Carl von Ossietzky
University of Oldenburg

Bachelor Thesis

# Combination of RTF Vector Estimates in Acoustic Sensor Networks

Bachelor of Engineering Program
Engineering Physics

**Submitted by**
Wiebke Middelberg

**Supervisor**
Prof. Dr. ir. Simon Doclo

**Co-Supervisor**
Nico Gößling, M. Sc.

Oldenburg, May 22, 2019

## Abstract

One of the main goals in audio signal processing is the suppression of background noise to increase the intelligibility of speech. A common noise reduction method is the so-called minimum variance distortionless response (MVDR) beamformer, which preserves the speech component in a reference microphone while minimizing the output noise power spectral density (PSD). In multi-channel noise reduction, this beamformer requires an estimate of the relative transfer function (RTF) vector, which contains the acoustic transfer functions (ATFs) relative to a reference microphone. In the past, several RTF vector estimators have been developed, amongst them, the spatial coherence (SC) -based RTF vector estimator, which uses a spatially separated external microphone to guide a local array's RTF vector estimate. In this thesis, an extension of the SC method is considered: A second external microphone is used and thus a second RTF vector estimate is obtained. Two approaches are proposed for an optimal linear combination, namely, the orthogonal projection of the true RTF vector on the plane spanned by the two estimates and the maximization of the output signal-to-noise ratio (SNR). The results show that the orthogonal projection (which is a purely theoretical, *ideal* solution) performs well in terms of both, noise reduction and speech distortion. The approach which maximizes the output SNR leads to similar results, but can be applied in practice when the biased output SNR is maximized.

## Zusammenfassung

Eines der Hauptziele der Audiosignalverarbeitung ist die Unterdrückung von Hintergrundgeräuschen, um die Sprachverständlichkeit zu erhöhen. Ein gängiges Verfahren zur Störgeräuschunterdrückung ist der so genannte Minimum Variance Distortionless Response (MVDR)-Beamformer, der die Sprachkomponente in einem Referenzmikrophon bewahrt und gleichzeitig die spektrale Leistungsdichte (PSD) des Störgeräuschs am Ausgang minimiert. Dieser Beamformer benötigt eine Schätzung des Vektors der relativen Übertragungsfunktionen (RTFs), in dem die akustischen Übertragungsfunktionen (ATFs) auf ein Referenz-Mikrophon bezogen sind. In der Vergangenheit wurden mehrere Ansätze der RTF-Schätzung entwickelt, darunter der auf räumlicher Kohärenz (SC) basierende RTF-Schätzer, der ein räumlich separiertes externes Mikrophon verwendet, um die RTF-Schätzung eines lokalen Arrays zu steuern. In dieser Arbeit wird eine Erweiterung der SC-Methode in Betracht gezogen: Ein zweites externes Mikrophon wird verwendet und somit eine zweite RTF-Vektor-Schätzung erhalten. Als Ansätze für eine optimale lineare Kombination werden die orthogonale Projektion des wahren RTF-Vektors auf der von den beiden Schätzungen aufgespannten Ebene und die Maximierung des Signal-Rausch-Abstandes (SNR) vorgeschlagen. Die Ergebnisse zeigen, dass die orthogonale Projektion (die eine rein theoretische, *ideale* Lösung ist) sowohl in Bezug auf Störgeräuschunterdrückung als auch

auf Sprachverzerrung gute Ergebnisse liefert. Der Ansatz, der den Ausgangs-SNR maximiert, führt zu ähnlichen Ergebnissen, kann allerdings in der Praxis angewendet werden, wenn der verzerrte Ausgangs-SNR maximiert wird.

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| **ATF** | acoustic transfer function |
| **ASN** | acoustic sensor network |
| **CS** | covariance subtraction |
| **CW** | covariance whitening |
| **E** | external microphone |
| **EVD** | eigenvalue decomposition |
| **FFT** | fast Fourier transform |
| **GEVD** | generalized eigenvalue decomposition |
| **L** | local array |
| **MVDR** | minimum variance distortionless response |
| **PSD** | power spectral density |
| **RIR** | room impulse response |
| **RTF** | relative transfer function |
| **S** | desired source |
| **SC** | spatial coherence |
| **SD** | speech distortion |
| **SNR** | signal-to-noise ratio |
| **SPP** | speech presence probability |
| **STFT** | short-time Fourier transform |
| **VAD** | voice activity detection |
| **WOLA** | weighted overlap add |

# Contents

# 1 Introduction

The intelligibility of speech in acoustic signals is often reduced by background noise. Especially at low signal-to-noise ratios (SNRs) or for people with hearing impairment, understanding a desired speaker can become difficult or even impossible. An example is the so-called ”cocktail-party” scenario [1], where a desired speech signal is corrupted by other interfering speakers. Noise reduction is therefore one of the essential tasks in audio signal processing to increase speech intelligibility.
In principle, it can be distinguished between single- and multi-channel noise reduction techniques. Using only one channel merely allows the use of spectro-temporal information, whereas multi-channel techniques also allow exploration of the spatial information about the target and/or noise component in the signal [2]. The latter yields, in general, better results than using only spectro-temporal information.
Since most modern devices like smart phones, hands-free systems, or hearing aids, are equipped with more than one microphone, multi-channel noise reduction is the method of choice in this thesis.

Especially in acoustic sensor networks (ASNs), where microphones are distributed over a larger area and can therefore gather even more diverse information about the ambient sound field, multi-channel algorithms for noise reduction, speech recognition and source localization are of high interest in current research [3].
One method of multi-channel noise reduction is the minimum variance distortionless response (MVDR) beamformer, which perfectly preserves the speech component in the reference microphone (i.e. distortionless) while minimizing the noise at the output. The beamformer can be steered by means of a relative transfer function (RTF) vector [4], which represents the relative position which is to be preserved. The RTF vector relates the desired source's acoustic transfer function (ATF) of all microphones to a reference microphone and is easier to estimate than ATFs [5]. The state-of-the-art RTF estimators are covariance subtraction (CS) and covariance whitening (CW) [6]. It has been shown that among these two estimators CW performs best in terms of the estimate's accuracy, but has a high computational complexity due to the required eigenvalue decomposition (EVD) [6, 7].
In [8, 9] a novel RTF estimator was proposed which exploits the spatial coherence (SC) properties of a diffuse noise field between a microphone array (e.g. binaural hearing aid devices) and a spatially separated external microphone. Under the assumption of a diffuse noise field and a sufficiently large distance between local array (L) and external microphone (E), only the speech components in the signals of L and E are correlated, which allows the use of an external microphone to guide

the RTF estimate of the local array. A configuration like this can be interpreted as an ASN with one (external) node and a fusion center where all information is collected and processed [10, 11]. This estimator obviously requires the availability of an external microphone, but yields comparable results to CW in terms of noise reduction, while exhibiting a much lower computational complexity. Furthermore, the real-time implementation of the SC method in [12] shows robust results and applicability in practice, even in dynamic scenarios with a moving desired source.

In this thesis, the SC estimator is extended with a second external microphone $E_2$. By making the same assumptions as with only one external microphone $E_1$ (ideally only the desired speech in L and $E_2$ is correlated), a second RTF vector estimate can be obtained using $E_2$. Hence, the question arises which estimate should be used or how they can be optimally combined. This problem is examined in the course of this thesis, where two (mostly theoretical) approaches are presented. Both optimal solutions are based on a linear combination of the two RTF vector estimates with a complex weighting factor. In the first approach (Subsection 5.1), the Hermitian angle - the angle between the true RTF vector (which is in general unknown) and the combination of estimates - is minimized, for which it can be shown that the orthogonal projection of the true RTF vector on the plane spanned by the estimates solves this problem. The second approach (Subsection 5.2) is the maximization of the output SNR. In principle, this problem is based on oracle knowledge about the speech component (i.e. information that is unknown in practice). However, the optimization of the biased output SNR yields, in theory, the same result and does not require information about the speech component directly. The solutions of these optimization problems are derived and evaluated.

This thesis is structured as follows: In Section 2, the notation used throughout this thesis, as well as the multi-channel signal model are introduced. Subsequently, the MVDR beamformer is introduced in Section 3, which is steered by the RTF vectors obtained from the SC method presented in Section 4. In Section 5, both the minimization of the Hermitian angle and the maximization of the output SNR are derived and discussed. The evaluation of the proposed optimal solutions is carried out in Section 6 using signals recorded with a binaural hearing aid device and several external microphones available. Firstly, the effect of the distance dependence of the two external microphones on the weighting of the RTF vector estimates is observed (Subsection 6.3). To this end, artificial levelling is required. The same holds for the experiment performed in Subsection 6.4, where the influence of the input SNR in

$E_1$ and $E_2$ is investigated. In Subsection 6.5, no artificial levelling is performed and thus both influences are taken into account. All experiments are performed with oracle knowledge. As performance measures, the output SNR of the filter and the speech distortion (SD), i.e. the undesired effects of the filter on the desired speech signal, are considered. Finally, the work presented in this thesis is concluded and an overview of further research is provided (Section 7).

# 2 Signal Model and Notation

In the following, an ASN with a local array (for simplicity here chosen to be linear) with $M$ microphones and two spatially separated external microphones ($\mathrm{E}_1$ and $\mathrm{E}_2$) is considered, as depicted in Fig. 2.1.

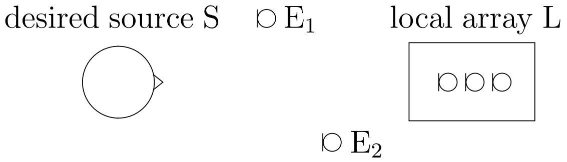

Fig. 2.1: Configuration with $M = 3$ local microphones and two external microphones.

The desired source (S) is placed some distance from the local array and is related to each microphone by the respective room impulse response (RIR). The recorded discrete time signal in the local array (indicated with the index L) $y_{\mathrm{L},m}[t]$ (with $m \in \{1, ..., M\}$) is given as the sum of the desired signal $x_{\mathrm{L},m}[t]$ and the noise component $n_{\mathrm{L},m}[t]$ with the sampling index $t$, that is

$$y_{\mathrm{L},m}[t] = x_{\mathrm{L},m}[t] + n_{\mathrm{L},m}[t] \,. \tag{2.1}$$

The signals for the external microphones $y_{\mathrm{E},1}$ and $y_{\mathrm{E},2}$ are defined similarly.
Using the short-time Fourier transform (STFT), the frequency domain signal $Y_{\mathrm{L},m}[k,l]$ is obtained as

$$Y_{\mathrm{L},m}[k,l] = \sum_{t=0}^{L_{\mathrm{FFT}}-1} y_{\mathrm{L},m}[l \cdot L_{\mathrm{R}} + t] \, w[t] \, e^{-j2\pi kt/L_{\mathrm{FFT}}} \,, \tag{2.2}$$

where $k$ is the frequency bin index, $l$ the frame index, $L_{\mathrm{FFT}}$ the frame length (here equal to the fast Fourier transform (FFT) length), $L_{\mathrm{R}}$ the hop size, and $j$ the imaginary unit ($j^2 = -1$). The samples in each frame are weighted with the window function $w[t]$. The frequency domain signals for the external microphones $Y_{\mathrm{E},1}[k,l]$ and $Y_{\mathrm{E},2}[k,l]$ are also obtained using (2.2) and are denoted with the index E. In the following, $k$ and $l$ are neglected for a simpler representation.

The frequency domain signals of all local microphones are stacked into one vector $\mathbf{y}$

$$\mathbf{y} = [Y_{\mathrm{L},1}, ..., Y_{\mathrm{L},M}]^T \tag{2.3}$$

and the extended vector $\bar{\bar{\mathbf{y}}}$ is defined as

$$\bar{\bar{\mathbf{y}}} = [\mathbf{y}^T, Y_{\mathrm{E},1}, Y_{\mathrm{E},2}]^T \,. \tag{2.4}$$

For the (extended) speech and noise the vectors $\mathbf{x}$ and $\mathbf{n}$ ($\bar{\bar{\mathbf{x}}}$ and $\bar{\bar{\mathbf{n}}}$ respectively) similar definitions are used. Here, $(\cdot)^T$ denotes the transpose operator and in the following $(\cdot)^H$ is used for the Hermitian operator.

Using (2.2) and the property that the speech vector $\mathbf{x}$ corresponds to the clean speech signal $S$ multiplied by the ATFs (of the source to each microphone of the local array) which are stacked in the vector $\mathbf{a}$ (similar to (2.3)), the signal model is given as

$$\begin{aligned} \mathbf{y} &= \mathbf{x} + \mathbf{n} \\ &= \mathbf{a}S + \mathbf{n} \,. \end{aligned} \tag{2.5}$$

The multiplication $\mathbf{a}S$ corresponds to the convolution of the RIR with the speech signal $s[t]$ in the time domain.
The ATF vector, which is in general not known and difficult to estimate, can be replaced with the RTF vector $\mathbf{h}$ by relating $\mathbf{a}$ to a reference microphone, which can be chosen freely without loss of generality, i.e.

$$\mathbf{h} = \frac{\mathbf{a}}{\mathbf{e}^T\mathbf{a}} \,, \tag{2.6}$$

with the selection vector $\mathbf{e}$ being a $M \times 1$-dimensional column vector which contains all zeros except for the entry corresponding to the reference microphone. With $X_{\mathrm{ref}}$ being the speech signal in the reference microphone, the speech component $\mathbf{x}$ in (2.5) is given by

$$\mathbf{x} = \mathbf{h}X_{\mathrm{ref}} \,. \tag{2.7}$$

Using the signal model in (2.5) and the assumption of statistical independence of $\mathbf{x}$ and $\mathbf{n}$, the covariance matrix of the noisy signal can be written as the sum of the covariance matrices of speech and noise, i.e.

$$\begin{aligned} \mathbf{R}_{\mathrm{y}} &= \mathcal{E}\{\mathbf{y}\mathbf{y}^H\} \\ &= \mathcal{E}\{\mathbf{x}\mathbf{x}^H\} + \mathcal{E}\{\mathbf{n}\mathbf{n}^H\} \\ &= \mathbf{R}_{\mathrm{x}} + \mathbf{R}_{\mathrm{n}} \,, \end{aligned} \tag{2.8}$$

where $\mathcal{E}\{\cdot\}$ denotes the expectation operator, for which the cross terms $\mathcal{E}\{\mathbf{x}\mathbf{n}^H\}$ and $\mathcal{E}\{\mathbf{n}\mathbf{x}^H\}$ fade due to the independence of $\mathbf{x}$ and $\mathbf{n}$. The covariance matrices extended with two external microphones are given by $\bar{\bar{\mathbf{R}}}_{\mathrm{y}}$, $\bar{\bar{\mathbf{R}}}_{\mathrm{x}}$ and $\bar{\bar{\mathbf{R}}}_{\mathrm{n}}$, respectively, and are defined similarly with the extended signal vectors. The dimensions of the extended covariance matrices are therefore $(M{+}2)\times(M{+}2)$ instead of $M\times M$. If only one external microphone is available (as in the original RTF estimator reviewed in Section 4), only one over-bar is used, i.e. $\bar{\mathbf{R}}_{\mathrm{y}}$, $\bar{\mathbf{R}}_{\mathrm{x}}$ and $\bar{\mathbf{R}}_{\mathrm{n}}$.

