BINAURAL BEAMFORMING ALGORITHMS AND PARAMETER ESTIMATION METHODS EXPLOITING EXTERNAL MICROPHONES

Von der Fakultät für Medizin und Gesundheitswissenschaften
der Carl von Ossietzky Universität Oldenburg
zur Erlangung des Grades und Titels eines
Doktors der Ingenieurwissenschaften (Dr.-Ing.)
angenommene Dissertation

von

Nico Gößling
geboren am 27. Februar 1990
in Hamburg (Deutschland)

Nico Gößling: *Binaural beamforming algorithms and parameter estimation methods exploiting external microphones*

ERSTGUTACHTER:
Prof. Dr. ir. Simon Doclo

WEITERE GUTACHTER:
Prof. Dr. ir. Emanuël Habets
Prof. Dr. Sharon Gannot

TAG DER DISPUTATION:
12. Oktober 2020

ACKNOWLEDGMENTS

This thesis has been written at the Signal Processing group in the Department of Medical Physics and Acoustics at the Carl von Ossietzky Universität Oldenburg in Oldenburg, Germany. I would like to take the opportunity to thank the many people who have accompanied me on this academic journey.

First of all, I would like to thank Simon Doclo, for the years of support, advice and guidance, for his confidence in me and my work, and of course for the many interesting and funny memories.

Furthermore, I would like to thank Emanuël Habets and Sharon Gannot for their reviews of this thesis and for their interest in my work, which I greatly appreciate, as well as Steven van der Par as a further member of the examination committee.

A special thanks to all current and former members of the Signal Processing group. Especially of course to Daniel Marquardt, the best supervisor, Dörte Fischer, the best office mate and Wiebke Middelberg, the best HiWi.

I would also like to thank the following, without whom this work would not have been possible. Arne Jacobsen, who has been my companion and friend since the beginning of my studies. Eva Wilk, who sparked my theoretical and practical interest in audio technology. Christopher Hauth, Henning Schepker, Mats Exter and Marvin Tammen for the many funny and interesting conversations (totally sober of course...). Elior Hadad for the nice and very productive collaboration. Birger Kollmeier, Volker Hohmann and all members of the SFB HAPPAA and the Cluster of Excellence Hearing4all for letting me be a small, humble part of something bigger.

Finally, I would like to express my gratitude to my family and friends for their years of support and encouragement. Especially to my parents Bine and Burki, Netti and Gunni, and my partner Kira.

Hamburg, October 2020
Nico Gößling

ABSTRACT

In everyday speech communication situations undesired acoustic sources, such as competing speakers and background noise, frequently lead to a decreased speech intelligibility. Over the last decades, hearing devices have evolved from simple sound amplification devices to more sophisticated devices with complex functionalities such as multi-microphone speech enhancement. Binaural beamforming algorithms are spatial filters that exploit the information captured by multiple microphones on both sides of the head of the listener. Besides reducing the undesired sources, another important objective of a binaural beamforming algorithm is the preservation of the binaural cues of all sound sources to preserve the listener's spatial impression of the acoustic scene.

The aim of this thesis is to develop and evaluate advanced binaural beamforming algorithms and to incorporate one or more external microphones in a binaural hearing device configuration. The first focus is to improve state-of-the-art binaural beamforming algorithms, more in particular to develop a novel algorithm that jointly preserves the binaural cues of a desired source, interfering sources and background noise. The second focus is the incorporation of one or more external microphones to improve the noise reduction and binaural cue preservation performance of binaural beamforming algorithms, without assuming any a-priori knowledge about the position of the external microphones.

First, we propose a novel binaural beamforming algorithm, called binaural LCMV beamformer with partial noise estimation (BLCMV-N), which allows to preserve the binaural cues of the interfering sources and allows to control the trade-off between noise reduction performance and binaural cue preservation of the background noise. We analytically derive the performance of the proposed BLCMV-N beamformer in terms of noise reduction and binaural cue preservation, show its advantages compared to state-of-the-art binaural beamforming algorithms, and validate the theoretical findings with realistic experiments and a perceptual listening test.

Second, we investigate the incorporation of an external microphone in the binaural MVDR beamformer with partial noise estimation (BMVDR-N) for an arbitrary noise field. We derive analytical expressions for the noise reduction and binaural cue preservation performance showing that incorporating an external microphone allows to significantly increase the noise reduction performance compared to using only the head-mounted microphones, while preserving the spatial impression of the background noise. The derived analytical expressions generalize the results obtained in previous work and are experimentally validated for realistic acoustic scenarios.

Finally, we propose computationally efficient methods to estimate the relative transfer function (RTF) vectors of the desired source, exploiting one or more external mi-

crophones that are spatially separated from the head-mounted microphones. Without requiring any a-priori knowledge about the positions of the external microphones and the desired source, these methods enable to incorporate external microphones in a binaural hearing device configuration and to steer binaural beamforming algorithms. We further propose several procedures to combine different RTF vector estimates, that can be obtained when multiple external microphones are available. Experimental results for a moving desired source in a reverberant environment show that the proposed methods are applicable in realistic and highly dynamic acoustic scenarios, and outperform state-of-the-art RTF vector estimation methods at a much lower computational complexity.

ZUSAMMENFASSUNG

In alltäglichen Situationen der Sprachkommunikation führen unerwünschte Schallquellen, wie z.B. konkurrierende Sprecher und Hintergrundgeräusch, regelmäßig zu einer verminderten Sprachverständlichkeit. Im Laufe der letzten Jahrzehnte haben sich Hörgeräte von einfachen Geräten zur Schallverstärkung zu anspruchsvolleren Geräten mit komplexen Funktionalitäten wie z.B. Multimikrofon-Sprachverbesserung entwickelt. Binaurale Beamforming-Algorithmen sind räumliche Filter, die die von mehreren Mikrofonen auf beiden Seiten des Kopfes des Hörers aufgenommenen Informationen ausnutzen. Neben der Reduktion der unerwünschten Quellen ist ein weiteres wichtiges Ziel eines binauralen Beamforming-Algorithmus der Erhalt der binauralen Cues aller Schallquellen, um den räumlichen Eindruck des Hörers der akustischen Szene zu erhalten.

Ziel dieser Thesis ist die Entwicklung und Evaluierung fortschrittlicher binauraler Beamforming-Algorithmen und die Einbindung eines oder mehrerer externer Mikrofone in eine binaurale Hörgerätekonfiguration. Der erste Schwerpunkt liegt auf der Verbesserung moderner binauraler Beamforming-Algorithmen, insbesondere auf der Entwicklung eines neuartigen Algorithmus, der sowohl die binauralen Cues einer gewünschten Quelle als auch die von Störquellen und Hintergrundgeräusch bewahrt. Der zweite Schwerpunkt ist die Einbindung eines oder mehrerer externer Mikrofone zur Verbesserung der Geräuschunterdrückung und des Erhalts der binauralen Cues durch binaurale Beamforming-Algorithmen, ohne dass a-priori Kenntnisse über die Position der externen Mikrofone vorausgesetzt werden.

Als erstes schlagen wir einen neuartigen binauralen Beamforming-Algorithmus vor, den so genannten binauralen LCMV Beamformer mit partieller Geräuschschätzung (BLCMV-N), der es ermöglicht, die binauralen Cues der Störquellen zu erhalten und den Trade-Off zwischen Geräuschunterdrückung und Erhalt der binauralen Cues des Hintergrundgeräuschs zu kontrollieren. Wir analysieren die Leistungsfähigkeit des vorgestellten BLCMV-N Beamformers in Bezug auf Geräuschunterdrückung und den Erhalt binauraler Cues, zeigen seine Vorzüge im Vergleich zu modernen binauralen Beamforming-Algorithmen auf und validieren die theoretischen Ergebnisse mit realistischen Experimenten und einem perzeptiven Hörtest.

Als zweites untersuchen wir die Einbindung eines externen Mikrofons in den binauralen MVDR Beamformer mit partieller Geräuschschätzung (BMVDR-N) für ein beliebiges Geräuschfeld. Wir leiten analytische Ausdrücke für die Leistungsfähigkeit im Bezug auf Geräuschunterdrückung und den Erhalt binauraler Cues her, die zeigen, dass durch die Einbindung eines externen Mikrofons die Geräuschunterdrückung im Vergleich zur ausschließlichen Verwendung der kopfgetragenen Mikrofone erheblich gesteigert werden kann, wobei der räumliche Eindruck des Hintergrundgeräuschs

erhalten bleibt. Die hergeleiteten analytischen Ausdrücke verallgemeinern die in früheren Arbeiten erzielten Ergebnisse und werden experimentell für realistische akustische Szenarien validiert.

Schließlich präsentieren wir rechnerisch effiziente Methoden zur Schätzung der Relative-Transfer-Function Vektoren (RTF Vektoren) der gewünschten Quelle, wobei ein oder mehrere externe Mikrofone ausgenutzt werden, die räumlich von den am Kopf getragenen Mikrofonen getrennt sind. Ohne a-priori Kenntnisse über die Positionen der externen Mikrofone und der gewünschten Quelle zu benötigen, ermöglichen diese Methoden die Einbindung externer Mikrofone in eine binaurale Hörgerätekonfiguration und die Steuerung von binauralen Beamforming-Algorithmen. Darüber hinaus schlagen wir mehrere Verfahren zur Kombination verschiedener Schätzungen der RTF Vektoren vor, die man erhält, wenn mehrere externe Mikrofone zur Verfügung stehen. Experimentelle Ergebnisse für eine sich bewegende, gewünschte Quelle in einer halligen Umgebung zeigen, dass die präsentierten Methoden in realistischen und sehr dynamischen akustischen Szenarien anwendbar sind und die modernsten Verfahren zur Schätzung von RTF Vektoren bei wesentlich geringerem Rechenaufwand übertreffen.

GLOSSARY

Acronyms and abbreviations

ANOVA	analysis of variance
ATF	acoustic transfer function
BLCMV	binaural linearly constrained minimum variance
BLCMV-N	binaural linearly constrained minimum variance with partial noise estimation
BMVDR	binaural minimum variance distortionless response
BMVDR-N	binaural minimum variance distortionless response with partial noise estimation
BRIR	binaural room impulse response
cf.	`confer` (see also)
CPSD	cross power spectral density
CS	covariance subtraction
CW	covariance whitening
D-BLCMV	desired BLCMV
DOA	direction-of-arrival
DRR	direct-to-reverberant ratio
DS	delay-and-sum
eBMVDR	extended binaural minimum variance distortionless response
eBMVDR-N	extended binaural minimum variance distortionless response with partial noise estimation
e.g.	`exempli gratia` (for example)
EVD	eigenvalue decomposition
FFT	fast Fourier transform
FS	filter-and-sum
GSC	generalized sidelobe canceller
HATS	head-and-torso simulator
I-BLCMV	interference BLCMV
IC	interaural coherence
i.e.	`id est` (that is)

ILD	interaural level difference
iSNR	intelligibility-weighted signal-to-noise ratio
ITD	interaural time difference
ITF	interaural transfer function
LCMV	linearly constrained minimum variance
MPDR	minimum power distortionless response
MSC	magnitude-squared coherence
MUSHRA	multi-stimulus test with hidden reference and anchor
MVDR	minimum variance distortionless response
PSD	power spectral density
RIR	room impulse response
RTF	relative transfer function
SC	spatial coherence
SD	superdirective
SIR	signal-to-interference ratio
SNR	signal-to-noise ratio
SPP	speech presence probability
SRT	speech reception threshold
STFT	short-time Fourier transform
VAD	voice activity detector

Mathematical notation

x	scalar x
$\mathbf{x}$	vector $\mathbf{x}$
$\mathbf{X}$	matrix $\mathbf{X}$
x^*	complex conjugate of scalar x
$\mathbf{x}^T$	transpose of vector $\mathbf{x}$
$\mathbf{x}^H$	conjugate transpose of vector $\mathbf{x}$
$\mathbf{X}^T$	transpose of matrix $\mathbf{X}$
$\mathbf{X}^H$	conjugate transpose of matrix $\mathbf{X}$
$\mathbf{X}^{-1}$	inverse of matrix $\mathbf{X}$
$\hat{\mathbf{x}}$	estimate of vector $\mathbf{x}$
$\hat{\mathbf{X}}$	estimate of matrix $\mathbf{X}$
$\mathbf{p}\{\mathbf{X}\}$	principal eigenvector of matrix $\mathbf{X}$
$\mathbb{C}$	set of complex numbers
$\mathbb{R}$	set of real numbers

j	imaginary unit, i.e., $j^2 = -1$
$\mathcal{E}\{\cdot\}$	expectation operator
$\lvert\cdot\rvert$	magnitude
$\Re\{\cdot\}$	real part of a complex number
$\angle(\cdot)$	unwrapped phase

Fixed symbols

M_L	number of microphones in the left hearing device
M_R	number of microphones in the right hearing device
M_H	number of head-mounted microphones
M_E	number of external microphones
M	total number of microphones
m	microphone index
f	frequency bin index
F	total number of frequency bins
t	time frame index
T_d	time frame size in the STFT framework

T_s	time frame shift in the STFT framework
ω	angular frequency
$y_{L,m}$	m-th microphone signal of the left hearing device
$y_{R,m}$	m-th microphone signal of the right hearing device
$x_{L,m}$	desired source component in the m-th microphone signal of the left hearing device
$x_{R,m}$	desired source component in the m-th microphone signal of the right hearing device
$u_{L,m}$	interfering source component in the m-th microphone signal of the left hearing device
$u_{R,m}$	interfering source component in the m-th microphone signal of the right hearing device
$n_{L,m}$	noise component in the m-th microphone signal of the left hearing device
$n_{R,m}$	noise component in the m-th microphone signal of the right hearing device
$v_{L,m}$	undesired component in the m-th microphone signal of the left hearing device
$v_{R,m}$	undesired component in the m-th microphone signal of the right hearing device
$\mathbf{y}$	noisy input vector
$\mathbf{x}$	desired source component vector
$\mathbf{u}$	interfering source component vector
$\mathbf{n}$	noise component vector
$\mathbf{v}$	undesired component vector
$\mathbf{a}$	ATF vector of the desired source
$\mathbf{b}$	ATF vector of the interfering source
$\mathbf{y}_e$	extended noisy input vector
$\mathbf{x}_e$	extended desired source component vector
$\mathbf{u}_e$	extended interfering source component vector
$\mathbf{n}_e$	extended noise component vector
$\mathbf{v}_e$	extended undesired component vector
$\mathbf{a}_e$	extended ATF vector of the desired source
$\mathbf{b}_e$	extended ATF vector of the interfering source
s_x	desired source signal
s_u	interfering source signal

$\mathbf{e}_L$	left selection vector
$\mathbf{e}_R$	right selection vector
$\mathbf{e}_E$	external microphone selection vector
$\mathbf{e}_{E,i}$	i-th external microphone selection vector
y_L	left reference microphone signal
y_R	right reference microphone signal
y_E	external microphone signal
$y_{E,i}$	i-th external microphone signal
x_L	desired source component in the left reference microphone signal
x_R	desired source component in the right reference microphone signal
x_E	desired source component in the external microphone signal
$x_{E,i}$	desired source component in the i-th external microphone signal
u_L	interfering source component in the left reference microphone signal
u_R	interfering source component in the right reference microphone signal
u_E	interfering source component in the external microphone signal
$u_{E,i}$	interfering source component in the i-th external microphone signal
n_L	noise component in the left reference microphone signal
n_R	noise component in the right reference microphone signal
n_E	noise component in the external microphone signal
$n_{E,i}$	noise component in the i-th external microphone signal
v_L	undesired component in the left reference microphone signal
v_R	undesired component in the right reference microphone signal
v_E	undesired component in the external microphone signal
$v_{E,i}$	undesired component in the i-th external microphone signal
a_L	ATF between the desired source and the left reference microphone
a_R	ATF between the desired source and the right reference microphone
a_E	ATF between the desired source and the external microphone
$a_{E,i}$	ATF between the desired source and the i-th external microphone
$\mathbf{a}_L$	RTF vector of the desired source with respect to the left reference microphone

$\mathbf{a}_R$	RTF vector of the desired source with respect to the right reference microphone
$\mathbf{a}_{L,e}$	extended RTF vector of the desired source with respect to the left reference microphone
$\mathbf{a}_{R,e}$	extended RTF vector of the desired source with respect to the right reference microphone
b_L	ATF between the interfering source and the left reference microphone
b_R	ATF between the interfering source and the right reference microphone
b_E	ATF between the interfering source and the external microphone
$b_{E,i}$	ATF between the interfering source and the i-th external microphone
$\mathbf{b}_L$	RTF vector of the interfering source with respect to the left reference microphone
$\mathbf{b}_R$	RTF vector of the interfering source with respect to the right reference microphone
$\mathbf{b}_{L,e}$	extended RTF vector of the interfering source with respect to the left reference microphone
$\mathbf{b}_{R,e}$	extended RTF vector of the interfering source with respect to the right reference microphone
$\mathbf{R}_y$	noisy input covariance matrix
$\mathbf{R}_x$	desired source covariance matrix
$\mathbf{R}_u$	interfering source covariance matrix
$\mathbf{R}_n$	noise covariance matrix
$\mathbf{R}_v$	undesired covariance matrix
$\mathbf{\Gamma}$	spatial coherence matrix
$\mathbf{R}_{y,e}$	extended noisy input covariance matrix
$\mathbf{R}_{x,e}$	extended desired source covariance matrix
$\mathbf{R}_{u,e}$	extended interfering source covariance matrix
$\mathbf{R}_{n,e}$	extended noise covariance matrix
$\mathbf{R}_{v,e}$	extended undesired covariance matrix
$\mathbf{r}_{n,E}$	cross correlation vector between the noise component in the head-mounted microphone signals and the noise component in the external microphone signal
$\mathbf{\Gamma}_e$	extended spatial coherence matrix
$\mathbf{I}_M$	$M \times M$-dimensional identity matrix

$\mathbf{0}_{M_H}$	M_H-dimensional zero vector
p_{s_x}	PSD of the desired source
p_{s_u}	PSD of the interfering source
p_{x_L}	PSD of the desired source component in the left reference microphone signal
p_{x_R}	PSD of the desired source component in the right reference microphone signal
$p_{x_{LR}}$	CPSD of the desired source component in the reference microphone signals
p_{x_E}	PSD of the desired source component in the external microphone signal
p_{u_L}	PSD of the interfering source component in the left reference microphone signal
p_{u_R}	PSD of the interfering source component in the right reference microphone signal
$p_{u_{LR}}$	CPSD of the interfering source component in the reference microphone signals
p_{u_E}	PSD of the interfering source component in the external microphone signal
p_{n_L}	PSD of the noise component in the left reference microphone signal
p_{n_R}	PSD of the noise component in the right reference microphone signal
$p_{n_{LR}}$	CPSD of the noise component in the reference microphone signals
p_{n_E}	PSD of the noise component in the external microphone signal
p_n	PSD of the noise components for a homogeneous noise field
p_{v_E}	PSD of the undesired component in the external microphone signal
$p_{x_L}^{\text{out}}$	PSD of the desired source component in the left output signal
$p_{x_R}^{\text{out}}$	PSD of the desired source component in the right output signal
$p_{u_L}^{\text{out}}$	PSD of the interfering source component in the left output signal
$p_{u_R}^{\text{out}}$	PSD of the interfering source component in the right output signal
$p_{n_L}^{\text{out}}$	PSD of the noise component in the left output signal
$p_{n_R}^{\text{out}}$	PSD of the noise component in the right output signal

$\mathrm{SNR}_L^{\mathrm{in}}$	SNR in the left reference microphone signal
$\mathrm{SNR}_R^{\mathrm{in}}$	SNR in the right reference microphone signal
$\mathrm{SNR}_E^{\mathrm{in}}$	SNR in the external microphone signal
$\mathrm{SNR}_{E,i}^{\mathrm{in}}$	SNR in the i-th external microphone signal
$\mathrm{SIR}_L^{\mathrm{in}}$	SIR in the left reference microphone signal
$\mathrm{SIR}_R^{\mathrm{in}}$	SIR in the right reference microphone signal
$\mathrm{SIR}_E^{\mathrm{in}}$	SIR in the external microphone signal
$\mathrm{SIR}_{E,i}^{\mathrm{in}}$	SIR in the i-th external microphone signal
$\mathrm{SNR}_L^{\mathrm{out}}$	SNR in the left output signal
$\mathrm{SNR}_R^{\mathrm{out}}$	SNR in the right output signal
$\mathrm{SIR}_L^{\mathrm{out}}$	SIR in the left output signal
$\mathrm{SIR}_R^{\mathrm{out}}$	SIR in the right output signal
$\Delta\mathrm{SNR}_L$	left SNR improvement
$\Delta\mathrm{SNR}_R$	right SNR improvement
$\Delta\mathrm{SIR}_L$	left SIR improvement
$\Delta\mathrm{SIR}_R$	right SIR improvement
ρ	output SNR of the BMVDR beamformer
ρ_e	output SNR of the eBMVDR beamformer

$\mathrm{ITF}_x^{\mathrm{in}}$	input ITF of the desired source
$\mathrm{ITF}_u^{\mathrm{in}}$	input ITF of the interfering source
$\mathrm{ITF}_x^{\mathrm{out}}$	output ITF of the desired source
$\mathrm{ITF}_u^{\mathrm{out}}$	output ITF of the interfering source
$\mathrm{IC}_x^{\mathrm{in}}$	input IC of the desired source component
$\mathrm{IC}_u^{\mathrm{in}}$	input IC of the interfering source component
$\mathrm{IC}_n^{\mathrm{in}}$	input IC of the noise component
$\mathrm{IC}_x^{\mathrm{out}}$	output IC of the desired source component
$\mathrm{IC}_u^{\mathrm{out}}$	output IC of the interfering source component
$\mathrm{IC}_n^{\mathrm{out}}$	output IC of the noise component
$\mathrm{MSC}_x^{\mathrm{in}}$	input MSC of the desired source component
$\mathrm{MSC}_u^{\mathrm{in}}$	input MSC of the interfering source component
$\mathrm{MSC}_n^{\mathrm{in}}$	input MSC of the noise component
$\mathrm{MSC}_x^{\mathrm{out}}$	output MSC of the desired source component
$\mathrm{MSC}_u^{\mathrm{out}}$	output MSC of the interfering source component
$\mathrm{MSC}_n^{\mathrm{out}}$	output MSC of the noise component
$\mathrm{MSC}_n^{\mathrm{des}}$	desired output MSC of the noise component
$\Delta\mathrm{MSC}$	MSC error

η	mixing parameter
η^{des}	mixing parameter that leads to $\text{MSC}_n^{\text{des}}$ for the BMVDR-N beamformer
η_e^{des}	mixing parameter that leads to $\text{MSC}_n^{\text{des}}$ for the eBMVDR-N beamformer
η_{opt}	SNR-optimal mixing parameter
δ	interference scaling parameter
$\bar{\delta}$	adjusted interference scaling parameter
$\delta_{\text{opt},L}$	SNR-optimal left interference scaling parameter
$\delta_{\text{opt},R}$	SNR-optimal right interference scaling parameter

$\mathbf{w}_L$	left filter vector
$\mathbf{w}_R$	right filter vector
z_L	output signal of the left hearing device
z_R	output signal of the right hearing device

$\mathbf{w}_{\text{BMVDR},L}$	left filter vector of the BMVDR beamformer
$\mathbf{w}_{\text{BMVDR},R}$	right filter vector of the BMVDR beamformer
$\mathbf{w}_{\text{BMVDR-N},L}$	left filter vector of the BMVDR-N beamformer
$\mathbf{w}_{\text{BMVDR-N},R}$	right filter vector of the BMVDR-N beamformer
$\mathbf{w}_{\text{BLCMV},L}$	left filter vector of the BLCMV beamformer
$\mathbf{w}_{\text{BLCMV},R}$	right filter vector of the BLCMV beamformer
$\mathbf{w}_{\text{BLCMV-N},L}$	left filter vector of the BLCMV-N beamformer
$\mathbf{w}_{\text{BLCMV-N},R}$	right filter vector of the BLCMV-N beamformer
$\mathbf{w}_{\text{eBMVDR},L}$	left filter vector of the eBMVDR beamformer
$\mathbf{w}_{\text{eBMVDR},R}$	right filter vector of the eBMVDR beamformer
$\mathbf{w}_{\text{eBMVDR-N},L}$	left filter vector of the eBMVDR-N beamformer
$\mathbf{w}_{\text{eBMVDR-N},R}$	right filter vector of the eBMVDR-N beamformer

$\mathbf{C}$	constraint matrix
$\mathbf{C}_L$	left constraint matrix
$\mathbf{C}_R$	right constraint matrix
$\mathbf{g}$	response vector
$\mathbf{g}_L$	left response vector
$\mathbf{g}_R$	right response vector

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

We are constantly exposed to undesired sound sources in many everyday situations of speech communication, such as family gatherings, in restaurants or in traffic. In complex acoustic scenarios where many speakers speak simultaneously from all directions, i.e., the so-called cocktail party scenario [1], speech intelligibility can be particularly impaired. While even normal-hearing persons may be affected by this issue, hearing-impaired persons may be completely excluded from communicating via speech. When speech intelligibility is affected by undesired sources such as interfering sources (e.g., competing speakers) and background noise (e.g., diffuse babble noise), a simple restoration of loudness is typically not sufficient. In such complex acoustic scenarios, beamforming algorithms for head-mounted assistive hearing devices (e.g., hearing aids, earbuds and hearables) are crucial to improve speech intelligibility and speech quality. In a binaural hearing device configuration, where the listener is equipped with one device on each ear and both devices exchange their microphone signals, the information captured by all microphones on both sides of the head can be exploited [2–4]. Besides reducing interfering sources and background noise, another important objective of a binaural beamforming algorithm is the preservation of the listener's spatial impression of the acoustic scene in order to exploit the spatial release from masking [5] and to prevent confusions due to a possible mismatch between acoustical and visual information. This can be achieved by preserving the binaural cues of all sound sources in the acoustic scene.

To combine noise reduction and binaural cue preservation, two different paradigms are typically adopted [3]. In the first paradigm, two microphone signals, i.e., one on each device, are filtered with the same (real-valued) spectro-temporal gain, which intrinsically guarantees binaural cue preservation for all sound sources [6, 7]. In the second paradigm, considered in this thesis, all available microphone signals from both devices are processed by different (complex-valued) spatial filters [8, 9]. Although the second paradigm in general allows for more degrees of freedom to achieve more noise and interference reduction and less speech distortion than the first paradigm, there is typically a trade-off between noise and interference reduction performance and binaural cue preservation. State-of-the-art binaural beamforming algorithms typically preserve the binaural cues of the desired speech source, but distort the binaural cues of the undesired sources (interfering sources and background noise). Hence, several extensions have been proposed that additionally aim

at preserving the binaural cues of the undesired sources [10–13]. Typically, these extensions are designed to additionally preserve the binaural cues of either interfering sources or background noise.

To improve the performance of hearing devices, it has been proposed to use one or more external microphones (e.g., lying on a table, attached to a person) in conjunction with the head-mounted microphones [14–16]. Such external microphones make it possible to not only locally sample the sound field (at the listener's head) but to increase spatial diversity by spatially distributing the microphones. Besides technical challenges (e.g., synchronization, bandwidth limitations, transmission loss), one of the main challenges posed by incorporating external microphones is the fact that the relative position of the external microphones to the head-mounted microphones and the sound sources is unknown and may be highly time-varying.

The main objective of this thesis is to **develop and evaluate advanced binaural beamforming algorithms** that aim at simultaneously preserving the binaural cues of all sound sources (i.e., desired source, interfering sources and background noise) and to **incorporate one or more external microphones in a binaural hearing device configuration** without assuming any a-priori knowledge about the position of the external microphones or the desired source.

The remainder of this chapter is organized as follows. In **Section 1.1** we describe the general acoustic scenario considered in this thesis. In **Section 1.2** we discuss binaural cues and their influence on speech intelligibility. In **Section 1.3** we provide an overview of general beamforming algorithms, binaural beamforming algorithms and algorithms incorporating external microphones.

1.1 Acoustic scenario

Figure 1.1 depicts the most general acoustic scenario considered in this thesis, consisting of a listener wearing binaural hearing devices with multiple microphones, a desired (speech) source, undesired sources (interfering source and background noise), and multiple external microphones at unknown positions. Section 1.1.1 discusses the specific character of the desired source, while Section 1.1.2 discusses the undesired sources. Section 1.1.3 briefly describes the influence of the acoustic environment, i.e., reverberation, on the sound sources.

1.1.1 *Desired speech source*

In this thesis the desired source signal is typically assumed to be a speech signal. Speech signals are highly non-stationary, i.e., their envelope but also their frequency content changes rapidly over time, such that short-time stationarity can typically only be assumed for about 20 to 30 ms [17, 18]. Moreover, a speaker is not always active in a typical conversation, resulting in longer pauses between spoken words and sentences. These pauses can be exploited to estimate algorithm parameters, e.g., by means of a voice activity detector (VAD) [19–23] or speech presence probabil-

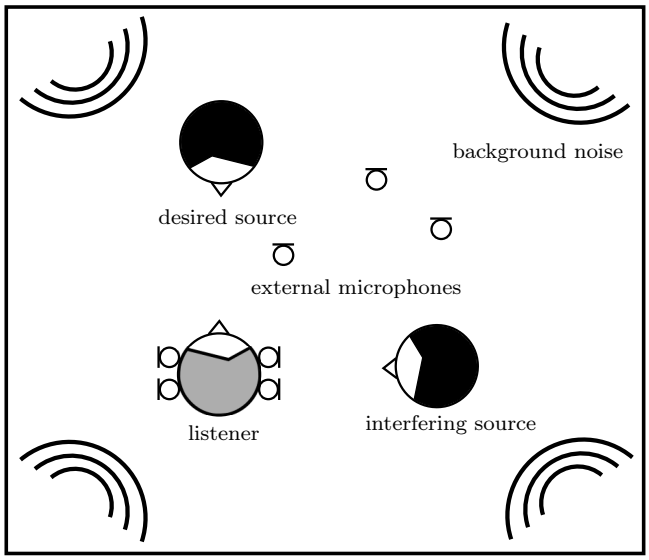

Fig. 1.1: General acoustic scenario with a listener wearing binaural hearing devices, a desired source, an interfering source and surrounding background noise, and multiple external microphones.

ity (SPP) estimator [24–27] that classifies speech-plus-noise periods and noise-only periods.

From a spatial perspective, it is typically assumed that the desired source is a spatially coherent (directional) source. More specifically, ignoring the influence of the acoustic environment, for a coherent source the signal components in two microphone signals are temporally shifted versions of each other, which may only differ in level. Although not always true, it is often assumed that the desired source is roughly in front of the listener. In this thesis we are not assuming a specific position of the desired source, i.e., we consider the position of the desired source as unknown and possibly even time-varying.

From a spectral perspective, the frequency content of speech signals is influenced by the different ways speech is produced [28]. Voiced speech, e.g., vowels, is produced by modulating the airflow in the throat by the vocal folds, has only little energy present above 4 kHz and has a mean frequency envelope that decreases by about 6 dB/octave. Unvoiced speech, e.g., consonants, is produced by turbulent airflows of the air at narrowings and has a broad, flat spectrum that can extend to about 12 kHz. For speech intelligibility the frequencies between 300 and 4000 Hz are of particular importance [29, 30].

1.1.2 *Undesired sources*

In this thesis we consider all sound sources that are detrimental to human communication, i.e., decrease speech intelligibility and speech quality, as undesired. We distinguish between two classes of undesired sources: Coherent (directional) interfering sources coming from a specific (unknown) direction, e.g., a competing speaker or an air conditioning unit, and incoherent (background) noise coming from all directions, e.g., microphone self-noise or a diffuse noise field produced by several speakers around the listener, i.e., the so-called cocktail-party scenario [1].

The spatially coherent interfering sources are typically assumed to be located at different positions than the desired source, which can be exploited by beamforming algorithms. While microphone self-noise is spatially totally incoherent, a diffuse noise field is typically modelled as a spherically or cylindrically isotropic noise field. The spatial coherence of such isotropic noise fields is frequency-dependent and depends on the distance between the measurement points, i.e., our ears [31] or microphones [32, 33]. As the distance between the measurement points increases, the spatial coherence typically decreases. In Chapter 6 we exploit this property for parameter estimation, using the external microphones that are spatially separated from the head-mounted microphones. Furthermore, if the background noise is homogeneous, it results in the same power at all measurement points.

1.1.3 *Acoustic environment*

In this thesis we assume that the acoustic environment, i.e., the room, itself can be modeled as a linear and time-invariant system [34], although the position of the sources and the microphones in the room may be time-varying. The acoustic path between a source and a microphone can be described by the so-called room impulse response (RIR). The RIR contains the direct path between the source and the microphone as well as all acoustic reflections, e.g., against walls or objects in the room, that are referred to as reverberation (sometimes further subdivided into early reflections and late reverberation). Figure 1.2 depicts an exemplary RIR. As can be observed, the dominant direct path arrives first, followed by a reverberation tail. Throughout this thesis we will consider several databases with measured RIRs for (binaural) hearing devices with and without external microphones, e.g., [35–38]. The database in [35] consists of measured RIRs for a binaural behind-the-ear hearing device configuration mounted on an artificial head without external microphones, either in an anechoic scenario or a reverberant cafeteria scenario. Additionally, recorded ambient noise is provided (cafeteria, courtyard, office) which was recorded using the same binaural hearing device configuration. The database in [36] consists of recorded signals in a real-world reverberant environment, where a listener with binaural behind-the-ear hearing devices was seated with three other persons at a circular table. Several external microphones were placed on the table (e.g., at a position where a person would probably lay down their smartphone) and the setup was surrounded by three layers of in total 56 seated persons producing realistic multi-talker babble noise. A more flexible option is to use simulated RIRs,

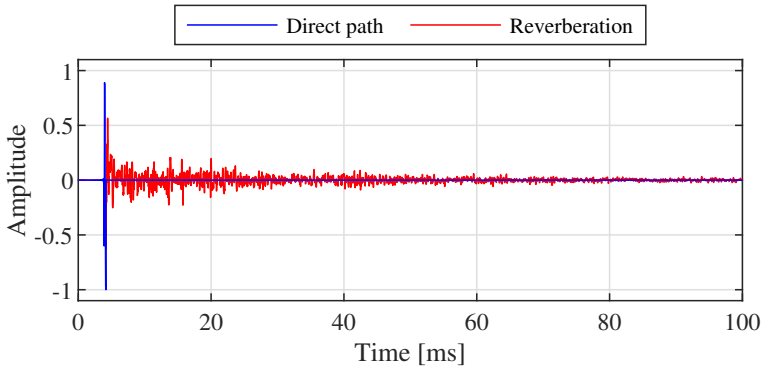

Fig. 1.2: Example of a room impulse response ($T_{60} \approx 300\,\text{ms}$, DRR $\approx 4.1\,\text{dB}$).

e.g., based on the image method for rectangular rooms [39]. An efficient implementation of the image method that also enables to consider a rigid sphere (approximating a head) can be found in [40, 41].

Some characteristic properties of the acoustic environment can be calculated directly from the RIR, e.g., the reverberation time (T_{60}), which is the time it takes for the reverberation to decrease by 60 dB [42], and the direct-to-reverberant ratio (DRR), which sets the energy of the direct sound in relation to the energy of the reverberation [43]. More reverberation typically leads to a smaller spatial coherence between the microphones and less sparsity in the time-frequency-domain due to an increased smearing over time.

In the time-domain the component in the microphone signals corresponding to a coherent source can be calculated by convolving the source signal with the RIRs between the source and the microphones. The equivalent of the RIR in the frequency-domain is defined as the acoustic transfer function (ATF). Assuming a multiplicative transfer function approximation in the short-time Fourier transform (STFT) domain [44], the signal component corresponding to a coherent source can be calculated in the frequency-domain by multiplying the source STFT coefficients with the ATFs. Another possibility is to consider so-called relative transfer functions (RTFs) [45, 46], which relate the ATFs between a coherent source and all microphones to a reference microphone. The component in the microphone signals corresponding to the incoherent noise can be calculated, e.g., using the method proposed in [47], where the microphone signals are generated under a predefined spatial coherence constraint.

1.2 Binaural cues

In this section we briefly discuss the binaural cues that are used by the human auditory system to localize sound sources and to determine, e.g., the width or diffuseness of a sound field. In Section 1.2.1 we discuss the binaural cues and how they relate to

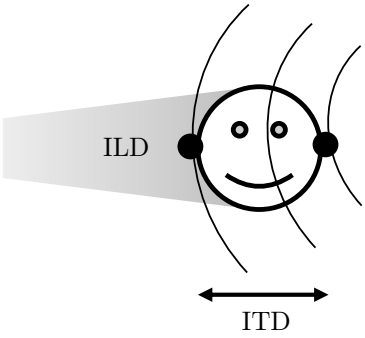

Fig. 1.3: Binaural cues: the path length difference between the left and the right ear causes the interaural time difference (ITD) and the shadowing of the head causes the interaural level difference (ILD).

spatial hearing. In Section 1.2.2 we briefly discuss the importance of binaural cues with regard to speech intelligibility due to spatial release from masking.

1.2.1 *Spatial hearing*

Binaural cues arise due to interaural differences, i.e., differences in the signals arriving at the left and the right ear of the listener [48]. Besides spectral cues, psychoacoustic studies have shown that the interaural time difference (ITD) and the interaural level difference (ILD) are particularly important cues for the localization of a single coherent source [48]. Figure 1.3 schematically illustrates the ITD and the ILD for a source originating from the side. The ITD is the time difference of arrival between sound waves at the left and the right ear. If a source is, e.g., directly in front of or directly behind the listener, then this time difference is approximately zero, because the path length to both ears is the same. If the source horizontally deviates from these positions, then the path lengths to the ears are different and the ITD becomes non-zero. In humans the highest occurring ITD is about 700 µs [48], which occurs when the source originates either from the left side or the right side and the path length difference to the ears is maximal. The ILD is the difference in sound pressure produced by a source at both ears. If a source originates from the side, then the sound pressure at the contralateral ear is lower than at the ipsilateral ear. This is especially the case at high frequencies and is mainly due to the shadowing of the head. Depending on the frequency, the position of the source and the exact shape of the head of the listener, the ILD can be up to 20 dB. It has been experimentally shown that psycho-acoustically the ITD plays a dominant role at low frequencies, whereas the ILD plays a dominant role at high frequencies for source localization [49].

While the spatial impression of a coherent source can be well described by the ITD and ILD cues, these cues cannot be used to describe the spatial impression of an incoherent sound field, where sound waves arrive at the ears of the listener from many

directions simultaneously. However, for incoherent sound fields it has been shown that the interaural coherence (IC) and the magnitude-squared coherence (MSC) can be used to describe, e.g., the perceived width or diffuseness of the sound field [50–52]. For a coherent source the MSC is equal to 1 for all frequencies, whereas for an incoherent sound field the MSC is frequency-dependent and is smaller than 1 for most frequencies [31–33]. Further, the IC and MSC can also be used to determine the reliability of the ITD and the ILD cues, especially in reverberant acoustic environments [53–55].

1.2.2 *Spatial release from masking*

Binaural cues (i.e., ITD, ILD, IC) play a significant role for speech intelligibility in terms of spatial release from masking [5, 48, 56–62]. Spatial release from masking refers to the increase in speech intelligibility that arises when the desired source and the undesired sources are spatially separated. For one interfering source, an improvement of the speech reception threshold (SRT) at 50 % speech intelligibility of over 10 dB has been reported if the desired source and the interfering source are spatially separated [5, 56]. It has been shown in [58] that for 4 interfering sources which are symmetrically placed around the listener, an SRT improvement of 1.9 dB can be achieved compared to when all interfering sources are placed in front of the listener. Further, for one desired source in a diffuse noise field, an improvement of the SRT up to 3.4 dB has been reported for binaural hearing compared to monaural hearing [63], whereas no such SRT improvement can be observed if the desired source and the noise both come from the same direction [59]. A detailed overview of studies concerned with the effect of spatial release from masking can be found in [5].

To preserve the spatial hearing of the listener and to allow the auditory system to take advantage of the spatial release from masking, it is desirable that beamforming algorithms for hearing devices preserve the binaural cues of all sound sources in the acoustic scene.

1.3 Overview of beamforming algorithms

In this section we provide a brief overview of beamforming algorithms that are of particular interest in the context of this thesis. In Section 1.3.1 we discuss general beamforming algorithms, i.e., not specifically for binaural hearing devices. In Section 1.3.2 we consider binaural beamforming algorithms, while in Section 1.3.3 we discuss algorithms incorporating one or more external microphones in a binaural hearing device configuration.

1.3.1 *General beamforming algorithms*

The main objective of general beamforming algorithms is to reduce all undesired sources (i.e., interfering sources and background noise) and limiting speech dis-

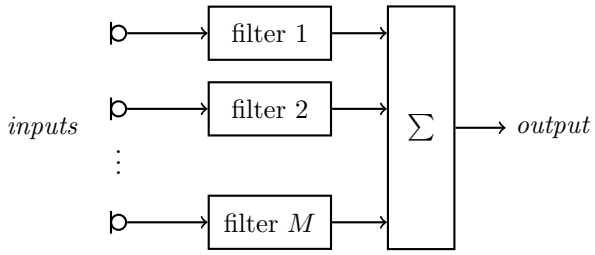

Fig. 1.4: Block diagram for the filter-and-sum (FS) structure.

tortion. In general, beamforming algorithms exploit spatial information by processing signals recorded with multiple microphones that are often arranged in a rather closely-spaced array configuration. Figure 1.4 depicts the filter-and-sum (FS) structure for M microphones, which is typically used to implement beamforming algorithms. In an FS structure all microphone signals are processed by different (complex-valued) filters and then summed to generate one output signal. One of the simplest beamforming algorithms is the delay-and-sum (DS) beamformer [64, 65], where the filters in the FS structure consist of simple delays. These delays are chosen in such a way that they spatially align the microphone signals to the assumed direction of the desired source, hence amplifying the desired source after the summation due to constructive interference. Considering these delays as algorithm parameters, the DS beamformer can be steered towards different directions [65]. A vector which can be used to steer beamforming algorithms, e.g., containing the delays, is referred to as steering vector in this thesis. The DS beamformer does not explicitly exploit information about the undesired sources, but maximizes the array gain in the case of spatially white noise [66]. Further, the superdirective (SD) beamformer has been proposed, which maximizes the array gain for a diffuse noise field [67, 68].

We distinguish between two classes of beamforming algorithms. Algorithms of the first class are fixed (data-independent) beamforming algorithms which are based on a-priori parameter and model assumptions that do not vary over time. A fixed implementation of the DS beamformer can be realized by assuming the direction-of-arrival (DOA) of the desired source to be fixed and choosing the parameters, i.e., the steering vector, accordingly. A fixed implementation of the SD beamformer can further be realized by modelling the coherence of the noise field as, e.g., spherically or cylindrically isotropic. Fixed beamforming algorithms usually require the microphone array topology to be known, i.e., the relative position of the microphones to each other. A fixed beamforming algorithm can therefore not be used for a microphone array where the relative position of the microphones to each other are unknown or can change over time, as it is the case for the external microphones considered in this thesis. Furthermore, the performance of fixed beamforming algorithms strongly depends on how well the a-priori assumptions and models match with reality. In complex acoustic scenarios, e.g., where the position of the desired source is difficult to predict or the noise field is not perfectly diffuse, fixed beamforming algorithms are therefore often not very performant and may even lead to a degradation of speech intelligibility and speech quality.

Algorithms of the second class are adaptive (data-dependent) beamforming algorithms which estimate the required parameters from the microphone signals. Adaptive beamforming algorithms are more flexible regarding the acoustic scenario, but typically require accurate parameter estimates to achieve high performance. To adaptively steer the DS and SD beamformer, the DOA of the desired source can be estimated [55, 69–75] and the steering vector can be set accordingly to spatially align the microphone signals to the estimated direction [76]. Exploiting the estimated signal statistics, i.e., covariance matrices, further leads to the widely-used class of adaptive beamforming algorithms that are based on constrained optimization problems [65]. First, the minimum power distortionless response (MPDR) beamformer [46, 65, 77] aims at minimizing the output variance subject to a single constraint to preserve the desired source component in a so-called reference microphone. Since minimizing the output variance may lead to target cancellation effects in case of DOA estimation errors of the desired source [65, 78], the minimum variance distortionless response (MVDR) beamformer [46, 65, 79] is typically used in practice, which aims at minimizing the noise variance in the output signal but requires an estimate of the noise covariance matrix. While the MVDR beamformer provides a good (background) noise reduction performance, it is not designed to explicitly control the reduction of interfering sources. Second, the linearly constrained minimum variance (LCMV) beamformer [46, 65, 80], as a more general version of the MVDR beamformer, aims at minimizing the noise variance subject to multiple constraints, e.g., to additionally preserve a scaled version of interfering source components in the reference microphone signal. Due to the additional constraints, the LCMV beamformer enables to control the reduction of interfering sources, but there are less degrees of freedom available for noise reduction, such that the noise reduction performance for the LCMV beamformer is lower or equal than for the MVDR beamformer. The MVDR and LCMV beamformers can be implemented in a so-called direct implementation, as considered in this thesis, or using the generalized sidelobe canceller (GSC) structure [46, 81, 82], which converts the constrained optimization problem into an unconstrained optimization problem. In a direct implementation the filter solving the respective constrained optimization problem is directly implemented based on estimates of the required parameters, e.g., a covariance matrix and steering vectors. A GSC structure only requires estimates of the steering vectors, from which a fixed beamformer and a so-called blocking matrix are constructed. The fixed beamformer generates a reference signal for the desired source, while the blocking matrix generates one or more noise reference signals. An adaptive filter [83] is then applied to the noise reference signals to reduce the residual noise in the reference signal for the desired source.

Since typically the ATFs between the sources and the microphones do not consist of simple delays, i.e., phase differences, but also include, e.g., reverberation, microphone characteristics or the acoustic influence of a head or device, ATFs should be used instead of mere delays in adaptive beamforming algorithms. However, since accurately estimating ATFs is a very difficult task in practice, especially in noisy and reverberant environments, it was proposed in [45, 84–86] to use RTFs instead of ATFs. In particular, to steer the MVDR beamformer an estimate of the RTFs of the desired source is required. Several RTF estimation methods have been pro-

posed in literature [85, 87–93]. The most popular methods are based either on covariance subtraction (CS) or covariance whitening (CW). These methods usually require an estimate of the microphone signal covariance matrix (e.g., estimated during speech-plus-noise periods) and the noise covariance matrix (e.g., estimated during noise-only periods). It should be noted that the computational complexity of the CW-based RTF estimation method is rather high due to the involved matrix operations (including an eigenvalue decomposition), which is especially relevant for a real-time implementation.

1.3.2 *Binaural beamforming algorithms*

Obviously, the general beamforming algorithms discussed in Section 1.3.1 can also be used for hearing devices. However, it should be realized that these beamforming algorithms only generate one output signal, whereas for the binaural hearing device configuration considered in this thesis two output signals are required, one for the left and one for the right ear of the listener. Compared to a bilateral hearing device configuration where both hearing devices operate independently, in a binaural hearing device configuration both hearing devices exchange their microphone signals, such that the information captured by all microphones on both sides of the head can be exploited [2–4]. Besides reducing all undesired sources (i.e., interfering sources and background noise) and limiting speech distortion, another important objective of a *binaural* beamforming algorithm is the preservation of the listener's spatial impression of the acoustic scene in order to exploit the spatial release from masking [5, 48, 56–62] and to prevent confusions due to a possible mismatch between acoustical and visual information. This is achieved by preserving the binaural cues (cf. Section 1.2) of all sound sources in the acoustic scene.

To combine noise reduction and binaural cue preservation, two different processing paradigms are typically adopted [3]. In the first processing paradigm, two microphone signals, i.e., one from each hearing device, are filtered with a common (real-valued) spectro-temporal gain, which intrinsically guarantees binaural cue preservation for all sound sources in the acoustic scene [6, 7, 94–102]. In the second processing paradigm, considered in this thesis, all available microphone signals from both devices are filtered and summed by different (complex-valued) filters [8–10, 12, 13, 103–108]. Figure 1.5 depicts the block diagram for the second processing paradigm applied to the considered binaural hearing device configuration.

As baseline binaural beamforming algorithm we consider the binaural minimum variance distortionless response (BMVDR) beamformer, which can be considered as the binaural version of the general MVDR beamformer discussed in Section 1.3.1. Compared to the general MVDR beamformer, the BMVDR beamformer aims at minimizing the noise variance in the output signals, while preserving the desired source component in two reference microphone signals, i.e., one on the left and one on the right hearing device. As reference microphones the frontal microphone on each hearing device is typically chosen. Since the reference microphones are usually relatively close to the ear canals of the listener, intuitively it can be seen that the spatial impression of the desired source can be preserved rather well by the BMVDR

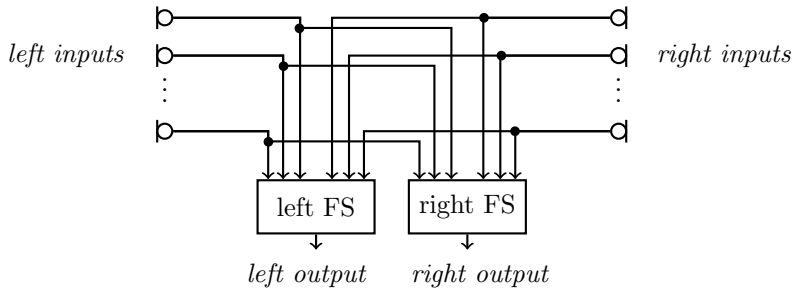

Fig. 1.5: Block diagram for the second binaural processing paradigm applied to the binaural hearing device configuration.

beamformer. For a single desired source, it has also been shown analytically that the BMVDR beamformer preserves the binaural cues of the desired source component but distorts the binaural cues of the undesired components (i.e., interfering sources and background noise) [2, 3, 10]. More in particular, after applying the BMVDR beamformer to the microphone signals, both output components exhibit the binaural cues of the desired source component and are hence perceived as coming from the direction of the desired source. This may not only lead to a mismatch between visual and auditory information, but also prevents the auditory system to take advantage of the spatial release from masking (cf. Section 1.2.2).

Aiming at additionally preserving the binaural cues of interfering sources or background noise, several extensions of the BMVDR beamformer have been proposed, e.g., by incorporating additional constraints into the spatial filter design [12, 105, 107–110] or by mixing with the scaled (noisy) reference microphone signals [9–11, 13, 111, 112]. In this thesis we consider two specific extensions of the BMVDR beamformer. The binaural linearly constrained minimum variance (BLCMV) beamformer [12] and the BMVDR beamformer with partial noise estimation (BMVDR-N) [11, 13] (cf. Chapter 3). The BLCMV beamformer can be considered as the binaural version of the general LCMV beamformer discussed in Section 1.3.1 and an extension of the BMVDR beamformer using additional constraints in the spatial filter design. In addition to preserving the desired source component in the reference microphone signals, the BLCMV beamformer preserves a scaled version of each interfering source component in the reference microphone signals by means of an *interference scaling parameter*, while minimizing the noise variance in the output signals. The additional constraints hence allow to preserve the binaural cues of interfering sources and enable to directly control the interference reduction performance. However, due to the additional constraints there are less degrees of freedom available for noise reduction, such that the noise reduction performance for the BLCMV beamformer is lower than for the BMVDR beamformer. Furthermore, the BLCMV beamformer does not allow to control the binaural cues of the background noise.

For the BMVDR-N beamformer it has been shown that the output signals can be interpreted as a mixture between the output signals of the BMVDR beamformer and the (noisy) reference microphone signals depending on a *mixing parameter*.

Hence, the BMVDR-N beamformer provides a trade-off between noise reduction performance and binaural cue preservation of the background noise. In order to achieve a desired output MSC and hence spatial impression of the noise component (cf. Section 1.2.1), in [13] a closed-form expression for the mixing parameter of the BMVDR-N beamformer has been derived. The desired output MSC of the noise component can, e.g., be psycho-acoustically motivated based on the IC discrimination ability of the human auditory system [13, 106], aiming for the spatial impression of the noise component in the reference microphone signals and the noise component in the output signals to be indistinguishable. While for (incoherent) background noise this approach showed promising results [13, 113], the effect of partial noise estimation on a (coherent) interfering source depends on the relative position of the interfering source to the desired source and is harder to control [10].

