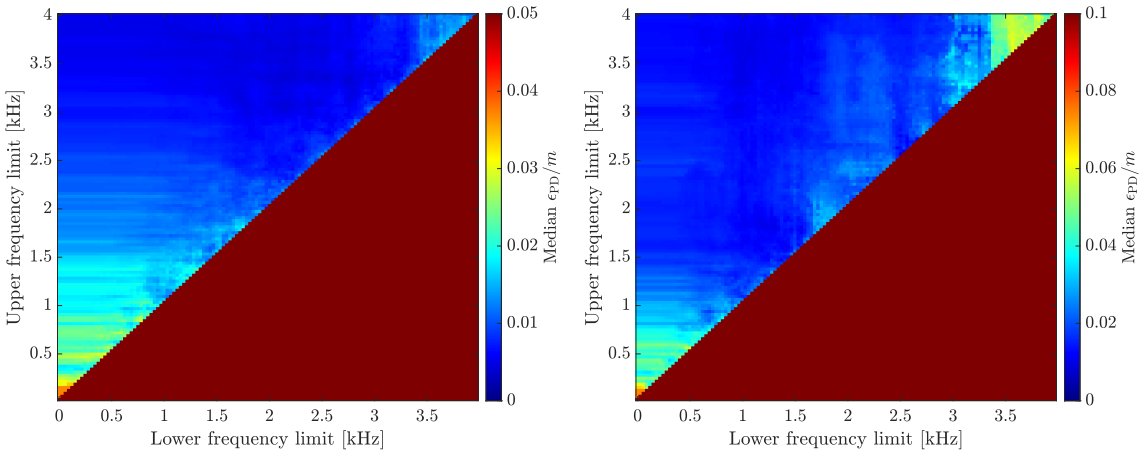
(4.7.5) $d_{a,b} = 40$ cm

(4.7.6) $d_{a,b} = 50$ cm

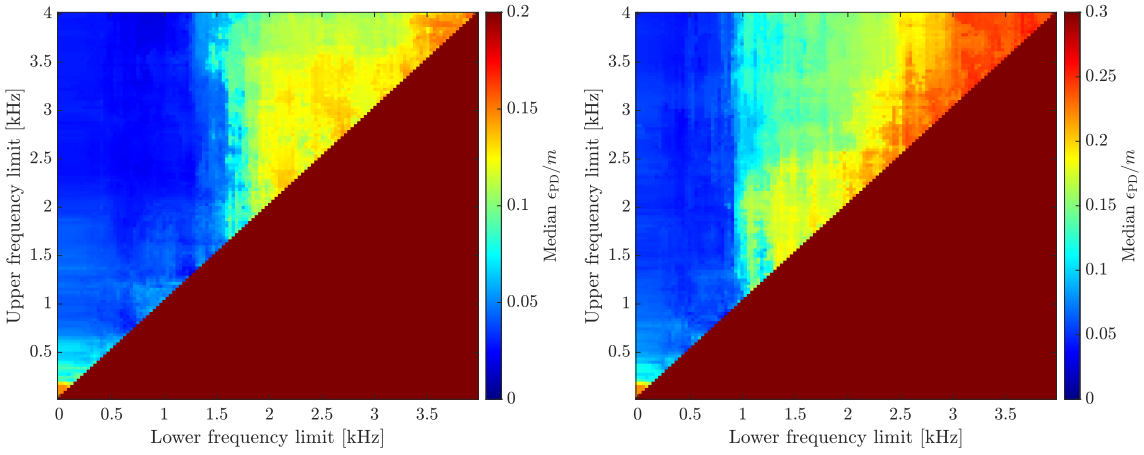
Fig. 4.7: PD estimation error for different frequency ranges and different PDs. Using an MPDR-S filter in the ECM initialization and initial coherence estimate $\hat{\mathbf{\Gamma}}_{(0),\mathbf{I}}[k]$.

The results in Fig. 4.8 show the PD estimation error at different PDs using $\hat{\mathbf{\Gamma}}_{(0),\mathbf{\Phi_x}}[k]$ as the coherence initialization together with the proposed MPDR-S filter. Overall,

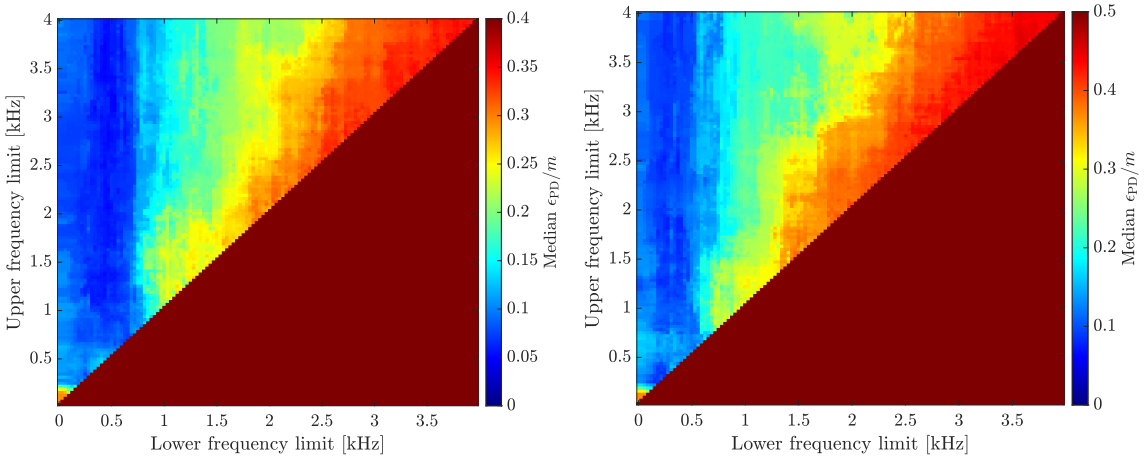
the results are similar to Figs. 4.6 and 4.7, i.e., the regions where the PD can be estimated to within 20% of the PD itself are very similar. This suggests that if the coherence estimate is good enough, varying between an MVDR filter and an MPDR-S filter does not change the PD estimation accuracy.



(4.8.1) $d_{a,b} = 5$ cm (4.8.2) $d_{a,b} = 10$ cm



(4.8.3) $d_{a,b} = 20$ cm (4.8.4) $d_{a,b} = 30$ cm



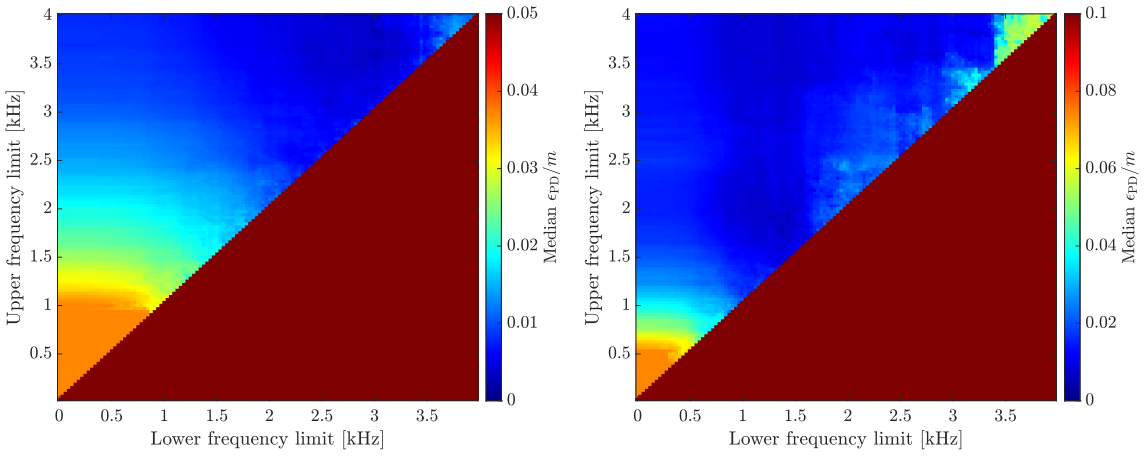
(4.8.5) $d_{a,b} = 40$ cm (4.8.6) $d_{a,b} = 50$ cm

Fig. 4.8: PD estimation error for different frequency ranges and different PDs. Using an MPDR-S filter in the ECM initialization and initial coherence estimate $\hat{\mathbf{\Gamma}}_{(0),\mathbf{\Phi_x}}[k]$.