The output signal $Z$ of the input signal $\mathbf{y}$, filtered with the filter vector $\mathbf{w} \in \mathbb{C}^M$, is defined as the sum of all filtered signals of the local microphones, that is

$$Z = \mathbf{w}^H \mathbf{y}\,. \tag{2.9}$$

In the next section, the MVDR beamformer is briefly reviewed, which yields a filter vector that depends on the noise covariance matrix and the RTF vector.

As an objective measure for noise reduction, the SNR improvement $\Delta \mathit{SNR}$, i.e.

$$\Delta \mathit{SNR} = \mathit{SNR}_{\mathrm{out}} - \mathit{SNR}_{\mathrm{in}}\,, \tag{2.10}$$

is considered, where $\mathit{SNR}_{\mathrm{in}}$ is the input SNR in the reference microphone, which is given by

$$\mathit{SNR}_{\mathrm{in}} = \frac{\mathbf{e}^T \mathbf{R}_{\mathrm{x}} \mathbf{e}}{\mathbf{e}^T \mathbf{R}_{\mathrm{n}} \mathbf{e}}\,. \tag{2.11}$$

The output SNR $\mathit{SNR}_{\mathrm{out}}$, which depends on the filter vector is given by

$$\mathit{SNR}_{\mathrm{out}} = \frac{\mathbf{w}^H \mathbf{R}_{\mathrm{x}} \mathbf{w}}{\mathbf{w}^H \mathbf{R}_{\mathrm{n}} \mathbf{w}}\,. \tag{2.12}$$

Another measure used in this thesis is speech distortion $\mathit{SD}$, which is the ratio of the speech power spectral density (PSD) $\phi_{\mathrm{x}}$ in the reference microphone and the speech PSD , $\mathbf{w}^H \mathbf{R}_{\mathrm{x}} \mathbf{w}$, in the output signal, i.e.

$$\mathit{SD} = \frac{\phi_{\mathrm{x}}}{\mathbf{w}^H \mathbf{R}_{\mathrm{x}} \mathbf{w}}\,. \tag{2.13}$$

# 3 MVDR Beamforming

The minimum variance distortionless response (MVDR) beamformer aims to minimize the output noise PSD, while preserving the speech component in the reference microphone. It must therefore be "steered" correctly towards the desired source, which can be done either by means of the ATF vector or with the RTF vector, which is easier to estimate. Therefore, in the following the RTF-steered MVDR beamformer is considered.

Mathematically, the MVDR beamformer is formulated as the constrained optimization problem [5]

$$\min_{\mathbf{w}} \mathbf{w}^H \mathbf{R}_\text{n} \mathbf{w} \qquad , \text{ s.t. } \mathbf{w}^H \mathbf{h} = 1 \,. \tag{3.1}$$

The constraint ensures the preservation of the desired source's signal in the reference microphone and the cost function itself is the residual noise PSD, which is to be minimized with respect to $\mathbf{w}$.

To derive the solution (as in [2, 13]) for the complex valued filter vector $\mathbf{w}$, the method of Lagrangian multipliers is used. The Lagrangian function then reads

$$\mathcal{L}(\mathbf{w}) = \mathbf{w}^H \mathbf{R}_\text{n} \mathbf{w} + \lambda(\mathbf{w}^H \mathbf{h} - 1) \,, \tag{3.2}$$

where $\lambda$ is the Lagrangian multiplier. According to the rules for vector derivation (for proofs see [2])

$$J = \mathbf{v}^H \mathbf{u}, \qquad \frac{dJ}{d\mathbf{v}} = \mathbf{u} \tag{3.3}$$

and

$$J = \mathbf{v}^H \mathbf{u} \mathbf{u}^H \mathbf{v}, \qquad \frac{dJ}{d\mathbf{v}} = 2\mathbf{u}\mathbf{u}^H \mathbf{v} \,, \tag{3.4}$$

where $\mathbf{u}$ and $\mathbf{v}$ are defined as complex valued vectors and $J$ is a scalar, taking the derivative of $\mathcal{L}(\mathbf{w})$ with respect to $\mathbf{w}$ gives

$$\frac{d\mathcal{L}(\mathbf{w})}{d\mathbf{w}} = 2\mathbf{R}_\text{n} \mathbf{w} + \lambda \mathbf{h} \,. \tag{3.5}$$

To find the stationary point, (3.5) is set equal to the $M \times 1$ zero vector $\mathbf{0}_{M\times 1}$, which yields

$$2\mathbf{R}_\text{n} \mathbf{w} + \lambda \mathbf{h} = \mathbf{0}_{M\times 1} \,. \tag{3.6}$$

Rearranging (3.6) to $\mathbf{w}$ gives

$$\mathbf{w} = -\frac{1}{2}\lambda\mathbf{R}_{\mathrm{n}}^{-1}\mathbf{h}\,, \tag{3.7}$$

with $(\cdot)^{-1}$ denoting the inverse operator for matrices. Substituting (3.7) into the constraint in (3.1) gives

$$(-\frac{1}{2}\lambda\mathbf{R}_{\mathrm{n}}^{-1}\mathbf{h})^{H}\mathbf{h} = 1\,. \tag{3.8}$$

From (3.8), the Lagrangian multiplier

$$\lambda = -2\frac{1}{\mathbf{h}^{H}\mathbf{R}_{\mathrm{n}}^{-1}\mathbf{h}} \tag{3.9}$$

is obtained. Re-substituting (3.9) into (3.7) then yields the solution for the filter vector

$$\mathbf{w} = \frac{\mathbf{R}_{\mathrm{n}}^{-1}\mathbf{h}}{\mathbf{h}^{H}\mathbf{R}_{\mathrm{n}}^{-1}\mathbf{h}}\,. \tag{3.10}$$

To use this filter in practice, an estimate of the noise covariance matrix $\mathbf{R}_{\mathrm{n}}$ and the RTF vector $\mathbf{h}$ is needed. Former can be obtained in noise-only frames, which requires the availability of a voice activity detection (VAD), such as the speech presence probability (SPP) estimator in [14].

For the RTF vector estimation, several approaches exist that either assume a fixed direction of the desired source (fixed beamforming) or adaptively estimate the RTF vector. In the next section, an adaptive RTF vector estimator is reviewed that assumes the availability of an external microphone and a diffuse noise field.

# 4 RTF Estimation Exploiting Spatial Coherence

The RTF estimation using the spatial coherence (SC) method (in [8, 9]) requires an external microphone to guide the local array's RTF estimate. In this section, this estimator is presented for the case that only one external microphone is available. The extended covariance matrices therefore have the dimensions $(M{+}1)\times(M{+}1)$.

The assumption under which this estimator was proposed are a diffuse noise field and a sufficiently large distance between the local array and the external microphone. The coherence of diffuse noise between two microphones can be modeled with the normalized Γ-function [15]

$$\Gamma(\nu, d) = \mathrm{sinc}\left(\frac{2\pi\nu d}{c}\right), \tag{4.1}$$

which is a function of the frequency $\nu$ and the distance $d$ between two microphones. $c$ denotes the speed of sound.

In Fig. 4.1, the Γ-function is shown as a function of frequency for three different distances. It is clearly visible that for larger spacing between microphones, the coherence is lower, especially at high frequencies.

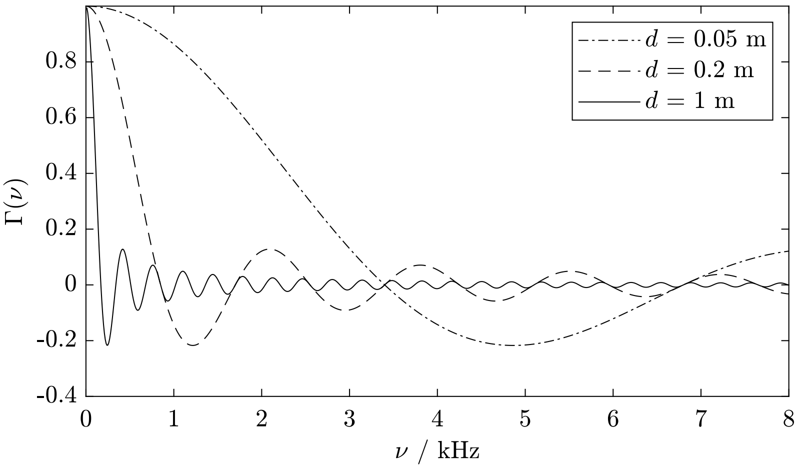

Fig. 4.1: Γ-function for the distances $d = \{0.05, 0.2, 1\}$ m in a frequency range of 0-8 kHz.

The last column of the extended noisy covariance matrix is simply the correlation between L and E, which can be written as

$$\begin{aligned}\mathcal{E}\{\mathbf{y}Y_\mathrm{E}^*\} &= \mathcal{E}\{(\mathbf{x}+\mathbf{n})(X_\mathrm{E}^*+N_\mathrm{E}^*)\} \\ &= \mathcal{E}\{\mathbf{x}X_\mathrm{E}^*\} + \mathcal{E}\{\mathbf{n}N_\mathrm{E}^*\},\end{aligned} \tag{4.2}$$

using the signal model in (2.5) and the assumption of statistical independence of noise and speech. Now assuming sufficient distance between L and E, the correlation of the noise signals becomes negligibly small and is thus assumed to be zero, i.e.

$$\mathcal{E}\{\mathbf{n}N_{\mathrm{E}}^{*}\} = \mathbf{0}_{M\times 1}\,, \tag{4.3}$$

according to the noise model in (4.1). Therefore, the extended noise covariance matrix can be written as

$$\bar{\mathbf{R}}_{\mathrm{n}} = \begin{bmatrix} \mathbf{R}_{\mathrm{n}} & \mathbf{0}_{M\times 1} \\ \mathbf{0}_{M\times 1}^{T} & \phi_{\mathrm{n}} \end{bmatrix}. \tag{4.4}$$

The last element of the last column is the noise PSD $\phi_{\mathrm{n}}$, which is the same in all microphones in a homogeneous diffuse noise field.
With (2.8) also holding for the extended covariance matrices and (4.4), the first $M$ entries of the last column of $\bar{\mathbf{R}}_{\mathrm{y}}$ only consist of the speech correlation of the local array and the external microphone, that is

$$[\mathbf{I}_{M\times M}, \mathbf{0}_{M\times 1}]\,\bar{\mathbf{R}}_{\mathrm{y}}\mathbf{e}_{\mathrm{E}} = \mathcal{E}\{\mathbf{x}X_{\mathrm{E}}^{*}\}\,. \tag{4.5}$$

$\mathbf{e}_{\mathrm{E}}$ is a selection vector corresponding to the external microphone, i.e. $\mathbf{e}_{\mathrm{E}} = [\mathbf{0}_{M\times 1}^{T}, 1]^{T}$, and $\mathbf{I}_{M\times M}$ denotes the $M\times M$ identity matrix. Dividing this column by the entry corresponding to the reference microphone, subsequently gives an RTF vector estimate

$$\begin{aligned}\hat{\mathbf{h}} &= [\mathbf{I}_{M\times M}, \mathbf{0}_{M\times 1}]\,\frac{\bar{\mathbf{R}}_{\mathrm{y}}\mathbf{e}_{\mathrm{E}}}{\bar{\mathbf{e}}^{T}\bar{\mathbf{R}}_{\mathrm{y}}\mathbf{e}_{\mathrm{E}}} \\ &= \frac{\mathcal{E}\{\mathbf{x}X_{\mathrm{E}}^{*}\}}{\mathcal{E}\{X_{1}X_{\mathrm{E}}^{*}\}}\,.\end{aligned} \tag{4.6}$$

Estimates are denoted by $\hat{(\cdot)}$ throughout this thesis. In (4.6), the RTF vector is estimated using the first microphone as the reference, and thus, the extended selection vector $\bar{\mathbf{e}}$ having a 1 as its first entry, that is $\bar{\mathbf{e}} = [1, \mathbf{0}_{M\times 1}^{T}]^{T}$.

# 5 Proposed Optimization Problems

In the following, an extension of the RTF estimator described in Section 4 is considered. Instead of using only one external microphone, a second one is used, which consequently yields a second RTF vector estimate for the local array. The problem arising from the availability of two estimates is in deciding which one to use or how to best combine them. This is especially relevant if one estimate is much less accurate than the other.

