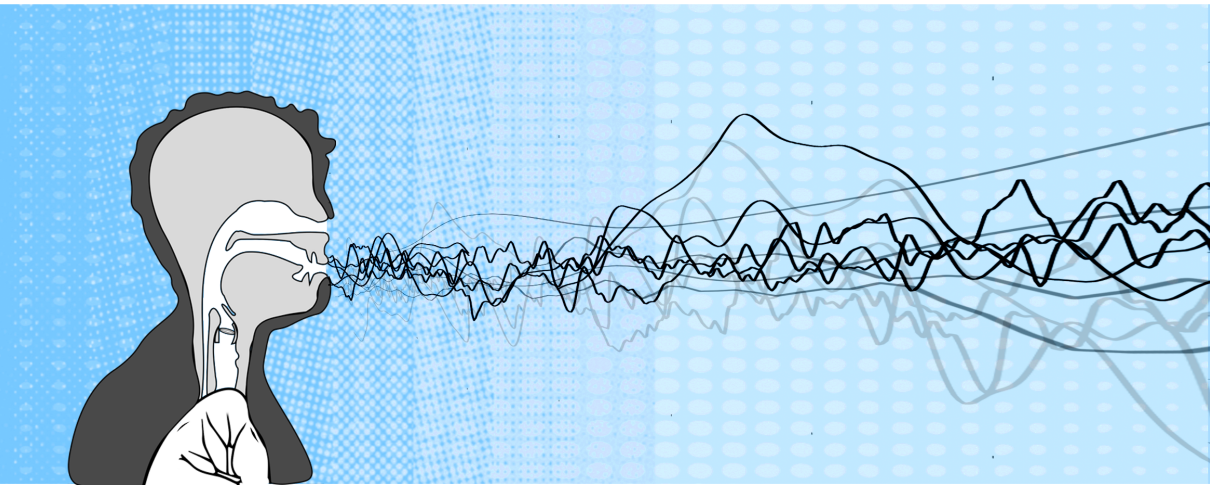
KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science
Department of Electrical Engineering

Multi-microphone speech enhancement

An integration of a priori and data-dependent spatial information



Randall Ali

Supervisors:
Prof. dr. ir. Marc Moonen
Prof. dr. ir. Toon van Waterschoot

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

November 2020

Multi-microphone speech enhancement

An integration of a priori and data-dependent spatial information

Randall ALI

Examination committee:
Prof. dr. ir. Herman Neuckermans, chairman
Prof. dr. ir. Marc Moonen, supervisor
Prof. dr. ir. Toon van Waterschoot, co-supervisor
Prof. dr. ir. Jan Wouters
Prof. dr. ir. Alexander Bertrand
Prof. Mads Græsbøll Christensen, M.Sc.E, Ph.D.
(Aalborg University, Denmark)
Priv.-Doz. Dr.-Ing. habil. Gerald Enzner
(Ruhr-Universität Bochum, Germany)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

November 2020

© 2020 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Randall Ali, Kasteelpark Arenberg 10, bus 2446, B-3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

Preface

This dissertation has been largely motivated by the prospect of further improving the technology of assistive hearing devices such as hearing aids and cochlear implants. With the increasing access to microphone signals in our sonic landscape, it is apparent that the acoustic signal processing technology in particular has the potential for additional improvement. The research set forth is focused on a specific aspect of acoustic signal processing technology: the task of enhancing speech after it has been corrupted by other acoustic disturbances and captured by multiple microphones, which is pertinent not only for assistive hearing devices, but also for a number of other applications. It is hoped that this dissertation will be useful to scientists and engineers who continue to develop the technology of assistive hearing devices, which will in turn eventually contribute to the well-being of those who have had to endure with the physical and mental challenges of hearing loss.

This research would not have been possible without the continued encouragement and assistance from a strong support network, to whom I would like to express my immense gratitude. Firstly to Marc Moonen and Toon van Waterschoot for their indispensable guidance, openness to discuss just about anything, which has led to countless insightful conversations, and for fostering a creative research atmosphere over the years. To the examination committee, Jan Wouters, Alexander Bertrand, Mads Græsbøll Christensen, and Gerald Enzner, for their extensive review and encouraging suggestions, which have shaped the final version of this manuscript. I am also thankful to Bas Van Dijk and Adam Hersbach from Cochlear, who steered the initial direction of this research. To all of the colleagues at ESAT, especially thanks to Giuliano, Niccolo, Maja, Elisa, Amin, and Thomas for assisting in getting the audio laboratory into a functional state from the ground up and for assistance with several measurements. To my family, as who knows where I would be without the branches of my family tree. A special thanks to María for the unconditional support, motivation, inspiration, and never-ending conversations on the ethics of technology. To Blinks, Ras, Ana, Oreste, Daryna (and all others I have forgotten only on paper) for the

stage dives that keep us alive and to impedance match the undulating rhythm of another culture. In short, thank you to everyone, everywhere, for everything in life.

Randall A. Ali,
Brussels, November 2020

Abstract

A speech signal captured by multiple microphones is often subject to a reduced intelligibility and quality due to the presence of noise and room acoustic interferences. Multi-microphone speech enhancement systems therefore aim at the suppression or cancellation of such undesired signals without substantial distortion of the speech signal. A fundamental aspect to the design of several multi-microphone speech enhancement systems is that of the spatial information which relates each microphone signal to the desired speech source. This spatial information is unknown in practice and has to be somehow estimated. Under certain conditions, however, the estimated spatial information can be inaccurate, which subsequently degrades the performance of a multi-microphone speech enhancement system.

This doctoral dissertation is focused on the development and evaluation of acoustic signal processing algorithms in order to address this issue. Specifically, as opposed to conventional means of estimating spatial information using only a priori knowledge or only observable microphone data, an integrated approach is pursued where both a priori and data-dependent spatial information are explicitly used. An initial investigation into such an a approach is firstly considered for the case of a microphone array from a confidence-based perspective, where a confidence metric is used to optimally combine a priori and data-dependent spatial information. The remainder of the dissertation is then dedicated to the study of a microphone array that has access to one or more external microphones. For this microphone configuration, a geometrically-based integration is investigated for the tasks of noise reduction, binaural speech enhancement, and speech dereverberation, where a priori spatial information is used for the microphone array(s) and data-dependent spatial information estimated from the observable microphone data is used for the external microphone(s). A final conception of an integrated approach is then explored for this microphone configuration by merging the confidence-based and geometrically-based integration techniques.

The mathematical framework for the integrated approach as applied to the different microphone configurations is presented, along with experimental evaluation using recorded audio data from various acoustic environments. The results have shown that by following an integrated approach, more spatially robust speech enhancement algorithms can be designed as opposed to relying solely on a priori spatial information or only data-dependent spatial information. Furthermore, the advantage of using a priori spatial knowledge was demonstrated as it served to provide contingency spatial information in cases when the data-dependent spatial information was deemed to be inaccurate. A number of experiments involving an assistive hearing device linked with external microphones have also shown that the proposed speech enhancement algorithms can improve speech intelligibility in comparison to only using the assistive hearing device or only listening to an external microphone signal.

.

Beknopte samenvatting

Een spraaksignaal dat door meerdere microfoons wordt opgevangen, is vaak onderhevig aan een verminderde verstaanbaarheid en kwaliteit vanwege de aanwezigheid van ruis en akoestische interferenties in de kamer. Meerkanaals spraakverbeteringssystemen richten zich daarom op het onderdrukken of verwijderen van dergelijke ongewenste signalen zonder het spraaksignaal aanzienlijk te vervormen. Een fundamenteel aspect van het ontwerp van verscheidene meerkanaals spraakverbeteringssystemen is de ruimtelijke informatie, dewelke ieder microfoonsignaal relateert aan de gewenste spraakbron. Deze ruimtelijke informatie is in de praktijk onbekend en moet op de een of andere manier worden geschat. Onder bepaalde omstandigheden kan de geschatte ruimtelijke informatie echter onnauwkeurig zijn, wat vervolgens de prestatie van een meerkanaals spraakverbeteringssysteem verslechtert.

Dit proefschrift is gericht op de ontwikkeling en evaluatie van algoritmen voor akoestische signaalverwerking om dit probleem aan te pakken. In het bijzonder wordt, in tegenstelling tot conventionele methoden om ruimtelijke informatie te schatten met alleen a priori kennis of alleen waarneembare microfoondata, een geïntegreerde benadering nagestreefd waarbij zowel a priori als data-afhankelijke ruimtelijke informatie expliciet wordt gebruikt. In een eerste onderzoek naar dergelijke benadering wordt een microfoonrooster vanuit een op vertrouwen gebaseerd perspectief bekeken, waarbij een betrouwbaarheidsmetriek wordt gebruikt om a priori en data-afhankelijke ruimtelijke informatie optimaal te combineren. De rest van het proefschrift is dan gewijd aan de studie van een microfoonrooster die toegang heeft tot een of meerdere externe microfoons. In deze microfoonconfiguratie wordt gezocht naar een geometrisch gebaseerde integratie voor de taken van ruisonderdrukking, binaurale spraakverbetering en dereverberatie van spraak, waarbij a priori ruimtelijke informatie wordt gebruikt voor het microfoonrooster(s) en data-afhankelijke ruimtelijke informatie, geschat op basis van de waarneembare microfoondata, wordt gebruikt voor de externe microfoon(s). Een laatste conceptie van een geïntegreerde benadering wordt dan bekomen voor deze microfoonconfiguratie door een combinatie van deze op

vertrouwen gebaseerde en geometrisch gebaseerde integratietechnieken.

Het wiskundige raamwerk voor de geïntegreerde benadering toegepast op de verschillende microfoonconfiguraties wordt gepresenteerd, samen met experimentele evaluatie gebruik makend van opgenomen audiogegevens uit verschillende akoestische omgevingen. De resultaten hebben aangetoond dat door het volgen van een geïntegreerde benadering, meer ruimtelijk robuuste spraakverbeteringsalgoritmen kunnen worden ontworpen in plaats van alleen te vertrouwen op a priori ruimtelijke informatie of alleen gegevensafhankelijke ruimtelijke informatie. Bovendien werd het voordeel van het gebruik van a priori ruimtelijke kennis aangetoond, aangezien het diende om onvoorziene ruimtelijke informatie te verschaffen in gevallen waarin de data-afhankelijke ruimtelijke informatie onnauwkeurig werd geacht. Een aantal experimenten met een gehoorapparaat gekoppeld met externe microfoons, hebben ook aangetoond dat de voorgestelde spraakverbeteringsalgoritmen de spraakverstaanbaarheid kunnen verbeteren in vergelijking met het alleen gebruiken van het gehoorapparaat of alleen luisteren naar een extern microfoonsignaal.

List of Abbreviations

AG	Array Gain
AIR	acoustic impulse response
APC	a priori constraint
ASR	automatic speech recognition
ATF	acoustic transfer function
BMVDR	binaural minimum variance distortionless response
CD	Cepstral distance
CI	cochlear implant
DANSE	distributed adaptive node-specific signal estimation
DAS	delay and sum
DDC	data-dependent constraint
DI	directivity index
DFT	discrete Fourier transform
EVD	eigenvalue decomposition
FM	frequency modulation
fwSegSNR	frequency weighted segmental signal-to-noise ratio
GEVD	generalised eigenvalue decomposition
GSC	generalised sidelobe canceller
HA	hearing aid

IDFT	inverse discrete Fourier transform
ILD	interaural level difference
ITD	interaural time difference
LCMV	linearly constrained minimum variance
LMA	local microphone array
LMS	least mean squares
LTI	linear time-invariant
MFB	matched filter beamformer
ML	maximum likelihood
MMSE	minimum mean square error
MPDR	minimum power distortionless response
MVDR	minimum variance distortionless response
MWF	multi-channel Wiener filter
NFMI	near field magnetic interference
NLMS	normalised least mean squares
OSI	open systems interconnection model
PDF	probability density function
PSD	power spectral density
PWT	pre-whitened-transformed
QCQP	quadratically constrained quadratic program
RF	radio frequency
RIR	room impulse response
RTF	relative transfer function
SDB	superdirective beamformer
SDW-MWF	speech distortion weighted multichannel Wiener filer
SI-SNR	speech intelligibility weighted signal-to-noise ratio

SNR	signal-to-noise ratio
SPP	speech presence probability
STFT	short-time Fourier transform
SRMR	speech-to-reverberation modulation energy ratio
STOI	short-time objective intelligibility
TDOA	time difference of arrival
VAD	voice activity detector
WASN	wireless acoustic sensor network
WOLA	weighted overlap and add
WNG	white noise gain
XM	external microphone

List of Symbols

ω	angular frequency	$\mathrm{rad} \cdot \mathrm{s}^{-1}$
ρ_m	fluid density	$\mathrm{kg} \cdot \mathrm{m}^{-3}$
c	speed of sound	$\mathrm{m} \cdot \mathrm{s}^{-1}$
f	frequency	Hz
f_{sch}	Schroeder frequency	Hz
f_s	sampling frequency	Hz
k	wavenumber	m^{-1}
L	total number of frames of an STFT	
P	number of samples in a time frame of the STFT	
R	number of samples to define hop-size for an STFT	
r	distance from sound source	m
$\mathbf{r}$	vector of 3-D cartesian coordinates	
r_c	critical distance	m
V	Volume	m^3
k	frequency bin of the STFT	
l	time frame of the STFT	
M	total number of microphones	
$\mathrm{M_a}$	number of microphones in the local microphone array	

M_e	number of external microphones	
n	discrete-time index	
S_T	total absorption coefficient	m^2
T_{60}	T-60 reverberation time	s

Continuous-time and frequency domain variables

ζ_i	damping constant for i-th mode	s^{-1}
G	Green's function between sound source and receiver	
h	impulse response between the sound source and receiver	
P	acoustic pressure	$Pa \cdot Hz^{-1}$
Ψ	frequency-wavenumber response, or beam pattern	
p	acoustic pressure	Pa

Discrete-time domain variables

g	acoustic impulse response
n	noise-only component of observed microphone signal
s	desired speech signal
x	speech-only component of observed microphone signal
y	observed microhone signal

Short time Fourier transform domain variables

$\widetilde{\mathbf{C}}_\mathbf{a}$	blocking matrix for local microphone array
$\mathbf{d}$	direct component of relative transfer function vector
$\mathbf{e}_1$	vector with all zeros except for a one as the first element
$\hat{\epsilon}$	maximum tolerable speech distortion for a data-dependent constraint
$\widetilde{\epsilon}$	maximum tolerable speech distortion for an a priori constraint
$\tilde{\mathbf{f}}_\mathbf{a}$	fixed beamformer for local microphone array

F	confidence metric
$\mathbf{g}$	acoustic transfer function vector
$\mathbf{h}$	relative transfer function vector
$\mathbf{L}$	spatial pre-whitening operation defined from a transformed noise-only correlation matrix (lower triangular matrix)
$\mathbf{n}$	vector of noise-only components from the observed microhone signals
$\mathbf{R_{nn}}$	noise-only spatial correlation matrix
$\mathbf{\Gamma_{nn}}$, $\mathbf{\Gamma}$	spatial coherence matrix
$\mathring{\mathbf{\Gamma}}_{\mathbf{nn}}$	spatial coherence matrix in a spherical diffuse field
$\mathbf{R}_{\mathbf{nn}}^{1/2}$	spatial pre-whitening operation from a noise-only correlation matrix (lower triangular matrix)
$\mathbf{R_{xx}}$	speech-only spatial correlation matrix
$\mathbf{R_{xx,r1}}$	rank-1 speech-only spatial correlation matrix
$\mathbf{R_{yy}}$, $\mathbf{\Phi_y}$	speech-plus-noise spatial correlation matrix
$\mathbf{\Gamma}^{1/2}$	spatial pre-whitening operation from a spatial coherence (lower triangular matrix)
$\sigma^2_{\mathrm{s_r}}$	power spectral density of the desired speech signal in a reference microphone
s	desired speech signal
s_r	desired speech signal in a reference microphone
$\mathbf{w}$	vector of complex-valued filters
$\mathbf{x}$	vector of speech-only components from the observed microhone signals
Φ_{s}	power spectral density of the direct component of the desired speech signal in a reference microphone
$\mathbf{\Phi_{x_d}}$	direct component speech-only correlation matrix
$\mathbf{x_d}$	vector of direct components of the desired speech signal from the observed microphone signals

Φ_{r}	power spectral density of the reverberant component of an observed microphone signal
$\mathbf{\Phi_{x_r}}$	reverberant-only spatial correlation matrix
$\mathbf{x_r}$	vector of reverberant components from the observed microphone signals
$\mathbf{y}$	vector of observed microhone signals
$\widetilde{\mathbf{\Upsilon}}$	Transformation matrix involving a blocking matrix and fixed beamformer
μ	parameter for trading off between noise reduction and speech distortion

Mathematical functions and other general notation

a (lower case)	scalar
a (bold lower case)	vector
A (bold upper case)	matrix
$\mathbf{I}$	identity matrix
$\mathbf{I}_\vartheta$	$\vartheta \times \vartheta$ identity matrix
$\mathbf{0}$	zero vector or matrix
$\mathbf{0}_{m\times n}$	zero vector or matrix with m-rows and n-columns
$\frac{\partial}{\partial x}$	partial derivative (first) w.r.t. x
$\frac{\partial^2}{\partial x^2}$	partial derivative (second) w.r.t. x
$\mathscr{D}$	Dual function
$\mathbb{C}$	the set of complex numbers
$\mathbb{R}$	the set of real numbers
j	imaginary unit, $\sqrt{-1}$
$\angle\{.\}$	argument (complex numbers)
$\approx$	approximately equal to
$\forall$	for all

$\gg$	much greater than
$\in$	belongs to
$\ll$	much less than
$\mathbb{E}\{.\}$	Expectation operator
$\rightarrow$	approaches
$\underset{\mathbf{a}}{\text{minimise}}\, f(\mathbf{a})$	minimise the function $f(\mathbf{a})$ over $\mathbf{a}$
$\{.\}^*$	complex conjugate
$\{.\}^{-1}$	matrix inverse
$\{.\}^H$	matrix Hermitian transpose
$\{.\}^T$	matrix transpose
$\|\|.\|\|$	Euclidean norm
$\|\|.\|\|_F$	Frobenius norm
blkdiag$\{\mathbf{A}_1, \ldots, \mathbf{A}_N\}$	block diagonal matrix with $\mathbf{A}_1, \ldots, \mathbf{A}_N$ on the diagonal
diag$\{a_1, \ldots, a_N\}$	diagonal matrix with $a_1, \ldots, a_N$ on the diagonal
ln(.)	natural logarithm
$\log_{10}(.)$	base-10 logarithm
sinc(.)	sinc function
trace$\{.\}$	trace operator
$\{.\}_{\mathbf{a}}$	variable pertaining to the local microphone array
$\{.\}_{\mathbf{e}}$	variable pertaining to external microphones
$\{.\}_{\mathrm{L}}$	variable pertaining to a left-ear hearing aid
$\{.\}_{\mathrm{R}}$	variable pertaining to a right-ear hearing aid
$\{\hat{.}\}$	data-dependent estimate
$\{\tilde{.}\}$	variable based on a priori knowledge
$\{\underline{.}\}$	variable in the pre-whitened-transformed domain

$\{\underline{\underline{\cdot}}\}$	variable in the pre-whitened-transformed domain with a dimension of the number of external microphones plus one
$\{\breve{\cdot}\}$	variable that has been transformed and partially pre-whitened
$\{\underline{\breve{\cdot}}\}$	variable that has been transformed and partially pre-whitened with a dimension of the number of external microphones plus one

Contents

Abstract iii

Beknopte samenvatting v

List of Abbreviations vii

List of Symbols xvi

Contents xvii

List of Figures xxiii

List of Tables xxxv

1 Introduction 1

1.1 Speech enhancement 1

1.1.1 Problem description 2

1.1.2 Existing solution architecture 4

1.2 Research Questions 6

1.2.1 Integrating a priori and data-dependent spatial information 7

1.2.2 The use of a priori spatial knowledge for a local microphone array and external microphones 8

1.3 Applications and societal relevance 10
1.4 Dissertation overview 12

2 From physical acoustics to speech enhancement 19
2.1 Speech Production 19
2.2 Acoustic propagation 22
2.2.1 Wave equation and free field propagation 22
2.2.2 Acoustic propagation in an enclosure 24
2.3 Beamforming 30
2.4 Acoustic impulse response in the discrete-time domain 35
2.5 Time-frequency domain signal enhancement 40
2.6 Signal model 46
2.7 Speech enhancement 50
2.7.1 Spatial filtering in the time-frequency domain 50
2.7.2 Spectral filtering in the time-frequency domain 62
2.8 Evaluation of speech enhancement algorithms 67
2.9 An interpretation of the signal model 71

3 An integrated MVDR beamformer for a microphone array 75
3.1 Introduction 75
3.2 Data Model 78
3.3 Relation to prior work: robust MVDR beamformers 79
3.4 Parameter estimation for the MVDR beamformer 83
3.4.1 Estimation of $\mathbf{R}_{\mathbf{nn}}$ 83
3.4.2 Estimation of $\mathbf{h}$ 85
3.5 LCMV with a priori and estimated constraints 90
3.6 Integrated approach 92
3.6.1 Formulation 92

3.6.2 Limiting cases of the tuning parameters 94
3.6.3 Equivalence to the Speech Distortion Weighted MWF . 95
3.7 Tuning Strategy 97
3.7.1 Metric of confidence 97
3.7.2 Tuning rules 100
3.8 Evaluation and Discussion 101
3.8.1 Simulated Data 102
3.8.2 Recorded Data 106
3.9 Conclusion 111

4 Extension of microphone array processing with an external microphone 115

4.1 Introduction 115
4.1.1 The use of external microphones 115
4.1.2 Signal processing challenges with external microphones 118
4.2 Beam patterns of a local microphone array with an external microphone 120
4.3 Signal model with a local microphone array and external microphones 124
4.4 Optimal beamforming with a local microphone array and external microphones 125
4.4.1 Array gain for an MVDR beamformer 126
4.4.2 Matched Filter and Superdirective Beamformers 127
4.5 Conclusion 133

5 Estimation of spatial information for external microphones using a priori spatial knowledge from a local microphone array 135

5.1 Introduction 135
5.2 Processing with a local microphone array using a priori spatial knowledge 138

5.2.1 LMA-based MVDR 138
5.2.2 LMA-based GSC 138
5.3 MVDR beamformer with an LMA and XMs using a priori spatial knowledge 140
5.3.1 MVDR formulation 140
5.3.2 Advantages of using partial a priori spatial knowledge . 141
5.4 Completing the Blocking Matrix 144
5.4.1 Cross Correlation RTF vector estimation 145
5.4.2 EVD RTF vector estimate 146
5.5 Rank-1 GEVD Method 150
5.5.1 Overview of the method 150
5.5.2 GEVD-based RTF vector estimation 152
5.5.3 MVDR-LMA-XM beamformer 157
5.6 Evaluation and Discussion 158
5.6.1 Evaluation in an office room 158
5.6.2 Performance in a public environment 169
5.6.3 Comparison to a fully data-dependent RTF vector . . . 175
5.7 Conclusion 184

6 Binaural speech enhancement using a priori spatial knowledge and external microphones 187

6.1 Introduction 187
6.2 Data Model 190
6.3 Data-dependent estimate of the entire RTF vector 192
6.4 Using a priori knowledge of the left and right RTF vector . . . 194
6.4.1 Binaural processing 194
6.4.2 Semi-bilateral processing 197
6.5 Evaluation with BTE-HAs 198

6.6 Conclusion 203

7 MWF-based speech dereverberation using a priori spatial knowledge and an external microphone 207

7.1 Introduction 207

7.2 Data model 210

7.3 Data-dependent estimation of the MWF parameters 212

7.4 Estimation of MWF parameters using a priori knowledge of the RTF vector for the LMA 213

7.5 Evaluation 216

7.5.1 Simulated Data 216

7.5.2 Recorded data from BTE hearing aid devices 222

7.6 Conclusion 227

8 An integrated MVDR beamformer for a local microphone array linked with external microphones 229

8.1 Introduction 229

8.2 Data Model 231

8.2.1 Unprocessed Signals 231

8.2.2 Pre-whitened-transformed (PWT) domain 232

8.3 MVDR with an LMA and XMs in the PWT domain 233

8.3.1 Using an a priori RTF vector 234

8.3.2 Using a data-dependent RTF vector 236

8.4 Integrated MVDR beamformer 238

8.4.1 Quadratically Constrained Quadratic Program 238

8.4.2 Effect of $\widetilde{\epsilon}$ and $\hat{\epsilon}$ 242

8.5 Confidence Metric and Tuning 245

8.5.1 Confidence Metric 245

8.5.2 Tuning strategy 245

8.6 Evaluation and discussion 250
8.6.1 Beam patterns for a linear microphone array 250
8.6.2 Effect of $\widetilde{\epsilon}$ and $\hat{\epsilon}$ 253
8.6.3 Performance of tuning strategies 258
8.7 Conclusion 267

9 Conclusion **269**

Bibliography **277**

Biographical sketch **305**

List of publications **307**

List of Figures

1.1 A speech enhancement scenario. Several microphones firstly observe a noisy environment containing the speech signal of interest. The microphone signals are then processed by a speech enhancement system which yields an enhanced version of the desired speech signal. 3

1.2 A common processing scheme for a speech enhancement system (such as the one depicted in Fig. 1.1). 5

2.1 A simplified cross-sectional view of the speech production anatomy. 20

2.2 Source filter model for speech production. The source (airflow at the vocal folds) can be a periodic or noisy excitation (or a combination of both), which is then spectrally shaped by a low pass filter with multiple resonances (formants of the vocal tract) to produce the speech signal. 21

2.3 Illustration of the direct and reverberant energy within an enclosure as a function of distance from the source. The direct energy density from the source decays as a function of the distance-squared, while the diffuse energy density is independent of the distance. The point at which both energies are equal is referred to as the critical distance, r_c. 29

2.4 Uniformly spaced linear microphone array. 32

2.5 Polar plots of $|\Psi(\omega, \theta)|$ as a result of a DAS beamformer for $\theta_o = 60°$, $M = 4$ (left column), $M = 10$ (right column), and ratios of $\frac{d}{\lambda} = \{0.01, 0.5, 2\}$. 34

2.6 First 250 ms of the AIR from the audio laboratory at ESAT-STADIUS, KU Leuven. 37

2.7 First 500 ms of the squared-amplitude of the AIR from Fig. 2.6, illustrating the decay of the AIR. 37

2.8 Schematic representation of the system model from the target speaker (source) to multiple microphones in the discrete-time domain. 38

2.9 (a) Input speech signal, (b) AIR between the source and the receiver, (c) discrete-time convolution of the speech and the AIR according to (2.38) without the noise term, (d) actual recording from the microphone signal. 39

2.10 (Left) displays the discrete-time domain signal that is to be transformed to a (Right) time-frequency representation. 40

2.11 Processing steps for the analysis stage of the WOLA method from the discrete-time domain to the STFT domain. 42

2.12 (Left) Spectrogram of a speech signal corrupted with the noise of interfering speakers and (right) the uncorrupted clean speech signal taken from [103]. The colour bar indicates the intensity of the STFT coefficients (in dB). Note that the y-axis is the actual frequency (kHz), and not the frequency bin index. 44

2.13 Processing steps for the synthesis stage of the WOLA method from the STFT domain to the discrete-time domain. 45

2.14 Coherence between two microphones within a spherical diffuse field of (2.62) for varying separation distance, r_{pq}. The first zero crossing of each curve occurs at $\frac{r_{pq}}{\lambda} = 0.5$. 53

2.15 DI and WNG performance metrics for the MFB and SDB beamformers. 57

2.16 Generalised Sidelobe Canceller (GSC). The upper branch consists of a fixed beamformer, $\mathbf{w}_{\mathrm{f}}$, and the lower branch consists of the blocking matrix, $\mathbf{B}$, and the unconstrained filter, $\mathbf{w}_{\mathrm{g}}$. The target source signal estimate is obtained from subtracting the signal at the end of the lower branch from that of the upper branch. . . 60

2.17 Trade-off between the speech distortion and residual noise distortion for the case when $\sigma^2_{\mathrm{s_r}} = \sigma^2_{\mathrm{n}_\theta}$. The dotted curve is the MSE defined in (2.95), whose minimum is given by the value of the Wiener filter from (2.96). 64

2.18 Comparison of the speech signal from a linear convolution of a speech signal and an AIR of length L_g, $y_{LC}[n]$ with the signal,$y_{CC}[n]$, synthesised from the narrowband approximation in the STFT domain for AIRs of length L_g and frames of length P. (Left) Relative error between both signals, (right) STOI between both signals. Additional information on the comparison is provided in the text. 73

3.1 Examples of logistic functions that can be obtained by varying ρ and σ_t in (3.68) in order to define the confidence metric. ρ controls the slope of the functions, while σ_t controls the threshold principal generalised eigenvalue, beyond which $\mathrm{F}(\mathrm{l}) \to 1$. 99

3.2 Plan view of simulation environment for an accurately estimated RTF vector with a reverberation time of 0.25 s. The arrow on the speech source indicates that simulations were done for different speech source angles, including that corresponding to $\widetilde{\mathbf{h}}$. Not shown is the simulated diffuse noise field. The room height was 2.6 m. 103

3.3 Performance of the MVDR-APR, MVDR-EST and MVDR-INT with different values of $\bar{\alpha}$ and $\bar{\beta}$, for a scenario where $\hat{\mathbf{h}}$ was accurately estimated. The input SNR was 4 dB. The left column, (a)-(b), displays the MVDR-INT with $\bar{\alpha} \geq \bar{\beta}$ and the right column, (c)-(d), displays the MVDR-INT with $\bar{\beta} > \bar{\alpha}$. The corresponding legend is shown in Table II. 105

3.4 Performance of the MVDR-APR, MVDR-EST and MVDR-INT with different values of $\bar{\alpha}$ and $\bar{\beta}$, for a scenario where $\hat{\mathbf{h}}$ was inaccurately estimated. The input SNR was -3 dB. The left column, (a)-(b), displays the MVDR-INT with $\bar{\alpha} \geq \bar{\beta}$ and the right column, (c)-(d), displays the MVDR-INT with $\bar{\beta} > \bar{\alpha}$. The corresponding legend is shown in Table II. 106

3.5 Scenario for recordings captured by a 4-element microphone array in an office room. The speech source was played through SS1, which was moved from 0°, -45°, -90°, and then back to 0°. The speech source remained in each position for approximately 20 s. Speech-shaped noise was played through the NS1 at 60°. 108

3.6 (Top) Input noisy signal of the reference microphone. The arrows above the plot indicate the 20 s intervals for which the speech source was at a particular angle in accordance with Fig. 3.5. (Bottom) Corresponding probability if speech is present in this signal after applying the SPP estimator from [192]. 109

3.7 Performance of the MVDR-APR, MVDR-EST, LCMV, MVDR-INT-sdw, and MVDR-INT-cnt when $\mu = 0.001$ for the scenario of Fig. 3.5. 110

3.8 Performance of the MVDR-APR, MVDR-EST, LCMV, MVDR-INT-sdw, and MVDR-INT-cnt when $\mu = 0.1$ for the scenario of Fig. 3.5. . 111

3.9 Metric of confidence, $\mathrm{F}(\mathrm{k}, \mathrm{l})$ from the results of Fig. 3.7 and Fig. 3.8. 112

4.1 Three different scenarios used to observe the beam patters, $|\breve{\Psi}(\omega, \mathbf{r}^{\mathrm{s}}_m)|$ as defined in (4.2). Details of each scenario are given in the text. 122

4.2 Beam patterns, $|\breve{\Psi}(\omega, \mathbf{r}^{\mathrm{s}}_m)|$, for different microphone configurations involving an LMA and an XM for the matched filter, $\boldsymbol{w}_{mf}(\omega)$ at different ratios of $\frac{d}{\lambda} = \{0.01, 0.5, 2\}$ (Further description in the text). The values of $|\breve{\Psi}(\omega, \mathbf{r}^{\mathrm{s}}_m)|$ are denoted by the colourbar. The x marker represents the speech source, the group of circles at the centre represents the LMA, and the highlighted circle in the middle and right columns represents the XMs. 123

4.3 Different scenarios (a),(b),(c) outlined in the text for evaluation the MFB and SDB for using an LMA and an XM. 128

4.4 DI and WNG performance metrics for the MFB and SDB beamformers when using a microphone array and an external microphone. 128

4.5 Plan view of simulation environment for observing the performance of the MFB and SDB as a function of source position. The arrow on the speech source indicates that simulations were done for different speech source angles. Not shown is the simulated diffuse noise field. The room height was 2.5 m. 129

4.6 (Top) Δ SI-SNR for the MFB and SDB with the LMA only and with the LMA and XM as a function of source angle. The input SI-SNR at the first microphone was -3 dB. (Bottom) Δ STOI for the MFB and SDB with the LMA only and with the LMA and XM as a function of source angle. The input STOI at the first microphone was 0.68. 130

4.7 Unweighted third-octave band Δ SNRs (indicated by the colour) as a function of incident source angle for the MFB (Left) using an LMA only and (Right) using an LMA and an XM. 132

4.8 Unweighted third-octave band Δ SNRs (indicated by the colour) as a function of incident source angle for the SDB (Left) using an LMA only and (Right) using an LMA and an XM. 132

5.1 LMA-based Generalised Sidelobe Canceller using a priori spatial knowledge $\mathrm{GSC}_{\mathrm{LMA}}^{\mathrm{AP}}$. 139

5.2 Array Gain as computed in (5.11) using two different RTF vector perturbation models from (5.14) illustrating the trade-off between using an a priori or a data-dependent approach. The vertical axis and the colourbar are both indicative of the array gain in dB. This plot may be re-created from the Matlab code provided in [95]. 143

5.3 $\mathrm{GSC}_{\mathrm{LMA\text{-}XM}}^{\mathrm{AP\text{-}CC}}$: The $\mathrm{GSC}_{\mathrm{LMA}}^{\mathrm{AP}}$ extended with XMs and using the cross corrleation method of estimating the missing RTF component (CC est.) for the XMs from (5.20). 147

5.4 $\mathrm{GSC}_{\mathrm{LMA\text{-}XM}}^{\mathrm{AP\text{-}EVD}}$: A $\mathrm{GSC}_{\mathrm{LMA}}^{\mathrm{AP}}$ extended with XMs and using the EVD method of estimating the missing RTF (EVD est.) component for the XMs from (5.34). 150