The results in Fig. 4.9 show the PD estimation error at different PDs using $\hat{\mathbf{\Gamma}}_{(0),\mathbf{I}}[k]$ as the coherence initialization together with the proposed MPDR-TV filter. This

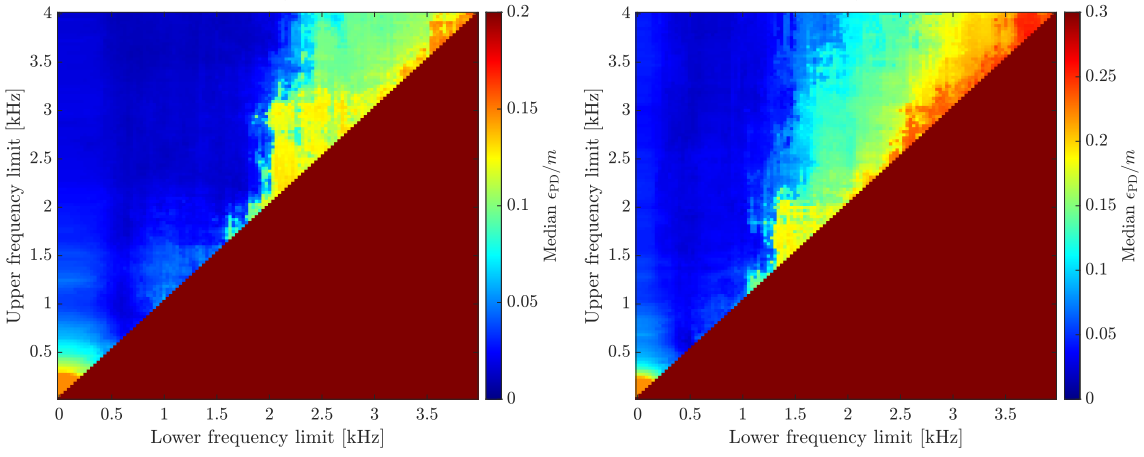
combination is robust at the same frequencies as when using the MPDR-S filter together with $\hat{\mathbf{\Gamma}}_{(0),\mathbf{I}}[k]$ and the PD estimation accuracy is similar.

If the lower frequency bound is set to $f_{\text{lower}} = 0.3$ kHz, and the upper frequency bound $f_{\text{upper}} \geq 3$ kHz then the median PD estimation error is in the region of 10% of the PD compared to a median PD estimation error around 20% of the PD when using the MPDR-S filter. Thus, incorporating the time-varying recorded speech covariance matrix into the MPDR can provide a 10% reduction in PD estimation error.



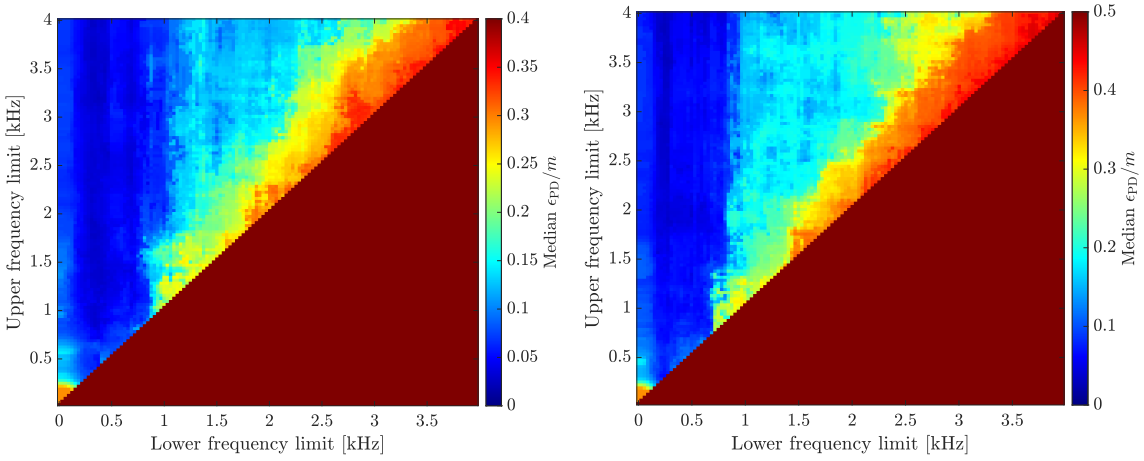
(4.9.1) $d_{a,b} = 5$ cm

(4.9.2) $d_{a,b} = 10$ cm



(4.9.3) $d_{a,b} = 20$ cm

(4.9.4) $d_{a,b} = 30$ cm



(4.9.5) $d_{a,b} = 40$ cm

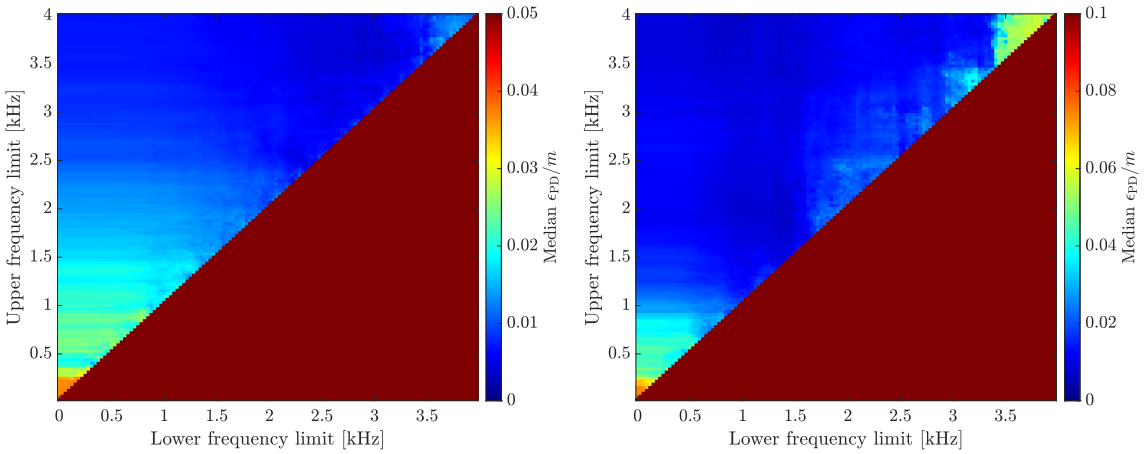
(4.9.6) $d_{a,b} = 50$ cm

Fig. 4.9: PD estimation error for different frequency ranges and different PDs. Using the time-varying MPDR-TV filter in the ECM initialization and initial coherence estimate $\hat{\mathbf{\Gamma}}_{(0),\mathbf{I}}[k]$.

The results in Fig. 4.10 show the PD estimation error at different PDs using $\hat{\mathbf{\Gamma}}_{(0),\Phi_{\mathbf{x}}}[k]$ as the coherence initialization together with the proposed MPDR-TV fil-

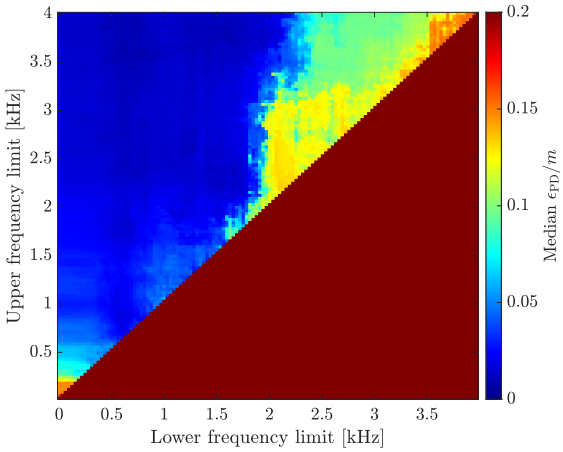
ter. If the lower frequency bound is set to $f_{\text{lower}} = 0.3$ kHz and the upper frequency bound $f_{\text{upper}} \geq 3$ kHz then the median PD estimation error is in the region of 10% of the PD compared to when using the MPDR-S filter an estimation of around 20% was obtained. Thus, using a time-varying estimate of the covariance matrix of the recorded speech in the MPDR filter instead of a stationary matrix can provide a 10% reduction in PD estimation error.

Similar performance improvements were obtained by using the MPDR-TV filter instead of the MPDR-S filter with the initial coherence estimate $\hat{\mathbf{\Gamma}}_{(0),\Phi_{\mathbf{x}}}[k]$, i.e., comparing the results in Fig. 4.10 and Fig. 4.8.