In this section, two optimization problems are proposed which aim to optimally combine the two RTF vector estimates $\hat{\mathbf{h}}_1$ and $\hat{\mathbf{h}}_2$ by means of a linear combination, i.e.

$$\hat{\mathbf{h}} = \alpha \hat{\mathbf{h}}_1 + (1 - \alpha)\hat{\mathbf{h}}_2 \,, \tag{5.1}$$

with the weight $\alpha$. Equivalently, in terms of the matrix $\hat{\mathbf{H}} = [\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2]$ containing the RTF vector estimates and the vector $\boldsymbol{\alpha} = [\alpha, 1 - \alpha]^T = [\alpha_1, \alpha_2]^T$ containing the weights, (5.1) can be written as

$$\hat{\mathbf{h}} = \hat{\mathbf{H}}\boldsymbol{\alpha} \,, \tag{5.2}$$

where the sum of all elements in $\boldsymbol{\alpha}$ is constrained to equal 1, that is

$$\mathbf{1}_{2\times1}^T \boldsymbol{\alpha} = 1 \,. \tag{5.3}$$

The idea of combining several RTF vectors (or respectively steering vectors) is motivated by the work presented in [16], where a linear combination of steering vectors (in a Bayesian framework) for MVDR beamforming was proposed. In the following, for the mathematical representation, only the matrix notation in (5.2) is used. The constraint in (5.3) is in either optimization problem not directly taken into account, but rather applied subsequently as a normalization.

## 5.1 Orthogonal Projection

Commonly, a measure for the accuracy of an RTF vector estimate is the so-called Hermitian angle $\Theta$ [17], which is the angle between the true RTF vector and its estimate. It follows that in general a smaller Hermitian angle leads to a more accurate (and subsequently an objectively better) RTF vector estimate. As a consequence of this, one approach to obtain an optimal combination of two estimates is to find the vector that minimizes $\Theta$. However, the Hermitian angle is difficult to optimize, since it is given by

$$\Theta = \arccos\left(\frac{|\mathbf{h}_{\text{true}}^H \hat{\mathbf{h}}|}{\|\mathbf{h}_{\text{true}}\|_2 \|\hat{\mathbf{h}}\|_2}\right), \tag{5.4}$$

where $|\cdot|$ denotes the absolute value and $\|\cdot\|_2$ is the 2-norm of a vector, with $\|\cdot\|_2 = \sqrt{(\cdot)^H(\cdot)}$.

This problem can, in fact, be solved as a generalized Rayleigh quotient (see Appendix A). However, to obtain a compact solution, orthogonal projection is considered (equivalence of the solutions is shown in Appendix A): The normalized orthogonal projection of the true RTF vector onto the plane spanned by the two estimates (referred to as the estimation plane) yields the vector with the smallest angle to the projected true RTF vector. An example for the orthogonal projection in the real-valued 3D-space is depicted in Fig. 5.1. The notation $\mathbf{h}(n)$ refers to the $n$-th element of the vector $\mathbf{h}$. The true RTF vector (red) is projected on the blue plane spanned by $\hat{\mathbf{h}}_1$ and $\hat{\mathbf{h}}_2$ (dark blue). The projection is represented by the green line in the plane and the rejection (error vector) is the line orthogonal to the plane. When the normalization constraint is fulfilled, the projection ends on the solution subspace, i.e. the blue line which connects the two estimates.

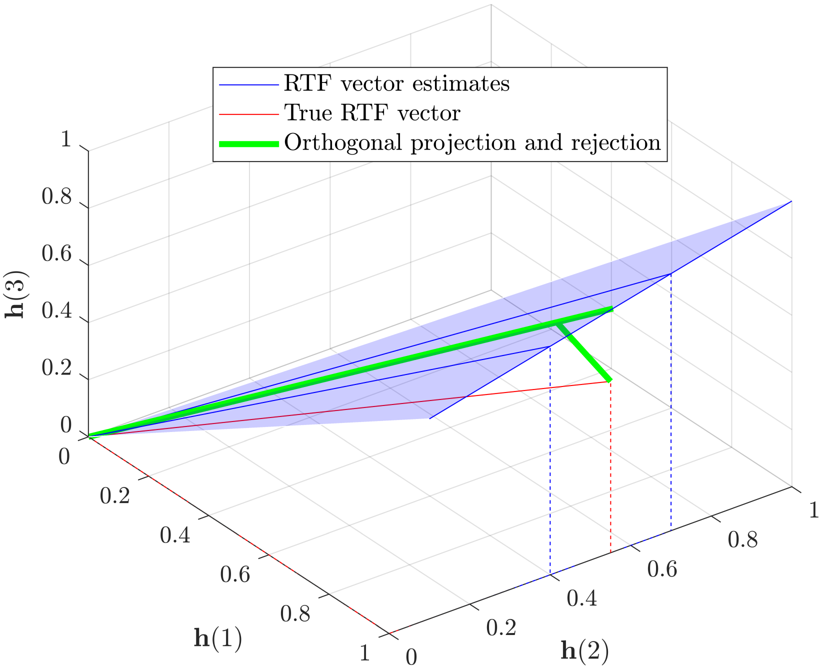

Fig. 5.1: Visualization of orthogonal projection (green) of $\mathbf{h}_{\text{true}}$ (red) on the plane spanned by the two RTF vector estimates (blue) in a real-valued 3D-space.

The weighting of the vectors (here $\hat{\mathbf{h}}_1$ and $\hat{\mathbf{h}}_2$) spanning the estimation plane for the orthogonal projection can be written as

$$\mathbf{v}_{\text{ortho}} = (\hat{\mathbf{H}}^H\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}^H\mathbf{h}_{\text{true}}\,, \tag{5.5}$$

according to [18]. The solution $\mathbf{v}_{\text{ortho}}$ represents the actual orthogonal projection. The resulting RTF vector, however, is in general not normalized to the entry corresponding to the reference microphone. Dividing $\mathbf{v}_{\text{ortho}}$ by the constraint in (5.3) then yields

$$\boldsymbol{\alpha}_{\text{ortho}} = \frac{\mathbf{v}_{\text{ortho}}}{\mathbf{1}_{2\times 1}^T\mathbf{v}_{\text{ortho}}} \tag{5.6}$$

as the normalized result.

## 5.2 Maximization of Output SNR

The maximization of the output SNR is motivated by one of the overall goals in speech enhancement: Reducing as much noise as possible while preserving the desired speech signal without distortion.

For the optimization, the solution of the MVDR beamformer in (3.10) is considered as the filter and substituted into the output SNR defined in (2.12), which yields

$$\mathit{SNR}_{\text{out}} = \frac{\mathbf{h}^H\mathbf{R}_{\text{n}}^{-1}\mathbf{R}_{\text{x}}\mathbf{R}_{\text{n}}^{-1}\mathbf{h}}{\mathbf{h}^H\mathbf{R}_{\text{n}}^{-1}\mathbf{h}}\,. \tag{5.7}$$

To this end, the property of $\mathbf{R}_{\text{n}}$ being Hermitian is used, i.e. $\mathbf{R}_{\text{n}} = \mathbf{R}_{\text{n}}^H$ and thus $(\mathbf{R}_{\text{n}}^{-1})^H = (\mathbf{R}_{\text{n}}^H)^{-1} = \mathbf{R}_{\text{n}}^{-1}$. Using the definition of the RTF vector estimate $\hat{\mathbf{h}}$ in (5.2), (5.7) can be written as

$$\mathit{SNR}_{\text{out}} = \frac{\boldsymbol{\alpha}^H\hat{\mathbf{H}}^H\mathbf{R}_{\text{n}}^{-1}\mathbf{R}_{\text{x}}\mathbf{R}_{\text{n}}^{-1}\hat{\mathbf{H}}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^H\hat{\mathbf{H}}^H\mathbf{R}_{\text{n}}^{-1}\hat{\mathbf{H}}\boldsymbol{\alpha}}\,. \tag{5.8}$$

With the condensed matrices

$$\mathbf{A} = \hat{\mathbf{H}}^H\mathbf{R}_{\text{n}}^{-1}\mathbf{R}_{\text{x}}\mathbf{R}_{\text{n}}^{-1}\hat{\mathbf{H}} \tag{5.9}$$

and

$$\mathbf{B} = \hat{\mathbf{H}}^H\mathbf{R}_{\text{n}}^{-1}\hat{\mathbf{H}}\,, \tag{5.10}$$

the output SNR can be written as

$$SNR_{\text{out}} = \frac{\boldsymbol{\alpha}^H \mathbf{A} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^H \mathbf{B} \boldsymbol{\alpha}}\,, \tag{5.11}$$

from which it becomes evident that the problem has the form of the generalized Rayleigh quotient of the matrices $\mathbf{A}$ and $\mathbf{B}$. The solution for stationary points of (5.11) is given by the generalized eigenvalue decomposition (GEVD) of $\mathbf{A}$ and $\mathbf{B}$, that is the EVD of $\mathbf{B}^{-1}\mathbf{A}$ [19]. A maximum then corresponds to the largest eigenvalue, thus, the solution for the weighting vector $\boldsymbol{\alpha}$ is given by the principle eigenvector $\mathbf{v}_{\text{max}}$ which corresponds to the principle eigenvalue, that is

$$\mathbf{v}_{\text{max}} = \mathcal{P}\{\mathbf{B}^{-1}\mathbf{A}\}, \tag{5.12}$$

where $\mathcal{P}\{\cdot\}$ is the principle eigenvector operator, which obtains the principle eigenvector of an EVD.

There are now three things to note about this solution: Firstly, it is not normalized, which means that the entry of the RTF vector in (5.2), corresponding to the reference microphone, is not necessarily equal to 1, such that the constraint in (5.3) is not fulfilled. Here, it is worth mentioning that for this specific constraint it does not matter if the optimization problem is solved as a constrained one or solved unconstrained (as done above) and then normalized afterwards. This can easily be argued, since the constraint would only be a division of numerator and denominator in the cost function with a constant, which would therefore cancel out, i.e. the problem is invariant to scaling. To ensure this normalization, the obtained solution is divided by the sum of its elements, i.e.

$$\boldsymbol{\alpha}_{\text{optSNR}} = \frac{\mathbf{v}_{\text{max}}}{\mathbf{1}_{2\times 1}^T \mathbf{v}_{\text{max}}}\,. \tag{5.13}$$

Secondly, the results are not constrained to be real-valued. Of course, the solutions for a combination of vector estimates are not explicitly required to be real-valued and can even be better if complex solutions are permitted, since complex-valued solutions introduce an additional degree of freedom. Nevertheless, if a real-valued solution is desired (which is not considered in this thesis), it must hold that $\boldsymbol{\alpha}^H = \boldsymbol{\alpha}^T$. Considering now the numerator and denominator of (5.11) separately, the numerator reads

$$\begin{aligned} \boldsymbol{\alpha}^H \mathbf{A} \boldsymbol{\alpha} &= \boldsymbol{\alpha}^T \mathbf{A} \boldsymbol{\alpha} \\ &= \boldsymbol{\alpha}^T \Re\{\mathbf{A}\} \boldsymbol{\alpha} + j \boldsymbol{\alpha}^T \Im\{\mathbf{A}\} \boldsymbol{\alpha}\,, \end{aligned} \tag{5.14}$$

and similarly for the denominator. $\Re\{\cdot\}$ and $\Im\{\cdot\}$ are the real- and imaginary-part operator, respectively. Since the matrices $\mathbf{A}$ and $\mathbf{B}$ are Hermitian, their real part is symmetric ($\Re\{\mathbf{A}\}^T = \Re\{\mathbf{A}\}$) and their imaginary part is anti-symmetric ($\Im\{\mathbf{A}\}^T = -\Im\{\mathbf{A}\}$). These properties are used in the following.
The expression in (5.14) is a scalar and must therefore be equal to its transpose. Applying the transpose operator, (5.14) yields

$$\begin{aligned}\boldsymbol{\alpha}^T\mathbf{A}\boldsymbol{\alpha} &= \left(\boldsymbol{\alpha}^T\mathbf{A}\boldsymbol{\alpha}\right)^T \\ &= \left(\boldsymbol{\alpha}^T\Re\{\mathbf{A}\}\boldsymbol{\alpha} + j\boldsymbol{\alpha}^T\Im\{\mathbf{A}\}\boldsymbol{\alpha}\right)^T \\ &= \left(\boldsymbol{\alpha}^T\Re\{\mathbf{A}\}\boldsymbol{\alpha}\right)^T + \left(j\boldsymbol{\alpha}^T\Im\{\mathbf{A}\}\boldsymbol{\alpha}\right)^T \\ &= \boldsymbol{\alpha}^T\Re\{\mathbf{A}\}^T\boldsymbol{\alpha} + j\boldsymbol{\alpha}^T\Im\{\mathbf{A}\}^T\boldsymbol{\alpha} \\ &= \boldsymbol{\alpha}^T\Re\{\mathbf{A}\}\boldsymbol{\alpha} - j\boldsymbol{\alpha}^T\Im\{\mathbf{A}\}\boldsymbol{\alpha}\,,\end{aligned} \tag{5.15}$$

which is only equal to (5.14) if $j\boldsymbol{\alpha}^T\Im\{\mathbf{A}\}\boldsymbol{\alpha} = -j\boldsymbol{\alpha}^T\Im\{\mathbf{A}\}\boldsymbol{\alpha} = 0$. Therefore, only the real parts of $\mathbf{A}$ and $\mathbf{B}$ must be considered for the real-valued solution of $\boldsymbol{\alpha}$, which is given by the principle eigenvector of the EVD of $\Re\{\mathbf{B}\}^{-1}\Re\{\mathbf{A}\}$.