5.5 $\mathrm{GSC}_{\mathrm{LMA\text{-}XM}}^{\mathrm{AP\text{-}GEVD}}$: A $\mathrm{GSC}_{\mathrm{LMA}}^{\mathrm{AP}}$ extended with XMs involving a rank-1 GEVD-based RTF vector estimation procedure. 151

5.6 Dummy behind-the-ear (BTE) hearing aid (HA) on a Neumann KU-100 dummy head. The BTE-HA consisted of two microphones spaced approximately 1.3 cm apart as shown on the leftmost photo. In some of the experiments, the BTE-HA was used on either the dummy head as shown, or on a test subject. 158

5.7 Acoustic scenario illustrating the spatial distribution of the speech source (SS1), the noise sources (NS1 - NS4), the LMA (LM1, LM2), and the XMs (XM1 - XM4). 159

5.8 Objective metrics for the various GSC algorithms, as well as for the XMs as a function of input SNR at LM1 using a perfect VAD (left) and an imperfect VAD (right). The scenario considered one male speech source (SS1), one mutli-talker babble noise source (NS1), the LMA, and one XM. Each of the three sub-plots within the particular metric uses a different XM as indicated, which corresponds to the XMs from Fig. 5.7. Each group of 2 bars for a particular algorithm represents the processing as performed with either $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ or $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$ as the a priori RTF vector for the LMA. 161

5.9 Objective metrics from the acoustic scenario of Fig. 5.7 with one speech source and four noise sources, as a function of various combinations of XMs when using a perfect VAD. 164

5.10 Objective metrics from the acoustic scenario of Fig. 5.7 with one speech source and four noise sources, as a function of various combinations of XMs when using an imperfect VAD. 165

5.11 Acoustic scenario analagous to that of Fig 5.7, except with two XMs. After 20 s, XM4 was switched to XM3 as indicated by the arrow XM-SWITCH-1, and after 40 s, XM1 was switched to XM1 as indicated by the arrow XM-SWITCH-2. 167

5.12 Performance of the algorithms when switching between different XMs over time using a perfect VAD (left) and an imperfect VAD (right). The uppermost plots display the clean speech reference signal in LM1 with the respective VAD superimposed. The middle plots show the Δ SI-SNR metric and the final row of plots show the Δ STOI metric. Both Δ SI-SNR and Δ STOI metrics were computed with 3 s frames with a 50 % overlap. The markers for the XMs correspond to the middle of these time frames. Markers were not used for the other plots for clarity. The vertical dotted line at 20 s indicate the switching of XMs from XM-SWITCH-1 and the vertical dotted line at 40 s indicate the switching of XMs from XM-SWITCH-2. 168

5.13 (Top) Listener wearing the BTE HAs. (Bottom) Conversation between the target speaker and the listener in a noisy cafe. The XM can also be seen on the table next to the target speaker. . 170

5.14 Plan view of the KU Leuven ESAT-STADIUS audio laboratory depicting the setup for performing an impulse response measurement from a source located at 0° (directly in front) with respect to the user equipped with the BTE HAs. 171

5.15 Measurement of the impulse response inside the KU Leuven ESAT-STADIUS audio laboratory corresponding to the depiction in Fig. 5.14. The BTE HAs can be seen on the listener's ear. . 172

5.16 (a) Impulse response from the loudspeaker to one of the BTE HA microphones from the measurement scenario of Fig. 5.14. (b) The truncated and smoothed impulse response of 256 samples used for computing the corresponding component in the a priori RTF vector. 172

5.17 Spectrograms of the (a) front left BTE HA signal (reference microphone), (b) XM signal, (c) $\mathrm{GSC}^{\mathrm{AP}}_{\mathrm{LMA}}$ with the optimal adaptive filters, (d) $\mathrm{GSC}^{\mathrm{AP}}_{\mathrm{LMA}}$ with the NLMS computed adaptive filters, (e) $\mathrm{GSC}^{\mathrm{AP\text{-}GEVD}}_{\mathrm{LMA\text{-}XM}}$ with the optimal adaptive filters, (f) $\mathrm{GSC}^{\mathrm{AP\text{-}GEVD}}_{\mathrm{LMA\text{-}XM}}$ with the NLMS computed adaptive filters. At 10 sec the target speaker moves the XM closer to them. The colour bar indicates the intensity of the STFT coefficients (in dB), which has been constrained to the range as shown for visual clarity (since the intensity of the XM STFT coefficients were substantially greater than those on the reference signal). 174

5.18 Plan view of the KU Leuven ESAT-STADIUS audio laboratory depicting the setup used in recording the data for comparing the use of an a priori RTF vector with a data-dependent RTF vector. 176

5.19 Effect of unweighted input SNR on the performance of the $\mathrm{MVDR}^{\mathrm{AP}}_{\mathrm{LMA}}$, $\mathrm{MVDR}^{\mathrm{AP}}_{\mathrm{LMA\text{-}XM}}$, and $\mathrm{MVDR}^{\mathrm{DD}}_{\mathrm{LMA\text{-}XM}}$. The left column (a)-(c) corresponds to using an LMA with $\mathrm{M_a} = 2$ and the right column (d)-(f) corresponds to using an LMA with $\mathrm{M_a} = 4$. . . . 179

5.20 Effect of the microphone used to compute the SPP on the performance of the $\mathrm{MVDR}^{\mathrm{AP}}_{\mathrm{LMA}}$, $\mathrm{MVDR}^{\mathrm{AP}}_{\mathrm{LMA\text{-}XM}}$, and $\mathrm{MVDR}^{\mathrm{DD}}_{\mathrm{LMA\text{-}XM}}$. The left column (a)-(c) corresponds to using an LMA with $\mathrm{M_a} = 2$ and the right column (d)-(f) corresponds to using an LMA with $\mathrm{M_a} = 4$. 180

5.21 Effect of the communication delay of the XM on the performance of the $\mathrm{MVDR}^{\mathrm{AP}}_{\mathrm{LMA}}$, $\mathrm{MVDR}^{\mathrm{AP}}_{\mathrm{LMA\text{-}XM}}$, and $\mathrm{MVDR}^{\mathrm{DD}}_{\mathrm{LMA\text{-}XM}}$. The left column (a)-(c) corresponds to using an LMA with $\mathrm{M_a} = 2$ and the right column (d)-(f) corresponds to using an LMA with $\mathrm{M_a} = 4$. 182

5.22 Effect of the location of the desired source on the performance of the $\mathrm{MVDR}^{\mathrm{AP}}_{\mathrm{LMA}}$, $\mathrm{MVDR}^{\mathrm{AP}}_{\mathrm{LMA\text{-}XM}}$, and $\mathrm{MVDR}^{\mathrm{DD}}_{\mathrm{LMA\text{-}XM}}$. The left column (a)-(c) corresponds to using an LMA with $\mathrm{M_a} = 2$ and the right column (d)-(f) corresponds to using an LMA with $\mathrm{M_a} = 4$. 183

6.1 Scenario with a user of a binaural assistive hearing device having access to XMs, listening to the target speaker. 190

6.2 Block scheme for a BMVDR that uses partial a priori knowledge of the RTF vectors. 193

6.3 Block scheme for a BMVDR that uses partial a priori knowledge of the RTF vectors. 196

6.4 Block scheme for a BMVDR that uses partial a priori knowledge of the RTF vectors. 198

6.5 Acoustic scenario in the KU Leuven ESAT-STADIUS audio laboratory used to evaluate the binaural processing schemes. . 199

6.6 Performance of the various BMVDR algorithms outlined in Table 6.1 when the correlation matrices have been estimated fairly accurately. 201

6.7 Performance of the various BMVDR algorithms outlined in Table 6.1 when the correlation matrices have been estimated less accurately than in Fig. 6.6. 203

7.1 Acoustic scenario consisting of a target speaker, an LMA, and an XM. 210

7.2 Simulated acoustic scenario with an LMA, a speech source, and several XMs placed at varying distances from the LMA and the speech source. The dereverberation algorithms were evaluated using the LMA and one of these XMs. The critical distance for the scenario was 0.66 m and defines the radius of the critical distance boundary (circumference) which is also depicted. . . . 218

7.3 Performance of the various dereverberation algorithms outlined in Table 7.1. The XM position number on the x-axis corresponds to the XM number as indicated on Fig. 7.2. 220

7.4 Spectrograms from two sentences using XM-5 that is within the critical distance. (a) Direct component of the target speech, (b) Unprocessed signal from the uppermost microphone of the LMA, (c) LMA-AP processed signal, (d) XM signal, (e) LMA-XM-AP processed signal, (f) LMA-XM-DD processed signal, (g) LMA-XM-AP-22 processed signal. The colour bar indicates the intensity of the STFT coefficients (in dB), which has been constrained to the range as shown for visual clarity (since the intensity of the XM STFT coefficients were substantially greater than those on the reference signal). 221

7.5 Acoustic scenario in the KU Leuven ESAT-STADIUS audio laboratory used to evaluate the dereverberation processing schemes. 223

7.6 Performance of the various dereverberation algorithms outlined in Table 7.2 using the BTE HAs. The XM position number on the x-axis corresponds to the XM number as indicated on Fig. 7.5. 225

7.7 Spectrograms from a 4 second sample of the unprocessed and processed audio with the dereverberation algorithms when using XM1. (a) Direct component of the target speech, (b) Unprocessed signal from the frontal left-ear BTE HA, (c) XM1 (d) HAL-AP, (e) HAL-XM-AP, (f) HAL-XM-DD, (g) HAL-XM-AP22. The colour bar indicates the intensity of the STFT coefficients (in dB), which has been constrained to the range as shown for visual clarity (since the intensity of the XM STFT coefficients were substantially greater than those on the reference signal). 226

8.1 Signal processing flow for obtaining a speech estimate in the PWT domain. 232

8.2 Signal processing flow for obtaining a speech estimate in the PWT domain using an a priori RTF vector. 236

8.3 Signal processing flow for obtaining a speech estimate in the PWT domain using a data-dependent RTF vector. 237

8.4 Signal processing flow for obtaining an integrated speech estimate in the PWT domain using both an a priori RTF vector and a data-dependent RTF vector. 240

8.5 Depiction of the four regions for which the APC and DDC may be active or inactive within the space spanned by the maximum tolerable speech distortion parameters, $\widetilde{\epsilon}$ and $\hat{\epsilon}$. Region I corresponds to complete suppression of the microphone signals since both constraints are inactive and hence $\underset{\sim}{\mathbf{w}}_{\text{int}} = \mathbf{0}$. Region II corresponds to using only a priori spatial information as the APC is active and the DDC inactive. Region III corresponds to using only data-dependent spatial information as the APC is inactive and the DDC is active. Region IV corresponds to the merging of both a priori and data-dependent spatial information as both the APC and DDC are active. The curve dividing the regions II and IV is the DDC bounding curve defined when the equality is satisfied from (8.45). The curve dividing the regions III and IV is the APC bounding curve defined when the equality is satisfied from (8.49). 243

8.6 Depiction of three different tuning strategies (a) trading off the maximum tolerable speech distortions between the APC and DDC, (b) fixed maximum tolerable speech distortion for the APC but variable maximum tolerable speech distortion for the DDC, and (c) fixed maximum tolerable speech distortion for the DDC but variable maximum tolerable speech distortion for the APC. 247

8.7 Beam patterns as a function of the confidence metric, F(l), for different tunings of the MVDR-INT as applied to a microphone configuration consisting of an LMA only. (Top) A tuning strategy similar to that depicted in Fig. 8.6 (a) and (Bottom) A tuning strategy similar to that depicted in Fig. 8.6 (b). F(l) = 0 corresponds to the position ϵ_{AP} and F(l) = 1 corresponds to the position ϵ_{DD} from Figs. 8.6. As F(l) increases, the path from ϵ_{AP} to ϵ_{DD} is followed resulting in the depicted beam patterns. . 251

8.8 Spatial scenario for the audio recordings. Separate recordings were made of speech signals from the loudspeakers positioned at 0° and 60°. These were then mixed with a re-created cocktail party type noise as explained in section 5.6.3 of Chapter 5. . . 254

8.9 Behaviour of the integrated $\text{MVDR}_{\text{LMA-XM}}$ beamformer as a function of $\widetilde{\epsilon}$ and $\hat{\epsilon}$ for the case when the desired speech source is at 0°, i.e., in the direction of the a priori constraint. (a) Lagrangian multiplier, $\log_{10}(\alpha)$, (b) Lagrangian multiplier, $\log_{10}(\beta)$, (c) ΔSNR, (d) Speech Distortion (SD). The APC and DDC bounding curves analogous to those from Fig. 8.5 are also shown. 256

8.10 Behaviour of the integrated $\text{MVDR}_{\text{LMA-XM}}$ beamformer as a function of $\widetilde{\epsilon}$ and $\hat{\epsilon}$ for the case when the source is at 60°, i.e., not in the direction of the a priori constraint. (a) Lagrangian multiplier, $\log_{10}(\alpha)$, (b) Lagrangian multiplier, $\log_{10}(\beta)$, (c) ΔSNR, (d) Speech Distortion (SD). The APC and DDC bounding curves analogous to those from Fig. 8.5 are also shown. 257

8.11 Performance of the MVDR-AP, MVDR-DD, and two tunings of the integrated MVDR beamformer, MVDR-INT-a and MVDR-INT-b along with XM1 and XM2 from Fig. 8.8 when SPP computed on XM2. 260

8.12 Confidence metrics of the evaluation performed in Fig. 8.11 for (Top) MVDR-INT-a and (Bottom) MVDR-INT-b. 261

8.13 Spectrograms corresponding to the signals evaluated from Fig. 8.11. Also included is the the reference speech-only signal used in computing Δ STOI (Ref) and the noisy reference microphone signal, $\text{y}_{\text{a},1}$. The colour bar indicates the intensity of the STFT coefficients (in dB), which has been constrained to the range as shown for visual clarity (since the intensity of the XM STFT coefficients were substantially greater than those on the reference signal). 262

8.14 Performance of the MVDR-AP, MVDR-DD, and two tunings of the integrated MVDR beamformer, MVDR-INT-a and MVDR-INT-b along with XM1 and XM2 from Fig. 8.8 when SPP computed on $\text{y}_{\text{a},1}$. 264

8.15 Confidence metric of the evaluation performed in Fig. 8.14 for (Top) MVDR-INT-a and (Bottom) MVDR-INT-b. 265

8.16 Spectrograms corresponding to the signals evaluated from Fig. 8.14. Also included is the the reference speech-only signal used in computing Δ STOI (Ref) and the noisy reference microphone signal, $\text{y}_{\text{a},1}$. The colour bar indicates the intensity of the STFT coefficients (in dB), which has been constrained to the range as shown for visual clarity (since the intensity of the XM STFT coefficients were substantially greater than those on the reference signal). 266

List of Tables

3.1 Table 1: Summary of speech enhancement strategies resulting from (3.56) for different tunings of $\bar{\alpha}$ and $\bar{\beta}$. 97

3.2 Global legend for the plots of the different noise reduction algorithms in Fig. 3.3 and Fig. 3.4. 104

5.1 Corresponding metrics for the individual XMs used for the algorithms corresponding to Fig. 5.9 and Fig. 5.10. 164

5.2 Summary of the experimental conditions used in evaluating the four different performance affecting criteria. 178

6.1 The various algorithms to be evaluated. 200

7.1 Dereverberation algorithms evaluated with a microphone array and an XM. 219

7.2 Dereverberation algorithms evaluated with BTE hearing aids and an XM. 224

Chapter 1

Introduction

1.1 Speech enhancement

Speech is the modality of human communication where language is conveyed through the articulation of sounds. Humans possess a unique ability (compared to other animals) to produce a wide range of distinct sounds, which in turn facilitates the communication of complex linguistic concepts. For instance, humans can produce highly distinctive vowel sounds important for quick and efficient communication such as /i/ in 'beet' or the /u/ in 'boot', whereas standard mammalians cannot [1]. Moreover, it has been demonstrated that speech merges acoustic cues in an encoding system so as to facilitate the transmission of linguistic data at very high rates [2].

Although the transmission of information via speech appears quite optimal in humans, communication is not complete without a consideration for the receiver of the information, the communication channel along which the information is transmitted, and any additional noise that may be present. A receiver of speech information is typically another human and therefore the receptor is a coupled system of the human ear and the brain. This receptor can be subdivided into two stages, namely a phonetic stage, which is related to the reception of information through the ear and permanent assembly of the brain, and a semantic stage, which takes into consideration the memory or long-term storage of information in the brain [3]. This receptor is further complemented by the existence of two ears, which makes for a powerful reception mechanism with exceptional properties [4].

Today, speech communication is widespread with the use of several devices such

as mobile telephones, videoconferecing systems, and hearing aids, where one or several microphones are used to capture the intended speech information before relaying it to the receiver[1]. In these cases, the communication channel is the path between the transmitter's voice and the microphone(s); and the noise can be several additional acoustic disturbances. Consider for instance, two persons engaging in a videoconference call, where the transmitter is a crowded cafe. As the transmitter speaks, their speech may be subject to several reflections within the room (communication channel) before arriving to the microphones. Additionally, the speech may also be corrupted by distracting sounds such as the speech generated by neighbouring persons, the clanking of cutlery, and background ambient noises. Consequently, the resulting sound that is heard on the receiving end has the potential to be a cacophony where the intended speech to be communicated becomes irretrievable.

It is in this unfortunate circumstance of events where *speech enhancement* finds its role, i.e., as a means to overcome the hindrance to efficient speech communication that is introduced by undesired signals as result of the communication channel and additional noise. In a general sense, speech enhancement can be defined as the extraction or recovery of one or more speech signals of interest by means of cancellation or suppression of undesired signals that would otherwise degrade or distort the speech signal(s) of interest [5, 6]. Hence upon capturing a set of corrupted microphone signals, they are then processed according to some rules or criteria set by a speech enhancement algorithm in order to (hopefully) yield an enhanced speech signal. Depending on the application at hand, the enhanced speech signal may be used for an auxiliary task such as transcription or transmitted to a human receiver.

1.1.1 Problem description

A generic speech enhancement scenario is illustrated in Fig. 1.2. The depiction on the left consists of a single speaker[2] (source) that generates the speech signal of interest[3], acoustic noise sources[4] and room acoustic interference (undesired

[1] In some applications, automatic speech recognition for instance, the microphones themselves may be the end receiver.

[2] Not to be confused with a loudspeaker. A speaker will refer to someone that is speaking, whereas a loudspeaker will refer specifically to the electroacoustic device. The speaker depiction in Fig. 1.2 is courtesy of [7].

[3] There may be more than one speech signal of interest in a speech enhancement scenario. The words "interest", "desired", and "target" will all be used interchangeably to refer to the signal(s) that is (are) to be enhanced.

[4] Other types of noise such as electronic noise from the microphones or quantisation noise resulting from analog to digital conversion are also relevant, but are intrinsic to the hardware devices.

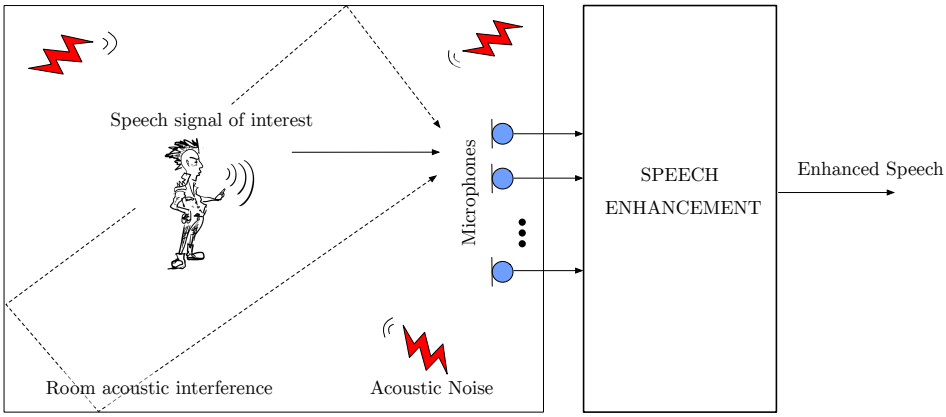


Figure 1.1: A speech enhancement scenario. Several microphones firstly observe a noisy environment containing the speech signal of interest. The microphone signals are then processed by a speech enhancement system which yields an enhanced version of the desired speech signal.

signals), and several microphones (receivers) in an enclosed space.

With respect to the undesired signals, examples of the acoustic noise source(s) may be an interfering speech signal from another speaker at a specific location, multiple speech signals from multiple speakers arriving from different locations, the passing of trains or buses, or the noise generated within a vehicle. The undesired room acoustic interference is due to the inevitable reflections (indicated by the dotted arrows in Fig. 1.1) from the boundaries of the enclosure (the communication channel), which can result in a distracting persistence of sound, known as *reverberation* (room acoustic effects are discussed in more detail in Chapter 2).

All of the microphones therefore capture a noisy signal consisting of the speech signal of interest superimposed with the several types of undesired signals within the room. These signals are sent to a speech enhancement system whose goal is then to produce an enhanced version of this noisy signal, where the speech signal of interest can be faithfully retrieved.

Interestingly enough, humans with normal hearing can comfortably focus on a speech signal of interest when confronted with the undesired signals as described in Fig. 1.1, even when there are multiple interfering speakers[5] [10, 11]. Furthermore, the spatial localisation properties of having two ears, the acoustic properties of speech, the use of prior linguistic knowledge, and the use of visual

[5]This phenomenon is referred to as the cocktail party effect [8, 9].

cues allow for humans to already do a fairly good job listening in adverse acoustic conditions [11].

Microphones, on the other hand are not so fortunate to have such exceptional properties and are essentially "brain-dead". The signals which the microphones receive are in fact quite different from what a human would perceive[6]. As a result, the desired speech signal(s) is(are) more susceptible to degradation in adverse acoustic conditions and this underscores the need for developing of effective speech enhancement systems.

Needless to say that the problem of designing a speech enhancement system is not trivial and faces several challenges. One of the major issues to address is that of *robustness* to changes and uncertainties in the acoustic environment, i.e., the speech enhancement system should be able to deliver a satisfactory performance in many if not all of the acoustic scenarios encountered. This poses a challenge primarily because it is difficult to predict the acoustic environment into which the speech enhancement will be deployed as well as the non-stationary nature of both the speech and undesired signals. Non-stationarity is considered in a broad context here meaning that signals may vary with respect to time (temporally), frequency (spectrally), and location (spatially).

1.1.2 Existing solution architecture

Depending on the application, a speech enhancement system will be designed in order to perform a number of tasks such as noise reduction [12], speech dereverberation [13], binaural speech enhancement [14], acoustic echo and feedback control [15], or a combination of such tasks. Even though countless speech enhancement algorithms have been developed to date, they commonly adhere to the architecture as depicted in Fig. 1.2 [5,6]. Further detail into the various blocks are discussed in Chapters 2 and 3, while in this section a higher level overview of the system is presented.

This system considers that there are multiple microphones which capture the noisy environment. Several multi-microphone configurations are used in practice such as a microphone array [16] where individual microphones are separated with fixed relative spacings, or what are referred to as more ad-hoc configurations [17], where microphones can be randomly distributed over a larger physical space. In either case, the first stage of several processing schemes involves a time-frequency

[6] A simple experiment can be done to make this evident. Have someone stand some distance away from you and record them speaking, for instance with a mobile phone. Upon listening to this recording, a difference should be noted between the captured signal and what is perceived when listening directly to the speaker.

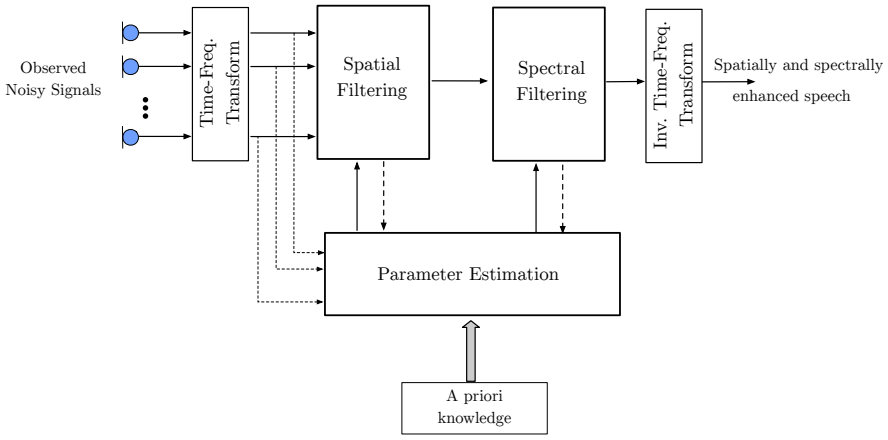


Figure 1.2: A common processing scheme for a speech enhancement system (such as the one depicted in Fig. 1.1).

transformation [18] in order to exploit both temporal and spectral variations of the microphone signals.

Since the acoustic signals in the noisy environment arrive to the different microphones at different times depending on their location in space and may be subject to different gains depending on the directivity of the microphones, the noisy microphone signals contain some spatial information or awareness of the acoustic environment. By exploiting this spatial information, the spatial filtering block of Fig. 1.2 is able to focus toward the specific location(s) of the desired speech signal(s). This is usually accomplished by applying a filter to each of the transformed microphone signals and then summing the filtered signals to produce the spatially enhanced speech signal in the time-frequency domain. This process of spatial filtering is also referred to as *beamforming* and has been the subject of an extensive field of research[7] [19, 20].

The spatially enhanced signal can then be further enhanced by a spectral filter, which applies a gain to the signal at particular frequencies. If the signal at a particular frequency is known to have speech, then applying a gain of one would maintain the speech signal without any distortion. On the other hand, if the signal at a particular frequency does not have speech (only undesired signals), then applying a gain of zero would suppress the undesired contribution at that particular frequency. Spectral filtering techniques are also applied for

[7]The concept of beamforming is quite general to any arrangement of sensors including antennas, hydrophones, accelerometers, etc. which makes it relevant for a number of research thematics. In chapter 2 and 3, more detail is provided on beamforming.

the case where only one microphone is available for speech enhancement (i.e., when spatial filtering is not possible) [21].

Critical to both spatial and spectral filtering is the parameter estimation block. While there are well-known theoretically optimal filters that can be used for the spatial and spectral processing, they are defined by a number of parameters which are unknown in practice and need to be estimated. Examples of such parameters are the statistics of the speech signal(s) and the undesired signals, as well as transfer functions between the desired speaker(s) and the microphones [22–26]. Consequently, a robust speech enhancement system is predicated on having a robust estimation of the corresponding parameters for the filters.

The parameter estimation can be split broadly into two classes. The first class relies on a priori knowledge (as depicted in Fig. 1.2) and can also be referred to as being guided[8] [27]. The second class is referred to as data-dependent, data-driven, or blind, meaning that the estimation of the parameters is done from the observed noisy signals themselves, without the need for any other prior information (as depicted by the dotted lines feeding into the parameter estimation block from the transformed noisy microphone signals in Fig. 1.2).

With the parameters estimated, they are then fed to the spatial and spectral filtering blocks as shown in Fig. 1.2 so that the filtering operations can be performed accordingly. The dotted lines that go from the spatial and spectral filtering blocks to the parameter estimation block are representative of certain algorithms which make use of iterative procedures to continuously update the filters and parameters until some optimal criterion is met. Furthermore, by performing the spatial filtering process firstly, the spatially enhanced speech signal can be used to also assist with the parameter estimation for the spectral filtering block. The final step in the entire process is then an inverse time-frequency transform which converts the processed signal back into a time-domain version which can be played back through a loudspeaker.

1.2 Research Questions

In this section, the main research questions addressed throughout the dissertation are highlighted, while each chapter will elaborate on more specific questions, methodologies, and results.

[8] More class divisions can be made for varying degrees of a priori knowledge [27].

1.2.1 Integrating a priori and data-dependent spatial information

As mentioned in the previous section, in order to have a robust speech enhancement system, robust estimation of the parameters which define the corresponding filters is required. In this dissertation, the focus is on the estimation of spatial parameters, i.e. parameters pertaining to the spatial information that relates each microphone signal to the desired speaker. There is an apparent dichotomy in the approach for the estimation of these parameters: either the guided approach is taken, where the a priori spatial information is based on a priori knowledge or a fully data-dependent approach is taken, where the spatial parameters are estimated directly from the observed microphone data.

With respect to the spatial filters, a priori knowledge[9] may include that of the microphone characteristics, relative spacings of microphones, desired speaker location, and room acoustics (e.g. no reverberation). It is quite common for instance, in assistive hearing devices to assume prior knowledge of the location of a desired speaker [30–36]. Recent technologies have also considered obtaining the location of a desired speaker through tracking eye-movements, which can also be used as a priori knowledge [37, 38]. With such a priori knowledge, theoretical models or prior measurements can be subsequently used to obtain an a priori estimate of the spatial parameters. Although the use of a priori spatial information can be robust in practice, if the assumptions used to define the a priori knowledge are broken, such as if a desired speaker is not in an a priori assumed direction, then the performance of the speech enhancement system will degrade.

Data-dependent parameter estimation approaches on the other hand have the immediate benefit of not relying on a priori assumptions. Since the spatial parameters are estimated directly from the data, then the parameters will be updated in accordance with any changes in the acoustic environment, such as the movement of a desired speaker(s). When acoustic conditions are favourable, the data-dependent approach can result in quite promising performance [25, 39–41]. However, in quite adverse acoustic conditions, data-dependent approaches are prone to poor estimation performance. For instance, in several algorithms, voice activity detectors are needed for differentiating between periods of speech and noise in order to estimate their respective statistics. In excessive noise and reverberation, especially among multiple competing speakers, this becomes

[9] With respect to temporal and spectral structure, methods for imposing a priori knowledge include the use of trained codebooks of the linear predictive or autoregressive coefficients of speech and noise signals [28, 29].

increasingly difficult [42] and hence there can be poor estimation leading to a degradation in the speech enhancement system.

As each of these approaches have particular merits and drawbacks, it is therefore hypothesised that one framework which integrates the approaches for both a priori and data-dependent spatial parameter estimation could improve the robustness of the speech enhancement system. Conceptually, from the perspective of a speech enhancement system based solely on data-dependent spatial information, incorporating a priori spatial information that is correct can be considered as a fall back or *contingency strategy* when acoustic conditions are adverse enough to significantly affect the accuracy of the estimated data-dependent spatial information. On the other hand, from the perspective of a speech enhancement system based solely on a priori spatial information, incorporating data-dependent spatial information can be viewed as an *opportunity* for an improved performance when the a priori spatial information is incorrect and the acoustic conditions are favourable. This leads to the first main research question:

How can the advantages of using both an a priori and a data-dependent spatial parameter estimation approach be integrated into one multi-microphone speech enhancement framework?

An initial investigation into this question is carried out in Chapter 3 with respect to a microphone array. It will be shown that a framework can indeed be derived to merge or integrate a priori and data-dependent spatial information from a *confidence-based* perspective. This means that a metric able to quantify the confidence there is in the accuracy of the data-dependent spatial information can be used to optimally combine both a priori and data-dependent spatial information.

1.2.2 The use of a priori spatial knowledge for a local microphone array and external microphones

The multi-microphone configurations considered for speech enhancement have been conventionally those of microphone arrays [43, 44]. Although several fruitful speech enhancement systems have been developed with these arrays, the one disadvantage is that a microphone array only provides a localised spatial sampling of the acoustic environment which places an inherent limitation on the performance that can be achieved.

Today, we are finding ourselves more frequently embedded in spaces laden with microphones, for instance, many persons have smartphones which are equipped with at least one microphone. This implies that a larger spatial sampling of

the acoustic environment can be achieved by alternative ad hoc microphone configurations, consisting of randomly placed microphones. The potential access to an increasing number of microphones has subsequently triggered significant research into the field of wireless acoustic sensor networks (WASNs) [17, 45–47], which are a specific class of wireless sensor networks [48].

A large part of this dissertation is dedicated to the study of the specific ad hoc microphone configuration comprised of a *local microphone array* (LMA) linked with one or more remote or *external microphones* (XMs). Here the terminology of a "local" microphone array is introduced since the array is considered to be confined or fixed within some area of the acoustic environment relative to the XMs which are subject to movement. Furthermore, such configurations are becoming increasingly popular in assistive hearing devices [49–51], where the microphone array on the assistive hearing device defines the reference location of the user, and hence is considered to be "local".

The previous research question related to the integration of an a priori and data-dependent spatial parameter estimation is therefore also investigated for such a microphone configuration. In this context, however, a priori spatial information is typically unavailable for the XMs. When XMs are used in conjunction with an LMA, there are no restrictions on what type of XMs can be used or where they must be placed in the acoustic environment. Hence, whereas an a priori estimate of the spatial parameters could be obtained for an LMA only due to prior knowledge of microphone spacings or some pre-calibration for instance, a similar procedure cannot be extended to the case when XMs are also used. As a result, the usual approach is to discard any available a priori spatial information for the LMA and favour a fully data-dependent spatial parameter estimation approach. Consequently, in order to be able to investigate the integration of the a priori and data-dependent spatial parameter estimation approaches for an LMA linked with XMs, it is firstly necessary to define an a priori spatial parameter estimation approach for such a microphone configuration. This forms the basis of the other central question throughout this dissertation:

How can a priori spatial knowledge be used for a microphone configuration involving a local microphone array and external microphones and to what benefit will it serve?

In Chapter 5 this question is investigated, where a speech enhancement technique is developed and evaluated such that a priori spatial knowledge is retained for the LMA and data-dependent spatial information is used for the XMs. This approach can also be interpreted as a means of integrating a priori and data-dependent spatial information, although from a different perspective than the confidence-based one in Chapter 3. The nature of the integration in this case stems from a more *geometrically-based* perspective as it is related to the

geometrical configuration of the microphones.

Following the development of this geometrically-based integration approach, it is then investigated if there are other speech enhancement tasks to which this approach could be applied and hence potentially benefit from the incorporation of a priori spatial knowledge when using XMs with an LMA. Chapters 6 and 7 discuss the results of this investigation, where the speech enhancement tasks are binaural speech enhancement and speech dereverberation respectively.

With the research question as posed in this section thoroughly investigated in Chapters 5 to 7, the research question from section 1.2.1 could then be revisited for the context of an LMA linked with XMs. Chapter 8 discusses the results of this inquiry where both the confidence-based and geometrically-based perspectives are both used to formulate a more global integration of the a priori and data-dependent spatial information, leading to a more insightful theory on how to bridge the benefits of using both classes of spatial information.