Since the BLCMV beamformer is able to control the binaural cues of the interfering sources but not of the background noise, whereas the BMVDR-N beamformer is able to control the binaural cues of the background noise but not of the interfering sources, one goal of this thesis is to **develop a binaural beamforming algorithm that is able to control the binaural cues of the interfering sources and the background noise**.

1.3.3 *Incorporation of external microphones*

As already mentioned in Section 1.3.1, microphone arrays often consist of closely-spaced microphones. In a hearing device the inter-microphone distance is a few millimetres up to a few centimetres. Due to the typical diameter of the listener's head, the maximum distance between two microphones in a binaural hearing device configuration is about 17 cm. Nevertheless, the sound field is only sampled locally at the listener's head.

Recent advances in communication system technology enable the deployment of so-called (wireless) acoustic sensor networks, consisting of several spatially distributed microphones or microphone arrays [114]. The spatial distribution of the microphones makes it possible to not only locally sample the sound field (at the listener's head) but to increase spatial diversity and hence the probability that one or more microphones are close to the desired source, thus leading to a larger signal-to-noise ratio (SNR) and DRR in the microphone signals. If no dedicated device is available for centralized processing (often called the fusion centre), distributed processing can be an option where the spatially distributed devices process their microphone signals locally and share the results with neighbouring devices [115–125]. Even if such a fusion centre is available for a centralized processing, one of the main technical challenges still is the synchronization of the microphone signals, since it cannot be guaranteed that all devices run at exactly the same sampling rate. Different methods for the synchronization of the microphone signals can be found, e.g., in [126–134]. In this thesis we assume that all microphone signals are available in a fusion centre for centralized processing, are transmitted (e.g., via a wireless link) without any transmission delay and are synchronized. However, besides the technical challenges (e.g., synchronization, bandwidth limitations, transmission loss), one of the

main challenges in acoustic sensor networks that we address in this thesis is the fact that the relative position of the microphones to each other and the sound sources is unknown and may be highly time-varying.

As a special case of acoustic sensor networks in this thesis, we consider hearing devices that are wirelessly linked to one or more external microphones. A typical use case is a classroom, where the teacher (as a well-defined desired source) wears a microphone which is then transmitted to the hearing devices of the hearing-impaired students [135, 136]. Others are hearing devices linked to, e.g., table microphones, smartphones or laptops, that assist the hearing device user in everyday noisy situations such as in restaurants, at family gatherings or in meetings [137–144]. Unfortunately, directly listening to the external microphone signal completely destroys the spatial impression of the acoustic scene, since the external microphone signal is a monaural signal that does not contain any binaural cues and hence leads to in-head perception. In [145] it was shown experimentally that the imprinting of binaural cues on an external microphone signal using a structural binaural model of the head and pinna helps to externalize the impression of the acoustic scene. It is hence desirable to use one or more external microphones in conjunction with the head-mounted microphones in a binaural beamforming algorithm, aiming at improving the noise reduction performance and the binaural cue preservation by exploiting the increased spatial diversity.

Several noise reduction and source localization algorithms have been recently proposed for (binaural) hearing devices incorporating one or more external microphones [14–16, 145–155]. In [14] a distributed noise reduction algorithm that has a reduced communication bandwidth and computational complexity has been evaluated for hearing devices with multiple external microphones, which is of particular interest when the hearing devices cannot be used as fusion centres for centralized processing. It was shown that the noise reduction performance in hearing devices can be significantly improved when the external microphones are incorporated in the hearing device configuration. Further, it was observed that the distributed algorithm was not able to achieve the same noise reduction performance as a centralized baseline algorithm. In [15] it has been shown for one interfering source that incorporating an external microphone in a partial noise estimation structure enables to improve both the noise reduction performance as well as the binaural cues, i.e., the ITD and ILD, of the interfering source compared to only using the head-mounted microphones. Assuming that an external microphone is placed very close to the desired source, such that the external microphone signal can be considered nearly noiseless, in [149] it has been shown that exploiting the external microphone signal can significantly improve the DOA estimation accuracy of the desired source. Although the results are encouraging, in practice it cannot always be assumed that the external microphone is placed very close to the desired source. As shown in [150], when the external microphone is placed close to the listener (e.g., lying on a table), the listener's body can be used to shield the external microphone from an interfering source in the back hemisphere. To address front-back ambiguity in single-microphone hearing devices, a frontal target source presence probability estimator was proposed that increased the interference reduction performance of a spectral filter. However, the assumption

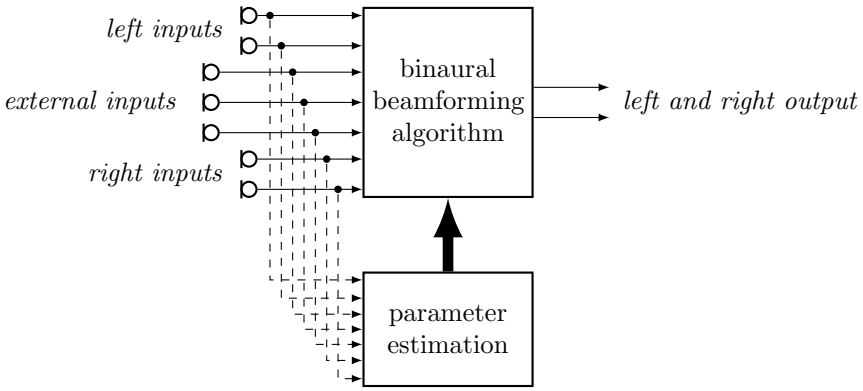

Fig. 1.6: Structure of a data-dependent binaural beamforming algorithm incorporating multiple external microphone.

that the external microphone is placed close to the listener may also not always be true in practice, and furthermore, this placement is not necessarily optimal for incorporating an external microphone in a binaural beamforming algorithm. In [16] three methods to incorporate multiple external microphones in a GSC structure were proposed, whereby the GSC structure was based on the head-mounted microphones and could not be changed. Further, it was assumed that the relative position of the desired source to the head-mounted microphones is known a-priori. Two of the three methods involved a procedure for completing an extended blocking matrix (including the head-mounted microphones and the external microphones), whereas the third method used the output of the head-mounted microphone-based GSC structure with an orthogonalized version of the external microphone signals for a generalized eigenvalue decomposition. All three methods were able to significantly improve the noise reduction performance compared to only using the head-mounted microphones, while the third method showed the best performance. The results in [16] are particularly interesting if the configuration to be extended cannot be changed and if the relative position of the desired source can be assumed a-priori (e.g., in front of the listener).

In this thesis we aim at jointly processing the signals of all microphones, i.e., all head-mounted microphones and all external microphones. We do not assume that the position of any external microphone or the desired source is known. The relative position of the external microphones or the desired source to the head-mounted microphones may change, e.g., if the listener moves or turns his head or if somebody moves the external microphones. Therefore, it is not possible to use fixed implementations of binaural beamforming algorithms when incorporating external microphones, since algorithm parameters, e.g., steering vectors containing the RTFs of the desired source, cannot be modelled a-priori. Although a variety of RTF estimation methods exists [84–93], they do not specifically exploit the increased spatial diversity due to the incorporation of the external microphones. It is hence one goal of this thesis to **develop binaural beamforming algorithms and parameter**

estimation methods that exploit one or more external microphones for a complex acoustic scenario where the positions of the external microphones and the desired source are unknown. Figure 1.6 depicts the general envisaged structure of a data-dependent implementation of a binaural beamforming algorithm that incorporates multiple external microphones. As can be observed, all microphone signals, i.e., all head-mounted microphone signals and external microphone signals, are used to estimate the required algorithm parameters (e.g., covariance matrices or the RTF vectors of the desired source). Furthermore, all microphone signals are jointly processed by the binaural beamforming algorithm to generate the output signal for the left and the right ear of the listener.

1.4 Outline of the thesis and main contributions

The main objective of this thesis is to develop and evaluate advanced binaural beamforming algorithms and to incorporate one or more external microphones in a binaural hearing device configuration. The first focus is to improve state-of-the-art binaural beamforming algorithms, more in particular to develop a binaural beamforming algorithm that jointly preserves the binaural cues of the desired source, interfering sources and background noise. The second focus is the incorporation of one or more external microphones to improve the noise reduction performance and binaural cue preservation of binaural beamforming algorithms, without assuming any a-priori knowledge about their position.

The main contributions in this thesis are threefold. First, we **propose a novel binaural beamforming algorithm**, called binaural LCMV beamformer with partial noise estimation (BLCMV-N), which merges the advantages of the BLCMV beamformer and the BMVDR-N beamformer. The BLCMV-N beamformer allows to preserve the binaural cues of interfering sources and allows to control the trade-off between noise reduction performance and binaural cue preservation of the background noise. Second, we **investigate the incorporation of an external microphone in the BMVDR-N beamformer for an arbitrary noise field**. We analytically and experimentally show that incorporating an external microphone allows to significantly increase the output SNR compared to using only the head-mounted microphones, while preserving the spatial impression of the background noise. Third, we **propose computationally efficient methods to estimate the RTF vectors of the desired source, exploiting one or more external microphones that are spatially separated from the head-mounted microphones**. Without requiring any a-priori knowledge about the positions of the external microphones and the desired source, experimental results for a moving source in a reverberant environment show that the proposed methods outperform state-of-the-art RTF vector estimation methods in terms of estimation accuracy and noise reduction performance when used to steer a BMVDR beamformer.

In the remainder of this section we provide a chapter-by-chapter overview of this thesis, describing the content and contribution of each chapter. A schematic overview of the thesis is depicted in Figure 1.7.

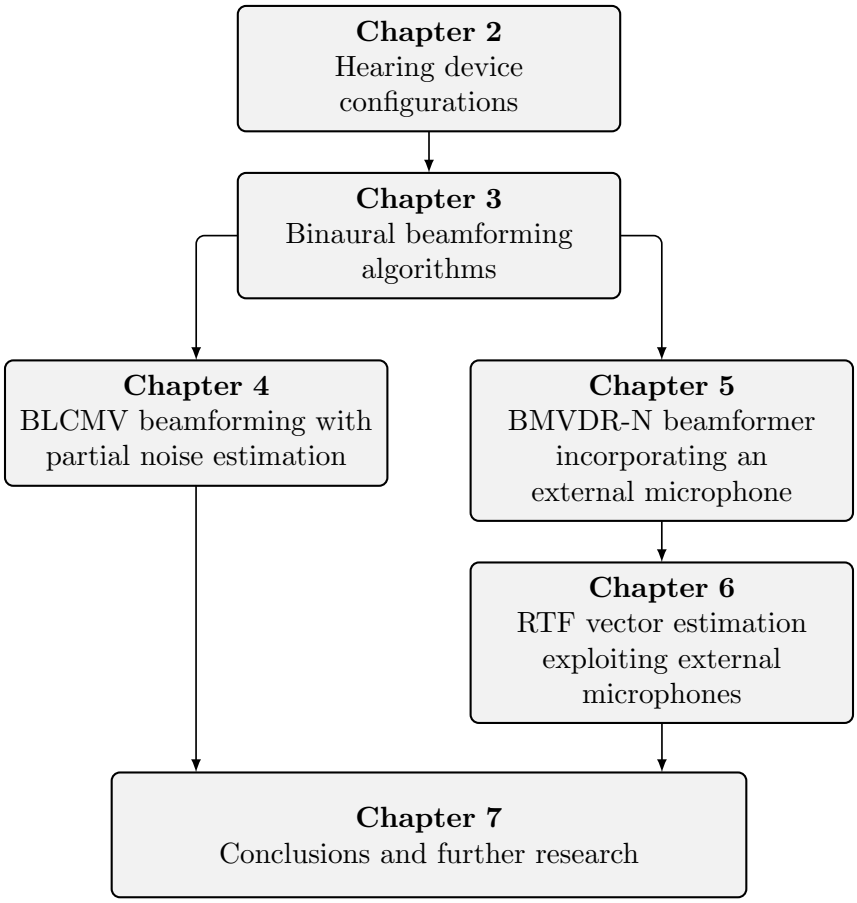

Fig. 1.7: Schematic overview of the thesis.

In **Chapter 2** we introduce the signal model for the binaural hearing device configuration, the extended binaural hearing device configuration (incorporating one external microphone) and the multi-extended binaural hearing device configuration (incorporating multiple external microphones), as well as the mathematical notation that is used throughout this thesis. Furthermore, we introduce the mathematical definitions of the objective performance measures and the binaural cues.

In **Chapter 3** we provide a detailed overview of state-of-the-art binaural beamforming algorithms. The considered binaural beamforming algorithms are the binaural minimum variance distortionless response (BMVDR) beamformer, the binaural linearly constrained minimum variance (BLCMV) beamformer and the BMVDR beamformer with partial noise estimation (BMVDR-N). We briefly review their performance in terms of noise reduction and binaural cue preservation, and point out the parameters that are required for their practical implementation. We further discuss several state-of-the-art methods to model or estimate these parameters, more in particular the covariance matrices and the steering vectors, i.e., the ATF or RTF vectors.

In **Chapter 4** we propose the BLCMV beamformer with partial noise estimation (BLCMV-N), merging the advantages of the BLCMV beamformer and the BMVDR-N beamformer, i.e., preserving the binaural cues of the interfering sources and controlling the reduction of the interfering sources as well as the binaural cues of the background noise. First, we derive two decompositions for the BLCMV-N beamformer which reveal differences and similarities between the BLCMV-N beamformer and the BLCMV beamformer. We show that the output signals of the BLCMV-N beamformer can be interpreted as a mixture between the noisy reference microphone signals and the output signals of a BLCMV beamformer using an adjusted interference scaling parameter. We then analytically derive the performance of the BLCMV-N beamformer in terms of noise and interference reduction performance and binaural cue preservation. We show that the output SNR of the BLCMV-N beamformer is smaller than or equal to the output SNR of the BLCMV beamformer and derive the optimal interference scaling parameter maximizing the output SNR of the BLCMV-N beamformer. The derived analytical expressions are first validated using measured anechoic ATFs. In addition, more realistic experiments are performed using recorded signals for a binaural hearing device in a reverberant cafeteria with one interfering source and multi-talker babble noise. Both the objective performance measures as well as the results of a perceptual listening test with 13 normal-hearing participants show that the proposed BLCMV-N beamformer is able to preserve the binaural cues and hence the spatial impression of the interfering source (like the BLCMV beamformer), while trading off between noise reduction performance and binaural cue preservation of the background noise (like the BMVDR-N beamformer). The publication related to this chapter is [156].

In **Chapter 5** we investigate the incorporation of one external microphone in the BMVDR and BMVDR-N beamformer, leading the to extended BMVDR (eBMVDR) beamformer and extended BMVDR-N (eBMVDR-N) beamformer, respectively. We consider an arbitrary noise field and derive analytical expressions for the output SNR and the binaural cues (more in particular the MSC) of the output noise compo-

nent when incorporating one external microphone in the eBMVDR-N beamformer. First, we show that an external microphone enables to obtain either a larger output SNR for the same mixing parameter or the same output SNR for a larger mixing parameter compared to using only the head-mounted microphones. Secondly, we show that the same desired output MSC of the noise component can be obtained for a smaller mixing parameter, implying that an external microphone enables to achieve the same spatial impression of the noise component compared to using only the head-mounted microphones while achieving a larger output SNR. The derived analytical expressions are first validated using simulated anechoic ATFs, where the listener's head is modelled as a rigid sphere [40]. In addition, experiments are performed using recorded signals for a binaural hearing device configuration in a reverberant environment with multiple interfering talkers as background noise [36]. For different positions of the external microphone and the desired source, the experimental results show that also in a realistic scenario incorporating an external microphone in the BMVDR-N beamformer significantly increases the output SNR and decreases the mixing parameter required to obtain a desired output MSC, i.e., spatial impression, of the noise component. The results generalize the results obtained in [15] assuming a coherent (directional) interference source, and the results in [157] assuming a homogeneous diffuse noise field and a desired source in front of the listener. The publications related to this chapter are [157, 158].

In **Chapter 6** we propose computationally efficient methods to estimate the RTF vectors of the desired source by exploiting one or multiple external microphones. The external microphones are assumed to be spatially separated from the head-mounted microphones, such that the spatial coherence (SC) between the noise component in the head-mounted microphone signals and the noise component in the external microphone signals is low. We first consider a binaural hearing device configuration with only one additional external microphone and propose an SC-based RTF vector estimation method, which estimates the RTF vectors of the desired source based on the (noisy) microphone signal covariance matrix. Assuming the SC between the noise components to be zero, we show that the SC-based method yields an unbiased estimate of the elements of the RTF vectors corresponding to the head-mounted microphones, while the element corresponding to the external microphone is biased. We provide a detailed bias analysis for an arbitrary noise field, a diffuse noise field and an interfering source. Next, we consider more than one external microphone and show that different RTF vector estimates can be obtained by using the proposed SC-based method for each external microphone. We propose several procedures to combine these RTF vector estimates, either by selecting the estimate corresponding to the highest input SNR, by averaging the estimates or by combining the estimates in order to maximize the output SNR of the eBMVDR beamformer filtering all microphone signals. Experimental results using recorded signals of a moving desired source for a binaural hearing device configuration with one or more external microphones in a reverberant environment show that the proposed SC-based method outperforms state-of-the-art RTF vector estimation methods in terms of noise reduction performance when used to steer the (e)BMVDR beamformer. In addition, the experimental results show that the output SNR-maximizing combination pro-

cedure of different RTF vector estimates yields the largest SNR improvement. The publications related to this chapter are [159–162].

In **Chapter 7** we summarize the main findings of the thesis and provide an outlook on potential further research.

2

HEARING DEVICE CONFIGURATIONS, NOTATION AND PERFORMANCE MEASURES

In this chapter we present the general notation and the signal models for the binaural hearing device configurations with and without external microphones (Section 2.1), as well as the mathematical definitions of the objective performance measures and the binaural cues (Section 2.2).

2.1 Hearing device configurations and signal models

Section 2.1.1 introduces the signal model for the binaural hearing device configuration, i.e., using only the head-mounted microphones. The signal models for the extended and the multi-extended binaural hearing device configurations, i.e., incorporating one or multiple external microphones, are introduced in Section 2.1.2 and Section 2.1.3, respectively. The binaural hearing device configuration, considered in Chapters 3 and 4, and the extended binaural hearing device configuration, considered in Chapters 5 and 6, can be seen as special cases of the multi-extended binaural hearing device configuration, considered in Chapter 6.

2.1.1 *Binaural hearing device configuration*

Consider a (head-mounted) binaural hearing device configuration as depicted in Figure 2.1, consisting of one hearing device with M_L microphones on the left side and one hearing device with M_R microphones on the right side of the head of the listener. The total number of head-mounted microphones is equal to $M_H = M_L + M_R$. In this thesis we generally consider an acoustic scenario with one desired source (target speaker) and one interfering source (competing speaker) in a noisy and reverberant acoustic environment, where the background noise is assumed to be incoherent (e.g., diffuse babble noise, sensor noise). In special cases we will also consider an acoustic scenario with only the desired source and the interfering source, or with only the desired source and the background noise.

To represent discrete-time signals in the frequency-domain, we use the short-time Fourier transform (STFT) [34], where a discrete-time signal $y_d(t_d) \in \mathbb{R}$, with t_d

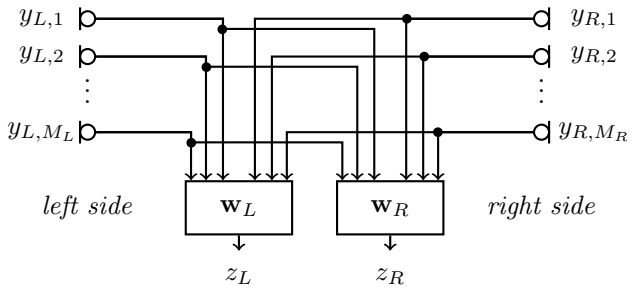

Fig. 2.1: Binaural hearing device configuration with M_L microphones on the left side and M_R microphones on the right side.

denoting the discrete time index, is split into overlapping time frames of size T_d and weighted by a sliding analysis window $w_{\mathrm{STFT}}(t_d)$. Using the discrete-time Fourier transform of size T_d, each weighted time frame is transformed to the frequency-domain as

$$y(f,t) = \sum_{t_d=0}^{T_d-1} y_d(tT_s + t_d) w_{\mathrm{STFT}}(t_d) e^{\frac{-j2\pi f t_d}{T_d}} \in \mathbb{C}\,, \tag{2.1}$$

with f the frequency bin index, t the time frame index, T_s the time frame shift, and j the imaginary unit (i.e., $j^2 = -1$). Using (2.1), in the STFT-domain the m-th microphone signal of the left hearing device $y_{L,m}(f,t) \in \mathbb{C}$ can be decomposed as

$$y_{L,m}(f,t) = x_{L,m}(f,t) + u_{L,m}(f,t) + n_{L,m}(f,t) = x_{L,m}(f,t) + v_{L,m}(f,t)\,, \tag{2.2}$$

with $m \in \{1, \ldots, M_L\}$, $x_{L,m}(f,t) \in \mathbb{C}$ the desired source component, $u_{L,m}(f,t) \in \mathbb{C}$ the interfering source component and $n_{L,m}(f,t) \in \mathbb{C}$ the (background) noise component. The (overall) undesired component $v_{L,m}(f,t) \in \mathbb{C}$ is defined as the sum of the interfering source component $u_{L,m}(f,t)$ and the noise component $n_{L,m}(f,t)$. Similarly, the m-th microphone signal of the right hearing device $y_{R,m}(f,t) \in \mathbb{C}$ can be decomposed as

$$y_{R,m}(f,t) = x_{R,m}(f,t) + u_{R,m}(f,t) + n_{R,m}(f,t) = x_{R,m}(f,t) + v_{R,m}(f,t)\,, \tag{2.3}$$

with $m \in \{1, \ldots, M_R\}$. For the sake of conciseness, we omit the variables f and t in the remainder of this thesis, except where specifically needed.

Stacking all microphone signals of both the left and the right hearing device in a vector, the M_H-dimensional noisy input vector is defined as

$$\mathbf{y} = [y_{L,1}, \ldots, y_{L,M_L}, y_{R,1}, \ldots, y_{R,M_R}]^T \in \mathbb{C}^{M_H}\,, \tag{2.4}$$

where $(\cdot)^T$ denotes the transpose. Using (2.2) this vector can be written as

$$\mathbf{y} = \mathbf{x} + \mathbf{u} + \mathbf{n} = \mathbf{x} + \mathbf{v}\,, \tag{2.5}$$

where $\mathbf{x}$, $\mathbf{u}$, $\mathbf{n}$ and $\mathbf{v} = \mathbf{u} + \mathbf{n}$ are defined similarly as $\mathbf{y}$ in (2.4).

The desired source component $\mathbf{x}$ and the interfering source component $\mathbf{u}$ in (2.5) can be written as

$$\mathbf{x} = \mathbf{a}s_x\,, \tag{2.6}$$

$$\mathbf{u} = \mathbf{b}s_u\,, \tag{2.7}$$

where $s_x \in \mathbb{C}$ and $s_u \in \mathbb{C}$ denote the desired source signal and the interfering source signal, respectively, and $\mathbf{a}$ and $\mathbf{b}$ denote the M_H-dimensional acoustic transfer function (ATF) vectors, containing the ATFs between the microphones and the desired source and the ATFs between the microphones and the interfering source, respectively. It should be noted that these ATFs include reverberation, the microphone characteristics and the head shadow effect.

Without loss of generality, we define the first microphone of each hearing device as the so-called reference microphone. To simplify the notation, the reference microphone signals $y_{L,1}$ and $y_{R,1}$ are denoted as y_L and y_R, i.e.,

$$y_L = \mathbf{e}_L^T\mathbf{y}\,, \tag{2.8}$$

$$y_R = \mathbf{e}_R^T\mathbf{y}\,, \tag{2.9}$$

where $\mathbf{e}_L$ and $\mathbf{e}_R$ denote M_H-dimensional selection vectors with all elements equal to 0, except one element equal to 1, i.e, $\mathbf{e}_L(1) = 1$ and $\mathbf{e}_R(M_L+1) = 1$. Using (2.5), (2.6), (2.7), (2.8) and (2.9), the left and the right reference microphone signals can be written as

$$y_L = x_L + \underbrace{u_L + n_L}_{v_L} = a_L s_x + b_L s_u + n_L\,, \tag{2.10}$$

$$y_R = x_R + \underbrace{u_R + n_R}_{v_R} = a_R s_x + b_R s_u + n_R\,. \tag{2.11}$$

The M_H-dimensional relative transfer function (RTF) vectors of the desired source $\mathbf{a}_L$ and $\mathbf{a}_R$ and the interfering source $\mathbf{b}_L$ and $\mathbf{b}_R$ are defined by relating the ATF vectors to the reference microphones [45, 46], i.e.,

$$\mathbf{a}_L = \frac{\mathbf{a}}{a_L}\,, \quad \mathbf{a}_R = \frac{\mathbf{a}}{a_R}\,, \tag{2.12}$$

$$\mathbf{b}_L = \frac{\mathbf{b}}{b_L}\,, \quad \mathbf{b}_R = \frac{\mathbf{b}}{b_R}\,. \tag{2.13}$$

The noisy input covariance matrix $\mathbf{R}_y \in \mathbb{C}^{M_H \times M_H}$, the desired source covariance matrix $\mathbf{R}_x \in \mathbb{C}^{M_H \times M_H}$, the interfering source covariance matrix $\mathbf{R}_u \in \mathbb{C}^{M_H \times M_H}$,

the noise covariance matrix $\mathbf{R}_n \in \mathbb{C}^{M_H \times M_H}$ and the undesired covariance matrix $\mathbf{R}_v \in \mathbb{C}^{M_H \times M_H}$ are defined as

$$\mathbf{R}_y = \mathcal{E}\{\mathbf{y}\mathbf{y}^H\}\,, \tag{2.14}$$

$$\mathbf{R}_x = \mathcal{E}\{\mathbf{x}\mathbf{x}^H\}\,, \tag{2.15}$$

$$\mathbf{R}_u = \mathcal{E}\{\mathbf{u}\mathbf{u}^H\}\,, \tag{2.16}$$

$$\mathbf{R}_n = \mathcal{E}\{\mathbf{n}\mathbf{n}^H\}\,, \tag{2.17}$$

$$\mathbf{R}_v = \mathcal{E}\{\mathbf{v}\mathbf{v}^H\}\,, \tag{2.18}$$

where $\mathcal{E}\{\cdot\}$ denotes the expected value operator and $(\cdot)^H$ denotes the conjugate transpose. Assuming statistical independence between the signal components $\mathbf{x}$, $\mathbf{u}$ and $\mathbf{n}$, the noisy input covariance matrix $\mathbf{R}_y$ can be written as

$$\mathbf{R}_y = \mathbf{R}_x + \underbrace{\mathbf{R}_u + \mathbf{R}_n}_{\mathbf{R}_v}\,. \tag{2.19}$$

Using (2.6), (2.7), (2.15) and (2.16), the desired source covariance matrix and the interfering source covariance matrix can be written as a rank-1 matrix, i.e.,

$$\mathbf{R}_x = p_{s_x}\mathbf{a}\mathbf{a}^H\,, \tag{2.20}$$

$$\mathbf{R}_u = p_{s_u}\mathbf{b}\mathbf{b}^H\,, \tag{2.21}$$

with $p_{s_x} = \mathcal{E}\{|s_x|^2\}$ the power spectral density (PSD) of the desired source and $p_{s_u} = \mathcal{E}\{|s_u|^2\}$ the PSD of the interfering source. The noise covariance matrix $\mathbf{R}_n$ and the undesired covariance matrix $\mathbf{R}_v$ are assumed to be full-rank, i.e., invertible and positive definite.

The PSD and the cross power spectral density (CPSD) of the desired source component in the reference microphone signals are equal to

$$p_{x_L} = \mathcal{E}\{|x_L|^2\} = \mathbf{e}_L^T\mathbf{R}_x\mathbf{e}_L = p_{s_x}|a_L|^2\,, \tag{2.22}$$

$$p_{x_R} = \mathcal{E}\{|x_R|^2\} = \mathbf{e}_R^T\mathbf{R}_x\mathbf{e}_R = p_{s_x}|a_R|^2\,, \tag{2.23}$$

$$p_{x_{LR}} = \mathcal{E}\{x_L x_R^*\} = \mathbf{e}_L^T\mathbf{R}_x\mathbf{e}_R = p_{s_x}a_L a_R^*\,, \tag{2.24}$$

where $(\cdot)^*$ denotes complex conjugation. Similarly, the PSD and the CPSD of the interfering source component in the reference microphone signals are equal to

$$p_{u_L} = \mathcal{E}\{|u_L|^2\} = \mathbf{e}_L^T\mathbf{R}_u\mathbf{e}_L = p_{s_u}|b_L|^2\,, \tag{2.25}$$

$$p_{u_R} = \mathcal{E}\{|u_R|^2\} = \mathbf{e}_R^T\mathbf{R}_u\mathbf{e}_R = p_{s_u}|b_R|^2\,, \tag{2.26}$$

$$p_{u_{LR}} = \mathcal{E}\{u_L u_R^*\} = \mathbf{e}_L^T\mathbf{R}_u\mathbf{e}_R = p_{s_u}b_L b_R^*\,. \tag{2.27}$$

Finally, the PSD and the CPSD of the noise component in the reference microphone signals are equal to

$$p_{n_L} = \mathcal{E}\{|n_L|^2\} = \mathbf{e}_L^T \mathbf{R}_n \mathbf{e}_L \,, \tag{2.28}$$

$$p_{n_R} = \mathcal{E}\{|n_R|^2\} = \mathbf{e}_R^T \mathbf{R}_n \mathbf{e}_R \,, \tag{2.29}$$

$$p_{n_{LR}} = \mathcal{E}\{n_L n_R^*\} = \mathbf{e}_L^T \mathbf{R}_n \mathbf{e}_R \,. \tag{2.30}$$

Using (2.12), (2.13), (2.22), (2.23), (2.25) and (2.26), the desired source covariance matrix and the interfering source covariance matrix in (2.20) and (2.21) can also be written in terms of RTF vectors, i.e.,

$$\mathbf{R}_x = p_{x_L} \mathbf{a}_L \mathbf{a}_L^H = p_{x_R} \mathbf{a}_R \mathbf{a}_R^H \,, \tag{2.31}$$

$$\mathbf{R}_u = p_{u_L} \mathbf{b}_L \mathbf{b}_L^H = p_{u_R} \mathbf{b}_R \mathbf{b}_R^H \,. \tag{2.32}$$

The narrowband input signal-to-noise ratio (SNR) in the left and the right reference microphone signal is defined as the ratio of the PSDs of the desired source and noise components, i.e.,

$$\mathrm{SNR}_L^{\mathrm{in}} = \frac{p_{x_L}}{p_{n_L}} \,, \tag{2.33}$$

$$\mathrm{SNR}_R^{\mathrm{in}} = \frac{p_{x_R}}{p_{n_R}} \,. \tag{2.34}$$

Similarly, the narrowband input signal-to-interference ratio (SIR) in the left and the right reference microphone signal is defined as the ratio of the PSDs of the desired source and interfering source components, i.e.,

$$\mathrm{SIR}_L^{\mathrm{in}} = \frac{p_{x_L}}{p_{u_L}} \,, \tag{2.35}$$

$$\mathrm{SIR}_R^{\mathrm{in}} = \frac{p_{x_R}}{p_{u_R}} \,. \tag{2.36}$$

For the binaural hearing device configuration in Figure 2.1, the output signals z_L and z_R of the left and the right hearing device are obtained by filtering and summing all head-mounted microphone signals, i.e.,

$$z_L = \mathbf{w}_L^H \mathbf{y} \,, \tag{2.37}$$

$$z_R = \mathbf{w}_R^H \mathbf{y} \,, \tag{2.38}$$

where $\mathbf{w}_L \in \mathbb{C}^{M_H}$ and $\mathbf{w}_R \in \mathbb{C}^{M_H}$ denote the left and the right filter vector, respectively.

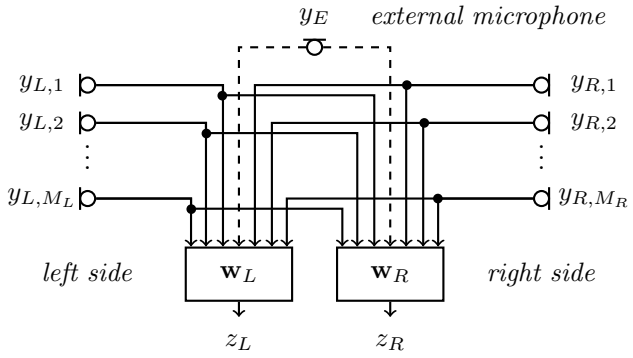

Fig. 2.2: Extended binaural hearing device configuration with M_L microphones on the left side, M_R microphones on the right side and one external microphone.

2.1.2 *Extended binaural hearing device configuration*

Figure 2.2 depicts the *extended* binaural hearing device configuration, incorporating $M_E = 1$ external microphone in addition to the head-mounted microphones. For the extended binaural hearing device configuration the total number of microphones is equal to $M = M_H + 1$. Similarly to (2.10) and (2.11), the external microphone signal y_E can be written as

$$y_E = x_E + \underbrace{u_E + n_E}_{v_E} = a_E s_x + b_E s_u + n_E \,, \tag{2.39}$$

where x_E, u_E, n_E and v_E denote the desired source component, the interfering source component, the noise component and the undesired component in the external microphone signal, respectively, while a_E and b_E denote the ATF between the desired source and the external microphone and the ATF between the interfering source and the external microphone, respectively. As mentioned in the introduction, we assume that the external microphone signal y_E is transmitted (e.g., via a wireless link) to the binaural hearing devices without any transmission delay and that the head-mounted microphone signals $\mathbf{y}$ and the external microphone signal y_E are perfectly synchronized. The M-dimensional extended noisy input vector is defined as

$$\mathbf{y}_e = \begin{bmatrix} \mathbf{y} \\ y_E \end{bmatrix} \in \mathbb{C}^M \,. \tag{2.40}$$

The extended desired source component $\mathbf{x}_e$, the extended interfering source component $\mathbf{u}_e$, the extended noise component $\mathbf{n}_e$, the extended undesired component $\mathbf{v}_e$, the extended ATF vector of the desired source $\mathbf{a}_e$ and the extended ATF vector of

the interfering source $\mathbf{b}_e$ are defined similarly as $\mathbf{y}_e$ in (2.40). Similarly to (2.8) and (2.9), the left and the right reference microphone signal can be selected as

$$y_L = \mathbf{e}_L^T \mathbf{y}_e \,, \tag{2.41}$$

$$y_R = \mathbf{e}_R^T \mathbf{y}_e \,, \tag{2.42}$$

where $\mathbf{e}_L$ and $\mathbf{e}_R$ now are M-dimensional selection vectors. Further, the external microphone signal can be selected as

$$y_E = \mathbf{e}_E^T \mathbf{y}_e \,, \tag{2.43}$$

where $\mathbf{e}_E$ is an M-dimensional selection vector with $\mathbf{e}_E(M) = 1$. Similarly to (2.12) and (2.13), the M-dimensional left and right extended RTF vectors of the desired source and the interfering source are defined as

$$\mathbf{a}_{L,e} = \frac{\mathbf{a}_e}{a_L} \,, \quad \mathbf{a}_{R,e} = \frac{\mathbf{a}_e}{a_R} \,, \tag{2.44}$$

$$\mathbf{b}_{L,e} = \frac{\mathbf{b}_e}{b_L} \,, \quad \mathbf{b}_{R,e} = \frac{\mathbf{b}_e}{b_R} \,. \tag{2.45}$$

Similarly to (2.15), (2.16), (2.17), (2.18), (2.31) and (2.32), the $M \times M$-dimensional extended desired source, interfering source, noise and undesired covariance matrices are defined as

$$\mathbf{R}_{x,e} = \mathcal{E}\{\mathbf{x}_e \mathbf{x}_e^H\} \;\; = p_{s_x} \mathbf{a}_e \mathbf{a}_e^H \;\; = p_{x_L} \mathbf{a}_{L,e} \mathbf{a}_{L,e}^H \;\; = p_{x_R} \mathbf{a}_{R,e} \mathbf{a}_{R,e}^H \,, \tag{2.46}$$

$$\mathbf{R}_{u,e} = \mathcal{E}\{\mathbf{u}_e \mathbf{u}_e^H\} \;\; = p_{s_u} \mathbf{b}_e \mathbf{b}_e^H = p_{u_L} \mathbf{b}_{L,e} \mathbf{b}_{L,e}^H = p_{u_R} \mathbf{b}_{R,e} \mathbf{b}_{R,e}^H \,, \tag{2.47}$$

$$\mathbf{R}_{n,e} = \mathcal{E}\{\mathbf{n}_e \mathbf{n}_e^H\} \,, \tag{2.48}$$

$$\mathbf{R}_{v,e} = \mathcal{E}\{\mathbf{v}_e \mathbf{v}_e^H\} \,. \tag{2.49}$$

Similarly to (2.19), the extended noisy input covariance matrix is defined as

$$\mathbf{R}_{y,e} = \mathcal{E}\{\mathbf{y}_e \mathbf{y}_e^H\} = \mathbf{R}_{x,e} + \underbrace{\mathbf{R}_{u,e} + \mathbf{R}_{n,e}}_{\mathbf{R}_{v,e}} \,. \tag{2.50}$$

The extended noise covariance matrix can be written as a block matrix, i.e.,

$$\mathbf{R}_{n,e} = \left[\begin{array}{c|c} \mathbf{R}_n & \mathbf{r}_{n,E} \\ \hline \mathbf{r}_{n,E}^H & p_{n_E} \end{array}\right] \,, \tag{2.51}$$

with

$$\mathbf{r}_{n,E} = \mathcal{E}\{\mathbf{n} n_E^*\} \in \mathbb{C}^{M_H} \,, \tag{2.52}$$

the cross correlation vector between the noise component in the head-mounted microphone signals and the noise component in the external microphone signal,

and $p_{n_E} = \mathcal{E}\{|n_E|^2\} = \mathbf{e}_E^T \mathbf{R}_{n,e} \mathbf{e}_E$ the PSD of the noise component in the external microphone signal.

The narrowband input SNR in the external microphone signal is defined as the ratio of the PSDs of the desired source and noise components, i.e.,

$$\mathrm{SNR}_E^{\mathrm{in}} = \frac{p_{x_E}}{p_{n_E}}\,, \tag{2.53}$$

with $p_{x_E} = \mathcal{E}\{|x_E|^2\} = \mathbf{e}_E^T \mathbf{R}_{x,e} \mathbf{e}_E = p_{s_x} |a_E|^2$ the PSD of the desired source component in the external microphone signal. Similarly, the narrowband input SIR in the external microphone signal is defined as the ratio of the PSDs of the desired source and interfering source components, i.e.,

$$\mathrm{SIR}_E^{\mathrm{in}} = \frac{p_{x_E}}{p_{u_E}}\,, \tag{2.54}$$

with $p_{u_E} = \mathcal{E}\{|u_E|^2\} = \mathbf{e}_E^T \mathbf{R}_{u,e} \mathbf{e}_E = p_{s_u} |b_E|^2$ the PSD of the interfering source component in the external microphone signal.

Similarly to (2.37) and (2.38), for the extended binaural hearing device configuration in Figure 2.2 the output signals z_L and z_R of the left and the right hearing device are obtained by filtering and summing all microphone signals, i.e., the head-mounted microphone signals and the external microphone signal, i.e.,

$$z_L = \mathbf{w}_L^H \mathbf{y}_e\,, \tag{2.55}$$

$$z_R = \mathbf{w}_R^H \mathbf{y}_e\,, \tag{2.56}$$

where $\mathbf{w}_L \in \mathbb{C}^M$ and $\mathbf{w}_R \in \mathbb{C}^M$ now are M-dimensional filter vectors.

2.1.3 *Multi-extended binaural hearing device configuration*

As a generalization of the extended binaural hearing device configuration, Figure 2.3 depicts the *multi-extended* binaural hearing device configuration, incorporating $M_E > 1$ external microphones in addition to the head-mounted microphones. For the multi-extended binaural hearing device configuration the total number of microphones is equal to $M = M_H + M_E$. Similarly to (2.39), the i-th external microphone signal $y_{E,i}$ can be written as

$$y_{E,i} = x_{E,i} + \underbrace{u_{E,i} + n_{E,i}}_{v_{E,i}} = a_{E,i} s_x + b_{E,i} s_u + n_{E,i}\,, \quad i \in \{1, \ldots, M_E\}\,, \tag{2.57}$$

where $x_{E,i}$, $u_{E,i}$, $n_{E,i}$ and $v_{E,i}$ denote the desired source component, the interfering source component, the noise component and the undesired component in the i-th external microphone signal, respectively, and $a_{E,i}$ and $b_{E,i}$ denote the ATF between the desired source and the i-th external microphone and the ATF between the interfering source and the i-th external microphone, respectively. For the sake of notational conciseness, we use the same notation for the multi-extended input vector,

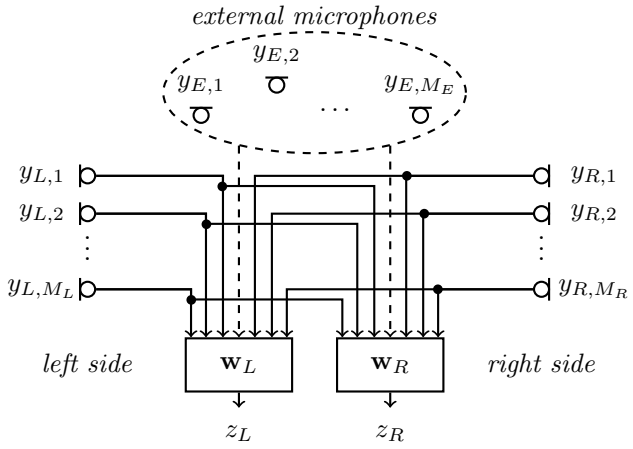

Fig. 2.3: Multi-extended binaural hearing device configuration consisting of M_L microphones on the left side, M_R microphones on the right side and M_E external microphones.

the multi-extended signal components and the multi-extended covariance matrices as for the extended binaural hearing device configuration with $M_E = 1$, e.g.,

$$\mathbf{y}_e = \left[\mathbf{y}^T,\, y_{E,1},\, \ldots,\, y_{E,M_E}\right]^T . \tag{2.58}$$

In this thesis it is clear from the context whether we are referring to an extended or a multi-extended binaural hearing device configuration. Similarly to (2.43), the i-th external microphone signal can be selected as

$$y_{E,i} = \mathbf{e}_{E,i}^T \mathbf{y}_e \,, \tag{2.59}$$

where $\mathbf{e}_{E,i}$ is an M-dimensional selection vector with $\mathbf{e}_{E,i}(M_H + i) = 1$.

The narrowband input SNR in the i-th external microphone signal is defined as the ratio of the PSDs of the desired source and noise components, i.e.,

$$\mathrm{SNR}_{E,i}^{\mathrm{in}} = \frac{p_{x_{E,i}}}{p_{n_{E,i}}} \,, \tag{2.60}$$

with $p_{x_{E,i}} = \mathcal{E}\{|x_{E,i}|^2\} = \mathbf{e}_{E,i}^T \mathbf{R}_{x,e} \mathbf{e}_{E,i} = p_{s_x} |a_{E,i}|^2$ and $p_{n_{E,i}} = \mathcal{E}\{|n_{E,i}|^2\} = \mathbf{e}_{E,i}^T \mathbf{R}_{n,e} \mathbf{e}_{E,i}$ the PSD of the desired source and noise components in the i-th external microphone signal. Similarly, the narrowband input SIR in the i-th external microphone signal is defined as the ratio of the PSDs of the desired source and interfering source components, i.e.,

$$\mathrm{SIR}_{E,i}^{\mathrm{in}} = \frac{p_{x_{E,i}}}{p_{u_{E,i}}} \,, \tag{2.61}$$

with $p_{u_{E,i}} = \mathcal{E}\{|u_{E,i}|^2\} = \mathbf{e}_{E,i}^T \mathbf{R}_{u,e} \mathbf{e}_{E,i} = p_{s_u} |b_{E,i}|^2$ the PSD of the interfering source component in the i-th external microphone signal.

As for the extended binaural hearing device configuration, for the multi-extended binaural hearing device configuration in Figure 2.3 the output signals z_L and z_R of the left and the right hearing device in (2.55) and (2.56) are obtained by filtering and summing all microphone signals, i.e., the head-mounted microphone signals and all external microphone signals.

2.2 Objective performance measures and binaural cues

The performance of the binaural beamforming algorithms presented in the next chapters are evaluated in terms of noise and interference reduction performance (Section 2.2.1) as well as binaural cue preservation for the different signal components (Section 2.2.2). Please note that all definitions of the objective performance measures and the binaural cues are given for the binaural hearing device configuration in Section 2.1.1. They can be similarly defined for the (multi-)extended binaural hearing device configuration using the (multi-)extended vectors and matrices, but are not given here for the sake of conciseness.

2.2.1 *Noise and interference reduction performance*

The output PSD of the desired source component in the left and the right output signal is defined as

$$p_{x_L}^{\text{out}} = \mathbf{w}_L^T \mathbf{R}_x \mathbf{w}_L = p_{s_x} \left| \mathbf{w}_L^H \mathbf{a} \right|^2 , \tag{2.62}$$

$$p_{x_R}^{\text{out}} = \mathbf{w}_R^T \mathbf{R}_x \mathbf{w}_R = p_{s_x} \left| \mathbf{w}_R^H \mathbf{a} \right|^2 . \tag{2.63}$$

Similarly, the output PSD of the interfering source component in the left and the right output signal is defined as

$$p_{u_L}^{\text{out}} = \mathbf{w}_L^T \mathbf{R}_u \mathbf{w}_L = p_{s_u} \left| \mathbf{w}_L^H \mathbf{b} \right|^2 , \tag{2.64}$$

$$p_{u_R}^{\text{out}} = \mathbf{w}_R^T \mathbf{R}_u \mathbf{w}_R = p_{s_u} \left| \mathbf{w}_R^H \mathbf{b} \right|^2 . \tag{2.65}$$

The output PSD of the noise component in the left and the right output signal is defined as

$$p_{n_L}^{\text{out}} = \mathbf{w}_L^T \mathbf{R}_n \mathbf{w}_L , \tag{2.66}$$

$$p_{n_R}^{\text{out}} = \mathbf{w}_R^T \mathbf{R}_n \mathbf{w}_R . \tag{2.67}$$

The narrowband output SNR in the left and the right output signal is defined as the ratio of the PSDs of the desired source and noise components, i.e.,

$$\mathrm{SNR}_L^{\mathrm{out}} = \frac{p_{x_L}^{\mathrm{out}}}{p_{n_L}^{\mathrm{out}}}\,, \tag{2.68}$$

$$\mathrm{SNR}_R^{\mathrm{out}} = \frac{p_{x_R}^{\mathrm{out}}}{p_{n_R}^{\mathrm{out}}}\,. \tag{2.69}$$

The left and the right SNR improvement (in dB) is hence given by

$$\Delta\mathrm{SNR}_L = 10\log_{10}\mathrm{SNR}_L^{\mathrm{out}} - 10\log_{10}\mathrm{SNR}_L^{\mathrm{in}}\,, \tag{2.70}$$

$$\Delta\mathrm{SNR}_R = 10\log_{10}\mathrm{SNR}_R^{\mathrm{out}} - 10\log_{10}\mathrm{SNR}_R^{\mathrm{in}}\,, \tag{2.71}$$

with $\mathrm{SNR}_L^{\mathrm{in}}$ and $\mathrm{SNR}_R^{\mathrm{in}}$ defined in (2.33) and (2.34). The narrowband output SIR in the left and the right output signal is defined as the ratio of the PSDs of the desired source and interfering source components, i.e.,

$$\mathrm{SIR}_L^{\mathrm{out}} = \frac{p_{x_L}^{\mathrm{out}}}{p_{u_L}^{\mathrm{out}}}\,, \tag{2.72}$$

$$\mathrm{SIR}_R^{\mathrm{out}} = \frac{p_{x_R}^{\mathrm{out}}}{p_{u_R}^{\mathrm{out}}}\,. \tag{2.73}$$

The left and the right SIR improvement (in dB) is hence given by

$$\Delta\mathrm{SIR}_L = 10\log_{10}\mathrm{SIR}_L^{\mathrm{out}} - 10\log_{10}\mathrm{SIR}_L^{\mathrm{in}}\,, \tag{2.74}$$

$$\Delta\mathrm{SIR}_R = 10\log_{10}\mathrm{SIR}_R^{\mathrm{out}} - 10\log_{10}\mathrm{SIR}_R^{\mathrm{in}}\,, \tag{2.75}$$

with $\mathrm{SIR}_L^{\mathrm{in}}$ and $\mathrm{SIR}_R^{\mathrm{in}}$ defined in (2.35) and (2.36).

2.2.2 *Binaural cues*

For the coherent sources in the considered acoustic scenario (i.e., the desired source and the interfering source) the main binaural cues that can be used to describe the spatial impression (cf. Section 1.2.1) are the interaural time difference (ITD) and the interaural level difference (ILD) [48], which can be computed from the so-called interaural transfer function (ITF). Using (2.20) and (2.21), the input ITFs of the desired source and the interfering source are given by [10]

$$\mathrm{ITF}_x^{\mathrm{in}} = \frac{\mathcal{E}\{|x_L|^2\}}{\mathcal{E}\{x_R x_L^*\}} = \frac{\mathbf{e}_L^T\mathbf{R}_x\mathbf{e}_L}{\mathbf{e}_R^T\mathbf{R}_x\mathbf{e}_L} = \frac{a_L}{a_R}\,, \tag{2.76}$$

$$\mathrm{ITF}_u^{\mathrm{in}} = \frac{\mathcal{E}\{|u_L|^2\}}{\mathcal{E}\{u_R u_L^*\}} = \frac{\mathbf{e}_L^T\mathbf{R}_u\mathbf{e}_L}{\mathbf{e}_R^T\mathbf{R}_u\mathbf{e}_L} = \frac{b_L}{b_R}\,. \tag{2.77}$$

Similarly, the output ITFs of the desired source and the interfering source are given by

$$\mathrm{ITF}_x^{\mathrm{out}} = \frac{\mathbf{w}_L^H \mathbf{R}_x \mathbf{w}_L}{\mathbf{w}_R^T \mathbf{R}_x \mathbf{w}_L} = \frac{\mathbf{w}_L^H \mathbf{a}}{\mathbf{w}_R^H \mathbf{a}}\,, \tag{2.78}$$

$$\mathrm{ITF}_u^{\mathrm{out}} = \frac{\mathbf{w}_L^H \mathbf{R}_u \mathbf{w}_L}{\mathbf{w}_R^T \mathbf{R}_u \mathbf{w}_L} = \frac{\mathbf{w}_L^H \mathbf{b}}{\mathbf{w}_R^H \mathbf{b}}\,. \tag{2.79}$$

From the ITF the ITD and ILD cues can be calculated as [10]

$$\mathrm{ITD} = \frac{\angle \mathrm{ITF}}{\omega}\,, \tag{2.80}$$

$$\mathrm{ILD} = |\mathrm{ITF}|^2\,, \tag{2.81}$$

with ω the angular frequency and where $\angle(\cdot)$ denotes the unwrapped phase.