Thirdly, the matrix $\mathbf{A}$ contains $\mathbf{R}_\mathrm{x}$. Therefore, information that is generally not given is still required to compute the solution. In addition to using the true $\mathbf{R}_\mathrm{x}$, the biased output SNR $\mathit{SNR}_\mathrm{out}^\mathrm{bias}$ can also be considered, i.e. $\mathbf{R}_\mathrm{y}$ is used instead of $\mathbf{R}_\mathrm{x}$. Maximizing $\mathit{SNR}_\mathrm{out}^\mathrm{bias}$ yields, in principle, the same result as the maximization of the output SNR, since

$$\begin{aligned}\mathit{SNR}_\mathrm{out}^\mathrm{bias} &= \frac{\mathbf{w}^H\mathbf{R}_\mathrm{y}\mathbf{w}}{\mathbf{w}^H\mathbf{R}_\mathrm{n}\mathbf{w}} \\ &= \frac{\mathbf{w}^H(\mathbf{R}_\mathrm{x} + \mathbf{R}_\mathrm{n})\mathbf{w}}{\mathbf{w}^H\mathbf{R}_\mathrm{n}\mathbf{w}} \\ &= \mathit{SNR}_\mathrm{out} + 1\ ,\end{aligned} \tag{5.16}$$

where the bias does not influence the stationary points, i.e. the optimum in the weighting vector $\boldsymbol{\alpha}$. This approach has also been proposed in [20].

# 6 Evaluation

In this section, the combinations of RTF vector estimates introduced in Section 5 are evaluated objectively in terms of noise reduction and speech distortion, using $\Delta SNR$ in (2.10), or here only in terms of $SNR_{\text{out}}$ in (2.12), and $SD$ in (2.13), respectively. Furthermore, the results for the true RTF vector $\mathbf{h}_{\text{true}}$ obtained from the RIR, each RTF vector estimate alone (i.e. the vectors obtained from $\text{E}_1$ and $\text{E}_2$ respectively), and the averaged RTF vector (i.e. both estimates are weighted with 0.5) are considered.

All in all, the RTF vector estimates, referred to as noted in Table 6.1, are evaluated. The respective weighting factors $\alpha_1$ and $\alpha_2$ corresponding to $\hat{\mathbf{h}}_1$ and $\hat{\mathbf{h}}_2$ (eventually stacked in the vector $\boldsymbol{\alpha}$), are stated additionally.

Table 6.1: Overview of used RTF estimates, including the corresponding values/equations for $\alpha_1$, $\alpha_2$, or $\boldsymbol{\alpha}$.

| **Notation** | **Method** | $\alpha_1$**,** $\alpha_2$**,** $\boldsymbol{\alpha}$ |
|---|---|---|
| $\mathbf{h}_{\text{true}}$ | True RTF vector obtained from the RIR | / |
| $\hat{\mathbf{h}}_1$ | RTF vector estimate obtained from $\text{E}_1$ | $\alpha_1 = 1$, $\alpha_2 = 0$ |
| $\hat{\mathbf{h}}_2$ | RTF vector estimate obtained from $\text{E}_2$ | $\alpha_1 = 0$, $\alpha_2 = 1$ |
| $\hat{\mathbf{h}}_{\text{AV}}$ | Averaged vector of $\hat{\mathbf{h}}_1$ and $\hat{\mathbf{h}}_2$ | $\alpha_1 = \alpha_2 = 0.5$ |
| $\hat{\mathbf{h}}_{\text{ortho}}$ | Orthogonal projection of $\mathbf{h}_{\text{true}}$ on estimation plane | From (5.6) |
| $\hat{\mathbf{h}}_{\text{optSNR}}$ | Maximized $SNR_{\text{out}}$ using true $\mathbf{R}_{\text{x}}$ in (5.7) | From (5.13) |
| $\hat{\mathbf{h}}_{\text{bias}}$ | Maximized biased $SNR_{\text{out}}$ using $\mathbf{R}_{\text{y}}$ in (5.16) | From (5.13) |

## 6.1 Set-up and Configurations

The recordings for the evaluation with real-world signals were done in the *Variable Acoustics Laboratory* at the Carl von Ossietzky University of Oldenburg at a reverberation time $T_{60}$ of about 300 ms. The noise signal was created by four loudspeakers facing the corners of the room and playing back different versions of multi-talker noise.

The used local microphone array L consisted of behind-the-ear hearing aid devices with two microphones (with a distance of 7 mm) at each side of the head mounted to a KEMAR (G.R.A.S.) dummy head. The external microphones (six in total) were placed in "look direction" of the dummy head with a distance $d$ ranging approximately from 0.7 m to 1.9 m. For the desired signal, a loudspeaker playing back a speech signal, uttered by an English speaking male, was placed about 15 cm away from the external microphone furthest away from the dummy head. In the

following, the first four channels (referred to as *Ch. 1-4*) are the microphones of the hearing aid devices, while *Ch. 5-10* are the external microphones, where *Ch. 5* is the microphone furthest away from the dummy head and each microphone up to *Ch. 10* is located closer to it, as depicted in Fig. 6.1.
Background noise and desired signal were recorded separately and mixed afterwards. The original recordings were sampled at 48 kHz, however, for the processing the signals were down-sampled to 16 kHz.

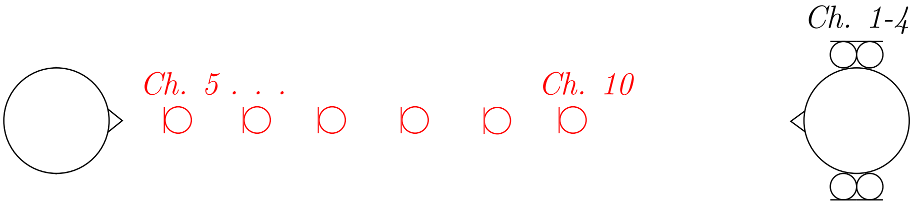

Fig. 6.1: Schematic set-up with dummy head (equipped with a binaural hearing aid device) and six available external microphones (indicated in red).

In the following, the different proposed RTF estimation methods are evaluated with regard to different aspects: The first experiment (Subsection 6.3) deals with the influence of the distance between the local array and the external microphones. To this end, the input SNRs in the two external microphones which were used, were leveled beforehand, adjusting the speech power and keeping the noise signal constant. In the second experiment (Subsection 6.4), the influence of the input SNR is investigated. There, the distance between the external microphones and the dummy head was held constant and the speech signals were scaled, so that different input SNRs were obtained in the two external microphones. In a last experiment (Subsection 6.5), a realistic scenario without artificial leveling is considered.

Please note that the artificial levelling only affects the external microphones and not the local microphone array. In all experiments, the input SNR in the reference microphone, here the first on the left side of the head (*Ch. 1*), was set to 0 dB and the other signals of *Ch. 2-4* were scaled accordingly. From this it follows that the true RTF vector is the same for *all* considered scenarios. The corresponding scores are therefore also constant, but they are still given as a reference value for *"ideal"* performance in all representations.

## 6.2 Implementation and Parameters

As the parameters of the weighted overlap add (WOLA) framework an FFT size $L_{\mathrm{FFT}}$ of 1024 samples per frame and an overlap of 50 %, that is a hop size $L_{\mathrm{R}}$ of

512 samples, were chosen. As the window function for both, analysis and synthesis, a square-root-Hann window was used.

The whole signal was analyzed *"batch"*, i.e. with oracle knowledge about speech and noise separately and with the complete signal being available prior to the processing. In practice, this means that the (extended) covariance matrices are calculated as the mean over all frames, from which it follows that there is only one RTF vector estimate (per method) over the whole signal and subsequently also only one filter vector. Speech presence was determined via an ideal VAD per frame. The noisy and speech covariance matrices, $\bar{\bar{\mathbf{R}}}_\mathrm{y}$ and $\bar{\bar{\mathbf{R}}}_\mathrm{x}$ (containing $\mathbf{R}_\mathrm{y}$ and $\mathbf{R}_\mathrm{x}$, respectively), were calculated in frames where speech was present, while the noise covariance matrix, $\bar{\bar{\mathbf{R}}}_\mathrm{n}$ (containing $\mathbf{R}_\mathrm{n}$), was calculated over all frames.

## 6.3 Experiment 1 - Distance Dependence

In this experiment, the input SNR in $\mathrm{E}_1$ and $\mathrm{E}_2$ was held fixed at 0 dB. To examine the influence of the distance of the external microphones to the local array, one microphone ($\mathrm{E}_1$) was set to a fixed position, once *Ch. 5* and once *Ch. 10* are used as $\mathrm{E}_1$, while $\mathrm{E}_2$ is varied over all other available channels.

In the following the results are presented. In Fig. 6.2 - 6.4, the results for for *Ch. 5* being set as $\mathrm{E}_1$ are shown, while the results for *Ch. 10* being set as $\mathrm{E}_1$ are shown in Fig. 6.5 - 6.7. For both conditions, the three graphs show the real part of the weighting factor $\alpha$ (i.e. the first entry of the vector $\boldsymbol{\alpha}$, favoring $\mathrm{E}_1$ if equal to 1 and favoring $\mathrm{E}_2$ if equal to 0), the *SD*, as well as the $\mathit{SNR}_\mathrm{out}$ scores. All scores are either averaged over frequency ($\alpha$ and *SD*) or computed via the shadow-filtered time signal of speech and noise signal ($\mathit{SNR}_\mathrm{out}$), where shadow-filtering means applying the filter vector to each signal component separately. For the weighting factor $\alpha$, additionally the standard deviation is represented by error bars. The imaginary part of $\alpha$ is not taken into consideration, since it weights both, $\mathrm{E}_1$ and $\mathrm{E}_2$ with the same absolute value but opposite sign and therefore only affects the phase.
For clarity, the performance in terms of *SD* is considered better, when lower scores are obtained, while higher scores in $\mathit{SNR}_\mathrm{out}$ are desired.

Fig. 6.2 depicts the averaged real part of the weighting factor $\alpha$ for the case that *Ch. 5* is fixed as $\mathrm{E}_1$ while $\mathrm{E}_2$ is varied (*Ch. 6 - Ch. 10*). It can be seen that for all three methods ($\hat{\mathbf{h}}_\mathrm{ortho}$, $\hat{\mathbf{h}}_\mathrm{optSNR}$, and $\hat{\mathbf{h}}_\mathrm{bias}$), an almost constant value of $\Re\{\alpha\}$ is obtained, which lies around 0.5. This result can be explained, when considering that for all available channels (with a minimal distance of 70 cm to the local array) the spacial coherence assumption from Section 4 is mostly fulfilled. The standard deviation, however, seems to decrease on average (especially for the orthogonal projection) over an increasing distance between $E_1$ and $E_2$.

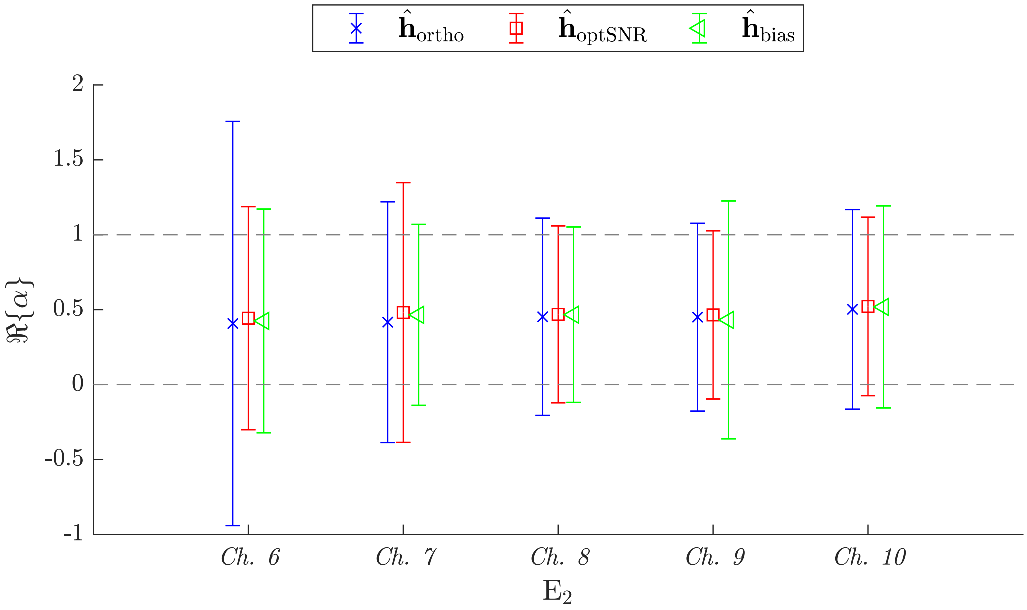

Fig. 6.2: Average of real part of the weighting factor $\alpha$ (i.e. weighting $E_1$) obtained by $\hat{\mathbf{h}}_{\text{ortho}}$, $\hat{\mathbf{h}}_{\text{optSNR}}$ and $\hat{\mathbf{h}}_{\text{bias}}$ over different $E_2$ with *Ch. 5* set as $E_1$, including the standard deviation as error bars and orientation lines at $\Re\{\alpha\}$=0 and $\Re\{\alpha\}$=1.

The weighting factor from Fig. 6.2 alone does not fully describe the performance of all channels, even though they are weighted almost equally for all channels for $E_2$. For further analysis of performance, *SD* and $\mathit{SNR}_{\text{out}}$ are considered (Fig. 6.3 and Fig. 6.4). As references, the constant performance of $E_1$ and the true RTF vector $\mathbf{h}_{\text{true}}$, respectively, is visualized as lines.
In Fig. 6.3, it can be seen that the true RTF vector yields by far the lowest *SD* scores (approximately 0.5 dB). $E_1$ alone shows in general the highest speech distortion. This result seems unexpected, since the microphone furthest away from L is expected to yield the best results, because a larger distance assures more safely that the SC assumption is fulfilled. However, even this "high" *SD* score (about 2.7 dB) is barely noticeable in informal listening tests.
The weighting approaches ($\hat{\mathbf{h}}_{\text{AV}}$, $\hat{\mathbf{h}}_{\text{ortho}}$, $\hat{\mathbf{h}}_{\text{optSNR}}$, $\hat{\mathbf{h}}_{\text{bias}}$) all yield lower *SD* scores than either external microphone alone (between 1.5 and 2 dB). The results show a rather constant performance, even with varying performance of $E_2$. The biased optimization of the output SNR has slightly higher *SD* scores than the orthogonal projection and the oracle optimization of the output SNR.