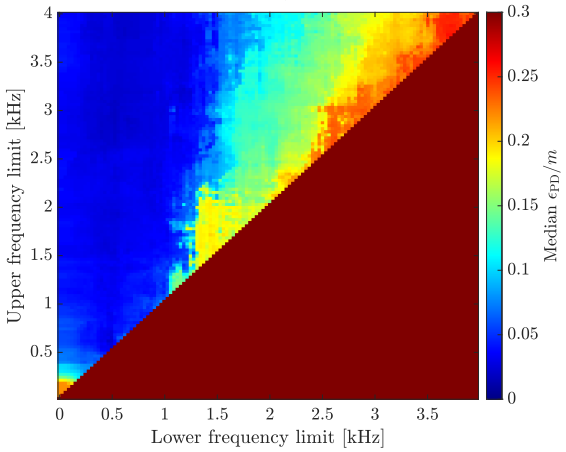


(4.10.1) $d_{a,b} = 5$ cm

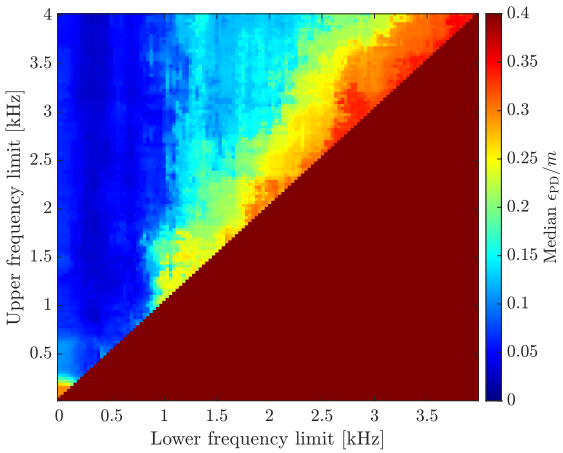
(4.10.2) $d_{a,b} = 10$ cm



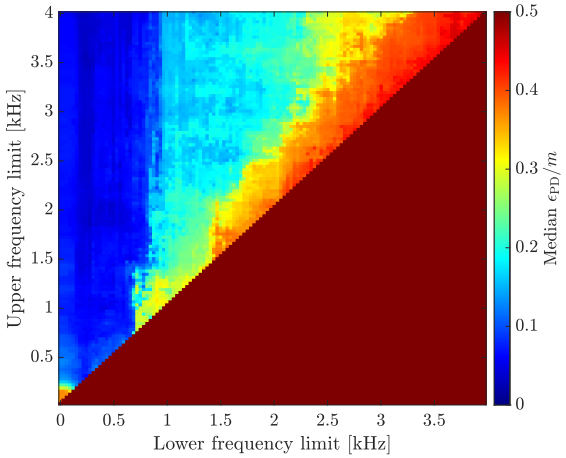
(4.10.3) $d_{a,b} = 20$ cm



(4.10.4) $d_{a,b} = 30$ cm



(4.10.5) $d_{a,b} = 40$ cm



(4.10.6) $d_{a,b} = 50$ cm

Fig. 4.10: PD estimation error for different frequency ranges and different PDs. Using the time-varying MPDR-TV filter in the ECM initialization and initial coherence estimate $\hat{\mathbf{\Gamma}}_{(0),\Phi_{\mathbf{x}}}[k]$.

Comparing the model coherence with the PD estimation errors (excluding the specific combination of initial coherence estimate $\hat{\mathbf{\Gamma}}_{(0),\mathbf{I}}[k]$ with the MVDR filter in Fig.

4.5), it can be seen that it is important to use frequencies bins corresponding to the main lobe of the sinc. Usually, the lowest PD estimation error is achieved when the lower frequency bound f_{lower} is set such that it is about halfway between the top of the main lobe of the sinc and the first zero-crossing which suggests that while including the main lobe of the sinc is important, there is a point at which including lower frequencies is detrimental to the PD estimation (e.g., this is clearly seen in Fig. 4.10 for lower frequencies below 150 Hz). For the tested PDs, a fixed lower frequency bound somewhere in the region $0.1\text{kHz} \geq f_{\text{lower}} \geq 0.4\text{kHz}$ results in a low PD estimation error corresponding to 10% of the PD.

The precise setting of the upper frequency bound f_{upper} is not so important. Using a frequency above 3 kHz works well for all tested PDs.

Overall, the best combination of blind estimators for all PDs is the proposed MPDR-TV filter together with the proposed coherence initialization $\hat{\mathbf{\Gamma}}_{(0),\mathbf{\Phi_x}}[k]$.

4.3 Influence of Erroneous RDTF Estimation

In Chapters 4.1 and 4.2 it was seen that the PD estimation is sensitive to the initialization of ECM, however, it was assumed that the true RDTF was known. In this Chapter, the influence of estimation errors in the true RDTF $\mathbf{g}[k]$ was analysed using the MPDR-TV filter in (3.4) in combination with the two best-performing blind, initial coherence estimates, $\mathbf{\Gamma}_{(0),\mathbf{I}}[k]$ and $\mathbf{\Gamma}_{(0),\mathbf{\Phi_x}}[k]$.

The mis-estimation of the RDTF vector $\mathbf{g}[k]$ was analysed in terms of the MAG estimation error, i.e., the average 2-norm $||.||_2$ of difference between the aligned (using (2.47)) and centered, estimated, coordinates and the centered, true coordinates of each microphone of the array, is defined as

$$\epsilon_{\text{MAG}} = \frac{1}{N}\sum_{n=1}^{N} ||\hat{\mathbf{M}}_{\text{c,rel}}\mathbf{Q}\mathbf{e}_n - \mathbf{M}_{\text{c}}\mathbf{e}_n||_2 \ . \tag{4.3}$$

for a randomly generated microphone array with $N = 5$ microphones. Other than the number of microphones, the simulation framework was the same as in Chapter 4.1. The microphone positions were randomly generated within a virtual cube of $0.3 \times 0.3 \times 0.3$ m, which means that the maximal possible PD was 0.52 m. Based on the results from Chapters 4.1 and 4.2, a fixed frequency range $f \in [0.2, 4]$ kHz was used.

To simulate estimation errors in the TDoA in (2.2), which could occur in practice

when using GCC-PHAT [8], Gaussian-distributed errors are applied to the TDoAs between the reference microphone and other microphones, with a zero-mean and standard-deviation corresponding to sample t, i.e., the TDoAs are erroneously estimated with a standard-deviation of t samples. The TDoAs were varied between 0.01 samples and 30 samples. Each plotted MAG estimation error is the median of 20 scenarios, with $N = 5$ and 10 s recorded speech per scenario, i.e., the median of 100 coordinate estimation errors.

The MAG estimation error is shown over different TDoA estimation erros in Fig. 4.11. Both coherence initializations show a very similar trend over all estimation errors. When the TDoAs are estimated with a low error (less than 0.01 samples), the median MAG estimation error for both initializations stays low, i.e. between $[1, 1.5]$ cm. Increasing the TDoA error to a standard-deviation $\sigma_\tau = 1$ sample slightly raises the estimation error to 2.5 cm. Interestingly, the MAG estimation error stays at a similar level for higher TDoA errors, all the way upto the highest tested sample error of $\sigma_\tau = 30$ samples.

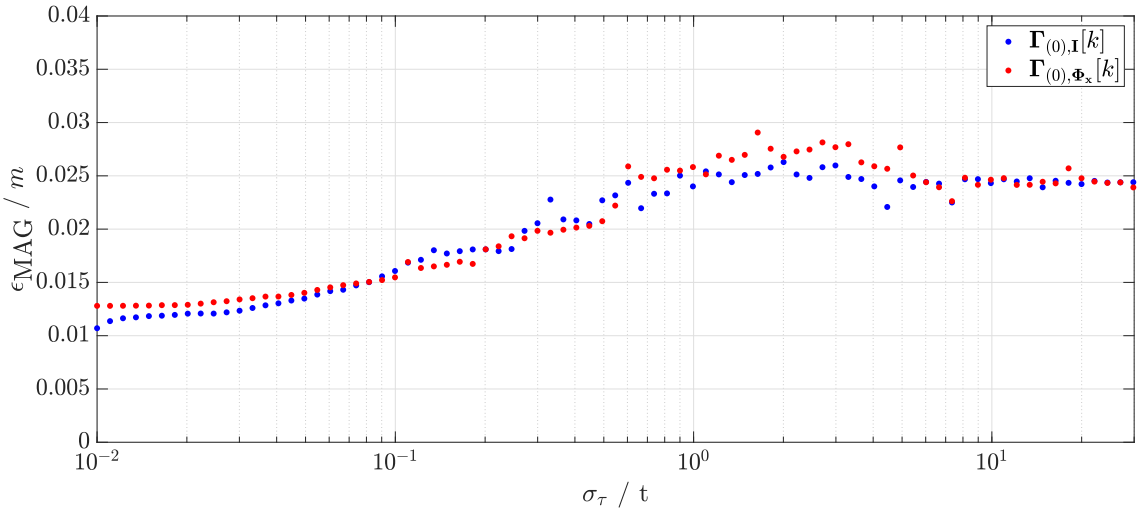


Fig. 4.11: MAG estimation error over TDoA estimation error in samples t.

Summarizing the results, using the MPDR-TV filter and either of the initial coherence estimates $\mathbf{\Gamma}_{(0),\mathbf{I}}[k]$ or $\mathbf{\Gamma}_{(0),\mathbf{\Phi}_\mathbf{x}}[k]$, the MAG estimation error is lowest for small TDoA estimation errors, however, ECM is quite robust to estimation errors in the RDTF (due to erroneously estimated TDoAs). This means that even in reverberant environments where the TDoA estimation using GCC-PHAT may suffer, the MAG can still be estimated using this framework.