1.3 Applications and societal relevance

To highlight the relevance of speech enhancement in our increasingly digital world, a few of the major applications are discussed in this section.

Assistive Hearing Devices

According to the World Health Organisation (WHO) [52], over 400 million people worldwide suffer with hearing loss, and it is projected that this number would double by 2050. The impact of hearing loss is multifaceted, ranging from effects on the individual suffering with the hearing loss to a more global economic problem. In children, for instance, hearing loss can result in delayed language development, which has negative long term consequences on their education, health, and integration into society. In relation to the economy, the WHO has estimated that unaddressed hearing loss can cost up to USD 750 billion [52] and in developing countries, individuals with hearing loss are faced with a higher unemployment rate compared to normal hearing individuals.

One of the main applications of speech enhancement, which is also the application considered in the majority of experiments in this dissertation, is for assistive hearing devices such as hearing aids (HAs) [32, 53–55] or cochlear implants (CIs) [33, 56–59]. These are medical devices which aspire to improve the hearing capability of an individual who suffers with hearing loss. An HA or CI is typically equipped with one or multiple microphones, which capture the acoustic

environment, i.e., a mixture of the desired signal(s) and other undesired signals. Simply amplifying the captured microphone signal(s) to compensate for the hearing loss would not be a viable strategy since in addition to amplifying the desired speech signal, all of the other undesired signals or noises will also be amplified. Speech enhancement therefore becomes quite relevant in order to provide some enhanced version of the captured microphone signal(s), after which the appropriate compensation can be applied. In the case of the HA this enhanced signal is transmitted to the middle and inner ear via a loudspeaker in the ear canal, whereas for a CI, the enhanced signal will be converted to electrical impulses which will directly stimulate the inner ear.

As evidence of the complex nature of a speech enhancement strategy for an assistive hearing device, Launer et. al [54] describes three main processing blocks used in modern digital HAs. The first can be described as "sound cleaning" or noise reduction. This is the process of identifying and preserving the sound that a user would like to focus on, and removing all other undesirable sounds. The second processing block of importance is that of environment recognition, which refers to the hearing aid device being able to detect whether the user is at home, in a restaurant, or in a car, for instance. With such recognition of the environment, certain settings can be applied to other processing blocks of the hearing aid. The third main processing block corresponds to the amplification of this (hopefully) cleaned sound of interest. This amplification stage can be carefully crafted as a function of frequency and according to the hearing profile of the user and their preferences.

In addition to the microphone arrays that are located on HAs and CIs, recent systems have also incorporated the use of an XM [49–51]. The advantage of an XM is that it is mobile and hence can be closer to the desired speaker, which can ultimately result in improved speech intelligibility in comparison to what would be achieved by the HA or CI alone. Much of the work in this dissertation, including several of the experiments have specifically considered such a configuration.

Hands-free communication

The importance of speech enhancement for hands-free communication became apparent within the context of hands-free telephony [60, 61], where the users on either end of the communication system do not have to hold a physical device (handset). Rather, free-standing microphones and loudspeakers embedded on wireless devices are the receivers and transmitters. Hands-free telephony is particularly applicable to in-vehicle communication as many countries have agreed to prohibit the use of hand-held phones while vehicles are in motion [62].

Today, the hands-free communication can also be extended to include video teleconferencing systems where traditional telephone equipment is replaced by computers equipped with microphones and loudspeakers. In all of the situations, the acoustic scenario is the same. A (near-end) user speaks on one side of the communication system and the speech signal captured by the microphone(s) is transmitted to the other (far-end) user of the communication system, where it is played back through a loudspeaker. When the far-end user speaks, the same process occurs but in the other direction. From the perspective of the near-end user, however, their microphone(s) capture not only their speech signal, but also background noise, reverberation, as well as the signal of the far-end speaker that would have been played through the near-end loudspeaker. Consequently, the signal transmitted to the far-end user is subject to a poor quality as they may hear echoes of their own voice and undesired noise. As such, several efforts in speech enhancement have been dedicated to improve hands-free communication [15, 63, 64].

Pre-processing for robust automatic speech recognition

Automatic speech recognition (ASR) refers to the process of converting speech signals captured by a microphone or microphone array to its corresponding texts and represents a technology by which speech is used to convey human intentions to machines [65]. ASR has been used for instance in the health care sector for medical dictation devices, assistive devices for disabled persons, personal assistants, car navigation systems, and language learning, just to name a few. However, many ASR systems are also susceptible to performance degradations from the regular suspects of background noise and room acoustic effects, especially if a speaker is too far away from the microphone(s). Hence speech enhancement techniques become applicable as a pre-processing step for developing robust ASR systems [66–69].

1.4 Dissertation overview

The following is a chapter by chapter overview of the dissertation.

Chapter 2 provides the relevant background material from which the speech enhancement techniques and concepts developed throughout this dissertation is founded. It begins the discussion of speech enhancement from physical acoustics, describing the speech production process and the propagation of sound in the free-field and within enclosures. It is then shown how this physical

acoustic perspective can be viewed from a signals and system theory point of view, modelling the response between a transmitter and receiver as a linear time-invariant system. This eventually leads into a discussion of the practical implementation of many speech enhancement techniques via a time-frequency transform. Some of the classical speech enhancement techniques, which are still used today, will be reviewed and serve as the basis of techniques that is explored throughout the dissertation. Details with respect to the signal model and the evaluation of speech enhancement algorithms are also discussed.

Chapter 3 begins with an overview of the development of robust speech enhancement techniques related to microphone array processing (an LMA only), in particular a minimum variance distortionless response (MVDR) beamformer. A review is also provided for methods of estimation of the parameters required for the MVDR beamformer, particularly that of a relative transfer function (RTF) vector (which constitutes the relevant spatial information). Inspired by the work in [70], a cost function that integrates RTF vectors based on a priori spatial knowledge and from a data-dependent estimate is investigated for an LMA. It is found that an optimal beamformer, referred to as an integrated MVDR beamformer, can be derived that is a convex combination of an MVDR beamformer based on the a priori spatial knowledge, an MVDR beamformer based on the data-dependent estimate, a linearly constrained minimum variance (LCMV) beamformer using both the a priori spatial knowledge and the data-dependent estimate, and an all-zero vector, with real, positive-valued coefficients.

From a confidence-based perspective, it is then proposed that optimal filter combinations can be made through the use of a confidence metric indicative of the accuracy of the data-dependent spatial information. Relations to the well-known speech distortion weighted multi-channel Wiener filter are also highlighted and used along with the confidence metric to suggest practical tuning rules for a practical implementation of the integrated MVDR beamformer. Performance as evaluated on simulated data as well as audio recordings demonstrate the spatially robust potential of this integrated MVDR beamformer. This chapter is partially based on the following publication:

- Ali R., van Waterschoot T., Moonen M., "Integration of a priori and estimated constraints into an MVDR beamformer for speech enhancement", *IEEE/ACM Trans. on Audio, Speech and Lang. Process.*, vol. 27, no. 12, Dec. 2019, pp. 2288-2300

Chapter 4 makes the transition from an LMA only to a microphone system that uses an LMA and XMs. A brief discussion of this technology and some of the challenges encountered is firstly discussed. The nature of beam patterns as

it relates to an LMA linked with an XM is also explored and the performance is compared to that of using an LMA only. Furthermore, the theoretical performance of some of the classical beamforming techniques discussed in Chapter 2 is observed for an LMA linked with an XM. These results provide further evidence of the improvements in the signal-to-noise ratio (SNR) that can be achieved from beamforming techniques when using an LMA linked with an XM.

Chapter 5 investigates how a priori spatial knowledge can be used when an LMA is linked with one or multiple XMs. This is considered once again in an MVDR beamformer framework, where the a priori spatial knowledge pertaining to the LMA, i.e. the part of the RTF vector pertaining to the LMA is explicitly retained. As a result, only a partial data-dependent estimation of the RTF vector would be needed, i.e. the part pertaining to the XMs. This is in direct contrast of having to estimate the entire RTF vector in a fully data-dependent approach. As opposed to the confidence-based perspective of integration in Chapter 3, the way in which both a priori and data-dependent spatial information are combined in this chapter can be interpreted as a geometrically-based integration as it is related to the geometrical configuration of the microphones.

Three methods are proposed for estimating the partially data-dependent RTF vector pertaining to the XMs, which is used to directly or indirectly extend a generalised sidelobe canceller (GSC) beamformer. It is found that one of the proposed methods facilitated access to two separate speech estimates, i.e. one that would have been obtained by using the LMA only with a priori spatial knowledge, and another which uses this same a priori spatial knowledge, but also includes the potential additional benefit from incorporating the XMs.

An extensive evaluation is also presented, which involves behind-the-ear hearing aid devices and multiple XMs, in a number of real-world scenarios as well as a comparison to a fully data-dependent approach. The proposed geometrically-based integration approach was found to demonstrate improvements over the LMA only, the XM only, and in some cases the fully data-dependent approach in terms of speech intelligibility. This chapter is partially based on the following publications:

- Ali R., van Waterschoot T., Moonen M., “Generalised Sidelobe Canceller for Noise Reduction in Hearing Devices using an external microphone”, in Proc. of the 2018 *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Calgary, Canada, Apr. 2018.
- Ali R., van Waterschoot T., Moonen M., “Completing the RTF vector for an MVDR beamformer as applied to a local microphone array and an

external microphone", in *Proc. of the Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, Sep. 2018.

- Ali R., Bernardi G., van Waterschoot T., Moonen M., "Methods of extending a generalised sidelobe canceller with external microphones", *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 27, no. 9, Sep. 2019, pp. 1349-1364

Chapter 6 investigates the task of binaural speech enhancement, where there are two LMA devices, namely two assistive hearing devices (one for the left ear and one for the right ear), each with their own separate and different a priori spatial information. This chapter considers a microphone system involving these two LMA devices linked with XMs and explores how the geometrically-based integration approach developed in Chapter 5 of incorporating a priori spatial information can be extended for this system. This is analysed within the framework of a binaural MVDR beamformer. It is demonstrated that this a priori approach is able to preserve the relevant binaural cues related to a desired speaker, while improving speech intelligibility over using only the LMAs. It is also shown that more intelligibility improvement can be achieved when preserving a priori spatial knowledge for the left and right ear LMAs in comparison to a fully data-dependent approach, when the relevant spatial parameters are poorly estimated. This chapter is partially based on the following publication:

- Ali R., van Waterschoot T., Moonen M., "Using partial a priori knowledge of relative transfer functions to design an MVDR beamformer for a binaural hearing assistive device with external microphones", in *Proc. of the 23rd Int. Congress on Acoustics (ICA)*, Aachen, Germany, Sep. 2019, pp. 1-8.

Chapter 7 investigates the task of speech dereverberation by using a multi-channel Wiener filter (MWF) with an LMA and a single XM. It is once again explored how the geometrically-based integration approach of incorporating a priori spatial information from Chapter 5 could be applied. Whereas in Chapters 5 and 6, part of the RTF vector corresponding to the XMs had to be estimated, in this Chapter, now that an MWF is being used, it will be shown how to obtain the additional estimates required for the speech power spectral density and the reverberant power spectral density. It will also be discussed how the imposition of a priori spatial knowledge can reduce the computational complexity of these estimates. Results are presented initially using simulated data to observe how the position of the XM in relation to the LMA affects the performance, followed by results involving behind-the-ear hearing aids as the LMA and an XM in a reverberant space. It is observed that the speech

intelligibility can be considerably improved when the XM is within the critical distance of the desired speaker. This chapter is partially based on the following publication:

- Ali R., van Waterschoot T., Moonen M., "MWF-based speech dereverberation with a local microphone array and an external microphone", in *Proc. of the 27th European Signal Processing Conference (EUSIPCO)*, A Coruña, Spain, Sep. 2019.

Chapter 8 returns to the concept of an integrated MVDR beamformer developed in Chapter 3, but now within the context of an LMA linked with multiple XMs, and therefore also relies on the technique developed in chapter 5 for obtaining an a priori RTF vector for such a microphone configuration. Consequently this chapter combines both perspectives of the confidence-based integration from Chapter 3 and the geometrically-based integration from Chapter 5 to formulate a more global integration of the a priori and data-dependent spatial information.

The procedures for obtaining an a priori RTF vector and data-dependent RTF vector for this microphone system are firstly considered in a pre-whitened-transformed (PWT) domain, which consists of an initial transformation of the microphone signals through a blocking matrix and a fixed beamformer, followed by a pre-whitening operation. Using these two RTF vectors, the integrated MVDR beamformer is then formulated as a quadratically constrained quadratic problem (QCQP), with two constraints, one of which is related to the maximum tolerable speech distortion for the imposition of the a priori RTF vector and the other related to the maximum tolerable speech distortion for the imposition of the data-dependent RTF vector. An analysis of how these maximum tolerable speech distortions affect the the activity of the respective constraints for the QCQP is also presented, which highlights the relationship with the SDW-MWF and also reveals the advantage over the SDW-MWF. Using the insights from this analysis, a general tuning framework is then discussed which assigns the speech distortions to the respective constraints in accordance with a confidence metric for the accuracy of the data-dependent RTF vector.

An evaluation with audio recordings involving behind-the-ear hearing aid microphones and XMs demonstrates the spatial robustness in terms of speech intelligibility of the integrated MVDR beamformer. This chapter is partially based on the following publication:

- Ali R., van Waterschoot T., Moonen M., "An integrated MVDR beamformer for speech enhancement using a local microphone array

and external microphones", Internal Report 20-76, ESAT-STADIUS, KU Leuven, *submitted for publication*, 2020.

Chapter 9 summarises the contributions of the dissertation and discusses directions for future work.

Chapter 2

From physical acoustics to speech enhancement

A more detailed description of the entire speech enhancement process as depicted in Fig. 1.1 and Fig. 1.2 is set forth in this chapter. The physical acoustics phenomena related to speech production, acoustic propagation and theoretical beamforming are discussed in sections 2.1, 2.2 and 2.3 respectively. Signals in the discrete-time domain and the acoustic impulse response are introduced in section 2.4. In section 2.5, the transition to processing in the time-frequency domain is presented and the corresponding details of the signal model in this domain are provided in section 2.6. Speech enhancement techniques are discussed in section 2.7, with an overview of several spatial and spectral filtering techniques. The objective metrics that will be used throughout the dissertation to evaluate the performance of a speech enhancement system are reviewed in section 2.8. As the signal model is the basis under which the speech enhancement algorithms have been developed in this dissertation, it is revisited in section 2.9, where a practical interpretation is discussed.

2.1 Speech Production

A cross-sectional view of the anatomy of speech production is depicted in Fig. 2.1, highlighting the three major areas responsible for producing speech: the lungs, the larynx, and the vocal tract. The lungs act as the main source of power, generating an airflow to be directed towards the larynx. While breathing, in the inhalation phase, the volume of the chest cavity increases and the pressure

inside the lungs decreases facilitating the flow of air through the vocal tract and down towards the lungs. In the exhalation phase, the opposite occurs where the volume of the chest cavity is reduced, increasing the pressure in the lungs, which produces an airflow towards the vocal tract. The act of speaking overrides this natural rhythm of breathing as short breaths of air are inhaled and the duration of exhalation corresponds to that of a sentence or a word[1] [72].

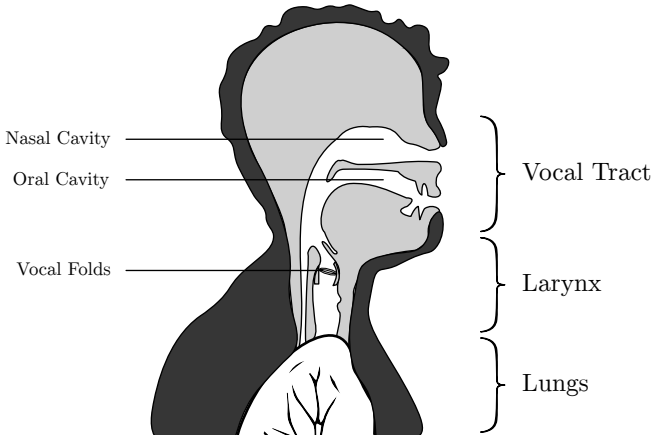


Figure 2.1: A simplified cross-sectional view of the speech production anatomy.

The role of the larynx is to control the vocal folds[2], which are two masses of tissue stretched across the larynx and are separated by an opening which is called the glottis. The vocal folds can be in one of three states: breathing, voiced, or unvoiced [72]. In the breathing state, the vocal folds are relaxed, maintaining a wide glottis so that there is minimal restriction of airflow. In the voiced state, such as in the voicing of a vowel, the vocal folds undergo a series of self sustained oscillations, due to the changes in velocity, pressure, and tension. As the lungs provide the airflow, there will be a high velocity at the glottis, which results in an air pressure decrease (Bernoulli's principle). This decrease in air pressure coupled with an increased tension of the vocal folds result in the closing of the vocal folds. As airflow is still provided from the lungs, the air pressure builds up behind the closed vocal folds, forcing it open, and the process repeats [72]. Therefore we can think of the larynx in its voiced

[1]The vocalisations that occur upon exhalation are called egressive sounds and account for the majority of human vocalisations. Ingressive sounds on the other hand, are produced during an inward airstream such as the word *jah* ('yeah') in Estonian [71].

[2]The vocal folds are also referred to as vocal cords. As pointed out by Leiberman [2], the term is derived from a theory proposed in French memoir in 1741, which postulated that the larynx could be modelled as a stringed instrument to produce sound. The French word for string is 'corde', hence the terminology vocal cords.

state as converting the steady flow of air from the lungs to periodic "puffs" of air that are directed towards the vocal tract [2, 72]. The rate at which these puffs of air occur is the fundamental frequency (Hz) and determines the pitch of the vocalisation. The final state of the vocal folds is the unvoiced state, where the vocal folds are closer together in comparison to the breathing state, and have more tension. The result is that the airflow is subject to more turbulence, which is also referred to as an aspiration [72]. This is evident when trying to voice the /h/ in 'hat' or even when whispering.

The airflow, originally generated by the lungs and subsequently modified by the larynx, then arrives to the vocal tract. The vocal tract includes both the oral and nasal cavities and extends from the larynx to the lips and nasal passage. The vocal tract acts as an acoustic resonator, analogous to the tubes of a vibraphone, which will amplify and suppress specific bands of frequencies. The effect of the resonator is that it adds a specific spectral characteristic or "colour" to the airflow coming from the larynx. Furthermore, depending on the shape of the vocal tract, which changes constantly during speaking, the characteristic of the acoustic resonator changes thereby giving rise to a wide range of sounds that can be produced. The resonant frequencies of the vocal tract are referred to as *formants*, which can be considered as the most important acoustic parameters in speech because of its quite extensive use in humans [1].

A simplified model using source filter theory [73, 74] summarises the speech production process as depicted in Fig. 2.2.

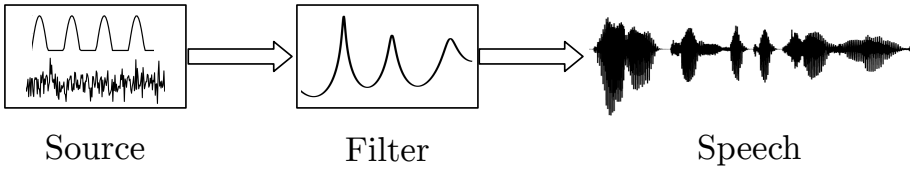


Figure 2.2: Source filter model for speech production. The source (airflow at the vocal folds) can be a periodic or noisy excitation (or a combination of both), which is then spectrally shaped by a low pass filter with multiple resonances (formants of the vocal tract) to produce the speech signal.

The larynx can be modelled as a source that produces either a periodic, or noisy airflow, which is then filtered by the vocal tract. The filter is modelled as a low pass filter with multiple resonances, where the frequencies of the resonant peaks are the formants. It can therefore be seen that through a combination of different source excitations and filters, a wide range of speech sounds can be obtained. In general these sounds will be quite non-stationary both spectrally and temporally.

This model, however, does not include any interaction between the larynx and the vocal tract, nor any additional sounds that can be generated due to the vocal tract as discussed in [72]. Nevertheless, it provides an appropriate model that is useful from a signal processing point of view.

2.2 Acoustic propagation

The speech generated is now considered as a pressure disturbance at some location in space and time. This section describes how such a pressure disturbance is affected as it acoustically propagates to some other location in space and time.

2.2.1 Wave equation and free field propagation

Luckily for us, the wave equation was derived over two centuries ago [75–77] and still provides us with some fundamental understanding of the nature of sound. The acoustic pressure[3], $p(\mathbf{r}, t)$ at position, $\mathbf{r} = \{x, y, z\}$, and time, t can be described by the wave equation in cartesian coordinates [78] [79]:

$$\nabla^2 p(\mathbf{r}, t) - \frac{1}{c^2} \frac{\partial^2 p(\mathbf{r}, t)}{\partial t^2} = 0 \tag{2.1}$$

where c is the speed of sound (m/s) and the Laplacian is defined as:

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{2.2}$$

The physical interpretation of (2.1) is that if the acoustic pressure is at a local maximum (or local minimum), then it is accelerated towards a decreasing (or increasing) pressure [80]. The speed at which the pressure disturbance travels is that of the speed of sound, c.

While (2.1) is quite simple and intuitive, it is important to acknowledge that its derivation has been a result of a linearisation of the underlying equation of state (for an adiabatic process) that describes the relationship between acoustic pressure and density, the Continuity equation that describes the conservation of mass, and the Navier-Stokes equation that describes fluid motion according to Newton's 2nd Law respectively [79]. In the Continuity and Navier-Stokes equations, the second-order terms can be neglected provided that the acoustic particle velocity is much less than the speed of sound, or if the *Mach number*

[3]The acoustic pressure in air is defined as a deviation from the mean atmospheric pressure.

$<< 1$, where the Mach number is defined as the ratio of acoustic particle velocity to the speed of sound. In terms of acoustic pressure, this linear approximation is valid when the acoustic pressure is much less than the mean atmospheric pressure. In a typical scenario of interest, a sound at 80 dB[4] is equivalent to 0.2 Pa, which is much less than the mean atmospheric pressure of 101,325 Pa. Hence, we can be contented that the acoustic propagation will be a linear process. Although this may at first seem somewhat disappointing that we do not have the capability to describe non-linear processes, we should be comforted in the fact that "superposition is the compensation we receive for enduring the limitations of linearity." [79] [81]

In general, the solutions to the wave equation can be expressed as harmonic waves [82]. Consequently, in the frequency domain, (2.1) can be expressed as the homogeneous Helmholtz equation for some angular frequency, $\omega = 2\pi f$, (f is the frequency in Hz):

$$(\nabla^2 + k^2)P(\mathbf{r}, \omega) = 0 \tag{2.3}$$

where

$$P(\mathbf{r}, \omega) = \int_{-\infty}^{\infty} p(\mathbf{r}, t)e^{-j\omega t}dt \tag{2.4}$$

is the Fourier transform of $p(\mathbf{r}, t)$, $k = \frac{\omega}{c} = \frac{2\pi}{\lambda}$ is the wavenumber, and λ is the wavelength (m^{-1}).

The Helmholtz equation now allows us to better understand the relationship between the observed acoustic pressure field at a receiver and a radiating source. If a *point source* of unit amplitude is considered, i.e. a radially oscillating sphere (compressing and expanding), whose radius is much smaller than the wavelength of the sound which is being emitted, then (2.3) can be modified to yield a generalised inhomogeneous Helmholtz equation [80]:

$$(\nabla^2 + k^2)P(\mathbf{r}, \omega) = -4\pi\delta(\mathbf{r} - \mathbf{r}_\mathrm{s}) \tag{2.5}$$

where $\mathbf{r}_\mathrm{s} = \{x_s, y_s, z_s\}$ is the source position and the 4π factor[5] results from the spherical symmetry of $P(\mathbf{r}, \omega)$ [80]. More conveniently, (2.5) may be expressed as [83] [80]:

$$(\nabla^2 + k^2)G(\mathbf{r}, \mathbf{r}_\mathrm{s}, \omega) = -4\pi\delta(\mathbf{r} - \mathbf{r}_\mathrm{s}) \tag{2.6}$$

where $G(\mathbf{r}, \mathbf{r}_\mathrm{s}, \omega)$ is the transfer function or Green's function between the source position at $\mathbf{r}_\mathrm{s}$ and the receiver position at $\mathbf{r}$ at angular frequency, ω. Hence, for

[4] Acoustic pressures in dB are in reference to 20 µPa.

[5] In some texts the 4π may not be included such as in [82]. However, it is not critical to understanding the underlying physical interpretation of the equations.

any given boundary conditions, the Green's function provides a full description of the relationship between a point source and the acoustic pressure at a receiver. In fact, an inverse Fourier transform of the Green's function results in the impulse response in the time domain between the source and the receiver (which will be discussed later on). In the simple case of an unbounded medium, i.e. a free-field condition, the Green's function is simply given by:

$$G(\mathbf{r}, \mathbf{r}_\mathrm{s}, \omega) = \frac{1}{||\mathbf{r}_\mathrm{s} - \mathbf{r}||} e^{-j\frac{\omega||\mathbf{r}_\mathrm{s}-\mathbf{r}||}{c}} \qquad \text{under free-field conditions} \tag{2.7}$$

where $||.||$ is the Euclidean vector norm and hence $||\mathbf{r}_\mathrm{s} - \mathbf{r}||$ is the distance between the source and the receiver. This expression reveals the *spherical spreading* nature of the wave propagation, i.e. the amplitude of the point source decreases with increasing distance away from the point source. It is also worth mentioning that the impulse response corresponding to the Green's function of (2.7) is a scaled and time-shifted delta function.

2.2.2 Acoustic propagation in an enclosure

While (2.7) provides an easily digestible notion of the wave propagation, we should not assimilate it too quickly as in many situations, it can be expected that the source and receiver are within some enclosure, for instance two people talking within a room. In such a case, the Green's function is given in a more general form [82]:

$$G(\mathbf{r}, \mathbf{r}_\mathrm{s}, \omega) = \sum_i A_i(\mathbf{r}_\mathrm{s}, \omega)\psi_i(\mathbf{r}) \tag{2.8}$$

which, for (2.6), has to satisfy the boundary condition $\frac{\partial G(\mathbf{r}, \mathbf{r}_\mathrm{s}, \omega)}{\partial n} = 0$, i.e. the normal component of the gradient of the Green's function is zero. $A_i(\mathbf{r}_\mathrm{s}, \omega)$ is a complex amplitude based on the source position and $\psi_i(\mathbf{r})$ is an orthogonal eigenfunction corresponding to the wavenumber, k_i, i.e. it satisfies the homogeneous equation:

$$(\nabla^2 + k_i^2)\psi_i(\mathbf{r}) = 0 \tag{2.9}$$

$\psi_i(\mathbf{r})$ can be normalised such that [82]:

$$\int_V \psi_i(\mathbf{r})\psi_j(\mathbf{r}) = \begin{cases} 0 & \text{for } i \neq j \\ V & \text{for } i = j \end{cases} \tag{2.10}$$

where V (m^3) is the volume of the enclosure. After expanding the right hand side of (2.6) into a sum of modes and finding an expression for $A_i(\mathbf{r}_\mathrm{s}, \omega)$, it can

finally be deduced that the Green's function is given by [82] [84]:

$$G(\mathbf{r}, \mathbf{r}_\mathrm{s}, \omega) = -\frac{4\pi}{V} \sum_i \frac{\psi_i(\mathbf{r})\psi_i(\mathbf{r}_\mathrm{s})}{k^2 - k_i^2} \tag{2.11}$$

where k_i is considered to be the eigenvalue corresponding to the particular eigenfunction, $\psi_i(\mathbf{r})$. Depending on the nature of the boundary of the enclosure, i.e. if it includes some absorption, then the eigenvalues can be complex [84] so that:

$$k_i = \frac{\omega_i}{c} + j\frac{\zeta_i}{c} \tag{2.12}$$

Substitution of (2.12) into (2.11), finally results in the expression for the Green's function as[6]:

$$G(\mathbf{r}, \mathbf{r}_\mathrm{s}, \omega) = -\frac{4\pi c^2}{V} \sum_i \frac{\psi_i(\mathbf{r})\psi_i(\mathbf{r}_\mathrm{s})}{\omega^2 - \omega_i^2 - 2j\zeta_i\omega_i + {\zeta_i}^2} \tag{2.13}$$

which is the direct relationship between the source at position $\mathbf{r}_\mathrm{s}$ and the receiver at position $\mathbf{r}$. It can already be seen that each term of the summation will become a maximum as $\omega \to \omega_i$, resulting in a resonance of the room and ζ_i will represent the damping of that particular resonance. Since the $(\omega^2 - \omega_i^2)$ term is responsible for the strong frequency dependence in the summation of (2.13), $2j\zeta_i\omega_i$ can be approximated by $2j\zeta_i\omega$ [84]. Then, grouping the constants and the values of the eigenfunctions into a single complex amplitude, $A_i(\mathbf{r}, \mathbf{r}_\mathrm{s}, \omega)$, (2.13) can be more compactly expressed as:

$$G(\mathbf{r}, \mathbf{r}_\mathrm{s}, \omega) = \sum_i \frac{A_i(\mathbf{r}, \mathbf{r}_\mathrm{s}, \omega)}{\omega^2 - \omega_i^2 - 2j\zeta_i\omega + {\zeta_i}^2} \tag{2.14}$$

With this compact expression, our picture of the behaviour of the sound field within an enclosure can now be updated. Simply put, it is the result of a sum of damped resonant components or what is often referred to as modes of the enclosure.

Finally, the impulse response in the time domain between the source and the receiver can be represented by the inverse Fourier transform of (2.14). The resulting solution is of the form [84]:

$$h(\mathbf{r}, \mathbf{r}_\mathrm{s}, t) = \sum_i c_i(\mathbf{r}, \mathbf{r}_\mathrm{s}) e^{-\bar{\zeta}_i t} \cos(\bar{\omega}_i t - \phi_i).\ u(t) \tag{2.15}$$

[6] In the denominator, ${\zeta_i}^2$ is not included in the equations given in [82] or [84] as it was assumed that $\zeta_i << \omega_i$, which is the condition for mostly rigid walls. In other words the complex contribution of the eigenvalues is assumed to be small.

where $u(t)$ is the Heaviside step function to maintain causality. It can be seen that (2.15) is essentially a sum of damped or decaying sinsusoids at different frequencies, each with its own damping coefficient. Depending on the degree of wall losses present in the room, $\bar{\zeta}_i$ and $\bar{\omega}_i$ are slightly different from ζ_i and ω.

While (2.14) provides a full description of the room acoustic behaviour in terms of modes, in practice it becomes more useful to treat acoustic behaviour within an enclosure into two classes according to the frequency range of interest. This can be understood by observing the density of modes (i.e. the number of modes per Hz) present for a three-dimensional enclosure [82] [79]:

$$\frac{dN}{df} \approx \frac{4\pi f^2 V}{c^3} \tag{2.16}$$

At lower frequencies, the behaviour is dominated by individual modes that are well separated from one another. However, at higher frequencies, it can be observed that the density of modes rapidly increases and there will be a significant overlap of modes. This classification in the higher frequency range of sound is typically referred to as a statistical description or model, where the sound field can be treated as isotropic and homogeneous, i.e. as a diffuse field [79].

Schroeder [85] suggested that a reasonable cross-over frequency to distinguish the individual modal behaviour at lower frequencies from the statistical behaviour at higher frequencies should be the point where 3 modes lie within the $-3\,\mathrm{dB}$ frequency bandwidth[7], Δf. Applying this criterion to (2.16) then results in the well-known *Schroeder frequency* [79] [82] :

$$f_{sch} = \sqrt{\frac{c^3 T_{60}}{4V \ln 10}} \tag{2.17}$$

For instance, in a room of dimensions $5\,\mathrm{m} \times 4\,\mathrm{m} \times 2\,\mathrm{m}$, $T_{60} = 0.5\,\mathrm{s}$, and $c = 343\,\mathrm{m/s}$, the Schroeder frequency, $f_{sch} = 234\,\mathrm{Hz}$. Hence a large part of the audible frequency spectrum for such a room can be treated according to the statistical analysis, which is inclusive of the frequency range important for speech.

The statistical model is now governed by new rules, namely that of an energy balance equation, which relates the amount of energy that is input into the room (such as that from the acoustic radiation of a loudspeaker) and the resulting losses due to absorption at the boundaries of the room[8], where the acoustic energy is converted to heat.

[7] The $-3\,\mathrm{dB}$ frequency bandwidth is defined as the bandwidth where the power of the resonance is within $-3\,\mathrm{dB}$ of its peak value.

[8] There is also absorption due to the fluid itself, but is neglected in the energy balance equation.

The energy density, ε (with units $\mathrm{kg} \cdot \mathrm{m}^{-1} \cdot \mathrm{s}^{-2}$), within a room can be given by its total potential energy density as follows [79]:

$$\varepsilon = \frac{\text{Potential Energy}}{V} = \frac{p^2}{\rho_m c^2} \tag{2.18}$$

where ρ_m is the density of the fluid (air in this case) and highlights the relationship between the acoustic energy density and the acoustic pressure within the room. The acoustic pressure is in fact independent of its location for a truly diffuse room [79]. Given this energy density, then the total energy incident per unit area and unit time at one of the boundaries of the room is given by [79] [84]:

$$B = \frac{c}{4}\varepsilon \tag{2.19}$$

which is independent of the location of the boundary[9]. The quantity of energy that is lost (converted to heat) at the boundary then depends on the specific nature of the surface, i.e., different materials exhibit different absorption characteristics. For instance, a material such as concrete will absorb the incident acoustic energy much less than a dense, foam-like material. In order to take this into account, an effective or total absorption area, S_T can be defined as follows:

$$S_T = \sum_{i}^{N_w} a_i S_i \tag{2.20}$$

where a_i is an absorption coefficient corresponding to the particular surface, S_i, for $i = 1, 2, \ldots, N_w$, where N_w is the number of surfaces within the enclosure. The total loss of acoustic energy within the enclosure is therefore the product of S_T with the incident energy defined in (2.19). Consequently, for a sound source that radiates with a constant acoustic power of Π, the energy balance equation can be formulated as [79] [84]:

$$\frac{d(\varepsilon V)}{dt} + \frac{cS_T}{4}\varepsilon = \Pi \tag{2.21}$$

which simply states that the acoustic power is balanced by any increase in acoustic energy, (εV), and the losses at the boundaries of the room [84]. The solution to (2.21) (with zero initial conditions and for a source turned on at $t = 0$) is then:

$$\varepsilon(t) = \frac{4\Pi}{cS_T}(1 - e^{-\frac{t}{\tau}}) \tag{2.22}$$

[9]The quantity B is referred to as the *irradiation strength* in [84]

where $\tau = \frac{4V}{cS_T}$ represents the time required for the energy of the diffuse sound field to reach $63.2\% = 1 - e^{-1}$ of its steady-state value after the source is turned off [79], where this steady-state value is given by:

$$\lim_{t \to \infty} \varepsilon(t) = \frac{4\Pi}{cS_T} \tag{2.23}$$

This steady-state value directly corresponds to the acoustic energy density in the diffuse field.