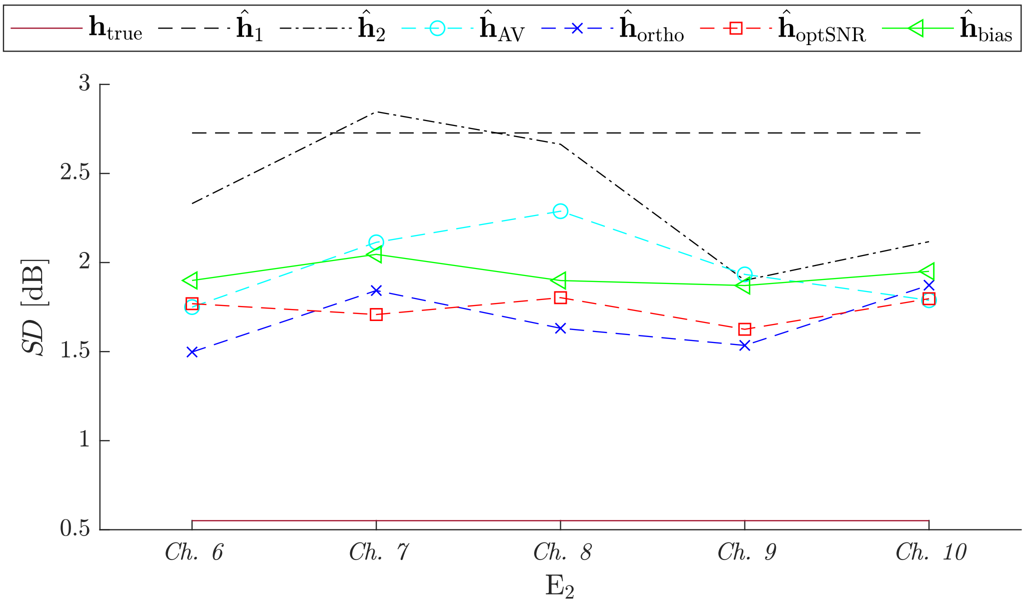

Fig. 6.3: *SD* in dB for all RTF vector estimates in Table 6.1 over different $\text{E}_2$ with *Ch. 5* set as $\text{E}_1$.

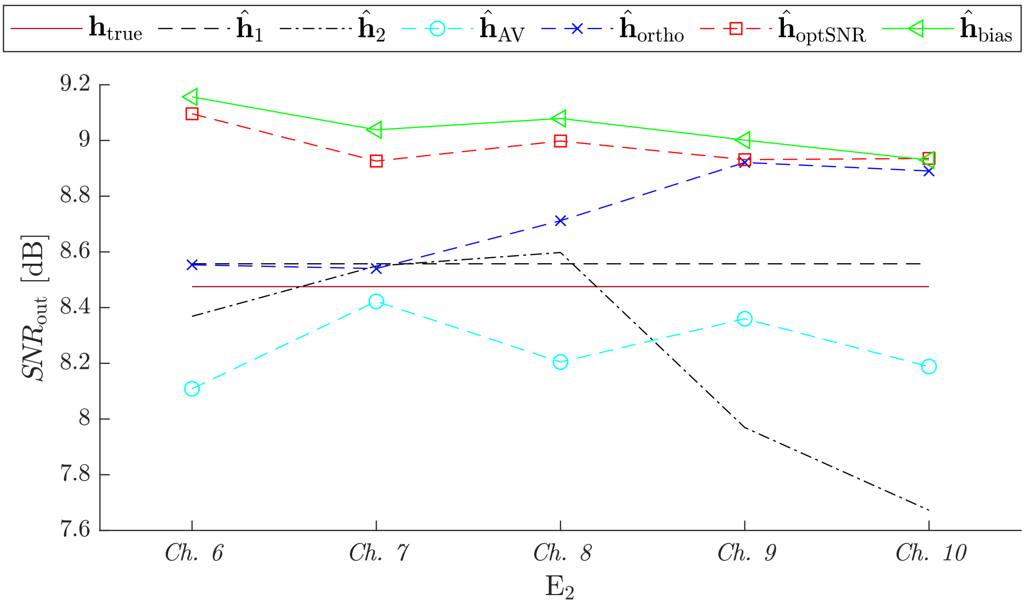

Fig. 6.4: $SNR_{\text{out}}$ in dB for all RTF vector estimates in Table 6.1 over different $\text{E}_2$ with *Ch. 5* set as $\text{E}_1$.

Considering $SNR_{\text{out}}$ in Fig. 6.4, it can clearly be seen that especially the SNR-based optimal solutions ($\hat{\mathbf{h}}_{\text{optSNR}}$ and $\hat{\mathbf{h}}_{\text{bias}}$) exhibit a good performance. Both yield scores of around 9 - 9.2 dB which is about 0.5 dB better than the best external microphone ($\text{E}_1$) alone. The orthogonal projection $\hat{\mathbf{h}}_{\text{ortho}}$ does not perform as well as the two other optimal solutions. When *Ch.6* - *Ch. 8* are used as $\text{E}_2$, the $SNR_{\text{out}}$ scores of $\hat{\mathbf{h}}_{\text{ortho}}$ are in the range of $\text{E}_1$ alone or just slightly better. For *Ch. 9* and *Ch. 10*, the performance of $\hat{\mathbf{h}}_{\text{ortho}}$ gets similar to those of $\hat{\mathbf{h}}_{\text{optSNR}}$ and $\hat{\mathbf{h}}_{\text{bias}}$. The averaging approach $\hat{\mathbf{h}}_{\text{AV}}$ always performs worse than the best external microphone alone and sometimes even worse than both (when the performances of $\text{E}_1$ and $\text{E}_2$ are similar). This is a first indicator that the averaging approach is not suitable and could be dismissed due to its poor performance.

In the following, a second scenario is considered: $\text{E}_1$ is set to *Ch. 10* and *Ch. 5* - *Ch. 9* are varied as $\text{E}_2$. The aim of this second step is to observe how the optimal solutions behave if one of the "worst performing" channels is combined with all others ("worst" in terms of $SNR_{\text{out}}$ as shown in Fig. 6.4).
In Fig. 6.5, the real part of $\alpha$ is depicted over the channels for $\text{E}_2$. Similar to the results in Fig. 6.2, the mean value for all weighting methods lies around 0.5 for all considered configurations. Another similarity is that the standard deviation is the highest if the channel next to the fixed one (here *Ch. 9*) is used. This implies that the weighting is less distinct and that large weighting factors are required to obtain for example phase corrections.

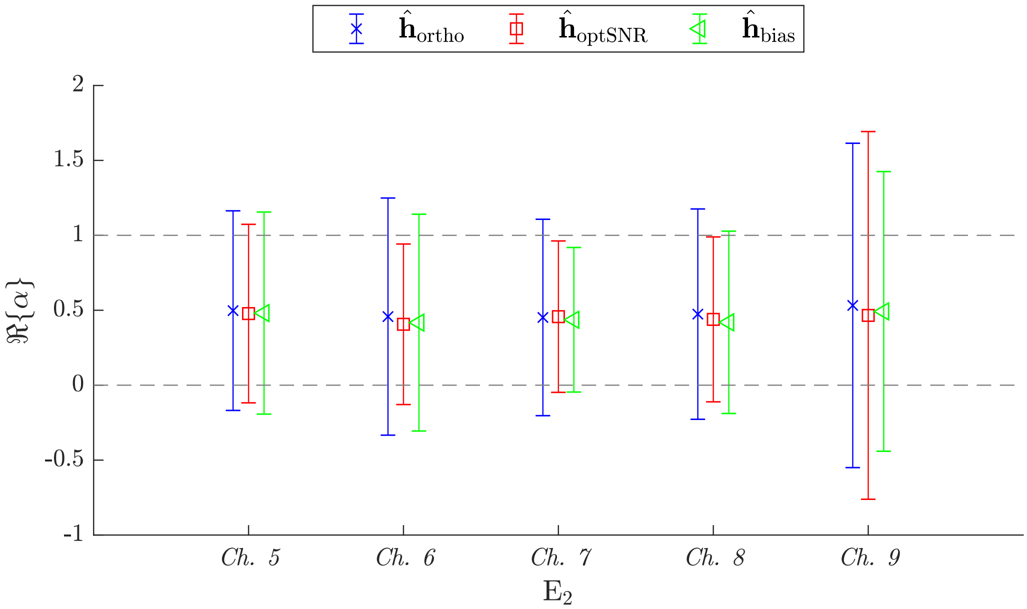

Fig. 6.5: Average of real part of the weighting factor $\alpha$ (i.e. weighting $\text{E}_1$) obtained by $\hat{\mathbf{h}}_{\text{ortho}}$, $\hat{\mathbf{h}}_{\text{optSNR}}$ and $\hat{\mathbf{h}}_{\text{bias}}$ over different $\text{E}_2$ with *Ch. 10* set as $\text{E}_1$, including the standard deviation as error bars and orientation lines at $\Re\{\alpha\}$=0 and $\Re\{\alpha\}$=1.

The resulting *SD* scores are depicted in Fig. 6.6. It can be seen that for most configurations the score is lower than both external microphones alone for all combination approaches (except for the averaging approach that once has a slightly higher score than $\text{E}_1$). Again, the results lie in a range of 1.5 - 2 dB.

The *SNR*$_{\text{out}}$ scores shown in Fig. 6.7 exhibit that, again, the averaging approach yields the worst performance, i.e. it always yields lower output SNRs than the better external microphone alone. The three other weighting approaches show much better results, even though for one configuration $\hat{\mathbf{h}}_{\text{optSNR}}$ and $\hat{\mathbf{h}}_{\text{bias}}$ perform slightly worse than the better external microphone. The orthogonal projection shows a similar performance.

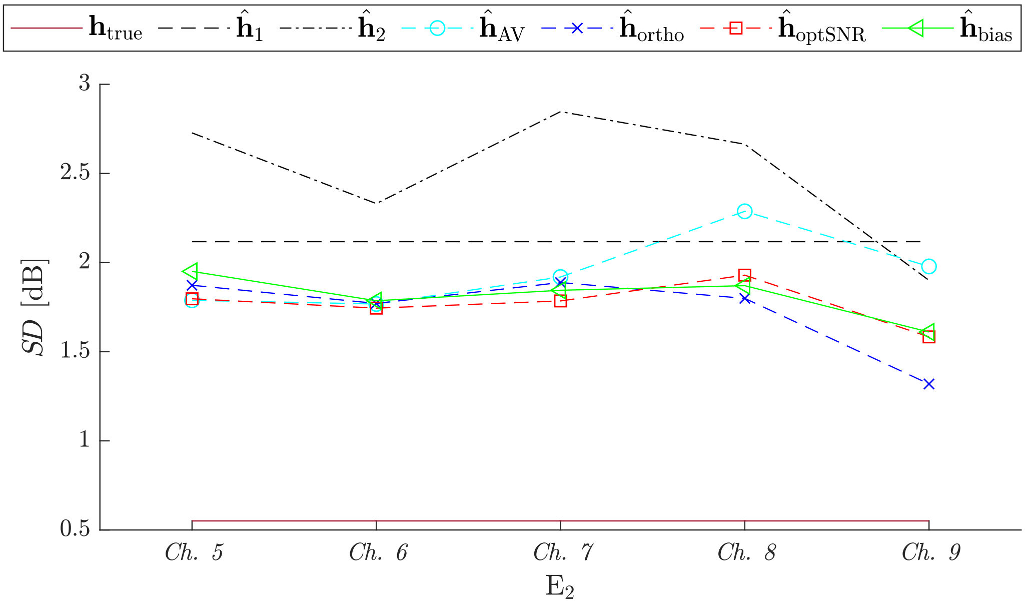

Fig. 6.6: *SD* in dB for all RTF vector estimates in Table 6.1 over different $\text{E}_2$ with *Ch. 10* set as $\text{E}_1$.

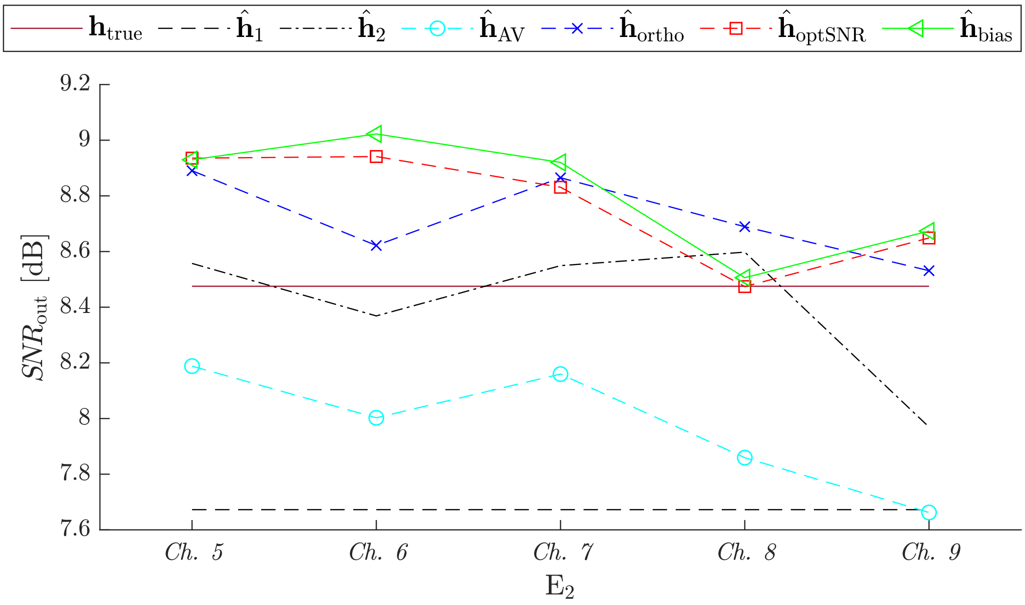

Fig. 6.7: *SNR*$_{\text{out}}$ in dB for all RTF vector estimates in Table 6.1 over different $\text{E}_2$ with *Ch. 10* set as $\text{E}_1$.