4.4 Simulated Practical Applications

In this Chapter, the MAG estimation capabilities are demonstrated for two types of simulated scenarios. In the first type of scenario in Chapter 4.4.1, an array of randomly positioned microphones was simulated. The microphones were positioned within a virtual cube with pre-defined dimensions and the size of this virtual cube was varied. In the second type of scenario in Chapter 4.4.2, the aim was to estimate the inter-aural PD of a pair of hearing-aids with a head in-between as well as the MAG of all of the available microphones in the hearing-aids.

4.4.1 Geometry Estimation of a Distributed 3D Microphone Array

To evaluate the generalizability in the context of a 3-dimensional microphone array with $N = 6$ microphones, both the PD estimation error between all microphone pairs in (4.2) as well as the MAG estimation error in (4.3) were analysed. In addition, to see how MDS influences the estimated PDs, the PDs in the PD matrix $\mathbf{P}_{\text{MDS}}$ were analysed, similarly to (4.2), however, the estiamted PDs $\hat{d}_{a,b}$ were replaced with the square-root of the entries of the reconstructed EDM in (2.42). This variation of the PD error is named $\epsilon_{\text{PD,MDS}}$

In Chapter 4.2 it was seen that the PD of the microphones plays a role in how well the PD can be estimated so here it is investigated how well the MAG can be estimated using the estimated PDs. The simulated scenario remained the same as in Chapter 4.1, except for the microphone geometry, which in this Chapter has $N = 6$ microphones, whose positions were randomly generated within a virtual cube of pre-defined dimensions. As the combination of parameters which led to the most robust performance of ECM and the lowest PD estimation error in Chapter 4.1, the initial coherence matrix estimate $\hat{\mathbf{\Gamma}}_{(0),\Phi_{\mathbf{x}}}[k]$ was used together with the MPDR-TV filter and frequency bins within the frequency range $f \in [0.2, 4]$ kHz were used. The RDTF $\mathbf{g}[k]$ was computed with the method described in [34] using oracle anechoic RIRs, i.e., as the principal eigenvector of the covariance matrix of the anechoic RTF (discrete Fourier transform (DFT) of the RIR using the same framework as the STFT framework), normalized by the first entry (the reference microphone).

The cubes were varied by the length of their sides, i.e., the cube-length (CL). In each scenario, 16 s recorded speech was available. 50 scenarios were analysed per CL and the CL was varied in the range CL $\in 0.1, 0.2, \ldots, 0.5$ m. For the case CL $= 0.5$, this equates to a maximal possible PD of 0.866 m. $N = 6$ MAG estimation errors and $\sum_{n=1}^{N-1} n = 15$ PD errors were evaluated. The resulting PD or MAG estimation errors were plotted in boxplots together with violin plots [35, 36] to show the distribution of the errors. The distribution was analysed to see if stays similar for each

Tab. 4.3: Figure guide for the analysis of PD and MAG estimation errors when estimating the geometry of a 3-dimensional MAG with $N = 6$ microphones.

ϵ_{PD}	ϵ_{PD}/CL	$\epsilon_{\mathrm{PD,MDS}}$	$\epsilon_{\mathrm{PD,MDS}}$/CL	ϵ_{MAG}	ϵ_{MAG}/CL
Fig. 4.12	Fig. 4.13	Fig. 4.14	Fig. 4.15	Fig. 4.16	Fig. 4.17

tested CL (or if, e.g., it changes for large arrays). It is possible for most PDs to be estimated accurately, with the exception of one or two, which, if estimated with a high error, could have a very detrimental effect on the overall MAG estimation of all microphones (not just those whose PD is erroneously estimated). This is why it is useful to compare both measures and their distributions, to see whether the how many PDs are estimated with a low error and the effect on the estimated MAG. The median is marked by the blue line, the 25% and 75% percentiles were marked by the bottom and top edges of the boxplot, respectively, and outliers were marked as blue circles.

Tab. 4.3 describes which error measures were plotted in which Figure. The PD estimation errors for the tested CLs are depicted in Fig. 4.12. The low medians in each boxplot as well as the wide violin plots at low errors indicate that most of the estimation errors are very low, for all CLs. As the CL increases, so does the spread of the PD estimation error ϵ_{PD}.

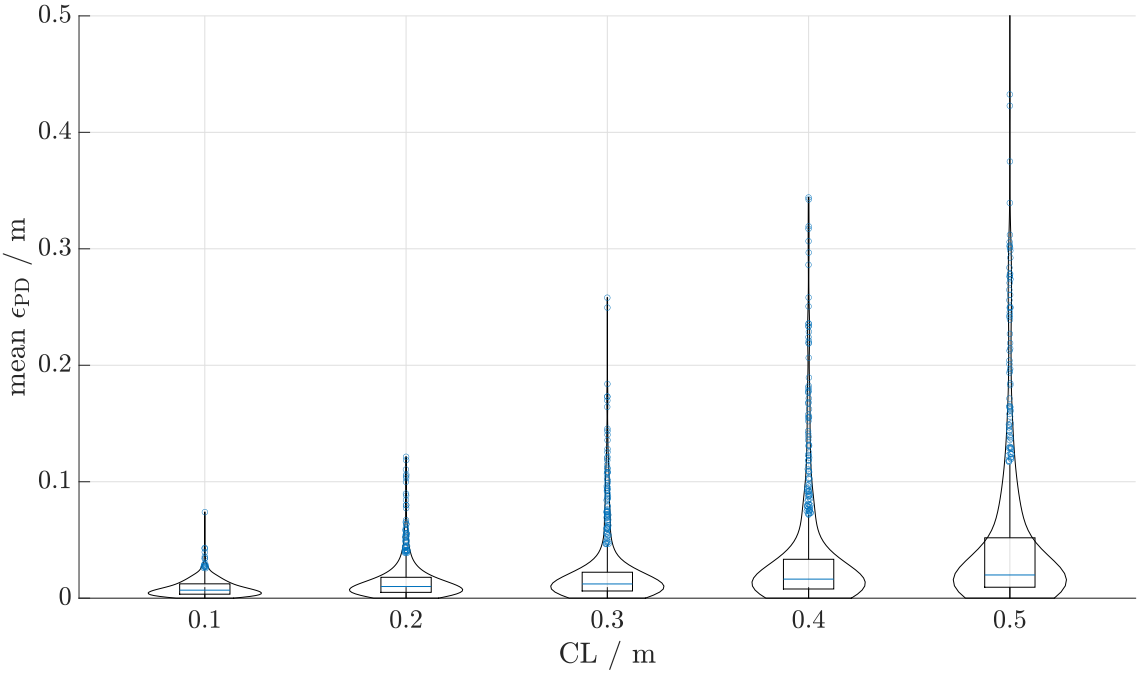


Fig. 4.12: PD estimation errors of randomly generated 3-dimensional arrays located within a virtual cube, whose size is determined by the CL.

To more easily compare how the PD estimation errors varies over CL, the PD esti-

mation error ϵ_{PD} is normalized by the respective CL in Fig. 4.13. For each CL, the normalized distribution of ϵ_{PD} looks similar which indicates that within the scope of the tested CLs, as the CL increases, the error is proportional to the size of the MAG. Comparing with Fig. 4.10, although the results are not strictly normalized in terms of amplitude, the colour axis is normalized to the PD which is being estimated. So by looking at the PD estimation error for the range of frequencies which was used in this evaluation (i.e., $f \in [0.2, 4]$ kHz), it is seen that the colour is similar across all tested PDs. This corresponds to the the trend seen in Fig. 4.13, that the PD estimation errors normalized by the CL are similar for different PDs.

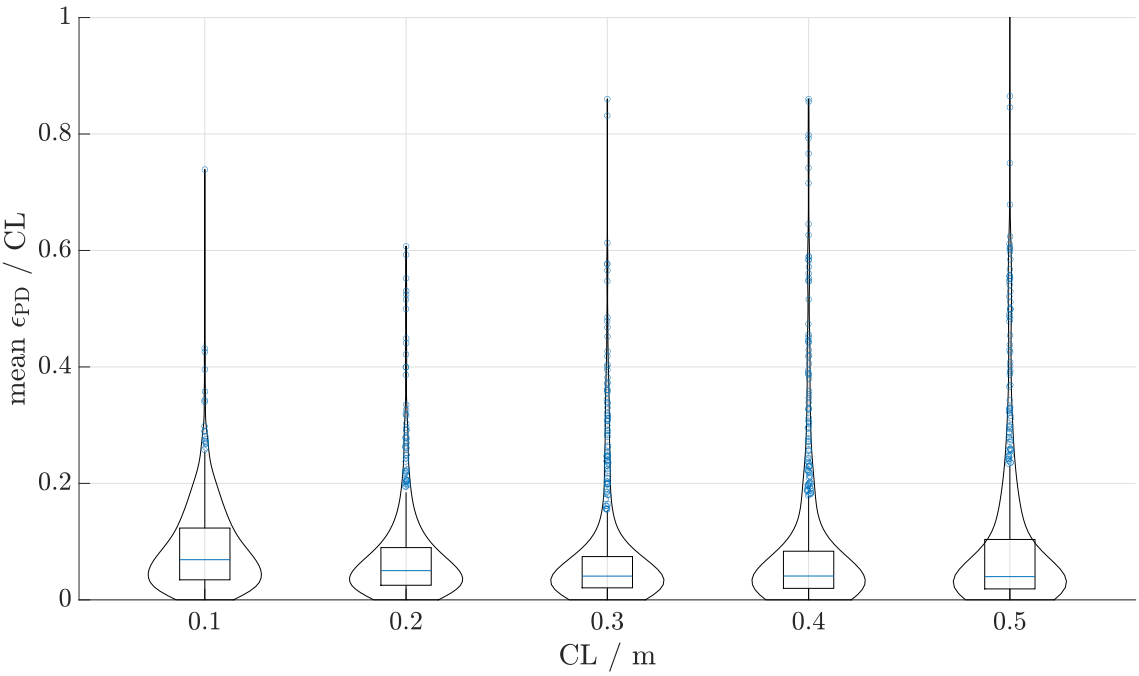


Fig. 4.13: PD estimation errors of randomly generated 3-dimensional arrays located within a virtual cube, whose size is determined by the CL, normalized by the respective CL.