The nature of this energy decay from (2.22) was further studied by Sabine [86], who discovered the relationship between the time it took for acoustic energy to decay within a room and its volume and effective absorption area. This was formalised to become known as the *reverberation time*, T_{60}, which is defined as the time that it takes for the steady-state sound field to decay by 60 dB. It is given by the Sabine equation [79] [86]:

$$T_{60} = \tau 6 \ln 10 = \frac{24V \ln 10}{cS_T} \tag{2.24}$$

The reverberation time is known to be strongly related to speech intelligibility, which was in fact the motivation for Sabine to embark upon this investigation. Later on in Chapter 7, the digital signal processing technique of speech dereverberation, i.e. attempting to reduce the perceived reverberation in order to improve speech intelligibility will be investigated.

While (2.22) provides an expression of the acoustic energy decay over time within the statistical model, it is still unclear as to which points in space from the sound source can be considered to be dominated by the diffuse field, and which points can be considered as dominated by the direct radiated sound from the source. For an omnidirectional point source, the energy density of the direct sound component is inversely proportional to the square of the distance, r from the sound source:

$$\varepsilon_d = \frac{\Pi}{4\pi c r^2} \tag{2.25}$$

By equating this direct energy density ε_d to the steady-state acoustic energy density from (2.23), we can find what is known as the *critical distance* [79] [84]:

$$r_c = \frac{1}{4}\sqrt{\frac{S_T}{\pi}} \approx 0.1\sqrt{\frac{V}{\pi T_{60}}} \tag{2.26}$$

which is defined as the distance from the source where the energy of the direct component of the source is equal to the diffuse sound field energy. Therefore, the perceived sound field between the source and the boundary of the critical

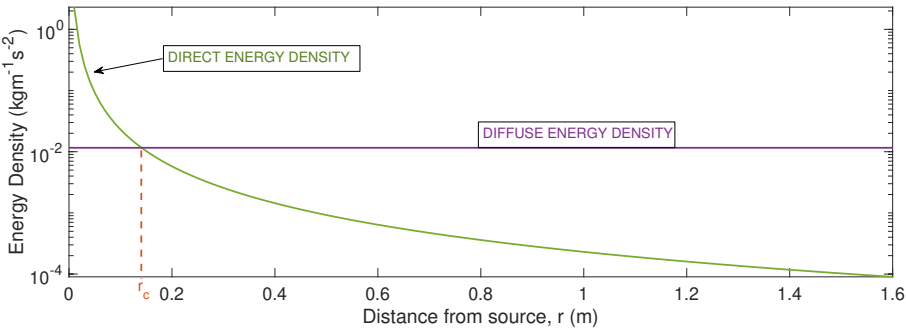


Figure 2.3: Illustration of the direct and reverberant energy within an enclosure as a function of distance from the source. The direct energy density from the source decays as a function of the distance-squared, while the diffuse energy density is independent of the distance. The point at which both energies are equal is referred to as the critical distance, r_c.

distance will be dominated by the direct component of the source, whereas beyond the critical distance, the perceived sound field will be dominated by the diffuse field. Fig. 2.3 illustrates the behaviour of the direct and diffuse energy as a function of distance from the source, with the intersection corresponding to the critical distance.

For an omnidirectional sound source, the critical distance may also be defined as [84]:

$$r_c \approx 0.1\sqrt{\frac{V}{\pi T_{60}}} \tag{2.27}$$

Finally, from (2.17) and (2.27), a relationship can be deduced between the critical distance and the Schroeder frequency:

$$r_c \approx \frac{c}{3f_{sch}} \tag{2.28}$$

which is testimony to the fact that both the critical distance and the Schroeder frequency are two aspects of the same phenomenon, which measure the dominance of the diffuse field with respect to that of the direct sound [79] and hence are characteristic descriptors of the behaviour of sound within an enclosure according to the statistical model.

2.3 Beamforming

In the speech enhancement systems, either single or multiple microphones are used as the receivers in capturing the acoustic environment. When multiple microphones are used, the *spatial information* encoded in each microphone signal will be different due to the different acoustic propagation paths from the speech source to all of the microphones as well as the different microphone responses. By exploiting this spatial information, the multiple microphone signals can be modified and combined so as to achieve a maximum response in particular directions and minimum in others. For instance, in an acoustic scenario consisting of multiple interfering speakers and one target speaker of interest, the spatial information encoded in multiple sensors within that acoustic field can be manipulated in order to attenuate the interfering speakers and maintain the target speaker. In the frequency domain[10], this is done by applying a linear time-invariant (LTI) filter, i.e. a complex weighting, to each of the sensor signals and then summing the resulting output. The two features that determine the performance of such spatial filters is their geometry and the design of the complex weightings [20]. Typical geometric configurations may consist of sensors in a linear array, a circular array, a spherical array, a planar array, or even randomly distributed. Each configuration has particular advantages and disadvantages and may be more suited depending on the application at hand. In section 2.7.1, a number of strategies are outlined for designing the complex weightings that can be applied to any particular sensor configuration. In the remaining of this section, the concept of spatial filtering in the frequency domain is introduced by analysing the beam pattern for a linear array of microphones.

For a distribution of microphones in the free-field, the spatial information can be compactly described using the Green's function from (2.7) in the following vector representation:

$$\boldsymbol{f}(\omega, \mathbf{r}_m^{\mathrm{s}}) = \begin{bmatrix} \frac{1}{||\mathbf{r}_{\mathrm{s}} - \mathbf{r}_1||} e^{-j\frac{\omega||\mathbf{r}_{\mathrm{s}} - \mathbf{r}_1||}{c}} \\ \frac{1}{||\mathbf{r}_{\mathrm{s}} - \mathbf{r}_2||} e^{-j\frac{\omega||\mathbf{r}_{\mathrm{s}} - \mathbf{r}_2||}{c}} \\ \vdots \\ \frac{1}{||\mathbf{r}_{\mathrm{s}} - \mathbf{r}_{\mathrm{M}}||} e^{-j\frac{\omega||\mathbf{r}_{\mathrm{s}} - \mathbf{r}_{\mathrm{M}}||}{c}} \end{bmatrix} \quad \text{under free-field conditions} \tag{2.29}$$

where $\mathbf{r}_m$ is the position of the mth microphone for $m \in [1, 2, \ldots, \mathrm{M}]$ and the index $\mathbf{r}_m^{\mathrm{s}}$ indicates that $\boldsymbol{f}$ is a function of the vector $\mathbf{r}_m^{\mathrm{s}} = [||\mathbf{r}_{\mathrm{s}} - \mathbf{r}_1|| \; ||\mathbf{r}_{\mathrm{s}} - \mathbf{r}_1|| \; \ldots \; ||\mathbf{r}_{\mathrm{s}} - \mathbf{r}_{\mathrm{M}}||]^T$, i.e. a vector of distances from the source position, $\mathbf{r}_{\mathrm{s}}$, to all microphone positions.

[10]The frequency domain here results from an observation interval long enough to be considered infinite. The latter sections of this chapter consider the processing in the time-frequency domain where shorter intervals are observed.

Applying a complex weighting to each of the elements in $\boldsymbol{f}(\omega, \mathbf{r}_m^{\mathrm{s}})$ and summing the output encapsulates the spatial filtering process by considering the term:

$$\Psi(\omega, \mathbf{r}_m^{\mathrm{s}}) = \boldsymbol{w}^H(\omega)\boldsymbol{f}(\omega, \mathbf{r}_m^{\mathrm{s}}) \tag{2.30}$$

where $\boldsymbol{w}(\omega) = [w_1(\omega)\ w_2(\omega)\ \ldots w_M(\omega)]^T$ is a vector of the complex weightings, commonly referred to as a *beamformer* [19]. The function, $\Psi(\omega, \mathbf{r}_m^{\mathrm{s}})$, is also commonly referred to as the *frequency-wavenumber response function* [20] when described in terms of its wavenumber, k and can also be interpreted as a transfer function for an LTI system [20].

In conventional array processing, sources are considered to be in the *far field*, i.e. they are much further in comparison to the distances that separate the microphones. In such a case, an added convenience is that we can treat the wave propagation as that of plane waves. For omnidirectional microphones, the amplitudes among the various microphones will be similar, and the differences are then only due to the phase shifts resulting from the time delay of propagation among the different microphones. Consequently, the magnitude term from (2.29) can be omitted. Furthermore, under the far field conditions, when the positions of the source and microphones are expressed in spherical coordinates, then (2.30) can be described in terms of angles of incidence in relation to the microphones.

In order to develop this notion further, the specific microphone arrangement of a uniformly spaced linear array by d (m) will be considered as in Fig. 2.4. With such an arrangement, since the source is in the far field, it can be perceived as impinging from some angle, θ in relation to the axis common to all microphones. If the direction of a source is in line with this axis, it is said to be in the *endfire* direction, whereas, if the direction of a source is perpendicular to this axis, it is said to be in the *broadside* direction. Denoting the first microphone as the uppermost microphone in Fig. 2.4, it can be observed that the additional distance that the plane waves travel to each successive microphone is simply a factor of $d\cos\theta$. (2.29) can therefore be rewritten as:

$$\boldsymbol{f}(\omega, \theta) = e^{-j\frac{\omega||\mathbf{r_S}-\mathbf{r_1}||}{c}} \underbrace{\begin{bmatrix} 1 \\ e^{-j\frac{\omega d\cos\theta}{c}} \\ \vdots \\ e^{-j\frac{\omega(M-1)d\cos\theta}{c}} \end{bmatrix}}_{\bar{\boldsymbol{f}}(\omega,\theta)} \tag{2.31}$$

where $\bar{\boldsymbol{f}}(\omega, \theta)$ is a relative *steering vector* among all of the microphones in relation to the first ($m = 1$). It is simply a vector of phase shifts as a result of the time delays, $\frac{(m-1)d\cos\theta}{c}$, for $m \in [1, 2, \ldots, \mathrm{M}]$. Therefore the spatial response can be determined as a function of angular frequency and incidence

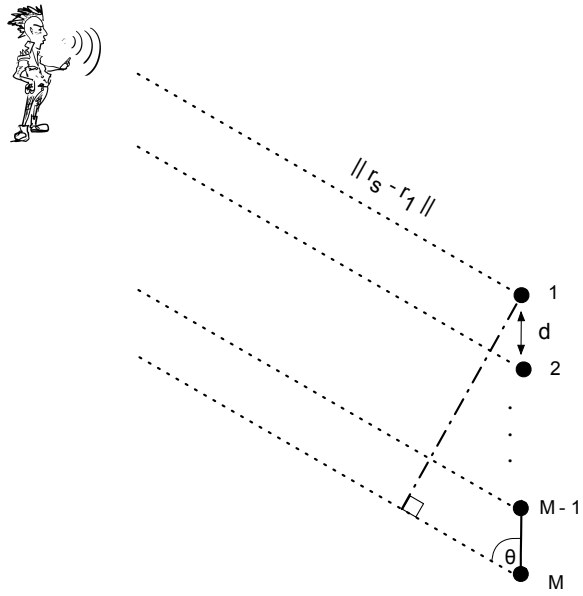


Figure 2.4: Uniformly spaced linear microphone array.

angle as:

$$\Psi(\omega, \theta) = \boldsymbol{w}^H(\omega)\bar{\boldsymbol{f}}(\omega, \theta) \tag{2.32}$$

which is also referred to as the *beam pattern*.

By adjusting the complex-valued weights, $\boldsymbol{w}(\omega)$, various beam patterns can be designed and optimised accordingly to some desired criterion. The adjustment of these complex weights is commonly referred to in the literature as beam shading (adjusting the magnitudes) and beam steering (adjusting the phases) [79] [20].

Perhaps the most conventional beamformer to use as an example to illustrate the beam patterns is that of a delay-and-sum (DAS) beamformer. As its name implies, time delays are introduced to each microphone in order to compensate for the delays experienced by each microphone as previously discussed, in relation to Fig. 2.4. This results in a coherent alignment of the source signal, which is then summed and averaged across all microphones. Such time delays can be introduced as equivalent phase shifts in $\boldsymbol{w}(\omega)$, which are derived as function of the incident source angle, and hence can be used to steer the microphone array to “look" at a particular direction. The DAS beamformer can consequently be

formulated as:

$$\boldsymbol{w}_{\text{DAS}}(\omega) = \frac{1}{M}\begin{bmatrix} 1 \\ e^{-j\frac{\omega d \cos\theta_o}{c}} \\ \vdots \\ e^{-j\frac{\omega(M-1)d\cos\theta_o}{c}} \end{bmatrix} \tag{2.33}$$

where θ_o is the desired angle to which the array is to be steered. Substitution of (2.33) in (2.32) then results in:

$$\Psi(\omega,\theta) = \frac{1}{M}\sum_{m=0}^{M-1} e^{-j\omega\frac{md(\cos\theta-\cos\theta_o)}{c}} \tag{2.34}$$

$$= \left[\frac{\sin\left(M\frac{\omega d}{2c}(\cos\theta-\cos\theta_o)\right)}{M\sin\left(\frac{\omega d}{2c}(\cos\theta-\cos\theta_o)\right)}\right] \tag{2.35}$$

$$= \left[\frac{\sin\left(M\frac{\pi d}{\lambda}(\cos\theta-\cos\theta_o)\right)}{M\sin\left(\frac{\pi d}{\lambda}(\cos\theta-\cos\theta_o)\right)}\right] \tag{2.36}$$

Hence, the beam pattern is a function of the number of microphones, M, the steering angle, θ_o, and the ratio $\frac{\omega d}{2c} = \frac{\pi d}{\lambda}$, where it is recalled that λ is the wavelength of the source signal.

Fig. 2.5 are plots of $|\Psi(\omega,\theta)|$, i.e. the magnitude of the closed-form expression in (2.35), for $\theta_o = 60°$, $M = 4$, $M = 10$, and various ratios of $\frac{d}{\lambda}$. In focusing on the left-hand column of plots for $M = 4$, the first thing to be observed is the variation in the shape of the beam pattern as a function of $\frac{d}{\lambda}$. When $\frac{d}{\lambda} = 0.01$, i.e. the case when $d << \lambda$, the beam pattern is quite omnidirectional. This is because for such long wavelengths, the phase differences, and hence the time delays, in the plane wave from any incident angle among the microphones is quite small. Consequently, there is not much steering that can be done and the array acts as an omnidirectional receiver.

As $\frac{d}{\lambda}$ increases, the beam pattern becomes more directional toward 60°. At $\frac{d}{\lambda} = 0.5$, although the beam points directly to 60°, it should be noted that due to the symmetrical nature of a linear array, another beam also appears at $-60°$. As far as the linear array is concerned, it observes the same phase difference from both of these positions, and hence cannot distinguish any difference.

This spatial ambiguity does not arise only from the geometry of the array, but also from when $d > \lambda$ as observed in Fig. 2.5 for the case of $\frac{d}{\lambda} = 2$. In such a case, as the wavelength is small in comparison to the microphone spacing, the same phase difference can be observed from a number of spatial positions,

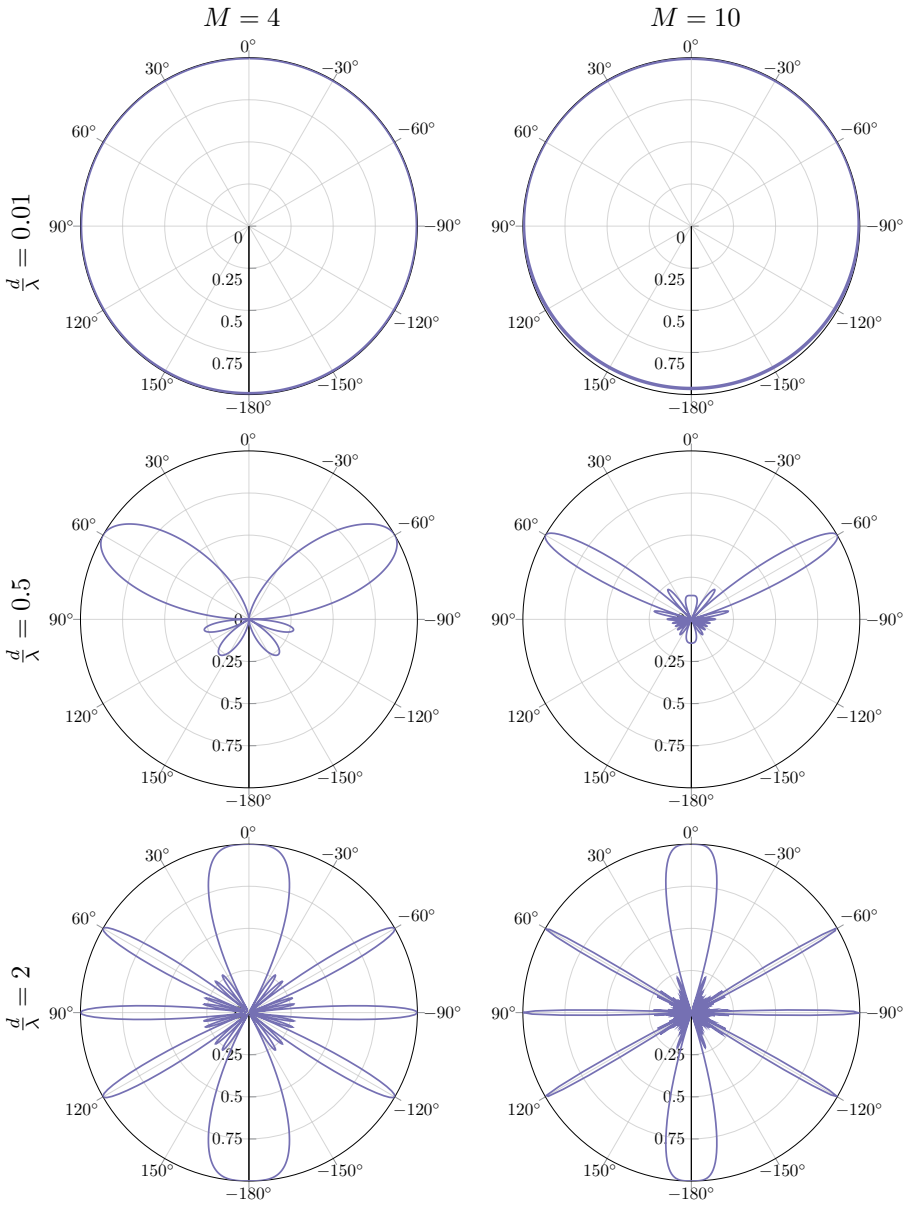


Figure 2.5: Polar plots of $|\Psi(\omega, \theta)|$ as a result of a DAS beamformer for $\theta_o = 60°$, $M = 4$ (left column), $M = 10$ (right column), and ratios of $\frac{d}{\lambda} = \{0.01, 0.5, 2\}$.

which result in the multiple beams that are produced, commonly referred to as *grating lobes* [20]. This can be interpreted as a spatial equivalent of aliasing as to what occurs in the time domain [87] and in fact the criterion under which this can be avoided is when $d \leq \frac{\lambda}{2}$ [20]. Finally, on the right-hand column of Fig. 2.5 for $M = 10$, the same trend due to the ratios of $\frac{d}{\lambda}$ can be observed, but however, now the beam widths are narrower than before.

Admittedly, it does come as a bit of a disappointment that the beam patterns are so frequency dependent for this linear array as the true beamforming or spatial selectivity occurs in a small range of frequencies. To overcome this, however, microphone configurations with varying, optimal spacing have been proposed to achieve a constant beam pattern across a wide range of frequencies [43]. Although, depending on the application at hand, such as in hearing aids, this type of approach may not be practical.

This section has given us a bit of insight into the theoretical nature of spatial processing as it relates to physical acoustics. However, in practice, the observation interval is usually not long enough to be considered infinite and as we've seen in Sec. 2.2.2, the Green's function is more complicated within an enclosure. Hence, spatial processing is commonly carried out in the time-frequency domain discussed in Sec. 2.5. After introducing the time-frequency domain, common techniques for designing spatial filters in this domain will then be discussed in section 2.7.1. Before getting here though, we must first make the transition into the discrete-time domain, which is discussed in the following section.

2.4 Acoustic impulse response in the discrete-time domain

In the previous sections, we considered the behaviour of the acoustic pressure in the continuous-time domain. However, when the acoustic environment is captured by microphones, the continuous-time domain signals must be sampled and hence we must now transition to the discrete-time domain. For a continuous-time domain signal, $x_c(t)$, its discrete-time equivalent is given by [88]:

$$x[n] = x_c(nT) = x_c\left(\frac{n}{f_s}\right) \qquad -\infty < n < \infty \tag{2.37}$$

where T is the sampling period (s), and f_s is the sampling frequency (Hz), with the relation that $T = \frac{1}{f_s}$. In the following, generally the notation of a discrete-time signal as "$x[n]$" will refer to the entire sequence of samples that

constitute the signal and it will be explicitly stated when the nth sample of the sequence is being referred to.

As we are able to consider the system between a static position of a sound source and a receiver within an enclosure as an LTI system, it can still be fully described by its impulse response [88]. Such an impulse response is related to (2.15), and in the speech enhancement context is referred to as the *acoustic impulse response* (AIR), $g[n]$.

Figure 2.6 displays the first 250 ms from an AIR measured using the exponential sine sweep method as outlined in [89]. With the help of the geometrical room acoustics perspective[11] from [84], we can make sense of this AIR by splitting it into three distinct sections. The first section is referred to as the *direct component* and represents the energy that arrives from the source directly (along the line of sight) to the receiver. The time taken for the direct component to arrive to the receiver from the source is determined by the source-receiver distance and the speed of sound. This time delay from the source to the receiver can be observed in the first few ms of Fig. 2.6, followed by the large peak corresponding to the direct component. This AIR is evident of a fairly significant direct component due to the close source-receiver distance. If the source-receiver distance is much larger, however, the direct component can be significantly lower in energy. In Fig. 1.1, the direct path is represented by the solid arrow from the speech source to one of the microphones.

After the direct component, strong reflections (depicted as the dotted arrows in Fig. 1.1) arrive from nearby surfaces such as walls, ceilings, or objects within the room near to the receiver. These are referred to as *early reflections* and it is argued that they are beneficial to speech intelligibility for normal hearing and hearing-impaired individuals [90–92]. Finally, after these strong reflections, the density of reflections begins to increase, but with progressively less energy as the sound goes through multiple reflections within the enclosure before arriving to the receiver. These *late reflections* can be considered as the sum of all of the reflections after the early reflections [84] and is referred to as the reverberation of the room and characterised by the previously discussed T_{60}. The time of transition between the early and late reflections is not fixed, but rather a function of the enclosure and is referred to as the *mixing time* [93].

Fig. 2.7 displays the first 500 ms of the square of the AIR from Fig. 2.6 on a logarithmic scale i.e. $10\log 10(g^2[n])$, where the decay[12] of the three distinct

[11] This perspective replaces the concept of a wave with a sound ray and is valid when the dimensions of the room and its walls are large in comparison to the wavelength of sound [84]. Even at a frequency of 500 Hz, the wavelength is ≈ 69 cm, which is typically less than the dimensions of a room.

[12] Energy decay curves (EDCs) are also a common way to observe the decay of the AIR [13].

sections of the AIR are more distinctly observed. The reverberant 'tail' of the AIR squared-amplitude decreases at an exponential rate (linear on the logarithmic scale), which corresponds to the exponential decay in (2.22). From Fig. 2.7, we can already obtain a back-of-the-envelope estimation[13] for the T_{60} as 1.5 s.

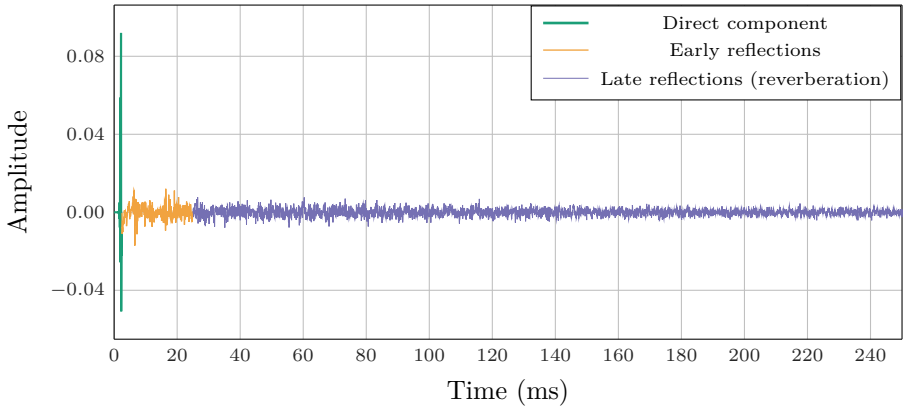


Figure 2.6: First 250 ms of the AIR from the audio laboratory at ESAT-STADIUS, KU Leuven.

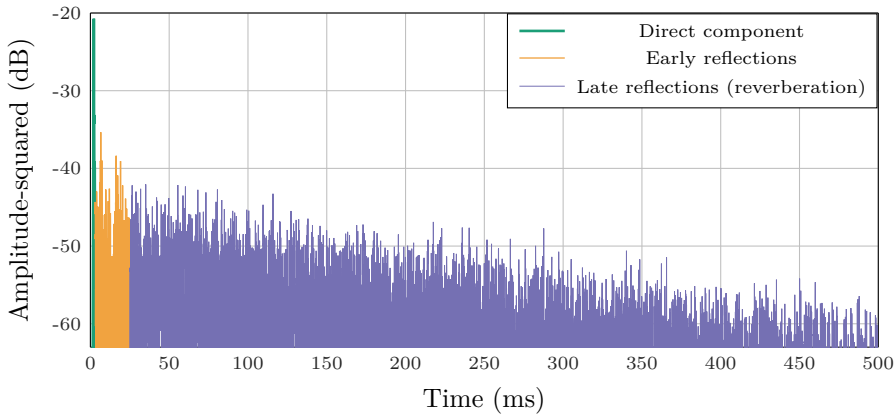


Figure 2.7: First 500 ms of the squared-amplitude of the AIR from Fig. 2.6, illustrating the decay of the AIR.

[13]Commonly associated with the renowned physicist, Enrico Fermi, a back-of-the-envelope estimation is a rough calculation done on any available scrap paper which is better than a guess, but less accurate than a mathematical proof.

Throughout this dissertation, the speech enhancement techniques that are used involve multiple microphones. Hence, for the case of a target speaker of interest, we can begin to conceptualise that a representation of the system of the source, receivers, and the AIRs may be depicted as in Fig. 2.8.

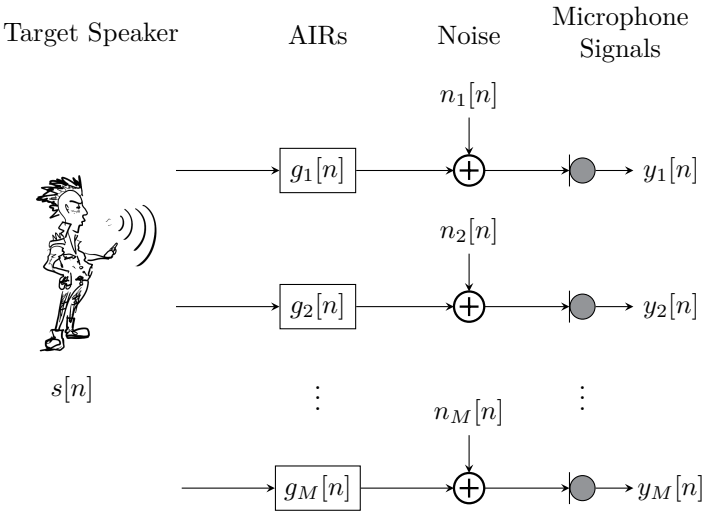


Figure 2.8: Schematic representation of the system model from the target speaker (source) to multiple microphones in the discrete-time domain.

Each microphone signal can be represented as a sum of two terms. The first consists of a linear convolution between the source signal, i.e. the target speaker signal, or the speech signal of interest[14], $s[n]$, and its corresponding AIR. The second term consists of an additive noise, which may include any combination of sensor noise, diffuse noise, localised noise, or competing target speaker signals[15]. Depending on the speech enhancement task at hand, a priori models may be used for this noise, or it may be estimated from observed data. Consequently, each microphone signal can be expressed as:

$$y_m[n] = \sum_{i=0}^{\infty} g_m[i]s[n-i] + n_m[n] \qquad m = 1, 2, \dots M \tag{2.38}$$

where M is the total number of microphones. Certainly, (2.38) can be extended if there are multiple target speakers by including an additional convolution

[14]All of these names are used interchangeably.

[15]For tasks which include speech dereverberation the AIR can be split such that a late reverberant component can be included as part of the noise term. This type of model is used in Chapter 7.

between each additional target source signal and its respective AIR. However, for simplicity in the following development, only one target speaker will be considered.

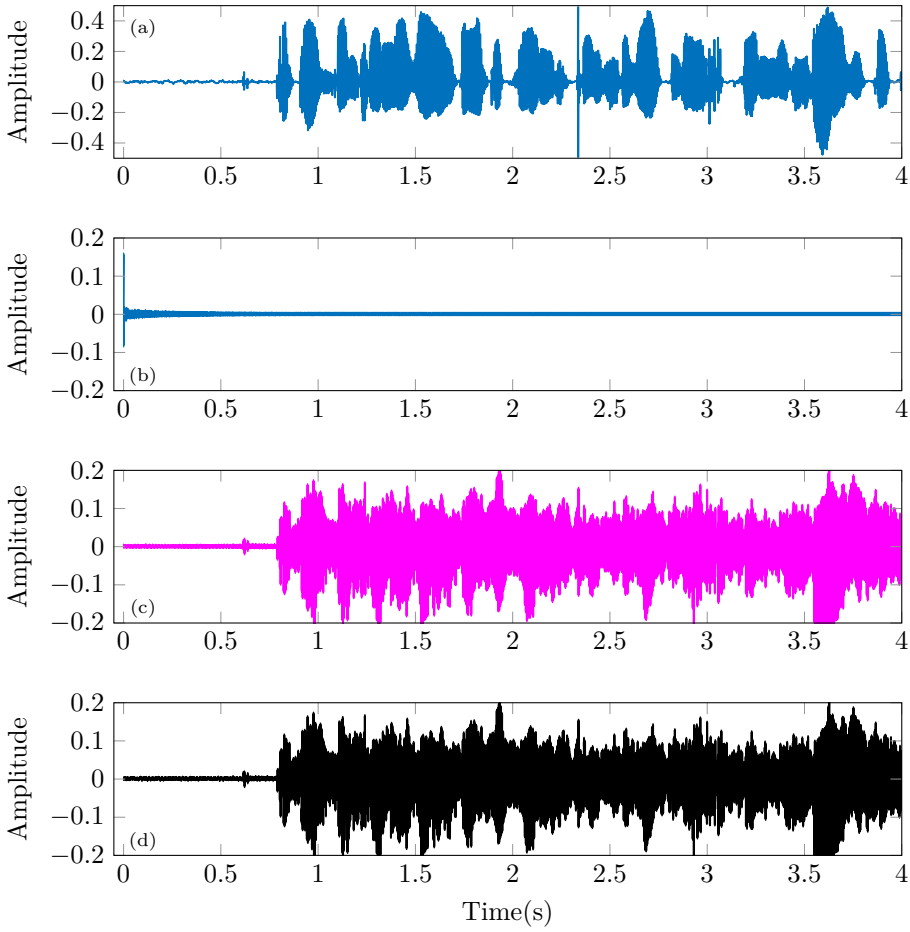


Figure 2.9: (a) Input speech signal, (b) AIR between the source and the receiver, (c) discrete-time convolution of the speech and the AIR according to (2.38) without the noise term, (d) actual recording from the microphone signal.

We can comfort ourselves that (2.38) is an accurate representation of the received signal at the microphone, without noise by observing Fig. 2.9. Fig. 2.9 (a) is a 4 s clip of an input speech signal [94] at the source position and Fig. 2.9 (b) is a measured AIR between the source and the microphone receiver that was truncated to the length shown. Fig. 2.9 (c) is the discrete-time convolution

between the speech signal and the AIR according to (2.38) without the noise term and Fig. 2.9 (d) is the actual recording as captured by the microphone. It can clearly be seen that the waveforms depicted in Figs. 2.9 (c) and Fig. 2.9 (d) are quite similar. Furthermore, upon listening to the signals provided in [95], there is perceptually no difference between the signals.

At this point, speech enhancement techniques performed in the discrete-time domain will need to incorporate some modelling of the AIR as it is unknown in practice, and all that is observable is the microphone signal. One simple model would be that of a Finite Impulse Response (FIR) filter, or an all-zero model. However, one of the disadvantages is that a large number of taps or weightings are required to model AIRs that correspond to a long T_{60}. Another option is a more compact pole-zero model or Infinite Impulse Response (IIR) filter [96–98], but estimation of the parameters can be challenging.

2.5 Time-frequency domain signal enhancement

As opposed to continuing with the enhancement of the captured signals in the discrete-time domain, it is often more common to transition to another domain where the signal can be described in terms of its spectral variations over time, i.e. a time-frequency representation as shown in Fig. 2.10. It is evident that there is more information that can be exploited, as there is spectral information in addition to the temporal information, which is represented by each "tile" in the time-frequency domain shown in Fig. 2.10. Furthermore, in such a domain, simpler signal models can often be used, upon which convenient mathematical tools can then be applied.