From this experiment, it can be concluded that there are distance dependencies for each external microphone alone, but that the optimized solution is rarely affected by this as long as there is one good estimate available. When both estimates are obtained by microphones that are close to L (i.e. the case where *Ch. 9* and *Ch. 10* are used), the results obtained by all weighting approaches get slightly worse. In general, it can be said that as long as the SC assumption is fulfilled for at least one microphone (here this was the case for all available external microphones), all weighting approaches, except for the simple averaging ($\hat{\mathbf{h}}_{\text{AV}}$), yield better results than the best external microphone alone. Furthermore, $\hat{\mathbf{h}}_{\text{optSNR}}$ and $\hat{\mathbf{h}}_{\text{bias}}$ always perform very similarly, which supports the theoretical finding that these two solutions are equivalent.

## 6.4 Experiment 2 - Input SNR Dependence

In this experiment, the influence of the input SNR in the external microphones is investigated. To this end, *Ch. 7* and *Ch. 8* are chosen as $\text{E}_1$ and $\text{E}_2$ respectively, since they performed most uniformly in terms of *SD* and $\mathit{SNR}_{\text{out}}$ at an identical input SNR of 0 dB, as shown in Subsection 6.3. A distance dependence of the results can therefore be neglected in the following.
The input SNR in $\text{E}_1$ is set to 0 dB, while it is varied from -30 to +30 dB in steps of 3 dB in $\text{E}_2$.

In Fig. 6.8, the dependence of the real part of the weighting factor $\alpha$ and its standard deviation are depicted. As expected, for lower input SNRs in $\text{E}_2$ than in $\text{E}_1$, latter is weighted stronger, i.e. $\Re\{\alpha\}$ goes to 1. The same behaviors can be observed when the input SNR in $\text{E}_2$ increases. At approximately 15 dB input SNR in $\text{E}_2$, the weighting favors this microphone almost completely, i.e. $\Re\{\alpha\}$ goes to 0. In the center of the graph, where the input SNRs in both channels only differ by some dB, the weighting is almost equal, i.e. in the range of 0.5. Interestingly, for input SNRs of around -5 dB, the standard deviation for the output SNR-based approaches becomes fairly large while strongly decreasing for high input SNRs in $\text{E}_2$. The standard deviation for the orthogonal projection exhibits an opposite behavior: It increases with increasing input SNR in $\text{E}_2$. An explanation for these observations can not easily be given.

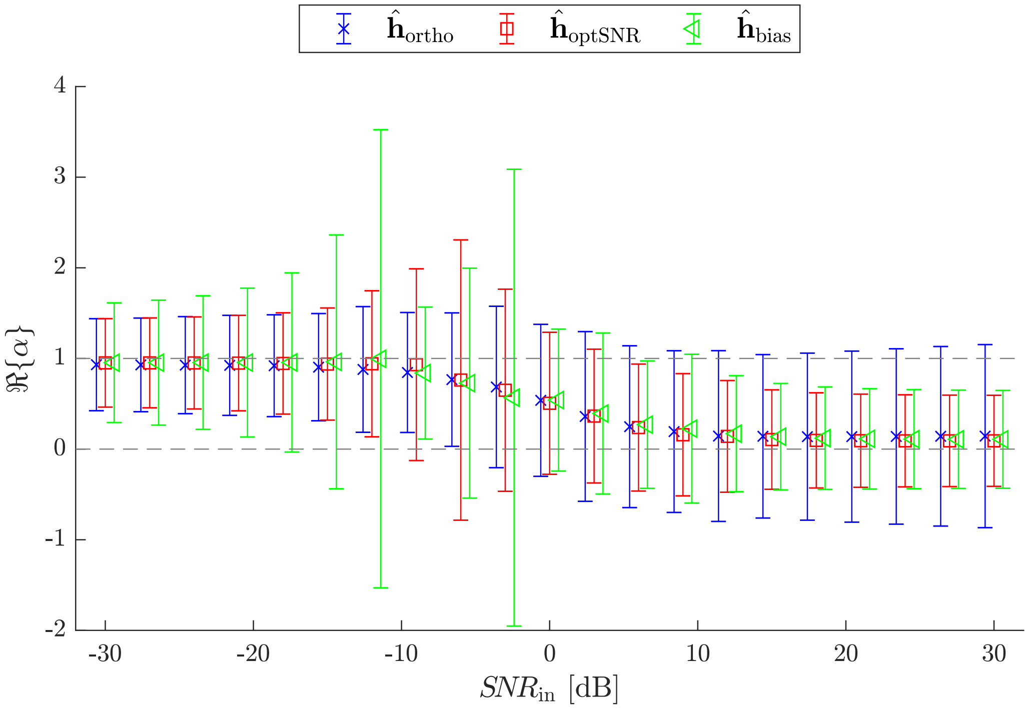

Fig. 6.8: Average of real part of the weighting factor $\alpha$ (i.e. weighting $E_1$) obtained by $\hat{\mathbf{h}}_{\text{ortho}}$, $\hat{\mathbf{h}}_{\text{optSNR}}$ and $\hat{\mathbf{h}}_{\text{bias}}$ over different input SNRs in $E_2$ with a fixed input SNR of 0 dB in $E_1$, including the standard deviation as error bars and orientation lines at $\Re\{\alpha\}$=0 and $\Re\{\alpha\}$=1.

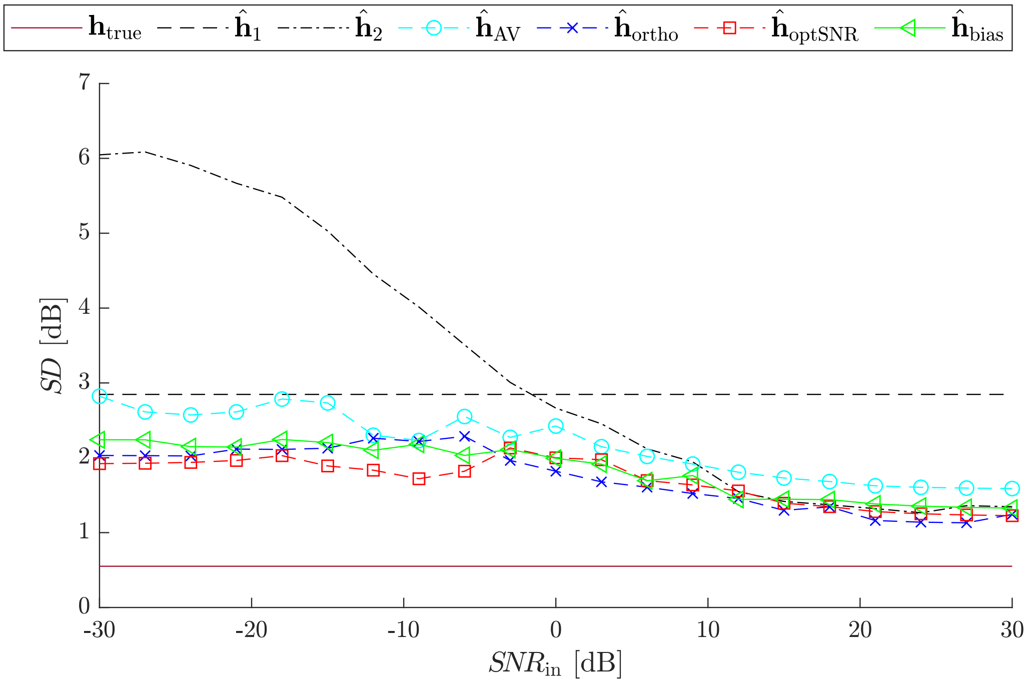

Fig. 6.9: $SD$ in dB for all RTF vector estimates in Table 6.1 over different input SNRs in $E_2$ with a fixed input SNR of 0 dB in $E_1$.

The *SD* scores are shown in Fig. 6.9. It is clearly visible that at low input SNRs the performance of $E_2$ strongly diminishes ($SD \approx 6$ dB at $SNR_{\text{in}} =$ -30 dB). The *SD* scores then improve to around 3 dB (i.e. the same as for $E_1$) at about 0 dB input SNR, until $E_2$ outperforms $E_1$ at high input SNRs. The optimized solutions ($\hat{\mathbf{h}}_{\text{ortho}}$, $\hat{\mathbf{h}}_{\text{optSNR}}$ and $\hat{\mathbf{h}}_{\text{bias}}$) always perform better than (or the same as) the best external microphone alone, although at high input SNRs, the gain is negligibly small. In that region, the averaging approach performs slightly worse than the best microphone alone.

The $SNR_{\text{out}}$ scores in Fig. 6.10 show similar tendencies as for *SD* above. $E_2$ performs much worse than $E_1$ at low input SNRs, but does not perform much better at high ones. The averaging approach always yields lower scores than $E_1$ alone and is therefore fairly sub-optimal. All approaches based on optimization yield comparable results that are slightly better than the best external microphone alone. However, the gain only measures around 0.5 dB and is thus not significant.

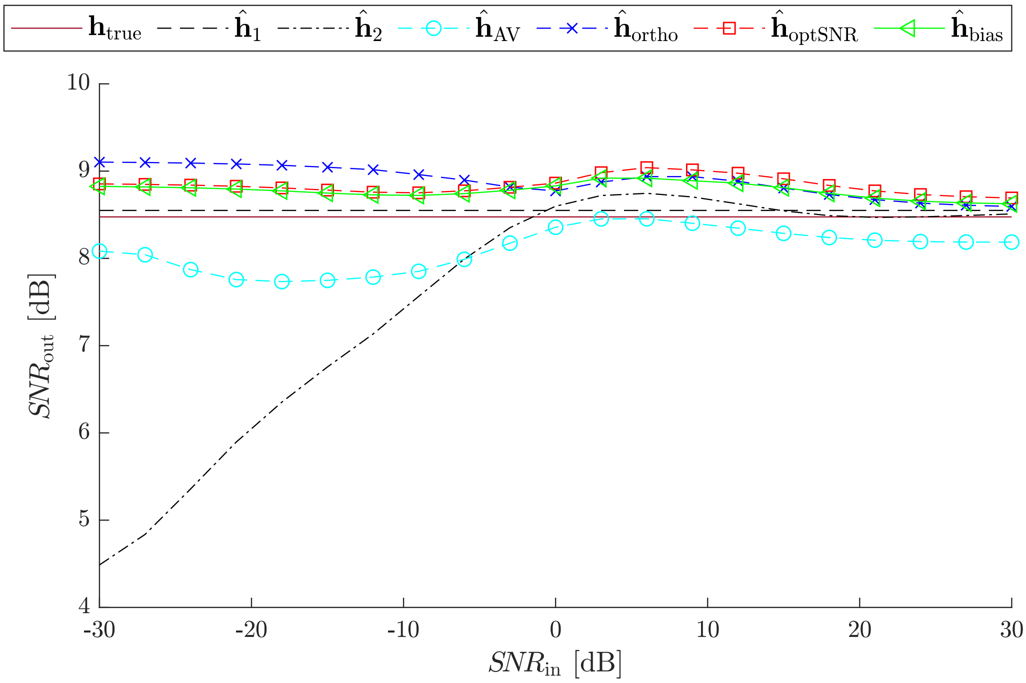

Fig. 6.10: $SNR_{\text{out}}$ in dB for all RTF vector estimates in Table 6.1 over different input SNRs in $E_2$ with a fixed input SNR of 0 dB in $E_1$.

From this experiment, it can be concluded that the input SNR in the external microphones has a much stronger influence on the weighting than the distance of the external microphones to the local array. Furthermore, it can be said that all three optimization problems ($\hat{\mathbf{h}}_{\text{ortho}}$, $\hat{\mathbf{h}}_{\text{optSNR}}$ and $\hat{\mathbf{h}}_{\text{bias}}$) yield reliable results if one good estimate is available, which is highly interesting in terms of applicability in practice, since the biased output SNR optimization is in principle also possible in blind implementations as long as a VAD is available. Moreover, in this experiment it was shown that the averaging approach, again, does not yield good results, especially if one estimate is significantly worse than the other.

## 6.5 Experiment 3 - Realistic Data

The last experiment, presented in this subsection, is performed similarly to the experiment in Subsection 6.3, but without artificial levelling, i.e. realistic signals are used. Again, $\mathrm{E}_1$ is set to *Ch. 10* and $\mathrm{E}_2$ is varied over all other available channels. In addition to the plots of $\Re\{\alpha\}$, *SD*, and $\mathit{SNR}_{\mathrm{out}}$, exemplary plots of the cost functions for the Hermitian angle and the output SNR are shown to visualize the respective positions of the different optimal solutions.

In Table 6.2, the input SNR and distance to L as well as the *SD* and $\mathit{SNR}_{\mathrm{out}}$ scores are given for all external microphones.
Intuitively, it is expected that $\mathrm{E}_2$ is always favored over $\mathrm{E}_1$ in terms of weighting, since the in- and output values favor *Ch. 5 - 9* over *Ch. 10*. Furthermore, the combination of vector estimates is expected to perform better when one external microphone is especially good, i.e. *Ch. 5* is used as $\mathrm{E}_2$.

Table 6.2: Specifications of input parameters and performance for each available external microphone.

| *Ch.* No. | $d$ / cm | $\mathit{SNR}_{\mathrm{in}}$ / dB | *SD* / dB | $\mathit{SNR}_{\mathrm{out}}$ / dB |
|---|---|---|---|---|
| *5* | 190.6 | 18.3 | 1.2 | 8.9 |
| *6* | 166.5 | 10.4 | 1.4 | 8.9 |
| *7* | 142.4 | 7.3 | 2.0 | 8.6 |
| *8* | 118.3 | 4.6 | 2.2 | 8.7 |
| *9* | 94.2 | 3.4 | 1.6 | 8.2 |
| *10* | 70.1 | 2.6 | 1.9 | 7.9 |

The expected behavior of $\Re\{\alpha\}$ can be seen in Fig. 6.11. For channels which yield much better results than *Ch. 10*, i.e. when *Ch. 5* is used as $\mathrm{E}_2$, the weighting factor takes values of around 0.15, indicating a clear weighting towards $\mathrm{E}_2$. For channels with similar performances to *Ch. 10*, especially *Ch. 9*, the weighting approaches 0.5, which means that the channels are weighted equally.