The PD estimation errors of the reconstructed EDM in (2.42) are shown in Fig. 4.14. Comparing these results with Fig. 4.12 it is seen that for each CL, the distribution of the errors is slightly widened, but the density of outliers reduces. This can be explained as large estimation errors being *averaged down* and small estimation errors being *averaged up* in MDS. The estimated PDs with a high estimation error (outliers) are brought closer to a PD which resembles a rank P geometry, however, this also affects the PDs which may be estimate with a lower estimation error because they are brought closer to PDs which form a rank P geometry with the estimated PDs which are outliers.

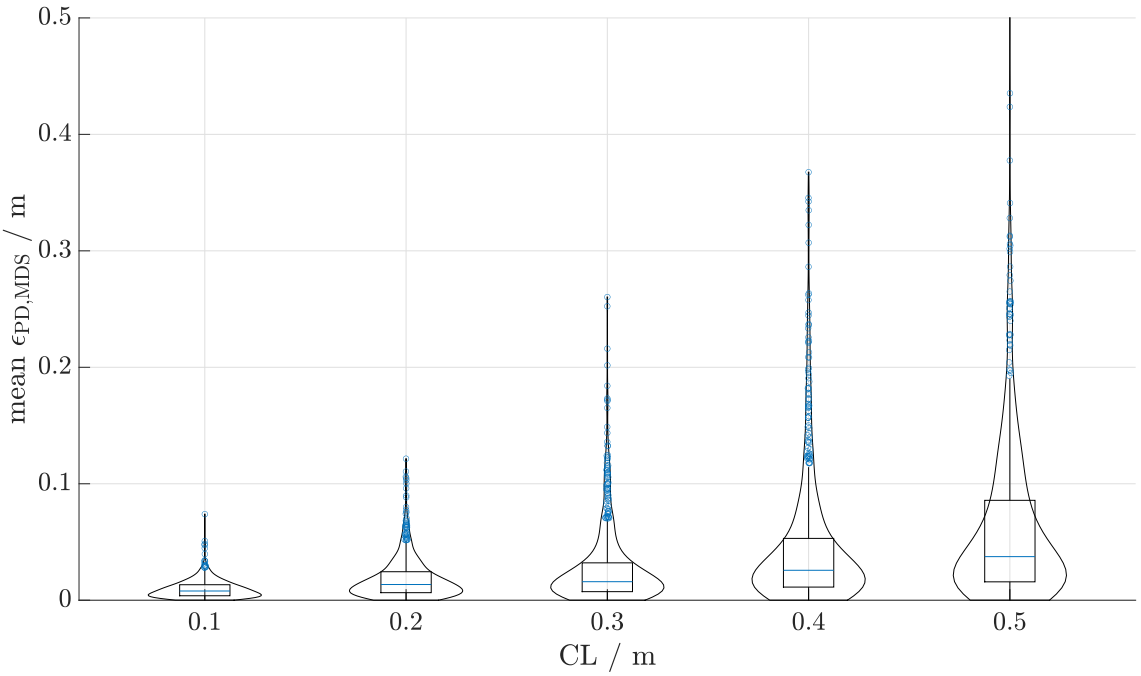


Fig. 4.14: PD estimation errors, after applying MDS, of randomly generated 3-dimensional arrays located within a virtual cube, whose size is determined by the CL.

Looking at the PD estimation error, normalized by the CL, in Fig. 4.15, compared to Fig. 4.13, the same trend is observed, but more clearly, as when comparing Fig. 4.14 with Fig. 4.12, i.e., the distribution of the errors is spread out. What is has in common with Fig. 4.13 is that the normalized distribution are similar for each CL.

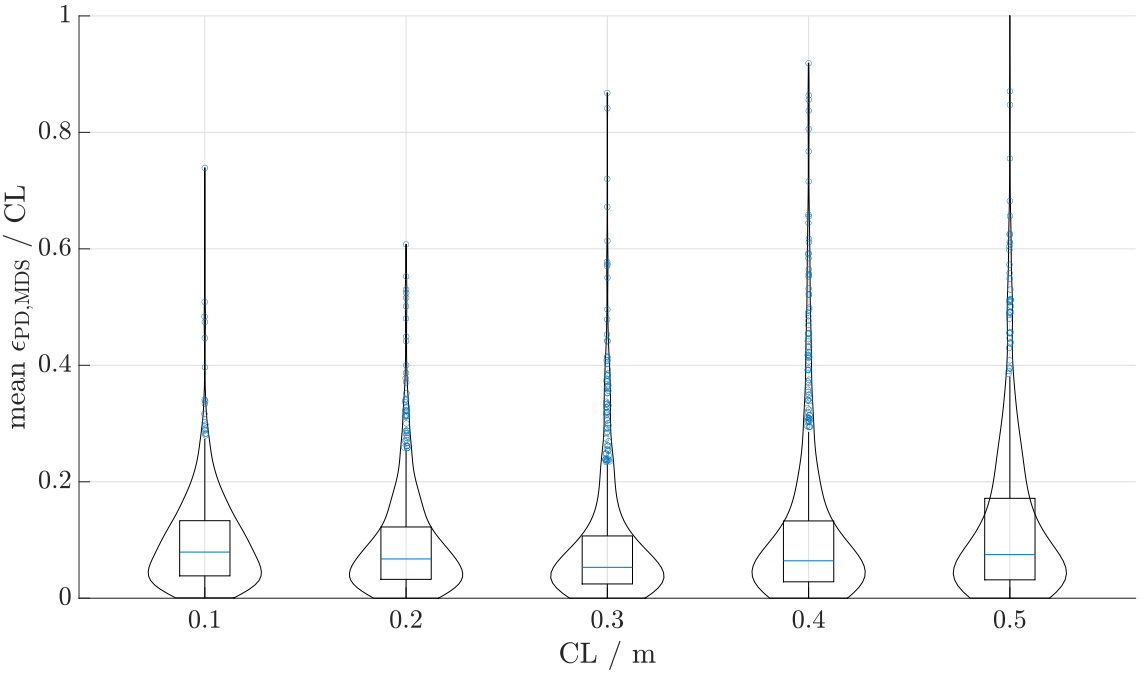


Fig. 4.15: PD estimation errors, after applying MDS, of randomly generated 3-dimensional arrays located within a virtual cube, whose size is determined by the CL, normalized by the respective CL.

The reconstructed MAG is evaluated in terms of MAG estimation error ϵ_{MAG} in Fig. 4.16. The spread of the error increases with CL and there are more errors closer to the median, compared to when looking at the PD estimation error in Fig. 4.12, where most of the PD errors which were below the median, were further below. Although the PD and MAG estimation errors are not directly comparable in terms of values, generally speaking, an array with low PD estimation errors should have low MAG estimation errors, and vice-versa. Comparing both measures, the results suggest that a few outliers in the PD estimation have an effect where they *average up* the MAG estimation errors, which would otherwise be low.