For an incoherent sound field (i.e., background noise), the ITD and ILD cues are not very descriptive, but the interaural coherence (IC) is known to play a major role for the spatial impression (e.g., spatial width or diffuseness) [48]. The input IC of the noise component is defined as

$$\mathrm{IC}_n^{\mathrm{in}} = \frac{\mathcal{E}\{n_L n_R^*\}}{\sqrt{\mathcal{E}\{|n_L|^2\}}\sqrt{\mathcal{E}\{|n_R|^2\}}} = \frac{\mathbf{e}_L^T \mathbf{R}_n \mathbf{e}_R}{\sqrt{\mathbf{e}_L^T \mathbf{R}_n \mathbf{e}_L}\sqrt{\mathbf{e}_R^T \mathbf{R}_n \mathbf{e}_R}}\,, \tag{2.82}$$

while the output IC of the noise component is defined as

$$\mathrm{IC}_n^{\mathrm{out}} = \frac{\mathbf{w}_L^T \mathbf{R}_n \mathbf{w}_R}{\sqrt{\mathbf{w}_L^T \mathbf{R}_n \mathbf{w}_L}\sqrt{\mathbf{w}_R^T \mathbf{R}_n \mathbf{w}_R}}\,. \tag{2.83}$$

Because the IC is typically complex-valued, the magnitude-squared coherence (MSC) is often used. The MSC is defined as the square of the absolute value of the IC, i.e.,

$$\mathrm{MSC}_n^{\mathrm{in}} = \left|\mathrm{IC}_n^{\mathrm{in}}\right|^2\,, \quad \mathrm{MSC}_n^{\mathrm{out}} = \left|\mathrm{IC}_n^{\mathrm{out}}\right|^2\,. \tag{2.84}$$

For the (coherent) desired and interfering sources the input IC can be defined similarly to (2.82). It can be shown using (2.20) and (2.21) that the input ICs for the desired source and the interfering source are equal to [11, 106]

$$\mathrm{IC}_x^{\mathrm{in}} = \frac{\mathbf{e}_L^T \mathbf{R}_x \mathbf{e}_R}{\sqrt{\mathbf{e}_L^T \mathbf{R}_x \mathbf{e}_L}\sqrt{\mathbf{e}_R^T \mathbf{R}_x \mathbf{e}_R}} = \frac{a_L a_R^*}{|a_L||a_R|} = e^{j\,a_L/a_R}\,, \tag{2.85}$$

$$\mathrm{IC}_u^{\mathrm{in}} = \frac{\mathbf{e}_L^T \mathbf{R}_u \mathbf{e}_R}{\sqrt{\mathbf{e}_L^T \mathbf{R}_u \mathbf{e}_L}\sqrt{\mathbf{e}_R^T \mathbf{R}_u \mathbf{e}_R}} = \frac{b_L b_R^*}{|b_L||b_R|} = e^{j\,b_L/b_R}\,, \tag{2.86}$$

such that their input MSCs are equal to 1, i.e.,

$$\mathrm{MSC}_x^{\mathrm{in}} = 1\,, \tag{2.87}$$

$$\mathrm{MSC}_u^{\mathrm{in}} = 1\,. \tag{2.88}$$

An MSC of 1 is perceived as a distinct point source, while smaller MSC values lead to a broader or even diffuse sound field perception [48]. The output IC of the desired source and the interfering source can be defined similarly to (2.83), i.e.,

$$\mathrm{IC}_x^{\mathrm{out}} = \frac{\mathbf{w}_L^T \mathbf{R}_x \mathbf{w}_R}{\sqrt{\mathbf{w}_L^T \mathbf{R}_x \mathbf{w}_L}\sqrt{\mathbf{w}_R^T \mathbf{R}_x \mathbf{w}_R}}\,, \tag{2.89}$$

$$\mathrm{IC}_u^{\mathrm{out}} = \frac{\mathbf{w}_L^T \mathbf{R}_u \mathbf{w}_R}{\sqrt{\mathbf{w}_L^T \mathbf{R}_u \mathbf{w}_L}\sqrt{\mathbf{w}_R^T \mathbf{R}_u \mathbf{w}_R}}\,. \tag{2.90}$$

2.3 Summary

In this chapter the signal model was introduced for three configurations that are used in the remainder of this thesis to derive and analyze binaural beamforming algorithms with and without external microphones: the binaural hearing device configuration using only the head-mounted microphones, the extended binaural hearing device configuration incorporating one external microphone, and the multi-extended binaural hearing device configuration incorporating multiple external microphones. Furthermore, we introduced the mathematical definitions of the objective performance measures and binaural cues that are used to evaluate the noise and interference reduction performance (i.e., speech intelligibility) and binaural cue preservation capabilities (i.e., spatial impression) of the considered algorithms.

3

BINAURAL BEAMFORMING ALGORITHMS AND PARAMETER ESTIMATION METHODS

In this chapter we briefly review three state-of-the-art binaural beamforming algorithms using only the head-mounted microphone signals, i.e., using the binaural hearing device configuration discussed in Section 2.1.1. Section 3.1 reviews the frequently-used binaural minimum variance distortionless response (BMVDR) beamformer that preserves the binaural cues of the desired source but distorts the binaural cues of the interfering source and the background noise. Section 3.2 reviews the binaural linearly constrained minimum variance (BLCMV) beamformer that uses an additional constraint to preserve the binaural cues of the interfering source. Section 3.3 reviews the BMVDR beamformer with partial noise estimation (BMVDR-N) that provides a trade-off between noise reduction performance and binaural cue preservation of the (background) noise. Section 3.4 discusses several methods to model or estimate the parameters that are required to implement the aforementioned binaural beamforming algorithms, more in particular, covariance matrices and ATF or RTF vectors.

3.1 Binaural Minimum Variance Distortionless Response (BMVDR) beamformer

The BMVDR beamformer aims at minimizing the PSD of the noise component in the output signals (2.66) and (2.67) while preserving the desired source component in the left and the right reference microphone signals. The constrained optimization problems for the left and the right filter vector $\mathbf{w}_L$ and $\mathbf{w}_R$ in (2.37) and (2.38) are given by [2, 3, 10]

$$\boxed{\min_{\mathbf{w}_L} \mathcal{E}\{|\mathbf{w}_L^H \mathbf{n}|^2\} \quad \text{subject to} \quad \mathbf{w}_L^H \mathbf{x} = x_L} \tag{3.1}$$

$$\boxed{\min_{\mathbf{w}_R} \mathcal{E}\{|\mathbf{w}_R^H \mathbf{n}|^2\} \quad \text{subject to} \quad \mathbf{w}_R^H \mathbf{x} = x_R} \tag{3.2}$$

Using (2.6) and (2.17), the left and the right filter vector of the BMVDR beamformer are given by [2, 46, 65]

$$\boxed{\mathbf{w}_{\text{BMVDR},L} = \frac{\mathbf{R}_n^{-1}\mathbf{a}}{\gamma_a} a_L^*} \tag{3.3}$$

$$\boxed{\mathbf{w}_{\text{BMVDR},R} = \frac{\mathbf{R}_n^{-1}\mathbf{a}}{\gamma_a} a_R^*} \tag{3.4}$$

with

$$\gamma_a = \mathbf{a}^H \mathbf{R}_n^{-1} \mathbf{a}\,. \tag{3.5}$$

The filter vectors in (3.3) and (3.4) can also be written using the RTF vectors of the desired source in (2.12), i.e.,

$$\mathbf{w}_{\text{BMVDR},L} = \frac{\mathbf{R}_n^{-1}\mathbf{a}_L}{\mathbf{a}_L^H \mathbf{R}_n^{-1} \mathbf{a}_L}\,, \tag{3.6}$$

$$\mathbf{w}_{\text{BMVDR},R} = \frac{\mathbf{R}_n^{-1}\mathbf{a}_R}{\mathbf{a}_R^H \mathbf{R}_n^{-1} \mathbf{a}_R}\,. \tag{3.7}$$

The parameters required to calculate the filter vectors of the BMVDR beamformer are hence the noise covariance matrix $\mathbf{R}_n$ and either the ATF vector of the desired source $\mathbf{a}$ or the RTF vectors of the desired source $\mathbf{a}_L$ and $\mathbf{a}_R$.

Please note that the BMVDR beamformer can also be defined as minimizing the PSD of the (overall) undesired component, i.e., using the undesired covariance matrix $\mathbf{R}_v$ instead of the noise covariance matrix $\mathbf{R}_n$. Alternatively, it is also possible to use the noisy input covariance matrix $\mathbf{R}_y$ instead of $\mathbf{R}_n$, which is referred to as the minimum power distortionless response (MPDR) beamformer [46, 65]. Since $\mathbf{R}_v$ is considerably more difficult to model or estimate in practice than $\mathbf{R}_n$ (cf. Section 3.4) and the MPDR beamformer may lead to severe target cancellation effects due to parameter estimation errors [78], in this thesis we only consider the BMVDR beamformer using $\mathbf{R}_n$.

By substituting (3.3) and (3.4) in (2.68) and (2.69), it has been shown in [3, 10] that the output SNR of the BMVDR beamformer for both the left and the right hearing device is equal to

$$\boxed{\rho = \text{SNR}^{\text{out}}_{\text{BMVDR},L} = \text{SNR}^{\text{out}}_{\text{BMVDR},R} = p_{s_x}\gamma_a} \tag{3.8}$$

which is always larger than or equal to the input SNR in (2.33) and (2.34). As can be observed from (3.8) and (3.5), ρ depends on the ATF vector $\mathbf{a}$ (i.e., the position of the desired source and the head-mounted microphones), the noise covariance matrix $\mathbf{R}_n$ and the PSD p_{s_x} of the desired source.

By substituting (3.3) and (3.4) in (2.72) and (2.73), it has been shown in [11] that the output SIR of the BMVDR beamformer for both the left and the right hearing device is equal to

$$\mathrm{SIR}^{\mathrm{out}}_{\mathrm{BMVDR},L} = \mathrm{SIR}^{\mathrm{out}}_{\mathrm{BMVDR},R} = \frac{p_{s_x}}{p_{s_u}} \frac{|\gamma_a|^2}{|\gamma_{ab}|^2}\,, \tag{3.9}$$

with γ_a defined in (3.5) and

$$\gamma_{ab} = \mathbf{a}^H \mathbf{R}_n^{-1} \mathbf{b}\,. \tag{3.10}$$

Although the BMVDR beamformer yields the largest output SNR among all distortionless binaural beamforming algorithms and typically also suppresses the interfering source to some extent, it does not allow to explicitly control the interference reduction. As can be seen from (3.9) and (3.10), the output SIR of the BMVDR beamformer depends on the ATF vectors $\mathbf{a}$ and $\mathbf{b}$ (i.e., the position of the desired and interfering sources and the head-mounted microphones), the noise covariance matrix $\mathbf{R}_n$ and the PSDs p_{s_x} and p_{s_u} of the desired and the interfering source. The interference reduction performance of the BMVDR beamformer hence depends on the relative position of the interfering source to the desired source. If both sources are co-located (i.e., $\mathbf{b} = \mathbf{a}$ and hence $\gamma_{ab} = \gamma_a$), the BMVDR beamformer obviously would not be able to perform any interference reduction.

As has been shown in [3, 10, 106], the BMVDR beamformer preserves the binaural cues of the desired source, i.e.,

$$\mathrm{ITF}^{\mathrm{out}}_{\mathrm{BMVDR},x} = \frac{a_L}{a_R} = \mathrm{ITF}^{\mathrm{in}}_{x}\,, \tag{3.11}$$

but distorts the binaural cues of the undesired sources, i.e., for the interfering source

$$\mathrm{ITF}^{\mathrm{out}}_{\mathrm{BMVDR},u} = \frac{a_L}{a_R} = \mathrm{ITF}^{\mathrm{in}}_{x}\,, \tag{3.12}$$

and for the background noise

$$\mathrm{IC}^{\mathrm{out}}_{\mathrm{BMVDR},n} = \mathrm{IC}^{\mathrm{in}}_{x} = e^{j\,a_L/a_R}\,, \tag{3.13}$$

$$\mathrm{MSC}^{\mathrm{out}}_{\mathrm{BMVDR},n} = 1\,. \tag{3.14}$$

Hence, at the output of the BMVDR beamformer both the interfering source and the background noise are perceived as coming from the direction of the desired source, which is obviously undesired in terms of sound quality and spatial awareness.

3.2 Binaural Linearly Constrained Minimum Variance (BLCMV) beamformer

In addition to preserving the desired source component in the reference microphone signals, the BLCMV beamformer preserves a scaled version of the interfering source

component in the reference microphone signals while minimizing the PSD of the noise component in the output signals [12, 105]. The constrained optimization problems for the left and the right filter vector are given by [12]

$$\boxed{\min_{\mathbf{w}_L} \mathcal{E}\{|\mathbf{w}_L^H \mathbf{n}|^2\} \quad \text{subject to} \quad \mathbf{w}_L^H \mathbf{x} = x_L\,,\ \mathbf{w}_L^H \mathbf{u} = \delta u_L} \tag{3.15}$$

$$\boxed{\min_{\mathbf{w}_R} \mathcal{E}\{|\mathbf{w}_R^H \mathbf{n}|^2\} \quad \text{subject to} \quad \mathbf{w}_R^H \mathbf{x} = x_R\,,\ \mathbf{w}_R^H \mathbf{u} = \delta u_R} \tag{3.16}$$

where $\delta \in \mathbb{R}$ denotes the interference scaling parameter, with $0 < \delta \leq 1$. Using (2.6), (2.7), (2.10), (2.11) and (2.17), the left and the right filter vector of the BLCMV beamformer are given by [12]

$$\boxed{\mathbf{w}_{\text{BLCMV},L} = \mathbf{R}_n^{-1}\mathbf{C}\left(\mathbf{C}^H\mathbf{R}_n^{-1}\mathbf{C}\right)^{-1}\mathbf{g}_L} \tag{3.17}$$

$$\boxed{\mathbf{w}_{\text{BLCMV},R} = \mathbf{R}_n^{-1}\mathbf{C}\left(\mathbf{C}^H\mathbf{R}_n^{-1}\mathbf{C}\right)^{-1}\mathbf{g}_R} \tag{3.18}$$

with the constraint matrix $\mathbf{C}$ and the left and the right response vector $\mathbf{g}_L$ and $\mathbf{g}_R$ defined as

$$\mathbf{C} = [\mathbf{a}\,,\ \mathbf{b}]\,, \tag{3.19}$$

$$\mathbf{g}_L = \begin{bmatrix} a_L^* \\ \delta b_L^* \end{bmatrix}\,, \tag{3.20}$$

$$\mathbf{g}_R = \begin{bmatrix} a_R^* \\ \delta b_R^* \end{bmatrix}\,. \tag{3.21}$$

The filter vectors in (3.17) and (3.18) can also be written using the RTF vectors of the desired source in (2.12) and the RTF vectors of the interfering source in (2.13) as [12]

$$\mathbf{w}_{\text{BLCMV},L} = \mathbf{R}_n^{-1}\mathbf{C}_L\left(\mathbf{C}_L^H\mathbf{R}_n^{-1}\mathbf{C}_L\right)^{-1}\mathbf{g}\,, \tag{3.22}$$

$$\mathbf{w}_{\text{BLCMV},R} = \mathbf{R}_n^{-1}\mathbf{C}_R\left(\mathbf{C}_R^H\mathbf{R}_n^{-1}\mathbf{C}_R\right)^{-1}\mathbf{g}\,, \tag{3.23}$$

with the response vector $\mathbf{g}$ and the left and the right constraint matrices $\mathbf{C}_L$ and $\mathbf{C}_R$ defined as

$$\mathbf{g} = \begin{bmatrix} 1 \\ \delta \end{bmatrix}\,, \tag{3.24}$$

$$\mathbf{C}_L = [\mathbf{a}_L,\, \mathbf{b}_L]\,, \tag{3.25}$$

$$\mathbf{C}_R = [\mathbf{a}_R,\, \mathbf{b}_R]\,. \tag{3.26}$$

The parameters required to calculate the filter vectors of the BLCMV beamformer are hence the noise covariance matrix $\mathbf{R}_n$, the ATF vector $\mathbf{a}$ or the RTF vectors $\mathbf{a}_L$ and $\mathbf{a}_R$ of the desired source, the ATF vector $\mathbf{b}$ or the RTF vectors $\mathbf{b}_L$ and $\mathbf{b}_R$ of the interfering source, and the interference scaling parameter δ. Please note that like the BMVDR beamformer the BLCMV beamformer could be defined using the undesired covariance matrix $\mathbf{R}_v$ or the noisy input covariance matrix $\mathbf{R}_y$ instead of the noise covariance matrix $\mathbf{R}_n$ [12]. For the same reasons as for the BMVDR beamformer and because it has been shown in [163] that using $\mathbf{R}_n$ is beneficial in practical scenarios with estimation errors, in this thesis we only consider the BLCMV beamformer using $\mathbf{R}_n$.

By substituting (3.17) and (3.18) in (2.68) and (2.69) it has been shown in [12] that the left and the right output SNR of the BLCMV beamformer are equal to

$$\mathrm{SNR}^{\mathrm{out}}_{\mathrm{BLCMV},L} = \frac{p_{s_x}|a_L|^2}{\mathbf{e}_L^T\mathbf{R}_{xu,1}\mathbf{e}_L}\,, \tag{3.27}$$

$$\mathrm{SNR}^{\mathrm{out}}_{\mathrm{BLCMV},R} = \frac{p_{s_x}|a_R|^2}{\mathbf{e}_R^T\mathbf{R}_{xu,1}\mathbf{e}_R}\,, \tag{3.28}$$

with

$$\mathbf{R}_{xu,1} = \frac{1}{1-\Psi}\left[\frac{\mathbf{a}\mathbf{a}^H}{\gamma_a} + \delta^2\frac{\mathbf{b}\mathbf{b}^H}{\gamma_b} - 2\Psi\delta\Re\left\{\frac{\mathbf{a}\mathbf{b}^H}{\gamma_{ab}^*}\right\}\right]\,, \tag{3.29}$$

$$\gamma_b = \mathbf{b}^H\mathbf{R}_n^{-1}\mathbf{b}\,, \quad \Psi = \frac{|\gamma_{ab}|^2}{\gamma_a\gamma_b}\,, \tag{3.30}$$

where $\Re\{\cdot\}$ denotes the real part of a complex number. The left and the right output SNR of the BLCMV beamformer in (3.27) and (3.28) are both smaller than or equal to the output SNR of the BMVDR beamformer in (3.8), since less degrees of freedom are available for noise reduction. In addition, it has been shown in [12] that the left and the right output SIR of the BLCMV beamformer are equal to

$$\mathrm{SIR}^{\mathrm{out}}_{\mathrm{BLCMV},L} = \mathrm{SIR}^{\mathrm{in}}_{L}\frac{1}{\delta^2}\,, \tag{3.31}$$

$$\mathrm{SIR}^{\mathrm{out}}_{\mathrm{BLCMV},R} = \mathrm{SIR}^{\mathrm{in}}_{R}\frac{1}{\delta^2}\,. \tag{3.32}$$

This means that the interference reduction of the BLCMV beamformer can be directly controlled by the interference scaling parameter δ, which is also intuitively clear from the constraints in (3.15) and (3.16).

It can also be directly seen from the constraints in (3.15) and (3.16) that the BLCMV beamformer preserves the binaural cues of both the desired source and the interfering source, i.e.,

$$\mathrm{ITF}^{\mathrm{out}}_{\mathrm{BLCMV},x} = \frac{a_L}{a_R} = \mathrm{ITF}^{\mathrm{in}}_{x}\,, \tag{3.33}$$

$$\mathrm{ITF}^{\mathrm{out}}_{\mathrm{BLCMV},u} = \frac{b_L}{b_R} = \mathrm{ITF}^{\mathrm{in}}_{u}\,. \tag{3.34}$$

In addition, it has been shown in [12] that the output IC of the noise component for the BLCMV beamformer is equal to

$$\mathrm{IC}^{\mathrm{out}}_{\mathrm{BLCMV},n} = \frac{\mathbf{e}_L^T\mathbf{R}_{xu,1}\mathbf{e}_R}{\sqrt{\mathbf{e}_L^T\mathbf{R}_{xu,1}\mathbf{e}_L}\sqrt{\mathbf{e}_R^T\mathbf{R}_{xu,1}\mathbf{e}_R}}\,. \tag{3.35}$$

Substituting (3.35) in (2.84), the output MSC of the noise component for the BLCMV beamformer is equal to

$$\mathrm{MSC}^{\mathrm{out}}_{\mathrm{BLCMV},n} = \frac{\left|\mathbf{e}_L^T\mathbf{R}_{xu,1}\mathbf{e}_R\right|^2}{\left(\mathbf{e}_L^T\mathbf{R}_{xu,1}\mathbf{e}_L\right)\left(\mathbf{e}_R^T\mathbf{R}_{xu,1}\mathbf{e}_R\right)}\,. \tag{3.36}$$

Because $\mathbf{R}_{xu,1}$ in (3.29) is a rank-2 matrix, it has been shown in [12] that the output MSC of the noise component is smaller than 1 but is not equal to the input MSC of the noise component. Furthermore, it should be noted that the output MSC of the noise component depends on the relative position of the interfering source to the desired source, cf. (3.29) and (3.30), such that it is not straightforward to control the binaural cues of the background noise.

3.3 BMVDR beamformer with partial noise estimation (BMVDR-N)

In addition to preserving the desired source component in the reference microphone signals, the BMVDR beamformer with partial noise estimation (BMVDR-N) aims at preserving a scaled version of the noise component in the reference microphone signals [3, 9, 10, 13]. The constrained optimization problems for the left and the right filter vector are given by

$$\boxed{\min_{\mathbf{w}_L} \mathcal{E}\left\{\left|\mathbf{w}_L^H\mathbf{n} - \eta n_L\right|^2\right\} \quad \text{subject to} \quad \mathbf{w}_L^H\mathbf{x} = x_L} \tag{3.37}$$

$$\boxed{\min_{\mathbf{w}_R} \mathcal{E}\left\{\left|\mathbf{w}_R^H\mathbf{n} - \eta n_R\right|^2\right\} \quad \text{subject to} \quad \mathbf{w}_R^H\mathbf{x} = x_R} \tag{3.38}$$

where $\eta \in \mathbb{R}$ denotes the mixing parameter, with $0 \leq \eta \leq 1$. It has been shown in [10, 13] that the resulting left and right filter vectors can be written as

$$\boxed{\mathbf{w}_{\mathrm{BMVDR-N},L} = (1-\eta)\mathbf{w}_{\mathrm{BMVDR},L} + \eta\mathbf{e}_L} \tag{3.39}$$

$$\boxed{\mathbf{w}_{\mathrm{BMVDR-N},R} = (1-\eta)\mathbf{w}_{\mathrm{BMVDR},R} + \eta\mathbf{e}_R} \tag{3.40}$$

with $\mathbf{w}_{\mathrm{BMVDR},L}$ and $\mathbf{w}_{\mathrm{BMVDR},R}$ defined in (3.3) and (3.4). Hence, the output signals of the BMVDR-N beamformer can be interpreted as a mixture between the noisy reference microphone signals (scaled with η) and the output signals of the BMVDR beamformer (scaled with $1-\eta$). For $\eta = 0$, the BMVDR-N beamformer in (3.39) and (3.40) is equal to the BMVDR beamformer in (3.3) and (3.4). For $\eta = 1$, the output signals are equal to the reference microphone signals, i.e., no beamforming is applied at all. The parameters required to calculate the filter vectors of the BMVDR-N beamformer are the mixing parameter η and the same parameters as for the BMVDR beamformer in Section 3.1.

In [10, 13] it has been shown that the left and the right output SNR of the BMVDR-N beamformer are equal to

$$\mathrm{SNR}^{\mathrm{out}}_{\mathrm{BMVDR-N},L} = \frac{\rho}{1+\eta^2\left(\frac{\rho}{\mathrm{SNR}^{\mathrm{in}}_L}-1\right)} \leq \rho\,, \tag{3.41}$$

$$\mathrm{SNR}^{\mathrm{out}}_{\mathrm{BMVDR-N},R} = \frac{\rho}{1+\eta^2\left(\frac{\rho}{\mathrm{SNR}^{\mathrm{in}}_R}-1\right)} \leq \rho\,, \tag{3.42}$$

with the output SNR of the BMVDR beamformer ρ defined in (3.8), such that

$$\mathrm{SNR}^{\mathrm{out}}_{\mathrm{BMVDR-N},L} \leq \mathrm{SNR}^{\mathrm{out}}_{\mathrm{BMVDR},L}\,, \tag{3.43}$$

$$\mathrm{SNR}^{\mathrm{out}}_{\mathrm{BMVDR-N},R} \leq \mathrm{SNR}^{\mathrm{out}}_{\mathrm{BMVDR},R}\,. \tag{3.44}$$

Since $\rho \geq \mathrm{SNR}^{\mathrm{in}}_L$ and $\rho \geq \mathrm{SNR}^{\mathrm{in}}_R$, it can easily be seen that (3.41) and (3.42) monotonically decrease with increasing η, i.e.,

$$\frac{\partial \mathrm{SNR}^{\mathrm{out}}_{\mathrm{BMVDR-N},L}}{\partial \eta} \leq 0\,, \tag{3.45}$$

$$\frac{\partial \mathrm{SNR}^{\mathrm{out}}_{\mathrm{BMVDR-N},R}}{\partial \eta} \leq 0\,, \tag{3.46}$$

such that a larger mixing parameter η leads to a smaller output SNR of the BMVDR-N beamformer [10, 13]. By substituting (3.39) and (3.40) in (2.72) and (2.73), it

can be shown that the left and the right output SIR of the BMVDR-N beamformer are equal to

$$\mathrm{SIR}^{\mathrm{out}}_{\mathrm{BMVDR-N},L} = \frac{p_{s_x}}{p_{s_u}} \frac{|a_L|^2}{\mathbf{e}_L^T \mathbf{R}_{xu,2} \mathbf{e}_L}\,, \tag{3.47}$$

$$\mathrm{SIR}^{\mathrm{out}}_{\mathrm{BMVDR-N},R} = \frac{p_{s_x}}{p_{s_u}} \frac{|a_R|^2}{\mathbf{e}_R^T \mathbf{R}_{xu,2} \mathbf{e}_R}\,, \tag{3.48}$$

with

$$\mathbf{R}_{xu,2} = (1-\eta)^2 \frac{|\gamma_{ab}|^2}{|\gamma_a|^2} \mathbf{a}\mathbf{a}^H + \eta^2 \mathbf{b}\mathbf{b}^H + (\eta - \eta^2) 2\Re\{\mathbf{a}\mathbf{b}^H \frac{\gamma_{ab}}{\gamma_a}\}\,, \tag{3.49}$$

with γ_a and γ_{ab} defined in (3.5) and (3.10), respectively. For $\eta = 0$, the left and the right output SIR of the BMVDR-N beamformer are equal to the left and the right output SIR of the BMVDR beamformer in (3.9). For $\eta = 1$, $\mathbf{R}_{xu,2} = \mathbf{b}\mathbf{b}^H$ and the left and the right output SIR of the BMVDR-N beamformer are equal to the left and the right input SIR in (2.35) and (2.36). As can be seen from (3.47), (3.48) and (3.49), the interference reduction of the BMVDR-N beamformer depends on the relative position of the interfering source to the desired source such that it is not straightforward to control using the mixing parameter η.

Similarly to the BMVDR and the BLCMV beamformer, the BMVDR-N beamformer preserves the binaural cues of the desired source [10, 13], i.e.,

$$\mathrm{ITF}^{\mathrm{out}}_{\mathrm{BMVDR-N},x} = \frac{a_L}{a_R} = \mathrm{ITF}^{\mathrm{in}}_x\,. \tag{3.50}$$

By substituting (3.39) and (3.40) in (2.79), it has been shown in [13, 164] that the output ITF of the interfering source is equal to

$$\mathrm{ITF}^{\mathrm{out}}_{\mathrm{BMVDR-N},u} = \frac{(1-\eta) a_L \frac{\gamma_{ab}}{\gamma_a} + \eta b_L}{(1-\eta) a_R \frac{\gamma_{ab}}{\gamma_a} + \eta b_R}\,. \tag{3.51}$$

As can be observed, for $\eta = 1$ the binaural cues of the interfering source are preserved, whereas for $\eta = 0$ the binaural cues of the interfering source are equal to the binaural cues of the desired source (as for the BMVDR beamformer). Due to γ_{ab} in (3.51), the output ITF of the interfering source for the BMVDR-N beamformer depends on the relative position of the interfering source to the desired source, such that it is not straightforward to control using the mixing parameter η, like the left and the right output SIR of the BMVDR-N beamformer.

By substituting (3.39) and (3.40) in (2.82) and (2.84), it has been shown in [13] that the output IC and the output MSC of the noise component for the BMVDR-N beamformer are equal to

$$\mathrm{IC}^{\mathrm{out}}_{\mathrm{BMVDR-N},n} = \frac{\frac{1-\eta^2}{\rho} p_{s_x} a_L a_R^* + \eta^2 p_{n_{LR}}}{\sqrt{\left(\frac{1-\eta^2}{\rho} p_{s_x} |a_L|^2 + \eta^2 p_{n_L}\right)\left(\frac{1-\eta^2}{\rho} p_{s_x} |a_R|^2 + \eta^2 p_{n_R}\right)}} \tag{3.52}$$

$$\mathrm{MSC}^{\mathrm{out}}_{\mathrm{BMVDR-N},n} = \frac{\left|\frac{1-\eta^2}{\rho} p_{s_x} a_L a_R^* + \eta^2 p_{n_{LR}}\right|^2}{\left(\frac{1-\eta^2}{\rho} p_{s_x} |a_L|^2 + \eta^2 p_{n_L}\right)\left(\frac{1-\eta^2}{\rho} p_{s_x} |a_R|^2 + \eta^2 p_{n_R}\right)}, \tag{3.53}$$

with p_{n_L}, p_{n_R} and $p_{n_{LR}}$ defined in (2.28), (2.29) and (2.30). For $\eta = 1$, the output MSC of the noise component in (3.53) is equal to the input MSC of the noise component. For $\eta = 0$, the output MSC of the noise component is equal to 1. Since a larger mixing parameter leads to a better MSC preservation of the noise component (i.e., a better preservation of the spatial impression) but a smaller output SNR, the mixing parameter allows to trade off between noise reduction performance and binaural cue preservation of the noise component.

In order to achieve a desired output MSC $\mathrm{MSC}^{\mathrm{des}}_n$ of the noise component, with

$$0 \leq \mathrm{MSC}^{\mathrm{in}}_n \leq \mathrm{MSC}^{\mathrm{des}}_n \leq 1\,, \tag{3.54}$$

in [13] a closed-form expression for the mixing parameter η^{des} has been derived, i.e.,

$$\eta^{\mathrm{des}} = \sqrt{\frac{\rho\left(\sqrt{\gamma^2 - \alpha\beta} - \gamma\right) + \alpha}{\rho^2\beta - 2\rho\gamma + \alpha}}\,, \tag{3.55}$$

with

$$\alpha = \left(\mathrm{MSC}^{\mathrm{des}}_n - 1\right) p^2_{s_x} |a_L|^2 |a_R|^2\,, \tag{3.56}$$

$$\beta = \left(\mathrm{MSC}^{\mathrm{des}}_n - \mathrm{MSC}^{\mathrm{in}}_n\right) p_{n_L} p_{n_R}\,, \tag{3.57}$$

$$\gamma = \mathrm{MSC}^{\mathrm{des}}_n \frac{p_{s_x} |a_L|^2 p_{n_L} + p_{s_x} |a_R|^2 p_{n_R}}{2} - \Re\{p_{s_x} a_L a_R^* p^*_{n_{LR}}\}\,. \tag{3.58}$$

Since $\mathrm{MSC}^{\mathrm{in}}_n \leq \mathrm{MSC}^{\mathrm{des}}_n \leq 1$ and all PSDs are positive (or zero), it can be easily seen that $\alpha \leq 0$ and $\beta \geq 0$. Aiming for the spatial impression of the noise component in the reference microphone signals and the noise component in the output signals of the BMVDR-N beamformer to be indistinguishable, it has been proposed in [13, 106] to define the desired output MSC of the noise component based on the IC discrimination ability of the human auditory system [165, 166]. A perceptual evaluation of the BMVDR-N beamformer using such psycho-acoustically motivated mixing parameters can be found in [113], showing that partially preserving the IC of a diffuse noise field, by using the BMVDR-N beamformer, significantly improves

spatial quality compared with the BMVDR beamformer while only marginally affecting speech intelligibility.

3.4 Parameter modelling and estimation

As shown in the previous sections, the parameters required to calculate the filter vectors of the binaural beamforming algorithms considered in this thesis are 1) covariance matrices and 2) steering vectors, i.e., ATF or RTF vectors of a coherent source (e.g., the desired source or the interfering source). In this section we discuss several methods to model or estimate these parameters in practice. Section 3.4.1 considers modelling and estimating covariance matrices, while Section 3.4.2 considers modelling and estimating steering vectors. The choice whether to use modelled or estimated parameters depends on the trade-off between estimation errors and the validity of the model assumptions for the considered acoustic scenario.

3.4.1 *Covariance matrices*

MODELLING OF COVARIANCE MATRICES On the one hand, the covariance matrices of the coherent sources (i.e., desired source and interfering source) in (2.20) and (2.21) may be highly time-varying both spectrally as well as spatially (moving sources). On the other hand, it can typically be assumed that the noise covariance matrix $\mathbf{R}_n$ in (2.17) is spatially more stationary and hence easier to model. Assuming a homogeneous noise field, where the PSDs of the noise component in all microphone signals are equal, the noise covariance matrix can be modelled as

$$\mathbf{R}_n = p_n \mathbf{\Gamma}\,, \tag{3.59}$$

with p_n the (possibly time-varying) PSD of the noise component in all microphone signals and $\mathbf{\Gamma} \in \mathbb{C}^{M_H \times M_H}$ the (time-invariant) spatial coherence matrix of the noise. Using (3.59) for the calculation of the filter vectors of the BMVDR beamformer in (3.3) and (3.4), it can be observed that the filter vectors are independent of p_n and hence only the spatial coherence matrix $\mathbf{\Gamma}$ is required, i.e.,

$$\mathbf{w}_{\text{BMVDR},L} = \frac{\mathbf{\Gamma}^{-1}\mathbf{a}}{\mathbf{a}^H\mathbf{\Gamma}^{-1}\mathbf{a}} a_L^*\,, \quad \mathbf{w}_{\text{BMVDR},R} = \frac{\mathbf{\Gamma}^{-1}\mathbf{a}}{\mathbf{a}^H\mathbf{\Gamma}^{-1}\mathbf{a}} a_R^*\,. \tag{3.60}$$

This also holds for the filter vectors of the BLCMV beamformer in (3.17) and (3.18) and the BMVDR-N beamformer in (3.39) and (3.40). In the following we discuss three common ways to model the spatial coherence matrix $\mathbf{\Gamma}$. First, assuming spatially white noise, which is frequently used to model sensor noise, the spatial coherence matrix is equal to the M_H-dimensional identity matrix $\mathbf{I}_{M_H}$, i.e.,

$$\mathbf{\Gamma}^{\text{white}} = \mathbf{I}_{M_H}\,. \tag{3.61}$$

Second, assuming diffuse background noise (more in particular a spherically isotropic noise field), which is frequently used to model multi-talker babble noise,

the (p,q)-th element of the spatial coherence matrix can be modelled, assuming no head between the microphones, as [32]

$$\mathbf{\Gamma}^{\text{diff}}(p,q) = \text{sinc}\left(\frac{\omega d_{p,q}}{c}\right), \tag{3.62}$$

with ω the angular frequency, $d_{p,q}$ the distance between the p-th and the q-th microphone and c the speed of sound. However, since it should be taken into account that the binaural hearing devices are head-mounted, a modified sinc-function should be used instead of (3.62) [11, 167], i.e.,

$$\mathbf{\Gamma}^{\text{diff,mod}}(p,q) = \text{sinc}\left(\alpha\frac{\omega d_{p,q}}{c}\right)\frac{1}{\sqrt{1+(\beta\frac{\omega d_{p,q}}{c})^4}}, \tag{3.63}$$

with $\alpha = 2.2$ and $\beta = 0.5$. Since the relative positions of the head-mounted microphones are (roughly) fixed, the spatial coherence matrix of the noise can be assumed stationary such that $\mathbf{\Gamma}^{\text{diff}}$ can be used as a time-invariant parameter to calculate the filter vectors of the binaural beamforming algorithms considered in this thesis. Third, if a database of (measured or simulated) ATFs is available (e.g., [35, 37]), the (p,q)-th element of the spatial coherence matrix can be modelled as [11–13]

$$\mathbf{\Gamma}^{\text{dat}}(p,q) = \frac{\sum_{k=1}^{K} h_p(\theta_k)h_q^*(\theta_k)}{\sqrt{\sum_{k=1}^{K}|h_p(\theta_k)|^2}\sqrt{\sum_{k=1}^{K}|h_q(\theta_k)|^2}}, \tag{3.64}$$

with $h(\theta_k)$ the anechoic ATF at angle θ_k and K the total number of angles in the database.

ESTIMATION OF COVARIANCE MATRICES To estimate covariance matrices, we distinguish between *batch* and *online* estimation. Batch estimation refers to utilizing the complete signal (i.e., all time frames t), hence yielding a time-invariant covariance matrix. Batch estimation should therefore only be performed if the acoustic scenario can be assumed to be spatially stationary. The batch estimate of the noisy input covariance matrix $\mathbf{R}_y$ can be calculated as

$$\hat{\mathbf{R}}_y^{\text{bat}}(f) = \frac{1}{|\mathcal{Y}|}\sum_{t\in\mathcal{Y}}\mathbf{y}(f,t)\mathbf{y}^H(f,t), \tag{3.65}$$

with $\mathcal{Y}$ the set of time frames where all sources are active and $|\mathcal{Y}|$ its cardinality. Similarly, the batch estimate of the noise covariance matrix $\mathbf{R}_n$ can be calculated as

$$\hat{\mathbf{R}}_n^{\text{bat}}(f) = \frac{1}{|\mathcal{N}|}\sum_{t\in\mathcal{N}}\mathbf{y}(f,t)\mathbf{y}^H(f,t), \tag{3.66}$$

with $\mathcal{N}$ the set of time frames where only the background noise is active and $|\mathcal{N}|$ its cardinality. The batch estimate of the undesired covariance matrix $\mathbf{R}_v$ can be calculated as

$$\hat{\mathbf{R}}_v^{\text{bat}}(f) = \frac{1}{|\mathcal{V}|} \sum_{t \in \mathcal{V}} \mathbf{y}(f,t)\mathbf{y}^H(f,t)\,, \tag{3.67}$$

with $\mathcal{V}$ the set of time frames where only the undesired sources (i.e., interfering source and background noise) are active and $|\mathcal{V}|$ its cardinality. Please note that the batch estimates of the desired source covariance matrix $\hat{\mathbf{R}}_x^{\text{bat}}$ and the interfering source covariance matrix $\hat{\mathbf{R}}_u^{\text{bat}}$ can be calculated similarly, but have rather theoretical relevance, since in practice it does not frequently happen that there are time frames where only the desired source of the interfering source are active.

Contrary to batch estimation, online estimation refers to utilizing current and past time frames to adaptively estimate a covariance matrix. Online estimation should therefore be performed when the acoustic scenario is time-varying, e.g., due to head movements or movements of the sources, and is typically implemented by recursive averaging [46]. If all sources are active in the current time frame, i.e., $t \in \mathcal{Y}$, the noisy input covariance matrix $\mathbf{R}_y$ can be recursively updated as

$$\hat{\mathbf{R}}_y^{\text{onl}}(f,t) = \alpha_y \hat{\mathbf{R}}_y(f,t-1) + (1-\alpha_y)\mathbf{y}(f,t)\mathbf{y}^H(f,t)\,, \tag{3.68}$$

with α_y the smoothing factor for the noisy input covariance matrix. Note that if $t \notin \mathcal{Y}$, typically the last estimate is used, i.e., $\hat{\mathbf{R}}_y^{\text{onl}}(f,t) = \hat{\mathbf{R}}_y^{\text{onl}}(f,t-1)$. If only the background noise is active in the current time frame, i.e., $t \in \mathcal{N}$, the noise covariance matrix $\mathbf{R}_n$ can be recursively updated as

$$\hat{\mathbf{R}}_n^{\text{onl}}(f,t) = \alpha_n \hat{\mathbf{R}}_n(f,t-1) + (1-\alpha_n)\mathbf{y}(f,t)\mathbf{y}^H(f,t)\,, \tag{3.69}$$

with α_n the smoothing factor for the noise covariance matrix and if $t \notin \mathcal{N}$, $\hat{\mathbf{R}}_n^{\text{onl}}(f,t) = \hat{\mathbf{R}}_n^{\text{onl}}(f,t-1)$. If only the undesired sources (i.e., interfering source and background noise) are active in the current time frame, i.e., $t \in \mathcal{V}$, the undesired covariance matrix $\mathbf{R}_v$ can be recursively updated as

$$\hat{\mathbf{R}}_v^{\text{onl}}(f,t) = \alpha_v \hat{\mathbf{R}}_v(f,t-1) + (1-\alpha_v)\mathbf{y}(f,t)\mathbf{y}^H(f,t)\,, \tag{3.70}$$

with α_v the smoothing factor for the undesired covariance matrix and if $t \notin \mathcal{V}$, $\hat{\mathbf{R}}_v^{\text{onl}}(f,t) = \hat{\mathbf{R}}_v^{\text{onl}}(f,t-1)$. The smoothing factors can be related to a time constant as

$$\alpha = e^{-\frac{T_s}{f_s \tau}}\,, \tag{3.71}$$

with T_s the frame shift in the STFT framework, f_s the sampling rate and τ the time constant. The time constant τ should be in the range of the assumed spatial stationarity of the respective signal component.

VOICE ACTIVITY DETECTOR (VAD) To estimate the covariance matrices, the time frames need to be classified into the different sets $\mathcal{Y}$, $\mathcal{N}$ and $\mathcal{V}$. In the time-domain, a common approach is to use a binary voice activity detector (VAD) [19–23], which indicates whether a sample or time frame consists of speech-plus-noise or noise-only. Assuming for now that no interfering source is present in the acoustic scene, the objective of a VAD is to indicate whether the desired source is active, i.e., $t \in \mathcal{Y}$ or not, i.e., $t \in \mathcal{N}$, in a specific sample or time frame. Figure 3.1 depicts an exemplary VAD output. As can be observed, short pauses occur between spoken words, allowing the noise covariance matrix only to be updated during these pauses. Contrary to a binary VAD in the time-domain, high-resolution speech presence probability (SPP) estimators in both time- and frequency-domain have been proposed [24–27]. An SPP estimate close to 1 indicates speech presence, whereas an SPP estimate close to 0 indicates speech absence. From an SPP estimate a high-resolution VAD in time- and frequency-domain can be obtained by applying upper and lower boundaries, i.e.,

$$\mathrm{VAD}(f,t) = \begin{cases} 1\,, & \text{if } \mathrm{SPP}(f,t) > \mathrm{SPP}_{\mathrm{upper}} \\ 0\,, & \text{if } \mathrm{SPP}(f,t) \leq \mathrm{SPP}_{\mathrm{lower}} \\ 0.5\,, & \text{else} \end{cases}, \tag{3.72}$$

where $\mathrm{SPP}_{\mathrm{upper}}$ denotes the upper SPP boundary and $\mathrm{SPP}_{\mathrm{lower}}$ denotes the lower SPP boundary. This approach also enables to define a region of uncertainty, i.e., $\mathrm{VAD}(f,t) = 0.5$, where no covariance matrix is updated. Figure 3.2 exemplarily depicts an SPP estimate (top) and the resulting high-resolution VAD (bottom) for a speech source in diffuse noise (broadband SNR of 5 dB) using an implementation of [27] with $\mathrm{SPP}_{\mathrm{upper}} = 0.7$ and $\mathrm{SPP}_{\mathrm{lower}} = 0.5$. Other approaches that, e.g., use a soft weighting in (3.68), (3.69) and (3.70) can be found in [46, 168].

It should be noted that when an interfering speech source is present in the acoustic scene, VADs or SPP estimators usually cannot distinguish between the desired speech source and the interfering speech source. As mentioned before, estimating the undesired covariance matrix $\mathbf{R}_v$ is hence a very difficult task in practice. Nevertheless, several approaches have been proposed to distinguish between multiple speech sources, e.g., based on position [169, 170] or based on auditory attention decoding [171–175].

3.4.2 *Steering vectors (ATF and RTF vectors)*

For the ATF and RTF vectors of the desired source and the interfering source, two approaches can be considered. In the first approach, the ATFs and RTFs are modelled by *anechoic* ATFs and RTFs, for which the relative position of the microphones needs to be either known or fixed. In the second approach, the *reverberant* RTFs are directly estimated from the microphone signals. It should be noted that we do not consider blind ATF estimation techniques [176–181] in this thesis, since accurately

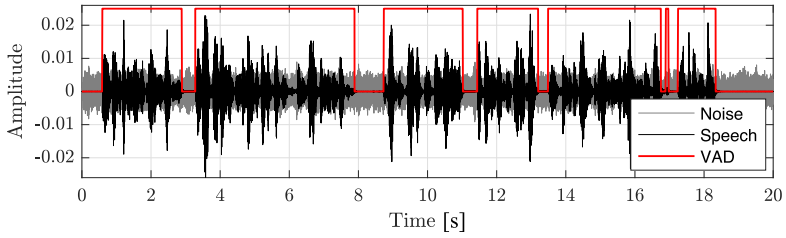

Fig. 3.1: Exemplary illustration of a VAD.

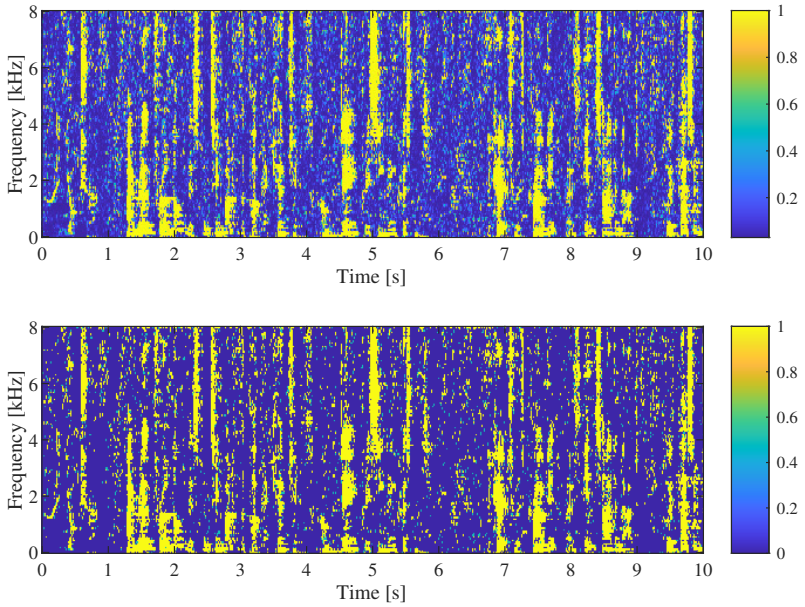

Fig. 3.2: Exemplary illustration of an SPP estimate (top) and the resulting high-resolution VAD (bottom) with $\mathrm{SPP}_{\mathrm{upper}} = 0.7$ and $\mathrm{SPP}_{\mathrm{lower}} = 0.5$.

estimating ATFs has proven to be very difficult in practice, especially for dynamic acoustic scenarios with background noise.

MODELLING OF ATF AND RTF VECTORS When modelling the ATFs and RTFs of a source using *anechoic* ATFs and RTFs, it is often assumed that these anechoic ATFs and RTFs are completely determined by the DOA of the source (assuming that the source is in the far-field and in the horizontal plane). The anechoic ATFs and RTFs can either be analytically computed based on a (spherical) head model [40] or selected from a database with measured binaural room impulse responses (BRIRs) for a specific microphone configuration [35, 37], which are transformed to the frequency-domain. In many hearing aid applications, it is simply assumed that the desired source is located in front of the listener. Since this assumption obviously does not always hold in practice, several approaches have been

proposed to estimate the DOA of one or multiple speakers from the binaural microphone signals, e.g., based on the dual delay line approach [69, 182], by using an auditory model [55], by modelling binaural cues using a Gaussian mixture model [183], by using blindly estimated impulse responses with an ITD/ILD model [70], by using blindly estimated RTFs [72, 74], by using a beamforming-based approach [73, 184], or by using classification-based methods [71, 185].

It should be realized that anechoic ATFs and RTFs obviously only contain the direct path of the source (including the shadowing of the head, cf. Section 1.2.1) but no reverberation. Another drawback is that the relative positions of the microphones need to be either known (when using an analytical model) or fixed (when using a database with measured BRIRs). This means that for the (multi-)extended binaural hearing device configuration with one or more external microphones, the usage of modelled ATFs and RTFs is typically not possible for the external microphones, since their relative position to the head-mounted microphones is usually not known and may even be time-varying. Assuming that the DOA of the desired source is known a-priori, in [16] methods to incorporate multiple external microphones in a GSC structure were proposed, whereby the GSC structure was based on the head-mounted microphones and could not be changed. The results in [16] are particularly interesting if the configuration to be extended cannot be changed and if the DOA of the desired source can be assumed a-priori (e.g., in front of the listener).

ESTIMATION OF RTF VECTORS Compared to using modelled RTFs, directly estimating the RTFs from the microphone signals has two main advantages. First, the relative positions of the microphones do not have to be known or fixed and can even be time-varying. Estimating the RTFs hence enables to incorporate one or more external microphones at unknown positions, compared to using modelled/measured RTFs. Second, estimated RTFs include early reflections or even late reverberation (depending on the time frame size of the STFT framework), which, e.g., leads to a more natural and wider perception for the desired source when used to steer the BMVDR beamformer. In this thesis we hence mainly consider direct estimation of the RTF vectors.

In the following we assume an acoustic scenario without an interfering source†, i.e., $\mathbf{R}_y = \mathbf{R}_x + \mathbf{R}_n$ and briefly review several state-of-the-art RTF vector estimation methods to estimate the RTF vectors of the desired source $\mathbf{a}_L$ and $\mathbf{a}_R$ in (2.12). Note that all methods discussed in this section can also be used to estimate the extended RTF vectors, e.g., $\mathbf{a}_{L,e}$ and $\mathbf{a}_{R,e}$ in (2.44). RTF vector estimation methods typically require estimates of the noisy input covariance matrix $\hat{\mathbf{R}}_y$ and the noise covariance

† While there are many RTF vector estimation methods available for a single source, it is not straightforward to jointly estimate the RTF vectors of multiple simultaneously active sources (e.g., the desired source and the interfering source).

matrix $\hat{\mathbf{R}}_n$ (see Section 3.4.1). Using (2.31), it can be easily shown that the left and the right RTF vector of the desired source in (2.12) can be written as

$$\mathbf{a}_L = \frac{\mathbf{R}_x \mathbf{e}_m}{\mathbf{e}_L^T \mathbf{R}_x \mathbf{e}_m}\,, \tag{3.73}$$

$$\mathbf{a}_R = \frac{\mathbf{R}_x \mathbf{e}_m}{\mathbf{e}_R^T \mathbf{R}_x \mathbf{e}_m}\,, \tag{3.74}$$

with $m \in \{1, \ldots, M_H\}$, i.e., selecting *any* column of the desired source covariance matrix $\mathbf{R}_x$ and normalizing by the element corresponding to the left or the right reference microphone. Even though it is not required, the column corresponding to the left and the right reference microphone is often used in the literature, i.e, $\mathbf{e}_m = \mathbf{e}_L$ in (3.73) and $\mathbf{e}_m = \mathbf{e}_R$ in (3.74).

Because the desired source covariance matrix $\mathbf{R}_x$ is unavailable in practice, biased RTF vector estimates can be obtained by using an estimate of the noisy input covariance matrix $\hat{\mathbf{R}}_y$, i.e.,

$$\boxed{\hat{\mathbf{a}}_L^{\mathrm{B}} = \frac{\hat{\mathbf{R}}_y \mathbf{e}_L}{\mathbf{e}_L^T \hat{\mathbf{R}}_y \mathbf{e}_L}} \tag{3.75}$$

$$\boxed{\hat{\mathbf{a}}_R^{\mathrm{B}} = \frac{\hat{\mathbf{R}}_y \mathbf{e}_R}{\mathbf{e}_R^T \hat{\mathbf{R}}_y \mathbf{e}_R}} \tag{3.76}$$

e.g., either using the batch estimate $\hat{\mathbf{R}}_y^{\mathrm{bat}}$ in (3.65) or the online estimate $\hat{\mathbf{R}}_y^{\mathrm{onl}}$ in (3.68).