The results for *SD* and $\mathit{SNR}_{\mathrm{out}}$ are depicted in Fig. 6.12 and 6.13. As expected, the combinations for *Ch. 5 - 7* yield the best results in terms of noise reduction and speech distortion. In terms of $SNR_{\text{out}}$, the averaging approach yields, as in experiments 1 and 2, results below the better external microphone alone. Surprisingly, this is also the case for the output SNR-based optimal weighting factors ($\hat{\mathbf{h}}_{\text{optSNR}}$ and $\hat{\mathbf{h}}_{\text{bias}}$) for the configuration where *Ch. 8* is used as $\mathrm{E}_2$. However, this is the only outlier throughout all conducted experiments. The *SD* scores show that all combinations yield comparable or even better results than the best external microphone alone (mostly in a range of 1 to 1.5 dB). Among the combinations, the orthogonal projection performs best in terms of *SD*, while the output SNR-based solutions perform best in terms of noise reduction (despite the single outlier).

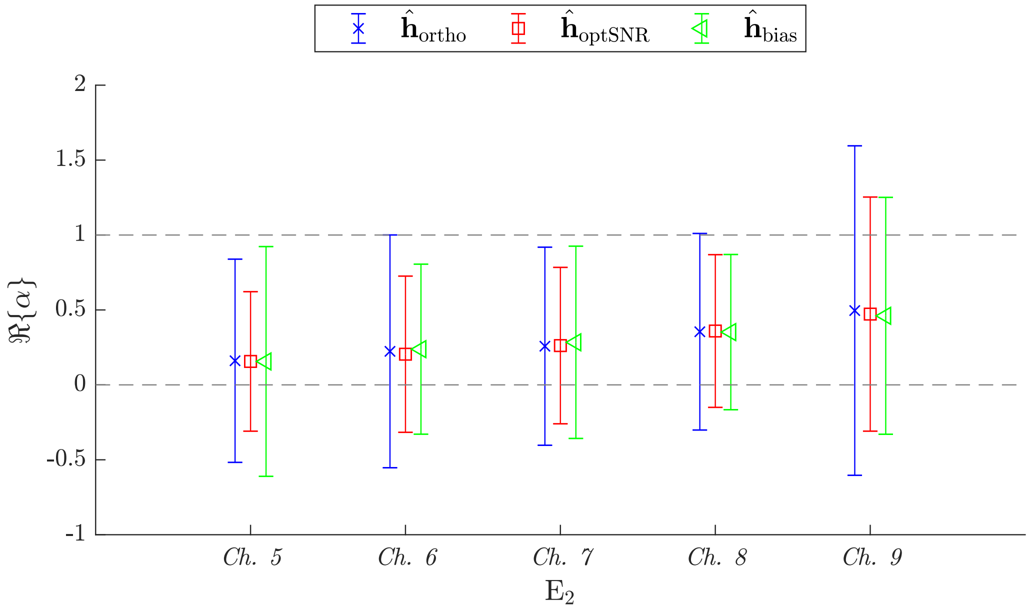

Fig. 6.11: Average of real part of the weighting factor $\alpha$ (i.e. weighting $\mathrm{E}_1$) obtained by $\hat{\mathbf{h}}_{\text{ortho}}$, $\hat{\mathbf{h}}_{\text{optSNR}}$ and $\hat{\mathbf{h}}_{\text{bias}}$ over different $\mathrm{E}_2$ with *Ch. 10* set as $\mathrm{E}_1$, including the standard deviation as error bars and orientation lines at $\Re\{\alpha\}$=0 and $\Re\{\alpha\}$=1.

In the following, exemplary cost functions for the output SNR and the Hermitian angle are depicted (Fig. 6.14 and 6.15) at a frequency of 500 Hz, for the configuration where $\mathrm{E}_1$ is set as *Ch. 10* and $\mathrm{E}_2$ is set as *Ch. 5*. These curves are meant to visualize the positions of the respective optimal solutions of the two cost functions which are to be optimized. It can clearly be seen that for $\hat{\mathbf{h}}_{\text{optSNR}}$ and $\hat{\mathbf{h}}_{\text{bias}}$ the solutions for $\alpha$ lie very close to each other and are both located at the maximum of the cost function of the $SNR_{\text{out}}$, while the weighting factor for the orthogonal projection differs from these solutions and does not lie directly at the maximum of this curve. In Fig. 6.15, it also becomes clear that the minimum of the Hermitian angle is not identical to the maximum of the output SNR and that the orthogonal projection, indeed, yields this minimum.

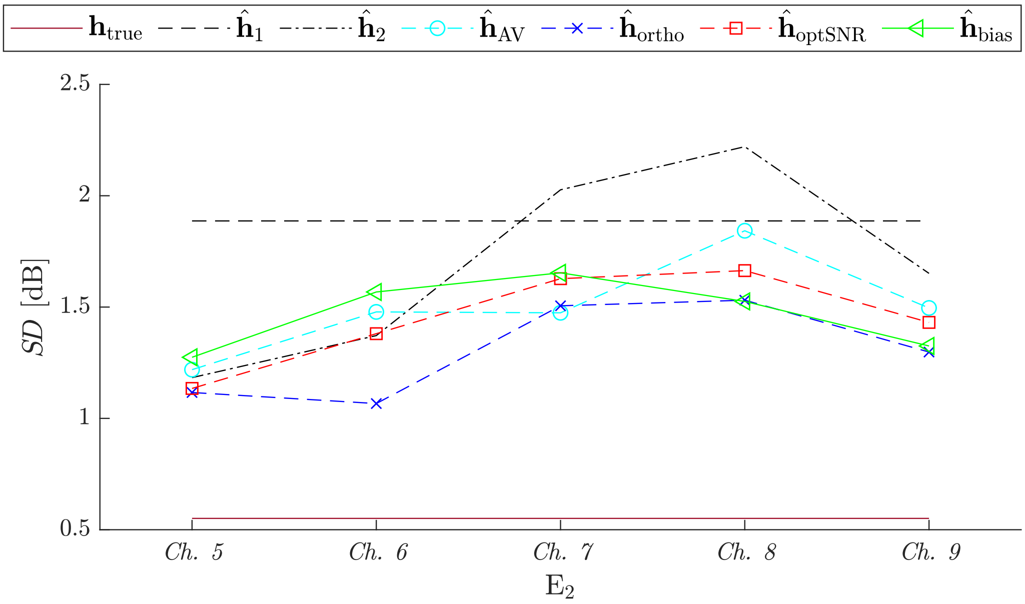

Fig. 6.12: *SD* in dB for all RTF vector estimates in Table 6.1 over different $\text{E}_2$ with *Ch. 10* set as $\text{E}_1$.

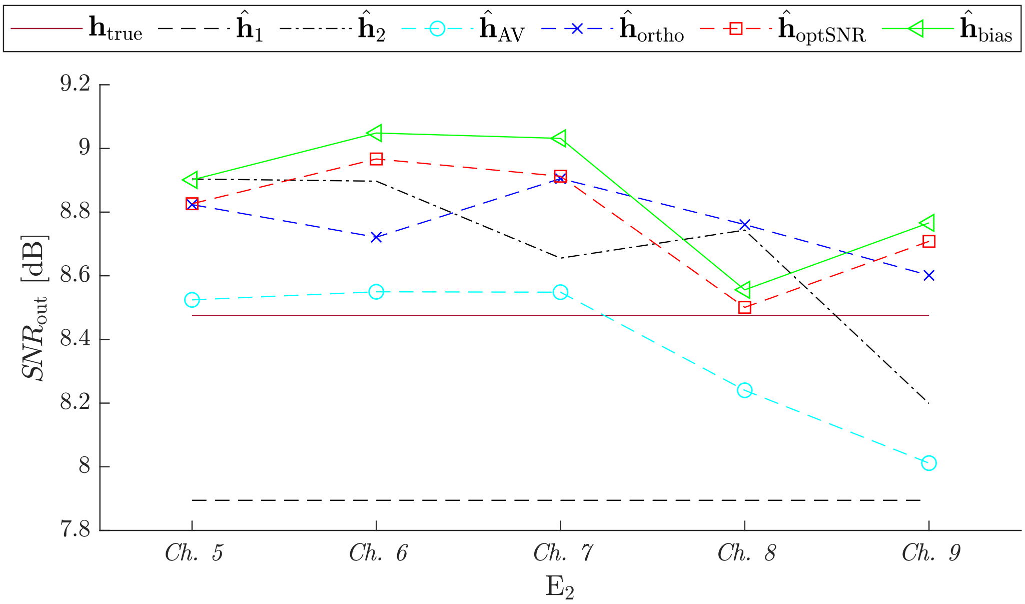

Fig. 6.13: $SNR_{\text{out}}$ in dB for all RTF vector estimates in Table 6.1 over different $\text{E}_2$ with *Ch. 10* set as $\text{E}_1$.

It is to be noticed that in this specific scenario, the weighting does not exhibit the expected behavior: Against the expectations, $\text{E}_1$ is weighted stronger than $\text{E}_2$, even though *Ch. 5* is used for the second external microphone and is therefore expected to perform better. These curves, however, are merely “snapshots” and are not representative for all frequencies.

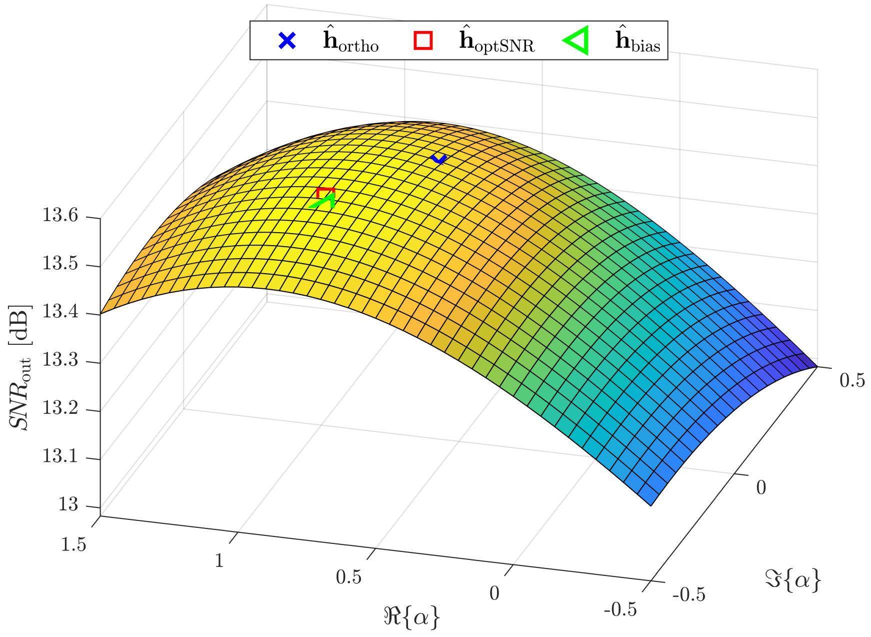

Fig. 6.14: Exemplary cost function of $SNR_{\mathrm{out}}$ at 500 Hz including the weighting factors for $\hat{\mathbf{h}}_{\mathrm{ortho}}$, $\hat{\mathbf{h}}_{\mathrm{optSNR}}$ and $\hat{\mathbf{h}}_{\mathrm{bias}}$, with *Ch. 10* as $\mathrm{E}_1$ and *Ch. 5* as $\mathrm{E}_2$.

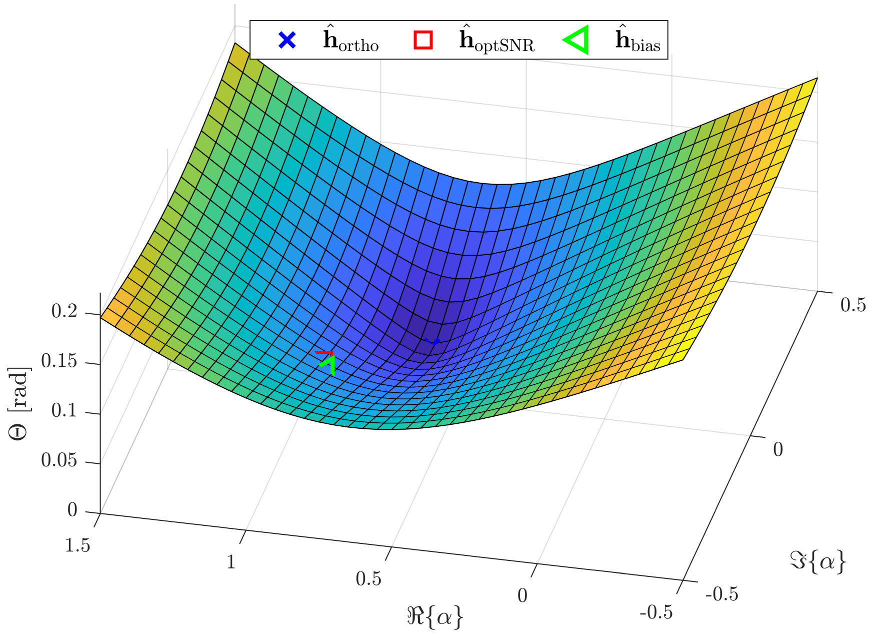

Fig. 6.15: Exemplary cost function of the Hermitian angle at 500 Hz including the weighting factors for $\hat{\mathbf{h}}_{\mathrm{ortho}}$, $\hat{\mathbf{h}}_{\mathrm{optSNR}}$ and $\hat{\mathbf{h}}_{\mathrm{bias}}$, with *Ch. 10* as $\mathrm{E}_1$ and *Ch. 5* as $\mathrm{E}_2$.