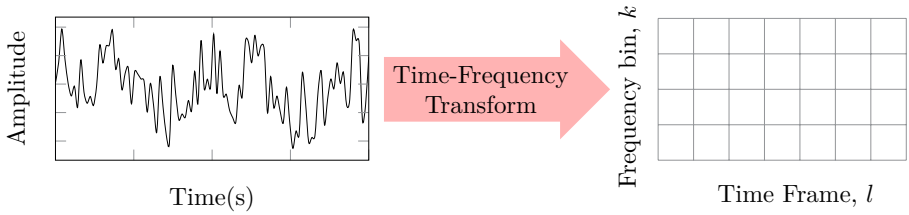


Figure 2.10: (Left) displays the discrete-time domain signal that is to be transformed to a (Right) time-frequency representation.

Recalling the overall speech enhancement framework depicted in Fig. 1.2, the time-frequency transform and its inverse was necessary for the first and last stages of the processing. In the remainder of this section, three main aspects of

the processing are described. Firstly the discrete-time domain signal is converted into its time-frequency representation (analysis). Secondly, within this domain, the signal is modified according to the enhancement algorithm, i.e., the spatial and spectral filtering. Finally, this modified time-frequency representation is converted back into the discrete-time domain (synthesis), which can then be sent to an appropriate electroacoustic system to be heard.

Throughout this dissertation, the time-frequency domain that will be used is the short-time Fourier transform (STFT), where the frequency scale is linear, meaning that there is a constant difference between adjacent frequency bins. Other time-frequency transforms may however use non-linear frequency scales such as the Mel scale [99] or the equivalent rectangular bandwidth (ERB) scale [100]. While such transforms may offer benefits in certain applications, the speech enhancement algorithms that are discussed within this dissertation are not specific to any one transform.

With respect to the analysis-synthesis procedure, it is the Weighted Overlap and Add (WOLA) [101] method that will be discussed and used throughout the dissertation. The WOLA method results in a cheaper processing scheme as compared to the overlap and add (OLA) or overlap and save (OLS) methods [102] that are used in frequency-domain adaptive filtering methods.

Analysis

Fig. 2.11 depicts the processing steps in the analysis stage of the WOLA method from the discrete-time domain signal to its STFT representation. The upper-most plot of Fig. 2.11 depicts the captured discrete-time domain signal from the mth microphone. In the second step, this discrete-time domain signal is segmented into fixed-length overlapping frames each consisting of P samples, which are then multiplied by the analysis window, $\upsilon_a[n]$. Therefore, the lth time frame depicted in the third sequence of Fig. 2.11 can be defined as:

$$\hat{y}_m[n, \mathrm{l}] = y_m[n + \mathrm{l}R]\, \upsilon_a[n] \tag{2.39}$$

for $n \in [0, 1, \ldots, P-1]$ and $\mathrm{l} \in [0, 1, \ldots, L-1]$, where L is the total number of frames and R is the hop-size in samples. R directly defines the amount of overlap, for instance, if $R = 256$ and $P = 512$, then the amount of overlap is defined as 50%.

The final step of the analysis processing then involves taking a discrete Fourier transform (DFT) on each of these windowed frames, $\hat{y}_m$, which yields complex-

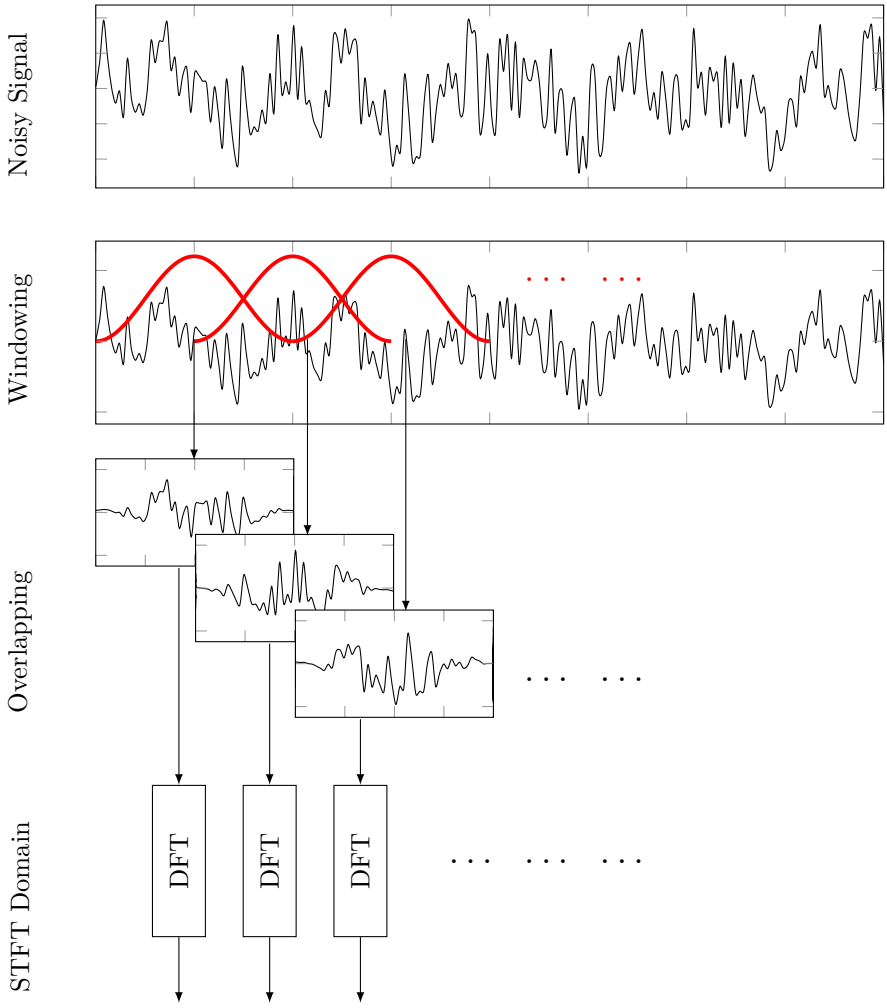


Figure 2.11: Processing steps for the analysis stage of the WOLA method from the discrete-time domain to the STFT domain.

valued STFT coefficients[16]:

$$\mathrm{y}_m(\mathrm{k},\mathrm{l}) = \sum_{n=0}^{P-1} \hat{y}_m(n,\mathrm{l}) e^{-j2\pi \frac{n\mathrm{k}}{P}} \tag{2.40}$$

[16]It should be noted that in the STFT domain, the notation for the signals is now in roman font

where P is the total number of frequency bins, and k is a frequency bin[17] corresponding to the specific frequency, f(Hz):

$$f = \frac{\mathrm{k} f_s}{P} \tag{2.41}$$

In the WOLA procedure, only the frequency bins, $\mathrm{k} \in [0, 1, \ldots, P/2]$ are processed as these correspond to the unique STFT coefficients. The STFT coefficients corresponding to the frequency bins, $\mathrm{k} \in [P/2+1, \ldots, P-1]$ are the complex conjugates of those from the frequency bins, $\mathrm{k} \in [P/2-1, \ldots, 1]$ and are hence disregarded.

One final comment to be made is that the application of the window before computing the DFT is an important step in this procedure as it helps to reduce the amount of spectral leakage between frequency bins [6,88], i.e. the amount of energy that "leaks" from one frequency bin to adjacent ones. Spectral leakage becomes more of an issue for smaller frame lengths since the DFT spectrum has to be "spread" across a smaller number of frequency bins. In general, it is recommended that a window whose magnitude response has a sufficiently narrow main lobe and low level of sidelobes be used [6].

Filtering

With the STFT coefficients computed in the analysis stage, the microphone signal at each time-frequency tile can be represented in terms of its magnitude, $|\mathrm{y}_m(\mathrm{k},\mathrm{l})|$ and its phase $\angle \mathrm{y}_m(\mathrm{k},\mathrm{l})$. A useful representation of the magnitude is that of a *spectrogram*, which is a plot of the intensity of the STFT coefficients as a function of time and frequency, where the intensity is computed as $10\log_{10}(|\mathrm{y}_m(\mathrm{k},\mathrm{l})|^2)$. Fig. 2.12 depicts the spectrogram of a speech sample corrupted with noise and the clean speech sample, where the colour is indicative of the intensity of the STFT coefficients in dB. These two images are markedly different, where the noisy spectrogram is representative of the data that is captured by the microphone, and the clean spectrogram is representative of the target signal. Hence the task of speech enhancement is to design some filter to be applied to the noisy STFT coefficients from each microphone signal in order to obtain an estimate of the clean speech STFT coefficients, $\hat{\mathrm{s}}(\mathrm{k},\mathrm{l})$. The speech enhancement techniques discussed throughout this dissertation take a filter-and-sum approach, where the noisy STFT coefficients from each

[17] Recall here once again that in the notation, k refers to a frequency bin and k is the wavenumber discussed in Sec. 2.2

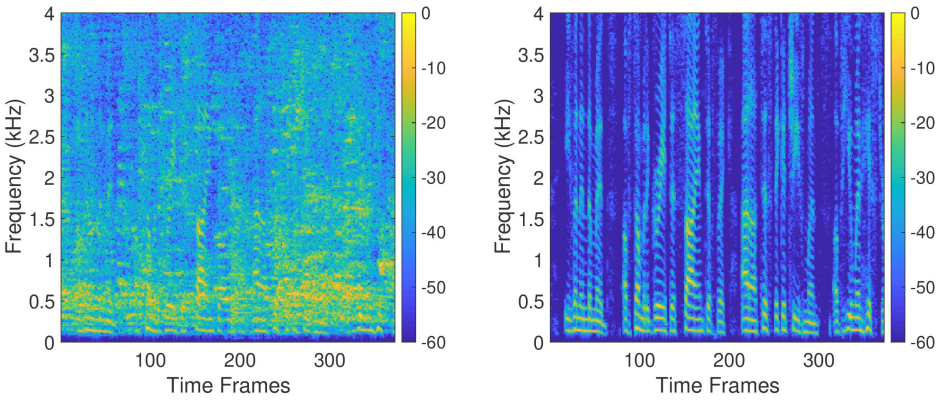


Figure 2.12: (Left) Spectrogram of a speech signal corrupted with the noise of interfering speakers and (right) the uncorrupted clean speech signal taken from [103]. The colour bar indicates the intensity of the STFT coefficients (in dB). Note that the y-axis is the actual frequency (kHz), and not the frequency bin index.

microphone signal are filtered and then summed to obtain $\hat{\mathrm{s}}(\mathrm{k},\mathrm{l})$:

$$\hat{\mathrm{s}}(\mathrm{k},\mathrm{l}) = \sum_{m=1}^{M} \mathrm{w}_m^*(\mathrm{k},\mathrm{l})\mathrm{y}_m(\mathrm{k},\mathrm{l}) \tag{2.42}$$

where $\{.\}^*$ is the complex-conjugate and $\mathrm{w}_m(\mathrm{k},\mathrm{l})$, is a complex-valued filter applied to time-frequency tile (k,l), for the microphone, m. In other words, some gain and phase adjustment is applied to each time-frequency tile for the particular microphone. A discussion on the strategies for designing such filters is left for section 2.7. In continuing with the overall processing scheme, now with an estimate of the clean speech STFT coefficients, the next step is to transform back into the discrete-time domain.

Synthesis

The process of returning to the discrete-time domain from the STFT domain is referred to as the synthesis and the steps involved are depicted in Fig. 2.13 as it pertains to the WOLA procedure. Each enhanced frame of STFT coefficients is firstly transformed back to the discrete-time domain by application of the

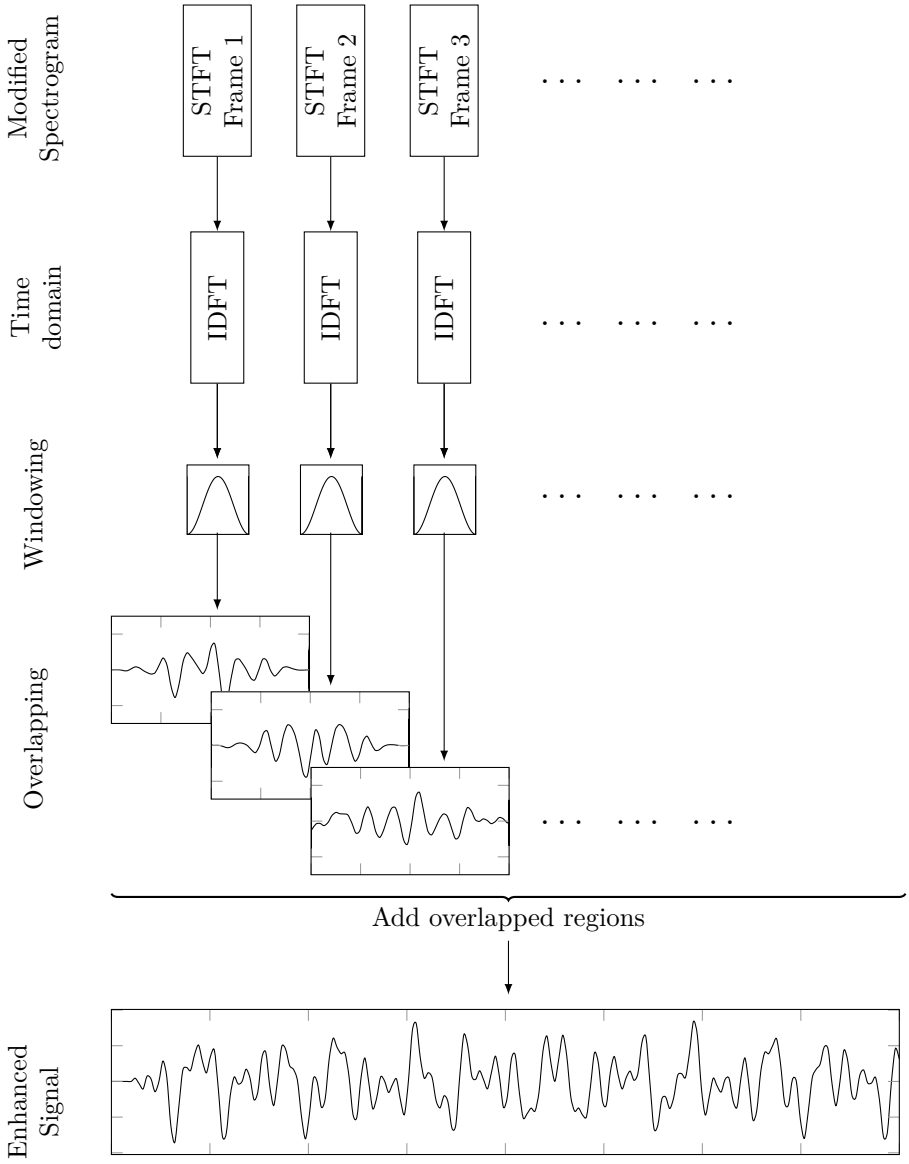


Figure 2.13: Processing steps for the synthesis stage of the WOLA method from the STFT domain to the discrete-time domain.

inverse discrete Fourier transform (IDFT):

$$\hat{s}[n,\mathrm{l}] = \frac{1}{P}\sum_{k=0}^{P-1} \hat{\mathrm{s}}(\mathrm{k},\mathrm{l}) e^{j2\pi\frac{n\mathrm{k}}{P}} \tag{2.43}$$

for $n \in [0, 1, \ldots, P-1]$, yielding a time-domain segment of P samples. In order to reduce artifacts that may be generated along the boundaries of the segments, another window, i.e. a synthesis window, $\upsilon_s[n]$ is applied to the segment, $\hat{s}[n, \mathrm{l}]$ [6]. Finally these windowed frames are overlapped (depicted in Fig. 2.13) according to the same hop size, R, as in the analysis stage, and then summed to yield the enhanced signal in the discrete-time domain:

$$\hat{s}[n] = \sum_{l=0}^{L-1} \hat{s}[\mathrm{l}, n - \mathrm{l}R]\, \upsilon_s[\mathrm{l}, n - \mathrm{l}R] \tag{2.44}$$

In absence of any modifications to the STFT coefficients, i.e., when $\hat{s}[n] = y_m[n]$, the discrete-time domain signal obtained after the synthesis procedure is:

$$y_m[n] = y_m[n] \sum_{l=-\infty}^{\infty} \upsilon_a[\mathrm{l}, n - \mathrm{l}R]\upsilon_s[\mathrm{l}, n - \mathrm{l}R] \tag{2.45}$$

which is a distortion of the original signal by the transformation of the analysis and synthesis windows. Therefore, these windows must be chosen such that perfect reconstruction can be achieved, i.e. under the absence of spectral modifications, there should be no distortion of the signal from the beginning of the analysis stage to the end of the synthesis stage. It is clear from (2.45) that the only way this perfect reconstruction property can be attained is when [6] [104][18]:

$$\sum_{l} \upsilon_a[\mathrm{l}, n - \mathrm{l}R]\upsilon_s[\mathrm{l}, n - \mathrm{l}R] = 1, \qquad \forall n \in \mathbb{Z} \tag{2.46}$$

An example of one such pair of windows that can be used for (υ_a, υ_s) is the square-root Hanning window [104][19].

2.6 Signal model

"Replacing the part of the universe under consideration by a model of simpler but similar structure" is what Rosenbleuth and Wiener [105] referred to as *abstraction*, without which it would be impossible to grasp or control the complex nature of the universe. Therefore models must constitute a central necessity of the scientific procedure [105]. Adhering to this philosophy, it is then natural that the question one has to ask within the speech enhancement context is *what is the signal model that can be used for each time-frequency tile?* (recall

[18] `http://ccrma.stanford.edu/~jos/sasp/WOLA_Processing_Steps.html`
[19] `http://ccrma.stanford.edu/~jos/sasp/Choice_WOLA_Window.html`

the picture from Fig. 2.10). Such a model will be the basis under which the filters are designed for the speech enhancement algorithms discussed in this dissertation.

From (2.38), we know that in the discrete-time domain, the mth microphone can be represented as sum of the discrete-time convolution between the AIR, $g_m[n]$, and the target speaker signal, $s[n]$; and the noise term. Considering only the convolution term, it can be shown that the discrete-time domain convolution is equivalent to an interframe and interband convolution in the STFT domain [106, 107]:

$$\mathrm{y}_m(\mathrm{k},\mathrm{l}) = \sum_{\mathrm{k}'=0}^{P-1} \sum_{l'=-\infty}^{\infty} \mathrm{g}_m(\mathrm{k},\mathrm{k}',\mathrm{l}')\mathrm{s}(\mathrm{k}',\mathrm{l}-\mathrm{l}') \tag{2.47}$$

where $\mathrm{s}(\mathrm{k},\mathrm{l})$ is the STFT representation[20] of $s[n]$ and $\mathrm{g}_m(\mathrm{k},\mathrm{k}',\mathrm{l}')$ is a set of coefficients related to $g_m[n]$ which take into account the analysis and synthesis windows (further details are provided in [107]). Hence, (2.47) essentially states that $\mathrm{y}_m(\mathrm{k},\mathrm{l})$ can be represented as a sum of individual convolutions in each frequency band, $\mathrm{k}' \in [0, 1, \ldots, P-1]$, between $\mathrm{s}(\mathrm{k}',\mathrm{l})$ and the filter $\mathrm{g}_m(\mathrm{k},\mathrm{k}',\mathrm{l}')$ [107]. $\mathrm{g}_m(\mathrm{k},\mathrm{k}',\mathrm{l}')$ is referred to as a band-to-band filter when $\mathrm{k} = \mathrm{k}'$ and a cross-band filter otherwise, which are used for cancelling aliasing effects [107]. Furthermore, the number of cross-band filters required for an accurate representation of an impulse response depends on the analysis and synthesis windows and the length of these filters are dependent on the length of the impulse response [107]. As the representation of (2.47) can result in quite computationally intensive algorithms, some approximations are usually made in practice.

Since the energy of the analysis and synthesis window frequency responses decreases for increasing separation in frequency bins, the first such approximation is to only consider the band-to-band filters, i.e. assuming that $\mathrm{g}_m(\mathrm{k},\mathrm{k}',\mathrm{l}') \approx 0$ when $\mathrm{k} \neq \mathrm{k}'$, resulting in the subband filtering operation [6] or what is referred to as the convolutive transfer function [108]:

$$\mathrm{y}_m(\mathrm{k},\mathrm{l}) = \sum_{l'=-\infty}^{\infty} \mathrm{g}_m(\mathrm{k},\mathrm{l}')\mathrm{s}(\mathrm{k},\mathrm{l}-\mathrm{l}') \tag{2.48}$$

An even simpler approximation can be made for cases when the analysis window is much longer than the filter length, i.e. when $P >> L_g$, where L_g is the length of the AIR, $g[n]$ in the discrete-time domain. In such a case, it can

[20] This STFT representation is computed as in (2.40) after an analysis window has been applied to $s[n]$. The notation of the primes, i.e. $\{.\}'$ is simply used for indexing across the different frequency bins and time frames.

then be assumed that $\mathrm{g}_m(\mathrm{k},\mathrm{l}') \approx 0$ when $\mathrm{l}' \neq 0$, resulting in the narrowband approximation [6]:

$$\mathrm{y}_m(\mathrm{k},\mathrm{l}) = \mathrm{g}_m(\mathrm{k},\mathrm{l})\mathrm{s}(\mathrm{k},\mathrm{l}) \tag{2.49}$$

where g_m is commonly referred to as the *acoustic transfer function* (ATF). This approximation is quite popular in speech enhancement methods and will be the model that is used throughout this dissertation. It allows for each frequency bin and time frame to be treated independently. Including the noise as an additional term yields the full signal model for each microphone as:

$$\mathrm{y}_m(\mathrm{k},\mathrm{l}) = \mathrm{g}_m(\mathrm{k},\mathrm{l})\mathrm{s}(\mathrm{k},\mathrm{l}) + \mathrm{n}_m(\mathrm{k},\mathrm{l}) \tag{2.50}$$

It is important to remark here that in many cases, the use of short time frames may be preferred in order to reduce processing delays and accommodate for non-stationary signals. However, this comes at the cost of a situation arising where $L_g > P$, i.e., breaking the narrowband assumption, especially in reverberant rooms.

To avoid confusion at this point, it is important to make a distinction here between *implementing* a discrete-time convolution in the STFT domain and *modelling* a discrete-time convolution process that has already occurred. An implementation typically involves an attempt to accurately perform the exact discrete-time convolution in the STFT domain [102] between some optimal filter and the signal to be filtered. Such an implementation is relevant in applications such as echo cancellation and feedback control [15] where filters involving long impulse responses are commonly required to properly remove undesired artifacts. Hence in such a context, if the narrowband assumption is broken, there will be an imperfect modelling of the correct filter length, which will result in imperfect cancellation and an overall performance degradation. The modelling on the other hand, is our imposition of how we can represent a discrete-time convolution process that has already occurred. This directly relates to our situation of interest as the discrete-time convolution from (2.38) has already occurred, and now we need to decide on how that process should be modelled in the STFT domain. Using this model, as will be seen later, various speech enhancement algorithms can then be subsequently designed. The degree to which the narrowband assumption is broken in this case now affects the performance in a different way as the enhancement strategy can now be interpreted as one of suppression as opposed to cancellation. Consequently, even if there is some undermodelling, but substantial suppression of unwanted signals, satisfactory speech enhancement can still be accomplished.

Recall from Fig. 2.8 that our system(s) of interest consist of multiple microphones, hence the signal model can be re-written in vector notation

by stacking all M microphone signals as follows:

$$\mathbf{y}(\mathrm{k},\mathrm{l}) = \mathbf{g}(\mathrm{k},\mathrm{l})\mathrm{s}(\mathrm{k},\mathrm{l}) + \mathbf{n}(\mathrm{k},\mathrm{l}) \tag{2.51}$$

where $\mathbf{y}(\mathrm{k},\mathrm{l}) = [\mathrm{y}_1(\mathrm{k},\mathrm{l})\ \mathrm{y}_2(\mathrm{k},\mathrm{l})\ \ldots \mathrm{y}_\mathrm{M}(\mathrm{k},\mathrm{l})]^T$, $\mathbf{g}(\mathrm{k},\mathrm{l}) = [\mathrm{g}_1(\mathrm{k},\mathrm{l})\ \mathrm{g}_2(\mathrm{k},\mathrm{l})\ \ldots \mathrm{g}_\mathrm{M}(\mathrm{k},\mathrm{l})]^T$, which is an ATF vector[21], and $\mathbf{n}(\mathrm{k},\mathrm{l}) = [\mathrm{n}_1(\mathrm{k},\mathrm{l})\ \mathrm{n}_2(\mathrm{k},\mathrm{l})\ \ldots \mathrm{n}_\mathrm{M}(\mathrm{k},\mathrm{l})]^T$. Once again (2.51) can be extended to multiple target speakers by replacing $\mathrm{s}(\mathrm{k},\mathrm{l})$ with a vector of all the target speaker signals, and replacing $\mathbf{g}(\mathrm{k},\mathrm{l})$ with a matrix, whose columns correspond to the respective ATF vectors from the particular target source to the microphones.

The ATF vector relates magnitude and phase changes directly from the target source to each of the microphones[22] and hence contains the relevant spatial information needed for many speech enhancement algorithms. As opposed to the ATF vector, however, it is more common to consider the relative magnitudes and phases among the different microphones, i.e. a relative transfer function (RTF) vector [53, 109], so that (2.51) can be rewritten as:

$$\mathbf{y}(\mathrm{k},\mathrm{l}) = \underbrace{\mathbf{h}(\mathrm{k},\mathrm{l})\mathrm{s}_r(\mathrm{k},\mathrm{l})}_{\mathbf{x}(\mathrm{k},\mathrm{l})} + \mathbf{n}(\mathrm{k},\mathrm{l}) \tag{2.52}$$

where $\mathbf{x}(\mathrm{k},\mathrm{l})$ is referred to as the speech contribution and the RTF vector is defined as:

$$\mathbf{h}(\mathrm{k},\mathrm{l}) = \frac{1}{\mathrm{g}_r(\mathrm{k},\mathrm{l})}\mathbf{g}(\mathrm{k},\mathrm{l}) \tag{2.53}$$

where $\mathrm{s}_r(\mathrm{k},\mathrm{l})$ and $\mathrm{g}_r(\mathrm{k},\mathrm{l})$ are respectively, the speech component and corresponding ATF in a reference microphone, r. Hence the RTF is simply a complex scaling of the ATF vector by the ATF from a reference microphone, so that the signal model of (2.52) is now representative of the relative differences between the different microphones in relation to the one chosen as the reference. This reference microphone is typically chosen as the one with the highest signal-to-noise ratio (SNR), however, in some applications such as with hearing devices, the reference microphone may be chosen for preserving spatial cues.

[21] In anechoic conditions, the ATF vector is referred to as the *steering vector* (recall (2.31)). However it is not uncommon for these two words to be used interchangeably during discussions on speech enhancement.

[22] This ATF vector also takes into consideration the frequency responses of the microphones as well as any data acquisition hardware.

For the signal model of (2.52), we may also define the $M \times M$ spatial correlation matrix for the speech-plus-noise, noise-only, and speech-only signals as follows:

$$\begin{aligned}\mathbf{R}_{\mathbf{yy}}(\mathrm{k},\mathrm{l}) &= \mathbb{E}\{\mathbf{y}(\mathrm{k},\mathrm{l})\mathbf{y}(\mathrm{k},\mathrm{l})^H\} \\ \mathbf{R}_{\mathbf{nn}}(\mathrm{k},\mathrm{l}) &= \mathbb{E}\{\mathbf{n}(\mathrm{k},\mathrm{l})\mathbf{n}(\mathrm{k},\mathrm{l})^H\} \\ \mathbf{R}_{\mathbf{xx}}(\mathrm{k},\mathrm{l}) &= \mathbb{E}\{\mathbf{x}(\mathrm{k},\mathrm{l})\mathbf{x}(\mathrm{k},\mathrm{l})^H\}\end{aligned} \tag{2.54}$$

where $\mathbb{E}\{.\}$ is the expectation operator and $\{.\}^H$ is the Hermitian transpose. An important assumption that will prove important later on in order to derive optimal filters is that the speech signal is uncorrelated with the noise signal so that:

$$\mathbf{R}_{\mathbf{yy}}(\mathrm{k},\mathrm{l}) = \mathbf{R}_{\mathbf{xx}}(\mathrm{k},\mathrm{l}) + \mathbf{R}_{\mathbf{nn}}(\mathrm{k},\mathrm{l}) \tag{2.55}$$

Furthermore, if there is one target speaker, then $\mathbf{R}_{\mathbf{xx}}$ can be modelled as a rank-1 matrix:

$$\mathbf{R}_{\mathbf{xx}}(\mathrm{k},\mathrm{l}) = \sigma^2_{\mathrm{s_r}}(\mathrm{k},\mathrm{l})\mathbf{h}(\mathrm{k},\mathrm{l})\mathbf{h}(\mathrm{k},\mathrm{l})^H \tag{2.56}$$

where $\sigma^2_{\mathrm{s_r}}(\mathrm{k},\mathrm{l}) = \mathbb{E}\{\mathrm{s}_r(\mathrm{k},\mathrm{l})\mathrm{s}^*_r(\mathrm{k},\mathrm{l})\}$ is the speech power spectral density (PSD) in the reference microphone.

2.7 Speech enhancement

Equipped with our understanding of the time-frequency domain processing and the signal model, we can now review several spatial and spectral filtering methods. In the following only the optimal filtering strategies are discussed and not the specific parameter estimation techniques. This dissertation primarily focuses on the spatial filtering methods and the relevant parameter estimation techniques will be discussed in the specific chapters.

2.7.1 Spatial filtering in the time-frequency domain

With the signal model of (2.52) in hand, spatial filtering can be accomplished via a filter-and-sum strategy, using a set of complex-valued filters, $\mathbf{w}(\mathrm{k},\mathrm{l})$ to modify the magnitude and phase of each time-frequency tile according to some optimisation criterion. The corresponding speech estimate[23], $\hat{\mathrm{s}}_{\theta,\mathrm{r}}(\mathrm{k},\mathrm{l})$, of the

[23]The subscript in the notation of $\hat{\mathrm{s}}_{\theta,\mathrm{r}}$ is to emphasise that this speech estimate was as a result of the spatial filtering process only, where θ is symbolic of the angle from which desired speech emanates.

clean speech STFT coefficients in the reference microphone is then given as:

$$\hat{\mathrm{s}}_{\theta,\mathrm{r}}(\mathrm{k},\mathrm{l}) = \mathbf{w}(\mathrm{k},\mathrm{l})^H \mathbf{y}(\mathrm{k},\mathrm{l}) \tag{2.57}$$

where $\mathbf{w}(\mathrm{k},\mathrm{l}) = [\mathrm{w}_1(\mathrm{k},\mathrm{l})\ \mathrm{w}_2(\mathrm{k},\mathrm{l})\ \ldots \mathrm{w}_\mathrm{M}(\mathrm{k},\mathrm{l})]^T$, is the complex-valued spatial filter or beamformer applied to each time-frequency tile. Therefore, as opposed to $\boldsymbol{w}(\omega)$ and $\bar{\boldsymbol{f}}(\omega)$ from Sec. 2.3, in the STFT domain, it is $\mathbf{w}(\mathrm{k},\mathrm{l})$ and $\mathbf{h}(\mathrm{k},\mathrm{l})$ that controls the spatial response.

In order to quantify the performance of some of the classical beamformers, some performance measures are firstly defined, followed by a review of the beamformer design procedures. In the remainder of this section, since the frequency bins and time frames are treated independently, the indices, (k,l), will be dropped, which will also remove some clutter of the equations.

Performance measures of an array

Array Gain The array gain (AG) is defined as the ratio of the output signal-to-noise ratio (SNR) to the input SNR at a reference microphone. Hence it is a metric that quantifies how much improvement can be gained when using an array as opposed to a single microphone and is defined as:

$$\mathrm{AG} = \frac{\mathrm{SNR}_\mathrm{out}}{\mathrm{SNR}_\mathrm{in}} \tag{2.58}$$

where recalling (2.54), the $\mathrm{SNR}_\mathrm{out}$ can be defined as:

$$\mathrm{SNR}_\mathrm{out} = \frac{\mathbf{w}^H \mathbf{R}_{\mathbf{xx}} \mathbf{w}}{\mathbf{w}^H \mathbf{R}_{\mathbf{nn}} \mathbf{w}} = \frac{\sigma^2_{\mathrm{s_r}} |\mathbf{w}^H \mathbf{h}|^2}{\mathbf{w}^H \mathbf{R}_{\mathbf{nn}} \mathbf{w}} \tag{2.59}$$

and the $\mathrm{SNR}_\mathrm{in} = \frac{\sigma^2_{\mathrm{s_r}}}{\sigma^2_{\mathrm{n_r}}}$, where $\sigma^2_{\mathrm{n_r}}$ is the noise PSD for the reference microphone. Hence the AG can be expressed as:

$$\mathrm{AG} = \frac{\sigma^2_{\mathrm{n_r}} |\mathbf{w}^H \mathbf{h}|^2}{\mathbf{w}^H \mathbf{R}_{\mathbf{nn}} \mathbf{w}} \tag{2.60}$$

A special case of the AG can also be considered when there is a homogeneous noise field. In such circumstances, the noise PSD in each diagonal element of $\mathbf{R}_{\mathbf{nn}}$ is equal and can be normalised such that $\mathbf{R}_{\mathbf{nn}} = \sigma^2_{\mathrm{n_r}} \mathbf{\Gamma}_{\mathbf{nn}}$, where $\mathbf{\Gamma}_{\mathbf{nn}}$ is referred to as a spatial coherence matrix, whose diagonal elements are equal to 1. Making these substitutions into (2.59), the AG under a homogeneous noise field is then:

$$\mathrm{AG}^{\mathbf{\Gamma}_{\mathbf{nn}}} = \frac{|\mathbf{w}^H \mathbf{h}|^2}{\mathbf{w}^H \mathbf{\Gamma}_{\mathbf{nn}} \mathbf{w}} \tag{2.61}$$

Directivity The directivity is a specific instance of the $\mathrm{AG}^{\mathbf{\Gamma_{nn}}}$ from (2.61) when the homogeneous noise field is considered to be a diffuse noise field[24]. Therefore the directivity is an indication of how well the particular beamformer is able to suppress a diffuse noise field [110]. Each element of the spatial coherence matrix in a spherical diffuse field can be computed as follows [111–113]:

$$\mathring{\mathbf{\Gamma}}_{\mathbf{nn}}(p,q) = \mathrm{sinc}\left(\frac{\omega\, r_{pq}}{c}\right) \tag{2.62}$$

$$p = 1, 2, \ldots M$$

$$q = 1, 2, \ldots M$$

where r_{pq} is the distance between the pth and q^{th} microphone. The directivity index metric is defined accordingly as:

$$\mathrm{DI} = \frac{|\mathbf{w}^H\mathbf{h}|^2}{\mathbf{w}^H\mathring{\mathbf{\Gamma}}_{\mathbf{nn}}\mathbf{w}} \tag{2.63}$$

The coherence between two microphones within this spherical diffuse field of (2.62) is displayed in Fig. 2.14 for varying separation distance, r_{pq}. The general trend of the coherence function for all separation distances is the same in that the signals are quite correlated for a lower frequency region, but eventually become uncorrelated as the frequency increases. The point demarcating these two regions can be considered as the first zero crossing, which corresponds to $\frac{r_{pq}}{\lambda} = 0.5$ [113]. For microphones that are closely spaced, this first zero crossing clearly occurs at a higher frequency as observed in Fig. 2.14. As will be seen shortly, this characteristic of the spherical diffuse noise field will have an impact on the directivity performance of a particular beamformer.