To reduce the bias, several methods have been proposed [85, 87, 89, 90, 93, 117]. In the *covariance subtraction* (CS) method [87, 90, 93] the desired source covariance matrix is first estimated as

$$\hat{\mathbf{R}}_x = \hat{\mathbf{R}}_y - \hat{\mathbf{R}}_n\,, \tag{3.77}$$

e.g., using the estimation methods discussed in Section 3.4.1. The RTF vectors of the desired source are then estimated as [87, 90, 93]

$$\boxed{\hat{\mathbf{a}}_L^{\mathrm{CS}} = \frac{\hat{\mathbf{R}}_x \mathbf{e}_L}{\mathbf{e}_L^T \hat{\mathbf{R}}_x \mathbf{e}_L}} \tag{3.78}$$

$$\boxed{\hat{\mathbf{a}}_R^{\mathrm{CS}} = \frac{\hat{\mathbf{R}}_x \mathbf{e}_R}{\mathbf{e}_R^T \hat{\mathbf{R}}_x \mathbf{e}_R}} \tag{3.79}$$

Since $\hat{\mathbf{R}}_x$ typically is not a rank-1 matrix, it has been proposed in [89, 90] to estimate the RTF vectors of the desired source as the principal eigenvector $\mathbf{p}\{\hat{\mathbf{R}}_x\}$

(corresponding to the largest eigenvalue), normalized by the element corresponding to the reference microphone, i.e.,

$$\boxed{\hat{\mathbf{a}}_L^{\mathrm{CS-R1}} = \frac{\mathbf{p}\{\hat{\mathbf{R}}_x\}}{\mathbf{e}_L^T\mathbf{p}\{\hat{\mathbf{R}}_x\}}} \tag{3.80}$$

$$\boxed{\hat{\mathbf{a}}_R^{\mathrm{CS-R1}} = \frac{\mathbf{p}\{\hat{\mathbf{R}}_x\}}{\mathbf{e}_R^T\mathbf{p}\{\hat{\mathbf{R}}_x\}}} \tag{3.81}$$

We refer to this method as the *CS method with rank-1 approximation* (CS-R1). It has been shown in [89] that the CS-R1 method outperforms the CS method, but obviously has a larger computational complexity due to the eigenvalue decomposition (EVD).

In the *covariance whitening* (CW) method, first a square-root decomposition (e.g., Cholesky decomposition [186]) of the estimated noise covariance matrix $\hat{\mathbf{R}}_n$ is computed, i.e.,

$$\hat{\mathbf{R}}_n = \hat{\mathbf{R}}_n^{H/2}\hat{\mathbf{R}}_n^{1/2}\,. \tag{3.82}$$

The estimated noisy input covariance matrix $\hat{\mathbf{R}}_y$ is then pre-whitened as

$$\hat{\mathbf{R}}_y^{\mathrm{w}} = \hat{\mathbf{R}}_n^{-1/2}\hat{\mathbf{R}}_y\hat{\mathbf{R}}_n^{-H/2}\,. \tag{3.83}$$

Based on the principal eigenvector $\mathbf{p}\{\hat{\mathbf{R}}_y^{\mathrm{w}}\}$, the RTF vectors of the desired source can be estimated as [85, 89, 90, 93, 117]

$$\boxed{\hat{\mathbf{a}}_L^{\mathrm{CW}} = \frac{\hat{\mathbf{R}}_n^{1/2}\mathbf{p}\{\hat{\mathbf{R}}_y^{\mathrm{w}}\}}{\mathbf{e}_L^T\hat{\mathbf{R}}_n^{1/2}\mathbf{p}\{\hat{\mathbf{R}}_y^{\mathrm{w}}\}}} \tag{3.84}$$

$$\boxed{\hat{\mathbf{a}}_R^{\mathrm{CW}} = \frac{\hat{\mathbf{R}}_n^{1/2}\mathbf{p}\{\hat{\mathbf{R}}_y^{\mathrm{w}}\}}{\mathbf{e}_R^T\hat{\mathbf{R}}_n^{1/2}\mathbf{p}\{\hat{\mathbf{R}}_y^{\mathrm{w}}\}}} \tag{3.85}$$

A performance analysis and comparison between the CS and the CW method can be found in [90, 93]. The results show that the CW method generally outperforms the CS method but comes with a significantly larger computational complexity due to the square-root decomposition and especially the EVD. Aiming at reducing the computational complexity, iterative methods for RTF vector estimation have been proposed by using the power iteration method (or von-Mises-Iteration) to calculate the principal eigenvector $\mathbf{p}\{\hat{\mathbf{R}}_x\}$ [92] or $\mathbf{p}\{\hat{\mathbf{R}}_y^{\mathrm{w}}\}$ [86]. The RTF vector estimation methods based on the *power method* (PM) are referred to as PM-CS and PM-CW, respectively. As mentioned in [86, 92], one iteration per frame is typically sufficient for an online implementation.

3.5 Summary

In this chapter we reviewed three state-of-the-art binaural beamforming algorithms that are used throughout this thesis and discussed noise and interference reduction performance and binaural cue preservation capabilities. First, we reviewed the frequently-used BMVDR beamformer, which preserves the binaural cues of the desired source but distorts the binaural cues of the undesired sources (i.e., interfering source and background noise). The BMVDR beamformer provides the best noise reduction performance among all considered distortionless binaural beamforming algorithms, but the interference reduction depends on the relative position of the interfering source to the desired source and is not controllable. Second, we reviewed the BLCMV beamformer, which preserves the binaural cues of both the desired source and the interfering source but distorts the binaural cues of the background noise, depending on the relative position of the interfering source to the desired source. Since less degrees of freedom are available for noise reduction, the BLCMV beamformer typically yields a lower noise reduction performance than the BMVDR beamformer, but enables to directly control the amount of interference reduction by means of an interference scaling parameter. Third, we reviewed the BMVDR-N beamformer, which allows to trade off between noise reduction performance and binaural cue preservation of the noise component by mixing the noisy reference microphone signals with the output signals of the BMVDR beamformer using a mixing parameter. While the BMVDR-N beamformer hence enables to control the background noise component in the output signals, the interference reduction and the binaural cue preservation of the interfering source depend on the relative position of the interfering source to the desired source and are not straightforward to control using the mixing parameter. We further reviewed several methods to model or estimate the parameters that are required to calculate the filter vectors of the considered binaural beamforming algorithms in practice, more in particular covariance matrices and RTF vectors. Using estimates of the RTF vectors, the relative positions of the microphones do not have to be known or fixed and can even be time-varying. Estimating the RTF vectors hence enables to incorporate one or more external microphones at unknown positions. The state-of-the-art RTF vector estimation methods that are used in the remainder of the thesis require estimates of both the noisy input covariance matrix and the noise covariance matrix. The CW method generally outperforms the CS method but comes with a significantly larger computational complexity, especially since the CW method requires an EVD.

In Chapter 4 we combine the advantages of the BLCMV beamformer and the BMVDR-N beamformer and propose the BLCMV beamformer with partial noise estimation (BLCMV-N). It is shown that the proposed BLCMV-N beamformer is able to preserve the binaural cues of both the desired source and the interfering source and further enables to trade off between noise reduction performance and binaural cue preservation of the noise component. In Chapter 5 we analyze the BMVDR beamformer and the BMVDR-N beamformer for the extended binaural hearing device configuration and show that incorporating an external microphone is beneficial both in terms of noise reduction performance and binaural cue preservation of the

noise component. Finally, in Chapter 6 we propose computationally efficient methods (i.e., not requiring an EVD) to estimate the RTF vectors of the desired source that only require an estimate of the noisy input covariance matrix by exploiting one or more external microphones and show that in practice the performance is similar (or even better) than the performance of the CW method.

4

BLCMV BEAMFORMER WITH PARTIAL NOISE ESTIMATION (BLCMV-N)

As discussed in Chapter 3, the BMVDR beamformer provides a good noise reduction performance and preserves the binaural cues of the desired source, but it does not allow to control the reduction of the interfering source and distorts the binaural cues of the undesired sources (interfering source and background noise). As the first extension discussed in Section 3.2, the binaural linearly constrained minimum variance (BLCMV) beamformer uses an additional interference reduction constraint, enabling to control the reduction of the interfering source while preserving the binaural cues of the interfering source in addition to the desired source by means of an interference scaling parameter [12, 105]. However, due to the additional constraint there are less degrees of freedom available for noise reduction, such that the noise reduction performance for the BLCMV beamformer is lower than for the BMVDR beamformer. Furthermore, it is not possible to explicitly trade off between noise reduction performance and binaural cue preservation of the background noise. As the second extension discussed in Section 3.3, the BMVDR beamformer with partial noise estimation (BMVDR-N) aims for the noise component in the output signals to be equal to a scaled version of the noise component in the reference microphone signals while preserving the desired source component in the reference microphone signals [3, 9, 10, 13]. It has been shown that the output signals of the BMVDR-N beamformer can be interpreted as a mixture between the output signals of the BMVDR beamformer and the noisy reference microphone signals, i.e., the BMVDR-N beamformer provides a trade-off between noise reduction performance and binaural cue preservation of the background noise. While for (incoherent) background noise the BMVDR-N beamformer showed promising results [13, 113], the effect of partial noise estimation on a (coherent) interfering source depends on the relative position of the interfering source to the desired source and is harder to control [10].

This chapter is partly based on:

[156] N. Gößling, E. Hadad, S. Gannot and S. Doclo, "Binaural LCMV beamforming with partial noise estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* in press, 2020.

Aiming at merging the advantages of the BLCMV beamformer and the BMVDR-N beamformer, i.e., preserving the binaural cues of the interfering source and controlling the reduction of the interfering source as well as the binaural cues of the background noise, in this chapter we propose the BLCMV beamformer with partial noise estimation (BLCMV-N). Compared to the BMVDR beamformer, the BLCMV-N beamformer uses an additional constraint to preserve a scaled version of the interfering source component in the reference microphone signals (like the BLCMV beamformer) and aims at preserving a scaled version of the noise component in the reference microphone signals. First, we derive two decompositions for the BLCMV-N beamformer which reveal differences and similarities between the BLCMV-N beamformer and the BLCMV beamformer. We show that the output signals of the BLCMV-N beamformer can be interpreted as a mixture between the noisy reference microphone signals and the output signals of a BLCMV beamformer using an adjusted interference scaling parameter. We then analytically derive the performance of the BLCMV-N beamformer in terms of noise and interference reduction performance and binaural cue preservation. We show that the output SNR of the BLCMV-N beamformer is smaller than or equal to the output SNR of the BLCMV beamformer and derive the optimal interference scaling parameter maximizing the output SNR of the BLCMV-N beamformer. The derived analytical expressions are first validated using measured anechoic ATFs. In addition, more realistic experiments are performed using recorded signals for a binaural hearing device configuration in a reverberant cafeteria with one interfering source and multi-talker babble noise. Both the objective performance measures as well as the results of a perceptual listening test with 13 normal-hearing participants show that the proposed BLCMV-N beamformer is able to preserve the binaural cues and hence the spatial impression of the interfering source (like the BLCMV beamformer), while trading off between noise reduction performance and binaural cue preservation of the background noise (like the BMVDR-N beamformer).

The remainder of this chapter is organized as follows. In Section 4.1 we present the BLCMV-N beamformer and derive two decompositions. In Section 4.2 we provide a detailed theoretical analysis of the proposed BLCMV-N beamformer in terms of noise and interference reduction performance and binaural cue preservation. In Section 4.3 we first validate the analytical expressions using anechoic ATFs, followed by simulations and a perceptual listening test using realistic recordings in a reverberant environment.

4.1 BLCMV beamformer with partial noise estimation

Aiming at merging the advantages of the BLCMV beamformer and the BMVDR-N beamformer, i.e., preserving the binaural cues of the interfering source and controlling the binaural cues of the background noise, in Section 4.1.1 we present the BLCMV beamformer with partial noise estimation (BLCMV-N). Similarly as for the BLCMV beamformer in [12], in Sections 4.1.2 and 4.1.3 we derive two decompositions for the BLCMV-N beamformer which reveal differences and similarities between the BLCMV-N beamformer and the BLCMV beamformer.

4.1.1 *BLCMV-N beamformer*

Compared to the BMVDR beamformer in (3.1) and (3.2), the BLCMV-N beamformer uses an additional constraint to preserve a scaled version of the interfering source component in the reference microphone signals, like the BLCMV beamformer in (3.15) and (3.16), and aims at preserving a scaled version of the noise component in the reference microphone signals, like the BMVDR-N beamformer in (3.37) and (3.38). The constrained optimization problem for the left and the right filter vector is given by

$$\min_{\mathbf{w}_L} \mathcal{E}\left\{\left|\mathbf{w}_L^H\mathbf{n} - \eta n_L\right|^2\right\} \quad \text{subject to} \quad \mathbf{w}_L^H\mathbf{x} = x_L\,,\ \mathbf{w}_L^H\mathbf{u} = \delta u_L \tag{4.1}$$

$$\min_{\mathbf{w}_R} \mathcal{E}\left\{\left|\mathbf{w}_R^H\mathbf{n} - \eta n_R\right|^2\right\} \quad \text{subject to} \quad \mathbf{w}_R^H\mathbf{x} = x_R\,,\ \mathbf{w}_R^H\mathbf{u} = \delta u_R \tag{4.2}$$

The solution of (4.1) and (4.2) is equal to (see Appendix A.1)

$$\mathbf{w}_{\text{BLCMV-N},L} = \eta\mathbf{e}_L + (1-\eta)\mathbf{R}_n^{-1}\mathbf{C}\left(\mathbf{C}^H\mathbf{R}_n^{-1}\mathbf{C}\right)^{-1}\begin{bmatrix} a_L^* \\ \bar{\delta}b_L^* \end{bmatrix} \tag{4.3}$$

$$\mathbf{w}_{\text{BLCMV-N},R} = \eta\mathbf{e}_R + (1-\eta)\mathbf{R}_n^{-1}\mathbf{C}\left(\mathbf{C}^H\mathbf{R}_n^{-1}\mathbf{C}\right)^{-1}\begin{bmatrix} a_R^* \\ \bar{\delta}b_R^* \end{bmatrix} \tag{4.4}$$

with $\mathbf{C}$ defined in (3.19) and the *adjusted interference scaling parameter* $\bar{\delta}$ equal to

$$\bar{\delta} = \frac{\delta - \eta}{1 - \eta}\,. \tag{4.5}$$

Hence, the output signals of the BLCMV-N beamformer can be interpreted as a mixture between the noisy reference microphone signals (scaled with η) and the output signals of a BLCMV beamformer (scaled with $1-\eta$) using the adjusted interference scaling parameter $\bar{\delta}$ in (4.5) instead of the interference scaling parameter δ. For $\eta = 0$, the BLCMV-N beamformer is equal to the BLCMV beamformer in (3.17) and (3.18) with $\bar{\delta} = \delta$, whereas for $\eta = 1$, it should be realized that only if $\delta = 1$ no beamforming is applied. Since mixing with the reference microphone signals not only affects the noise component but also the interfering source component, the adjusted interference scaling parameter $\bar{\delta}$ depends on both the interference scaling parameter δ as well as the mixing parameter η due to the interference reduction constraint in (4.1) and (4.2). Figure 4.1 depicts $\bar{\delta}$ as a function of η for different values of δ. It can be seen that

$$\bar{\delta}(\eta,\delta) = \begin{cases} > 0\,, & \text{for } \delta > \eta \\ < 0\,, & \text{for } \delta < \eta \\ 0\,, & \text{for } \delta = \eta \end{cases}. \tag{4.6}$$

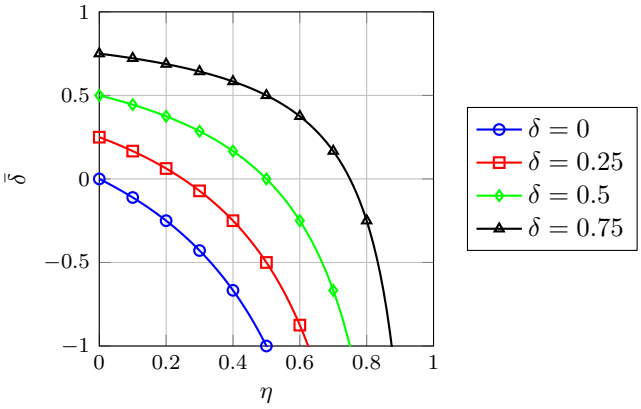

Fig. 4.1: Adjusted interference scaling parameter $\bar{\delta}$ as a function of η for different values of δ.

As shown in more detail in the following sections, using the parameters δ and η it is possible to control the noise reduction performance, the interference reduction performance and the binaural cues of the background noise for the BLCMV-N beamformer.

4.1.2 *Decomposition into two BLCMV beamformers*

In [12] it has been shown that the BLCMV beamformer in (3.17) and (3.18) can be decomposed as the sum of two sub-BLCMV beamformers, i.e.,

$$\mathbf{w}_{\mathrm{BLCMV},L} = \mathbf{w}_{x,L} + \delta \mathbf{w}_{u,L}\,, \tag{4.7}$$

$$\mathbf{w}_{\mathrm{BLCMV},R} = \mathbf{w}_{x,R} + \delta \mathbf{w}_{u,R}\,, \tag{4.8}$$

with

$$\mathbf{w}_{x,L} = \mathbf{R}_n^{-1}\mathbf{C}\left(\mathbf{C}^H\mathbf{R}_n^{-1}\mathbf{C}\right)^{-1}\mathbf{g}_{x,L}\,, \tag{4.9}$$

$$\mathbf{w}_{u,L} = \mathbf{R}_n^{-1}\mathbf{C}\left(\mathbf{C}^H\mathbf{R}_n^{-1}\mathbf{C}\right)^{-1}\mathbf{g}_{u,L}\,, \tag{4.10}$$

$$\mathbf{w}_{x,R} = \mathbf{R}_n^{-1}\mathbf{C}\left(\mathbf{C}^H\mathbf{R}_n^{-1}\mathbf{C}\right)^{-1}\mathbf{g}_{x,R}\,, \tag{4.11}$$

$$\mathbf{w}_{u,R} = \mathbf{R}_n^{-1}\mathbf{C}\left(\mathbf{C}^H\mathbf{R}_n^{-1}\mathbf{C}\right)^{-1}\mathbf{g}_{u,R}\,, \tag{4.12}$$

and the respective response vectors

$$\mathbf{g}_{x,L} = \begin{bmatrix} a_L^* \\ 0 \end{bmatrix}, \quad \mathbf{g}_{u,L} = \begin{bmatrix} 0 \\ b_L^* \end{bmatrix}, \tag{4.13}$$

$$\mathbf{g}_{x,R} = \begin{bmatrix} a_R^* \\ 0 \end{bmatrix}, \quad \mathbf{g}_{u,R} = \begin{bmatrix} 0 \\ b_R^* \end{bmatrix}. \tag{4.14}$$

The sub-BLCMV beamformer $\mathbf{w}_{x,L}$ and $\mathbf{w}_{x,R}$ in (4.9) and (4.11) preserves the desired source component in the reference microphone signals and steers a null towards the interfering source, whereas the sub-BLCMV beamformer $\mathbf{w}_{u,L}$ and $\mathbf{w}_{u,R}$ in (4.10) and (4.12) preserves the interfering source component in the reference microphone signals and steers a null towards the desired source. It can therefore be shown that [12]

$$\mathbf{w}_{x,L}^H \mathbf{a} = a_L\,, \quad \mathbf{w}_{x,L}^H \mathbf{b} = 0\,, \tag{4.15}$$

$$\mathbf{w}_{u,L}^H \mathbf{a} = 0\,, \quad \mathbf{w}_{u,L}^H \mathbf{b} = b_L\,, \tag{4.16}$$

$$\mathbf{w}_{x,R}^H \mathbf{a} = a_R\,, \quad \mathbf{w}_{x,R}^H \mathbf{b} = 0\,, \tag{4.17}$$

$$\mathbf{w}_{u,R}^H \mathbf{a} = 0\,, \quad \mathbf{w}_{u,R}^H \mathbf{b} = b_R\,. \tag{4.18}$$

Using (4.3), (4.4) and (4.15)–(4.18), it can be easily seen that the proposed BLCMV-N beamformer can be decomposed as

$$\boxed{\mathbf{w}_{\mathrm{BLCMV-N},L} = \eta \mathbf{e}_L + (1-\eta)\mathbf{w}_{x,L} + (\delta-\eta)\mathbf{w}_{u,L}} \tag{4.19}$$

$$\boxed{\mathbf{w}_{\mathrm{BLCMV-N},R} = \eta \mathbf{e}_R + (1-\eta)\mathbf{w}_{x,R} + (\delta-\eta)\mathbf{w}_{u,R}} \tag{4.20}$$

Figure 4.2 depicts this decomposition of the BLCMV-N beamformer using two sub-BLCMV beamformers. Hence, the BLCMV-N beamformer can be interpreted as a mixture between the reference microphone signals (scaled with η), a BLCMV beamformer that preserves the desired source and rejects the interfering source (scaled with $1-\eta$) and a BLCMV beamformer that preserves the interfering source and rejects the desired source (scaled with $\delta-\eta$). Since the scaling of the sub-BLCMV beamformer $\mathbf{w}_{x,L}$ controls the desired source component without affecting the interfering source component and the scaling of the sub-BLCMV beamformer $\mathbf{w}_{u,L}$ controls the interfering source component without affecting the desired source component [12], it can be directly observed from the scaling factors in (4.19) and (4.20) that the desired source component is not distorted and the interfering source component is scaled with δ.

4.1.3 *Decomposition using binauralization postfilters*

In [12] it has also been shown that the sub-BLCMV beamformer $\mathbf{w}_{x,L}$ in (4.9) for the left hearing device and the sub-BLCMV beamformer $\mathbf{w}_{x,R}$ in (4.11) for the right

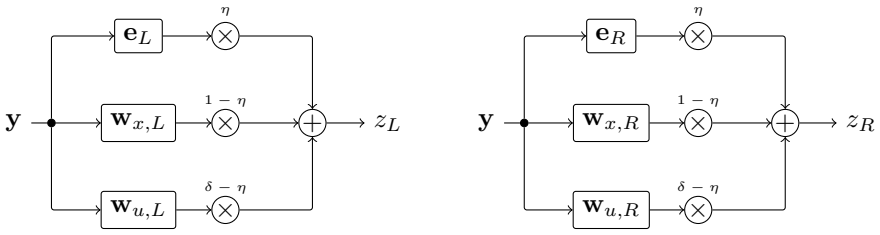

Fig. 4.2: Decomposition of the BLCMV-N beamformer into a mixture between the reference microphone signal and two sub-BLCMV beamformers.

hearing device can be written using a common spatial filter and two binauralization postfilters as

$$\mathbf{w}_{x,L} = \mathbf{w}_x a_L^* \,, \tag{4.21}$$

$$\mathbf{w}_{x,R} = \mathbf{w}_x a_R^* \,, \tag{4.22}$$

with the common desired BLCMV beamformer (D-BLCMV) given by

$$\mathbf{w}_x = \frac{1}{1-\Psi}\left(\frac{\mathbf{R}_n^{-1}\mathbf{a}}{\gamma_a} - \Psi\frac{\mathbf{R}_n^{-1}\mathbf{b}}{\gamma_{ab}}\right) , \tag{4.23}$$

and the ATFs a_L and a_R between the desired source and the reference microphones used as binauralization postfilters. Similarly, the sub-BLCMV beamformer $\mathbf{w}_{u,L}$ in (4.10) and the sub-BLCMV beamformer $\mathbf{w}_{u,R}$ in (4.12) can be written as

$$\mathbf{w}_{u,L} = \mathbf{w}_u b_L^* \,, \tag{4.24}$$

$$\mathbf{w}_{u,R} = \mathbf{w}_u b_R^* \,, \tag{4.25}$$

with the common interference BLCMV beamformer (I-BLCMV) given by

$$\mathbf{w}_u = \frac{1}{1-\Psi}\left(\frac{\mathbf{R}_n^{-1}\mathbf{b}}{\gamma_b} - \Psi\frac{\mathbf{R}_n^{-1}\mathbf{a}}{\gamma_{ab}^*}\right) , \tag{4.26}$$

and the ATFs b_L and b_R between the interfering source and the reference microphones used as binauralization postfilters.

Using (4.21), (4.22), (4.24) and (4.25) in (4.19) and (4.20), the BLCMV-N beamformer can be decomposed as

$$\boxed{\mathbf{w}_{\mathrm{BLCMV-N},L} = \eta\mathbf{e}_L + (1-\eta)a_L^*\mathbf{w}_x + (\delta-\eta)b_L^*\mathbf{w}_u} \tag{4.27}$$

$$\boxed{\mathbf{w}_{\mathrm{BLCMV-N},R} = \eta\mathbf{e}_R + (1-\eta)a_R^*\mathbf{w}_x + (\delta-\eta)b_R^*\mathbf{w}_u} \tag{4.28}$$

Figure 4.3 depicts this decomposition of the BLCMV-N beamformer using common spatial filters and binauralization postfilters. The output signals of the BLCMV-N

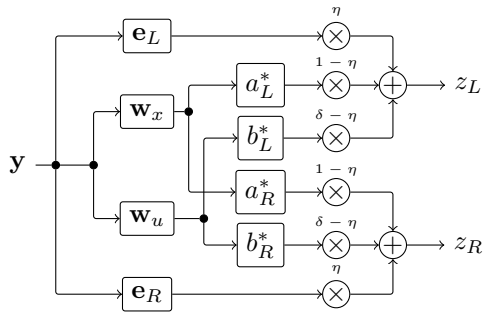

Fig. 4.3: Decomposition of the BLCMV-N beamformer into a mixture between the reference microphone signals and two BLCMV beamformers with binauralization postfilters.

beamformer can hence be interpreted as a mixture between the reference microphone signals (scaled with η), the binauralized output signals of the D-BLCMV beamformer (scaled with $1-\eta$) and the binauralized output signals of the I-BLCMV beamformer (scaled with $\delta - \eta$).

Due to the constraints in (4.1) and (4.2), the BLCMV-N beamformer perfectly preserves the desired source component and scales the interfering source component with δ, i.e.,

$$\mathbf{w}_{\mathrm{BLCMV-N},L}^H \mathbf{x} = x_L \,, \tag{4.29}$$

$$\mathbf{w}_{\mathrm{BLCMV-N},L}^H \mathbf{u} = \delta u_L \,, \tag{4.30}$$

$$\mathbf{w}_{\mathrm{BLCMV-N},R}^H \mathbf{x} = x_R \,, \tag{4.31}$$

$$\mathbf{w}_{\mathrm{BLCMV-N},R}^H \mathbf{u} = \delta u_R \,. \tag{4.32}$$

Using (4.27) and (4.28), the noise component in the output signals of the BLCMV-N beamformer are equal to

$$\mathbf{w}_{\mathrm{BLCMV-N},L}^H \mathbf{n} = \eta n_L + (1-\eta) n_x a_L + (\delta - \eta) n_u b_L \,, \tag{4.33}$$

$$\mathbf{w}_{\mathrm{BLCMV-N},R}^H \mathbf{n} = \eta n_R + (1-\eta) n_x a_R + (\delta - \eta) n_u b_R \,, \tag{4.34}$$

with $n_x = \mathbf{w}_x^H \mathbf{n}$ and $n_u = \mathbf{w}_u^H \mathbf{n}$ the noise component in the output signal of the D-BLCMV beamformer and the I-BLCMV beamformer, respectively. The noise component in the output signals of the BLCMV-N beamformer can hence be interpreted as a mixture between the noise component in the reference microphone signals (scaled with η), a coherent residual noise source (n_x) coming from the direction of the desired source (scaled with $1-\eta$) and a coherent residual noise source (n_u) coming from the direction of the interfering source (scaled with $\delta - \eta$).

4.2 Performance of the BLCMV-N beamformer

In this section we provide a performance analysis of the proposed BLCMV-N beamformer. In Section 4.2.1 we derive the output PSDs of the signal components. In Sections 4.2.2 and 4.2.3 we analyze the noise and interference reduction performance and the binaural cue preservation. Finally, in Section 4.2.4 we discuss the setting of the mixing parameter η and the interference scaling parameter δ.

4.2.1 *Output power spectral densities*

Due to the constraints in (4.1) and (4.2), the output PSD of the desired source and interfering source components in the left and the right output signal of the BLCMV-N beamformer is equal to, cf. (2.22)–(2.26),

$$p^{\text{out}}_{\text{BLCMV-N},x_L} = p_{x_L} = p_{s_x}|a_L|^2\,, \tag{4.35}$$

$$p^{\text{out}}_{\text{BLCMV-N},u_L} = \delta^2 p_{u_L} = \delta^2 p_{s_u}|b_L|^2\,, \tag{4.36}$$

$$p^{\text{out}}_{\text{BLCMV-N},x_R} = p_{x_R} = p_{s_x}|a_R|^2\,, \tag{4.37}$$

$$p^{\text{out}}_{\text{BLCMV-N},u_R} = \delta^2 p_{u_R} = \delta^2 p_{s_u}|b_R|^2\,. \tag{4.38}$$

The PSD of the desired source component obviously remains unchanged, while the PSD of the interfering source component is directly scaled with the interference scaling parameter δ. Furthermore, the output PSD of the noise component in the left and the right output signal of the BLCMV-N beamformer is equal to (see Appendix A.2)

$$p^{\text{out}}_{\text{BLCMV-N},n_L} = \mathbf{e}_L^T\left(\eta^2\mathbf{R}_n + \mathbf{R}_{xu,3}\right)\mathbf{e}_L\,, \tag{4.39}$$

$$p^{\text{out}}_{\text{BLCMV-N},n_R} = \mathbf{e}_R^T\left(\eta^2\mathbf{R}_n + \mathbf{R}_{xu,3}\right)\mathbf{e}_R\,, \tag{4.40}$$

with

$$\mathbf{R}_{xu,3} = \frac{1}{1-\Psi}\left[(1-\eta^2)\frac{\mathbf{a}\mathbf{a}^H}{\gamma_a} + (\delta^2-\eta^2)\frac{\mathbf{b}\mathbf{b}^H}{\gamma_b} - 2\Psi(\delta-\eta^2)\Re\left\{\frac{\mathbf{a}\mathbf{b}^H}{\gamma_{ab}^*}\right\}\right], \tag{4.41}$$

with γ_a defined in (3.5), γ_{ab} defined in (3.10), and γ_b and Ψ defined in (3.30). It can be seen that the output PSD of the noise component for the BLCMV-N beamformer is a quadratic function in both the mixing parameter η and the interference scaling parameter δ. By comparing (4.41) to (3.29), it can be observed that

$$\boxed{\mathbf{R}_{xu,3} = \mathbf{R}_{xu,1} - \eta^2\mathbf{R}_{xu,1}^{\delta=1}} \tag{4.42}$$

where $\mathbf{R}_{xu,1}^{\delta=1}$ denotes the expression for the BLCMV beamformer in (3.29) with $\delta = 1$, corresponding to no suppression of the interfering source. Please note that

for $\eta = 0$, $\mathbf{R}_{xu,3} = \mathbf{R}_{xu,1}$, and for $\eta = 1$ and $\delta = 1$, $\mathbf{R}_{xu,3} = \mathbf{0}_{M_H}$. By using (4.41) in (4.39) and (4.40), it follows that

$$p^{\text{out}}_{\text{BLCMV-N},n_L} = \eta^2 \left(p^{\text{in}}_{n_L} - p^{\text{out},\delta=1}_{\text{BLCMV},n_L} \right) + p^{\text{out}}_{\text{BLCMV},n_L} , \tag{4.43}$$

$$p^{\text{out}}_{\text{BLCMV-N},n_R} = \eta^2 \left(p^{\text{in}}_{n_R} - p^{\text{out},\delta=1}_{\text{BLCMV},n_R} \right) + p^{\text{out}}_{\text{BLCMV},n_R} . \tag{4.44}$$

4.2.2 *Noise and interference reduction performance*

By substituting (4.35), (4.37), (4.39) and (4.40) in (2.68) and (2.69), the left and the right output SNR of the BLCMV-N beamformer is equal to

$$\text{SNR}^{\text{out}}_{\text{BLCMV-N},L} = \frac{p_{s_x} |a_L|^2}{\mathbf{e}_L^T \left(\eta^2 \mathbf{R}_n + \mathbf{R}_{xu,3} \right) \mathbf{e}_L} , \tag{4.45}$$

$$\text{SNR}^{\text{out}}_{\text{BLCMV-N},R} = \frac{p_{s_x} |a_R|^2}{\mathbf{e}_R^T \left(\eta^2 \mathbf{R}_n + \mathbf{R}_{xu,3} \right) \mathbf{e}_R} , \tag{4.46}$$

which depends on both the mixing parameter η and the interference scaling parameter δ. Using (4.43) and (4.44) and realizing that the output PSD of the noise component in the left and the right output signal of the BLCMV beamformer (for any value for δ) is smaller than or equal to the PSD of the noise component in the left and the right reference microphone signal, respectively, the output SNR of the BLCMV-N beamformer in (4.45) and (4.46) is smaller than or equal to the output SNR of the BLCMV beamformer in (3.27) and (3.28), i.e.,

$$\boxed{\text{SNR}^{\text{out}}_{\text{BLCMV-N},L} \leq \text{SNR}^{\text{out}}_{\text{BLCMV},L} \leq \text{SNR}^{\text{out}}_{\text{BMVDR},L}} \tag{4.47}$$

$$\boxed{\text{SNR}^{\text{out}}_{\text{BLCMV-N},R} \leq \text{SNR}^{\text{out}}_{\text{BLCMV},R} \leq \text{SNR}^{\text{out}}_{\text{BMVDR},R}} \tag{4.48}$$

By substituting (4.35) and (4.36) in (2.72) and (2.73), the left and the right output SIR of the BLCMV-N beamformer is equal to

$$\text{SIR}^{\text{out}}_{\text{BLCMV-N},L} = \frac{1}{\delta^2} \text{SIR}^{\text{in}}_L , \tag{4.49}$$

$$\text{SIR}^{\text{out}}_{\text{BLCMV-N},R} = \frac{1}{\delta^2} \text{SIR}^{\text{in}}_R , \tag{4.50}$$

which is equal to the left and the right output SIR of the BLCMV beamformer in (3.31) and (3.32) and solely controlled by the interference scaling parameter δ. For $\eta = 0$, the left and the right output SNR of the BLCMV-N beamformer is equal to the left and the right output SNR of the BLCMV beamformer in (3.27) and (3.28), while for $\eta = 1$ and $\delta = 1$, the left and the right output SNR of the BLCMV-N beamformer is equal to the left and the right input SNR because no beamforming is applied.

4.2.3 *Binaural cue preservation*

Similarly as for the BLCMV beamformer, due to the constraints in (4.1) and (4.2) the BLCMV-N beamformer preserves the binaural cues of both the desired source and the interfering source, i.e.,

$$\mathrm{ITF}^{\mathrm{out}}_{\mathrm{BLCMV-N},x} = \frac{a_L}{a_R} = \mathrm{ITF}^{\mathrm{in}}_{x}\,, \tag{4.51}$$

$$\mathrm{ITF}^{\mathrm{out}}_{\mathrm{BLCMV-N},u} = \frac{b_L}{b_R} = \mathrm{ITF}^{\mathrm{in}}_{u}\,. \tag{4.52}$$

Using (2.83), the output IC of the noise component for the BLCMV-N beamformer is equal to (see Appendix A.2 for derivation of components)

$$\mathrm{IC}^{\mathrm{out}}_{\mathrm{BLCMV-N},n} = \frac{\mathbf{e}_L^T\left(\eta^2\mathbf{R}_n + \mathbf{R}_{xu,3}\right)\mathbf{e}_R}{\sqrt{\mathbf{e}_L^T\left(\eta^2\mathbf{R}_n + \mathbf{R}_{xu,3}\right)\mathbf{e}_L}\sqrt{\mathbf{e}_R^T\left(\eta^2\mathbf{R}_n + \mathbf{R}_{xu,3}\right)\mathbf{e}_R}}\,, \tag{4.53}$$

with $\mathbf{R}_{xu,3}$ defined in (4.41). Since $\mathbf{R}_{xu,3}$ depends on both the mixing parameter η and the interference scaling parameter δ, also the output IC of the noise component in (4.53) depends on both parameters. Using (2.84), the output MSC of the noise component for the BLCMV-N beamformer is equal to

$$\mathrm{MSC}^{\mathrm{out}}_{\mathrm{BLCMV-N},n} = \left|\mathrm{IC}^{\mathrm{out}}_{\mathrm{BLCMV-N},n}\right|^2. \tag{4.54}$$

Since for $\eta = 0$ the BLCMV-N beamformer is equal to the BLCMV beamformer, the output MSC of the noise component is smaller than 1, see Section 3.2. It should however be realized that in contrast to the BMVDR-N beamformer discussed in Section 3.3, for $\eta = 1$ the BLCMV-N beamformer does not always preserve the MSC of the noise component. Only for $\eta = 1$ and $\delta = 1$ the binaural cues of all signal components are preserved because no beamforming is applied. Table 4.1 summarizes the noise and interference reduction performance and binaural cue preservation of all considered binaural beamforming algorithms.

4.2.4 *Parameter settings*

Maximizing the left output SNR in (4.45) corresponds to minimizing the denominator, i.e., using (4.42),

$$D(\eta,\delta) = \mathbf{e}_L^T\left[\eta^2\left(\mathbf{R}_n - \mathbf{R}^{\delta=1}_{xu,1}\right) + \mathbf{R}_{xu,1}\right]\mathbf{e}_L\,. \tag{4.55}$$

Setting the derivative of (4.55) with respect to the mixing parameter η equal to zero, yields

$$\eta_{\mathrm{opt}} = 0 \tag{4.56}$$

Table 4.1: Noise and interference reduction performance and binaural cue preservation of all considered binaural beamforming algorithms. †: Depends on relative position of interfering source to desired source.

Algorithm	**SNR**	**SIR**	**ITF$_x$**	**ITF$_u$**	**MSC$_n$**
BMVDR	+++	†	preserved	not preserved	not preserved
BLCMV	++	controlled by δ	preserved	preserved	not preserved, †
BMVDR-N	++	†	preserved	not preserved, †	controlled by η
BLCMV-N	+	controlled by δ	preserved	preserved	controlled by η

as the optimal mixing parameter η in terms of left (and right) output SNR. The derivative of (4.55) with respect to the interference scaling parameter δ is equal to, using (3.29),

$$\frac{\partial D(\eta,\delta)}{\partial \delta} = \frac{1}{1-\Psi}\left(2\delta\frac{|b_L|^2}{\gamma_b} - 2\Psi\Re\left\{\frac{a_L b_L^*}{\gamma_{ab}^*}\right\}\right). \tag{4.57}$$

Setting (4.57) to zero and solving for δ yields the optimal interference scaling parameter in terms of left output SNR, i.e.,

$$\delta_{\text{opt},L} = \frac{\alpha_L}{\beta_L}, \tag{4.58}$$

with

$$\alpha_L = \Psi\Re\left\{\frac{a_L b_L^*}{\gamma_{ab}^*}\right\}, \quad \beta_L = \frac{|b_L|^2}{\gamma_b}. \tag{4.59}$$

As can be seen from (4.49), the output SIR is not affected by the mixing parameter η but is solely determined by the interference scaling parameter δ.

4.3 Simulations

In Section 4.3.1 we first validate the expressions derived in the previous sections using measured anechoic ATFs. In Section 4.3.2 we then experimentally compare the performance of the proposed BLCMV-N beamformer with the BMVDR beamformer, BLCMV beamformer and BMVDR-N beamformer using recorded signals in a reverberant environment with a competing speaker and multi-talker babble noise. Finally, in Section 4.3.3 we compare the spatial impression of the considered binaural beamforming algorithms using a perceptual listening test.

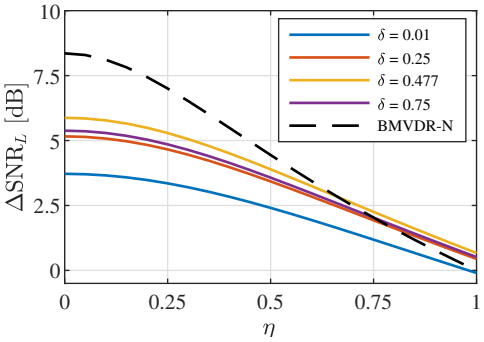

Fig. 4.4: Left SNR improvement for the BLCMV-N beamformer and the BMVDR-N beamformer at 500 Hz.

4.3.1 *Validation using measured anechoic ATFs*

To validate the derived expressions for the considered algorithms we used measured anechoic ATFs of two behind-the-ear hearing aids mounted on a head-and-torso-simulator (HATS) [35]. Each hearing aid has two microphones ($M_H = 4$) with an inter-microphone distance of about 14 mm. We chose the front microphone on each hearing aid as reference microphone. The ATFs were calculated from anechoic RIRs using a 512-point FFT at a sampling rate of 16 kHz.

The desired source was placed at 0° (in front) and the interfering source was placed at −35° (to the left), both at a distance of 3 m from the HATS. The desired source covariance matrix $\mathbf{R}_x$ and the interfering source covariance matrix $\mathbf{R}_u$ were constructed using the ATF vector of the desired source $\mathbf{a}$ and the ATF vector of the interfering source $\mathbf{b}$ according to (2.20) and (2.21), respectively, where the PSD of the desired source p_{s_x} and the PSD of the interfering source p_{s_u} were both set to 1. As background noise we considered a combination of spatially white and cylindrically isotropic noise, i.e., the noise covariance matrix $\mathbf{R}_n$ was constructed as

$$\mathbf{R}_n = p_n^{\text{white}}\mathbf{\Gamma}^{\text{white}} + p_n^{\text{dat}}\mathbf{\Gamma}^{\text{dat}}\,, \tag{4.60}$$

with p_n^{white} the PSD of the spatially white noise, $\mathbf{\Gamma}^{\text{white}}$ defined in (3.61), p_n^{dat} the PSD of the cylindrically isotropic noise and $\mathbf{\Gamma}^{\text{dat}}$ its spatial coherence matrix. The spatial coherence matrix $\mathbf{\Gamma}^{\text{dat}}$ of the cylindrically isotropic noise was calculated as in (3.64) using all available anechoic ATFs (72 for the database in [35]). The PSD of the spatially white noise p_n^{white} was set to −55 dB, while the PSD of the cylindrically isotropic noise p_n^{dat} was set to 1.

4.3.1.1 *Noise and interference reduction performance*

Using (2.33) and (4.45) in (2.70), Figure 4.4 depicts the left SNR improvement at 500 Hz for the BLCMV-N beamformer for different values of the mixing parameter η and the interference scaling parameter δ and the BMVDR-N beamformer for

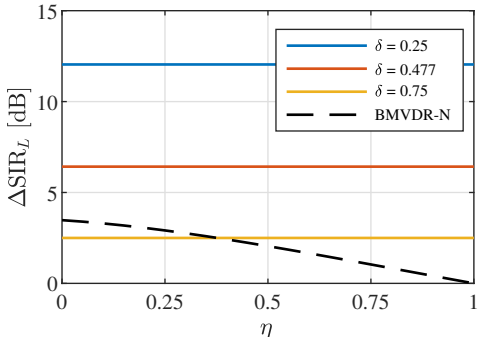

Fig. 4.5: Left SIR improvement for the BLCMV-N beamformer and the BMVDR-N beamformer at 500 Hz.

different values of the mixing parameter η. As expected, the BMVDR beamformer (i.e., BMVDR-N beamformer for $\eta = 0$) yields the largest SNR improvement (cf. (4.47)). Since the BMVDR-N beamformer mixes the output signals of the BMVDR beamformer with the noisy reference microphone signals, it can be observed that increasing the mixing parameter η reduces the SNR improvement of the BMVDR-N beamformer compared to the BMVDR beamformer ($\eta = 0$). For the BLCMV-N beamformer, both η and δ affect the SNR improvement, which is in line with (4.45). Similarly to the BMVDR-N beamformer, the BLCMV-N beamformer mixes the output signals of a BLCMV beamformer with the noisy reference microphone signals. Hence, it can be observed that for any value of the interference scaling parameter δ, increasing the mixing parameter η reduces the SNR improvement of the BLCMV-N beamformer compared to the BLCMV beamformer ($\eta = 0$), which is in line with (4.47). Since less degrees of freedom are available for noise reduction, the BLCMV beamformer ($\eta = 0$) yields a smaller SNR improvement compared to the BMVDR beamformer ($\eta = 0$), as discussed in Chapter 3. Using (4.58), the interference scaling parameter δ maximizing the left output SNR was equal to $\delta_{\mathrm{opt},L} = 0.477$ for the considered acoustic scenario. As expected, it can be observed that using $\delta_{\mathrm{opt},L}$ leads to the largest SNR improvement of all considered values of δ. For large values of the mixing parameter η, it can be observed that the BLCMV-N beamformer yields a larger SNR improvement than the BMVDR-N beamformer. It should be noted that the exact behaviour depends on the interference scaling parameter δ and the relative position of the interfering source to the desired source.

Using (2.35) and (4.49) in (2.74), Figure 4.5 depicts the left SIR improvement at 500 Hz for the BLCMV-N beamformer for different values of the mixing parameter η and the interference scaling parameter δ and the BMVDR-N beamformer for different values of the mixing parameter η. As expected from (3.31) and (4.49), both the BLCMV-N beamformer and the BLCMV beamformer ($\eta = 0$) yield the same SIR improvement, which is solely controlled by the interference scaling parameter δ. Hence, increasing the interference scaling parameter δ reduces the SIR improvement for both the BLCMV-N beamformer and the BLCMV beamformer. For the BMVDR-N beamformer it can be observed that increasing the mixing parameter η

reduces the SIR improvement. It should be noted that the exact behaviour depends on the relative position of the interfering source to the desired source, as can be seen from (3.47) and (3.49).

4.3.1.2 *Binaural cue preservation of background noise*

For different frequencies, Figure 4.6 depicts the input MSC in (2.84) of the noise component (**Input**) and the output MSC of the noise component for the BLCMV beamformer in (3.36) for different values of the interference scaling parameter δ, the BMVDR-N beamformer in (3.53) for different values of the mixing parameter η and the BLCMV-N beamformer in (4.54) for different values of the mixing parameter η and the interference scaling parameter δ. Although the BLCMV beamformer is not designed to preserve the MSC of the noise component, it can be observed that an output MSC smaller than 1 is obtained, especially for large values of δ [12]. However, since the output MSC of the noise component depends on the relative position of the interfering source to the desired source, it cannot be easily controlled. Since the BMVDR-N beamformer mixes the output signals of the BMVDR beamformer with the noisy reference microphone signals, it can be observed that the output MSC of the noise component is smaller than 1, and for $\eta = 1$ the MSC is perfectly preserved (but no beamforming is applied). For the BLCMV-N beamformer, it can be observed that both η and δ influence the output MSC of the noise component, as discussed in Section 4.2.3. For $\eta = 0$, the output MSC of the noise component for the BLCMV-N beamformer is obviously equal to the output MSC of the noise component for the BLCMV beamformer. For a fixed value of δ, it can be observed that the output MSC of the noise component approaches the input MSC of the noise component for increasing η, although it should be realized that perfect preservation of the MSC of the noise component is only possible for $\delta = 1$ (cf. Section 4.2.3).

For several values of the mixing parameter η, Figure 4.7 depicts the MSC error of the noise component for the BLCMV-N beamformer and the BMVDR-N beamformer, averaged over all frequencies, i.e.,

$$\Delta\text{MSC} = \frac{1}{F-1}\sum_{f=1}^{F-1} \left|\text{MSC}_n^{\text{in}}(f) - \text{MSC}_n^{\text{out}}(f)\right|, \tag{4.61}$$

with f the frequency bin index and F the total number of frequency bins. As expected, the BMVDR beamformer ($\eta = 0$) yields the largest MSC error of the noise component and increasing the mixing parameter η reduces the frequency-averaged MSC error of the noise component for the BMVDR-N beamformer [13]. For the considered acoustic scenario, it can be observed for the BLCMV-N beamformer that for any value of the interference scaling parameter δ, increasing the mixing parameter η reduces the frequency-averaged MSC error of the noise component compared to the BLCMV beamformer ($\eta = 0$). Further, it can be observed that for small values of the interference scaling parameter δ, the effect of the mixing parameter η is larger than for large values of the interference scaling parameter δ, for which the frequency-averaged MSC error is relatively small for all values of

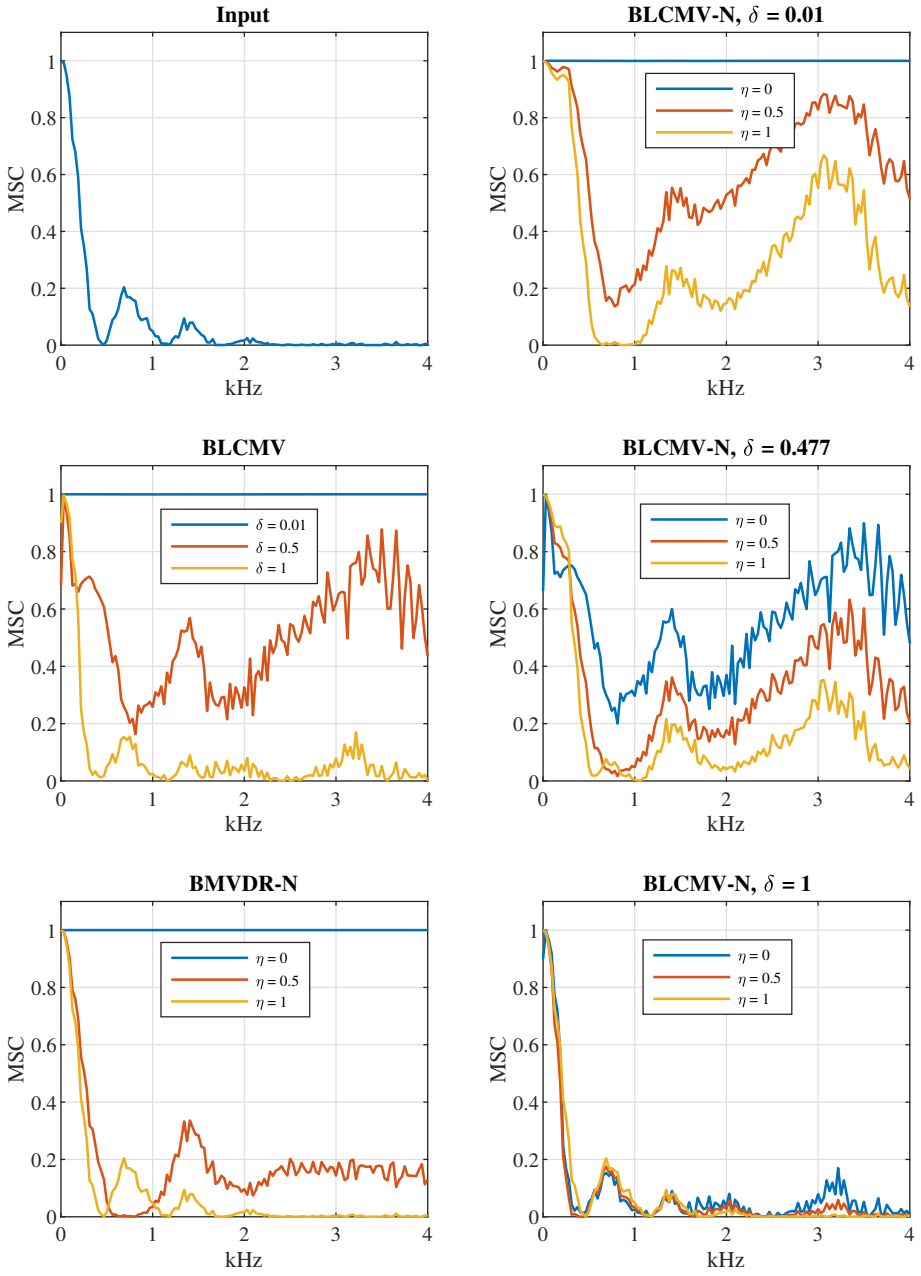

Fig. 4.6: The MSC of the noise component in the reference microphone signals (**Input**), in the output signals of the BLCMV beamformer for different values of the interference scaling parameter δ, the BMVDR-N beamformer for different values of the mixing parameter η and the BLCMV-N beamformer for different values of the mixing parameter η and the interference scaling parameter δ.