Finally, the conclusion can be drawn that the input SNR has a stronger influence on the weighting of two RTF vector estimates obtained by two external microphones using the SC method. The obtained results from $\hat{\mathbf{h}}_{\text{ortho}}$, $\hat{\mathbf{h}}_{\text{optSNR}}$ and $\hat{\mathbf{h}}_{\text{bias}}$ all outperform the simple averaging approach in almost all considered scenarios. These three optimal solutions all yield comparable results. Especially $\hat{\mathbf{h}}_{\text{optSNR}}$ and $\hat{\mathbf{h}}_{\text{bias}}$ exhibit quite similar performances which implies that their equivalence found in theory also applies in practice. This finding makes the biased optimization promising, since it can also be used in blind scenarios (i.e. without knowledge about $\mathbf{R}_{\text{x}}$). It could also be shown in the conducted experiments that the orthogonal projection and the output SNR-based approaches do not yield the same, but fairly similar results, even though they are optimizing different cost functions. Lastly, it can be said that all optimal solutions perform better than (or at least as well as) the best available external microphone in all considered configurations. Therefore, the proposed solutions are more robust than only using one (possibly sub-optimal) external microphone.

# 7 Conclusion and Outlook

To summarize the work done in this thesis, an ASN with a fusion center, i.e. a local microphone array, and two external nodes, each equipped with one microphone, is considered. Assuming uncorrelated noise, e.g. a diffuse noise filed, and a single desired source, the SC method yields two RTF vector estimates for the local array. To combine these two estimates, two approaches were proposed, namely the orthogonal projection of the true RTF vector on the plane spanned by the two estimates, which is a purely theoretical approach, and maximization of the output SNR. For the latter, it has been shown that either the output SNR or the biased output SNR can be maximized, since both yield, in theory, the same maximum. This theoretical equivalence is supported by the experimental results, where is was shown that both yield very similar results for all conditions. The fact that instead of the (unknown) speech covariance matrix also the noisy covariance matrix can be used, makes this approach usable in practice.

In the experiments for the dependence of distance and input SNR in the external microphones, it was shown that the input SNR plays a much bigger role for the weighting of the two external microphones than the distance of $\mathrm{E}_1$ and $\mathrm{E}_2$ to the local array. This can be concluded since the real part of weighting factor $\alpha$ is almost constant at a value of about 0.5, i.e. identical weighting, when the input SNR in the external microphones is fixed and the distance is varied. For different input SNRs, however, a strong dependence becomes observable: The external microphone with the higher input SNR is weighted significantly more than the one with the lower input SNR, especially for large differences in input SNR, while yielding a weighting factor of around 0.5 if the input SNR is set to similar values in both external microphones.

The experimental results, furthermore, showed that a simple averaging approach yields sub-optimal results which are even below the performance of $\mathrm{E}_1$ and $\mathrm{E}_2$ separately. It was also shown that the performance of the output SNR-based optimal solutions and the orthogonal projection yield comparable results, even though they optimize different cost functions which, in principle, have different positions of their optima. Finally, both optimization approaches yield better results in terms of noise reduction and speech distortion than either external microphone alone, in almost all conditions, even if one of the available external microphones yields significantly worse performances than the other. This implies robustness of the proposed optimal solutions.

In conclusion, the output SNR-based solution using the noisy covariance matrix is the most promising approach to use in practice, since it does not depend on oracle knowledge.

The extension of the work done in this thesis which has already been done in [20] can be seen as an outlook. There, more than two external microphones were used and the external microphones were also filtered, since the bias in the entry of the RTF vector estimate corresponding to the respective external microphones can be neglected in practice [21].

As a practical next step, the RTF estimation method using several external microphones will be implemented in a real-time framework (as in [12]) for a more realistic evaluation of the performance.
Furthermore, it is planned to observe the influences of coherent interfering speakers, since the SC method was originally developed only for diffuse noise, which is in practice often not the case. The influences of a violated coherence assumption will be analyzed both analytically and experimentally.

# References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one or two ears," *The Journal of the Acoustic Society of America*, vol. 25, pp. 975–979, Sep. 1953.

[2] S. Doclo, *Multi-Microphone Noise Reduction and Dereveberation Techniques for Speech Applications.* PhD thesis, Katholike Universiteit Leuven, 2003.

[3] A.Bertrand, "Applications and trands in wireless acoustic sensor networks: a signal processing perspective," in *Proc. IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, (Ghent, Begium), 2011.

[4] S. Gannot, D. Burshtein, and E. Weinstein, "Signal Enhancement Using Beamforming and Non-Stationarity with Applications to Speech," *IEEE Transactions on Signal Processing*, vol. 49, pp. 1614–1626, Aug. 2001.

[5] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, pp. 269–302, Wiley, 2010.

[6] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Brisbane, Australia), pp. 544–548, Apr. 2015.

[7] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 1071–1086, Aug. 2009.

[8] N. Gößling and S. Doclo, "RTF-based binaural MVDR beamformer exploiting an external microphone in a diffuse noise field," in *Proc. ITG Conference on Speech Communication*, (Oldenburg, Germany), pp. 106–110, Oct. 2018.

[9] N. Gößling and S. Doclo, "Relative transfer function estimation exploiting spatially separated microphones in a diffuse noise field," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, (Tokyo, Japan), pp. 146–150, Sep. 2018.

[10] A. Bertrand, *Signal Processing Algorithms for Wireless Acoustic Sensor Networks.* PhD thesis, Katholieke Universiteit Leuven, 2011.

[11] A. Bertrand and M. Moonen, “Robust Distributed Noise Reduction in Hearing Aids with External Acoustic Sensor Nodes,” *EURASIP Journal on Advances in Signal Processing*, vol. 2009, p. 14 pages, Jan. 2009.

[12] W. Middelberg, N. Gößling, and S. Doclo, “Real-time evaluation of an RTF steered MVDR beamformer incorporating an external microphone,” in *Proc. Fortschritte der Akustik - DAGA*, (Rostock, Germany), Mar. 2019. (Abstract).

[13] D. Marquardt, *Development and Evaluation of Psychoacoustically Motivated Binaural Noise Reduction and Cue Preservation Techniques.* PhD thesis, Carl von Ossietzky Universität Oldenburg, 2015.

[14] T. Gerkmann and R. C. Hendriks, “Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 1383–1393, May 2012.

[15] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, “A consolidated perspective on multi-microphone speech enhancement and source separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 25, pp. 692–730, Apr. 2017.

[16] K. Bell, Y. Ephraim, and H. van Trees, “Robust adaptive beamforming under uncertainty in source direction-of-arrival,” in *Proc. 8th Workshop on Statistical Signal and Array Processing*, pp. 546–549, Aug. 1996.

[17] R. Varzandeh, M. Taseska, and E. A. P. Habets, “An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation,” in *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, (San Francisco, USA), pp. 11–15, Mar. 2017.

[18] S. M. Kay, *Fundamentals of Statistical Signal Processing.* TBS, 1993.

[19] G. Strang, *Linear Algebra and its Applications.* Thomson Brooks/Cole, 2006.

[20] N. Gößling, W. Middelberg, and S. Doclo, “RTF-steered binaural MVDR beamforming incorporating multiple external microphones,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz, USA), Oct. 2019. (Submitted).

[21] N. Gößling and S. Doclo, “RTF-steered binaural MVDR beamforming incorporating an external microphone for dynamic acoustic scenarios,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Brighton, UK), May 2019.

# Appendix A

To show that the weighting vector $\boldsymbol{\alpha}$ that minimizes the Hermitian angle in (5.4) is equal to the orthogonal projection of $\mathbf{h}_{\text{true}}$ on the estimation plane, spanned by the estimates $\hat{\mathbf{h}}_1$ and $\hat{\mathbf{h}}_2$, the argument of the arccos-function (the arccos- function has its minimum for an argument equal to 1) is considered, which can never exceed 1 according to the Cauchy-Schwarz inequality, i.e

$$\frac{|\mathbf{h}_{\text{true}}^H \hat{\mathbf{h}}|}{||\mathbf{h}_{\text{true}}||_2 ||\hat{\mathbf{h}}||_2} \leqslant 1\,. \tag{A.1}$$

Hence, the Hermitian angle is minimal if the argument is maximal, which leads to the maximization problem

$$\max_{\boldsymbol{\alpha}} \left( \frac{|\mathbf{h}_{\text{true}}^H \hat{\mathbf{H}} \boldsymbol{\alpha}|}{||\mathbf{h}_{\text{true}}||_2 ||\hat{\mathbf{H}} \boldsymbol{\alpha}||_2} \right)^2 , \tag{A.2}$$

where $\hat{\mathbf{h}}$ is substituted by its definition in (5.2) and the squared argument is maximized for the sake of simplicity. The cost function $c(\boldsymbol{\alpha})$ then reads

$$c(\boldsymbol{\alpha}) = \frac{1}{\mathbf{h}_{\text{true}}^H \mathbf{h}_{\text{true}}} \frac{\boldsymbol{\alpha}^H \hat{\mathbf{H}}^H \mathbf{h}_{\text{true}} \mathbf{h}_{\text{true}}^H \hat{\mathbf{H}} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^H \hat{\mathbf{H}}^H \hat{\mathbf{H}} \boldsymbol{\alpha}}\,. \tag{A.3}$$

The factor $1/(\mathbf{h}_{\text{true}}^H \mathbf{h}_{\text{true}})$ is neglected in the following, since it has no influence on the position of the maximum. For conciseness, the matrices in numerator and denominator are condensed, such that

$$\mathbf{A} = \hat{\mathbf{H}}^H \mathbf{h}_{\text{true}} \mathbf{h}_{\text{true}}^H \hat{\mathbf{H}}\,, \qquad \mathbf{B} = \hat{\mathbf{H}}^H \hat{\mathbf{H}}\,. \tag{A.4}$$

The derivative of (A.3) with respect to $\boldsymbol{\alpha}$ is then given by

$$\frac{dc(\boldsymbol{\alpha})}{d\boldsymbol{\alpha}} = \frac{2\mathbf{A}\boldsymbol{\alpha}(\boldsymbol{\alpha}^H \mathbf{B}\boldsymbol{\alpha}) - 2\mathbf{B}\boldsymbol{\alpha}(\boldsymbol{\alpha}^H \mathbf{A}\boldsymbol{\alpha})}{(\boldsymbol{\alpha}^H \mathbf{B}\boldsymbol{\alpha})^2}\,. \tag{A.5}$$

For the roots of (A.5), i.e. the solution of the equation when set to zero, the denominator is irrelevant. Therefore, it must only hold that

$$2\mathbf{A}\boldsymbol{\alpha}\boldsymbol{\alpha}^H \mathbf{B}\boldsymbol{\alpha} = 2\mathbf{B}\boldsymbol{\alpha}\boldsymbol{\alpha}^H \mathbf{A}\boldsymbol{\alpha}\,. \tag{A.6}$$

Considering the orthogonal projection and its solution for $\boldsymbol{\alpha}$ in (5.5), it can be seen that this vector solves the equation in (A.6)

$$\begin{split}&\hat{\mathbf{H}}^H\mathbf{h}_{\text{true}}\mathbf{h}_{\text{true}}^H\hat{\mathbf{H}}(\hat{\mathbf{H}}^H\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}^H\mathbf{h}_{\text{true}}\mathbf{h}_{\text{true}}^H\hat{\mathbf{H}}\,\underbrace{(\hat{\mathbf{H}}^H\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}^H\hat{\mathbf{H}}}_{\mathbf{I}}(\hat{\mathbf{H}}^H\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}^H\mathbf{h}_{\text{true}} = \\ &\underbrace{\hat{\mathbf{H}}^H\hat{\mathbf{H}}(\hat{\mathbf{H}}^H\hat{\mathbf{H}})^{-1}}_{\mathbf{I}}\hat{\mathbf{H}}^H\mathbf{h}_{\text{true}}\mathbf{h}_{\text{true}}^H\hat{\mathbf{H}}(\hat{\mathbf{H}}^H\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}^H\mathbf{h}_{\text{true}}\mathbf{h}_{\text{true}}^H\hat{\mathbf{H}}(\hat{\mathbf{H}}^H\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}^H\mathbf{h}_{\text{true}}\,,\end{split} \tag{A.7}$$

which then yields

$$\begin{split}&\hat{\mathbf{H}}^H\mathbf{h}_{\text{true}}\mathbf{h}_{\text{true}}^H\hat{\mathbf{H}}(\hat{\mathbf{H}}^H\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}^H\mathbf{h}_{\text{true}}\mathbf{h}_{\text{true}}^H\hat{\mathbf{H}}(\hat{\mathbf{H}}^H\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}^H\mathbf{h}_{\text{true}} = \\ &\hat{\mathbf{H}}^H\mathbf{h}_{\text{true}}\mathbf{h}_{\text{true}}^H\hat{\mathbf{H}}(\hat{\mathbf{H}}^H\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}^H\mathbf{h}_{\text{true}}\mathbf{h}_{\text{true}}^H\hat{\mathbf{H}}(\hat{\mathbf{H}}^H\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}^H\mathbf{h}_{\text{true}}\,.\end{split} \tag{A.8}$$

Both sides in (A.8) are obviously the same, which shows that the orthogonal projection of $\mathbf{h}_{\text{true}}$ on the column space of $\hat{\mathbf{H}}$ also minimizes the Hermitian angle.

## Selbstständigkeitserklärung

Hiermit erkläre ich an Eides statt, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutz habe. Außerdem versichere ich, dass ich die allgemeinen Prinzipien wissenschaftlicher Arbeit und Veröffentlichung, wie sie in den Leitlinien guter wissenschaftlicher Praxis der Carl von Ossietzky Universität Oldenburg festgelegt sind, befolgt habe.

Ort, Datum

Unterschrift Wiebke Middelberg