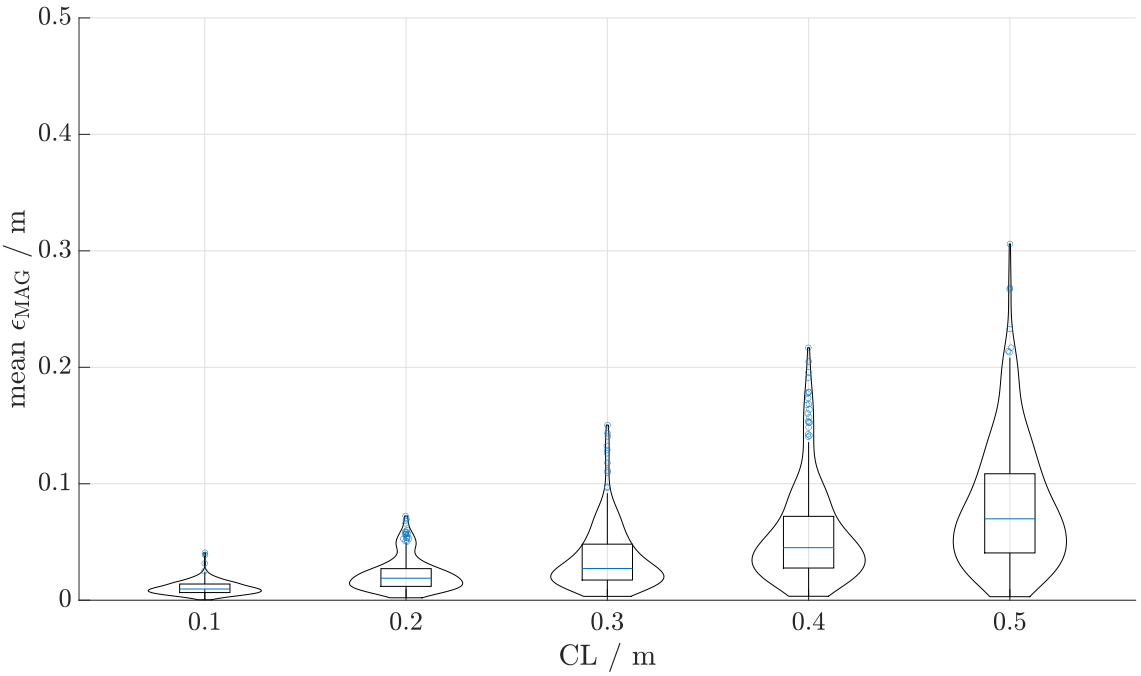


Fig. 4.16: MAG estimation error of randomly generated 3-dimensional arrays located within a virtual cube, whose size is determined by the CL.

The MAG estimation error ϵ_{MAG}, normalized by the CL, is plotted in Fig. 4.17. These results show the same trend is the same as Fig. 4.13 and Fig. 4.15, i.e., that the error normalized by the CL does not vary much. With larger distances, however, the distribution of errors at larger distances spreads out, suggesting that there are more outliers when larger PDs need to be estimated, and that as a result, they increase the MAG estimation error relative to the size of the array.

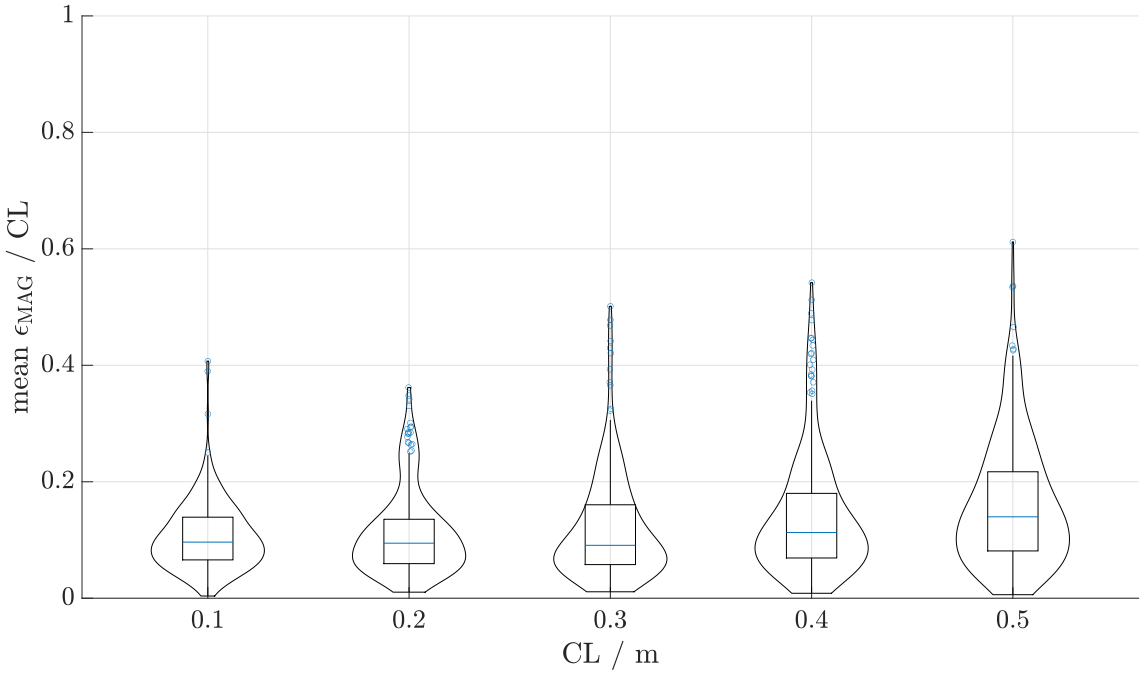


Fig. 4.17: MAG estimation errors of randomly generated 3-dimensional arrays located within a virtual cube, whose size is determined by the CL, normalized by the respective CL.

In summary, the MAG estimation framework from [14] with the proposed MPDR-TV filter and initial coherence estimate $\mathbf{\Gamma}_{(0),\mathbf{\Phi_x}}[k]$ can be used to estimate MAGs of various sizes, however, as the PDs increase past a certain distance, the likelihood that a PD is estimated with a high error increases, which can detrimentally affect the estimation of the whole MAG.

4.4.2 Estimating the Geometry of a Pair of Hearing Aids

In this Chapter, the MAG of a pair of BTE hearing-aids is estimated using the proposed modification to the coherence model in (3.12). The MAG estimation using MDS, presented in Chapter 2.4 is compared with the proposed modification incorporating prior knowledge about the hearing-aid geometry presented in Chapter 3.4.

To simulate the head-related impulse response (HRIR) of a head, recorded impulse-responses from the HRIR database [4] were used, from [4], which were measured with BTE hearing-aids on a dummy-head, 16.4 cm apart, with three microphones per device. HRIRs measured in a reverberant office scenario, with a DRR = 5 dB, were used and the source DoA (1 m from the dummy-head, on which the hearing-aids were sitting) was varied over the DoAs $\{0°, -10°, \ldots, -90°\}$. For each DoA, the RDTF $\mathbf{g}[k]$ was computed with oracle anechoic HRIRs for that DoA using the method described in [34], i.e., as the principal eigenvector of the covariance matrix of the anechoic RTF normalized by the first entry (the reference microphone). As-

Tab. 4.4: Figure guide for the evaluation of MAG and inter-aural PD estimation errors of a pair of hearing-aids

ϵ_{PD}	ϵ_{MAG}
Fig. 4.18	Fig. 4.19

suming that in spherical coordinates the elevation angle is 0°, in practice the RDTF could be estimated by estimating the DoAs based on time-differences between individual microphones of the same hearing-aid (upto a reflection, since it can't be determined from one hearing-aid whether a source is to the *left* or *right*) and estimating the inter-aural time difference (ITD) using GCC-PHAT [8], mapping the ITDs to a TDoA which corresponds to this ITD.

The PD estimation error ϵ_{PD} in (4.2) and MAG estimation error ϵ_{MAG} in (4.3) were evaluated for each DoA and grouped in a boxplot and violin plot to show the error distribution, similarly to Chapter 4.4.1. To see whether it is beneficial to incorporate more microphones to estimate the PD or MAG, the PD and MAG errors were evaluated for the $N \in \{2, 4, 6\}$ microphone cases. For the $N = 4$ and $N = 6$ cases, the modification to MDS was also compared and is named $\text{MDS}_{\text{rank1}}$, so overall, five implementations were compared in terms of PD and MAG estimation error. The $N = 2$ case, and for the $N = 4$ and $N = 6$ cases, both MDS and modified MDS ($\text{MDS}_{\text{rank1}}$) were compared.

In the cases where $N > 2$, the inter-aural PD between hearing-aids was estimated in two ways. Either the average of the square-root of the (squared-PD) entries of $\mathbf{D}_{\text{MDS}}$, corresponding to the PDs between adjacent microphones in opposite hearing-aids, was used (method name: MDS), or the average of the square-root of the corresponding entries of the reconstructed EDM from 2.42 were used, with the proposed modification to MDS in Chapter 3.4 (method name: $\text{MDS}_{\text{rank1}}$). In all tested implementations, wherever the PD could realistically be known beforehand (i.e., between microphones of the same hearing-aid), the prior knowledge was incorporated into the PD matrix to estimate the MAG.

The same frequency range as in Chapter 4.4.1, i.e., $f \in [0.2, 4]$ kHz, was used in the PD estimation and the ECM algorithm was initialized using the MPDR-TV filter and the initial coherence matrix estimate $\hat{\mathbf{\Gamma}}_{(0),\mathbf{I}}[k]$ (because it seemed like a better fit to the modified coherence in (3.12) than using $\hat{\mathbf{\Gamma}}_{(0),\mathbf{\Phi_x}}[k]$).