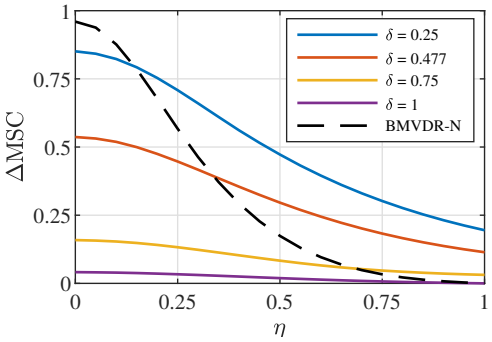

Fig. 4.7: Frequency-averaged MSC error of the noise component for the BLCMV-N beamformer and the BMVDR-N beamformer.

the mixing parameter η. The results clearly show that the mixing parameter η in the BLCMV-N beamformer enables to control the binaural cues of the background noise.

4.3.2 *Experimental results using reverberant recordings*

For a more realistic evaluation, we compare the performance of the considered binaural beamforming algorithms using reverberant recordings. Similarly to Section 4.3.1, the experimental setup consists of two hearing aids, each with two microphones, mounted on a HATS in a cafeteria with a reverberation time of approximately 1.25 s [35]. The desired source was again placed at 0° (at a distance of about 102 cm), while the interfering source was again placed at −35° (at a distance of about 118 cm), see [35] for more details. The desired and interfering source components were generated by convolving clean speech signals with the measured reverberant room impulse responses corresponding to the desired source and interfering source positions. The desired source was a male German speaker, speaking eight sentences with a pause of 1 s between the sentences. The interfering source was a male Dutch speaker, speaking seven sentences with a pause of 0.25 s between the sentences. As background noise we used realistic recordings [35], consisting of multi-talker babble noise, clacking plates and temporally dominant competing speakers. The used background noise hence clearly differed from the perfectly diffuse noise in Section 4.3.1. The entire signal had a length of about 28 s. The desired source and the background noise were active the entire time, whereas the interfering source only became active after about 14 s. The desired source component, the interfering source component and the noise component were mixed at an input SNR of 10 dB and input SIR of 5 dB in the right reference microphone. Again, we chose the front microphone on each hearing aid as reference microphone.

The processing was performed at a sampling rate of 16 kHz in the STFT domain with a time frame size T_d of 8192 samples and a square-root Hann window with 50 % overlap, i.e., $T_s = 4096$. We used an oracle VAD (i.e., using the desired source

Table 4.2: Objective performance measures for all considered binaural beamforming algorithms in the reverberant environment.

	BMVDR	BLCMV	BMVDR-N	BLCMV-N
ΔSNR_L [dB]	13.0	10.1	8.6	7.6
ΔSNR_R [dB]	12.9	9.2	8.6	7.0
ΔSIR_L [dB]	-0.1	9.7	0.82	9.8
ΔSIR_R [dB]	-4.3	8.7	-2.4	8.9
ΔMSC	0.86	0.64	0.10	0.19

and interfering source signals) to batch-estimate the noise covariance matrix $\mathbf{R}_n$, the undesired covariance matrix $\mathbf{R}_v$ (interfering source plus background noise) and $\mathbf{R}_{xn} = \mathbf{R}_x + \mathbf{R}_n$ (desired source plus background noise) over the entire signal. See Section 3.4.1 for details. All binaural beamforming algorithms were implemented using RTF vectors, as discussed in Chapter 3. Using the CW method in (3.84) and (3.85), the RTF vectors of the desired source and the interfering source were estimated based on the generalised eigenvalue decomposition of the batch estimates of $\mathbf{R}_{xn}$ and $\mathbf{R}_n$ or $\mathbf{R}_v$ and $\mathbf{R}_n$, respectively. The mixing parameter was set to $\eta = 0.3$ and the interference scaling parameter was set to $\delta = 0.3$.

As objective performance measures for noise and interference reduction performance, we used the left and the right SNR improvement (ΔSNR_L, ΔSNR_R) in (2.70) and (2.71) and the left and the right SIR improvement (ΔSIR_L, ΔSIR_R) in (2.74) and (2.75). As objective performance measure for binaural cue preservation of the background noise we used the frequency-averaged MSC error of the noise component (ΔMSC) as defined in (4.61). All objective performance measures were computed using the reference microphone signals and the output signals of all considered algorithms. Table 4.2 presents the objective performance measures for all considered algorithms.

4.3.2.1 *Noise and interference reduction performance*

In terms of noise reduction performance, it can be observed that – as expected – the BMVDR beamformer yields the highest SNR improvement (13.0 dB for the left and 12.9 dB for the right side). All other algorithms yield a lower SNR improvement, for the BLCMV beamformer due to the additional constraint for the interfering source, for the BMVDR-N beamformer due to the mixing with the noisy reference microphone signals, and for the BLCMV-N beamformer due to both effects. The partial noise estimation for the BLCMV-N beamformer seems to result in a smaller drop in noise reduction performance compared to the BLCMV beamformer (2.5 dB for the left side, 2.2 dB for the right side) than for the BMVDR-N beamformer compared to the BMVDR beamformer (4.4 dB for the left side, 4.3 dB for the right side). Please note that both for the BMVDR-N beamformer as well as for the BLCMV-N beamformer this drop in noise reduction performance depends on the relative position of the interfering source to the desired source.

In terms of interference reduction performance, it can be observed that both the BLCMV beamformer and the BLCMV-N beamformer approximately lead to the same SIR improvement (for the left and the right side), which is in line with the theoretical SIR improvement in (3.31), (3.32), (4.49) and (4.50), i.e., $10\log\frac{1}{\delta^2} \approx 10.5\,\text{dB}$. The fact that this theoretical SIR improvement is not reached and the fact that the SIR improvements for the BLCMV and BLCMV-N beamformers are not exactly the same is due to estimation errors in the covariance matrices, which was also already noted in [12, 163]. In addition, it can be observed that the BMVDR beamformer and BMVDR-N beamformer lead to very low (even negative) SIR improvements, which is presumably due to the fact that the interfering source is relatively close to the desired source.

4.3.2.2 *Binaural cue preservation of background noise*

As expected, the BMVDR beamformer yields the largest MSC error of the noise component ΔMSC. As discussed in Section 3.2, the output MSC of the noise component for the BLCMV beamformer is typically smaller than 1, hence leading to a smaller MSC error compared to the BMVDR beamformer. Due to the mixing with the noisy reference microphone signals, both the BMVDR-N beamformer and the BLCMV-N beamformer yield a much smaller MSC error of the noise component than the BMVDR beamformer and the BLCMV beamformer, where the MSC error is slightly smaller for the BMVDR-N beamformer than for the BLCMV-N beamformer.

In conclusion, the objective performance measures show that the BLCMV-N beamformer leads to a very similar interference reduction as the BLCMV beamformer, while providing a trade-off between noise reduction performance (slightly worse than the BLCMV beamformer) and binaural cue preservation of the background noise (much better than the BLCMV beamformer).

4.3.3 *Perceptual listenting test*

To further investigate the spatial impression of the different output signal components for the four considered algorithms, we conducted a perceptual listening test similarly to [113]. The desired source was now placed at $-35°$ and the interfering source was placed at 90°, in order to enhance the perceived spatial differences between both sources. The desired source component, the interfering source component and the noise component were mixed at an input SNR of 0 dB and input SIR of 0 dB in the right reference microphone. Thirteen self-reported normal-hearing subjects participated in the perceptual listening test, where none of the authors participated. All subjects can be considered expert listeners, i.e., they were familiar with similar perceptual listening tests, and gave informed consent. The listening test was conducted in a sound proof listening booth using an RME Fireface UCX sound card with Sennheiser HD 580 headphones.

Using a procedure similar to the MUlti-Stimulus Test with Hidden Reference and Anchor (MUSHRA) [187], the task was to rate the perceived spatial difference with

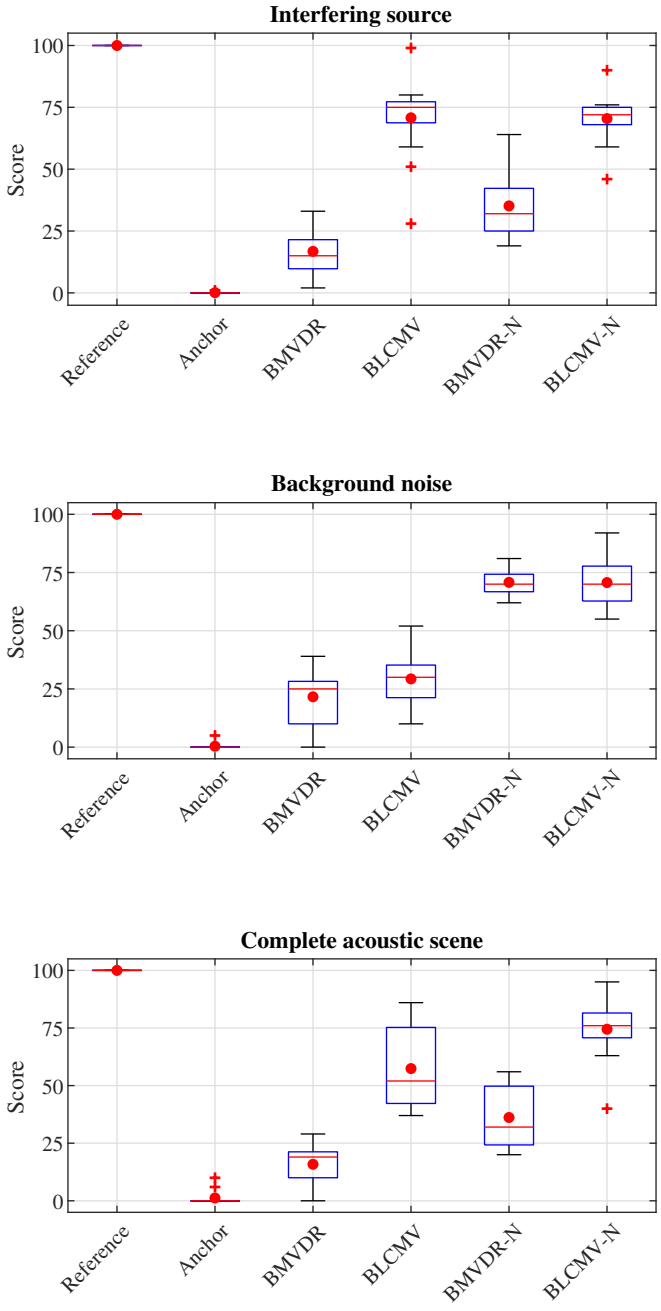

Fig. 4.8: Boxplot of the MUSHRA scores for all three evaluations. The plot depicts the median score (red line), the mean score (red dot), the first and third quartiles (blue boxes) and the interquartile ranges (whiskers). Outliers are indicated by red + markers.

respect to a reference signal. For a coherent source (e.g., interfering source), this corresponds to rating differences in perceived source location, whereas for a diffuse noise field this corresponds to rating differences in perceived diffuseness. A score of 0 is associated with a large perceived spatial difference, whereas a score of 100 is associated with no perceived spatial difference. As reference signal we used the (unprocessed) reference microphone signals, while as anchor signal we used the left reference microphone signal, played back to both ears. The anchor signal was hence a monaural signal with no binaural cues, which is perceived in the center of the head.

We conducted three evaluations, where only some components were active in the output signals, the reference signal and the anchor signal. In the first evaluation, only the desired source component and the interfering source component (i.e., no noise component) were active and the task was to rate the spatial difference for the interfering source. In the second evaluation, only the desired source component and the noise component (i.e., no interfering source component) were active and the task was to rate the spatial difference for the background noise. In the third evaluation, all signal components were active and the task was to rate the spatial difference for the interfering source and the background noise simultaneously. To familiarize the subjects with the tasks and the sound material, a training round was performed. Audio samples for all binaural beamforming algorithms and the unprocessed input signals are available online[†].

The MUSHRA scores for the three evaluations are shown in Figure 4.8. A one-way repeated-measures analysis of variance (ANOVA) was performed. The analysis revealed a significant within-subjects effect for all three evaluations. Hence, post-hoc comparison t-tests with Bonferroni correction were performed [188].

INTERFERING SOURCE: The within-subjects effect was significant [$F(2.098, 25.176) = 219.2$, $p < .001$, Greenhouse-Geisser correction]. As expected, the BLCMV beamformer and the BLCMV-N beamformer preserved the spatial impression of the interfering source significantly better than the BMVDR beamformer and the BMVDR-N beamformer ($p < .001$). The BMVDR-N beamformer performed significantly better than the BMVDR beamformer ($p < .001$), which is not unexpected since the interfering source component is also mixed with the mixing parameter η. No significant difference was found between the BLCMV beamformer and the BLCMV-N beamformer ($p = 1$).

BACKGROUND NOISE: The within-subjects effect was significant [$F(3.072, 36.869) = 332.066$, $p < .001$, Greenhouse-Geisser correction]. As expected, the BMVDR-N beamformer and the BLCMV-N beamformer, both using partial noise estimation, preserved the spatial impression of the background noise significantly better than the BMVDR beamformer and the BLCMV beamformer ($p < .001$). No significant difference was found between the BMVDR-N beamformer

† https://uol.de/en/sigproc/research/audio-demos/binaural-noise-reduction/blcmv-n-beamformer

and the BLCMV-N beamformer ($p = 1$) and between the BMVDR beamformer and the BLCMV beamformer ($p = .614$).

COMPLETE ACOUSTIC SCENE: The within-subjects effect was significant [$F(2.905, 34.858) = 171.783$, $p < .001$, Greenhouse-Geisser correction]. In terms of preservation of the spatial impression of the complete acoustic scene, the BMVDR-N beamformer scored significantly higher than the BMVDR beamformer ($p < .001$), the BLCMV beamformer scored significantly higher than the BMVDR-N beamformer ($p = .014$), and the proposed BLCMV-N beamformer scored significantly higher than the BLCMV beamformer ($p = .025$).

In summary, the results of the perceptual listening test showed that the BLCMV-N beamformer is capable of preserving the spatial impression of an interfering source and background noise in a realistic acoustic scenario, outperforming all other considered binaural beamforming algorithms in terms of spatial impression.

4.4 Summary

In this chapter we proposed the BLCMV-N beamformer, merging the advantages of the BLCMV beamformer and the BMVDR-N beamformer, i.e., preserving the binaural cues of the interfering source and controlling the reduction of the interfering source as well as the binaural cues of the background noise. We showed that the output signals of the BLCMV-N beamformer can be interpreted as a mixture between the noisy reference microphone signals and the output signals of a BLCMV beamformer using an adjusted interference scaling parameter. We provided a theoretical comparison between the BMVDR beamformer, the BLCMV beamformer, the BMVDR-N beamformer and the proposed BLCMV-N beamformer in terms of noise and interference reduction performance and binaural cue preservation. The obtained analytical expressions were first validated using measured anechoic acoustic transfer functions. Experimental results using recorded signals in a realistic reverberant environment showed that the BLCMV-N beamformer leads to a very similar interference reduction as the BLCMV beamformer, while providing a trade-off between noise reduction performance (slightly worse than the BLCMV beamformer) and binaural cue preservation of the background noise (much better than the BLCMV beamformer). In addition, the results of a perceptual listening test with 13 normal-hearing participants showed that the proposed BLCMV-N beamformer is capable of preserving the spatial impression of an interfering source and background noise in a realistic acoustic scenario, outperforming all other considered binaural beamforming algorithms in terms of spatial impression.

5

PERFORMANCE ANALYSIS OF THE EXTENDED BMVDR-N BEAMFORMER

While in Chapter 4 we only took into account the use of the head-mounted microphones, in this chapter we consider the extended binaural hearing device configuration in Figure 2.2 and investigate the incorporation of one external microphone in the BMVDR-N beamformer (Section 3.3).

In this chapter we consider an arbitrary noise field and derive analytical expressions for the output SNR and the binaural cues (more in particular the MSC) of the output noise component when incorporating an external microphone in the BMVDR-N beamformer. First, we show that an external microphone enables to obtain either a larger output SNR for the same mixing parameter or the same output SNR for a larger mixing parameter compared to using only the head-mounted microphones. Secondly, we show that the same desired output MSC of the noise component can be obtained for a smaller mixing parameter, implying that an external microphone enables to achieve the same spatial impression of the noise component compared to using only the head-mounted microphones while achieving a larger output SNR. The derived analytical expressions are first validated using simulated anechoic ATFs, where the listener's head is modelled as a rigid sphere [40]. In addition, experiments are performed using recorded signals for a binaural hearing device configuration in a reverberant environment with multiple interfering talkers as background noise [36]. For different positions of the external microphone and the desired source, the experimental results show that also in a realistic scenario incorporating an external microphone in the BMVDR-N beamformer significantly increases the output SNR and decreases the mixing parameter required to obtain a desired output MSC, i.e., spatial impression, of the noise component. The results generalize the results obtained in [15] assuming a coherent (directional) interference source, and the results

This chapter is partly based on:

[158] N. Gößling, D. Marquardt and S. Doclo, "Performance analysis of the extended binaural MVDR beamformer with partial noise estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, manuscript accepted with minor revisions, 2020.

in [157] assuming a homogeneous diffuse noise field and a desired source in front of the listener.

The remainder of this chapter is organized as follows. In Section 5.1 the considered signal model is briefly repeated for clarity. In Section 5.2 we define the extended BMVDR (eBMVDR) beamformer and the extended BMVDR-N (eBMVDR-N) beamformer, incorporating the external microphone. We then derive analytical expressions for the output SNR (Section 5.3) and the output MSC (Section 5.4) of the noise component for the eBMVDR-N beamformer for an arbitrary noise field and without assuming a specific position of the desired source. In Section 5.5 we provide simulation results using simulated anechoic ATFs as well as using recorded signals in a reverberant environment.

5.1 Signal model

We consider the extended binaural hearing device configuration as introduced in Section 2.1.2. In this chapter we do not distinguish between the interfering source and the background noise, but consider an arbitrary noise field. The extended noisy input vector (including the head-mounted microphone signals and the external microphone signal) is hence equal to

$$\mathbf{y}_e = \mathbf{x}_e + \mathbf{n}_e\,, \tag{5.1}$$

with $\mathbf{x}_e$ the extended desired source component and $\mathbf{n}_e$ the extended noise component. The extended noisy input covariance matrix $\mathbf{R}_{y,e}$ is then equal to

$$\mathbf{R}_{y,e} = \mathbf{R}_{x,e} + \mathbf{R}_{n,e}\,. \tag{5.2}$$

Since it is particularly important for this chapter, please note again that the output SNR of the BMVDR beamformer (as introduced in Section 3.1) for both the left and the right hearing device is equal to [3, 10]

$$\rho = \mathrm{SNR}^{\mathrm{out}}_{\mathrm{BMVDR},L} = \mathrm{SNR}^{\mathrm{out}}_{\mathrm{BMVDR},R} = p_{s_x}\mathbf{a}^H\mathbf{R}_n^{-1}\mathbf{a}\,. \tag{5.3}$$

5.2 Extended BMVDR (eBMVDR) and extended BMVDR-N (eBMVDR-N) beamformers

The BMVDR beamformer incorporating the external microphone is referred to as the *extended* BMVDR (eBMVDR) beamformer. Similarly to (3.3)–(3.7), by replacing the noise covariance matrix $\mathbf{R}_n$ with the extended noise covariance matrix $\mathbf{R}_{n,e}$ in (2.51) and the RTF vectors $\mathbf{a}_L$ and $\mathbf{a}_R$ with the extended RTF vectors $\mathbf{a}_{L,e}$ and

$\mathbf{a}_{R,e}$ in (2.44) the left and the right filter vector of the eBMVDR beamformer is equal to

$$\mathbf{w}_{\text{eBMVDR},L} = \frac{\mathbf{R}_{n,e}^{-1}\mathbf{a}_e}{\mathbf{a}_e^H\mathbf{R}_{n,e}^{-1}\mathbf{a}_e}a_L^* = \frac{\mathbf{R}_{n,e}^{-1}\mathbf{a}_{L,e}}{\mathbf{a}_{L,e}^H\mathbf{R}_{n,e}^{-1}\mathbf{a}_{L,e}}\,, \tag{5.4}$$

$$\mathbf{w}_{\text{eBMVDR},R} = \frac{\mathbf{R}_{n,e}^{-1}\mathbf{a}_e}{\mathbf{a}_e^H\mathbf{R}_{n,e}^{-1}\mathbf{a}_e}a_R^* = \frac{\mathbf{R}_{n,e}^{-1}\mathbf{a}_{R,e}}{\mathbf{a}_{R,e}^H\mathbf{R}_{n,e}^{-1}\mathbf{a}_{R,e}}\,. \tag{5.5}$$

The BMVDR beamformer with partial noise estimation incorporating the external microphone signal is referred to as the *extended* BMVDR-N (eBMVDR-N) beamformer. Similarly to (3.39) and (3.40), the left and the right filter vector of the eBMVDR-N beamformer is equal to

$$\mathbf{w}_{\text{eBMVDR-N},L} = (1-\eta)\mathbf{w}_{\text{eBMVDR},L} + \eta\mathbf{e}_L\,, \tag{5.6}$$

$$\mathbf{w}_{\text{eBMVDR-N},R} = (1-\eta)\mathbf{w}_{\text{eBMVDR},R} + \eta\mathbf{e}_R\,, \tag{5.7}$$

where $\eta \in \mathbb{R}$ again denotes the mixing parameter, with $0 \leq \eta \leq 1$. The output signals of the eBMVDR-N beamformer are again equal to a mixture between the output signals of the eBMVDR beamformer (scaled with $1-\eta$) and the (noisy) reference microphone signals (scaled with η).

Similarly to (5.3) by substituting (5.6) and (5.7) in (2.68) and (2.69), the output SNR of the eBMVDR beamformer is equal to [157]

$$\boxed{\rho_e = \text{SNR}_{\text{eBMVDR},L}^{\text{out}} = \text{SNR}_{\text{eBMVDR},R}^{\text{out}} = p_{s_x}\mathbf{a}_e^H\mathbf{R}_{n,e}^{-1}\mathbf{a}_e} \tag{5.8}$$

Similarly to the mixing parameter η^{des} for the BMVDR-N beamformer in (3.55), the mixing parameter η_e^{des} for the eBMVDR-N beamformer leading to a desired output MSC, $\text{MSC}_n^{\text{des}}$, of the noise component is equal to

$$\boxed{\eta_e^{\text{des}} = \sqrt{\frac{\rho_e\left(\sqrt{\gamma^2-\alpha\beta}-\gamma\right)+\alpha}{\rho_e^2\beta - 2\rho_e\gamma + \alpha}}} \tag{5.9}$$

with α, β and γ defined in (3.56)–(3.58) and ρ_e defined in (5.8).

5.3 Output SNR with an external microphone

The inverse of the extended noise covariance matrix $\mathbf{R}_{n,e}$ can be written in terms of $\mathbf{R}_n^{-1}$ as [186]

$$\mathbf{R}_{n,e}^{-1} = \left[\begin{array}{c|c} \mathbf{R}_n^{-1} + \frac{1}{\xi}\mathbf{R}_n^{-1}\mathbf{r}_{n,E}\mathbf{r}_{n,E}^H\mathbf{R}_n^{-1} & -\frac{1}{\xi}\mathbf{R}_n^{-1}\mathbf{r}_{n,E} \\ \hline -\frac{1}{\xi}\mathbf{r}_{n,E}^H\mathbf{R}_n^{-1} & \frac{1}{\xi} \end{array}\right], \tag{5.10}$$

with

$$\xi = p_{n_E} - \mathbf{r}_{n,E}^H \mathbf{R}_n^{-1} \mathbf{r}_{n,E}\,, \tag{5.11}$$

the Schur complement of $\mathbf{R}_n$ in (2.51). It can be shown that $\xi > 0$, since $\mathbf{R}_{n,e}$ is assumed to be positive definite [186]. By substituting (5.10) in (5.8) and using (5.3), the output SNR of the eBMVDR beamformer can be written as

$$\rho_e = p_{s_x} \left(\mathbf{a}^H \mathbf{R}_n^{-1} \mathbf{a} + \frac{1}{\xi} \left| \mathbf{r}_{n,E}^H \mathbf{R}_n^{-1} \mathbf{a} - a_E \right|^2 \right)\,, \tag{5.12}$$

$$= \rho + p_{s_x} \frac{\left| \mathbf{r}_{n,E}^H \mathbf{R}_n^{-1} \mathbf{a} - a_E \right|^2}{p_{n_E} - \mathbf{r}_{n,E}^H \mathbf{R}_n^{-1} \mathbf{r}_{n,E}}\,. \tag{5.13}$$

Hence, as expected, the output SNR ρ_e of the eBMVDR beamformer is always larger than or equal to the output SNR ρ of the BMVDR beamformer (without an external microphone), i.e.,

$$\boxed{\rho_e \geq \rho} \tag{5.14}$$

As can be observed from (5.13), the SNR improvement due to incorporating the external microphone depends on the ATF a_E between the desired source and the external microphone. In addition, the SNR improvement depends on the PSD p_{n_E} of the noise component in the external microphone signal and the spatial correlation $\mathbf{r}_{n,E}$ between the noise component in the head-mounted microphones signals and the external microphone signal. This implies that the SNR improvement obviously depends on the position of the external microphone relative to the head-mounted microphones and the desired source.

Similarly to (3.41) and (3.42), the left and the right output SNR of the eBVMDR-N beamformer is equal to

$$\mathrm{SNR}_{\mathrm{eBMVDR-N},L}^{\mathrm{out}} = \frac{\rho_e}{1 + \eta^2 \left(\frac{\rho_e}{\mathrm{SNR}_L^{\mathrm{in}}} - 1 \right)}\,, \tag{5.15}$$

$$\mathrm{SNR}_{\mathrm{eBMVDR-N},R}^{\mathrm{out}} = \frac{\rho_e}{1 + \eta^2 \left(\frac{\rho_e}{\mathrm{SNR}_R^{\mathrm{in}}} - 1 \right)}\,. \tag{5.16}$$

Since $\rho_e \geq \mathrm{SNR}_L^{\mathrm{in}}$ and $\rho_e \geq \mathrm{SNR}_R^{\mathrm{in}}$, (5.15) and (5.16) also monotonically decrease with increasing η, such that a larger mixing parameter η leads to a smaller output SNR of the eBMVDR-N beamformer. Figure 5.1 depicts the left output SNR of the eBMVDR-N beamformer in (5.15) as a function of the output SNR ρ_e of the eBMVDR beamformer for different values of the mixing parameter η. For a given binaural hearing device configuration (i.e., positions of the head-mounted microphones), desired source position and noise field, the output SNR ρ in (5.3) of the BMVDR beamformer is a constant.

Now consider different positions of the external microphone, such that the output SNR ρ_e of the eBMVDR beamformer in (5.8) can be considered as a variable. Based

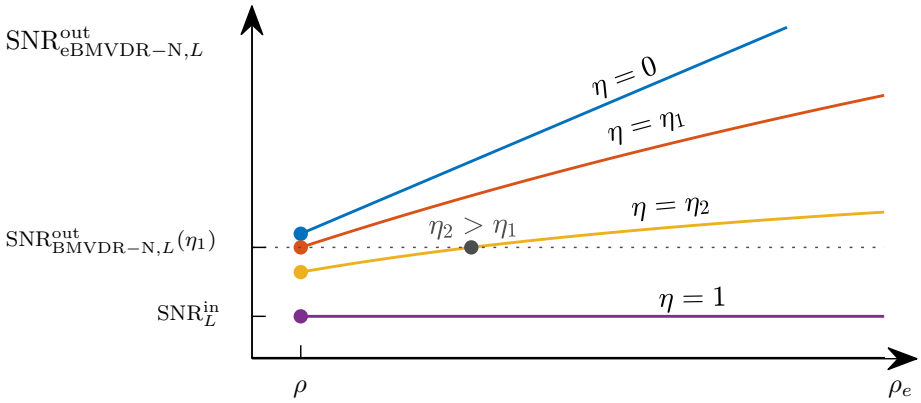

Fig. 5.1: Left output SNR of the eBMVDR-N beamformer as a function of the output SNR ρ_e of the eBMVDR beamformer for different values of the mixing parameter η. Please note that $\eta = 0$ corresponds to the eBMVDR beamformer and $\eta = 1$ corresponds to the left reference microphone signal.

on (5.14), the smallest possible value for the output SNR ρ_e of the eBMVDR beamformer is equal to ρ, i.e., the output SNR of the BMVDR beamformer (without external microphone). For a fixed value of the mixing parameter η, it can be easily shown that the partial derivative of (5.15) and (5.16) with respect to ρ_e is equal to

$$\frac{\partial \mathrm{SNR}^{\mathrm{out}}_{\mathrm{eBMVDR-N},L}}{\partial \rho_e} = \frac{1-\eta^2}{\left(1+\eta^2(\frac{\rho_e}{\mathrm{SNR}^{\mathrm{in}}_{L}}-1)\right)^2} \geq 0\,, \tag{5.17}$$

$$\frac{\partial \mathrm{SNR}^{\mathrm{out}}_{\mathrm{eBMVDR-N},R}}{\partial \rho_e} = \frac{1-\eta^2}{\left(1+\eta^2(\frac{\rho_e}{\mathrm{SNR}^{\mathrm{in}}_{R}}-1)\right)^2} \geq 0\,, \tag{5.18}$$

since $0 \leq \eta \leq 1$. Hence, for each value of the mixing parameter, e.g., $\eta = \eta_1$ (see Figure 5.1), the left and the right output SNR of the eBMVDR-N beamformer monotonically increases with ρ_e, and using (5.14), is always larger than or equal to the left and the right output SNR of the BMVDR-N beamformer, i.e.,

$$\boxed{\mathrm{SNR}^{\mathrm{out}}_{\mathrm{eBMVDR-N},L}(\eta_1) \geq \mathrm{SNR}^{\mathrm{out}}_{\mathrm{BMVDR-N},L}(\eta_1)} \tag{5.19}$$

$$\boxed{\mathrm{SNR}^{\mathrm{out}}_{\mathrm{eBMVDR-N},R}(\eta_1) \geq \mathrm{SNR}^{\mathrm{out}}_{\mathrm{BMVDR-N},R}(\eta_1)} \tag{5.20}$$

In addition, the same output SNR of the BMVDR-N beamformer is obtained when using a larger mixing parameter $\eta_2 > \eta_1$ for the eBMVDR-N beamformer, i.e.,

$$\boxed{\mathrm{SNR}^{\mathrm{out}}_{\mathrm{eBMVDR-N},L}(\eta_2) = \mathrm{SNR}^{\mathrm{out}}_{\mathrm{eBMVDR-N},L}(\eta_1)} \tag{5.21}$$

$$\boxed{\mathrm{SNR}^{\mathrm{out}}_{\mathrm{eBMVDR-N},R}(\eta_2) = \mathrm{SNR}^{\mathrm{out}}_{\mathrm{eBMVDR-N},R}(\eta_1)} \tag{5.22}$$

This means that incorporating an external microphone allows to use a larger mixing parameter, i.e., achieve a better spatial impression of the noise component, to obtain the same output SNR compared to only using the head-mounted microphone signals.

5.4 Output MSC with an external microphone

As discussed in Section 3.3, the mixing parameter η controls the binaural cues of the noise component at the output of the BMVDR-N beamformer. Since a larger mixing parameter leads to a lower output SNR, it is hence desirable to achieve the desired binaural cues of the noise output component using a small mixing parameter. For the special case of a coherent (directional) noise source, it has been experimentally shown in [15] that the same binaural cues, i.e., ILD and ITD, of the output noise component can be achieved using a smaller mixing parameter when incorporating an external microphone compared to using only the head-mounted microphones. Further, for the special case of a homogeneous noise field and a desired source in front of the listener, it has been analytically shown in [157] that the same desired output MSC of the noise component can be achieved using a smaller mixing parameter in the eBMVDR-N beamformer than in the BMVDR-N beamformer. In this section we generalize the analytical expressions derived in [157] without making any assumption about the noise field and the position of the desired source.

Since it was not straightforward to directly prove that η_e^{des} in (5.9) is always smaller than (or equal to) η^{des} in (3.55), we will take an indirect approach. Since $\rho_e \geq \rho$, showing that $\eta_e^{\mathrm{des}} \leq \eta^{\mathrm{des}}$ corresponds to showing that η_e^{des} monotonically decreases with ρ_e, i.e.,

$$\frac{\partial \eta_e^{\mathrm{des}}}{\partial \rho_e} \leq 0\,. \tag{5.23}$$

Since

$$\frac{\partial (\eta_e^{\mathrm{des}})^2}{\partial \rho_e} = 2\eta_e^{\mathrm{des}} \frac{\partial \eta_e^{\mathrm{des}}}{\partial \rho_e}\,, \tag{5.24}$$

and $\eta_e^{\mathrm{des}} \geq 0$, it is sufficient to show that

$$\frac{\partial (\eta_e^{\mathrm{des}})^2}{\partial \rho_e} \leq 0 \tag{5.25}$$

in order to show (5.23). The squared mixing parameter in (5.9) can be written as

$$\left(\eta_e^{\mathrm{des}}\right)^2 = \frac{\nu_1(\rho_e)}{\nu_2(\rho_e)}\,, \tag{5.26}$$

with

$$\nu_1(\rho_e) = \rho_e(\kappa - \gamma) + \alpha\,, \tag{5.27}$$

$$\nu_2(\rho_e) = \rho_e^2\beta - 2\rho_e\gamma + \alpha\,, \tag{5.28}$$

with

$$\kappa = \sqrt{\gamma^2 - \alpha\beta}\,, \tag{5.29}$$

and with α, β and γ defined in (3.56)–(3.58). Using the quotient rule to compute the partial derivative of (5.26) with respect to ρ_e gives

$$\frac{\partial(\eta_e^{\text{des}})^2}{\partial\rho_e} = \frac{\frac{\partial\nu_1(\rho_e)}{\partial\rho_e}\nu_2(\rho_e) - \frac{\partial\nu_2(\rho_e)}{\partial\rho_e}\nu_1(\rho_e)}{\nu_2^2(\rho_e)}\,. \tag{5.30}$$

Hence, since $\nu_2^2(\rho_e) \geq 0$, it is sufficient to show that

$$\zeta(\rho_e) = \frac{\partial\nu_1(\rho_e)}{\partial\rho_e}\nu_2(\rho_e) - \frac{\partial\nu_2(\rho_e)}{\partial\rho_e}\nu_1(\rho_e) \leq 0 \tag{5.31}$$

in order to proof that (5.25) holds. Computing the partial derivatives of (5.27) and (5.28) with respect to ρ_e gives

$$\frac{\partial\nu_1(\rho_e)}{\partial\rho_e} = \kappa - \gamma\,, \tag{5.32}$$

$$\frac{\partial\nu_2(\rho_e)}{\partial\rho_e} = 2\rho_e\beta - 2\gamma\,. \tag{5.33}$$

By substituting (5.27), (5.28), (5.32) and (5.33) in (5.31), $\zeta(\rho_e)$ can be written as a quadratic function of ρ_e, i.e.,

$$\zeta(\rho_e) = \psi_1\rho_e^2 - 2\psi_2\rho_e + \psi_3\,, \tag{5.34}$$

with

$$\psi_1 = (\gamma - \kappa)\beta\,, \tag{5.35}$$

$$\psi_2 = \alpha\beta = \gamma^2 - \kappa^2\,, \tag{5.36}$$

$$\psi_3 = \alpha(\kappa + \gamma)\,. \tag{5.37}$$

The extremum of the quadratic function $\zeta(\rho_e)$ in (5.34) can be found by setting $\frac{\partial\zeta(\rho_e)}{\partial\rho_e} = 0$, leading to

$$\tilde{\rho}_e = \frac{\psi_2}{\psi_1} = \frac{\alpha}{\gamma - \kappa} = \frac{\kappa + \gamma}{\beta}\,. \tag{5.38}$$

Substituting (5.38) in (5.34) yields

$$\zeta(\tilde{\rho}_e) = \frac{\psi_1\psi_3 - \psi_2^2}{\psi_1} = 0\,. \tag{5.39}$$

The second-order partial derivative of $\zeta(\rho_e)$ in (5.34) is equal to

$$\frac{\partial^2\zeta(\rho_e)}{\partial\rho_e^2} = 2\psi_1 = 2(\gamma-\kappa)\beta\,. \tag{5.40}$$

Since $\alpha \leq 0$ and $\beta \geq 0$ (see Section 3.3), using (5.29) it follows that

$$\alpha\beta = (\gamma-\kappa)(\gamma+\kappa) \leq 0\,. \tag{5.41}$$

We now consider two cases:

1. $\gamma \geq 0$: Since $\kappa \geq 0$, it follows that $\gamma+\kappa \geq 0$, such that $\gamma-\kappa \leq 0$ in order to satisfy (5.41).
2. $\gamma \leq 0$: Since $\kappa \geq 0$, it directly follows that $\gamma-\kappa \leq 0$.

Since $\gamma-\kappa \leq 0$ and $\beta \geq 0$, the second-order partial derivative in (5.40) is always negative (or equal to zero). Since the extremum is hence a maximum with function value 0, cf. (5.39), the quadratic function $\zeta(\rho_e)$ in (5.34) is negative (or zero) for all values of ρ_e. Hence, $\frac{\partial\eta_e^{\text{des}}}{\partial\rho_e} \leq 0$, such that

$$\boxed{\eta_e^{\text{des}} \leq \eta^{\text{des}}} \tag{5.42}$$

i.e., to achieve the same desired output MSC of the noise component a smaller mixing parameter can be used in the eBMVDR-N beamformer (incorporating an external microphone) than in the BMVDR-N beamformer (using only the head-mounted microphones). Together with the SNR results obtained in Section 5.3, this implies that for any arbitrary noise field and position of the desired source an external microphone enables to achieve the same spatial impression of the noise component while achieving a larger output SNR.

5.5 Experimental results

In Section 5.5.1 we first validate the analytical expressions derived in the previous sections using simulated anechoic ATFs for various positions of the external microphone. In Section 5.5.2 we provide experimental results using recorded signals in a reverberant environment with multiple interfering speakers as background noise, showing that also in a realistic scenario incorporating an external microphone enables to significantly increase the output SNR and decrease the mixing parameter required to obtain a desired output MSC of the noise component.

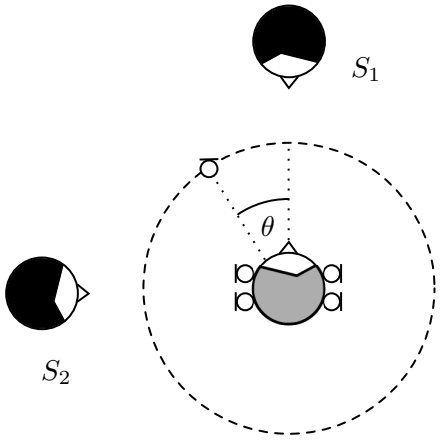

Fig. 5.2: Anechoic validation setup using 2 microphones on each side of the head. The external microphone was placed at 3 m distance to the listener for different angles θ. The desired source was placed at 3.5 m distance at two different angles, i.e., S_1 at 0° and S_2 at −90°.

5.5.1 *Validation using anechoic ATFs*

To validate our theoretical findings from Section 5.3 and Section 5.4 we simulated an anechoic acoustic scenario, where the head of the listener was modelled as a rigid sphere with diameter 17 cm [40]. Without considering any hearing devices, we considered 2 microphones on each side of the head, i.e., $M_H = 4$, with an inter-microphone distance of 7 mm, such that including the external microphone the total number of microphones was $M = 5$. The sample rate was equal to 16 kHz and all ATFs were simulated using an FFT length of 256 samples. Figure 5.2 depicts the validation setup. The external microphone was placed at a distance of 3 m to the listener, where the azimuth angle θ was varied from −180° to 180°. The desired source was placed at a distance of 3.5 m to the listener at two different angles, i.e., S_1 at 0° (in front) and S_2 at −90° (to the left). Hence, the smallest distance between the external microphone and the desired source was equal to 0.5 m, whereas the largest distance was equal to 6.5 m.

Using the simulated ATF vectors $\mathbf{a}_e$, the extended desired source covariance matrix $\mathbf{R}_{x,e}$ was calculated as in (2.46) with $p_{s_x} = 1$, i.e., assuming a flat spectrum. As background noise we considered 8 mutually independent noise sources with equal power at angles {−140°, −100°, −60°, −20°, 20°, 60°, 100°, 140°}, resulting in a diffuse-like noise field which is neither coherent nor perfectly diffuse. The extended noise covariance matrix $\mathbf{R}_{n,e}$ was calculated as the sum of the 8 corresponding (rank-1) covariance matrices, constructed using the simulated ATF vectors of the noise sources. As reference microphones we considered the front microphones on the left and the right side. The input SNR in the left reference microphone signal was set to 0 dB (averaged over all frequencies), leading to the input SNR in the external microphone

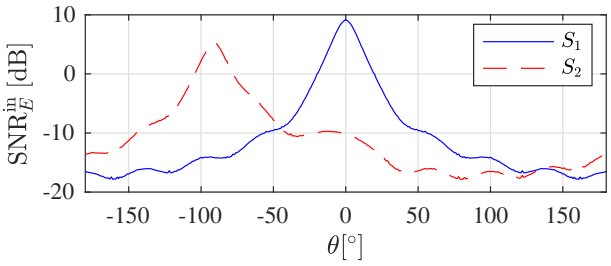

Fig. 5.3: Input SNR in the external microphone signal (averaged over all frequencies) for different angles θ of the external microphone for both considered positions of the desired source S_1 and S_2.

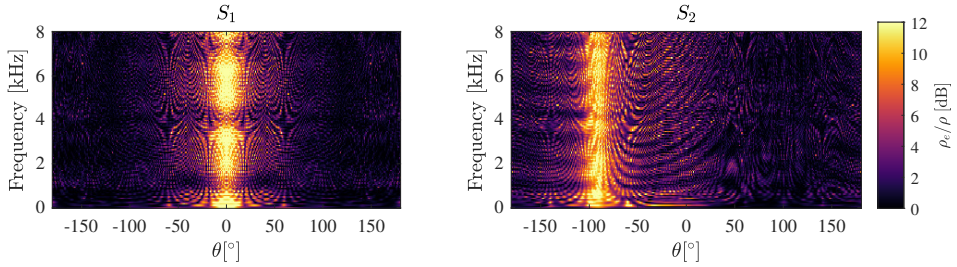

Fig. 5.4: Benefit of incorporating an external microphone in terms of output SNR (ρ_e/ρ) for different angles θ of the external microphone for (left) position S_1 and (right) position S_2.

signal (averaged over all frequencies) as depicted in Figure 5.3. As can be observed, the input SNR in the external microphone signal varied within a range of nearly 30 dB, with the highest input SNR occuring when the external microphone is closest to the desired source (i.e., 0° for S_1 and $-90°$ for S_2).

BMVDR BEAMFORMER VS. EBMVDR BEAMFORMER In Section 5.3, we showed that the output SNR of the eBMVDR beamformer ρ_e in (5.8) is always larger than or equal to the output SNR of the BMVDR beamformer ρ in (5.3). Figure 5.4 depicts the benefit of incorporating an external microphone in terms of the output SNR ratio, i.e.,

$$\frac{\rho_e}{\rho} = \frac{\mathbf{a}_e^H \mathbf{R}_{n,e}^{-1} \mathbf{a}_e}{\mathbf{a}^H \mathbf{R}_n^{-1} \mathbf{a}} \tag{5.43}$$

for different angles of the external microphone and for both considered positions of the desired source. As can be observed, for all positions of the external microphone and the desired source and for all frequencies $\rho_e \geq \rho$, hence satisfying (5.14). Moreover, the benefit of incorporating the external microphone is larger for small distances between the desired source and the external microphone, in this case leading to an SNR improvement of more than 12 dB.

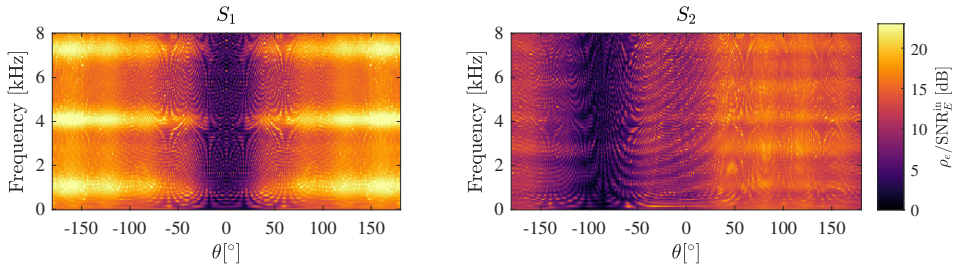

Fig. 5.5: Benefit of incorporating the external microphone in the BMVDR beamformer compared to directly using the external microphone signal for different angles θ of the external microphone for (left) position S_1 and (right) position S_2.

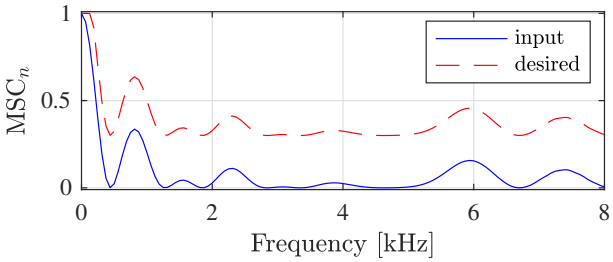

Fig. 5.6: Input MSC and desired output MSC of the noise component, limiting the MSC error to 0.3.

EBMVDR BEAMFORMER VS. EXTERNAL MICROPHONE SIGNAL Figure 5.5 depicts the benefit of incorporating the external microphone in the BMVDR beamformer compared to directly using the external microphone signal in terms of the SNR ratio $\rho_e/\mathrm{SNR}_E^{\mathrm{in}}$. As can be observed, for all positions of the external microphone and the desired source and for all frequencies it was beneficial to incorporate the external microphone in the BMVDR beamformer, i.e., $\rho_e > \mathrm{SNR}_E^{\mathrm{in}}$. The benefit is largest if the external microphone is far away from the desired source and hence the input SNR in the external microphone signal is low, but even for smaller distances the benefit is in the range of a few dB and hence valuable.

BMVDR-N BEAMFORMER VS. EBMVDR-N BEAMFORMER In Section 5.4, we showed that to achieve the same desired output MSC of the noise component the mixing parameter η_e^{des} in (5.9) of the eBMVDR-N beamformer is smaller than (or equal to) the mixing parameter η^{des} in (3.55) of the BMVDR-N beamformer, hence leading to a larger output SNR. Here, we set the desired output MSC of the noise component equal to

$$\mathrm{MSC}_n^{\mathrm{des}} = \min\left(1,\, \mathrm{MSC}_n^{\mathrm{in}} + 0.3\right) , \tag{5.44}$$

hence limiting the MSC error to 0.3 for all frequencies and satisfying (3.54). Figure 5.6 depicts the input MSC of the noise component, computed using (2.82) and

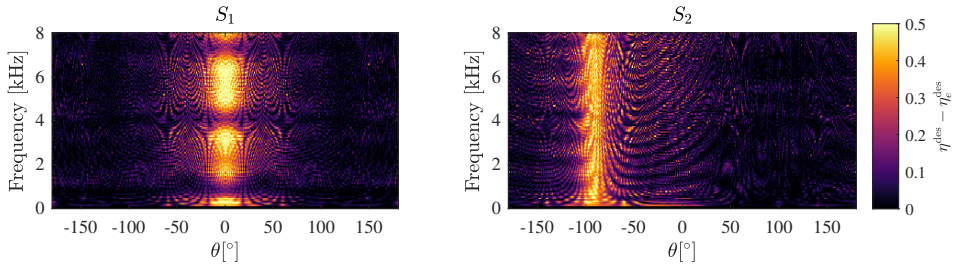

Fig. 5.7: Difference between the mixing parameter η^{des} of the BMVDR-N beamformer and the mixing parameter η_e^{des} of the eBMVDR-N beamformer, leading to the same desired output MSC of the noise component, for different angles θ of the external microphone for (left) position S_1 and (right) position S_2.

(2.84), and the desired output MSC of the noise component in (5.44). It can be observed that the input MSC of the noise component resembles a squared (modified) sinc function, as expected for a diffuse-like noise field and modelling the head as a rigid sphere [31] (cf. (3.63)). Figure 5.7 depicts the benefit of incorporating the external microphone in terms of the difference between the mixing parameters, i.e., $\eta^{\text{des}} - \eta_e^{\text{des}}$, both leading to the same desired output MSC of the noise component. As can be observed, for all positions of the external microphone and the desired source and for all frequencies $\eta_e^{\text{des}} \leq \eta^{\text{des}}$. As proven in Section 5.4, by comparing Figure 5.4 and Figure 5.7 it can be observed that a larger output SNR ρ_e of the eBMVDR beamformer leads to a smaller mixing parameter η_e^{des} of the eBMVDR-N beamformer and hence an improved trade off compared to the mixing parameter η^{des} of the BMVDR-N beamformer. In other words, one has to mix less with the noisy reference microphone signals for the eBMVDR-N beamformer than for the BMVDR-N beamformer, while both beamformers lead to the same spatial impression of the noise component. This effect is larger when the external microphone is close to the desired source, leading to mixing parameter differences that are larger than 0.5.

5.5.2 *Experimental results*

For a more realistic evaluation we used a database with recorded signals in a real-world reverberant environment [36]. The experimental setup is depicted in Figure 5.8, where a listener and two different speakers were sitting at a circular table with a diameter of 106 cm in a room with size $12.7 \times 10 \times 3.6$ m^3 and a reverberation time of about 620 ms. The setup was surrounded by three layers of in total 56 seated persons producing realistic multi-talker babble noise. Hence, the noise component was diffuse-like, but also contained temporally coherent sources and sensor noise (from the microphones and the recording equipment). The (male) speaker S_1 sat in front of the listener at the other end of the table, while the (female) speaker S_2 sat to the right of the listener. The listener was wearing $M_H = 4$ head-mounted hearing aid microphones, i.e., two microphones on each side. For the external microphone we

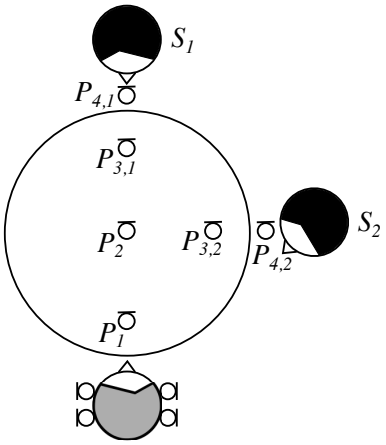

Fig. 5.8: Experimental realistic setup with a listener wearing head-mounted hearing aid microphones, two different speaker positions (S_1 and S_2) and several possible positions of the external microphone. The setup was surrounded by 56 persons producing realistic multi-talker babble noise.

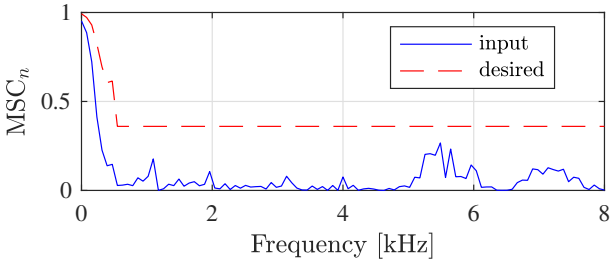

Fig. 5.9: Measured input MSC of the noise component and (psycho-acoustically motivated) desired output MSC of the noise component.

selected several realistic positions from the database, e.g., representing the listener's smartphone on the table (P_1), a microphone of a conference system in the center of the table (P_2), the smartphone of each speaker placed on the table ($P_{3,1}$ for speaker S_1 and $P_{3,2}$ for speaker S_2) and a headset worn by each speaker ($P_{4,1}$ for speaker S_1 and $P_{4,2}$ for speaker S_2). Only one speaker was active at a time and read 12 sentences for about 25 s, while the listener tried to sit as still as possible (but small movements occurred).