White noise gain The white noise gain (WNG) is a special case of the $\mathrm{AG}^{\mathbf{\Gamma_{nn}}}$ when $\mathbf{\Gamma_{nn}} = \mathbf{I}$, where $\mathbf{I}$ is an identity matrix. In other words, for the special case of spatially uncorrelated noise. Hence the WNG is given as:

$$\mathrm{WNG} = \frac{|\mathbf{w}^H\mathbf{h}|^2}{\mathbf{w}^H\mathbf{w}} \tag{2.64}$$

The WNG is in fact a measure of numerical robustness of the beamformer as it quantifies how resistant the beamformer is to uncorrelated errors that may be present from microphone to microphone [114]. Such uncorrelated errors could be

[24]Homogeneous quantities refer to those that are the same in all locations, while isotropic quantities refer to those that are the same from all directions. Hence a diffuse noise field is both homogeneous and isotropic with respect to some geometry. For instance, a diffuse noise field may be cylindrically or spherically isotropic.

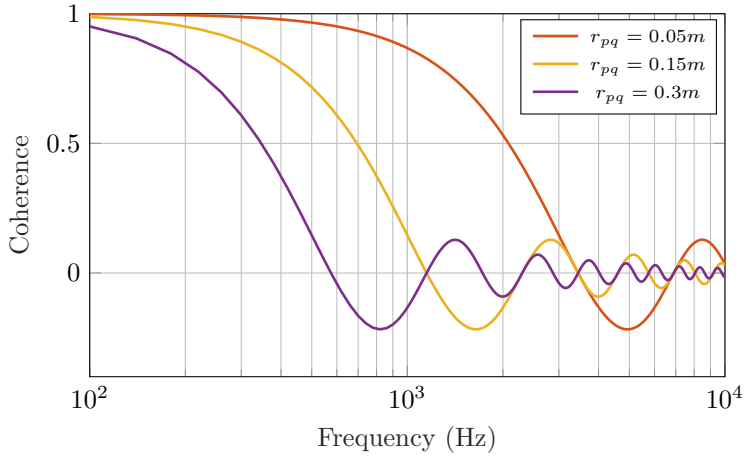


Figure 2.14: Coherence between two microphones within a spherical diffuse field of (2.62) for varying separation distance, r_{pq}. The first zero crossing of each curve occurs at $\frac{r_{pq}}{\lambda} = 0.5$.

related to the gain, phase, or even positions of the individual microphones [20]. The sensitivity of the beamformer to uncorrelated errors among microphones is given by the reciprocal of the WNG, which is a relation than can be used in beamformer design [114].

Optimal beamformers

Continuing with our signal model from (2.52) (without the indexing again), we can finally discuss the design of optimal beamformers that are commonly used in applications (beyond audio) that require spatial processing. These will also serve as the foundation methods which will be built upon throughout the dissertation. One main advantage of the signal model developed thus far is that the M-dimensional microphone signal vector, $\mathbf{y}$, can be considered as an element of an M-dimensional vector space (or more precisely, an inner product space). This facilitates the use of linear algebra concepts such as bases, subspaces, orthogonality, and projection [115] in deriving optimal filters for speech enhancement.

In the following, it will be seen that the optimal beamformers are typically functions of the second-order statistics, i.e. the speech-plus-noise correlation matrix, $\mathbf{R}_{\mathbf{yy}}$, the noise-only correlation matrix, $\mathbf{R}_{\mathbf{nn}}$, and the RTF vector, $\mathbf{h}$. In practice, it should be recalled that these are unknown and the only signal

that we can observe is $\mathbf{y}$. Hence, some sort of estimation for these parameters is also required, which can make use of a priori models and/or the observed data. A further discussion into estimation of the relevant parameters is left for the particular chapters which apply these beamformers, while only a discussion related to the optimisation criteria for design and some properties ensues for when these quantities are known in theory.

Minimum Variance Distortionless Response Beamformer One of the most prominent and still widely used beamformers is the minimum variance distortionless response (MVDR) beamformer. As the name states, the optimisation criterion is such that the PSD of the noise after filtering should be minimised (minimum variance), with a constraint that the target signal should not be subject to any distortion (distortionless response). This type of estimation problem was originally considered in [116] and applied quite extensively in the field of seismic array processing [117]. The formulation which we are familiar with today was derived in [118] and is sometimes referred to as a Capon beamformer in the literature.

The filtering of the signal model from (2.52) can be firstly expressed as:

$$\hat{\mathrm{s}}_{\theta,\mathrm{r}} = \mathbf{w}^H\mathbf{y} = \mathbf{w}^H\mathbf{h}\,\mathrm{s}_r + \mathbf{w}^H\mathbf{n} \tag{2.65}$$

In the absence of noise, it is clear from this equation that when the constraint $\mathbf{w}^H\mathbf{h} = 1$ is imposed, a unity gain is applied to s_r so that it is completely undistorted, i.e. $\hat{\mathrm{s}}_{\theta,\mathrm{r}} = \mathrm{s}_r$. In the presence of noise, however, there will be an additional output noise term, $\mathbf{w}^H\mathbf{n}$, whose PSD or variance is to be minimised.

An alternative interpretation of this design criterion is that $\hat{\mathrm{s}}_{\theta,\mathrm{r}}$ must be a minimum variance unbiased estimate of s_r [20], which implies that $\mathbb{E}\{\hat{\mathrm{s}}_r\} = \mathbb{E}\{\mathbf{w}^H\mathbf{h}\,\mathrm{s}_r\}+\mathbb{E}\{\mathbf{w}^H\mathbf{n}\} = \mathbb{E}\{\mathrm{s}_r\}$. This is then equivalent once again to minimising the output noise PSD with the distortionless constraint. Recalling (2.54), the output noise PSD is:

$$\mathbb{E}\{|\mathbf{w}^H\mathbf{n}|^2\} = \mathbf{w}^H\mathbf{R}_{\mathbf{nn}}\mathbf{w} \tag{2.66}$$

Hence the optimisation problem for the MVDR beamformer becomes:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimise}} \quad & \mathbf{w}^H\mathbf{R}_{\mathbf{nn}}\mathbf{w} \\ \text{subject to} \quad & \mathbf{w}^H\mathbf{h} = 1 \end{aligned} \tag{2.67}$$

Using a Lagrangian multiplier to impose the constraint, the optimal MVDR beamformer is then [20]:

$$\mathbf{w}_{\mathrm{MVDR}} = \frac{\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{h}}{\mathbf{h}^H\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{h}} \tag{2.68}$$

Hence it can be observed that the optimal weights depend on the inverse of the noise-only correlation matrix and the RTF vector. In fact, the MVDR beamformer can be interpreted as a spatial decorrelation process, followed by a matched filter for the target signal [110]. The denominator of (2.68) is a scaling to ensure that a unity gain (distortionless response) will be applied to the target signal in the direction corresponding to $\mathbf{h}$.

Substituting (2.68) into (2.59), along with the distortionless constraint, $\mathbf{w}^H\mathbf{h} = 1$, results in the output SNR of the MVDR beamformer in the direction of the target source as:

$$\mathrm{SNR}_{\mathrm{out}}^{\mathrm{MVDR}} = \sigma_{\mathrm{s_r}}^2 (\mathbf{h}^H \mathbf{R}_{\mathbf{nn}}^{-1} \mathbf{h})^{-1} \tag{2.69}$$

and hence the AG follows as:

$$\mathrm{AG}^{\mathrm{MVDR}} = \sigma_{\mathrm{n_r}}^2 \mathbf{h}^H \mathbf{R}_{\mathbf{nn}}^{-1} \mathbf{h} \tag{2.70}$$

which is now independent of the input SNR.

A close variant on the MVDR beamformer that is also encountered is the minimum power distortionless response (MPDR) beamformer. As opposed to minimising the output noise PSD, the criterion of the MPDR beamformer is to minimise the total output PSD, i.e. the output PSD of both the target source and the noise. This corresponds to a simple replacement of $\mathbf{R}_{\mathbf{nn}}$ with $\mathbf{R}_{\mathbf{yy}}$, and hence the MPDR beamformer is given as:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimise}} \quad & \mathbf{w}^H \mathbf{R}_{\mathbf{yy}} \mathbf{w} \\ \text{subject to} \quad & \mathbf{w}^H \mathbf{h} = 1 \end{aligned} \tag{2.71}$$

Recall that in practice, $\mathbf{h}$ is not known, however, it can be shown that in cases where $\mathbf{h}$ has been accurately estimated, then the MPDR and the MVDR beamformers are equivalent [53]. In general however, it is known that the performance of the MPDR beamformer is significantly compromised when there is mismatch between the estimate and true RTF vector [119].

Matched Filter and Superdirective Beamformers Two special cases of the MVDR filter can be derived by imposing particular noise models for $\mathbf{R}_{\mathbf{nn}}$: (i) the matched filter beamformer (MFB) , and (ii) superdirective beamformer (SDB) . By analysing these particular beamformers, we can understand the inherent trade-off that exists between the directivity of the beamformers and their white noise gain.

For the MFB, it is assumed that $\mathbf{R}_{\mathbf{nn}} = \sigma_{\mathrm{n_r}}^2 \mathbf{I}$, which means that it consists of only spatially white or uncorrelated noise between the sensors with the same

PSD, $\sigma^2_{\mathrm{n_r}}$. Without loss of generality, for the case of $\sigma^2_{\mathrm{n_r}} = 1$, substitution of $\mathbf{R_{nn}} = \mathbf{I}$ into the optimisation criterion of (2.67) is equivalent to a maximisation of the WNG defined in (2.64), and hence the solution is:

$$\mathbf{w}_{\mathrm{MFB}} = \frac{\mathbf{h}}{\mathbf{h}^H\mathbf{h}} \tag{2.72}$$

which is referred to as the matched filter beamformer [53]. Such a beamformer is also classified as a data-independent beamformer as it does not rely on any observed data. On one hand, this suggests some degree of robustness and predictability of the performance of the MFB, but on the other hand, makes it not very effective in a dynamic acoustic scenario.

From (2.63) and (2.64), along with the distortionless constraint, the DI and WNG for the MFB in the direction of the target source is then given as[25]:

$$\mathrm{DI}^{\mathrm{MFB}} = \frac{(\mathbf{h}^H\mathbf{h})^2}{\mathbf{h}^H\mathring{\mathbf{\Gamma}}^{-1}_{\mathbf{nn}}\mathbf{h}} \qquad \mathrm{WNG}^{\mathrm{MFB}} = \mathbf{h}^H\mathbf{h} \tag{2.73}$$

As opposed to the model of $\mathbf{R_{nn}} = \sigma^2_{\mathrm{n_r}}\mathbf{I}$, the SDB considers that the noise field is diffuse as discussed previously, hence $\mathbf{R_{nn}} = \sigma^2_{\mathrm{n_r}}\mathring{\mathbf{\Gamma}}_{\mathbf{nn}}$. W.l.o.g., for $\sigma^2_{\mathrm{n_r}} = 1$, substitution of $\mathbf{R_{nn}} = \mathring{\mathbf{\Gamma}}_{\mathbf{nn}}$ into (2.67) is equivalent to a maximisation now of the directivity of the beamformer defined in (2.63). Hence the SDB is given as:

$$\mathbf{w}_{\mathrm{SDB}} = \frac{\mathring{\mathbf{\Gamma}}^{-1}_{\mathbf{nn}}\mathbf{h}}{\mathbf{h}^H\mathring{\mathbf{\Gamma}}^{-1}_{\mathbf{nn}}\mathbf{h}} \tag{2.74}$$

Accordingly, the DI and WNG in the direction of the target source is:

$$\mathrm{DI}^{\mathrm{SDB}} = \mathbf{h}^H\mathring{\mathbf{\Gamma}}^{-1}_{\mathbf{nn}}\mathbf{h} \qquad \mathrm{WNG}^{\mathrm{SDB}} = \frac{(\mathbf{h}^H\mathring{\mathbf{\Gamma}}^{-1}_{\mathbf{nn}}\mathbf{h})^2}{\mathbf{h}^H\mathring{\mathbf{\Gamma}}^{-1}_{\mathbf{nn}}\mathring{\mathbf{\Gamma}}^{-1}_{\mathbf{nn}}\mathbf{h}} \tag{2.75}$$

Fig. 2.15 illustrates the general trade-off between DI and WNG for the MFB and SDB for a system with $M = 4$, with a microphone spacing of 5 cm. For both beamformers, $\mathbf{h}$ was computed according to the relative steering vector model of $\bar{\boldsymbol{f}}$ from (2.31). In general, it can be observed that at lower frequencies, the SDB is able to reject interfering sources outside of the target source direction due to its increasing directivity in comparison with the MFB. For an end-fire array, it was shown in [120] that the DI for a SDB even tends to M^2. This high directivity, however, comes at the expense of a significantly reduced WNG which

[25] If $\mathbf{h}$ is based on the relative steering vector model of $\bar{\boldsymbol{f}}$ from (2.31) then the MFB is equivalent to the DAS beamformer. In this case the WNG will be equal to the number of microphones, M.

is due to the matrix, $\mathring{\mathbf{\Gamma}}_{\mathbf{nn}}$ becoming rank-deficient. For a spherically diffuse noise field for example, recall from Fig. 2.14 that at lower frequencies, the microphone signals are more correlated and hence can result in a rank-deficient matrix. The MFB on the other hand, while having a poor directivity at lower frequencies, maintains a constant WNG across the entire frequency region.

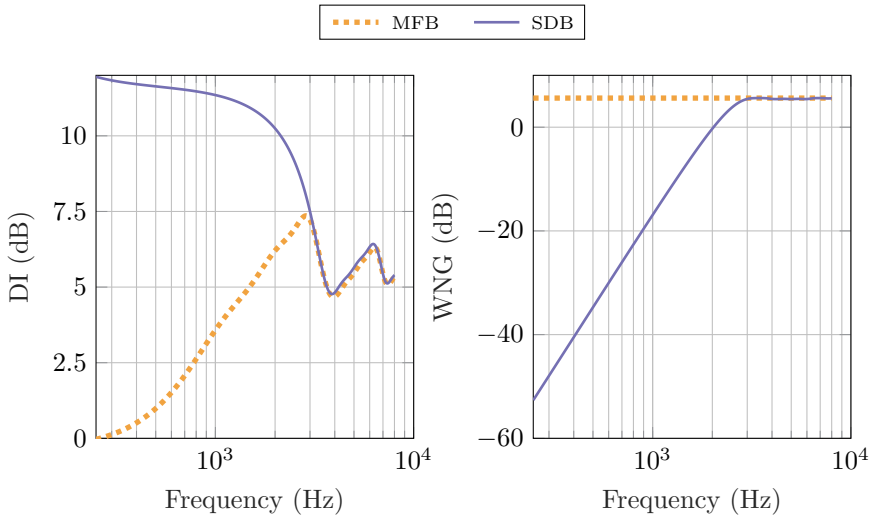


Figure 2.15: DI and WNG performance metrics for the MFB and SDB beamformers.

At higher frequencies, the performance of the SDB eventually converges to the performance of the MFB since $\mathring{\mathbf{\Gamma}}_{\mathbf{nn}} \to \mathbf{I}$ for increasing frequency as can be deduced from (2.62). This can also be seen from Fig. 2.14, as beyond the limit of $\frac{r_{pq}}{\lambda} = 0.5$, the noise field is uncorrelated. As a means to arrive at a compromise between the two extremes of performance exhibited by each of these beamformers, it is common to add a scaled identity matrix to $\mathring{\mathbf{\Gamma}}_{\mathbf{nn}}$ [110] [119].

Linearly constrained minimum variance (LCMV) Beamformer A more general case to the MVDR beamformer is known as the linearly constrained minimum variance (LCMV) beamformer [121], which facilitates a greater degree of control on the response of the beamformer. As opposed to the single distortionless constraint in the MVDR beamformer, the LCMV allows for the imposition of multiple constraints, while still minimising the output noise PSD. These constraints may be distortionless so as to preserve target signals incoming from different directions, for instance different target speaker signals. Additionally, other constraints may be included so as to steer nulls or zeros in

other directions, for instance, to reduce a jammer or noise signal. The latter of these types of constraints would be of the form, $\mathbf{w}^H\mathbf{h_n} = 0$ for steering a null into a specific direction defined by the RTF vector, $\mathbf{h_n}$. Another class of constraints is referred to as derivative constraints [122], which can be used to control the beamwidth of the beamformer. By grouping all of these linear constraints together, the optimisation criterion for the LCMV beamformer can be defined as:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimise}} \quad & \mathbf{w}^H\mathbf{R_{nn}}\mathbf{w} \\ \text{subject to} \quad & \mathbf{H}^H\mathbf{w} = \mathbf{b} \end{aligned} \tag{2.76}$$

where $\mathbf{H}$ is an $M \times N_c$ matrix consisting of N_c RTF vectors which define the constraints, and $\mathbf{b}$ is a vector of length N_c, whose entries are typically ones or zeros depending on the type of constraints.

The corresponding solution to this equality constrained least squares problem can be found by defining the dual function with a Lagrange multiplier vector, $\boldsymbol{\nu}$, and employing the Karush-Kuhn-Tucker (KKT) conditions to solve the following linear set of equations:

$$\begin{bmatrix} \mathbf{R_{nn}} & \mathbf{H} \\ \mathbf{H}^H & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \boldsymbol{\nu} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{b} \end{bmatrix} \tag{2.77}$$

where the left-hand side coefficient matrix is known as the KKT matrix [123]. The result is the well-known LCMV beamformer:

$$\mathbf{w}_{\text{LCMV}} = \mathbf{R}_{\mathbf{nn}}^{-1}\,\mathbf{H}\,[\mathbf{H}^H\,\mathbf{R}_{\mathbf{nn}}^{-1}\,\mathbf{H}]^{-1}\mathbf{b} \tag{2.78}$$

While such a beamformer seems too good to be true, it must be acknowledged that there is a limitation to the maximum allowable constraints such that $N_c \leq M$, i.e. the number of constraints allowed is limited by the number of microphones available in the system. In the limit case, when $N_c = M$, there are no more degrees of freedom left for minimisation of the output noise PSD and hence a data independent beamformer will be obtained [19]. Furthermore it should be noted that the MVDR beamformer from (2.68) is a special case of (2.78) for the case when there is only one distortionless constraint.

Generalised sidelobe canceller (GSC) A practical implementation of the LCMV beamformer was developed in [124], which transforms the constrained optimisation problem of (2.76) to an unconstrained one, and is referred to as the generalised sidelobe canceller (GSC). The first step for deriving the GSC formulation is to split the M dimensional microphone signal space into two subspaces. The first subspace is referred to as the *constraint subspace*, defined by the $M \times N_c$ matrix of constraints, $\mathbf{H}$. The second subspace, referred to as the

orthogonal subspace is defined by the *blocking matrix*, $\mathbf{B}$, i.e. an $M \times (M - N_c)$ matrix, whose columns are defined as a basis for the orthogonal complement of the space spanned by the columns of $\mathbf{H}$. Therefore:

$$\mathbf{H}^H\mathbf{B} = \mathbf{0}_{N_c \times (M-N_c)} \tag{2.79}$$

where the subscript on the right hand side denotes the dimension.

The LCMV beamformer can be split into two orthogonal components as follows:

$$\mathbf{w}_{\text{LCMV}} = \mathbf{w}_{\text{f}} - \mathbf{w}_{\text{b}} \tag{2.80}$$

where $\mathbf{w}_{\text{f}}$ is a projection of $\mathbf{w}_{\text{LCMV}}$ into the constraint subspace, $\mathbf{H}$ and $\mathbf{w}_{\text{b}}$ is the projection of $\mathbf{w}_{\text{LCMV}}$ into the orthogonal subspace, $\mathbf{B}$. Therefore, making use of (2.78), $\mathbf{w}_{\text{f}}$ is:

$$\begin{aligned}\mathbf{w}_{\text{f}} &= \mathbf{H}(\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H\mathbf{w}_{\text{LCMV}} \\ &= \mathbf{H}(\mathbf{H}^H\mathbf{H})^{-1}\mathbf{b}\end{aligned} \tag{2.81}$$

where $\mathbf{H}(\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H$ is a projection matrix. $\mathbf{w}_{\text{f}}$ is also referred to as the quiescent component [20] [125] of $\mathbf{w}_{\text{LCMV}}$ because if $\mathbf{R_{nn}} = \mathbf{I}$, then $\mathbf{w}_{\text{LCMV}} = \mathbf{w}_{\text{f}}$.

The vector, $\mathbf{w}_{\text{b}}$ can be expressed in terms of a linear combination of the columns of $\mathbf{B}$, hence $\mathbf{w}_{\text{b}} = \mathbf{B}\mathbf{w}_{\text{g}}$, where $\mathbf{w}_{\text{g}}$ is an $(M - N_c) \times 1$ vector and the LCMV beamformer can now be expressed as $\mathbf{w}_{\text{LCMV}} = \mathbf{w}_{\text{f}} - \mathbf{B}\mathbf{w}_{\text{g}}$. Since $\mathbf{w}_{\text{f}}$ and $\mathbf{B}$ already satisfies the constraints, which is independent of $\mathbf{w}_{\text{g}}$, the constrained LCMV problem reduces to an unconstrained one [126], which is the GSC optimisation criterion:

$$\underset{\mathbf{w}_{\text{g}}}{\text{minimise}} \quad (\mathbf{w}_{\text{f}} - \mathbf{B}\mathbf{w}_{\text{g}})^H\mathbf{R_{nn}}(\mathbf{w}_{\text{f}} - \mathbf{B}\mathbf{w}_{\text{g}}) \tag{2.82}$$

whose solution is as follows:

$$\mathbf{w}_{\text{g}} = (\mathbf{B}^H\,\mathbf{R_{nn}}\,\mathbf{B})^{-1}\mathbf{B}^H\,\mathbf{R_{nn}}\mathbf{w}_{\text{f}} \tag{2.83}$$

A block depiction of the GSC is shown in Fig. 2.16 which consists of two branches. The upper branch consists of a fixed beamformer, $\mathbf{w}_{\text{f}}$, defined in (2.81). The lower branch consists of the blocking matrix, $\mathbf{B}$, and the unconstrained filter, $\mathbf{w}_{\text{g}}$. Such an implementation is quite desirable as $\mathbf{w}_{\text{g}}$ can also be computed with adaptive filtering techniques such as Least-Mean-Squares (LMS) or Normalised Least Mean Squares (NLMS) [125].

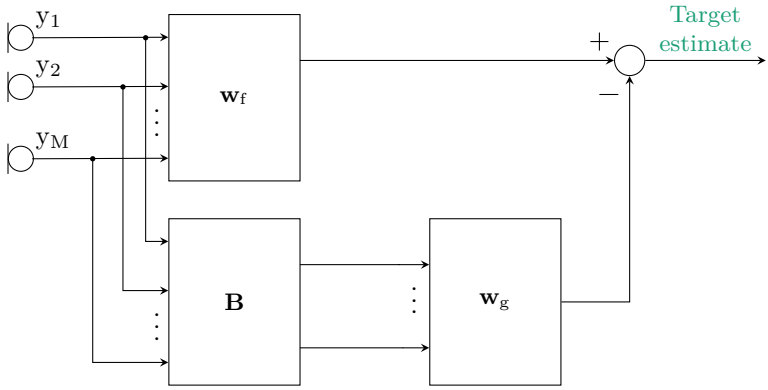


Figure 2.16: Generalised Sidelobe Canceller (GSC). The upper branch consists of a fixed beamformer, $\mathbf{w}_{\mathrm{f}}$, and the lower branch consists of the blocking matrix, $\mathbf{B}$, and the unconstrained filter, $\mathbf{w}_{\mathrm{g}}$. The target source signal estimate is obtained from subtracting the signal at the end of the lower branch from that of the upper branch.

Minimum mean square error (MMSE) estimator The final technique that will be discussed is that of a minimum mean square error (MMSE) estimator, which as its name implies finds the minimum of the mean square error. Returning to the signal model of (2.52), the mean square error can be defined as the square of the difference between the true target signal in the reference microphone, s_r, and the filtered microphone signals, $\mathbf{w}^H\mathbf{y}$. Hence the MMSE optimisation criterion is:

$$\underset{\mathbf{w}}{\text{minimise}} \quad \mathbb{E}\{|\mathrm{s}_r - \mathbf{w}^H\mathbf{y}|^2\} \tag{2.84}$$

whose solution is the *multichannel Wiener filter* (MWF):

$$\begin{aligned}\mathbf{w}_{\mathrm{MWF}} &= \mathbf{R}_{\mathbf{yy}}^{-1}\mathbf{R}_{\mathbf{ys_r}} \\ &= (\mathbf{R}_{\mathbf{xx}} + \mathbf{R}_{\mathbf{nn}})^{-1}\mathbf{R}_{\mathbf{xx}}\mathbf{e}_{\mathrm{r}}\end{aligned}$$

where $\mathbf{R}_{\mathbf{ys_r}} = \mathbb{E}\{\mathbf{y}\mathrm{s}_r^*\}$ and $\mathbf{e}_{\mathrm{r}}$ is an $M \times 1$ vector of all zeros except for a one in the position of the reference microphone in the vector. Recalling the assumption of the speech component being uncorrelated with the noise component and the relation of the correlation matrices from (2.55), the second expression in (2.85) is derived by recognising that $\mathbb{E}\{\mathbf{y}\mathrm{s}_r^*\} = \mathbb{E}\{(\mathbf{x}+\mathbf{n})\mathrm{s}_r^*\} = \sigma_{\mathrm{s_r}}^2\mathbf{h} = \sigma_{\mathrm{s_r}}^2\mathbf{h}\mathbf{h}^H\mathbf{e}_{\mathrm{r}} = \mathbf{R}_{\mathbf{xx}}\mathbf{e}_{\mathrm{r}}$ since $\mathbf{h}^H\mathbf{e}_{\mathrm{r}} = 1$.

In (2.84), the mean square error can be rewritten as follows:

$$\mathbb{E}\{|\mathrm{s}_r - \mathbf{w}^H\mathbf{y}|^2\} = \underbrace{\mathbb{E}\{|\mathrm{s}_r - \mathbf{w}^H\mathbf{x}|^2\}}_{\text{speech distortion}} + \underbrace{\mathbb{E}\{|\mathbf{w}^H\mathbf{n}|^2\}}_{\text{residual noise distortion}} \tag{2.85}$$

consisting of two distinct components, one relating to the resulting speech distortion, and the other related to the residual noise from the filtering process. Since both the target speech signal as well as the noise are included in this mean square error, it means that they both cannot be simultaneously minimised. Consequently, an alternative to (2.84) would be to have individual control over the speech distortion and the residual noise so that they can be optimally traded off for the application of interest. This concept was originally proposed in [127] and the resulting estimator has become to be known as the *speech distortion weighted multichannel Wiener filter* (SDW-MWF) [128] [53]. Hence (2.84) is alternatively posed with a parameter, μ, becomes:

$$\underset{\mathbf{w}}{\text{minimise}} \quad \mathbb{E}\{|\mathrm{s}_r - \mathbf{w}^H\mathbf{y}|^2\} + \mu\mathbb{E}\{|\mathbf{w}^H\mathbf{n}|^2\} \tag{2.86}$$

resulting in the solution for the SDW-MWF:

$$\mathbf{w}_{\text{SDW-MWF}} = (\mathbf{R}_{\mathbf{xx}} + \mu\mathbf{R}_{\mathbf{nn}})^{-1}\mathbf{R}_{\mathbf{xx}}\mathbf{e}_{\mathrm{r}} \tag{2.87}$$

where $\mu \in [0 \ \ \infty)$ is a parameter to trade-off the amount of speech distortion against the amount of noise reduction. Smaller values of μ put more emphasis on reducing the speech distortion, whereas larger values of μ put more emphasis on reducing the residual noise. For the special case when $\mu = 1$, the SDW-MWF is equivalent the MMSE estimator from (2.85). In [127], it was suggested that μ could be tuned in accordance with the SNR of the noisy signal.

Given that a rank-1 assumption from (2.56) has been imposed on $\mathbf{R}_{\mathbf{xx}}$, it is expected that a direct computation of (2.87) can have some numerical instabilities especially for the case as $\mu \to 0$. However, an alternative expression can be derived by firstly substituting (2.56) into (2.87) so that:

$$\mathbf{w}_{\text{SDW-MWF}} = (\mu\mathbf{R}_{\mathbf{nn}} + \sigma^2_{\mathrm{s_r}}\mathbf{h}\mathbf{h}^H)^{-1}\sigma^2_{\mathrm{s_r}}\mathbf{h} \tag{2.88}$$

Making use of the matrix inversion lemma, the inverse term can be re-written as:

$$(\mu\mathbf{R}_{\mathbf{nn}} + \sigma^2_{\mathrm{s_r}}\mathbf{h}\mathbf{h}^H)^{-1} = \frac{1}{\mu}\mathbf{R}^{-1}_{\mathbf{nn}} - \frac{\frac{1}{\mu}\mathbf{R}^{-1}_{\mathbf{nn}}\mathbf{h}\mathbf{h}^H\mathbf{R}^{-1}_{\mathbf{nn}}}{\frac{\mu}{\sigma^2_{\mathrm{s_r}}} + \mathbf{h}^H\mathbf{R}^{-1}_{\mathbf{nn}}\mathbf{h}} \tag{2.89}$$

After substitution back into (2.88) and some rearranging, it can then be shown that the SDW-MWF filter can be decomposed as [129] [53]:

$$\mathbf{w}_{\text{SDW-MWF}} = \underbrace{\frac{\mathbf{R}^{-1}_{\mathbf{nn}}\mathbf{h}}{\mathbf{h}^H\mathbf{R}^{-1}_{\mathbf{nn}}\mathbf{h}}}_{\mathbf{w}_{\text{MVDR}}} \underbrace{\frac{\sigma^2_{\mathrm{s_r}}}{\sigma^2_{\mathrm{s_r}} + \mu(\mathbf{h}^H\mathbf{R}^{-1}_{\mathbf{nn}}\mathbf{h})^{-1}}}_{\text{single-channel post filter}} \tag{2.90}$$

Hence with a rank-1 approximation to $\mathbf{R_{xx}}$, it can be said that the SDW-MWF is a product of an MVDR beamformer followed by a single-channel Wiener type post (spectral) filter. It can also be observed that if $\mu \to \infty$, then the SDW-MWF is simply the MVDR beamformer. Furthermore, recalling (2.69), (2.90) can also be expressed as:

$$\mathbf{w}_{\text{SDW-MWF}} = \underbrace{\frac{\mathbf{R_{nn}^{-1}h}}{\mathbf{h}^H\mathbf{R_{nn}^{-1}h}}}_{\mathbf{w}_{\text{MVDR}}} \underbrace{\frac{\text{SNR}_{\text{out}}^{\text{MVDR}}}{\mu + \text{SNR}_{\text{out}}^{\text{MVDR}}}}_{\text{single-channel post filter}} \tag{2.91}$$

so that the single-channel post filter can be interpreted in terms of the output SNR of the MVDR beamformer. For the design of these post filters, and the SDW-MWF in general, it is clear from (2.90) that estimates are needed for the parameters $\sigma_{\text{s}_\text{r}}^2$, $\mathbf{R_{nn}}$, and $\mathbf{h}$.

2.7.2 Spectral filtering in the time-frequency domain

Upon recalling the speech enhancement framework of Fig. 1.2, it should be noted that the MMSE filter or SDW-MWF previously discussed encompasses both the spatial and spectral filtering blocks. As there is an extensive amount of literature dedicated to the design of spectral filters, however, this section briefly highlights some of the popular methods.

Following the enhancement from any of the spatial filters discussed in the previous section, the spatially enhanced signal can now be modelled as[26]:

$$\hat{\text{s}}_{\theta,\text{r}}(\text{k},\text{l}) = \text{s}_\text{r}(\text{k},\text{l}) + \text{n}_\theta(\text{k},\text{l}) \tag{2.92}$$

which consists of the complex STFT coefficient corresponding to the desired speech component, $\text{s}_\text{r}(\text{k},\text{l})$, and some residual noise, $\text{n}_\theta(\text{k},\text{l})$. In general a particular gain function, $\zeta(\text{k},\text{l})$ can be applied to $\hat{\text{s}}_{\theta,\text{r}}(\text{k},\text{l})$ in order to yield the final enhanced STFT coefficient:

$$\hat{\text{s}}_\text{r}(\text{k},\text{l}) = \zeta(\text{k},\text{l})\ \hat{\text{s}}_{\theta,\text{r}}(\text{k},\text{l}) \tag{2.93}$$

In the following, two classes of spectral filters are discussed. The first is that of a Wiener filter, which estimates the complex speech STFT coefficient, $\text{s}_\text{r}(\text{k},\text{l})$, and is an optimal linear MMSE estimator when the speech and noise STFT coefficients ($\text{s}_\text{r}(\text{k},\text{l})$ and $\text{n}_\theta(\text{k},\text{l})$ respectively) are assumed to be independent,

[26]This model assumes that there has been no distortion from the spatial filter. In practice, however, there will be some distortion as the parameters of the spatial filter have to be estimated. Consequently, the actual complex speech STFT coefficients in this model are not entirely equal to the true speech STFT coefficients.