The PD estimation error results in Fig. 4.18 show that the PD estimation performance is very similar, regardless of the method used to estimate the PD. Comparing the results with those from Chapter 4.4.1, the difference in distribution can be at-

tributed to the change in coherence model (also in Fig. 4.18 the axes are much closer to the range of the distribution). Incorporating prior knowledge into the PD estimation has virtually no effect because the small PDs between microphones on the same device can already be estimated quite accurately. With each method, the PD is estimated with an error between 0 and 2.5 cm, corresponding to an error between 0 - 15% of the PD.

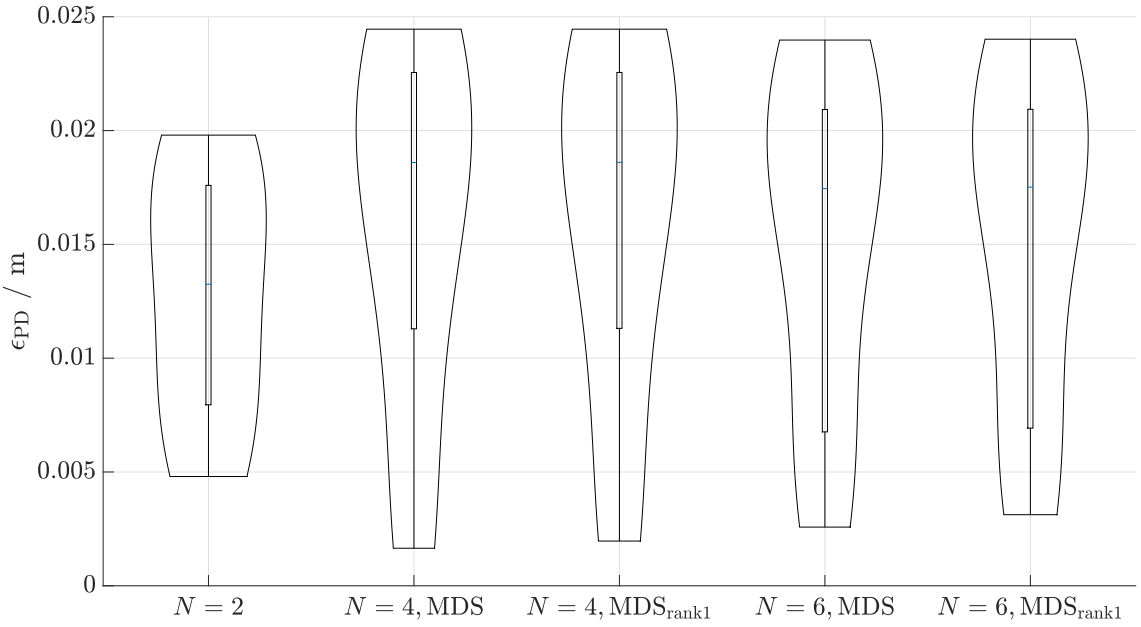


Fig. 4.18: PD estimation error of estimated inter-aural PD between hearing-aids

The MAG estimation error results are depicted in Fig. 4.19. In the case of $N = 2$, the geometry is very simple to estimate because it is only 1-dimensional, hence the estimation error is relatively low, i.e., 6.6 mm. In the case of $N = 4$, the geometry becomes slightly more complex, i.e. it is 2-dimensional (instead of 1-dimensional). The increased number of PDs to estimate, relative to the $N = 2$ geometry, introduce a small amount of error in the MAG estimation, however, the estimation error is still low, i.e., with a median error of 9 mm. For this geometry, the proposed modification to MDS slightly decreases the median estimation error to 8.3 mm. In the case of $N = 6$, the median MAG estimation error is 9.5 mm using MDS and the proposed modification to MDS reduces the median MAG estimation error to 8.8 mm and eliminates MAG estimation errors above 12 mm, changing the shape of the error distribution.

Comparing the results in Figs. 4.18 and 4.19, the proposed modification to incorporate prior knowledge in MDS provides virtually no benefit when only estimating only the inter-aural PD, however, the MAG estimation error can be reduced when estimating the MAG of an array with $N > 2$ microphones. In Chapter 4.4 it was

seen that outliers in the estimated PD can *average up* the errors of the MAG estimation. By incorporating prior knowledge and enforcing certain coordinates, those coordinates which would have been negatively affected in MDS are rectified without negatively impacting the PDs which are estimated accurately, hence why in Fig. 4.19 the distribution for the condition $N = 4, \mathrm{MDS}_{\mathrm{rank1}}$ and $N = 6, \mathrm{MDS}_{\mathrm{rank1}}$ becomes wider for lower errors.

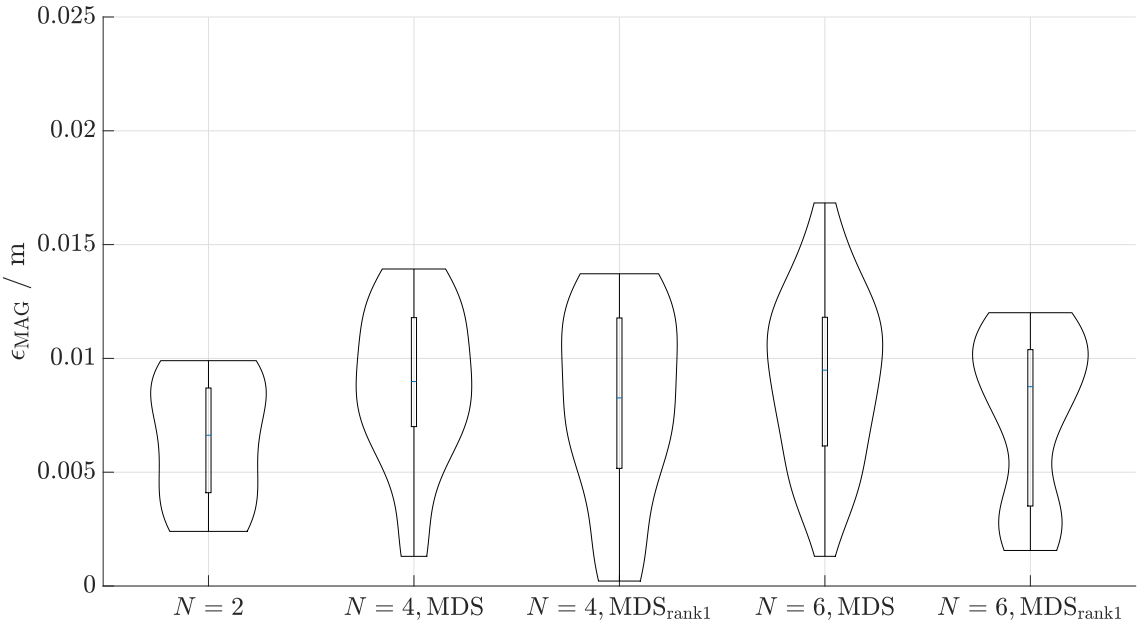


Fig. 4.19: MAG estimation error of estimated hearing-aid geometry

Summarizing the results of the hearing-aid PD and MAG estimation, the state-of-the-art framework with the proposed MPDR-TV filter in ECM and psychoacoustically motivated, modified coherence function and frequency range $f \in [0.2, 4]$ kHz to estimate the PD, can be used to estimate the distance between hearing-aids with an error below 2.5 cm. Increasing the number of microphones does not bring any benefits in terms of accurately estimating the PD or MAG. Incorporating prior knowledge (i.e., from the PDs between microphones belonging to the same hearing-aid) can *fix* outliers in the MAG estimation.

5 Conclusions and Outlook

Noisy and/or reverberant scenarios are a common problem in virtual conferencing or hearing-aid users in live scenarios. Using distributed microphones (such as hearing-aids or phones), spatial and temporal properties of the recorded speech signals at each microphone can be exploited in multi-microphone speech enhancement via denoising and/or dereverberation. Conveniently, the acoustical signal at the microphones can be used to estimate the MAG. In the case that the signal arriving at the microphones is reverberant speech, the coherence is a useful signal component which can be used to estimate the MAG, only requiring one (reverberant) source, unlike time-based MAG estimation methods which require an impractical, minimum number of sources. The state-of-the-art framework employs an ECM algorithm to estimate the coherence, then estimates the PDs by finding the model coherence which best fits the estimated coherence, and the MAG is estimated from the PDs. In this thesis, modifications were proposed to each of these steps and the modified implementations were compared with the state-of-the-art.

It was seen that ECM was sensitive to the initialization of the speech and reverberation PSDs and the coherence. It was also seen that the frequency range affected the PD estimation and the optimal frequency range was dependent on the PD, which makes it difficult to select blindly. An alternative blind initial coherence estimate was proposed which was more robust for a wider range of frequencies, requiring less careful tuning of the frequency range. The largest improvement in the PD estimation accuracy was obtained by replacing the MVDR filter which was used to estimate the initial speech and reverberation with a proposed, time-varying MPDR filter. This is because the coherence matrix in the MVDR filter relies on a very poor estimate of the initial coherence, whereas the covariance matrix in the MPDR filter can be estimated from the available recorded signal. Although the MPDR filter still requires an estimate of the RDTF vector, whose accurate blind estimation can be difficult in highly reverberant environments, it was shown for a simulated 3-dimensional microphone array that the MAG estimation error does not increase much with erroneously estimated RDTFs.