We used separate recordings of the speakers and the background noise and mixed them at an intelligibility-weighted SNR (iSNR) for the right reference microphone signal $\mathrm{iSNR}_R^{\mathrm{in}} = 0\,\mathrm{dB}$. The iSNR is computed by weighting the SNR in each frequency band with a function that takes into account the importance of each frequency band for speech intelligibility [30, 189]. We used a sample rate of 16 kHz and processed the signals in an STFT framework with $T_d = 1024$, $T_s = 512$ and a square-root Hann window. For the extended BMVDR beamformers using all micro-

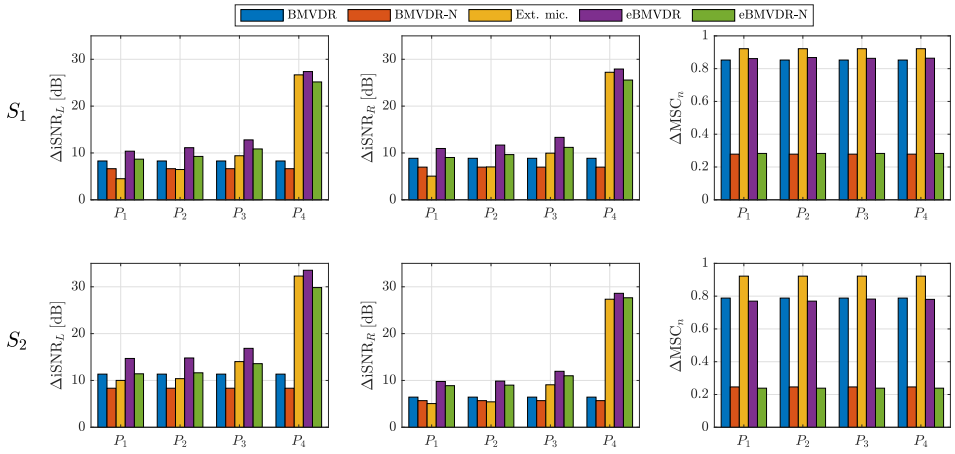

Fig. 5.10: Intelligibility-weighted SNR improvement ΔiSNR_L and ΔiSNR_R, and the MSC error ΔMSC_n of the noise component for the BMVDR beamformer, the BMVDR-N beamformer, the external microphone signal, the eBMVDR beamformer and the eBMVDR-N beamformer, for the different positions of the external microphone. The results are shown for speaker S_1 (top row) and speaker S_2 (bottom row).

phones (i.e., eBMVDR in (5.4) and (5.5), and eBMVDR-N in (5.6) and (5.7)) the extended noisy input covariance matrix $\mathbf{R}_{y,e}$ and the extended noise covariance matrix $\mathbf{R}_{n,e}$ were recursively averaged as in (3.68) and (3.69), using a smoothing factor of $\alpha = 0.9$, corresponding to about 300 ms (cf. (3.71)). To distinguish between speech-plus-noise and noise-only frames we used a thresholded speech presence probability (SPP) estimate in the right reference microphone signal, where we used the SPP estimation method proposed in [27] (cf. Section 3.4.1). The (time-varying) extended RTF vectors $\mathbf{a}_{L,e}$ and $\mathbf{a}_{R,e}$ were estimated from the estimated extended covariance matrices $\mathbf{R}_{y,e}$ and $\mathbf{R}_{n,e}$ using the covariance whitening method, i.e., as the principal eigenvector of the pre-whitened extended noisy input covariance matrix $\mathbf{R}_{n,e}^{-1}\mathbf{R}_{y,e}$ [85, 89, 90, 93] (cf. Section 3.4.2). For the BMVDR beamformers using only the head-mounted microphones (i.e., BMVDR in (3.3) and (3.4), and BMVDR-N in (3.39) and (3.40)), the covariance matrices $\mathbf{R}_y$ and $\mathbf{R}_n$ were constructed by discarding the last row and the last column of $\mathbf{R}_{y,e}$ and $\mathbf{R}_{n,e}$, respectively. The (time-varying) RTF vectors $\mathbf{a}_L$ and $\mathbf{a}_R$ were estimated from the estimated covariance matrices $\mathbf{R}_y$ and $\mathbf{R}_n$, also using the covariance whitening method.

For the beamformers with partial noise estimation (BMVDR-N and eBMVDR-N), the desired output MSC of the noise component $\text{MSC}_n^{\text{des}}$ was set in a psychoacoustically motivated way by constraining the output MSC of the noise component by means of frequency-dependent lower and upper boundaries such that the listener's spatial impression of a diffuse noise field should not be altered [11, 13, 106]. These boundaries were defined based on the IC discrimination ability of the human auditory system in diffuse noise fields [165, 166]. Below 500 Hz, the MSC boundaries were chosen as a function of the desired output MSC of the noise component

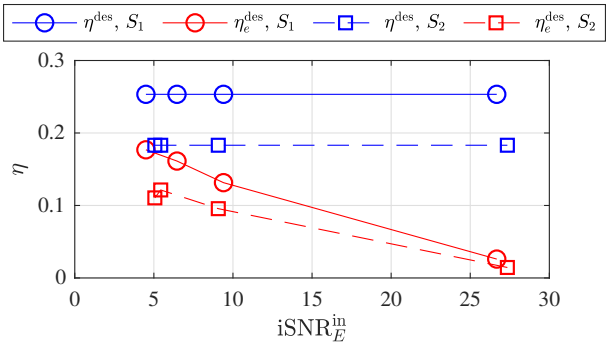

Fig. 5.11: The mixing parameters η_e^{des} and η^{des} (averaged over all frequencies) leading to the desired output MSC of the noise component for the different external microphone positions, mapped to the respective input iSNRs in the external microphone signal.

$\text{MSC}_n^{\text{des}}$, whereas above 500 Hz as a fixed lower MSC boundary of 0 and fixed upper MSC boundary of 0.36. Figure 5.9 depicts the frequency-dependent long-term input MSC of the noise component and the (psycho-acoustically motivated) frequency-dependent desired output MSC $\text{MSC}_n^{\text{des}}$ of the noise component. Similarly as in Figure 5.6, it can be observed that the input MSC of the noise component resembles a squared (modified) sinc function. The mixing parameters η^{des} and η_e^{des} of the BMVDR-N and eBMVDR-N beamformers were computed using (3.55) and (5.9) based on the estimated (extended) RTF vectors and the estimated (extended) noise covariance matrix.

To evaluate the algorithms in terms of noise reduction performance and preservation of the spatial impression of the noise component, based on the output signals we used the left and the right iSNR improvement, relating the left and the right output iSNR to the left and the right input iSNR, i.e.,

$$\Delta\text{iSNR}_L = \text{iSNR}_L^{\text{out}} - \text{iSNR}_L^{\text{in}}\,, \tag{5.45}$$

$$\Delta\text{iSNR}_R = \text{iSNR}_R^{\text{out}} - \text{iSNR}_R^{\text{in}}\,, \tag{5.46}$$

and the broadband MSC error ΔMSC_n of the noise component, i.e.,

$$\Delta\text{MSC}_n = \frac{1}{F}\sum_{f=1}^{F} |\text{MSC}_n^{\text{out}}(f) - \text{MSC}_n^{\text{in}}(f)|\,, \tag{5.47}$$

where f is the frequency bin index and F is the total number of frequency bins.

Figure 5.10 depicts the iSNR improvement and the MSC error of the noise component for the BMVDR beamformer, the BMVDR-N beamformer, the external microphone signal, the eBMVDR beamformer and the eBMVDR-N beamformer, for the different positions of the external microphone. The top row shows the results for speaker S_1, while the bottom row shows the results for speaker S_2. Considering

the iSNR improvements for speaker S_1, it can be observed that incorporating the external microphone signal in the binaural noise reduction algorithms is beneficial for all positions of the external microphone. More in particular, in terms of iSNR improvement the eBMVDR beamformer always outperforms the BMVDR beamformer and the eBMVDR-N beamformer always outperforms the BMVDR-N beamformer. The iSNR improvement is similar in both hearing devices due to the symmetric scenario for speaker S_1. As expected, the iSNR improvement increases for decreasing distance between the external microphone and speaker S_1 with a very large iSNR improvement for position P_4 (headset microphone). The eBMVDR beamformer outperforms the external microphone signal for all considered positions, whereas the eBMVDR-N beamformer outperforms the external microphone signal for all considered positions except for position P_4. In contrast, the BMVDR beamformer and the BMVDR-N beamformer outperform the external microphone signal only for positions P_1 and P_2, i.e., for the position close to the listener and in the center of the table. Comparing the eBMVDR beamformer to the eBMVDR-N beamformer, it appears that the drop in iSNR improvement for the eBMVDR-N beamformer due to mixing with the noisy reference microphone signals is approximately the same for all positions of the external microphone.

Considering the MSC error ΔMSC_n of the noise component for speaker S_1, as expected only the binaural noise reduction algorithms with partial noise estimation, i.e., the BMVDR-N beamformer and the eBMVDR-N beamformer, are able to yield a low MSC error and hence preserve the spatial impression of the noise component. The external microphone signal obviously shows the worst performance in terms of MSC error of the noise component since the external microphone signal is just a monaural signal that does not include any binaural cues, hence leading to in-head localization.

Considering the results for speaker S_2 (bottom row), it can be observed that similar results as for speaker S_1 are obtained. However, due to the asymmetric setup, the iSNR improvement at the right side (better ear with larger input iSNR) is always smaller than the iSNR improvement at the left side. In addition, the drop in iSNR improvement for the eBMVDR-N beamformer compared to the eBMVDR beamformer is different for the left and the right side but remains approximately constant for the different positions of the external microphone.

For both speakers S_1 and S_2, Figure 5.11 depicts the mixing parameters η_e^{des} and η^{des} (averaged over all frequencies) of the eBMVDR-N beamformer and the BMVDR-N beamformer, which lead to the desired output MSC $\text{MSC}_n^{\text{des}}$ of the noise component. The mixing parameters are plotted as a function of the input iSNR in the external microphone signal $\text{iSNR}_E^{\text{in}}$ for the different external microphone positions. It can be observed that the mixing parameter is always smaller for the eBMVDR-N beamformer than for the BMVDR-N beamformer and decreases with increasing input iSNR in the external microphone signal. Further, the mixing parameter is always smaller for speaker S_2 than for speaker S_1, i.e., if the speaker is not positioned in front of the listener.

In conclusion, the experimental results in this section showed that for all considered positions of the external microphone and the speaker the iSNR improvement is larger for the eBMVDR-N beamformer (incorporating the external microphone) than for the BMVDR-N beamformer (using only the head-mounted microphones) and for the external microphone signal (except for P_4). In addition, the mixing parameter leading to the same desired output MSC of the noise component is always smaller for the eBMVDR-N beamformer than for the BMVDR-N beamformer. All experimental results in this section are in line with the theoretical findings of the previous sections.

5.6 Summary

In this chapter we analytically showed for an arbitrary noise field and without making any assumptions about the position of the desired source that by incorporating an external microphone in the BMVDR-N beamformer 1) a larger output SNR can be obtained for the same mixing parameter, 2) the same output SNR can be obtained for a larger mixing parameter, and 3) the same desired output MSC of the noise component can be obtained for a smaller mixing parameter. The obtained analytical expressions were first validated using simulated anechoic acoustic transfer functions. In addition, experimental results using recorded signals in a realistic reverberant environment showed that incorporating an external microphone in the BMVDR-N beamformer enables to significantly increase the output SNR compared to using only the head-mounted microphone signals while preserving the spatial impression of the noise component. While in this chapter we analyzed and experimentally investigated the incorporation of an external microphone in the BMVDR and BMVDR-N beamformers, in the next chapter we propose computationally efficient methods to estimate the RTF vectors of the desired source by exploiting one or multiple external microphones.

6

RTF VECTOR ESTIMATION EXPLOITING EXTERNAL MICROPHONES

As discussed in Chapter 3, an important parameter required for calculating the filter vectors of the considered binaural beamforming algorithms (with and without external microphones) is the steering vector, e.g., the relative transfer function (RTF) vector of the desired source. In this chapter we consider the multi-extended binaural hearing device configuration (cf. Chapter 2) and propose computationally efficient methods to estimate the RTF vectors of the desired source by exploiting one or multiple external microphones. The external microphones are assumed to be spatially separated from the head-mounted microphones, such that the spatial coherence (SC) between the noise component in the head-mounted microphone signals and the noise component in the external microphone signals is low. We first consider the extended binaural hearing device configuration with only one external microphone and propose an SC-based RTF vector estimation method, which estimates the RTF vectors of the desired source as the last column of the extended noisy input covariance matrix (corresponding to the external microphone), normalized by the element corresponding to the reference microphone. Assuming the SC between the

This chapter is partly based on:

[159] 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, Sep. 2018, pp. 146–150. **Best student paper award**

[160] 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, Oct. 2018, pp. 106-110.

[161] 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, pp. 416–420.

[162] 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, NY, USA, Oct. 2019, pp. 368–372.

noise components to be zero, we show that the SC-based method yields an unbiased estimate of the elements of the RTF vectors corresponding to the head-mounted microphones, while the element corresponding to the external microphone is biased. We provide a detailed bias analysis for an arbitrary noise field, a diffuse noise field and an interfering source. Next, we consider the multi-extended binaural hearing device configuration with more than one external microphone and show that different RTF vector estimates can be obtained by using the SC-based method for each external microphone. We propose several procedures to combine these RTF vector estimates, either by selecting the estimate corresponding to the highest input SNR, by averaging the estimates or by combining the estimates in order to maximize the output SNR of the eBMVDR beamformer filtering all microphone signals (see Chapter 5). Experimental results using recorded signals of a moving desired source for a binaural hearing device configuration with one or more external microphones in a reverberant environment show that the proposed SC-based method outperforms state-of-the-art RTF vector estimation methods in terms of noise reduction performance when used to steer the (e)BMVDR beamformer. In addition, the experimental results show that the output SNR-maximizing combination procedure of different RTF vector estimates yields the largest SNR improvement.

This chapter is structured as follows. In Section 6.1 we consider one external microphone and introduce the SC-based method to estimate the RTF vectors of the desired source. In Section 6.2 we present a bias analysis of the SC-based method for different noise fields. In Section 6.3 we present an extension of the SC-based method for multiple external microphones by linearly combining multiple RTF vector estimates. In Section 6.4 we evaluate the performance of the proposed RTF vector estimation methods for realistic acoustic scenarios with, e.g., a moving desired source and diffuse-like background noise.

6.1 SC-based RTF vector estimation using one external microphone

In this section we consider the extended binaural hearing device configuration with one external microphone, as depicted in Figure 2.2. In addition, let us for now assume that the acoustic scenario only consists of a desired source and background noise, i.e.,

$$\mathbf{y}_e = \mathbf{x}_e + \mathbf{n}_e \,, \tag{6.1}$$

such that the extended noisy input covariance matrix in (2.50) is equal to

$$\mathbf{R}_{y,e} = \mathbf{R}_{x,e} + \mathbf{R}_{n,e} \,. \tag{6.2}$$

Similarly as the left and the right RTF vector of the desired source in (3.73) and (3.74), the left and the right extended RTF vector of the desired source $\mathbf{a}_{L,e}$ and $\mathbf{a}_{R,e}$ in (2.44) can be written as

$$\mathbf{a}_{L,e} = \frac{\mathbf{R}_{x,e}\mathbf{e}_m}{\mathbf{e}_L^T\mathbf{R}_{x,e}\mathbf{e}_m}\,, \tag{6.3}$$

$$\mathbf{a}_{R,e} = \frac{\mathbf{R}_{x,e}\mathbf{e}_m}{\mathbf{e}_R^T\mathbf{R}_{x,e}\mathbf{e}_m}\,, \tag{6.4}$$

with $m \in \{1, \ldots, M\}$, i.e., as *any* column of the extended desired source covariance matrix $\mathbf{R}_{x,e}$, normalized by the element corresponding to the respective reference microphone.

We now make the fundamental assumption that the noise component in the head-mounted microphone signals $\mathbf{n}$ is uncorrelated with the noise component in the external microphone signal n_E, i.e.,

$$\boxed{\mathbf{r}_{n,E} = \mathcal{E}\{\mathbf{n}n_E^*\} = \mathbf{0}_{M_H}} \tag{6.5}$$

with $\mathbf{0}_{M_H}$ an M_H-dimensional zero vector. Hence, the extended noise covariance matrix $\mathbf{R}_{n,e}$ in (2.51) is equal to

$$\mathbf{R}_{n,e} = \left[\begin{array}{c|c} \mathbf{R}_n & \mathbf{0}_{M_H} \\ \hline \mathbf{0}_{M_H}^H & p_{n_E} \end{array}\right], \tag{6.6}$$

such that

$$\mathbf{R}_{n,e}\mathbf{e}_E = p_{n_E}\mathbf{e}_E\,, \tag{6.7}$$

with $\mathbf{e}_E$ and M-dimensional selection vector with $\mathbf{e}_E(M) = 1$ and p_{n_E} the PSD of the noise component in the external microphone signal. If the assumption in (6.5) holds, it can be easily shown that the covariance between the head-mounted microphone signals and the external microphone signal is equal to the covariance between the desired source components in these microphone signals, i.e.,

$$\mathcal{E}\{\mathbf{y}y_E^*\} = \mathcal{E}\{(\mathbf{x}+\mathbf{n})(x_E^* + n_E^*)\} = \mathcal{E}\{\mathbf{x}x_E^*\}\,. \tag{6.8}$$

Using (6.6) and (6.8) in (6.2), the extended noisy input covariance matrix $\mathbf{R}_{y,e}$ can be written as

$$\mathbf{R}_{y,e} = \left[\begin{array}{c|c} \mathbf{R}_x + \mathbf{R}_n & \mathcal{E}\{\mathbf{x}x_E^*\} \\ \hline \mathcal{E}\{\mathbf{x}^H x_E\} & p_{x_E} + p_{n_E} \end{array}\right]. \tag{6.9}$$

It can be observed that the last column (corresponding to the external microphone) contains only information about the desired source component, except for the last

element (corresponding to the external microphone). Using (6.2) and (6.6), it can easily be shown that

$$\mathbf{R}_{y,e}\mathbf{e}_E = \mathbf{R}_{x,e}\mathbf{e}_E + \mathbf{R}_{n,e}\mathbf{e}_E = \mathbf{R}_{x,e}\mathbf{e}_E + p_{n_E}\mathbf{e}_E\,, \tag{6.10}$$

$$\mathbf{e}_L^T\mathbf{R}_{y,e}\mathbf{e}_E = \mathbf{e}_L^T\mathbf{R}_{x,e}\mathbf{e}_E + \mathbf{e}_L^T p_{n_E}\mathbf{e}_E = \mathbf{e}_L^T\mathbf{R}_{x,e}\mathbf{e}_E\,, \tag{6.11}$$

$$\mathbf{e}_R^T\mathbf{R}_{y,e}\mathbf{e}_E = \mathbf{e}_R^T\mathbf{R}_{x,e}\mathbf{e}_E + \mathbf{e}_R^T p_{n_E}\mathbf{e}_E = \mathbf{e}_R^T\mathbf{R}_{x,e}\mathbf{e}_E\,, \tag{6.12}$$

such that, using (6.3) and (6.4) with $m = M$,

$$\frac{\mathbf{R}_{y,e}\mathbf{e}_E}{\mathbf{e}_L^T\mathbf{R}_{y,e}\mathbf{e}_E} = \frac{\mathbf{R}_{x,e}\mathbf{e}_E + p_{n_E}\mathbf{e}_E}{\mathbf{e}_L^T\mathbf{R}_{x,e}\mathbf{e}_E} = \mathbf{a}_{L,e} + \frac{p_{n_E}}{\mathbf{e}_L^T\mathbf{R}_{x,e}\mathbf{e}_E}\mathbf{e}_E\,, \tag{6.13}$$

$$\frac{\mathbf{R}_{y,e}\mathbf{e}_E}{\mathbf{e}_R^T\mathbf{R}_{y,e}\mathbf{e}_E} = \frac{\mathbf{R}_{x,e}\mathbf{e}_E + p_{n_E}\mathbf{e}_E}{\mathbf{e}_R^T\mathbf{R}_{x,e}\mathbf{e}_E} = \mathbf{a}_{R,e} + \frac{p_{n_E}}{\mathbf{e}_R^T\mathbf{R}_{x,e}\mathbf{e}_E}\mathbf{e}_E\,. \tag{6.14}$$

We now define the SC-based estimates of the left and the right extended RTF vector of the desired source $\mathbf{a}_{L,e}$ and $\mathbf{a}_{R,e}$ (including the external microphone) as

$$\boxed{\mathbf{a}_{L,e}^{\mathrm{SC}} = \frac{\mathbf{R}_{y,e}\mathbf{e}_E}{\mathbf{e}_L^T\mathbf{R}_{y,e}\mathbf{e}_E}} \tag{6.15}$$

$$\boxed{\mathbf{a}_{R,e}^{\mathrm{SC}} = \frac{\mathbf{R}_{y,e}\mathbf{e}_E}{\mathbf{e}_R^T\mathbf{R}_{y,e}\mathbf{e}_E}} \tag{6.16}$$

i.e., as the *last* column of the extended noisy input covariance matrix $\mathbf{R}_{y,e}$, normalized by the element corresponding to the respective reference microphone. Using (6.13) and (6.14), it can be easily shown that the first M_H elements of $\mathbf{a}_{L,e}^{\mathrm{SC}}$ and $\mathbf{a}_{R,e}^{\mathrm{SC}}$ in (6.15) and (6.16) are equal to the left and the right RTF vector of the desired source $\mathbf{a}_L$ and $\mathbf{a}_R$ (without the external microphone), i.e.,

$$\mathbf{a}_L^{\mathrm{SC}} = [\mathbf{I}_{M_H},\ \mathbf{0}_{M_H}]\,\mathbf{a}_{L,e}^{\mathrm{SC}} = \mathbf{a}_L\,, \tag{6.17}$$

$$\mathbf{a}_R^{\mathrm{SC}} = [\mathbf{I}_{M_H},\ \mathbf{0}_{M_H}]\,\mathbf{a}_{R,e}^{\mathrm{SC}} = \mathbf{a}_R\,, \tag{6.18}$$

with $\mathbf{I}_{M_H}$ the $(M_H \times M_H)$-dimensional identity matrix. However, from (6.13) and (6.14) it can also be seen that the last element of $\mathbf{a}_{L,e}^{\mathrm{SC}}$ and $\mathbf{a}_{R,e}^{\mathrm{SC}}$ in (6.15) and (6.16) is not equal to the last element of $\mathbf{a}_{L,e}$ and $\mathbf{a}_{R,e}$, but is corrupted by a bias term (even when the assumption in (6.5) perfectly holds). A more general analysis of the bias of the SC-based estimates in (6.15) and (6.16) is provided in the next section.

In practice, an estimate of the extended noisy input covariance matrix $\hat{\mathbf{R}}_{y,e}$ is used in (6.15) and (6.16), i.e.,

$$\hat{\mathbf{a}}_{L,e}^{\mathrm{SC}} = \frac{\hat{\mathbf{R}}_{y,e}\mathbf{e}_E}{\mathbf{e}_L^T\hat{\mathbf{R}}_{y,e}\mathbf{e}_E}\,, \tag{6.19}$$

$$\hat{\mathbf{a}}_{R,e}^{\mathrm{SC}} = \frac{\hat{\mathbf{R}}_{y,e}\mathbf{e}_E}{\mathbf{e}_R^T\hat{\mathbf{R}}_{y,e}\mathbf{e}_E}\,, \tag{6.20}$$

where $\hat{\mathbf{R}}_{y,e}$ can be easily estimated from the microphone signals, e.g., similar to the online estimator in (3.68). Using (6.17) and (6.18), the first M_H elements of $\hat{\mathbf{a}}_{L,e}^{\mathrm{SC}}$ and $\hat{\mathbf{a}}_{R,e}^{\mathrm{SC}}$ in (6.19) and (6.20) correspond to an SC-based estimate of the left and the right RTF vector of the desired source $\hat{\mathbf{a}}_L^{\mathrm{SC}}$ and $\hat{\mathbf{a}}_R^{\mathrm{SC}}$ (without external microphone), i.e.,

$$\hat{\mathbf{a}}_L^{\mathrm{SC}} = [\mathbf{I}_{M_H},\ \mathbf{0}_{M_H}]\,\hat{\mathbf{a}}_{L,e}^{\mathrm{SC}}\,, \tag{6.21}$$

$$\hat{\mathbf{a}}_R^{\mathrm{SC}} = [\mathbf{I}_{M_H},\ \mathbf{0}_{M_H}]\,\hat{\mathbf{a}}_{R,e}^{\mathrm{SC}}\,. \tag{6.22}$$

It should be noted that the proposed SC-based estimators have a low computational complexity and do not require an estimate of a noise covariance matrix, but obviously require an external microphone signal to be transmitted to the head-mounted hearing devices. As already mentioned, transmission aspects such as synchronization are outside the scope of this thesis.

6.2 Bias analysis of the SC-based RTF vector estimates

In this section we provide a theoretical bias analysis of the SC-based estimates in (6.15) and (6.16). Section 6.2.1 derives general expressions of the multiplicative and additive bias for an arbitrary noise field. Section 6.2.2 considers the special case of a diffuse noise field, while Section 6.2.3 considers the special case of an interfering source.

6.2.1 *Arbitrary noise field*

In this section we consider an acoustic scenario with an arbitrary noise field (e.g., combination of diffuse noise and an interfering speaker), i.e.,

$$\mathbf{y}_e = \mathbf{x}_e + \mathbf{v}_e\,, \tag{6.23}$$

such that the extended noisy input covariance matrix is equal to

$$\mathbf{R}_{y,e} = \mathbf{R}_{x,e} + \mathbf{R}_{v,e}\,. \tag{6.24}$$

Substituting (6.24) in (6.15) and (6.16), the SC-based estimates of the left and the right extended RTF vector $\mathbf{a}_{L,e}^{\mathrm{SC}}$ and $\mathbf{a}_{R,e}^{\mathrm{SC}}$ are equal to

$$\mathbf{a}_{L,e}^{\mathrm{SC}} = \frac{(\mathbf{R}_{x,e}+\mathbf{R}_{v,e})\mathbf{e}_E}{\mathbf{e}_L^T(\mathbf{R}_{x,e}+\mathbf{R}_{v,e})\mathbf{e}_E} = \frac{p_{x_L}\mathbf{a}_{L,e}\mathbf{a}_{L,e}^H\mathbf{e}_E + \mathbf{R}_{v,e}\mathbf{e}_E}{\mathbf{e}_L^T\mathbf{R}_{x,e}\mathbf{e}_E + \mathbf{e}_L^T\mathbf{R}_{v,e}\mathbf{e}_E}\,, \tag{6.25}$$

$$\mathbf{a}_{R,e}^{\mathrm{SC}} = \frac{(\mathbf{R}_{x,e}+\mathbf{R}_{v,e})\mathbf{e}_E}{\mathbf{e}_R^T(\mathbf{R}_{x,e}+\mathbf{R}_{v,e})\mathbf{e}_E} = \frac{p_{x_R}\mathbf{a}_{R,e}\mathbf{a}_{R,e}^H\mathbf{e}_E + \mathbf{R}_{v,e}\mathbf{e}_E}{\mathbf{e}_R^T\mathbf{R}_{x,e}\mathbf{e}_E + \mathbf{e}_R^T\mathbf{R}_{v,e}\mathbf{e}_E}\,, \tag{6.26}$$

which can be written as

$$\mathbf{a}_{L,e}^{\mathrm{SC}} = \mathbf{a}_{L,e}\epsilon_L^{\mathrm{mult}} + \boldsymbol{\epsilon}_L^{\mathrm{add}}\,, \tag{6.27}$$

$$\mathbf{a}_{R,e}^{\mathrm{SC}} = \mathbf{a}_{R,e}\epsilon_R^{\mathrm{mult}} + \boldsymbol{\epsilon}_R^{\mathrm{add}}\,, \tag{6.28}$$

where the left and the right multiplicative bias factor are equal to

$$\epsilon_L^{\mathrm{mult}} = \frac{1}{1+\frac{\mathbf{e}_L^T\mathbf{R}_{v,e}\mathbf{e}_E}{\mathbf{e}_L^T\mathbf{R}_{x,e}\mathbf{e}_E}}\,, \tag{6.29}$$

$$\epsilon_R^{\mathrm{mult}} = \frac{1}{1+\frac{\mathbf{e}_R^T\mathbf{R}_{v,e}\mathbf{e}_E}{\mathbf{e}_R^T\mathbf{R}_{x,e}\mathbf{e}_E}}\,, \tag{6.30}$$

and the left and the right additive bias vector are equal to

$$\boldsymbol{\epsilon}_L^{\mathrm{add}} = \frac{\mathbf{R}_{v,e}\mathbf{e}_E}{\mathbf{e}_L^T\mathbf{R}_{x,e}\mathbf{e}_E + \mathbf{e}_L^T\mathbf{R}_{v,e}\mathbf{e}_E}\,, \tag{6.31}$$

$$\boldsymbol{\epsilon}_R^{\mathrm{add}} = \frac{\mathbf{R}_{v,e}\mathbf{e}_E}{\mathbf{e}_R^T\mathbf{R}_{x,e}\mathbf{e}_E + \mathbf{e}_R^T\mathbf{R}_{v,e}\mathbf{e}_E}\,. \tag{6.32}$$

The multiplicative bias factors in (6.29) and (6.30) depend on the ratio between the CPSD of the undesired component in the reference microphone signals and the external microphone signal and the CPSD of the desired source component in the reference microphone signals and the external microphone signal. The additive bias vectors in (6.31) and (6.32) also depend on this CPSD ratio but additionally on the overall correlation between the undesired component in the head-mounted microphone signals and the undesired component in the external microphone signal, i.e., $\mathbf{R}_{v,e}\mathbf{e}_E$.

If the undesired component in the head-mounted microphone signals is uncorrelated with the undesired component in the external microphone signal, as assumed in the derivation of the SC-based method in Section 6.1, i.e., $\mathbf{R}_{v,e}\mathbf{e}_E = p_{v_E}\mathbf{e}_E$, it can be easily shown that the multiplicative bias factors are equal to 1, i.e.,

$$\epsilon_L^{\mathrm{mult}} = \epsilon_R^{\mathrm{mult}} = 1\,, \tag{6.33}$$

and the additive bias vectors are equal to

$$\boldsymbol{\epsilon}_L^{\text{add}} = \frac{p_{v_E}}{\mathbf{e}_L^T \mathbf{R}_{x,e} \mathbf{e}_E} \mathbf{e}_E \,, \tag{6.34}$$

$$\boldsymbol{\epsilon}_R^{\text{add}} = \frac{p_{v_E}}{\mathbf{e}_R^T \mathbf{R}_{x,e} \mathbf{e}_E} \mathbf{e}_E \,. \tag{6.35}$$

This means that only the last element of the SC-based RTF vector estimates, corresponding to the external microphone, is biased. Using (6.33), (6.34) and (6.35) in (6.27) and (6.28), it can be shown that this element is in this case corrupted by a real-valued multiplicative bias [161], i.e.,

$$\mathbf{e}_E^T \mathbf{a}_{L,e}^{\text{SC}} = \mathbf{e}_E^T \mathbf{a}_{L,e} \left(1 + \frac{p_{v_E}}{p_{x_L} |\mathbf{e}_E^T \mathbf{a}_{L,e}|^2} \right) = \mathbf{e}_E^T \mathbf{a}_{L,e} \left(1 + \frac{p_{v_E}}{p_{x_E}} \right) , \tag{6.36}$$

$$\mathbf{e}_E^T \mathbf{a}_{R,e}^{\text{SC}} = \mathbf{e}_E^T \mathbf{a}_{R,e} \left(1 + \frac{p_{v_E}}{p_{x_R} |\mathbf{e}_E^T \mathbf{a}_{R,e}|^2} \right) = \mathbf{e}_E^T \mathbf{a}_{R,e} \left(1 + \frac{p_{v_E}}{p_{x_E}} \right) . \tag{6.37}$$

Hence, the bias only affects the amplitude but not the phase of the RTF estimate between the left and the right reference microphone and the external microphone. Note that the bias is the same for the left and the right RTF estimate and depends on the inverse SNR† in the external microphone signal. The element of the RTF vector estimate corresponding to the external microphone hence is amplified if the SNR in the external microphone signal is low.

6.2.2 *Diffuse noise field*

In this section we assume that the undesired component is a (homogeneous) diffuse noise component (e.g., a spherically isotropic noise field), i.e.,

$$\mathbf{y}_e = \mathbf{x}_e + \mathbf{n}_e \,, \tag{6.38}$$

such that the extended noisy input covariance matrix can be written as

$$\mathbf{R}_{y,e} = \mathbf{R}_{x,e} + p_n \boldsymbol{\Gamma}_e^{\text{diff}} \,, \tag{6.39}$$

where $\boldsymbol{\Gamma}_e^{\text{diff}}$ denotes the extended spatial coherence matrix (i.e., including the external microphone). $\boldsymbol{\Gamma}_e^{\text{diff}}$ can be modelled similar to $\boldsymbol{\Gamma}^{\text{diff}}$ in (3.62). Substituting $p_n \boldsymbol{\Gamma}_e^{\text{diff}}$

† It should be noted that the term SNR is used here, although a (general) undesired component is assumed in this section.

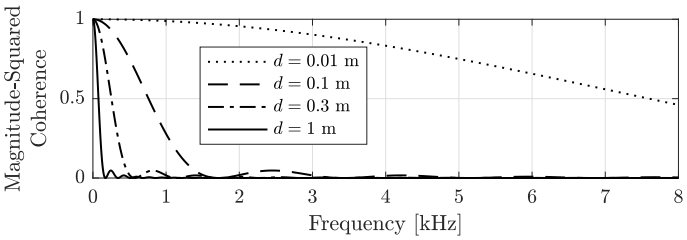

Fig. 6.1: Magnitude-squared coherence between two microphones in a spherically isotropic noise field for $d \in \{0.01,\ 0.1,\ 0.3,\ 1\}$ m and $c = 343\,\mathrm{m\,s^{-1}}$.

for $\mathbf{R}_{v,e}$ in (6.29), (6.30), (6.31) and (6.32), the left and the right multiplicative bias factor for a diffuse noise field are equal to

$$\epsilon_L^{\mathrm{mult}} = \frac{1}{1 + \frac{p_n \mathbf{e}_L^T \mathbf{\Gamma}_e^{\mathrm{diff}} \mathbf{e}_E}{\mathbf{e}_L^T \mathbf{R}_{x,e} \mathbf{e}_E}}\,, \tag{6.40}$$

$$\epsilon_R^{\mathrm{mult}} = \frac{1}{1 + \frac{p_n \mathbf{e}_R^T \mathbf{\Gamma}_e^{\mathrm{diff}} \mathbf{e}_E}{\mathbf{e}_R^T \mathbf{R}_{x,e} \mathbf{e}_E}}\,, \tag{6.41}$$

and the left and the right additive bias vector for a diffuse noise field are equal to

$$\boldsymbol{\epsilon}_L^{\mathrm{add}} = \frac{p_n \mathbf{\Gamma}_e^{\mathrm{diff}} \mathbf{e}_E}{\mathbf{e}_L^T \mathbf{R}_{x,e} \mathbf{e}_E + p_n \mathbf{e}_L^T \mathbf{\Gamma}_e^{\mathrm{diff}} \mathbf{e}_E}\,, \tag{6.42}$$

$$\boldsymbol{\epsilon}_R^{\mathrm{add}} = \frac{p_n \mathbf{\Gamma}_e^{\mathrm{diff}} \mathbf{e}_E}{\mathbf{e}_R^T \mathbf{R}_{x,e} \mathbf{e}_E + p_n \mathbf{e}_R^T \mathbf{\Gamma}_e^{\mathrm{diff}} \mathbf{e}_E}\,. \tag{6.43}$$

Assuming a spherically isotropic noise field, similarly to the spatial coherence matrix in (3.62), the (p,q)-th element of the extended spatial coherence matrix can be modelled as

$$\mathbf{\Gamma}_e^{\mathrm{diff}}(p,q) = \mathrm{sinc}\left(\frac{\omega d_{p,q}}{c}\right)\,, \tag{6.44}$$

with $d_{p,q}$ the distance between the p-th and the q-th microphone and c the speed of sound. Figure 6.1 depicts the magnitude-squared coherence $|\mathbf{\Gamma}_e^{\mathrm{diff}}(p,q)|^2$ between two microphones in a spherically isotropic noise field for $d \in \{0.01,\ 0.1,\ 0.3,\ 1\}$ m and $c = 343\,\mathrm{m\,s^{-1}}$. It can be observed that for large distances between the microphones the coherence tends to be very small and hence the assumption in (6.5) approximately holds, especially for high frequencies.

Assuming that the external microphone is sufficiently far away from the head-mounted microphones such that

$$\mathbf{\Gamma}_e^{\mathrm{diff}} \mathbf{e}_E = \mathbf{e}_E\,, \tag{6.45}$$

the multiplicative bias factors are equal to 1, i.e.,

$$\epsilon_L^{\text{mult}} = \epsilon_R^{\text{mult}} = 1\,, \tag{6.46}$$

and the additive bias vectors are equal to

$$\boldsymbol{\epsilon}_{L,n}^{\text{add}} = \frac{p_n}{\mathbf{e}_L^T \mathbf{R}_{x,e} \mathbf{e}_E} \mathbf{e}_E\,, \tag{6.47}$$

$$\boldsymbol{\epsilon}_{R,n}^{\text{add}} = \frac{p_n}{\mathbf{e}_R^T \mathbf{R}_{x,e} \mathbf{e}_E} \mathbf{e}_E\,. \tag{6.48}$$

In a (homogeneous) diffuse noise field, the fundamental assumption of the SC-based method in (6.5) can hence be satisfied when the distance between the external microphone and the head-mounted microphones is large enough (cf. simulations in Section 6.4).

6.2.3 *Interfering source*

In this section we assume that the extended noisy input vector consists of a desired source component and an interfering source component, i.e.,

$$\mathbf{y}_e = \mathbf{x}_e + \mathbf{u}_e\,. \tag{6.49}$$

Using (2.46) and (2.47), the extended noisy input covariance matrix is then equal to

$$\mathbf{R}_{y,e} = \mathbf{R}_{x,e} + \mathbf{R}_{u,e} = p_{x_L} \mathbf{a}_{L,e} \mathbf{a}_{L,e}^H + p_{u_L} \mathbf{b}_{L,e} \mathbf{b}_{L,e}^H\,. \tag{6.50}$$

Substituting (6.50) in (6.15) and (6.16), the multiplicative bias factors for an interfering source are equal to

$$\epsilon_L^{\text{mult}} = \frac{1}{1 + \frac{p_{u_L}}{p_{x_L}} \frac{\mathbf{b}_{L,e}^H \mathbf{e}_E}{\mathbf{a}_{L,e}^H \mathbf{e}_E}}\,, \tag{6.51}$$

$$\epsilon_R^{\text{mult}} = \frac{1}{1 + \frac{p_{u_R}}{p_{x_R}} \frac{\mathbf{b}_{R,e}^H \mathbf{e}_E}{\mathbf{a}_{R,e}^H \mathbf{e}_E}}\,, \tag{6.52}$$

and the additive bias vectors for an interfering source are equal to

$$\boldsymbol{\epsilon}_L^{\text{add}} = \mathbf{b}_{L,e} \frac{1}{1 + \frac{p_{x_L}}{p_{u_L}} \frac{\mathbf{a}_{L,e}^H \mathbf{e}_E}{\mathbf{b}_{L,e}^H \mathbf{e}_E}}\,, \tag{6.53}$$

$$\boldsymbol{\epsilon}_R^{\text{add}} = \mathbf{b}_{R,e} \frac{1}{1 + \frac{p_{x_R}}{p_{u_R}} \frac{\mathbf{a}_{R,e}^H \mathbf{e}_E}{\mathbf{b}_{R,e}^H \mathbf{e}_E}}\,. \tag{6.54}$$

Hence, the SC-based estimate of the left and the right extended RTF vector of the desired source is equal to a mixture between the left and the right extended RTF vector of the desired source $\mathbf{a}_{L,e}$ and $\mathbf{a}_{R,e}$ and the left and the right extended RTF vector of the interfering source $\mathbf{b}_{L,e}$ and $\mathbf{b}_{R,e}$, depending on the CPSD ratios

$$\frac{\mathbf{e}_L^T\mathbf{R}_{u,e}\mathbf{e}_E}{\mathbf{e}_L^T\mathbf{R}_{x,e}\mathbf{e}_E} = \frac{p_{u_L}\mathbf{b}_{L,e}^H\mathbf{e}_E}{p_{x_L}\mathbf{a}_{L,e}^H\mathbf{e}_E}\,, \tag{6.55}$$

$$\frac{\mathbf{e}_R^T\mathbf{R}_{u,e}\mathbf{e}_E}{\mathbf{e}_R^T\mathbf{R}_{x,e}\mathbf{e}_E} = \frac{p_{u_R}\mathbf{b}_{R,e}^H\mathbf{e}_E}{p_{x_R}\mathbf{a}_{R,e}^H\mathbf{e}_E}\,. \tag{6.56}$$

Hence, in an acoustic scenario with multiple speakers and assuming sparsity in the STFT domain, i.e., one speaker is assumed to be dominant in each time-frequency bin, the SC-based RTF vector estimation method will estimate the RTF vectors corresponding to the dominant speaker in each time-frequency bin. When these RTF vector estimates are used in, e.g., an eBMVDR beamformer (see Chapter 5), this means that all speakers are considered as desired sources and will be enhanced.

6.3 RTF vector estimation exploiting multiple external microphones

In this section we consider the multi-extended binaural hearing device configuration (cf. Figure 2.3) with M_H head-mounted microphones and M_E external microphones, i.e., $M = M_H + M_E$ microphones in total. In Section 6.3.1 we show that M_E different SC-based RTF vector estimates can be obtained, i.e., one for each external microphone. In Section 6.3.2 we propose three procedures to linearly combine the different RTF vector estimates. In the first procedure we select the RTF vector estimate corresponding to the external microphone with the highest narrowband input SNR. In the second procedure we simply average the different RTF vector estimates. In the third procedure we linearly combine the different RTF vector estimates such that the narrowband output SNR of the eBMVDR beamformer is maximized.

6.3.1 *SC-based RTF vector estimation per external microphone*

Assuming that the noise component in each external microphone signal is uncorrelated with the noise component in all other (head-mounted and external) microphone signals, i.e.,

$$\mathcal{E}\{\mathbf{n}_e n_{E,i}^*\} = p_{n_{E,i}}\,\mathbf{e}_{E,i}\,, \quad i \in \{1, \ldots, M_E\}\,, \tag{6.57}$$

the SC-based RTF vector estimation method proposed in Section 6.1 can be used with each external microphone to estimate the M-dimensional RTF vectors of the desired source, similarly as in (6.19) and (6.20), i.e.,

$$\hat{\mathbf{a}}_{L,e}^{\mathrm{SC}-i} = \frac{\hat{\mathbf{R}}_{y,e}\mathbf{e}_{E,i}}{\mathbf{e}_L^T\hat{\mathbf{R}}_{y,e}\mathbf{e}_{E,i}}, \quad i \in \{1, \ldots, M_E\}, \tag{6.58}$$

$$\hat{\mathbf{a}}_{R,e}^{\mathrm{SC}-i} = \frac{\hat{\mathbf{R}}_{y,e}\mathbf{e}_{E,i}}{\mathbf{e}_R^T\hat{\mathbf{R}}_{y,e}\mathbf{e}_{E,i}}, \quad i \in \{1, \ldots, M_E\}. \tag{6.59}$$

In practice M_E different RTF vector estimates are obtained, since 1) the assumption in (6.57) is not perfectly satisfied, 2) a different element (corresponding to the i-th external microphone) is biased, and 3) $\hat{\mathbf{R}}_{y,e} \neq \hat{\mathbf{R}}_{x,e} + \hat{\mathbf{R}}_{n,e}$. Aiming at obtaining one RTF vector estimate from the M_E different RTF vector estimates, in the next section we propose several procedures to linearly combine (or select) these estimates. The resulting left and right RTF vector estimates could then, e.g., be used in the BMVDR beamformer (cf. Section 3.1) or eBMVDR beamformer (cf. Section 5.2).

6.3.2 *Combination of SC-based RTF vector estimates*

By linearly combining the different RTF vector estimates (per frequency), the (normalized) combined left and right RTF vector estimate is given by

$$\boxed{\hat{\mathbf{a}}_{L,e}^{\mathrm{SC-C}} = \frac{\mathbf{A}_L^{\mathrm{SC}}\mathbf{c}}{\mathbf{e}_{L,e}^T\mathbf{A}_L^{\mathrm{SC}}\mathbf{c}}} \tag{6.60}$$

$$\boxed{\hat{\mathbf{a}}_{R,e}^{\mathrm{SC-C}} = \frac{\mathbf{A}_R^{\mathrm{SC}}\mathbf{c}}{\mathbf{e}_{R,e}^T\mathbf{A}_R^{\mathrm{SC}}\mathbf{c}}} \tag{6.61}$$

where $\mathbf{A}_L^{\mathrm{SC}}$ and $\mathbf{A}_R^{\mathrm{SC}}$ denote $M \times M_E$-dimensional matrices, containing the M_E SC-based left and right RTF vector estimates in (6.58) and (6.59), i.e.,

$$\mathbf{A}_L^{\mathrm{SC}} = \left[\hat{\mathbf{a}}_{L,e}^{\mathrm{SC}-1}, \ldots, \hat{\mathbf{a}}_{L,e}^{\mathrm{SC}-M_E}\right], \tag{6.62}$$

$$\mathbf{A}_R^{\mathrm{SC}} = \left[\hat{\mathbf{a}}_{R,e}^{\mathrm{SC}-1}, \ldots, \hat{\mathbf{a}}_{R,e}^{\mathrm{SC}-M_E}\right], \tag{6.63}$$

and $\mathbf{c}$ denotes the M_E-dimensional (complex-valued) combination vector. In the following we propose three different procedures to determine the combination vector $\mathbf{c}$ in practice.

The first procedure, denoted as **inSNR**, is to select the left and the right RTF vector estimate (per frequency) corresponding to the external microphone with the

highest narrowband input SNR, similarly to [190]. Since the input SNR in the i-th external microphone signal can be written as

$$\mathrm{SNR}_{E,i}^{\mathrm{in}} = \frac{\mathbf{e}_{E,i}^T \mathbf{R}_{x,e} \mathbf{e}_{E,i}}{\mathbf{e}_{E,i}^T \mathbf{R}_{n,e} \mathbf{e}_{E,i}} = \frac{\mathbf{e}_{E,i}^T \mathbf{R}_{y,e} \mathbf{e}_{E,i}}{\mathbf{e}_{E,i}^T \mathbf{R}_{n,e} \mathbf{e}_{E,i}} - 1\,, \tag{6.64}$$

this procedure only requires an estimate of the extended noisy input covariance matrix $\mathbf{R}_{y,e}$ (and not $\mathbf{R}_{x,e}$), i.e.,

$$\boxed{\mathbf{c}^{\mathrm{inSNR}} = \mathbf{e}_{E,\hat{i}}\,, \quad \hat{i} = \arg\max_i \frac{\mathbf{e}_{E,i}^T \hat{\mathbf{R}}_{y,e} \mathbf{e}_{E,i}}{\mathbf{e}_{E,i}^T \hat{\mathbf{R}}_{n,e} \mathbf{e}_{E,i}}} \tag{6.65}$$

Especially for a complex acoustic scenario with a moving desired source, where the distance of the source to each external microphone and hence the input SNR in each external microphone is time-varying, the inSNR-based selection procedure is expected to outperform the SC-based method only using one external microphone.

Assuming a uniform distribution of the estimation errors for the M_E SC-based RTF vector estimates, in the second procedure, denoted as **AV**, we propose to simply average the estimates, i.e.,

$$\boxed{\mathbf{c}^{\mathrm{AV}} = \left[\frac{1}{M_E}, \ldots, \frac{1}{M_E}\right]^T} \tag{6.66}$$

Intuitively, this procedure is sub-optimal, especially when the estimation errors are very different.

In the third procedure, denoted as **maxSNR**, we propose to combine the SC-based left and right RTF vector estimates (per frequency) such that the narrowband output SNR of the eBMVDR beamformer is maximized. Please note again that the left and the right output SNRs of the eBMVDR beamformer are equal (cf. (5.8)). Using (6.60) and (6.61) in (5.4) and (5.5), the left and the right output SNR in (2.68) and (2.69) can be written as the generalized Rayleigh quotient

$$\mathrm{SNR}_{\mathrm{eBMVDR},L}^{\mathrm{out}} = \mathrm{SNR}_{\mathrm{eBMVDR},R}^{\mathrm{out}} = \frac{\mathbf{c}^H \mathbf{\Lambda}_1 \mathbf{c}}{\mathbf{c}^H \mathbf{\Lambda}_2 \mathbf{c}} - 1\,, \tag{6.67}$$

with

$$\mathbf{\Lambda}_1 = (\mathbf{A}_L^{\mathrm{SC}})^H \mathbf{R}_{n,e}^{-1} \mathbf{R}_{y,e} \mathbf{R}_{v,e}^{-1} \mathbf{A}_L^{\mathrm{SC}}\,, \tag{6.68}$$

$$\mathbf{\Lambda}_2 = (\mathbf{A}_L^{\mathrm{SC}})^H \mathbf{R}_{n,e}^{-1} \mathbf{A}_L^{\mathrm{SC}}\,. \tag{6.69}$$

Aiming at maximizing the output SNR of the eBMVDR beamformer in (6.67), the SNR-maximizing combination vector $\mathbf{c}^{\text{maxSNR}}$ is equal to the principal eigenvector $\mathbf{p}\{\mathbf{\Lambda}_2^{-1}\mathbf{\Lambda}_1\}$ of the $(M_E \times M_E)$-dimensional matrix $\mathbf{\Lambda}_2^{-1}\mathbf{\Lambda}_1$, i.e.,

$$\boxed{\mathbf{c}^{\text{maxSNR}} = \arg\max_{\mathbf{c}} \ \text{SNR}^{\text{out}}_{\text{eBMVDR},\{L,R\}} = \mathbf{p}\{\mathbf{\Lambda}_2^{-1}\mathbf{\Lambda}_1\}} \tag{6.70}$$

which hence also only requires an estimate of the extended noisy input covariance matrix $\mathbf{R}_{y,e}$ (and not $\mathbf{R}_{x,e}$). Although constructing the matrices $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ comes with some computational complexity, the computational complexity of the $(M_E \times M_E)$-dimensional eigenvalue decomposition (EVD) is always smaller than the computational complexity of the $(M \times M)$-dimensional EVD required for the CW method (cf. Section 3.4.2).

6.4 Simulations

In this section we present two experiments, each of which deals with different aspects of the performance of the proposed SC-based method, more in particular when using the RTF vector estimates in the BMVDR beamformer (only filtering the head-mounted microphone signals) or the eBMVDR beamformer (filtering the head-mounted and the external microphone signals). In Section 6.4.1 we present experiment 1 (published in [161]), where we consider using *one* external microphone with the proposed RTF estimation methods in Section 6.1, and investigate the noise reduction performance and binaural cue preservation of the eBMVDR beamformer in comparison to the BMVDR beamformer for a moving desired source. In Section 6.4.2 we present experiment 2 (published in [162]), where we consider using *multiple* external microphones with the proposed RTF estimation methods and combination procedures in Section 6.3, and investigate the noise reduction performance of the eBMVDR beamformer for a moving desired source. In Appendix B we present two additional experiments (published in [159] and [160]), where we consider a static desired source and one external microphone and investigate the influence of the input SNR, reverberation time and the time constants used for covariance matrix estimation on the RTF vector estimation accuracy, and the noise reduction performance and binaural cue preservation when using the RTF vector estimates in a BMVDR beamformer.