Gaussian random variables [130]. The second class is a statistical-model-based approach and takes into account the probability density functions (PDFs) of the speech and noise STFT coefficients. These methods commonly estimate the magnitude of the speech STFT coefficient as opposed to its actual complex value and use the phase from the "noisy" complex STFT coefficient, i.e., the phase from $\hat{\mathrm{s}}_{\theta,\mathrm{r}}(\mathrm{k},\mathrm{l})$.

Wiener filtering

The linear MMSE estimator of the complex speech STFT coefficients is such that a linear gain, $\zeta \in [0,1]$ is applied to $\hat{\mathrm{s}}_{\theta,\mathrm{r}}$, which can be found by minimising the mean square error:

$$\underset{\zeta}{\text{minimise}} \quad \mathbb{E}\{|\mathrm{s}_\mathrm{r} - \zeta\, \hat{\mathrm{s}}_{\theta,\mathrm{r}}|^2\} \tag{2.94}$$

Therefore at each time frame, different weights can be applied to different frequency bins according to some particular tuning rule defined by the function ζ. Substitution of (2.92) in (2.94) then results in:

$$\mathbb{E}\{|\mathrm{s}_\mathrm{r} - \zeta\hat{\mathrm{s}}_{\theta,\mathrm{r}}|^2\} = \underbrace{(1-\zeta)^2\sigma^2_{\mathrm{s}_\mathrm{r}}}_{\text{speech distortion}} + \underbrace{\zeta^2\sigma^2_{\mathrm{n}_\theta}}_{\substack{\text{residual noise}\\ \text{distortion}}} \tag{2.95}$$

which consists of a speech distortion term and a residual noise distortion (similar to the MMSE estimator in the multi-channel case) where $\sigma^2_{\mathrm{n}_\theta} = \mathbb{E}\{\mathrm{n}_\theta \mathrm{n}_\theta{}^*\}$ is the residual noise PSD. The solution to (2.94) yields the classic Wiener filter [131, 132]:

$$\zeta_{\mathrm{WF}} = \frac{\sigma^2_{\mathrm{s}_\mathrm{r}}}{\sigma^2_{\mathrm{s}_\mathrm{r}} + \sigma^2_{\mathrm{n}_\theta}} = \frac{\xi}{\xi + 1} \tag{2.96}$$

where $\xi = \sigma^2_{\mathrm{s}_\mathrm{r}}/\sigma^2_{\mathrm{n}_\theta}$ and is referred to as the a priori SNR. Therefore in order to apply the Wiener filter, estimates are needed for the speech and residual noise PSDs or the a priori SNR. A concise review of several methods for estimation these parameters can be found in [21].

Speech distortion weighted Wiener filter Figure 2.17 is a plot of the speech distortion and residual noise distortion as a function of the linear spectral gain for the case when $\sigma^2_{\mathrm{s}_\mathrm{r}} = \sigma^2_{\mathrm{n}_\theta}$. The trade-off[27] is clear with values of $\zeta \to 0$ resulting in less residual noise distortion but more speech distortion and values of $\zeta \to 1$ resulting in more residual noise distortion but less speech distortion. The dotted curve is the sum of both distortions, i.e., the mean square error from (2.95), whose minimum corresponds to the Wiener filter from (2.96).

[27] Depending on the values of $\sigma^2_{\mathrm{s}_\mathrm{r}}$ and $\sigma^2_{\mathrm{n}_\theta}$, these curves may be skewed differently.

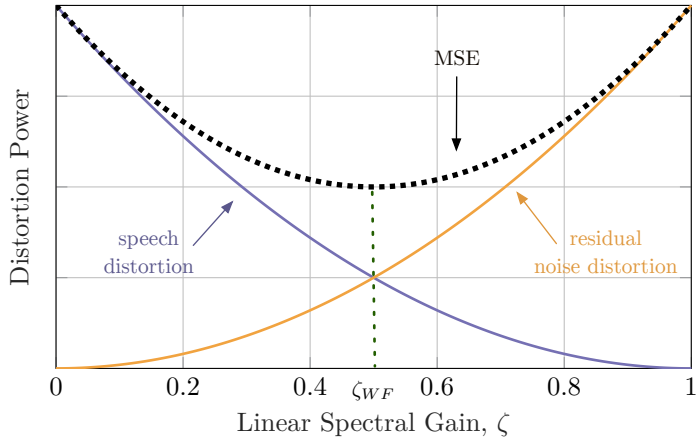


Figure 2.17: Trade-off between the speech distortion and residual noise distortion for the case when $\sigma^2_{\mathrm{s_r}} = \sigma^2_{\mathrm{n}_\theta}$. The dotted curve is the MSE defined in (2.95), whose minimum is given by the value of the Wiener filter from (2.96).

Nevertheless, these curves suggest that either the speech or residual noise distortion can always be constrained to be below some threshold value. For instance, by placing a constraint on the residual noise distortion from (2.95), the following optimisation problem can be formulated [127, 133]:

$$\begin{aligned} &\underset{\zeta}{\text{minimise}} \quad (1-\zeta)^2\sigma^2_{\mathrm{s_r}} \\ &\text{subject to} \quad \zeta^2\sigma^2_{\mathrm{n}_\theta} \leq \delta_{\mathrm{n}} \end{aligned} \tag{2.97}$$

where δ_{n} is the maximum tolerable residual noise distortion. Solving (2.97) with a Lagrangian multiplier, μ, [123], then results in a speech distortion weighted Wiener filter:

$$\zeta_{\mathrm{SDW}} = \frac{\sigma^2_{\mathrm{s_r}}}{\sigma^2_{\mathrm{s_r}} + \mu\sigma^2_{\mathrm{n}_\theta}} = \frac{\xi}{\xi + \mu} \tag{2.98}$$

which is equivalent to the single-channel post filter from (2.90) if $\mathbf{w}_{\mathrm{MVDR}}$ was the spatial filter used to obtain $\hat{\mathrm{s}}_{\theta,\mathrm{r}}$. One of the issues with this spectral filter, however, is that it is not clear how to choose μ in practice. Nevertheless, a reasonable strategy as suggested in [127, 133] is a tuning based on an estimated SNR.

Psychoacoustic spectral weighting. These methods [134, 135] exploit the phenomenon of auditory masking [136] where sounds are perceived to be

inaudible in the presence of others. Auditory masking is also known to be used in speech coding applications [137]. In [135] the residual noise distortion was perceptually shaped in order to render it inaudible while in [134], auditory masking thresholds were used to constrain the residual noise distortion. In [138], the masking thresholds were also used in the context of an SDW-MWF.

Codebooks In (2.92), the PSD of $\hat{\mathrm{s}}_{\theta,\mathrm{r}}$ is equal to the sum of the speech and residual noise PSDs. In a codebook strategy [28], the speech and residual noise PSDs are modelled by autoregressive (AR) processes. Each AR process consists of a prediction gain and a set of linear prediction coefficients. A codebook of AR processes for a set of particular speakers and noises can therefore be trained and applied as a Wiener type gain.

Statistical-model-based filters

Several of the early statistical-model-based filters were based on non-linear estimators of the magnitude of the speech DFT coefficients, where the magnitude is denoted as $\mathrm{a_r} = |\mathrm{s_r}|$. For instance, Mcaulay and Malpass [139] used a a maximum likelihood (ML) framework [130], where $\mathrm{a_r}$ was assumed to be deterministic, but unknown. Ephraim and Malah [140] then proposed to use a Bayesian framework [130], which alternatively models $\mathrm{a_r}$ as a random variable, whose realisation needs to be estimated. Much of the development of statistical-model-based spectral filters has emerged from this approach and therefore it is briefly reviewed in the following.

In the Bayesian framework, a Bayesian MMSE estimator of $\mathrm{a_r}$ can be found as [130]:

$$\begin{aligned}\hat{\mathrm{a}}_\mathrm{r} &= \int \mathrm{a_r}\, p(\mathrm{a_r} \mid \hat{\mathbf{s}}_{\boldsymbol{\theta},\boldsymbol{r}}(\mathrm{l}))\, d\mathrm{a_r} \\ &= \mathbb{E}[\,\mathrm{a_r} \mid \hat{\mathbf{s}}_{\boldsymbol{\theta},\boldsymbol{r}}(\mathrm{l})\,] \end{aligned} \tag{2.99}$$

where $\hat{\mathbf{s}}_{\boldsymbol{\theta},\boldsymbol{r}}(\mathrm{l}) = [\hat{\mathrm{s}}_{\theta,\mathrm{r}}(1,\mathrm{l}), \hat{\mathrm{s}}_{\theta,\mathrm{r}}(2,\mathrm{l}), \ldots \hat{\mathrm{s}}_{\theta,\mathrm{r}}(P,\mathrm{l})]^T$ is the vector of the corresponding PSDs at all frequency bins at a particular time frame, l. The quantity $p(\mathrm{a_r} \mid \hat{\mathbf{s}}_{\boldsymbol{\theta},\boldsymbol{r}}(\mathrm{l}))$ is known as the posterior PDF and summarises the state of knowledge of $\mathrm{a_r}$ after the observation of the data [130]. From (2.99), it is then evident that the Bayesian MMSE estimator for $\mathrm{a_r}$ is given by the mean of this posterior PDF. Applying Bayes' rule, the posterior PDF can also be expressed as:

$$p(\mathrm{a_r} \mid \hat{\mathbf{s}}_{\boldsymbol{\theta},\boldsymbol{r}}(\mathrm{l})) = \frac{p(\hat{\mathbf{s}}_{\boldsymbol{\theta},\boldsymbol{r}}(\mathrm{l}) \mid \mathrm{a_r})p(\mathrm{a_r})}{\int p(\hat{\mathbf{s}}_{\boldsymbol{\theta},\boldsymbol{r}}(\mathrm{l}) \mid \mathrm{a_r})p(\mathrm{a_r})} \tag{2.100}$$

where $p(\mathrm{a_r})$ is the PDF of $\mathrm{a_r}$ and the denominator is a normalising term independent of $\mathrm{a_r}$ so that $p(\mathrm{a_r} \mid \hat{\mathbf{s}}_{\boldsymbol{\theta},\boldsymbol{r}}(\mathrm{l}))$ integrates to one [130]. The posterior PDF expressed in (2.100) demonstrates that the Bayesian framework allows for the imposition of prior knowledge by specifying $p(\mathrm{a_r})$, which highlights the main advantage for this approach. Furthermore, alternative cost functions can be used within the Bayesian framework [141], some of which include psychoacoustic considerations [142, 143].

It should be noted that a Baysian MMSE estimator for the complex speech STFT coefficients can also be derived by finding the mean of the posterior PDF, $p(\mathrm{s_r} \mid \hat{\mathbf{s}}_{\boldsymbol{\theta},\boldsymbol{r}}(\mathrm{l}))$ [144]. From this perspective, it can also be shown that if the PDFs of both $\mathrm{s_r}$ and n_θ are Gaussian, then the Wiener filter from (2.96) is obtained [144].

As opposed to finding the mean of the posterior PDF, other methods [145, 146] have also considered a maximum a posteriori (MAP) approach, where the estimator corresponds to the value which maximises the posterior PDF [130]. Having obtained either the ML, Bayes MMSE, or MAP magnitude estimator, the next task would be to obtain an estimate for the phase of the speech STFT coefficients in order to obtain the final estimate for $\mathrm{s_r}$. In [140], an optimal MMSE phase estimator was found to be that of the noisy phase of $\hat{\mathrm{s}}_{\theta,\mathrm{r}}$. Therefore, a spectral gain function will be of the form:

$$\zeta_{\mathrm{mag}} = \frac{\hat{\mathrm{a}}_{\mathrm{r}}}{|\hat{\mathrm{s}}_{\theta,\mathrm{r}}|} \tag{2.101}$$

where $\hat{\mathrm{a}}_{\mathrm{r}}$ is the corresponding amplitude estimator. In general, these gains are functions of the a priori SNR as previously defined, as well as an a posteriori SNR, which are both parameters that need to be estimated.

While the approach of [140] is quite popular, the underlying assumptions have since been the subject of much discussion. For instance, the modelling of the real and imaginary STFT coefficients as independent Gaussian random variables as in [140] has been debated in [144, 146, 147]. In [144], for example, the prior PDFs of the clean speech DFT coefficients were modelled by either a complex Laplace or by a complex bilateral Gamma density. Ephraim [148] has however stated that the Gaussian or other models are not contradictory since non-Gaussian models can be derived using a conditional Gaussian PDF and an appropriately chosen PDF for the variance of the STFT coefficient. Another assumption is that of the independence of the real and imaginary parts of the DFT coefficients, which has been addressed in [149, 150]. The most recent debate, however, concerns the use of the noisy phase for the estimators. The results from the study in [151] suggest that accurate estimates of the phase could be beneficial and has triggered research into this area. A more complete discussion on the topic and methods for estimating the phase is provided in [152].

2.8 Evaluation of speech enhancement algorithms

After filtering the microphone signals, the final step is to apply the synthesis step to return to the time domain. Once the enhanced signal is obtained, the question is then *how do we quantify the extent to which the speech signal has really been enhanced?* In other words, how to do we evaluate the enhanced speech?

There are two methods which can be used to perform the evaluation. The first is that of *subjective evaluation*, where the enhanced speech signal is aurally presented to a test group of listeners, who then assign a score to some particular attribute of the sound. If a large enough sample of scores is collected, statistics extracted from the scores can be indicative of how well the speech has been enhanced. The other method for evaluation is that of *objective evaluation*, where the enhanced signal is compared mathematically to some reference signal, which is the target or desired signal. In some cases, however, such a reference may not be available, and hence other meaningful information will have to be extracted from the enhanced signal in order to quantify the extent of the enhancement. The advantage of subjective evaluation is that it can correlate with what is physically meaningful, however it comes at the cost of carrying out cumbersome testing procedures to large groups. The objective evaluation, on the other hand can be used for quick performance checks and development of algorithms. However, the objective metrics should be truly representative of a subjective evaluation, for which much research has been devoted to [153–156].

With the choice of either method, the next obstacle to overcome is what attribute of the enhanced signal should be evaluated? Even more generally, are there even attributes that can be sufficiently defined for the evaluation? Two distinct attributes which are commonly discussed in the evaluation of enhanced speech signals are the *quality* and the *intelligibility* [154, 157]. The intelligibility is an attribute that is related to "what" has been said, i.e., the degree to which the words have been understood. The quality, on the other hand, is an attribute related to "how" the signal sounds. Compared to the intelligibility, the quality is more difficult to concretely define as it is an interaction between the listener and the signal [158], which can be highly subjective and can vary from person to person.

Another issue to highlight in the evaluation of the enhanced speech signals is that of normal and impaired hearing listeners. Although a signal may sound acceptable for someone with normal hearing, it may not be the case for someone with impaired hearing. In [154], a number of objective metrics have been reviewed which take this into account.

Throughout this dissertation, objective metrics have been used to evaluate

the performance of the algorithms focusing more on performance in terms of intelligibility rather than quality. The specific ones used are briefly mentioned here to serve as a consolidated reference.

Recall that the speech enhancement algorithms aim to obtain an estimate of the speech signal ("output") in a reference microphone. Therefore this reference microphone signal ("input") and the enhanced speech signal is denoted in the discrete time domain respectively as:

$$y_r[n] = x_r[n] + n_r[n] \qquad \text{Input reference signal} \tag{2.102}$$

$$\hat{s}_r[n] = s_f[n] + n_f[n] \qquad \text{Output enhanced signal} \tag{2.103}$$

where $x_r[n]$ and $n_r[n]$ are the reference speech and noise components respectively, and $s_f[n]$ and $n_f[n]$ are the speech and noise components after filtering respectively. In practice, there is only access to the signal $y_r[n]$. However, if the noisy signals have been constructed from separately scaled speech and noise components (in controlled laboratory experiments for instance), then there is access to $x_r[n]$ and $n_r[n]$. Consequently, the filtering operation can be applied to these signals individually to obtain $s_f[n]$ and $n_f[n]$ for evaluation purposes.

Intrusive metrics

The following metrics are referred to as intrusive metrics as there is access to the signals $x_r[n]$ and $n_r[n]$, which can be used as reference signals for comparison with the enhanced signal.

Unweighted SNR The most basic objective metric that can be used is the SNR of the enhanced signal. As opposed to the SNRs previously considered in the STFT domain, which were narrowband (i.e. for specific frequencies), a broadband SNR can be defined, which takes into consideration all frequencies. Such an SNR can either be frequency-weighted in order to emphasise the influence of certain frequency bands or simply unweighted. Firstly, the unweighted, broadband input and output SNRs can be defined respectively as follows:

$$\mathrm{SNR}_{\mathrm{in}} = 10\log_{10}\left(\frac{\mathbb{E}\{|x_r[n]|^2\}}{\mathbb{E}\{|n_r[n]|^2\}}\right) \tag{2.104}$$

$$\mathrm{SNR}_{\mathrm{out}} = 10\log_{10}\left(\frac{\mathbb{E}\{|s_f[n]|^2\}}{\mathbb{E}\{|n_f[n]|^2\}}\right) \tag{2.105}$$

Speech Intelligibility-weighted SNR (SI-SNR) The SI-SNR was proposed in [159], where the SNR was frequency-weighted according to a set of frequencies important for intelligibility. The SI-SNR can be computed as:

$$\text{SI-SNR} = \sum_j I_j \, \text{SNR}_{j,\text{out}} \tag{2.106}$$

where the band importance function I_j expresses the importance of the j-th one-third octave band with centre frequency, f_j^c for intelligibility, and $\text{SNR}_{j,\text{out}}$ is the output SNR (dB) in the j-th one-third octave band. The centre frequencies, f_j^c and the values for I_j are defined in [160]. The change in SI-SNR from input to output is defined as $\Delta\text{SI-SNR} = \sum_j I_j \, (\text{SNR}_{j,\text{out}} - \text{SNR}_{j,\text{in}})$ where $\text{SNR}_{j,\text{in}}$ is the input SNR (dB) in the j-th one-third octave band.

Frequency-weighted segmental SNR Another frequency-weighted SNR, referred to as the Frequency-weighted segmental SNR (fwSNRseg) [153] applies a frequency-weighting to segmented frames as follows:

$$\text{fwSNRseg} = \frac{10}{Ł} \sum_{l=0}^{Ł-1} \frac{\sum_{j=1}^{F} W(j,l) \log_{10} \frac{X_r(j,l)^2}{(X_r(j,l) - \hat{S}_r(j,l))^2}}{\sum_{j=1}^{F} W(j,l)} \tag{2.107}$$

where $W(j,l)$ is the weight placed on the jth frequency band, F is the number of frequency bands, Ł is the total number of frames, $X_r(j,l)$ is the critical-band magnitude of the clean speech component in the jth frequency band and l^{th} frame, and $\hat{S}_r(j,l)$ is the corresponding spectral magnitude of the enhanced signal in the same band and frame. The weighting can be set such that $W(j,l) = X_r(j,l)^p$, where p is an exponent that can be used to control the emphasis on spectral peaks and/or valleys, and where lower values result in a more uniform weight being applied to all spectral components [153].

Short-time Objective Intelligibility (STOI) One of the most recent popular metrics of for speech intelligibility which has found good correlation to subjective evaluation is the STOI [155]. In one-third octave bands, short-time temporal envelope segments of the speech reference signal and the enhanced signal are compared using a correlation coefficient. Averaging over all frames and frequencies results in a value between 0 and 1, with larger values corresponding to increasing intelligibility.

Cepstral Distance (CD) The CD [161] is a linear predictive coding (LPC) distance measure related to speech quality. It essentially measures a log spectral

distance as follows:

$$\mathrm{CD} = \frac{10}{\ln 10}\sqrt{2\sum_{r=0}^{R}[c_x(r) - \hat{c}_s(r)]^2} \tag{2.108}$$

where $c_x(r)$ and $\hat{c}_s(r)$ are LPC coefficients of $x_r[n]$ and $\hat{s}_r[n]$ respectively, and R is the maximum order of the coefficients.

Non-intrusive metrics

In several practical cases, access to a reference signal may not be possible, therefore a non-intrusive metric needs to be used, i.e., an evaluation using only the enhanced signal has to be done. Furthermore, it can also be ambiguous as to what is the reference signal that should be used with the intrusive metrics. For instance, for a signal recorded in a highly reverberant room, the speech signal recorded in a reference microphone, i.e., the $x_r[n]$ signal in such a case may not be a suitable reference.

Several proposals have been made on non-intrusive metrics [162–165]. The Speech-to-reverberation modulation energy ratio (SRMR) [162, 163] was used for some evaluations in this dissertation. The SRMR, originally developed for a dereverberation context [162], computes the ratio of modulation energy in lower frequency bands to the modulation energy in higher frequency bands. The rationale for this metric was that most of the speech modulation energy was concentrated in lower frequencies whereas this energy became smeared across the frequency spectrum in the presence of reverberation [162]. Recently, this metric has been adapted for assistive hearing devices, and particularly for cochlear implant devices [163].

A final remark

While the objective metrics have provided an efficient way to evaluate the performance, they should not be taken as the absolute truth. Rather they should be interpreted as a guide regarding the potential of the algorithms to perform better or worse as the case may be. The evaluation is not fully complete until a satisfactory degree of subjective testing is done to truly determine the benefit in practice. As a means of supplementing the performance in terms of the objective metrics throughout this dissertation, several accompanying audio samples are provided [95] for listening for a personal subjective evaluation.

2.9 An interpretation of the signal model

We should always be attentive to the fact that all of the previously described optimal beamformers are built upon the narrowband signal model of (2.50). As we have seen from Sec. 2.6, this narrowband signal model is only correct under the assumption that $P >> L_g$. It is therefore worthwhile to ponder for a moment on how this narrowband signal model can be interpreted when the assumption of $P >> L_g$ is broken, which happens quite often in practice.

The multiplication of two DFT sequences corresponds to a circular convolution in the time domain, which is equivalent to a linear convolution followed by a time aliasing [88]. Only when $P >> L_g$ is the effect of the time aliasing approximately removed and hence the narrowband model will be appropriate to represent the discrete-time linear convolution in (2.38) when using the WOLA method. However, when $L_g > P$, the model is not accurate any more, but nevertheless has been used for decades in many speech enhancement tasks with acceptable results. Therefore the question at hand is - for cases when the narrowband assumption is broken, what is (2.49) representative of in terms of what we hear in the time-domain?

In order to investigate this, an audio signal was resynthesized for different frame lengths, P, and impulse response lengths, L_g, according to the narrowband approximation of (2.49), using the same speech signal and truncated versions of the AIR depicted in Fig. 2.9. For a particular L_g, regardless of the frame length, a linear convolution was done between the entire speech sequence and the AIR from Fig. 2.9 truncated to a length L_g. This signal will be denoted as $y_{LC}[n]$[28] and will serve as a reference on the true representation of the signal against which a comparison will be done.

For the generation of signals from the narrowband model of (2.49), for each L_g, a different frame length, P was chosen. Firstly, if $P > L_g$, then the discrete-time domain AIR was padded with zeros up to that frame length P , otherwise, if $P < L_g$, then the AIR was truncated to that frame length, P. Following this, a DFT was taken of the AIR and a STFT of the corresponding discrete-time domain speech signal in order to yield $\mathrm{g(k)}$[29] and $\mathrm{s(k,l)}$ respectively for $\mathrm{k} \in [0, 1, \ldots, P-1]$ and $\mathrm{l} \in [0, 1, \ldots, P-1]$. In each frame, the multiplication operation as in (2.49) was then computed over all frequency bins, yielding a sequence which was subsequently circularly convolved with the DFT of an analysis window in order to replicate the analysis window multiplication done

[28] These synthesized audio signals are now of length, N, as the original speech signal, i.e. $n \in [0, 1, \ldots, N-1]$

[29] the AIR is assumed to be time invariant, hence the removal of the index, l. Since only one microphone is considered the subscript $_m$ is also dropped.

in the discrete-time domain of the WOLA method. The synthesis procedure of the WOLA method was then applied to this recreated STFT sequence in order to obtain the resulting discrete-time domain signal. This discrete-time domain signal will be denoted as $y_{CC}[n]$ and would therefore be representative of what an accurate or inaccurate modelling with respect to the narrowband assumption (i.e. the relationship between P and L_g) would yield.

Fig. 2.18 displays two metrics that were computed from a comparison of $y_{LC}[n]$ for each value of L_g (recall $y_{LC}[n]$ is the same for all frame lengths for the particular L_g) with $y_{CC}[n]$ synthesised for each L_g and each P. The first metric of Fig. 2.18 (left) is a relative error between the two signals computed as:

$$\text{Relative Error} = 20\log_{10}\left(\frac{||y_{CC}[n] - y_{LC}[n]||}{||y_{LC}[n]||}\right) \tag{2.109}$$

From this metric, it is clearly seen that as $P > L_g$, then the relative error between $y_{LC}[n]$ and $y_{CC}[n]$ decreases, indicating that the modelled circular convolution is a reasonable representation to the true linear convolution process. However, this is not the case for $P < L_g$ as there are indeed errors between $y_{LC}[n]$ and $y_{CC}[n]$. This is an expected result from the previous discussion. The additional information that Fig. 2.18 (left) provides would be best understood by observing a particular point, say for instance $L_g = 16384$ and $P = 2048$. In such a case, $P < L_g$, and hence there is a large relative error between $y_{CC}[n]$ and $y_{LC}[n]$ as indicated by the upper-most plot. However, this relative error is much lower for cases when $L_g < 2048$, which suggests that although there is imperfect modelling of the true convolution with a longer AIR, the same model would still be a reasonable one for shorter AIRs.

This can also be observed in Fig. 2.18 (right), which shows the same comparison, but in terms of the STOI metric. Therefore, in cases where $P < L_g$, it can be inferred that one interpretation of the narrowband model from (2.50) is that only a linear convolution with an AIR of length $< P$ will be modelled correctly and all other imperfections will be grouped into the noise term. Practically, depending on the length of L_g in some enclosure, the interpretation can be further extended to just modelling some direct and early components of the speech signal plus additional noise, which will include the late reverberation and any other artifacts from the imperfect modelling.

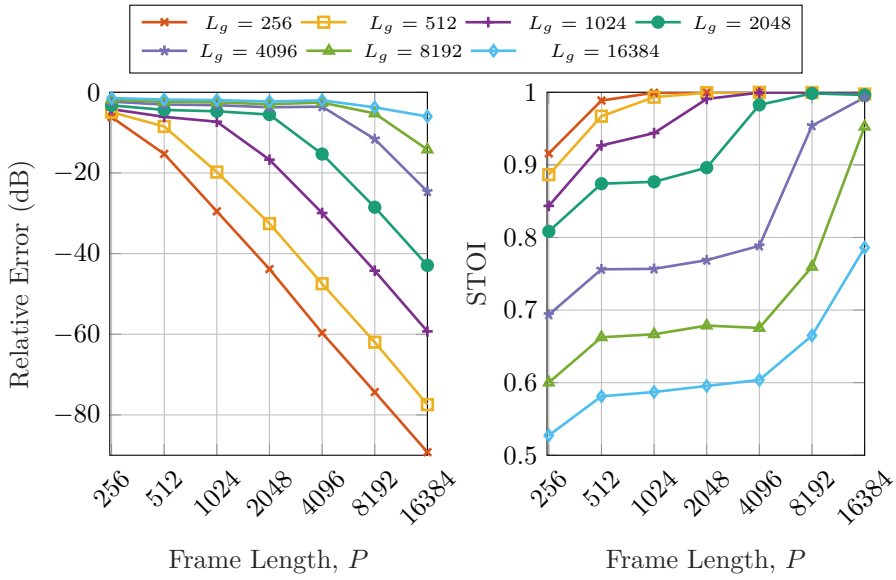


Figure 2.18: Comparison of the speech signal from a linear convolution of a speech signal and an AIR of length L_g, $y_{LC}[n]$ with the signal,$y_{CC}[n]$, synthesised from the narrowband approximation in the STFT domain for AIRs of length L_g and frames of length P. (Left) Relative error between both signals, (right) STOI between both signals. Additional information on the comparison is provided in the text.

Chapter 3

An integrated MVDR beamformer for a microphone array

3.1 Introduction

As discussed in Chapter 2, the single constraint in a minimum variance distortionless response (MVDR) beamformer is in the form of an acoustic transfer function (ATF) vector or more commonly a relative transfer function (RTF) vector, where either of these vectors essentially captures all of the relevant spatial information pertaining to the acoustic scenario. As these constraints are unknown in practice, conventional techniques for estimating them can be grouped into two broad approaches. Either they can be based on a priori knowledge (a priori constraint) or estimated from the observed microphone data (data-dependent constraint). In this chapter[1], as opposed to choosing one of these approaches, the concept of integrating or merging both a priori and data-dependent constraints in an MVDR beamformer is introduced, analysed, and evaluated.

In the a priori approach, the constraint is based on a priori knowledge or assumptions such as microphone characteristics, position, speech source location,

[1]This chapter is partially based on the following publication: Ali R., van Waterschoot T., Moonen M., "Integration of a priori and estimated constraints into an MVDR beamformer for speech enhancement", *IEEE/ACM Trans. on Audio, Speech and Lang. Process.*, vol. 27, no. 12, Dec. 2019, pp. 2288-2300.

and room acoustics (e.g. no reverberation). Therefore, a fixed model is used to characterise the ATF vector from the speech source to the microphone array or RTF vector among all of the microphones. For instance, it is not uncommon in hearing devices to assume knowledge of the speech source location [30–36].

The data-dependent approach on the other hand does not involve any such a priori knowledge and alternatively, the observable microphone data can be used to estimate for an RTF vector [25, 39, 109, 166–169]. For instance, in subspace-based approaches [25, 39], such an estimate will usually depend on the second order statistics of speech and noise signals (i.e. the speech-plus-noise correlation matrix and the noise-only correlation matrix).

Each of the mentioned approaches has its merits and drawbacks and may be more suitable for specific acoustic environments. In the a priori approach, if the speech source is indeed in the direction as defined by the a priori assumptions and the microphone array is properly calibrated, then a robust solution can be obtained. However, this approach may not be so effective when the assumptions are broken, particularly if there is a mismatch between the a priori assumed direction and the actual speech source location [170]. The data-dependent approach compensates for this shortcoming as it is independent of any such a priori assumptions and hence not subject to such degradations. The caveat however is that for a good performance, the observable microphone data must be such that it can guarantee an accurate estimation. For instance, in the subspace-based methods, it is essential that an accurate distinction is made between periods of speech and non-speech activity (for the accurate estimation of the speech-plus-noise correlation matrix and the noise-only correlation matrix). While a fairly accurate distinction can be made in favourable acoustic conditions, in adverse acoustic environments such as for a low signal-to-noise-ratio (SNR) and excessive reverberation, the ability to make such a distinction becomes increasingly difficult [6]. The result is that the MVDR beamformer with an RTF vector data-dependent estimate in this case can result in an undesirable performance.

Therefore the question remains as to how can the mismatch between an assigned RTF vector and the true RTF vector be minimised, or in other words, how can the problem of spatial robustness be addressed? In section 3.3, existing techniques for this problem are reviewed. Solutions have also been proposed in the context of blind source separation [171] for alternative means of estimating the RTF vector without the need for distinguishing between periods of speech and non-speech activity. In the remaining of the chapter, however, the alternative strategy proposed in [172] is considered, where the constraint from a priori knowledge and the constraint from a data dependent estimate were integrated into one cost function by softening these two hard constraints, thereby yielding two tuning parameters. In [172], only a limited analysis was performed on such

a cost function and a solution was presented that did not provide much insight into how such a strategy could be applied in practice.

It is therefore the intention of this chapter to further analyse the potential of the proposed strategy in [172] both theoretically and experimentally. In this chapter, the cost function proposed in [172] is reframed in terms of an a priori RTF vector and a data-dependent RTF vector and will be subsequently referred to as an integrated MVDR beamformer. There are two interpretations that such an integrated approach offers. For an MVDR beamformer that relies solely on a data-dependent RTF vector, the integrated strategy provides a *contingency* option of resorting to the use of an a priori RTF vector in cases where the data-dependent RTF vector is believed to be inaccurate. On the contrary, for an MVDR beamformer that relies solely on a a priori RTF vector, the integrated strategy provides an *opportunity* to use a data-dependent RTF vector in cases where the a priori RTF vector is deemed inaccurate. Throughout this chapter (and the dissertation as a whole), the perspective of a contingency option is taken as there is the general preference for applying the data-dependent methods.

It will be shown in this chapter that this integrated MVDR beamformer can be expressed as a convex combination of the conventional MVDR beamformers, a linearly constrained minimum variance (LCMV) beamformer, and an all-zero vector, with real, positive-valued coefficients. An analysis of how the tuning parameters affect these coefficients both theoretially and through simulations will also demonstrate that the integrated MVDR can encompass a wide range of speech enhancement filters.

With a framework established for an optimal combination of a priori and data-dependent spatial information, the next question would then be what combinations should be used in practice? If there is a metric that can be used to express the confidence in the accuracy of the data-dependent RTF vector, then such a metric can be used to dictate what combinations are made. In other words, the metric can be used to design tuning rules for the integrated MVDR beamformer, where more or less weight is given to the imposition of a particular constraint. Consequently this approach of integrating a priori and data-dependent spatial information can be described as a *confidence-based* one. In this chapter, two such tuning rules are proposed, which make use of a metric of confidence defined by a posterior SNR and the relationship between the integrated MVDR and the speech-distortion-weighted MWF (SDW-MWF).

Finally, an evaluation with recorded data demonstrates that the integrated MVDR beamformer can indeed be more robust in terms of SNR and speech intelligibility for changes in the speech source location and inaccurately estimated correlation matrices in comparison to relying on either of the conventional

MVDR beamformers.

The remaining sections of this chapter are organised as follows. In Section 3.2, the data model and some definitions are restated for convenience. In Section 3.3, a review is provided on existing techniques for addressing the issue of spatial robustness, i.e. robust beamformers, in order to contextualise where the concept of the integrated MVDR beamformer fits into the literature. In Section 3.4, techniques for computing the unknown parameters in an MVDR beamformer are discussed, including the two conventional approaches to prescribe the distortionless constraints of the MVDR beamformer. In Section 3.5, an LCMV strategy using the two conventional approaches for prescribing the distortionless constraints is discussed. In Section 3.6, the integrated approach is introduced and analysed. In Section 3.7, two tuning rules for a practical implementation are proposed. In Section 3.8, the integrated MVDR approach is evaluated with simulated and recorded data from an office room. Some final remarks on the integrated MVDR beamformer are given in Section 3.9.