For 3-dimensional microphone arrays of various sizes, it was shown that most of the PDs were estimated with a low error, and that relative to the array size, the PD estimation error was constant for the tested sizes. However, for applications such as delay-and-sum beamforming, knowing only the PDs is not enough - the MAG must be known (or the inter-microphone spacing relative to the source, which first requires knowing the MAG). It was shown that the MAG could be estimated quite accurately using the state-of-the-art framework with proposed modifications, however, because of outliers in the PD estimation, the distribution of the MAG estimation errors was

more spread out than for the PD estimation errors, since the smaller PD errors were *averaged up*.

The state-of-the-art framework with proposed modifications was also applied to estimate the PD between hearing-aids. For this, a psycho-acoustically motivated coherence model was suggested to model the coherence between hearing-aids. The PD estimation error was low, regardless of the number of microphones used from the hearing-aids. In addition to estimating the PD, the MAG estimation was also analysed, investigating the influence of incorporating prior-knowledge. Using a proposed-modification to the MAG estimation, it was shown that the MAG could be estimated slightly more accurately than when using the prior knowledge, but not in the proposed way.

It was seen that the ECM initialization plays a large role in the robustness of the MAG estimation framework. Perhaps a more fitting application for this framework would be if it were initialized based on a known geometry and iterated online, updating changes to the MAG over time. The framework should only require minor adjustments such as an iteratively updated RDTF estimate, a coherence estimate which is recursively smoothed or averaged over a short time window (instead of averaged over all of the available data), and also the PD matrix and MAG should be estimated online.

In order to see how robust the MAG estimation is in more or less reverberant environments, the MAG estimation should be analysed for different DRRs. The influence of the signal type (e.g., coherent noise instead of speech) on the frequency range and smoothing in the time-varying MPDR filter could also be investigated.

The results of the frequency-range analysis showed that selection of the lower frequency bound was crucial for determining the PD accurately. While the proposed modifications to the framework made it more robust, especially for lower frequencies than the state-of-the-art, it could be seen that the lower frequency bound which led to the most accurate PD estimation was dependent on the PD itself. This motivates further research - perhaps it is possible to determine the optimal frequency-range on-the-fly based on the amplitude of the estimated coherence or in an alternating way, estimating the PD, then the optimal frequency range for that PD, then estimating the PD again, etc..

Although the psycho-acoustically motivated sinc coherence worked quite well for estimating the distance between hearing-aids, perhaps a better model can be found or derived to estimate the distance between hearing-aids more accurately.

References

[1] O. Schwartz, S. Gannot, and E. Habets, "An expectation-maximization algorithm for multimicrophone speech dereverberation and noise reduction with coherence matrix estimation," *TASLP*, vol. 24, no. 9, pp. 1495–1510, 2016.

[2] I. Lindevald and A. Benade, "Two-ear correlation in the statistical sound fields of rooms," *JASA*, vol. 80, no. 2, pp. 661–664, 1986.

[3] E. Habets, *RIR-Generator*. Available at `https://github.com/ehabets/RIR-Generator`.

[4] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP JASP*, pp. 1–10, 2009.

[5] S. Doclo, S. Gannot, M. Moonen, A. Spriet, S. Haykin, and K. Liu, "Acoustic beamforming for hearing aid applications," *Handbook on array processing and sensor networks*, pp. 269–302, 2010.

[6] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *TSP*, vol. 50, no. 9, pp. 2230–2244, 2002.

[7] A. Plinge, F. Jacob, R. Haeb-Umbach, and G. Fink, "Acoustic microphone geometry calibration: An overview and experimental evaluation of state-of-the-art algorithms," *SPM*, vol. 33, no. 4, pp. 14–29, 2016.

[8] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *TASSP*, vol. 24, no. 4, pp. 320–327, 1976.

[9] M. Pollefeys and D. Nister, "Direct computation of sound and microphone locations from time-difference-of-arrival data," in *ICASSP*, pp. 2445–2448, IEEE, 2008.

[10] N. Gaubitch, B. Kleijn, and R. Heusdens, "Auto-localization in ad-hoc microphone arrays," in *ICASSP*, pp. 106–110, IEEE, 2013.

[11] I. Dokmanić, L. Daudet, and M. Vetterli, "How to localize ten microphones in one finger snap," in *EUSIPCO*, pp. 2275–2279, IEEE, 2014.

[12] B. Champagne, S. Bédard, and A. Stéphenne, "Performance of time-delay estimation in the presence of room reverberation," *TSAP*, vol. 4, no. 2, pp. 148–152, 1996.

[13] I. McCowan, M. Lincoln, and I. Himawan, "Microphone array shape calibration in diffuse noise fields," *TASLP*, vol. 16, no. 3, pp. 666–670, 2008.

[14] O. Schwartz, A. Plinge, E. Habets, and S. Gannot, "Blind microphone geometry calibration using one reverberant speech event," in *WASPAA*, pp. 131–135, IEEE, 2017.

[15] X.-L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ecm algorithm: A general framework," *Biometrika*, vol. 80, no. 2, pp. 267–278, 1993.

[16] S. T. Birchfield, "Geometric microphone array calibration by multidimensional scaling," in *ICASSP*, vol. 5, pp. V–157, IEEE, 2003.

[17] P. Pertilä, M. Mieskolainen, and M. Hämäläinen, "Passive self-localization of microphones using ambient sounds," in *EUSIPCO*, pp. 1314–1318, IEEE, 2012.

[18] E. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *JASA*, vol. 122, no. 6, pp. 3464–3470, 2007.

[19] R. Cook, R. Waterhouse, R. Berendt, S. Edelman, and M. Thompson Jr, "Measurement of correlation coefficients in reverberant sound fields," *JASA*, vol. 27, no. 6, pp. 1072–1077, 1955.

[20] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *TASLP*, vol. 25, no. 4, pp. 692–730, 2017.

[21] Y. Avargel and I. Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *TASLP*, vol. 15, no. 4, pp. 1305–1319, 2007.

[22] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *TSAP*, vol. 13, no. 5, pp. 845–856, 2005.

[23] H. Ye and R. DeGroat, "Maximum likelihood doa estimation and asymptotic cramér-rao bounds for additive unknown colored noise," *TSP*, vol. 43, no. 4, pp. 938–949, 1995.

[24] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids," in *EUSIPCO*, pp. 61–65, IEEE, 2014.

[25] B. Van Veen and K. Buckley, "Beamforming: A versatile approach to spatial filtering," *ASSP*, vol. 5, no. 2, pp. 4–24, 1988.

[26] J. Li, P. Stoica, and Z. Wang, "On robust capon beamforming and diagonal loading," *TSP*, vol. 51, no. 7, pp. 1702–1715, 2003.

[27] A. Elnashar, S. M. Elnoubi, and H. A. El-Mikati, "Further study on robust adaptive beamforming with optimum diagonal loading," *TAP*, vol. 54, no. 12, pp. 3647–3658, 2006.

[28] I. Dokmanic, R. Parhizkar, J. Ranieri, and M. Vetterli, "Euclidean distance matrices: essential theory, algorithms, and applications," *SPM*, vol. 32, no. 6, pp. 12–30, 2015.

[29] J. C. Gower, "Euclidean distance geometry," *Math. Sci*, vol. 7, no. 1, pp. 1–14, 1982.

[30] P. Schoenemann, *A solution of the orthogonal Procrustes problem with applications to orthogonal and oblique rotation*. PhD thesis, University of Illinois at Urbana-Champaign, 1964.

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

[32] P. Kabal, *TSP Speech Database*. Available at `http://www-mmsp.ece.mcgill.ca/Documents/Data/`.

[33] P. Naylor and N. Gaubitch, *Speech Dereverberation*. Berlin, Germany: Springer-Verlag, 2010.

[34] N. Gößling and S. Doclo, "Relative transfer function estimation exploiting spatially separated microphones in a diffuse noise field," in *IWAENC*, pp. 146–150, IEEE, 2018.

[35] C. Hummersone and T. Prätzlich, *IoSR Matlab Toolbox*. Available at `https://github.com/IoSR-Surrey/MatlabToolbox`.

[36] J. Hintze and R. Nelson, "Violin plots: a box plot-density trace synergism," *The American Statistician*, vol. 52, no. 2, pp. 181–184, 1998.

Eigenständigkeitserklärung:

I hereby confirm that this thesis is entirely my own work. I confirm that no part of the document has been copied from either a book or any other source – including the internet – except where such sections are clearly shown as quotations and the sources have been correctly identified within the text or in the list of references. Moreover I confirm that I have taken notice of the 'Leitlinien guter wissenschaftlicher Praxis' of the University of Oldenburg.

Oldenburg, April 6, 2021 Klbrümann

Ort, Datum Unterschrift