All signals were recorded in a variable acoustics laboratory located at the University of Oldenburg, for which the room dimensions are about $7 \times 6 \times 2.7$ m^3, and where the reverberation time T_{60} can be easily changed by closing and opening absorber panels mounted to the walls and the ceiling. All reverberation times were measured using the broadband energy decay curve of the measured impulse responses. At approximately the center of the room a KEMAR head-and-torso simulator (HATS) was placed. Two behind-the-ear hearing aid dummies with two microphones each, i.e., $M_H = 4$, were placed on the ears of the HATS, with an inter-microphone distance of about 14 mm. We chose the frontal microphone on each hearing aid as the reference microphone. No a-priori information about the position of any external

microphone were used in the experiments, i.e., the external microphones were placed at unknown, arbitrary positions.

To generate diffuse-like background noise, in both experiments we placed four loudspeakers facing the corners of the laboratory, playing back different multi-talker recordings. Due to small number of loudspeakers and temporally dominant interfering speakers in the multi-talker recordings, the resulting noise field was neither coherent nor perfectly diffuse. All microphone signals, i.e., the head-mounted microphone signals and the external microphone signal(s), were recorded synchronously, thereby neglecting synchronization and latency aspects.

All microphone signals were processed in an STFT framework (cf. (2.1)) with a frame size of $T_d = 512$ samples, corresponding to 32 ms, and a frame shift of $T_s = 256$ samples, corresponding to 50 % overlap, and a square-root Hann window. Similarly to (3.68) and (3.69), the extended noisy input covariance matrix $\hat{\mathbf{R}}_{y,e}^{\text{onl}}$ and the extended noise covariance matrix $\hat{\mathbf{R}}_{n,e}^{\text{onl}}$ were recursively estimated during detected speech-plus-noise and noise-only bins, respectively. If required, the covariance matrices $\hat{\mathbf{R}}_{y}^{\text{onl}}$ and $\hat{\mathbf{R}}_{n}^{\text{onl}}$ (only including the head-mounted microphones) were estimated using a subset of the estimates $\hat{\mathbf{R}}_{y,e}^{\text{onl}}$ and $\hat{\mathbf{R}}_{n,e}^{\text{onl}}$, respectively. The (time-varying) estimates of the covariance matrices were then used in the different RTF vector estimation methods and in the calculation of the (time-varying) (e)BMVDR beamformers. All performance measures were computed in the time-domain using the shadow filter approach, i.e., using the individual signal components in the reference microphone signals and the output signals.

6.4.1 *Experiment 1 – One external microphone*

In the first experiment we consider one external microphone, i.e., $M_E = 1$, and compare the performance of the eBMVDR beamformer (filtering the head-mounted and the external microphone signals) and the BMVDR beamformer (only filtering the head-mounted microphone signals) for a moving desired source, either using the CW method or the proposed SC-based method.

Using the absorber panels that are mounted to the walls and the ceiling of the laboratory, the reverberation time was set to approximately 350 ms. The experimental setup is depicted in Figure 6.2. The external microphone was placed at about 1.5 m in front of the HATS. The desired source was a male German speaker played back by a loudspeaker which was placed at about 2 m from the HATS at same height. Initially, the loudspeaker was placed at an angle of 0°, i.e., in front of the HATS (at a distance of about 0.5 m to the external microphone). During the first 5 s, the loudspeaker remained static in its initial position. During the following 5 s, the loudspeaker was moved (by hand) to an angle of about 75° to the right side of the HATS (at a distance of about 1.5 m to the external microphone), where it remained for another 5 s. The desired source and the diffuse-like background noise were recorded separately and mixed afterwards to an average intelligibility-weighted SNR (iSNR) [189] of 0 dB in the right reference microphone signal. The iSNR is defined as the sum of the SNRs in all frequency bins weighted with a frequency-dependent band

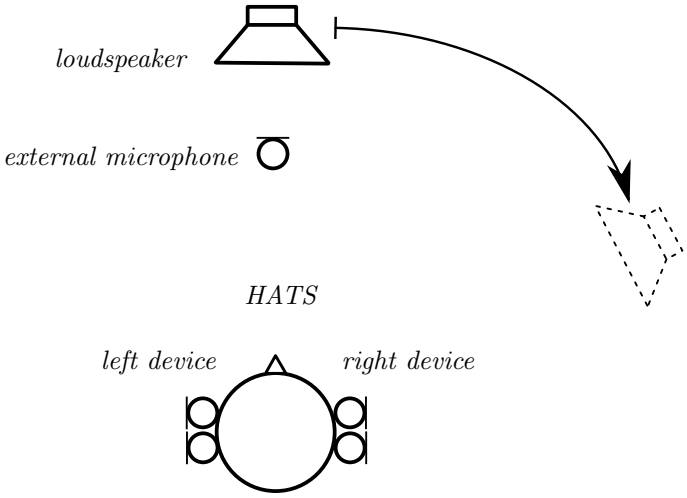

Fig. 6.2: Experimental setup for experiment 1 with $M_H = 4$ head-mounted microphones and $M_E = 1$ external microphone. The loudspeaker (as desired source) was moved from its initial position in front of the listener to the right side.

importance function, for which we used the same weights as for the speech intelligibility index in [30] (based on one-third octaves). The average iSNR in the external microphone was equal to about 14 dB. The complete signal had a length of 15 s with 0.5 s of noise-only at the beginning.

To distinguish between speech-plus-noise and noise-only bins, we estimated a high-resolution VAD from an SPP estimate in every time-frequency bin [27] (cf. (3.72), with $\mathrm{SPP}_{\mathrm{upper}} = 0.6$ and $\mathrm{SPP}_{\mathrm{lower}} = 0.4$) using the external microphone signal. The smoothing factors for the online estimation of the covariance matrices were chosen as $\alpha_y = 0.852$ and $\alpha_n = 0.984$, corresponding to time constants of 100 ms for speech-plus-noise and 1 s for noise-only, respectively.

In this experiment we considered five different versions of the (e)BMVDR beamformer, either using the RTF vectors $\hat{\mathbf{a}}_L$ and $\hat{\mathbf{a}}_R$ (i.e., filtering only the head-mounted microphone signals but not the external microphone signal) or the extended RTF vectors $\hat{\mathbf{a}}_{L,e}$ and $\hat{\mathbf{a}}_{R,e}$ (i.e., filtering all available microphone signals), i.e.,

- **FIX**: Fixed BMVDR beamformer using *anechoic* RTF vectors calculated from measured impulse responses [35] corresponding to a position in front of the listener (cf. Section 3.4.2).
- **CW** and **CWE**: RTF-steered (e)BMVDR beamformer using the CW RTF vector estimation method in (3.84) and (3.85), without and with incorporating the external microphone.
- **SC** and **SCE**: RTF-steered (e)BMVDR beamformer using the proposed SC-based RTF vector estimation method in (6.21) and (6.22), and (6.19) and (6.20), respectively.

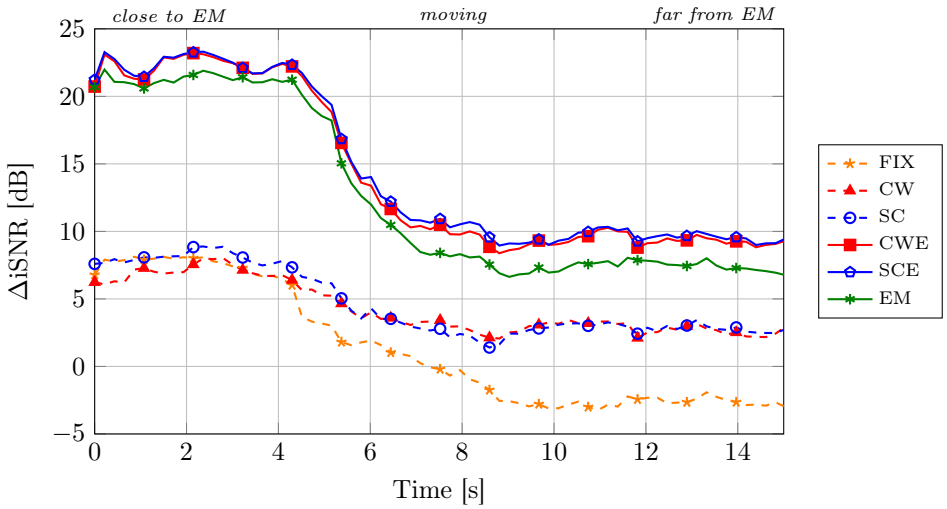

Fig. 6.3: Intelligibility-weighted SNR improvement (plotted over time) for all considered (e)BMVDR beamformers and the external microphone.

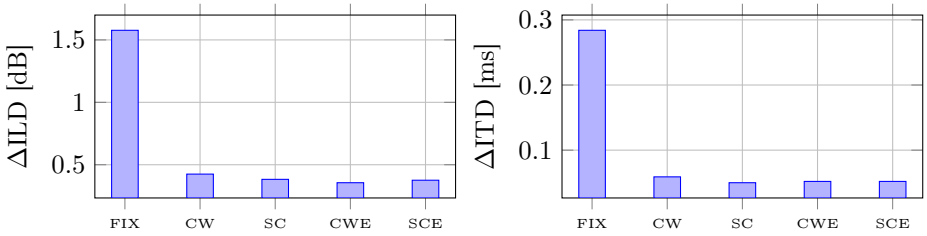

Fig. 6.4: Reliable binaural cue errors (averaged over time) for all considered (e)BMVDR beamformers.

In addition, we considered the external microphone (EM) signal without applying any noise reduction.

As performance measure for noise reduction, in this experiment we used the iSNR improvement (ΔiSNR) in blocks of 1 s between the right reference microphone signal and the right output signal or the external microphone signal. As performance measure for binaural cue preservation (cf. Section 2.2.2) we used the reliable binaural cue errors of the desired source component, i.e., ΔILD and ΔITD, based on an auditory model [55] and averaged over time and all frequencies. This auditory model mimics physiological aspects of the human auditory system and hence enables to only consider binaural cues errors that are perceivable by human listeners.

Figure 6.3 depicts the iSNR improvement for all considered (e)BMVDR beamformers and the external microphone. The plot is divided into three phases. First, where the desired source remains statically in the initial position close to the external microphone (*close to EM*). Second, where the desired source moves to the right side of

the HATS (*moving*). Third, where the desired source remains statically in its final position (*far from EM*). As expected, using anechoic RTF vectors corresponding to a position in front of the listener (FIX) leads to the worst performance (even negative ΔiSNR) of all considered (e)BMVDR beamformers, since it does not track the movement of the desired source. The RTF-steered CW and SC beamformers (both not filtering the external microphone signal) show a similar iSNR improvement between 2 and 9 dB. Averaging the iSNR improvements over time shows that SC outperforms CW by about 0.31 dB. The RTF-steered CWE and SCE beamformers (both filtering all available microphone signals) outperform all other considered beamformers and the EM. At the initial position of the desired source (0 to 5 s, *close to EM*), the CWE and SCE beamformers outperform the CW and SC beamformers by about 14 dB and the EM by about 2 dB. At the final position of the desired source (10 to 15 s, *far from EM*), the CWE and SCE beamformers outperform the CW and SC beamformers by about 6 dB and the EM by about 3 dB. Averaging the iSNR improvement over time shows that SCE outperforms CWE by about 0.3 dB.

Figure 6.4 depicts the (reliable) ILD and ITD errors for all considered (e)BMVDR beamformers. It should be stressed that directly using the EM signal does not provide any binaural cues to the user, hence leading to in-head localization. As expected, FIX shows the worst performance, i.e., the largest binaural cue errors. All RTF-steered (e)BMVDR beamformers show similar, small binaural cue errors, indicating that the desired source is perceived as coming from the correct direction.

In conclusion, these results show that the SC-based RTF vector estimation method (cf. Section 6.1) can be used to steer an (e)BMVDR beamformer in an acoustic scenario with a moving desired source and yields a similar (even slightly better) iSNR improvement and similar binaural cues as the state-of-the-art CW RTF vector estimation method at much lower computational complexity. The results further indicate that the considered distance between the external microphone and the head-mounted microphones is large enough, such that the bias of the SC-based method (cf. Section 6.2) is not affecting the noise reduction performance and the binaural cue preservation in practice.

6.4.2 *Experiment 2 – Multiple external microphones*

Contrary to the previous experiment, in this experiment we consider multiple external microphones. For a moving desired source we compare the performance of the eBMVDR beamformer (filtering the head-mounted and the external microphone signals) using the CW RTF vector estimation method in (3.84) and (3.85) (using the head-mounted microphones and the external microphones), the proposed SC-based RTF vector estimation method in (6.15) and (6.16) per external microphone, and the three proposed procedures to combine SC-based RTF vector estimates (cf. Section 6.3.2).

Using the absorber panels that are mounted to the walls and the ceiling of the laboratory, the reverberation time was set to approximately 400 ms. In addition to the $M_H = 4$ head-mounted microphones, $M_E = 3$ external microphones were

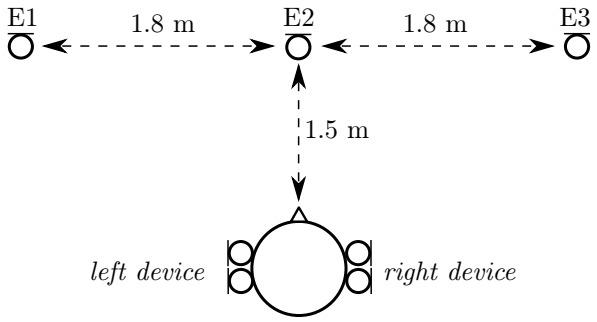

Fig. 6.5: Experimental setup for experiment 2 with $M_H = 4$ head-mounted microphones and $M_E = 3$ external microphones. The desired source moved from E1 to E3.

placed in front of the HATS as depicted in Figure 6.5. Hence, in total $M = 7$ microphones were used in the eBMVDR beamformer. The desired source was a male speaker, walking from the first external microphone (E1) to the third external microphone (E3) while speaking ten German sentences with pauses of about half a second between the sentences. The desired source and the diffuse-like background noise were recorded separately and mixed afterwards. Due to the moving desired source, the input SNR in the head-mounted reference microphone signals varied between approximately 0 and 6 dB, while the input SNR in the external microphone signals varied approximately between 0 and 11 dB.

To distinguish between speech-plus-noise and noise-only time-frequency bins, we estimated the SPP [27] in the three (noisy) external microphone signals, and averaged and thresholded them afterwards (cf. (3.72), with $\mathrm{SPP}_{\mathrm{upper}} = 0.4$ and $\mathrm{SPP}_{\mathrm{lower}} = 0.4$). The time constants for the online estimation of the covariance matrices were chosen as 250 ms for speech-plus-noise and 1 s for noise-only.

In this experiment we considered seven different versions of the eBMVDR beamformer using the multi-extended RTF vectors $\hat{\mathbf{a}}_{L,e}$ and $\hat{\mathbf{a}}_{R,e}$ (i.e., filtering all available microphone signals), i.e.,

- **CW**: The state-of-the-art CW method in (3.84) and (3.85), incorporating the three external microphones, i.e., using all microphones.
- **SC-1**, **SC-2** and **SC-3**: The proposed SC method in (6.58) and (6.59) using each external microphone separately, i.e., either using E1, E2 or E3.
- **inSNR**, **AV** and **maxSNR**: The proposed procedure to linearly combine SC-based RTF vector estimates in (6.60) and (6.61) using the three proposed combination vectors in (6.65), (6.66) and (6.70), respectively.

As performance measure, in this experiment we used the binaural SNR improvement (ΔBSNR), which is defined as [11]

$$\Delta\mathrm{BSNR} = 10\log\left(\frac{p_{x_L}^{\mathrm{out}} + p_{x_R}^{\mathrm{out}}}{p_{n_L}^{\mathrm{out}} + p_{n_R}^{\mathrm{out}}}\right) - 10\log\left(\frac{p_{x_L} + p_{x_R}}{p_{n_L} + p_{n_R}}\right). \tag{6.71}$$

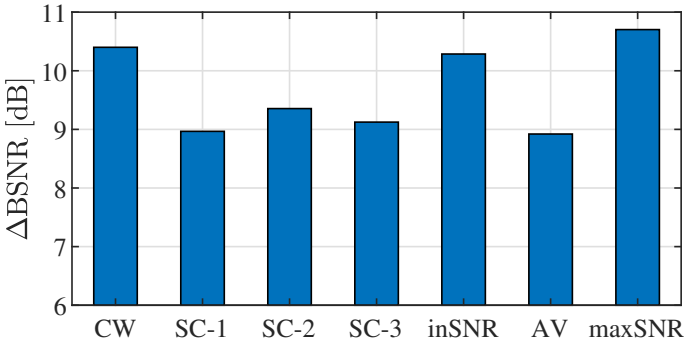

Fig. 6.6: Binaural SNR improvement for all considered RTF vector estimation methods, averaged over time and frequency.

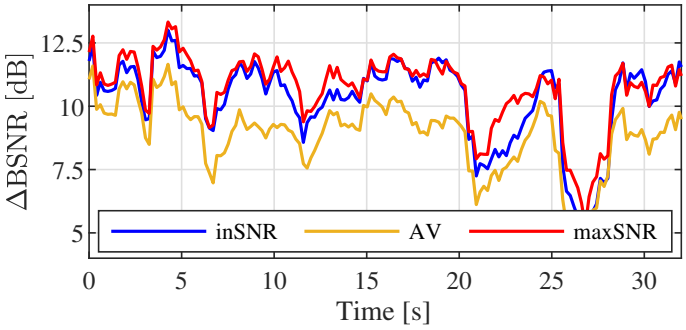

Fig. 6.7: Binaural SNR improvement over time for the inSNR, AV and maxSNR combination procedures, averaged over all frequencies.

The binaural SNR improvement is especially useful when the better ear changes, e.g., when the desired source is not always on the same side of the HATS.

Figure 6.6 depicts the ΔBSNR (averaged over time and all frequencies) for all considered RTF vector estimation methods. The CW method as a state-of-the-art benchmark yields an average ΔBSNR of 10.4 dB. The SC-based method using one external microphone, i.e., SC-1, SC-2 and SC-3, yields an average ΔBSNR of about 9 dB and hence could not reach the performance of the CW method. The input SNR-based combination (inSNR) yields an average ΔBSNR of 10.3 dB, which is similar to the CW method. The averaging combination (AV) yields an average ΔBSNR of only 8.9 dB, which is even worse than the SC-based method per external microphone. This can probably be explained by the rather different RTF vector estimation errors for the three external microphones. The SNR-maximizing combination (maxSNR) yields an average ΔBSNR of 10.7 dB, hence outperforming all other combination procedures and RTF vector estimation methods. Comparing the computational complexity of the best three methods, the CW method has the largest complexity due to the 7-dimensional EVD, whereas the maxSNR combination only requires a 3-dimensional EVD and the inSNR combination does not even

require an EVD. Nevertheless, the maxSNR combination enables to improve the ΔBSNR by about 0.5 dB compared to the inSNR combination. Figure 6.7 depicts the ΔBSNR over time (averaged over all frequencies) for the three RTF vector estimate combination procedures using the proposed combination vectors in more detail. It can be observed that the maxSNR combination outperforms the inSNR and AV combination for almost all time instances. The sound files of the input and output signals are available online[†].

In conclusion, the results indicate that in a scenario with multiple external microphones at unknown positions, the SC-based RTF vector estimation method should not be used with just any external microphone, but with all external microphones such that the different SC-based RTF vector estimates can be linearly combined afterwards. The input SNR-based combination yields a similar noise reduction performance as the state-of-the-art CW method at much lower computational complexity, while the output SNR-maximizing combination outperformed all other combination procedures and RTF vector estimation methods.

6.5 Summary

In this chapter we proposed computationally efficient SC-based methods to estimate the RTF vectors of the desired source, by exploiting one or more external microphones that are spatially separated from the head-mounted microphones. Assuming that the SC between the noise component in the head-mounted microphone signals and the noise component in the external microphone signal is zero, we showed that the elements of the RTF vectors corresponding to the head-mounted microphones are unbiased, while the element corresponding to the external microphone is corrupted by a bias term. We provided a bias analysis of the SC-based method for an arbitrary undesired component, a diffuse noise field and an interfering source. The analysis showed that for an arbitrary undesired component the bias depends on a CPSD ratio and the overall correlation between the undesired component in the head-mounted microphone signals and the undesired component in the external microphone signal. If the undesired component in the head-mounted microphone signals is uncorrelated with the undesired component in the external microphone signal, only the last element of the RTF vector (corresponding to the external microphone) is affected by real-valued bias that is inversely proportional to the SNR in the external microphone signal. We further showed that in a diffuse noise field, the assumption that the noise component in the head-mounted microphone signals is uncorrelated with noise component in the external microphone signal is approximately satisfied when the distance between the external microphone and the head-mounted microphones is large enough. For an interfering source, the bias analysis finally showed that the resulting RTF vector estimate is equal to a mixture between the RTF vectors of the desired source and the interfering source, depending on a

† https://uol.de/en/sigproc/research/audio-demos/binaural-noise-reduction/multiple-external-microphones

CPSD ratio. When multiple external microphones are available, we showed that different RTF vector estimates can be obtained by using the SC-based method for each external microphone. We proposed several procedures to combine these RTF vector estimates, either by selecting the estimate corresponding to the highest input SNR, by averaging the estimates or by combining the estimates in order to maximize the output SNR of the eBMVDR beamformer. We conducted two experiments with a moving desired source and diffuse-like background noise in a reverberant environment. The results of the first experiment showed that the SC-method can be used to steer an (e)BMVDR beamformer in an acoustic scenario with a moving desired source and yields a similar (even slightly better) noise reduction performance and binaural cue preservation as the state-of-the-art CW RTF vector estimation method at much lower computational complexity. Further, the bias of the SC-based method does not seem to affect the noise reduction performance and binaural cue preservation in practice, when the external microphone is spatially separated from the head-mounted microphones. The results of the second experiment showed that in a scenario with multiple external microphones, the SC-based RTF vector estimation method should not be used with just any external microphone, but with all external microphones such that the different SC-based RTF vector estimates can be linearly combined afterwards. The input SNR-based combination yields a similar noise reduction performance as the state-of-the-art CW method at much lower computational complexity, while the output SNR-maximizing combination outperformed all other combination procedures and RTF vector estimation methods.

7

CONCLUSIONS AND FURTHER RESEARCH

In this chapter we summarize the main results of the thesis and indicate possible directions for further research.

7.1 Conclusions

Beamforming algorithms for head-mounted assistive hearing devices are crucial to improve speech intelligibility and speech quality in complex acoustic scenarios, where speech communication is affected by undesired sources such as an interfering source (e.g., a competing speaker) and background noise (e.g., diffuse babble noise). Besides reducing the interfering source and the background noise, another important objective of a binaural beamforming algorithm is the preservation of the listener's spatial impression of the acoustic scene, which can be achieved by preserving the binaural cues of all sound sources. Although, most state-of-the-art binaural beamforming algorithms preserve the binaural cues of the desired source, they distort the binaural cues of either the interfering source, the background noise or both.

To improve the performance of hearing devices, one or more external microphones (e.g., lying on a table, attached to a person) can be used in conjunction with the head-mounted microphones, enabling to not only locally sample the sound field at the listener's head but to increase the spatial diversity by spatially distributing the external microphones. Besides technical challenges (e.g., synchronization, bandwidth limitations, transmission loss), one of the main algorithmic challenges posed by incorporating external microphones is the fact that the relative position of the external microphones to the head-mounted microphones and the sound sources is unknown and may be time-varying.

The main objective of this thesis was to develop and evaluate advanced binaural beamforming algorithms and to incorporate one or more external microphones in a binaural hearing device configuration. The first focus was to improve state-of-the-art binaural beamforming algorithms, more in particular to develop a binaural beamforming algorithm that jointly preserves the binaural cues of the desired source, the interfering source and the background noise. The second focus was the incorporation of one or more external microphones to improve the noise reduction performance and the binaural cue preservation of binaural beamforming algorithms, without assuming any a-priori knowledge about the position of the external microphones.

After presenting the signal models for the binaural hearing device configurations with and without external microphones, in **Chapter 2**, we briefly reviewed three state-of-the-art binaural beamforming algorithms using only the head-mounted microphone signals in **Chapter 3**. First, we reviewed the frequently-used BMVDR beamformer, which preserves the binaural cues of the desired source but distorts the binaural cues of the undesired sources (i.e., interfering source and background noise). The BMVDR beamformer provides the best noise reduction performance among all considered distortionless binaural beamforming algorithms, but the interference reduction depends on the relative position of the interfering source to the desired source and is not controllable. Second, we reviewed the BLCMV beamformer, which preserves the binaural cues of both the desired source and the interfering source but distorts the binaural cues of the background noise, depending on the relative position of the interfering source to the desired source. Since less degrees of freedom are available for noise reduction, the BLCMV beamformer typically yields a lower noise reduction performance than the BMVDR beamformer, but enables to directly control the amount of interference reduction by means of an interference scaling parameter. Third, we reviewed the BMVDR-N beamformer, which allows to trade off between noise reduction performance and binaural cue preservation of the noise component by mixing the noisy reference microphone signals with the output signals of the BMVDR beamformer using a mixing parameter. While the BMVDR-N beamformer hence enables to control the background noise component in the output signals, the interference reduction and the binaural cue preservation of the interfering source depend on the relative position of the interfering source to the desired source and are not straightforward to control using the mixing parameter. We further discussed several methods to estimate the parameters that are required to calculate the filter vectors of the considered binaural beamforming algorithms in practice, more in particular covariance matrices and RTF vectors. Directly estimating the RTF vectors, the relative positions of the microphones do not have to be known or fixed and can even be time-varying. Estimating the RTF vectors hence enables to incorporate one or more external microphones at unknown positions, compared to using modelled/measured RTFs. Most state-of-the-art RTF vector estimation methods require estimates of both the noisy input covariance matrix and the noise covariance matrix.

To address the first focus of the thesis, in **Chapter 4** we proposed a novel binaural beamforming algorithm, which merges the advantages of the BLCMV beamformer and the BMVDR-N beamformer, i.e., it allows to preserve the binaural cues of the interfering source and control the trade-off between noise reduction performance and binaural cue preservation of the background noise. To address the second focus of the thesis, we first investigated the incorporation of an external microphone in the BMVDR-N beamformer for an arbitrary noise field in **Chapter 5**. We then proposed computationally efficient methods to estimate the RTF vectors of the desired source in **Chapter 6**, which exploit one or more external microphones that are spatially separated from the head-mounted microphones, and only require an estimate of the noisy input covariance matrix.

In **Chapter 4** we proposed the BLCMV-N beamformer, a novel binaural beamforming algorithm that merges the advantages of the BLCMV beamformer and the BMVDR-N beamformer, i.e., preserving the binaural cues of the interfering source and controlling the reduction of the interfering source as well as the binaural cues of the background noise. Compared to the BMVDR beamformer, the BLCMV-N beamformer uses an additional constraint to preserve a scaled version of the interfering source component in the reference microphone signals (like the BLCMV beamformer) and aims at preserving a scaled version of the noise component in the reference microphone signals. First, we showed that the output signals of the BLCMV-N beamformer can be interpreted as a mixture between the noisy reference microphone signals and the output signals of a BLCMV beamformer using an adjusted interference scaling parameter. We then derived two decompositions for the BLCMV-N beamformer which revealed differences and similarities between the BLCMV-N beamformer and the BLCMV beamformer. Furthermore, we provided a theoretical comparison between the BMVDR beamformer, the BLCMV beamformer, the BMVDR-N beamformer and the proposed BLCMV-N beamformer in terms of noise and interference reduction performance and binaural cue preservation. We showed that the output SNR of the BLCMV-N beamformer is smaller than or equal to the output SNR of the BLCMV beamformer and derived the optimal interference scaling parameter maximizing the output SNR of the BLCMV-N beamformer. The obtained analytical expressions were first validated using measured anechoic ATFs. In addition, more realistic experiments were performed using recorded signals for a binaural hearing device configuration in a reverberant cafeteria with one interfering source and multi-talker babble noise. The results showed that the BLCMV-N beamformer

- leads to a very similar interference reduction as the BLCMV beamformer
- provides a trade-off between noise reduction performance (slightly worse than the BLCMV beamformer) and binaural cue preservation of the background noise (much better than the BLCMV beamformer).

In addition, the results of a perceptual listening test with 13 normal-hearing participants showed that the proposed BLCMV-N beamformer

- is able to preserve the binaural cues and hence the spatial impression of the interfering source (like the BLCMV beamformer)
- provides a trade-off between noise reduction performance and binaural cue preservation of the background noise (like the BMVDR-N beamformer).

While in Chapter 4 we only took into account the use of the head-mounted microphones, in **Chapter 5** we investigated the incorporation of an external microphone in the BMVDR-N beamformer. We analytically showed for an arbitrary noise field and without making any assumptions about the position of the desired source that by incorporating an external microphone in the BMVDR-N beamformer

- a larger output SNR can be obtained for the same mixing parameter
- the same output SNR can be obtained for a larger mixing parameter

- the same desired output MSC of the noise component can be obtained for a smaller mixing parameter.

These results imply that an external microphone enables to achieve the same spatial impression of the noise component compared to using only the head-mounted microphones, while achieving a larger output SNR. The obtained analytical expressions were first validated using simulated anechoic ATFs. In addition, experiments were performed using recorded signals for a binaural hearing device configuration in a realistic reverberant environment with multiple competing talkers as background noise. For different positions of the external microphone and the desired source, the experimental results showed that also in a realistic acoustic scenario incorporating an external microphone in the BMVDR-N beamformer

- significantly increases the output SNR
- decreases the mixing parameter required to obtain a desired output MSC, i.e., spatial impression, of the noise component.

Finally, in **Chapter 6** we proposed computationally efficient methods to estimate the RTF vectors of the desired source, by exploiting one or more external microphones that are spatially separated from the head-mounted microphones. We first considered the extended binaural hearing device configuration with only one external microphone and proposed an SC-based RTF vector estimation method, which estimates the RTF vectors of the desired source as the last column of the extended noisy input covariance matrix (corresponding to the external microphone), normalized by the element corresponding to the reference microphone. Assuming that the SC between the noise component in the head-mounted microphone signals and the noise component in the external microphone signal is zero, we showed that

- the elements of the RTF vectors corresponding to the head-mounted microphones are unbiased
- the element corresponding to the external microphone is corrupted by a bias term.

We provided a bias analysis of the proposed SC-based RTF vector estimates for an arbitrary undesired component, a diffuse noise field and an interfering source.

- For an arbitrary undesired component, the bias depends on a CPSD ratio and the overall correlation between the undesired component in the head-mounted microphone signals and the undesired component in the external microphone signal.
- For a diffuse noise field, the assumption that the noise component in the head-mounted microphone signals is uncorrelated with the noise component in the external microphone signal is approximately satisfied when the distance between the external microphone and the head-mounted microphones is large enough.
- For an interfering source, the SC-based RTF vector estimate is equal to a mixture between the RTF vectors of the desired source and the interfering source, depending on a CPSD ratio.

When multiple external microphones are available, we showed that different RTF vector estimates can be obtained by using the SC-based method for each external microphone. We proposed several procedures to combine these RTF vector estimates, either by selecting the estimate corresponding to the highest input SNR, by averaging the estimates or by combining the estimates in order to maximize the output SNR of the eBMVDR beamformer (filtering the head-mounted and the external microphone signals). We conducted two experiments with a moving desired source and diffuse-like background noise in a reverberant environment. The results of the first experiment with one external microphone at an unknown position showed that the SC-based method

- can be used to steer an (e)BMVDR beamformer in a dynamic acoustic scenario with a moving desired source
- yields a similar (even slightly better) noise reduction performance and binaural cue preservation as the state-of-the-art CW method at a much lower computational complexity.

The results of the second experiment with three external microphones at unknown positions showed that, the SC-based RTF vector estimation method should not be used with just any external microphone, but with all external microphones such that the different SC-based RTF vector estimates can be linearly combined afterwards. The input SNR-based combination yields a similar noise reduction performance as the state-of-the-art CW method at a much lower computational complexity, while the output SNR-maximizing combination outperformed all other combination procedures and RTF vector estimation methods.

7.2 Suggestions for further research

In Chapter 4 we proposed the BLCMV-N beamformer, which merges the advantages of the BLCMV beamformer and the BMVDR-N beamformer. We showed that the output signals of the BLCMV-N beamformer can be interpreted as a mixture between the noisy reference microphone signals and the output signals of a BLCMV beamformer using an adjusted interference scaling parameter. In order to achieve a desired output MSC of the noise component, in [13] a closed-form expression for the mixing parameter of the BMVDR-N beamformer has been derived. The desired output MSC of the noise component can, e.g., be psycho-acoustically motivated based on the IC discrimination ability of the human auditory system [13, 106], aiming for the spatial impression of the noise component in the reference microphone signals and the noise component in the output signals to be indistinguishable. A similar approach could probably be applied to derive a psycho-acoustically optimal mixing parameter for the BLCMV-N beamformer. It is expected that the relative position of the interfering source to the desired source influences the psycho-acoustically optimal mixing parameter.

Furthermore, we presented a perceptual listening test for the BLCMV-N beamformer in which we subjectively evaluated the preservation of the spatial impression of individual signal components and the complete signal. Further research is however required with respect to the resulting speech intelligibility of the binaural

beamforming algorithms evaluated in this thesis, since the relation between binaural cues of all signal components (i.e., desired source component, interfering source component, noise component) and speech intelligibility is not trivial. Therefore, in the future a large scale hearing study as in [113] should be conducted, where both speech intelligibility and spatial impression are evaluated for the BLCMV-N beamformer and state-of-the-art binaural beamforming algorithms in anechoic and reverberant acoustic scenarios with an interfering source and background noise.

The last point that should be considered is the incorporation of one or more external microphones in the BLCMV-N beamformer. The incorporation of external microphones would lead to the extended BLCMV-N beamformer and raises similar questions as the incorporation of an external microphone in the BMVDR-N beamformer in Chapter 5. The theoretical and practical influence of external microphones on the noise and interference reduction performance and the optimal interference scaling parameter maximizing the output SNR of the BLCMV-N beamformer are of particular interest.

In Chapter 6 we proposed SC-based methods to estimate the RTF vectors of the desired source, by exploiting one or more external microphones that are spatially separated from the head-mounted microphones. Using the SC-based RTF vector estimates in the (e)BMVDR beamformer, the experimental results showed that the proposed methods enable to significantly reduce diffuse-like background noise in realistic acoustic scenarios with a moving desired source. However, the analytical bias analysis for an interfering source in Chapter 6 implies that the SC-based RTF vector estimates are equal to a mixture between the RTF vectors of the desired source and the interfering source (depending on a CPSD ratio). Although using the SC-based RTF vector estimates in the (e)BMVDR beamformer could hence potentially enhance multiple desired sources, the reduction of an interfering source remains a challenge, especially since it remains a challenge to determine the desired source in multi-speaker scenarios. Contrary to using the SC-based RTF vector estimates in the (e)BMVDR beamformer, a GSC structure as in [16] could be utilized where it is expected that the adaptive filter stage can better deal with an interfering source. Since in [16] the relative position of the desired source to the head-mounted microphones is assumed a-priori, the SC-based RTF vector estimates could be used instead to steer the beamformer stage and to construct the blocking matrix. By combining the approach in [16] with the proposed RTF estimation methods in this thesis, a novel scheme could be developed that can either enhance all coherent sources or just one (i.e., the desired source).

As already shown in [149], the availability of external microphones can improve the performance of DOA estimators. Since the DOA is included as information in the RTFs, RTFs were used, e.g., in [72, 74], to estimate the DOA of the desired source. It is therefore reasonable to use the RTF estimation methods proposed in this thesis with existing or novel methods for DOA estimation.

Furthermore, in this thesis we have only considered the incorporation of individual external microphones whose positions can be completely random. Instead, it is also possible to incorporate external microphone arrays, where the output signals

of these arrays could be considered as enhanced external microphone signals and novel distributed algorithms could be developed (e.g., see [14]). Moreover, experiments should be conducted in which the microphones of explicit devices such as smartspeakers, smartphones or laptops are used as external microphones or external microphone arrays with the proposed methods in realistic scenarios to prove their suitability for everyday use.

Finally, some practical and technical questions still remain unanswered, e.g., the influence of synchronization between the external microphone signals and the head-mounted microphone signals on the performance of the binaural beamforming algorithms and RTF vector estimation methods presented in this thesis. If it turns out that synchronization has a significant impact on the performance of the proposed methods, existing synchronization method (e.g., [127–134]) can be a first starting point to develop novel and customized methods.

To conclude, we have seen in this thesis what tremendous possibilities the incorporation of external microphones offers. These possibilities are not limited to binaural hearing device configurations, meaning that the incorporation of external microphones and the interconnection of several microphones to form an acoustic sensor network will provide great potential in many areas of acoustic signal processing in the future.

A

APPENDIX TO CHAPTER 4

In Appendix A.1 we derive the BLCMV beamformer with partial noise estimation (BLCMV-N). In Appendix A.2 we derive the left and the right output noise PSD, and the output noise CPSD for the BLCMV-N beamformer.

A.1 Derivation of the BLCMV-N beamformer

Using (2.6), (2.7), (2.10), (3.19) and (3.20), the constrained optimization problem in (4.1) and (4.2) can be reformulated as

$$\min_{\mathbf{w}_L} \mathcal{E}\left\{\left|\mathbf{w}_L^H\mathbf{n} - \eta n_L\right|^2\right\} \quad \text{subject to} \quad \mathbf{C}^H\mathbf{w}_L = \mathbf{g}_L\,, \tag{A.1}$$

$$\min_{\mathbf{w}_R} \mathcal{E}\left\{\left|\mathbf{w}_R^H\mathbf{n} - \eta n_R\right|^2\right\} \quad \text{subject to} \quad \mathbf{C}^H\mathbf{w}_R = \mathbf{g}_R\,. \tag{A.2}$$

This constrained optimization problem can be solved using the method of Lagrange multipliers, where the Lagrangian function is given by

$$\begin{aligned}\mathcal{L}(\mathbf{w}_L, \boldsymbol{\lambda}_L) = {} & \mathbf{w}_L^H\mathbf{R}_n\mathbf{w}_L - \eta\mathbf{e}_L^T\mathbf{R}_n\mathbf{w}_L - \eta\mathbf{w}_L^H\mathbf{R}_n\mathbf{e}_L \\ & + \eta^2 p_{n_L} + \boldsymbol{\lambda}_L^H\left(\mathbf{C}^H\mathbf{w}_L - \mathbf{g}_L\right) + \left(\mathbf{w}_L^H\mathbf{C} - \mathbf{g}_L^H\right)\boldsymbol{\lambda}_L\,,\end{aligned} \tag{A.3}$$

$$\begin{aligned}\mathcal{L}(\mathbf{w}_R, \boldsymbol{\lambda}_R) = {} & \mathbf{w}_R^H\mathbf{R}_n\mathbf{w}_R - \eta\mathbf{e}_R^T\mathbf{R}_n\mathbf{w}_R - \eta\mathbf{w}_R^H\mathbf{R}_n\mathbf{e}_R \\ & + \eta^2 p_{n_R} + \boldsymbol{\lambda}_R^H\left(\mathbf{C}^H\mathbf{w}_R - \mathbf{g}_R\right) + \left(\mathbf{w}_R^H\mathbf{C} - \mathbf{g}_R^H\right)\boldsymbol{\lambda}_R\,,\end{aligned} \tag{A.4}$$

with $\boldsymbol{\lambda}_L$ and $\boldsymbol{\lambda}_R$ denoting the left and the right 2-dimensional vector of Lagrangian multipliers. Setting the gradient with respect to $\mathbf{w}_L$ and $\mathbf{w}_R$

$$\nabla_{\mathbf{w}_L}\mathcal{L}(\mathbf{w}_L, \boldsymbol{\lambda}_L) = 2\mathbf{R}_n\mathbf{w}_L - 2\eta\mathbf{R}_n\mathbf{e}_L + 2\mathbf{C}\boldsymbol{\lambda}_L\,, \tag{A.5}$$

$$\nabla_{\mathbf{w}_R}\mathcal{L}(\mathbf{w}_R, \boldsymbol{\lambda}_R) = 2\mathbf{R}_n\mathbf{w}_R - 2\eta\mathbf{R}_n\mathbf{e}_R + 2\mathbf{C}\boldsymbol{\lambda}_R \tag{A.6}$$

equal to $\mathbf{0}$ yields

$$\mathbf{w}_L = \eta\mathbf{e}_L - \mathbf{R}_n^{-1}\mathbf{C}\boldsymbol{\lambda}_L\,, \tag{A.7}$$

$$\mathbf{w}_R = \eta\mathbf{e}_R - \mathbf{R}_n^{-1}\mathbf{C}\boldsymbol{\lambda}_R\,. \tag{A.8}$$

Substituting (A.7) and (A.8) into the constraint $\mathbf{C}^H\mathbf{w}_L = \mathbf{g}_L$ and $\mathbf{C}^H\mathbf{w}_R = \mathbf{g}_R$, respectively, and solving for the Lagrangian multiplier $\boldsymbol{\lambda}_L$ and $\boldsymbol{\lambda}_R$ yields

$$\boldsymbol{\lambda}_L = \left(\mathbf{C}^H\mathbf{R}_n^{-1}\mathbf{C}\right)^{-1}\left(\eta\mathbf{C}^H\mathbf{e}_L - \mathbf{g}_L\right) , \tag{A.9}$$

$$\boldsymbol{\lambda}_R = \left(\mathbf{C}^H\mathbf{R}_n^{-1}\mathbf{C}\right)^{-1}\left(\eta\mathbf{C}^H\mathbf{e}_R - \mathbf{g}_R\right) . \tag{A.10}$$

Substituting (A.9) and into (A.7) and (A.8), the solution to (4.1) and (4.2) is given by

$$\mathbf{w}_{\mathrm{BLCMV-N},L} = \eta\mathbf{e}_L + \mathbf{R}_n^{-1}\mathbf{C}\left(\mathbf{C}^H\mathbf{R}_n^{-1}\mathbf{C}\right)^{-1}\left(\mathbf{g}_L - \eta\mathbf{C}^H\mathbf{e}_L\right) , \tag{A.11}$$

$$\mathbf{w}_{\mathrm{BLCMV-N},R} = \eta\mathbf{e}_R + \mathbf{R}_n^{-1}\mathbf{C}\left(\mathbf{C}^H\mathbf{R}_n^{-1}\mathbf{C}\right)^{-1}\left(\mathbf{g}_R - \eta\mathbf{C}^H\mathbf{e}_R\right) , \tag{A.12}$$

where, using (3.19), (3.20) and (3.21),

$$\mathbf{g}_L - \eta\mathbf{C}^H\mathbf{e}_L = \begin{bmatrix}(1-\eta)a_L^* \\ (\delta-\eta)b_L^*\end{bmatrix} , \tag{A.13}$$

$$\mathbf{g}_R - \eta\mathbf{C}^H\mathbf{e}_R = \begin{bmatrix}(1-\eta)a_R^* \\ (\delta-\eta)b_R^*\end{bmatrix} . \tag{A.14}$$

A.2 Output noise PSD for the BLCMV-N beamformer

Using (4.27) in (2.66), the left output PSD of the noise component for the BLCMV-N beamformer is given by

$$\begin{aligned}\mathbf{w}_L^H\mathbf{R}_n\mathbf{w}_L = {} & \eta^2\mathbf{e}_L^T\mathbf{R}_n\mathbf{e}_L \\ & + \eta(1-\eta)\left[a_L\mathbf{w}_x^H\mathbf{R}_n\mathbf{e}_L + \mathbf{e}_L^T\mathbf{R}_n\mathbf{w}_x a_L^*\right] \\ & + \eta(\delta-\eta)\left[b_L\mathbf{w}_u^H\mathbf{R}_n\mathbf{e}_L + \mathbf{e}_L^T\mathbf{R}_n\mathbf{w}_u b_L^*\right] \\ & + (1-\eta)^2|a_L|^2\mathbf{w}_x^H\mathbf{R}_n\mathbf{w}_x \\ & + (\delta-\eta)(1-\eta)\left[a_L^*b_L\mathbf{w}_u^H\mathbf{R}_n\mathbf{w}_x\right. \\ & + \left. a_Lb_L^*\mathbf{w}_x^H\mathbf{R}_n\mathbf{w}_u\right] \\ & + (\delta-\eta)^2|b_L|^2\mathbf{w}_u^H\mathbf{R}_n\mathbf{w}_u .\end{aligned} \tag{A.15}$$

Using (4.23) and (4.26), the components in (A.15) are given by

$$\mathbf{e}_L^T\mathbf{R}_n\mathbf{w}_x = \frac{1}{1-\Psi}\left(\frac{a_L}{\gamma_a} - \Psi\frac{b_L}{\gamma_{ab}}\right), \tag{A.16}$$

$$\mathbf{w}_x^H\mathbf{R}_n\mathbf{w}_x = \frac{1}{(1-\Psi)\gamma_a}, \tag{A.17}$$

$$\mathbf{e}_L^T\mathbf{R}_n\mathbf{w}_u = \frac{1}{1-\Psi}\left(\frac{b_L}{\gamma_b} - \Psi\frac{a_L}{\gamma_{ab}^*}\right), \tag{A.18}$$

$$\mathbf{w}_u^H\mathbf{R}_n\mathbf{w}_u = \frac{1}{(1-\Psi)\gamma_b}, \tag{A.19}$$

$$\mathbf{w}_x^H\mathbf{R}_n\mathbf{w}_u = \frac{\Psi}{(1-\Psi)\gamma_{ab}^*}. \tag{A.20}$$

Substituting (A.16)–(A.20) in (A.15) yields

$$\begin{aligned}\mathbf{w}_L^H\mathbf{R}_n\mathbf{w}_L &= \eta^2 p_{n_L} \qquad\qquad (A.21)\\ &\quad + \frac{1}{1-\Psi}\left[(1-\eta^2)\frac{|a_L|^2}{\gamma_a} + (\delta^2-\eta^2)\frac{|b_L|^2}{\gamma_b} - 2\Psi(\delta-\eta^2)\Re\left\{\frac{a_L b_L^*}{\gamma_{ab}^*}\right\}\right],\\ &= \mathbf{e}_L^T(\eta^2\mathbf{R}_n + \mathbf{R}_{xu,3})\mathbf{e}_L, \qquad\qquad (A.22)\end{aligned}$$

with $\mathbf{R}_{xu,3}$ defined in (4.41). Similarly, it can be shown that the right output PSD of the noise component for the BLCMV-N beamformer is equal to

$$\mathbf{w}_R^H\mathbf{R}_n\mathbf{w}_R = \mathbf{e}_R^T(\eta^2\mathbf{R}_n + \mathbf{R}_{xu,3})\mathbf{e}_R, \tag{A.23}$$

and that the output CPSD of the noise component for the BLCMV-N beamformer is equal to

$$\mathbf{w}_L^H\mathbf{R}_n\mathbf{w}_R = \mathbf{e}_L^T(\eta^2\mathbf{R}_n + \mathbf{R}_{xu,3})\mathbf{e}_R. \tag{A.24}$$

B

APPENDIX TO CHAPTER 6

In this appendix we present two additional experiments for Chapter 6, where we consider a static desired source and one external microphone. In Appendix B.1 we present an experiment (published in [159]), where we investigate the influence of the input SNR and the time constants used for covariance matrix estimation on the RTF vector estimation accuracy and the noise reduction performance when using the RTF vector estimates in a BMVDR beamformer. In Appendix B.2 we present an experiment (published in [160]), where we investigate the influence of input SNR and reverberation time on the noise reduction performance and binaural cue preservation of the BMVDR beamformer.

B.1 Experiment – RTF vector estimation accuracy and noise reduction performance of BMVDR beamformer

In this appendix we consider a static desired source and one external microphone and compare the performance of the proposed SC-based RTF vector estimation method in (6.17) and (6.18) (using the head-mounted microphones and one external microphone) with the state-of-the-art CW, CS, PM-CW, PM-CS and CS-R1 RTF vector estimation methods discussed in Section 3.4.2 (using only the head-mounted microphones). It should be noted that in this experiment we will only use the external microphone to estimate the M_H-dimensional RTF vector $\mathbf{a}_L$ corresponding to the head-mounted microphones. First, we describe the experimental setup and the algorithmic parameters. Second, we evaluate the RTF vector estimation accuracy and the noise reduction performance when using the RTF vector estimates in a BMVDR beamformer, i.e., not including the external microphone signal.

For the simulations we used the database of real-world recordings (sampling rate of $f_s = 16\,\text{kHz}$) described in [36]. The room dimensions were about $12.7 \times 10 \times 3.6\ \text{m}^3$ with a reverberation time of about 620 ms. We used $M_H = 4$ head-mounted microphones, i.e., two microphones on each hearing device. As reference microphone we chose the front microphone on the left hearing device. The external microphone was located on a table in front of the desired source at about 60 cm from the reference microphone (cf. position P_1 in Figure 5.8). The desired source was an English-speaking female speaker who was sitting to the right of the hearing device user at an angle of about 45° (cf. speaker S_2 in Figure 5.8). Both the hearing device user and the desired

source were seated at a circular table with a diameter of 106 cm. In addition, 56 other speakers which were also seated at tables generated a realistic babble noise. The noise component hence contained mainly diffuse but also directional components from temporally dominant interfering speakers. To generate the noisy microphone signals separate recordings of the desired source component and the noise component were mixed together at different input SNRs $\{-10, -5, 0, 5, 10\}$ dB. The SNR in the external microphone signal was about 13 dB higher than in the reference microphone signal (due to distance and head shadow effect). In this experiment we calculated all SNRs using the intelligibility-weighted SNR [189]. The total signal length was about 23 s.

We used an STFT framework with a frame size of $T_d = 512$ samples and a frame shift of $T_s = 256$ samples and a square-root Hann window. To estimate the extended noisy input covariance matrix $\mathbf{R}_{y,e}$ using speech-plus-noise frames and the noise covariance matrix $\mathbf{R}_n$ using noise-only frames, we used a simple oracle broadband energy-based VAD calculated from the desired source component in the reference microphone signal. To recursively estimate these covariance matrices, we used the time constants τ_y and τ_n, respectively. The corresponding smoothing factors (cf. Section 3.4.1) were computed using (3.71). Please note that a smaller time constant corresponds to a smaller smoothing factor and hence to a faster adaptation to possible changes, but may also lead to less accurate estimates of the covariance matrices. Especially in a scenario where the microphones or the desired source may change their position, a small time constant is desirable to be able to track changes fast enough. Because the background noise can be assumed to be rather stationary, we set the corresponding time constant to $\tau_n = 500$ ms. The time constant used to recursively estimate the extended noisy input covariance matrix $\mathbf{R}_{y,e}$ was chosen as $\tau_y \in \{50, 100, 150, 200\}$ ms. The noisy input covariance matrix $\mathbf{R}_y$ (without the external microphone signal) was estimated by using a subset of the extended noisy input covariance matrix $\mathbf{R}_{y,e}$. The covariance matrix estimates were initialised using the corresponding batch estimates in (3.65) and (3.66).

As suggested in [92], to evaluate the RTF vector estimation accuracy we used the Hermitian angle between a reference RTF vector $\bar{\mathbf{a}}$ and an estimated RTF vector $\hat{\mathbf{a}}$, i.e.,

$$\Theta(f,t) = \arccos \frac{|\bar{\mathbf{a}}^H(f,t)\hat{\mathbf{a}}(f,t)|}{\|\bar{\mathbf{a}}(f,t)\|_2 \|\hat{\mathbf{a}}(f,t)\|_2} \,. \tag{B.1}$$

The reference RTF vector $\bar{\mathbf{a}}$ was calculated as the principal eigenvector of the batch desired source covariance matrix $\hat{\mathbf{R}}_x^{\text{bat}}$ (estimated using all available speech frames, cf. Section 3.4.1), normalised by its first element (corresponding to the reference microphone).

Figure B.1 depicts the results (averaged over all frequencies and time frames) for different time constants τ_y as a function of input SNR. As expected, the performance of all RTF vector estimation methods improves by increasing the input SNR and the time constant. It can be observed that the proposed SC method generally outperforms the other methods for all input SNRs and time constants. The CS method showed worse performance, which is in line with the literature [90, 93]. Only for

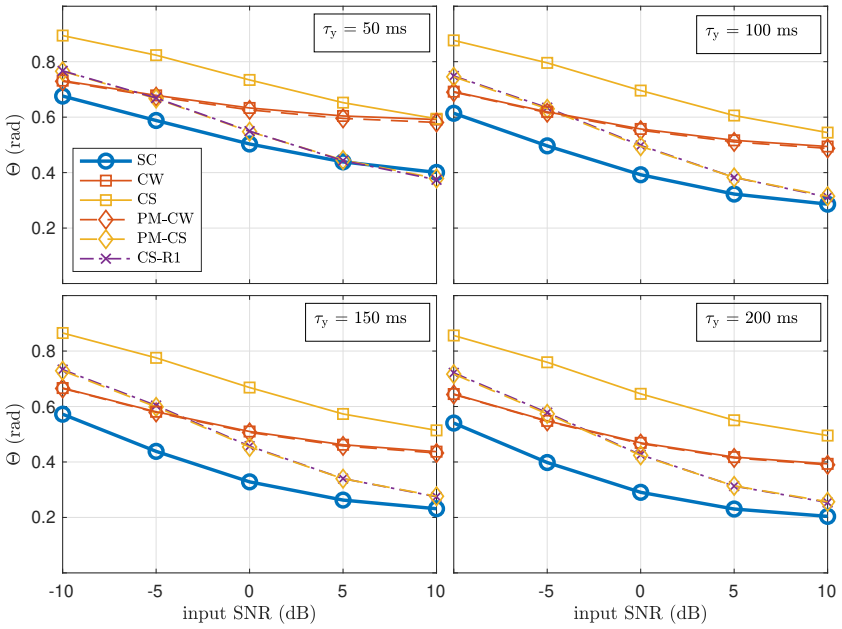

Fig. B.1: Hermitian angle Θ between the reference RTF vector $\bar{\mathbf{a}}$ and the estimated RTF vectors (averaged all frequencies and time frames) for different input SNRs and different time constants τ_y.

a time constant of $\tau_y = 50\,\text{ms}$ and a high input SNR of 10 dB, the CS-R1 and PM-CS methods slightly outperformed the proposed SC method. For an exemplary input SNR of 0 dB and a time constant of 50 ms, Figure B.2 depicts the Hermitian angles (averaged over all frequencies) for the first 100 time frames. The proposed SC method starts to adapt after about 22 frames because this is the first frame where the desired source was active. All other methods rely on estimates of both the noisy input and noise covariance matrices and hence adapt during noise-only and speech-plus-noise frames. The CS-R1 and CW estimators both seem to benefit from the long-term (batch) initializations in the first frames but perform worse than the proposed estimator afterwards.

We evaluated the noise reduction performance when using the time-varying estimated RTF vectors to steer a BMVDR beamformer (only the left output signal was considered) and the time-varying estimate of $\mathbf{R}_n$ in (3.6). Please note that for all RTF vector estimation methods the BMVDR beamformer is M_H-dimensional. Figure B.3 depicts the SNR improvement (ΔSNR) calculated by applying the BMVDR beamformer to the desired source and noise components separately. As can be observed, the proposed SC estimator clearly outperforms all other estimators for all input SNRs and time constants.

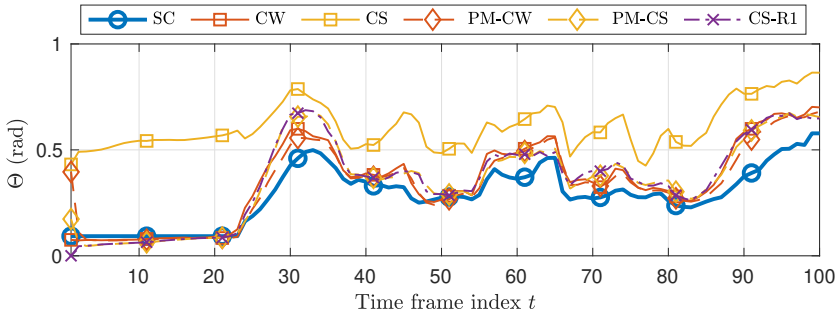

Fig. B.2: Hermitian angle Θ between the reference RTF vector $\bar{\mathbf{a}}$ and the estimated RTF vectors (averaged over all frequencies) for an input SNR of 0 dB and $\tau_y = 50\,\text{ms}$.

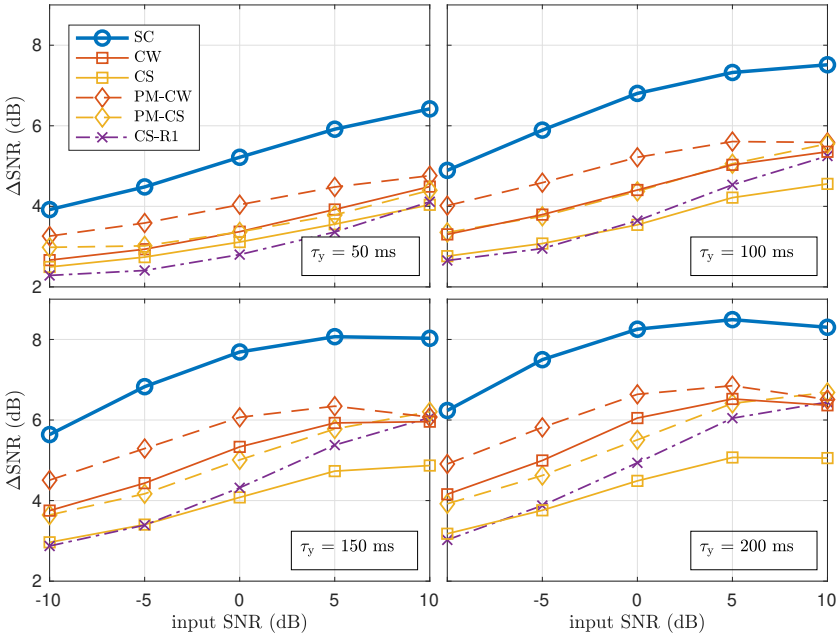

Fig. B.3: SNR improvement ΔSNR of a BMVDR beamformer (left output) steered by using the estimated RTF vectors for different time constants τ_y.

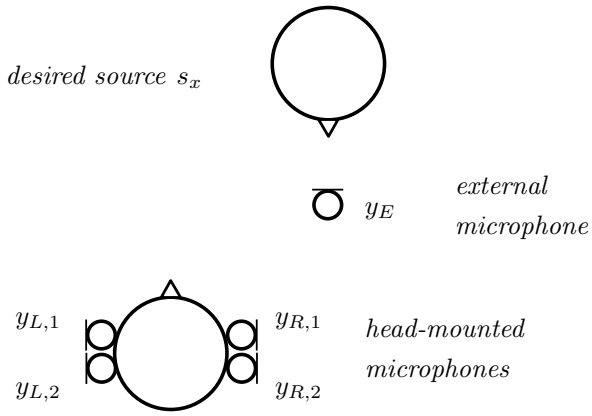

Fig. B.4: Experimental setup with $M_H = 4$ head-mounted microphones, $M_E = 1$ external microphone and one static desired source.

B.2 Experiment – Influence of input SNR and reverberation time

In this appendix we consider a static desired source and one external microphone and compare the noise reduction performance and binaural cue preservation of the BMVDR beamformer in (3.6) and (3.7) using the proposed SC method in (6.19) and (6.19) (only using the first M_H elements, cf. (6.17) and (6.18)) with the biased (B) and covariance whitening (CW) RTF vector estimation methods discussed in Section 3.4.2. Since in practice the assumption in (6.5) does not perfectly hold, in this experiment we also consider an oracle version ($\mathrm{SC}_{\mathrm{opt}}$) of the proposed SC method, which uses the clean desired speech signal s_x as the external microphone signal, such that (6.5) perfectly holds, i.e.,

$$\hat{\mathbf{a}}_L^{\mathrm{SC}_{\mathrm{opt}}} = \frac{\mathcal{E}\{\mathbf{y}s_x^*\}}{\mathcal{E}\{y_L s_x^*\}}\,, \tag{B.2}$$

$$\hat{\mathbf{a}}_R^{\mathrm{SC}_{\mathrm{opt}}} = \frac{\mathcal{E}\{\mathbf{y}s_x^*\}}{\mathcal{E}\{y_R s_x^*\}}\,. \tag{B.3}$$

All signals were recorded in a variable acoustics laboratory located at the University of Oldenburg, where the reverberation time can be easily changed by closing and opening absorber panels mounted to the walls and the ceiling. The room dimensions are about $7 \times 6 \times 2.7$ m^3, where the reverberation time was set approximately to the three different values $T_{60} \in \{250,\ 500,\ 750\}$ ms. The reverberation times were measured using the broad band energy decay curve of the measured impulse responses. At the center of the laboratory a KEMAR head-and-torso simulator (HATS) was placed. Two behind-the-ear hearing aid dummies with two microphones

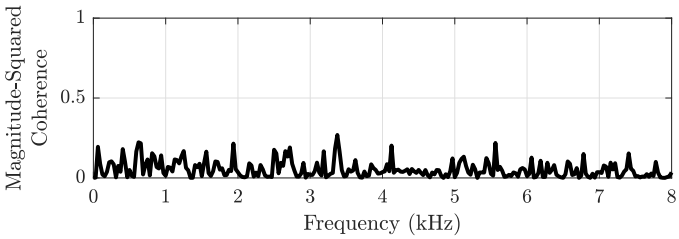

Fig. B.5: Measured long-term magnitude-squared coherence between the noise component in the left reference microphone signal and the noise component in the external microphone signal.

each, i.e., $M_H = 4$, were placed on the ears of the HATS, with an inter-microphone distance of about 14 mm. We chose the frontal microphone on each hearing aid as the reference microphone.

The desired source s_x was a male English speaker played back by a loudspeaker which was placed at about 2 m from the center of the head at the same height and at an angle of about 35°, i.e., to the right side of the HATS (cf. Figure B.4). The external microphone was placed at about 0.5 m from the desired source, leading to a distance of about 1.5 m to the HATS. To generate diffuse-like background noise, we used four loudspeakers facing the corners of the laboratory, playing back different multi-talker recordings. Figure B.5 shows the long-term magnitude-squared coherence between the recorded noise component in the left reference microphone signal and the noise component in the external microphone signal. It can be observed that while the assumption in (6.5) does not perfectly hold, the coherence is fairly small. The desired source component and the diffuse-like background noise component were recorded separately in order to be able to mix them together at different input SNRs $\in \{-5, 0, 5\}$ dB. The SNR in the external microphone signal was about 9.6 dB higher than in the head-mounted microphone signals. Please note that streaming and directly using the external microphone signal would not include any binaural cues. The complete signal had a length of 20 s with 0.5 s of noise-only at the beginning.

All signals were processed at a sampling rate of 16 kHz. We used an STFT framework with time frame size $T_d = 256$, corresponding to 16 ms, a time frame shift of $T_s = 128$ and a square-root Hann window. To distinguish between speech-plus-noise and noise-only frames, we used an oracle broad band VAD, based on the energy of the desired source component in the right reference microphone signal (cf. Section 3.4.1). Using this VAD, the extended noisy input covariance matrix $\mathbf{R}_{y,e}(f,t)$ and the noise covariance matrix $\mathbf{R}_n(f,t)$ were recursively estimated as (cf. (3.68) and (3.69))

$$\hat{\mathbf{R}}_{y,e}^{\text{onl}}(f,t) = \alpha_y \hat{\mathbf{R}}_{y,e}^{\text{onl}}(f,t-1) + (1-\alpha_y)\mathbf{y}_e(f,t)\mathbf{y}_e^H(f,t)\,, \tag{B.4}$$

$$\hat{\mathbf{R}}_{n}^{\text{onl}}(f,t) = \alpha_n \hat{\mathbf{R}}_{n}^{\text{onl}}(f,t-1) + (1-\alpha_n)\mathbf{y}(f,t)\mathbf{y}^H(f,t)\,, \tag{B.5}$$

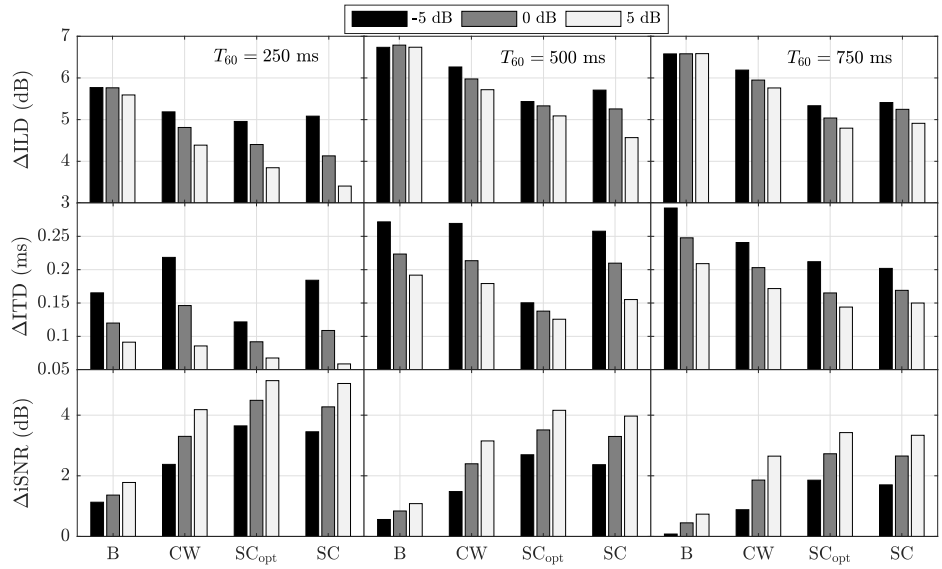

Fig. B.6: Binaural cue errors and iSNR improvement for the RTF vector estimators for different reverberation times (250 ms, 500 ms, 750 ms) and different input SNRs (-5 dB, 0 dB, 5 dB).

during detected speech-plus-noise frames and noise-only frames, respectively. The smoothing factors were chosen as $\alpha_y = 0.8521$ and $\alpha_n = 0.9841$, corresponding to time constants of 50 ms for speech-pluse-noise and 500 ms for noise-only, respectively. The noisy input covariance matrix $\hat{\mathbf{R}}_y^{\text{onl}}$ (without the external microphone signal) was calculated as a subset of the extended noisy input covariance matrix $\hat{\mathbf{R}}_{y,e}^{\text{onl}}$ (with the external microphone signal). The corresponding batch estimates of the covariance matrices were used to initialize these matrices.

The (time-varying) estimates of the covariance matrices were then used to compute the RTF vectors of the desired source, either using the biased method (B) in (3.75) and (3.76), the covariance-whitening method (CW) in (3.84) and (3.85), the oracle SC-based method (SC$_{\text{opt}}$) in (B.2) and (B.3) and the proposed SC-based method (SC) in (6.17) and (6.18). We then computed the (time-varying) BMVDR beamformer in (3.6) and (3.7) using the estimated RTF vectors and the estimated noise covariance matrix $\hat{\mathbf{R}}_n^{\text{onl}}(f,t)$ in (3.69).

The performance of the BMVDR beamformer using the considered RTF vector estimation methods was evaluated in terms of noise reduction and binaural cue preservation. As a performance measure for noise reduction we used the intelligibility-weighted SNR improvement (ΔiSNR) [189] between the right reference microphone signal and the output signal of the right hearing device. The ΔiSNR is defined as the sum of the SNR improvements (cf. Section 2.2.1) in all frequency bins weighted with a frequency-dependent band importance function, for which we used the same weights as for the speech intelligibility index in [30] (based on one-third octaves).

As a performance measure for binaural cue preservation (cf. Section 2.2.2) we used the reliable binaural cue errors of the direct sound of the desired source component in the left and the right output signals, i.e., ΔILD and ΔITD, based on an auditory model [55] and averaged over all frequencies.

Figure B.6 depicts the results for all four considered RTF vector estimation methods for different reverberation times and input SNRs. As expected, the biased method generally shows worst performance in terms of binaural cue preservation and noise reduction performance. Considering the ILD error, it can be observed for all methods that the ILD errors generally increase for increasing T_{60} and decreasing input SNR. In addition, it can be observed that the SC method consistently outperforms the CW method, especially for large T_{60}. Moreover, almost no difference can be observed between the SC method and the oracle $\mathrm{SC}_{\mathrm{opt}}$ method, for all T_{60} and input SNRs. Considering the ITD errors, it can be observed for all methods that the ITD errors generally increase for increasing T_{60} and decreasing input SNRs. Contrary to the ILD error, the SC method typically leads to larger ITD errors than the oracle $\mathrm{SC}_{\mathrm{opt}}$ method, especially for $T_{60} = 250\,\mathrm{ms}$ and $500\,\mathrm{ms}$. Informal listening tests showed that when using the SC method (and $\mathrm{SC}_{\mathrm{opt}}$) the desired source is perceived as a point source and sounded slightly less reverberated than the reference microphone signals. For the biased and CW methods, the binaural cue error sometimes showed large variations over all frequencies, which may lead to strange sounding artefacts, such that some frequencies are perceived as coming from another direction and the desired source sounds slightly diffuse.

Considering the iSNR improvement, it can be observed that for all estimation methods the SNR improvement generally decreases for increasing T_{60} and decreasing input SNR. In addition, it can be observed that the SC method consistently outperforms the CW method for all T_{60} and input SNRs. Moreover, almost no difference can be observed between the SC method and the oracle $\mathrm{SC}_{\mathrm{opt}}$ method.

From these results, it can be concluded that the SC method (exploiting the external microphone in addition to the head-mounted microphones) outperforms the CW method (only using the head-mounted microphones). It should be noted that although the SC methods needs an external microphone, it comes with a much lower computational complexity compared with the CW method, since the SC method only relies on an estimate of the extended noisy input covariance matrix and does not need to perform an EVD. Moreover, for the considered scenario, i.e., the external microphone about 0.5 m from the desired source and about 1.5 m from the head-mounted microphones, the overall performance of the (practically implementable) SC method is very similar to the oracle $\mathrm{SC}_{\mathrm{opt}}$ method, showing that the spatial coherence assumption in (6.5) is valid for the considered scenario. It can be expected that placing the external microphone closer to the desired source would slightly improve the performance of the SC method, especially in terms of binaural cue preservation.

BIBLIOGRAPHY

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, Sep. 1953.

[2] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[3] S. Doclo, S. Gannot, D. Marquardt, and E. Hadad, "Binaural speech processing with application to hearing devices," in *Audio Source Separation and Speech Enhancement*, Wiley, 2018, ch. 18, pp. 413–442.

[4] V. Hamacher, U. Kornagel, T. Lotter, and H. Puder, "Binaural signal processing in hearing aids: Technologies and algorithms," in *Advances in Digital Speech Transmission*, Wiley, 2008, ch. 14, pp. 401–429.

[5] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, Jan. 2000.

[6] T. Wittkop and V. Hohmann, "Strategy-selective noise reduction for binaural digital hearing aids," *Speech Communication*, vol. 39, no. 1-2, pp. 111–138, Jan. 2003.

[7] K. Reindl, Y. Zheng, A. Schwarz, S. Meier, R. Maas, A. Sehr, and W. Kellermann, "A stereophonic acoustic signal extraction scheme for noisy and reverberant environments," *Computer Speech and Language*, vol. 27, no. 3, pp. 726–745, May 2013.

[8] D. P. Welker, J. E. Greenberg, J. G. Desloge, and P. M. Zurek, "Microphone-array hearing aids with binaural output. II. A two-microphone adaptive system," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 6, pp. 543–551, Nov. 1997.

[9] T. J. Klasen, T. van den Bogaert, M. Moonen, and J. Wouters, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Transactions on Signal Processing*, vol. 55, no. 4, pp. 1579–1585, Apr. 2007.

[10] B. Cornelis, S. Doclo, T. van den Bogaert, J. Wouters, and M. Moonen, "Theoretical analysis of binaural multi-microphone noise reduction techniques," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 342–355, Feb. 2010.

[11] D. Marquardt, "Development and evaluation of psychoacoustically motivated binaural noise reduction and cue preservation techniques," PhD thesis, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany, Nov. 2015.

[12] E. Hadad, S. Doclo, and S. Gannot, "The binaural LCMV beamformer and its performance analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 543–558, Mar. 2016.

[13] D. Marquardt and S. Doclo, "Interaural coherence preservation for binaural noise reduction using partial noise estimation and spectral postfiltering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1261–1274, Jul. 2018.

[14] 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, 14 pages, Jan. 2009.

[15] J. Szurley, A. Bertrand, B. van Dijk, and M. Moonen, "Binaural noise cue preservation in a binaural noise reduction system with a remote microphone signal," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 952–966, May 2016.

[16] R. Ali, G. Bernardi, T. van Waterschoot, and M. Moonen, "Methods of extending a generalised sidelobe canceller with external microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 9, pp. 1349–1364, May 2019.

[17] J. L. Flanagan, *Speech analysis synthesis and perception*, 3rd ed. Springer Science & Business Media, 2013.

[18] P. C. Loizou, *Speech enhancement: Theory and practice*. CRC Press, 2013.

[19] S. Van Gerven and F. Xie, "A comparative study of speech detection methods," in *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, Rhodos, Greece, Sep. 1997, pp. 1095–1098.

[20] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[21] J.-H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 1965–1976, Jun. 2006.

[22] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, Apr. 2013.

[23] S. Mousazadeh and I. Cohen, "Voice activity detection in presence of transient noise using spectral clustering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1261–1271, Jun. 2013.

[24] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[25] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[26] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1072–1077, Jul. 2010.

[27] 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, no. 4, pp. 1383–1393, May 2012.

[28] G. Fant, *Acoustic theory of speech production*, 2nd ed. Walter de Gruyter, 1970.

[29] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *The Journal of the Acoustical Society of America*, vol. 19, no. 1, pp. 90–119, Jan. 1947.

[30] ANSI S3.5-1997, *Methods for calculation of the speech intelligibility index*, 1997.

[31] M. Jeub, M. Dörbecker, and P. Vary, "A semi-analytical model for the binaural coherence of noise fields," *IEEE Signal Processing Letters*, vol. 18, no. 3, pp. 197–200, Mar. 2011.

[32] B. F. Cron and C. H. Sherman, "Spatial-correlation functions for various noise models," *The Journal of the Acoustical Society of America*, vol. 34, pp. 1732–1736, Nov. 1962.

[33] H. Cox, "Spatial correlation in arbitrary noise fields with application to ambient sea noise," *The Journal of the Acoustical Society of America*, vol. 54, no. 5, pp. 1289–1301, Nov. 1973.

[34] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing*. Pearson Education, 2014.

[35] 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 Journal on Advances in Signal Processing*, vol. 2009, pp. 1–10, Jan. 2009.

[36] W. S. Woods, E. Hadad, I. Merks, B. Xu, S. Gannot, and T. Zhang, "A real-world recording database for ad hoc microphone arrays," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2015, pp. 1–5.

[37] F. Denk, S. M. Ernst, S. D. Ewert, and B. Kollmeier, "Adapting hearing devices to the individual ear acoustics: Database and target response correction functions for various device styles," *Trends in Hearing*, vol. 22, pp. 1–19, Jun. 2018.

[38] R. M. Corey, N. Tsuda, and A. C. Singer, "Acoustic impulse responses for wearable audio devices," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 216–220.

[39] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[40] D. P. Jarret, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: Algorithm and application," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1462–1472, Sep. 2012.

[41] *SMIR Generator*. [Online]. Available: `https://www.audiolabs-erlangen.de/fau/professor/habets/software/smir-generator` (visited on 04/03/2020).

[42] H. Kuttruff, *Acoustics: An introduction*. CRC Press, 2006.

[43] P. A. Naylor and N. D. Gaubitch, Eds., *Speech dereverberation*. Springer, 2010.

[44] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337–340, May 2007.

[45] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[46] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[47] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.

[48] J. Blauert, *Spatial hearing: The psychophysics of human sound localization*. MIT press, 1997.

[49] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1648–1661, Mar. 1992.

[50] J. S. Bradley and G. A. Soulodre, "Objective measures of listener envelopment," *The Journal of the Acoustical Society of America*, vol. 98, no. 5, pp. 2590–2597, Nov. 1995.

[51] K. Kurozumi and K. Ohgushi, "The relationship between the cross-correlation coefficient of two-channel acoustic signals and sound image quality," *The Journal of the Acoustical Society of America*, vol. 74, no. 6, pp. 1726–1733, Dec. 1983.

[52] M. R. Schroeder, D. Gottlob, and K. F. Siebrasse, "Comparative study of European concert halls: correlation of subjective preference with geometric and acoustic parameters," *The Journal of the Acoustical Society of America*, vol. 56, no. 4, pp. 1195–1201, Oct. 1974.

[53] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *The Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3075–3089, Nov. 2004.

[54] B. Rakerd and W. M. Hartmann, "Localization of sound in rooms. V. Binaural coherence and human sensitivity to interaural time differences in noise," *The Journal of the Acoustical Society of America*, vol. 128, no. 5, pp. 3052–3063, Nov. 2010.

[55] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Communication*, vol. 53, no. 5, pp. 592–605, May 2011.

[56] D. D. Dirks and R. H. Wilson, "The effect of spatially separated sound sources on speech intelligibility," *Journal of Speech and Hearing Research*, vol. 12, no. 1, pp. 5–38, Mar. 1969.

[57] A. W. Bronkhorst and R. Plomp, "The effect of head-induced interaural time and level differences on speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 83, no. 4, pp. 1508–1516, Jun. 1988.

[58] ——, "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing," *The Journal of the Acoustical Society of America*, vol. 92, no. 6, pp. 3132–3139, Dec. 1992.

[59] M. L. Hawley, R. Y. Litovsky, and J. F. Culling, "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," *The Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 833–843, Feb. 2004.

[60] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 331–342, Jul. 2006.

[61] M. Lavandier and J. F. Culling, "Prediction of binaural speech intelligibility against noise in rooms," *The Journal of the Acoustical Society of America*, vol. 127, no. 1, pp. 387–399, Jan. 2010.

[62] M. T. Pastore and W. A. Yost, "Spatial release from masking with a moving target," *Frontiers in Psychology*, vol. 8, no. 2238, Dec. 2017.

[63] I. Arweiler and J. M. Buchholz, "The influence of spectral characteristics of early reflections on speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 130, no. 2, pp. 996–1005, Aug. 2011.

[64] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *The Journal of the Acoustical Society of America*, vol. 78, no. 5, pp. 1508–1518, Nov. 1985.

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

[66] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer, 2008.

[67] H. Cox, R. M. Zeskind, and T. Kooji, "Practical supergain," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 3, pp. 393–398, Jun. 1986.

[68] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone arrays*, 2001, ch. 2, pp. 19–38.

[69] S. Goetze, T. Rohdenburg, V. Hohmann, B. Kollmeier, and K.-D. Kammeyer, "Direction of arrival estimation based on the dual delay line approach for binaural hearing aid microphone arrays," in *Proc. International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, Xiamen, China, Nov. 2007, pp. 84–87.

[70] I. Merks, G. Enzner, and T. Zhang, "Sound source localization with binaural hearing aids using adaptive blind channel identification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 438–442.

[71] H. Kayser and J. Anemüller, "A discriminative learning approach to probabilistic acoustic source localization," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, France, Sep. 2014, pp. 99–103.

[72] S. Braun, W. Zhou, and E. A. P. Habets, "Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2015, pp. 1–5.

[73] M. Zohourian, G. Enzner, and R. Martin, "On the use of beamforming spproaches for binaural speaker localization," in *Proc. ITG Conference on Speech Communication*, Paderborn, Germany, Oct. 2016, pp. 1–5.

[74] D. Marquardt and S. Doclo, "Noise power spectral density estimation for binaural noise reduction exploiting direction of arrival estimates," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2017, pp. 234–238.

[75] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, Mar. 2019.

[76] W. Kellermann, "A self-steering digital microphone array," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, Apr. 1991, pp. 3581–3584.

[77] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[78] L. Ehrenberg, S. Gannot, A. Leshem, and E. Zehavi, "Sensitivity analysis of MVDR and MPDR beamformers," in *Proc. IEEE Convention of Electrical and Electronics Engineers in Israel*, Eilat, Israel, Dec. 2010, pp. 416–420.

[79] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the MVDR beamformer in room acoustics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 158–170, Jan. 2010.

[80] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[81] L. J. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[82] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, pp. 2677–2684, Oct. 1999.

[83] S. S. Haykin, *Adaptive filter theory*. Pearson Education, 2005.

[84] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1529–1539, Jul. 2007.

[85] 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, no. 6, pp. 1071–1086, Aug. 2009.

[86] A. Krueger, E. Warsitz, and R. Haeb-Umbach, "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 206–219, Jan. 2011.

[87] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, Sep. 2004.

[88] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, May 2009.

[89] R. Serizel, M. Moonen, B. van Dijk, and J. Wouters, "Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 785–799, Apr. 2014.

[90] 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, Apr. 2015, pp. 544–548.

[91] R. Giri, B. D. Rao, F. Mustiere, and T. Zhang, "Dynamic relative impulse response estimation using structured sparse Bayesian learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 514–518.

[92] 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, Mar. 2017, pp. 11–15.

[93] S. Markovich-Golan, S. Gannot, and W. Kellermann, "Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function," in *Proc. European Signal Processing Conference (EUSIPCO)*, Rome, Italy, Sep. 2018, pp. 2499–2503.

[94] B. Kollmeier, J. Peissig, and V. Hohmann, "Binaural noise-reduction hearing-aid scheme with real-time processing in the frequency-domain," *Scandinavian Audiology. Supplementum*, vol. 38, pp. 28–38, Jan. 1993.

[95] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–14, Dec. 2006.

[96] G. Grimm, V. Hohmann, and B. Kollmeier, "Increase and subjective evaluation of feedback stability in hearing aids by a binaural coherence-based noise reduction scheme," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1408–1419, Sep. 2009.

[97] H. Kamkar-Parsi and M. Bouchard, "Improved noise power spectrum density estimation for binaural hearing aids operating in a diffuse noise field environment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 521–533, May 2009.

[98] K. Reindl, Y. Zheng, and W. Kellermann, "Speech enhancement for binaural hearing aids based on blind source separation," in *Proc. International Symposium on Communications, Control and Signal Processing (ISCCSP)*, Limassol, Cyprus, Mar. 2010, pp. 1–6.

[99] M. Jeub, C. Nelke, H. Krüger, C. Beaugeant, and P. Vary, "Robust dual-channel noise power spectral density estimation," in *Proc. European Signal Processing Conference (EUSIPCO)*, Barcelona, Spain, Aug. 2011, pp. 2304–2308.

[100] H. Kamkar-Parsi and M. Bouchard, "Instantaneous binaural target PSD estimation for hearing aid noise reduction in complex acoustic environments," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 4, pp. 1141–1154, Apr. 2011.

[101] R. Baumgärtel, M. Krawczyk-Becker, D. Marquardt, C. Völker, H. Hu, T. Herzke, G. Coleman, K. Adiloglu, S. M. A. Ernst, T. Gerkmann, S. Doclo, B. Kollmeier, V. Hohmann, and M. Dietz, "Comparing binaural signal processing strategies I: Instrumental evaluation.," *Trends in Hearing*, vol. 19, pp. 1–16, Dec. 2015.

[102] G. Enzner, M. Azarpour, and J. Siska, "Cue-preserving MMSE filter for binaural speech enhancement," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Xi'an, China, Sep. 2016, pp. 1–5.

[103] R. Aichner, H. Buchner, M. Zourub, and W. Kellermann, "Multi-channel source separation preserving spatial information," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu HI, USA, Apr. 2007, pp. 5–8.

[104] E. Hadad, S. Gannot, and S. Doclo, "Binaural linearly constrained minimum variance beamformer for hearing aid applications," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany, Sep. 2012, pp. 4–6.

[105] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2449–2464, Dec. 2015.

[106] D. Marquardt, V. Hohmann, and S. Doclo, "Interaural coherence preservation in multi-channel Wiener filtering based noise reduction for binaural hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2162–2176, Dec. 2015.

[107] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Relaxed binaural LCMV beamforming," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 137–152, Jan. 2017.

[108] H. As'ad, M. Bouchard, and H. Kamkar-Parsi, "A robust target linearly constrained minimum variance beamformer with spatial cues preservation for binaural hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1549–1563, Oct. 2019.

[109] D. Marquardt, E. Hadad, S. Gannot, and S. Doclo, "Theoretical analysis of linearly constrained multi-channel Wiener filtering algorithms for combined noise reduction and binaural cue preservation in binaural hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2384–2397, Dec. 2015.

[110] W. Pu, J. Xiao, T. Zhang, and Z.-Q. Luo, "A penalized inequality-constrained minimum variance beamformer with applications in hearing aids," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2017, pp. 175–179.

[111] J. Thiemann, M. Müller, D. Marquardt, S. Doclo, and S. van de Par, "Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, Feb. 2016.

[112] A. I. Koutrouvelis, J. Jensen, M. Guo, R. C. Hendriks, and R. Heusdens, "Binaural speech enhancement with spatial cue preservation utilising simultaneous masking," in *Proc. European Signal Processing Conference (EUSIPCO)*, Kos, Greece, Aug. 2017, pp. 628–632.

[113] N. Gößling, D. Marquardt, and S. Doclo, "Perceptual evaluation of binaural MVDR-based algorithms to preserve the interaural coherence of diffuse noise fields," *Trends in Hearing*, vol. 24, pp. 1–18, Apr. 2020.

[114] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *Proc. IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, Ghent, Belgium, Nov. 2011, pp. 1–6.

[115] S. Doclo, M. Moonen, T. van den Bogaert, and J. Wouters, "Reduced-bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 38–51, Jan. 2009.

[116] A. Bertrand and M. Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks–Part I: Sequential node updating," *IEEE Transactions on Signal Processing*, vol. 58, pp. 5277–5291, Oct. 2010.

[117] ——, "Distributed node-specific LCMV beamforming in wireless sensor networks," *IEEE Transactions on Speech and Audio Processing*, vol. 60, no. 1, pp. 233–246, Jan. 2012.

[118] S. Markovich-Golan, S. Gannot, and I. Cohen, "Low-complexity addition or removal of sensors/constraints in LCMV beamformers," *IEEE Transactions on Signal Processing*, vol. 60, no. 3, pp. 1205–1214, Mar. 2012.

[119] ——, "Performance of the SDW-MWF with randomly located microphones in a reverberant enclosure," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1513–1523, Jul. 2013.

[120] M. Souden, K. Kinoshita, M. Delcroix, and T. Nakatani, "Location feature integration for clustering-based speech separation in distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 354–367, Feb. 2014.

[121] Y. Zeng and R. C. Hendriks, "Distributed delay and sum beamformer for speech enhancement via randomized gossip," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 260–273, Jan. 2014.

[122] S. Markovich-Golan, A. Bertrand, M. Moonen, and S. Gannot, "Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks," *Signal Processing*, vol. 107, pp. 4–20, Feb. 2015.

[123] A. I. Koutrouvelis, T. W. Sherson, R. Heusdens, and R. C. Hendriks, "A low-cost robust distributed linearly constrained beamformer for wireless acoustic sensor networks with arbitrary topology," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1434–1448, Aug. 2018.

[124] J. Zhang, A. I. Koutrouvelis, R. Heusdens, and R. C. Hendriks, "Distributed rate-constrained LCMV beamforming," *IEEE Signal Processing Letters*, vol. 26, no. 5, pp. 675–679, May 2019.

[125] J. Zhang, R. Heusdens, and R. C. Hendriks, "Relative acoustic transfer function estimation in wireless acoustic sensor networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1507–1519, Oct. 2019.

[126] S. Wehr, I. Kozintsev, R. Lienhart, and W. Kellermann, "Synchronization of acoustic sensors for distributed ad-hoc audio networks and its use for blind source separation," in *Proc. IEEE International Symposium on Multimedia Software Engineering*, Miami, FL, USA, Dec. 2004, pp. 18–25.

[127] S. Markovich-Golan, S. Gannot, and I. Cohen, "Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany, Sep. 2012, pp. 1–4.

[128] D. Cherkassky and S. Gannot, "Blind synchronization in wireless sensor networks with application to speech enhancement," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, France, Sep. 2014, pp. 184–188.

[129] Y. Zeng, R. C. Hendriks, and N. Gaubitch, "On clock synchronization for multi-microphone speech processing in wireless acoustic sensor networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 231–235.

[130] D. Cherkassky, S. Markovich-Golan, and S. Gannot, "Performance analysis of MVDR beamformer in WASN with sampling rate offsets and blind synchronization," in *Proc. European Signal Processing Conference (EUSIPCO)*, Nice, France, Aug. 2015, pp. 245–249.

[131] J. Schmalenstroeer, P. Jebramcik, and R. Haeb-Umbach, "A combined hardware-software approach for acoustic sensor network synchronization," *Signal Processing*, vol. 107, pp. 171–184, Feb. 2015.

[132] L. Wang and S. Doclo, "Correlation maximization based sampling rate offset estimation for distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 571–582, Mar. 2016.

[133] M. H. Bahari, A. Bertrand, and M. Moonen, "Blind sampling rate offset estimation for wireless acoustic sensor networks through weighted least-squares coherence drift estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 674–686, Mar. 2017.

[134] D. Cherkassky and S. Gannot, "Blind synchronization in wireless acoustic sensor networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 651–661, Mar. 2017.

[135] A. Boothroyd and F. Iglehart, "Experiments with classroom FM amplification," *Ear and Hearing*, vol. 19, no. 3, pp. 202–217, Jun. 1998.

[136] E. C. Schafer, K. Sanders, D. Bryant, K. Keeney, and N. Baldus, "Effects of voice priority in FM systems for children with hearing aids," *Journal of Educational Audiology*, vol. 19, pp. 12–24, 2013.

[137] A. Boothroyd, "Hearing aid accessories for adults: The remote FM microphone," *Ear and Hearing*, vol. 25, no. 1, pp. 22–33, Feb. 2004.

[138] L. Thibodeau, "Comparison of speech recognition with adaptive digital and FM remote microphone hearing assistance technology by listeners who use hearing aids," *American Journal of Audiology*, vol. 23, no. 2, Jun. 2014.

[139] G. De Ceulaer, J. Bestel, H. E. Mülder, F. Goldbeck, S. P. J. de Varebeke, and P. J. Govaerts, "Speech understanding in noise with the Roger Pen, Naida CI Q70 processor, and integrated Roger 17 receiver in a multi-talker network," *European Archives of Oto-Rhino-Laryngology*, vol. 273, pp. 1107–1114, May 2016.

[140] E. A. Lopez, O. A. Costa, and D. V. Ferrari, "Development and technical validation of the mobile based assistive listening system: A smartphone-based remote microphone," *American Journal of Audiology*, vol. 25, no. 3S, pp. 288–294, Oct. 2016.

[141] C. R. Benítez-Barrera, G. P. Angley, and A. M. Tharpe, "Remote microphone system use at home: Impact on caregiver talk," *Journal of Speech, Language, and Hearing Research*, vol. 61, no. 2, pp. 399–409, Feb. 2018.

[142] Y.-C. Lin, Y.-H. Lai, H.-W. Chang, Y. Tsao, Y. Chang, and R. Y. Chang, "SmartHear: A smartphone-based remote microphone hearing assistive system using wireless technologies," *IEEE Systems Journal*, vol. 12, no. 1, pp. 20–29, Mar. 2018.

[143] K. C. Wagener, M. Vormann, M. Latzel, and H. E. Mülder, "Effect of hearing aid directionality and remote microphone on speech intelligibility in complex listening situations," *Trends in Hearing*, vol. 22, no. 1-12, Oct. 2018.

[144] T. Wesarg, Y. Stelzig, D. Hilgert-Becker, B. Kathage, K. Wiebe, A. Aschendorff, S. Arndt, and I. Speck, "Application of digital remote wireless microphone technology in single-sided deaf cochlear implant recipients," *Journal of the American Academy of Audiology*, vol. 31, no. 4, pp. 246–256, Apr. 2020.

[145] J. M. Kates, K. H. Arehart, M. R. Kumar, and K. Sommerfeldt, "Externalization of remote microphone signals using a structural binaural model of the head and pinna," *The Journal of the Acoustical Society of America*, vol. 143, no. 5, pp. 2666–2677, May 2018.

[146] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, "Sound source localization for hearing aid applications using wireless microphones," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 16–20.

[147] D. Yee, H. Kamkar-Parsi, H. Puder, and R. Martin, "A speech enhancement system using binaural hearing aids and an external microphone," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 246–250.

[148] D. Yee, H. Kamkar-Parsi, R. Martin, and H. Puder, "A model-based placement strategy for a nearby external microphone for speech enhancement in hearing aids," in *Proc. ITG Conference on Speech Communication*, Paderborn, Germany, Oct. 2016, pp. 140–144.

[149] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, "Informed sound source localization using relative transfer functions for hearing aid applications," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 611–623, Mar. 2017.

[150] D. Yee, H. Kamkar-Parsi, R. Martin, and H. Puder, "A noise reduction postfilter for binaurally-linked single-microphone hearing aids utilizing a nearby external microphone," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 5–18, Jan. 2018.

[151] R. Ali, T. van Waterschoot, and M. Moonen, "Generalised sidelobe canceller for noise reduction in hearing devices using an external microphone," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 521–525.

[152] ——, "Completing the RTF vector for an MVDR beamformer as applied to a local microphone array and an external microphone," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, Sep. 2018, pp. 211–215.

[153] M. Farmani, M. S. Pedersen, and J. Jensen, "Sound source localization for hearing aid applications using wireless microphones," in *Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, Sheffield, UK, Jul. 2018, pp. 455–459.

[154] D. Yee, "Improved noise reduction in binaural hearing aids using and external microphone," PhD thesis, Ruhr-Universität Bochum, Bochum, Germany, Jun. 2018.

[155] J. M. Kates and K. H. Arehart, "Integrating a remote microphone with hearing-aid processing," *The Journal of the Acoustical Society of America*, vol. 145, no. 6, pp. 3551–3566, Jun. 2019.

[156] N. Gößling, E. Hadad, S. Gannot, and S. Doclo, "Binaural LCMV beamforming with partial noise estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, in press, 2020.

[157] N. Gößling, D. Marquardt, and S. Doclo, "Performance analysis of the extended binaural MVDR beamformer with partial noise estimation in a homogeneous noise field," in *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, San Francisco, USA, Mar. 2017, pp. 1–5.

[158] ——, "Performance analysis of the extended binaural MVDR beamformer with partial noise estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, manuscript accepted with minor revisions, 2020.

[159] 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, Sep. 2018, pp. 146–150.

[160] ——, "RTF-based binaural MVDR beamformer exploiting an external microphone in a diffuse noise field," in *Proc. ITG Conference on Speech Communication*, Oldenburg, Germany, Oct. 2018, pp. 106–110.

[161] ——, "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, pp. 416–420.

[162] 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, NY, USA, Oct. 2019, pp. 368–372.

[163] N. Gößling, D. Marquardt, I. Merks, T. Zhang, and S. Doclo, "Optimal binaural LCMV beamforming in complex acoustic scenarios: Theoretical and practical insights," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, Sep. 2018, pp. 381–385.

[164] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Comparison of binaural multichannel Wiener filters with binaural cue preservation of the interferer," in *Proc. IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, Eilat, Israel, Nov. 2016, pp. 1–5.

[165] M. Tohyama and A. Suzuki, "Interaural cross-correlation coefficients in stereo-reproduced sound fields," *The Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 780–786, Feb. 1989.

[166] A. Walther and C. Faller, "Interaural correlation discrimination from diffuse field reference correlations," *The Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1496–1502, Mar. 2013.

[167] I. M. Lindevald and A. H. Benade, "Two-ear correlation in the statistical sound fields of rooms," *The Journal of the Acoustical Society of America*, vol. 80, no. 2, pp. 661–664, Aug. 1986.

[168] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2159–2169, Sep. 2011.

[169] M. Taseska and E. A. P. Habets, "Spotforming using distributed microphone arrays," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2013, pp. 1–4.

[170] ——, "Informed spatial filtering for sound extraction using distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 7, pp. 1195–1207, Jul. 2014.

[171] C. Horton, R. Srinivasan, and M. D'Zmura, "Envelope responses in single-trial EEG indicate attended speaker in a cocktail party," *Journal of Neural Engineering*, vol. 11, no. 4, p. 046 015, Jun. 2014.

[172] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, Jul. 2015.

[173] A. Aroudi, B. Mirkovic, M. De Vos, and S. Doclo, "Impact of different acoustic components on EEG-based auditory attention decoding in noisy and reverberant conditions," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 4, pp. 652–663, Apr. 2019.

[174] A. Aroudi and S. Doclo, "Cognitive-driven binaural beamforming using EEG-based auditory attention decoding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 862–875, Jan. 2020.

[175] T. de Taillez, B. Kollmeier, and B. Meyer, "Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech," *European Journal of Neuroscience*, vol. 51, no. 5, pp. 1234–1241, Mar. 2020.

[176] Y. Huang and J. Benesty, "Adaptive multi-channel least mean square and Newton algorithms for blind channel identification," *Signal Processing*, vol. 82, pp. 1127–1138, Aug. 2002.

[177] ——, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Transactions on Signal Processing*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[178] Y. Huang, J. Benesty, and J. Chen, "Optimal step size of the adaptive multichannel LMS algorithm for blind SIMO identification," *IEEE Signal Processing Letters*, vol. 12, no. 3, pp. 173–176, Mar. 2005.

[179] M. K. Hasan and P. A. Naylor, "Analyzing effect of noise on LMS-type approaches to blind estimation of SIMO channels: robustness issue," in *Proc. European Signal Processing Conference (EUSIPCO)*, Florence, Italy, Sep. 2006, pp. 1–4.

[180] G. Enzner, I. Merks, and T. Zhang, "Adaptive filter algorithms and misalignment criteria for blind binaural channel identification in hearing-aids," in *Proc. European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, Aug. 2012, pp. 315–319.

[181] M. Hu, D. Sharma, S. Doclo, M. Brookes, and P. A. Naylor, "Blind adaptive SIMO acoustic system identification using a locally optimal step-size," in *Proc. Audio Engineering Society (AES) Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)*, Leuven, Belgium, Feb. 2016, pp. 1–6.

[182] C. Liu, B. C. Wheeler, D. O'Brien Jr., R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," *The Journal of the Acoustical Society of America*, vol. 108, no. 4, pp. 1888–1905, Oct. 2000.

[183] T. May, S. van de Par, and A. Kohlrausch, "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2016–2030, Sep. 2012.

[184] M. Zohourian, G. Enzner, and R. Martin, "Binaural speaker localization integrated into an adaptive beamformer for hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 515–528, Mar. 2018.

[185] R. Varzandeh, K. Adiloglu, S. Doclo, and V. Hohmann, "Exploiting periodicity features for joint detection and DOA estimation of speech sources using convolutional neural networks," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 566–570.

[186] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[187] ITU-R BS.1534-1, *Method for the subjective assessment of intermediate quality level of coding systems*. 2003.

[188] B. R. Kirkwood and J. A. C. Sterne, *Essential medical statistics*. John Wiley & Sons, 2010.

[189] J. E. Greenberg, P. M. Peterson, and P. M. Zurek, "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance," *The Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 3009–3010, Nov. 1993.

[190] T. C. Lawin-Ore and S. Doclo, "Reference microphone selection for MWF-based noise reduction using distributed microphone arrays," in *Proc. ITG Conference on Speech Communication*, Braunschweig, Germany, Sep. 2012, pp. 1–4.

LIST OF PUBLICATIONS

Peer-reviewed Journal Papers

[J1] **N. Gößling**, D. Marquardt, and S. Doclo, "Performance analysis of the extended binaural MVDR beamformer with partial noise estimation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, manuscript accepted with minor revisions, 2020.

[J2] **N. Gößling**, E. Hadad, S. Gannot, and S. Doclo, "Binaural LCMV beamforming with partial noise estimation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, in press, 2020.

[J3] **N. Gößling**, D. Marquardt, and S. Doclo, "Perceptual evaluation of binaural MVDR-based algorithms to preserve the interaural coherence of diffuse noise fields," *Trends in Hearing*, vol. 24, pp. 1–18, Apr. 2020.

Peer-reviewed Conference Papers

[C1] **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, pp. 368–372.

[C2] **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, pp. 416–420.

[C3] **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, Oct. 2018, pp. 106–110.

[C4] C. F. Hauth, **N. Gößling**, and T. Brand, "Performance prediction of the binaural MVDR beamformer with partial noise estimation using a binaural speech intelligibility model," in *Proc. ITG Conference on Speech Communication*, Oldenburg, Germany, Oct. 2018, pp. 301–305.

[C5] J. Klug, D. Marquardt, **N. Gößling**, and S. Doclo, "Evaluation of signal-dependent partial noise estimation algorithms for binaural hearing aids," in *Proc. ITG Conference on Speech Communication, Oldenburg*, Germany, Oct. 2018, pp. 236–240.

[C6] **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, Sep. 2018, pp. 146–150. **Best student paper award**

[C7] **N. Gößling**, D. Marquardt, I. Merks, T. Zhang, and S. Doclo, "Optimal binaural LCMV beamforming in complex acoustic scenarios: theoretical and practical insights," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, Sep. 2018, pp. 381–385.

[C8] **N. Gößling**, D. Marquardt, and S. Doclo, "Comparison of RTF estimation methods between a head-mounted binaural hearing device and an external microphone," in *Proc. International Workshop on Challenges in Hearing Assistive Technology (CHAT)*, Stockholm, Sweden, Aug. 2017, pp. 101–106.

[C9] **N. Gößling**, D. Marquardt, and S. Doclo, "Performance analysis of the extended binaural MVDR beamformer with partial noise estimation in a homogeneous noise field," in *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, San Francisco, USA, Mar. 2017, pp. 1–5.

Conference Abstracts

[A1] **N. Gößling** and S. Doclo, "Binaural MVDR beamforming incorporating external microphones in complex acoustic scenarios," *47th Erlanger Kolloquium*, Erlangen, Germany, Feb. 2020.

[A2] **N. Gößling** and S. Doclo, "RTF-steered binaural MVDR beamforming incorporating an external microphone for dynamic acoustic scenarios," *International Congress on Acoustics (ICA)*, Aachen, Germany, Sep. 2019.

[A3] W. Middelberg, **N. Gößling**, and S. Doclo, "Real-time evaluation of an RTF-steered binaural MVDR beamformer incorporating an external microphone," *Deutsche Jahrestagung für Akustik (DAGA)*, Rostock, Germany, Mar. 2019.

[A4] **N. Gößling** and S. Doclo, "Comparison of binaural MVDR-based beamforming algorithms using an external microphone," *11th Speech in Noise Workshop (SPiN)*, Ghent, Belgium, Jan. 2019.

[A5] J. Klug, **N. Gößling**, and S. Doclo, "Subjective evaluation of signal-dependent partial noise estimation algorithms for binaural hearing aids," *11th Speech in Noise Workshop (SPiN)*, Ghent, Belgium, Jan. 2019.

[A6] C. F. Hauth, **N. Gößling**, and S. Doclo, T. Brand, "Performance prediction of the binaural MVDR beamformer with partial noise estimation using a binaural speech intelligibility model," *11th Speech in Noise Workshop (SPiN)*, Ghent, Belgium, Jan. 2019.

[A7] **N. Gößling** and S. Doclo, "RTF-based binaural MVDR beamformer exploiting an external microphone for dynamic acoustic scenarios," *IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, Eilat, Israel, Dec. 2018.

[A8] **N. Gößling** and S. Doclo, "Comparison of binaural MVDR-based beamforming algorithms using an external microphone," *International Hearing Aid Research Conference (IHCON)*, Lake Tahoe, USA, Aug. 2018.

[A9] **N. Gößling**, T. Wendt, and S. D. Ewert, "Perceptually-plausible simulation of diffuse reflections and sound scattering," *Deutsche Jahrestagung für Akustik (DAGA)*, Aachen, Germany, Mar. 2016.