3.2 Data Model

In the following, the microphone configuration consists of an array of M microphones. Recalling (2.52), in the short-time Fourier transform (STFT) domain, the vector of observed microphone signals at frequency bin, k, and time frame, l, is represented as:

$$\mathbf{y}(\mathrm{k},\mathrm{l}) = \underbrace{\mathbf{h}(\mathrm{k},\mathrm{l})\mathrm{s}_1(\mathrm{k},\mathrm{l})}_{\mathbf{x}(\mathrm{k},\mathrm{l})} + \mathbf{n}(\mathrm{k},\mathrm{l}) \tag{3.1}$$

where (dropping the dependency on k and l for brevity[2]) $\mathbf{y} = [\mathrm{y}_1\ \mathrm{y}_2\ \ldots \mathrm{y}_M]^T$ is a vector containing the respective microphone signals, $\mathbf{x}$ is the speech contribution, represented by s_1, the speech signal in the reference microphone of the array[3], filtered with $\mathbf{h} = [1\ \mathrm{h}_2\ \ldots \mathrm{h}_M]^T$, the RTF vector for the array (with the first microphone used as the reference, which translates to the first component of $\mathbf{h}$ being equal to 1). $\mathbf{n} = [\mathrm{n}_1\ \mathrm{n}_2\ \ldots \mathrm{n}_M]^T$ is the noise contribution. The PSD of the speech signal in the first microphone is also given by $\sigma^2_{\mathrm{s}_1} = \mathbb{E}\{\mathrm{s}_1(\mathrm{k},\mathrm{l})\mathrm{s}_1^*(\mathrm{k},\mathrm{l})\}$.

[2]Although this dependency is usually dropped to simplify reading, it will be sometimes reintroduced to emphasise the dependency on frequency or time frame.

[3]The microphone in the array closest to the source will be used as the reference and will also be referred to as the "first" microphone for convenience. In general however, any microphone may be used as the reference microphone and would be dependent on the application. For instance, in hearing aid devices, a microphone located on the device itself may be more appropriate as a reference microphone as it is closest to the human ear and can be useful to preserve some aspects of the spatial auditory perception of the acoustic environment.

An estimate of the speech component in the first microphone of the array, s_1, is then obtained through the linear filtering of the microphone signals, such that:

$$\mathrm{z}_1 = \mathbf{w}^H \mathbf{y} \tag{3.2}$$

where $\mathbf{w} = [\mathrm{w}_1\ \mathrm{w}_2\ \ldots \mathrm{w}_M]^T$ is the complex-valued filter (beamformer) to be designed. Recall from section 2.7.1 that one such beamformer is the MVDR beamformer, repeated here for convenience (recall $\mathbf{R_{nn}} = \mathbb{E}\{\mathbf{n}\mathbf{n}^H\}$):

$$\mathbf{w}_{\mathrm{MVDR}} = \frac{\mathbf{R_{nn}}^{-1}\mathbf{h}}{\mathbf{h}^H\mathbf{R_{nn}}^{-1}\mathbf{h}} \tag{3.3}$$

3.3 Relation to prior work: robust MVDR beamformers

A formal definition of a robust beamformer is one which attempts to deliver good performance in the presence of mismatch and whose minimum goal is to ensure that its performance should never degrade below the performance of a delay-and-sum beamformer [20]. A multitude of robust beamformers have thus been proposed [173] [174], and it is the purpose of this section to briefly review some of them. Furthermore, this section allows us to better contextualise how the integrated MVDR beamformer discussed throughout the rest of the chapter differs from these existing robust beamforming techniques and where it fits into the literature.

Firstly, as the MVDR parameters, $\mathbf{R_{nn}}$ and $\mathbf{h}$ from (3.3) are unknown in practice, they need to be estimated[4] (see Sec. 3.4). The discussion on robust beamformers typically considers the case when $\mathbf{R_{nn}}$ is estimated by the sample covariance matrix, $\hat{\mathbf{R}}_{\mathbf{sc}} = \frac{1}{L}\sum_{l=1}^{L}\mathbf{y}(\mathrm{l})\mathbf{y}(\mathrm{l})^H$, where L is some finite number of observed time frames. Denoting any estimate of $\mathbf{h}$ as $\mathbf{h_s}$[5], the MVDR beamformer from (3.3) now becomes:

$$\mathbf{w}_{\mathrm{SCM}} = \frac{\hat{\mathbf{R}}_{\mathbf{sc}}^{-1}\mathbf{h_s}}{\mathbf{h_s}^H\hat{\mathbf{R}}_{\mathbf{sc}}^{-1}\mathbf{h_s}} \tag{3.4}$$

When either $\hat{\mathbf{R}}_{\mathbf{sc}} \neq \mathbf{R_{nn}}$ (due to finite observations for instance) or $\mathbf{h_s} \neq \mathbf{h}$ (due to incorrect a priori knowledge for instance), a mismatch is said to occur, which can result in severe performance degradation [20, 119, 175].

[4] In relation to the main block diagram from Fig. 1.2, the estimation of $\mathbf{R_{nn}}$ and $\mathbf{h}$ corresponds to the parameter estimation block.

[5] The $\hat{(\cdot)}$ notation is left for the specific case when $\mathbf{h}$ is estimated from observed data, see Sec. 3.4.

An early remedy to this problem was the approach of diagonal loading of the estimated correlation matrix [114, 176, 177]. Since the white noise gain (WNG) is related to a measure of robustness to uncorrelated errors [114], a constraint could be imposed such that it does not decrease below a minimum level. Recalling from (2.64) that WNG is proportional to the inverse of the squared-norm of the filter coefficients, then the MVDR optimisation problem can be regularised [178] with a quadratic penalty term as follows:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimise}} \quad & \mathbf{w}^H \hat{\mathbf{R}}_{\mathbf{sc}} \mathbf{w} + \gamma ||\mathbf{w}||^2 \\ \text{subject to} \quad & \mathbf{w}^H \mathbf{h}_{\mathbf{s}} = 1 \end{aligned} \tag{3.5}$$

resulting in the diagonally-loaded robust beamformer:

$$\mathbf{w}_{\text{DL}} = \frac{(\hat{\mathbf{R}}_{\mathbf{sc}} + \gamma \mathbf{I})^{-1} \mathbf{h}_{\mathbf{s}}}{\mathbf{h}_{\mathbf{s}}^H (\hat{\mathbf{R}}_{\mathbf{sc}} + \gamma \mathbf{I})^{-1} \mathbf{h}_{\mathbf{s}}} \tag{3.6}$$

where the parameter, γ is referred to as the diagonal loading factor. This beamformer is simply that of the MVDR beamformer, except that a scaled diagonal matrix has been added to the correlation matrix. This diagonally-loaded beamformer improves the original MVDR beamformer performance, but it is unclear as to how γ should be chosen in practice. For this reason, such an approach is sometimes referred to as an ad-hoc one [173]. Furthermore, beamformers of this nature are referred to as being numerically robust, i.e., robust to the sensitivity of the array gain to mismatches in the microphone gains and phases and the beamformer weights [5].

An alternative approach to robust beamforming considers the projection of $\mathbf{h}_{\mathbf{s}}$ onto the subspace spanned by the significant set of eigenvectors of $\hat{\mathbf{R}}_{\mathbf{sc}}$ [175]. Consider the set of orthonormal eigenvectors of $\hat{\mathbf{R}}_{\mathbf{sc}}$ that have been ordered according to decreasing eigenvalues:

$$\mathbf{E} = [\mathbf{E}_{\mathbf{s}} \;\; \mathbf{E}_{\mathbf{n}}] \tag{3.7}$$

where $\mathbf{E}_{\mathbf{s}}$ represent the set of significant eigenvectors corresponding to the largest eigenvalues and $\mathbf{E}_{\mathbf{n}}$ corresponding to the set of eigenvectors corresponding to the smallest eigenvalues and hence those that span the noise subspace. Therefore a projection of $\mathbf{h}_{\mathbf{s}}$ into the subspace spanned by $\mathbf{E}_{\mathbf{s}}$ is $\mathbf{h}_{\mathbf{sub}} = \mathbf{E}_{\mathbf{s}} \mathbf{E}_{\mathbf{s}}{}^H \mathbf{h}_{\mathbf{s}}$, so that the projection-based robust beamformer is given by:

$$\mathbf{w}_{\text{PROJ}} = \frac{\hat{\mathbf{R}}_{\mathbf{sc}}^{-1} \mathbf{h}_{\mathbf{sub}}}{\mathbf{h}_{\mathbf{sub}}^H \hat{\mathbf{R}}_{\mathbf{sc}}^{-1} \mathbf{h}_{\mathbf{sub}}} \tag{3.8}$$

This beamformer results in an improved performance in comparison to that of (3.4), but however may still suffer some degradation in cases of low SNR as the subspaces required for the projection may be swapped [174].

Another type of robust beamformers consider the case of a moving undesired interfering source, and propose beamformers that broaden the beam pattern nulls or notches toward these interfering sources. In [179], it is proposed to impose data-driven derivative constraints, which corresponds to adding another matrix to $\hat{\mathbf{R}}_{\mathbf{sc}}$ (see [179] for details) with a parameter that allows for controlling the trade-off between the null depth and null width.

With the same intent of broadening the beam pattern nulls, alternative methods [180–182] propose that $\hat{\mathbf{R}}_{\mathbf{sc}}$ is replaced with a tapered version such that:

$$\hat{\mathbf{R}}_{\text{tap}} = \hat{\mathbf{R}}_{\mathbf{sc}} \odot \hat{\mathbf{T}} \tag{3.9}$$

where $\odot$ is the Hadamard (i.e. element-wise) product and $\hat{\mathbf{T}}$ is a positive semi-definite matrix of taper coefficients, which may be for instance defined by an appropriate set of sinc functions for a desired null shape [182].

A different class of techniques considers a less ad-hoc approach and alternatively modifies the MVDR optimisation problem to include a worst-case distortionless constraint that needs to be maintained [183]. The uncertainty of the prescribed distortionless constraint, $\mathbf{h_s}$, is defined as $\delta = \mathbf{h} - \mathbf{h_s}$ for $||\delta|| \leq \varepsilon$, where ε is an a priori known bound, so that the uncertainty region can be considered as spherical. The modified MVDR optimisation problem is then:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimise}} \quad & \mathbf{w}^H \hat{\mathbf{R}}_{\mathbf{sc}} \mathbf{w} \\ \text{subject to} \quad & |\mathbf{w}^H(\mathbf{h_s} + \delta)| \geq 1 \quad \forall\, ||\delta|| \leq \varepsilon \end{aligned} \tag{3.10}$$

which states that the distortionless response is maintained by the inequality constraints which take into account all of the possible mismatched distortionless constraints within the spherical uncertainty region defined by δ and ε. This is in direct contrast to the single constraint that was used in the MVDR problem of (2.67) and hence it can already be seen that this worst-case robust beamformer has the potential to have an improved performance to cases of mismatch once the true distortionless constraint is within in the uncertainty region. However, as $|\mathbf{w}^H(\mathbf{h_s} + \delta)| \geq 1$ is a nonlinear and nonconvex constraint on $\mathbf{w}$ and there are infinite such constraints, (3.10) is a nonconvex quadratic program [183]. Nevertheless, it was shown in [183] that (3.10) can be reformulated as a second-order cone program (SOCP) as follows, which is now convex and can be solved by interior point methods [184, 185]:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimise}} \quad & \mathbf{w}^H \hat{\mathbf{R}}_{\mathbf{sc}} \mathbf{w} \\ \text{subject to} \quad & \mathbf{w}^H \mathbf{h_s} \geq \varepsilon ||\mathbf{w}|| + 1 \end{aligned} \tag{3.11}$$

It was further shown in [183] that the solution to (3.11) also belongs to the class of diagonally-loaded robust beamformers, where the optimal value of the

diagonal loading is related to ε and the Lagrangian multiplier. As it is unclear how to compute the Lagrangian multiplier in closed form, the interior point methods are preferred for finding a solution.

Several extensions to this approach have also been proposed. For instance, in [186], the non-stationarity of interfering signals was taken into consideration by adding uncertainty into a matrix of observed data. In [187], an elliptical uncertainty set for the worst-case distortionless response constraint was considered as opposed to the spherical set. A formulation involving a general-rank signal model was also proposed in [188]. This technique was also framed as a quadratically constrained quadratic problem (QCQP) in [34, 36, 189] incorporating priori knowledge of multiple ATFs.

A more recent technique for robust beamforming takes a statistical perspective. In [190], von Mises and von Mises-Fisher distributions were used to model the the variation in direction of arrival of a source. Hence, different distributions accommodating for more or less uncertainty around a particular direction of arrival can be assigned. Such a model could then be used to obtain the relevant correlation matrices in order to design the beamformers, which has shown to be robust.

The remaining work presented in this chapter considers an alternative aspect related to the robustness of beamformers, which is in terms of the underlying spatial information used to obtain estimates for the RTF vector. As will be detailed in the following section, data-dependent methods for estimating the correlation matrices and the RTF vector can result in robust MVDR beamformers when acoustic conditions are favourable. Additionally, a data-dependent RTF vector will take into consideration room acoustics, which can be desirable for certain applications. However, as already mentioned, in poor acoustic conditions, the data-dependent RTF vector estimate becomes less accurate and the performance of the MVDR beamformer degrades. Even if the worst-case performance optimisation approach were to be used in this situation, performance may still not improve if ε is not large enough. In response to such a precarious scenario, an approach may be to completely abandon a data-dependent method and simply use a priori knowledge to obtain the RTF vector, such as a far-field model or a pre-measured impulse response. As opposed to this binary option, the work that follows is an investigation on how it may be possible to integrate or merge the use of both an a priori and a data-dependent RTF vector. In relation to the worst case approach, instead of having regions of uncertainty around a particular constraint, the integrated approach uses two constraints, each of which makes use of different spatial information and may not necessarily lie in the same region of uncertainty as would have been specified by ε.

3.4 Parameter estimation for the MVDR beamformer

Before developing the integrated MVDR beamformer approach, in this section some more recent techniques are firstly presented for approximating or estimating the unknown MVDR parameters, $\mathbf{R_{nn}}$ and $\mathbf{h}$, from (3.3) that can be applied in practice.

3.4.1 Estimation of $\mathrm{R_{nn}}$

In Section 3.3, one option to estimate $\mathbf{R_{nn}}$ was that of the sample covariance matrix. In this section however, some alternative methods that in general, perform better, are presented. One popular method of computing $\mathbf{R_{nn}}$ is by recursive recursive averaging, along with a voice activity detector (VAD) [191] or a speech presence probability (SPP) estimator [192]. Using either the VAD or SPP, a threshold can be defined above which, speech is regarded to be present. Consequently, an estimator for $\mathbf{R_{nn}}$[6] can be computed as:

$$\hat{\mathbf{R}}_{\mathbf{nn}}^{\mathbf{ht}}(\mathrm{k},\mathrm{l}) = \begin{cases} \lambda \hat{\mathbf{R}}_{\mathbf{nn}}^{\mathbf{ht}}(\mathrm{k},\mathrm{l}-1) + (1-\lambda)\mathbf{y}(\mathrm{k},\mathrm{l})\mathbf{y}^H(\mathrm{k},\mathrm{l}) & \text{speech absent} \\ \hat{\mathbf{R}}_{\mathbf{nn}}^{\mathbf{ht}}(\mathrm{k},\mathrm{l}-1) & \text{speech present} \end{cases} \tag{3.12}$$

where the superscript, $\{.\}^{\mathbf{ht}}$ refers to 'hard threshold', $\lambda \in [0\ 1]$ is commonly referred to as a *forgetting factor* and controls how fast $\hat{\mathbf{R}}_{\mathbf{nn}}$ is updated. The update equation when speech is absent also corresponds to an exponential weighting [125] of the outer product term over all time frames as follows:

$$\hat{\mathbf{R}}_{\mathbf{nn}}^{\mathbf{ht}}(\mathrm{k},\mathrm{l}) = (1-\lambda)\sum_{j=0}^{\infty} \lambda^j \mathbf{y}(\mathrm{k},\mathrm{l}-j)\mathbf{y}^H(\mathrm{k},\mathrm{l}-j) \quad \text{speech absent} \tag{3.13}$$

so that the largest weighting applied to the current time frame and exponentially decreasing weightings are applied to the previous frames. Hence a relationship exists between λ and a time constant, τ, where τ would correspond to the time at which the exponential weight decreases to $1/e$ ($\approx 37\%$) of its initial value. So, for a stationary signal with some mean of the outer product, the previous τ seconds of data would be used to make up approximately 63% of the mean, and 5τ seconds of data would be used to make up approximately 99% of the mean. The relationship between the forgetting factor and τ in terms of blocks

[6] Note that $\mathbf{R_{yy}}$ can be estimated in a similar manner, but without distinguishing periods of speech absence and speech presence in (3.12)

of samples is given by [193] [194]:

$$\tau = \frac{-R}{f_s \ln(\lambda)} \approx \frac{1}{1-\lambda}\frac{R}{f_s} \tag{3.14}$$

where the second expression is an approximation from the first-order Taylor series, $\ln(\lambda) \approx (\lambda - 1)$. It is to be recalled that R is the hop-size[7] in samples and f_s is the sampling frequency (Hz). From (3.14), it can be now be observed that smaller values of λ would correspond to smaller time constants, and hence a faster updating of the correlation matrices, which may be more convenient for highly non-stationary environments.

As opposed to the hard thresholding, a soft threshold can also be used where the SPP is directly incorporated into the forgetting factor. Hence the soft threshold estimator of $\mathbf{R_{nn}}$ is:

$$\hat{\mathbf{R}}^{\mathbf{st}}_{\mathbf{nn}}(\mathrm{k},\mathrm{l}) = \bar{\lambda}(\mathrm{k},\mathrm{l})\hat{\mathbf{R}}^{\mathbf{st}}_{\mathbf{nn}}(\mathrm{k},\mathrm{l}-1) + (1-\bar{\lambda}(\mathrm{k},\mathrm{l}))\mathbf{y}(\mathrm{k},\mathrm{l})\mathbf{y}^H(\mathrm{k},\mathrm{l}) \tag{3.15}$$

where the superscript, $\{.\}^{\mathbf{st}}$ refers to 'soft threshold', $\bar{\lambda}(\mathrm{k},\mathrm{l})$ is a time-frequency dependent forgetting factor defined as [23][8]:

$$\bar{\lambda}(\mathrm{k},\mathrm{l}) = \lambda + (1-\lambda)P(\mathcal{H}_1|\mathrm{y}_1)(\mathrm{k},\mathrm{l}) \tag{3.16}$$

where $P(\mathcal{H}_1|\mathrm{y}_1)(\mathrm{k},\mathrm{l})$[9] is the probability that speech is present in the reference microphone for each time-frequency segment.

Another technique that can be used to estimate $\mathbf{R_{nn}}$ is by a fixed spatial coherence matrix, $\mathbf{\Gamma_{nn}}$, scaled with an estimated noise PSD:

$$\hat{\mathbf{R}}^{\mathbf{\Gamma}}_{\mathbf{nn}}(\mathrm{k},\mathrm{l}) = \hat{\sigma}^2_{\mathrm{n}}(\mathrm{k},\mathrm{l})\mathbf{\Gamma_{nn}} \tag{3.17}$$

where $\hat{\sigma}^2_{\mathrm{n}}(\mathrm{k},\mathrm{l})$ is an estimate of the true the noise PSD, $\sigma^2_{\mathrm{n}}(\mathrm{k},\mathrm{l})$ in each time-frequency segment. $\mathbf{\Gamma_{nn}}$ can be computed according to a pre-defined diffuse noise field, such as a spherical diffuse noise field from (2.62). For the estimation of the noise power, an extensive variety of methods have been proposed in the literature [21, 23, 24, 195].

Denoting any estimate of $\mathbf{R_{nn}}$ as $\hat{\mathbf{R}}_{\mathbf{nn}}$ using any of the discussed techniques, the MVDR beamformer now becomes:

$$\mathbf{w}_{\mathrm{MVDR}} = \frac{\hat{\mathbf{R}}^{-1}_{\mathbf{nn}}\mathbf{h}}{\mathbf{h}^H\hat{\mathbf{R}}^{-1}_{\mathbf{nn}}\mathbf{h}} \tag{3.18}$$

[7] In [194], the size of the DFT window is used in (3.14). It has been modified here to include the overlap introduced by the hop-size.

[8] This was developed within the context of single-channel speech enhancement to obtain a noise PSD estimate, but may also be used as mentioned here for obtaining the estimate of the noise-correlation matrix.

[9] The first (reference) microphone is used here for convenience and any other microphone may be used if so desired.

3.4.2 Estimation of h

Two conventional approaches are usually taken for the estimation of $\mathbf{h}$, one of which relies on a priori knowledge, and the other on the real-time, observed microphone data. These two methods are discussed in the following and will be eventually merged for the integrated MVDR beamformer.

A priori knowledge (MVDR-APR)

One method of estimating $\mathbf{h}$ is by the imposition of a priori knowledge. As mentioned earlier, this a priori knowledge may be based on a priori assumptions such as microphone characteristics, position, speaker location, and room acoustics (e.g. no reverberation). For instance, if the angle of arrival, θ, of the desired speaker and the inter-microphone spacing, d, of a microphone array is known, then this spatial information can be used to estimate an RTF vector based on a far field model: $\widetilde{\mathbf{h}}^{\mathrm{ff}} = [1\ e^{-j\frac{\omega d \cos\theta}{c}}\ \ldots e^{-j\frac{\omega(M-1)d\cos\theta}{c}}]^T$ where c is the speed of sound. Alternatively, a priori knowledge can be in the form of a set of pre-measured impulse responses between a desired speaker from different angles of arrival and the microphones on a particular device [34, 189]. If the angle of arrival is assumed to be known, then one of these impulse responses can be used to obtain an priori RTF vector as opposed to the far field model. Denoting the an a priori RTF vector in general as $\widetilde{\mathbf{h}} = [1\ \widetilde{\mathrm{h}}_2\ \ldots \widetilde{\mathrm{h}}_{\mathrm{M}}]^T$, the optimal MVDR beamformer is then given by:

$$\widetilde{\mathbf{w}} = \frac{\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\widetilde{\mathbf{h}}}{\widetilde{\mathbf{h}}^H\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\widetilde{\mathbf{h}}} \tag{3.19}$$

which will be referred to as the MVDR-APR. The speech estimate in the first microphone of the array is subsequently calculated as:

$$\widetilde{\mathrm{z}}_1 = \widetilde{\mathbf{w}}^H\mathbf{y} \tag{3.20}$$

Parameters denoted as $(\widetilde{\cdot})$ will be associated with parameters that make use of a priori knowledge.

In the case where the assumptions are satisfied, for instance if the speech source truly lies in the direction defined by $\widetilde{\mathbf{h}}$, the MVDR-APR can be regarded as robust in adverse acoustic conditions. In other scenarios, however, it can be expected that the speech signal does not always adhere to the assumptions, and hence the MVDR-APR will degrade in performance and potentially suffer from distortions.

Data-dependent (MVDR-EST)

In order to accommodate for the shortcomings of defining an RTF vector based on a priori knowledge, the alternative is to make direct use of the observable microphone data. One popular class of methods uses a subspace-based approach, where the RTF vector is estimated from the speech-plus-noise correlation matrix and noise-only correlation matrix. In particular, the method of covariance whitening or equivalently that which involves a Generalised Eigenvalue Decomposition (GEVD) has demonstrated promising performance [25, 39, 41]. As this will be the data-dependent method used throughout the chapter, it is briefly reviewed in the following. While there is a preference for this method, it should be noted that the integrated MVDR beamformer to be developed is not dependent on this specific method. In fact, any data-dependent estimate of an RTF vector can be used in the integrated MVDR beamformer.

The GEVD of the matrix pencil $(\hat{\mathbf{R}}_{\mathbf{yy}}, \hat{\mathbf{R}}_{\mathbf{nn}})$ is defined by [196]:

$$\hat{\mathbf{R}}_{\mathbf{yy}}\hat{\mathbf{U}} = \hat{\mathbf{R}}_{\mathbf{nn}}\hat{\mathbf{U}}\hat{\mathbf{\Sigma}} \tag{3.21}$$

where $\hat{\mathbf{R}}_{\mathbf{yy}}$ is an estimate of $\mathbf{R}_{\mathbf{yy}}$, $\hat{\mathbf{\Sigma}} = \text{diag}\{\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_{\text{M}}\}$, and $\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_{\text{M}}$ are generalised eigenvalues (GEVLs) arranged in descending order so that $\hat{\sigma}_1$ is the largest GEVL, referred to as the principal GEVL. The matrix, $\hat{\mathbf{U}} = [\mathbf{u}_1\ \mathbf{u}_2\ \ldots\ \mathbf{u}_{\text{M}}]$, where $\mathbf{u}_1\ \mathbf{u}_2\ \ldots\ \mathbf{u}_{\text{M}}$ are the corresponding generalised eigenvectors (GEVCs) to the GEVLs, and hence $\mathbf{u}_1$ will be the principal GEVC. Since $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$ are hermitian-symmetric and $\hat{\mathbf{R}}_{\mathbf{nn}}$ is assumed to be positive-definite, it follows that the GEVLs are all real and the GEVCs, $\mathbf{u}_i$ and $\mathbf{u}_j$ are $\hat{\mathbf{R}}_{\mathbf{nn}}$-orthogonal if $\hat{\sigma}_i \neq \hat{\sigma}_j$ [197], meaning that:

$$\mathbf{u}_i^H\hat{\mathbf{R}}_{\mathbf{nn}}\mathbf{u}_j = 0 \tag{3.22}$$

If $\hat{\sigma}_i = \hat{\sigma}_j$ however, then $\mathbf{u}_i$ and $\mathbf{u}_j$ may be chosen to be $\hat{\mathbf{R}}_{\mathbf{nn}}$-orthogonal [197]. Therefore, it follows that the GEVCs can be normalised as follows:

$$\hat{\mathbf{U}}^H\hat{\mathbf{R}}_{\mathbf{nn}}\hat{\mathbf{U}} = \mathbf{I}_M \tag{3.23}$$

The GEVD of this matrix pencil $(\hat{\mathbf{R}}_{\mathbf{yy}}, \hat{\mathbf{R}}_{\mathbf{nn}})$, is also equivalent to a joint diagonalisation of $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$ such that:

$$\hat{\mathbf{R}}_{\mathbf{yy}} = \hat{\mathbf{Q}}\hat{\mathbf{\Sigma}}_{\mathbf{y}}\hat{\mathbf{Q}}^H; \quad \hat{\mathbf{R}}_{\mathbf{nn}} = \hat{\mathbf{Q}}\mathbf{I}_M\hat{\mathbf{Q}}^H \tag{3.24}$$

where $\hat{\mathbf{Q}}$ is a full rank $M \times M$ matrix (not necessarily unitary). Substitution of (3.24) in (3.21) then reveals that $\hat{\mathbf{\Sigma}}_{\mathbf{y}} = \hat{\mathbf{\Sigma}}$ and $\hat{\mathbf{Q}} = \hat{\mathbf{U}}^{-H}$.

Recalling the assumption from (2.55) that $\mathbf{R_{yy}} = \mathbf{R_{xx}} + \mathbf{R_{nn}}$, an estimate for the speech-only correlation matrix, $\mathbf{R_{xx}}$, then follows as:

$$\hat{\mathbf{R}}_{\mathbf{xx}} = \hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{nn}} = \hat{\mathbf{Q}} \underbrace{(\hat{\mathbf{\Sigma}} - \mathbf{I}_M)}_{\hat{\mathbf{\Sigma}}_{\mathbf{x}}} \hat{\mathbf{Q}}^H \tag{3.25}$$

with the definition of the elements in $\hat{\mathbf{\Sigma}}_{\mathbf{x}}$ as $\hat{\mathbf{\Sigma}}_{\mathbf{x}} = \text{diag}\{\hat{\sigma}_{\text{x1}}, \hat{\sigma}_{\text{x2}} \dots \hat{\sigma}_{\text{xM}}\}$. A rank-1 approximation of $\mathbf{R_{xx}}$ would therefore be given as:

$$\begin{aligned} \hat{\mathbf{R}}_{\mathbf{xx,r1}} &= \hat{\sigma}_{\text{x1}} \hat{\mathbf{Q}} \mathbf{e}_1 (\hat{\mathbf{Q}} \mathbf{e}_1)^H \\ &= \underbrace{|\mathbf{e}_1^T \hat{\mathbf{Q}} \mathbf{e}_1|^2 \hat{\sigma}_{\text{x1}}}_{\text{estimated speech PSD}} \hat{\mathbf{h}} \hat{\mathbf{h}}^H \end{aligned} \tag{3.26}$$

where, $\mathbf{e}_1 = [1\ 0 \dots 0]^T$, and $\hat{\mathbf{h}}$ is the data-dependent RTF vector estimator given as:

$$\hat{\mathbf{h}} = \frac{\hat{\mathbf{Q}} \mathbf{e}_1}{\mathbf{e}_1^T \hat{\mathbf{Q}} \mathbf{e}_1} \tag{3.27}$$

For infinite numerical precision and when $\mathbf{R_{nn}}$ is invertible and well-conditioned, the GEVD problem of (3.21) can also be equivalently viewed as three different eigenvalue decompositions (EVDs), all of which yield the same data-dependent RTF vector estimator of (3.27).

1. Spatial pre-whitening

   A spatial pre-whitening operation[10] is firstly defined from $\hat{\mathbf{R}}_{\mathbf{nn}}$ using a Cholesky decomposition as follows:

   $$\hat{\mathbf{R}}_{\mathbf{nn}} = \hat{\mathbf{R}}_{\mathbf{nn}}^{1/2} \hat{\mathbf{R}}_{\mathbf{nn}}^{H/2} \tag{3.28}$$

   where $\hat{\mathbf{R}}_{\mathbf{nn}}^{1/2}$ is a lower triangular matrix (also referred to as the square root of $\hat{\mathbf{R}}_{\mathbf{nn}}$). Spatial pre-whitening is then performed by multiplying a signal vector of interest by $\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2}$. Hence, spatial pre-whitening of a correlation matrix is performed by pre-multiplying with $\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2}$ and post-multiplying by $\hat{\mathbf{R}}_{\mathbf{nn}}^{-H/2}$. The data-dependent RTF vector can then be obtained from an EVD of the spatially pre-whitened $\hat{\mathbf{R}}_{\mathbf{yy}}$ [25, 41]:

   $$\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2} \hat{\mathbf{R}}_{\mathbf{yy}} \hat{\mathbf{R}}_{\mathbf{nn}}^{-H/2} = \hat{\mathbf{P}} \hat{\mathbf{\Lambda}} \hat{\mathbf{P}}^H \tag{3.29}$$

[10] It is the use of this spatial pre-whitening operator why this procedure is also referred to as covariance whitening in [25, 41].

where $\hat{\mathbf{\Lambda}}$ is a diagonal matrix of eigenvalues arranged in descending order and $\hat{\mathbf{P}}$ is a unitary matrix of the corresponding eigenvectors. Recalling (2.55) and (2.56), $\hat{\mathbf{R}}_{\mathbf{yy}}$ will consist of a rank-1 speech-only correlation matrix plus $\hat{\mathbf{R}}_{\mathbf{nn}}$. Hence it means that the RTF-vector has also been spatially pre-whitened, and this operation needs to be undone. Consequently, the data-dependent RTF vector can be computed as follows using the principal (first in this case) eigenvector in $\hat{\mathbf{P}}$ [25, 41]:

$$\hat{\mathbf{h}}_{\text{spw}} = \frac{\hat{\mathbf{R}}_{\mathbf{nn}}^{1/2}\hat{\mathbf{P}}\mathbf{e}_1}{\mathbf{e}_1^T\hat{\mathbf{R}}_{\mathbf{nn}}^{1/2}\hat{\mathbf{P}}\mathbf{e}_1} \tag{3.30}$$

where the $\hat{\mathbf{R}}_{\mathbf{nn}}^{1/2}$ matrix is used to undo the spatial pre-whitening and the denominator ensures the appropriate scaling. The relation to the GEVD problem of (3.21) can be observed by the substitution of (3.24) in (3.29), which shows that $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Lambda}}$ and $\hat{\mathbf{Q}} = \hat{\mathbf{R}}_{\mathbf{nn}}^{1/2}\hat{\mathbf{P}}$. Hence (3.27) and (3.30) are equivalent. Throughout the dissertation, it is this relationship between spatial pre-whitening and the GEVD problem of (3.21) which are used to formulate various optimisation problems. Numerical errors may nevertheless arise by directly inverting the Cholesky factors and explicitly forming the matrix product of (3.29) for finite precision and ill-conditioning of $\hat{\mathbf{R}}_{\mathbf{nn}}$. Hence more numerically stable algorithms incorporating the Cholesky factors [198] may be preferred.

2. Pre-multiplication of $\hat{\mathbf{R}}_{\mathbf{yy}}$ with $\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}$

   The GEVD problem of (3.21) may also be replaced by considering an EVD of the matrix product $\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\hat{\mathbf{R}}_{\mathbf{yy}}$ [199] (which is not necessarily hermitian-symmetric):

$$\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\hat{\mathbf{R}}_{\mathbf{yy}} = \hat{\mathbf{V}}\hat{\mathbf{\Omega}}\hat{\mathbf{V}}^{-1} \tag{3.31}$$

   where $\hat{\mathbf{\Omega}}$ is a diagonal matrix of the eigenvalues arranged in descending order, and $\hat{\mathbf{V}}$ contains the corresponding eigenvectors. Substitution of (3.24) in (3.31) shows that $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Omega}}$ and $\hat{\mathbf{Q}} = \hat{\mathbf{V}}^{-H}$. Hence upon comparing to (3.27), the data-dependent RTF vector estimator is then given by:

$$\hat{\mathbf{h}}_{\text{prm}} = \frac{\hat{\mathbf{R}}_{\mathbf{nn}}\hat{\mathbf{V}}\mathbf{e}_1}{\mathbf{e}_1^T\hat{\mathbf{R}}_{\mathbf{nn}}\hat{\mathbf{V}}\mathbf{e}_1} \tag{3.32}$$

   The clear disadvantage of this method however, is that the matrix product $\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\hat{\mathbf{R}}_{\mathbf{yy}}$ has to be formed, which can introduce numerical errors if $\hat{\mathbf{R}}_{\mathbf{nn}}$ is ill-conditioned and there is finite numerical precision.

3. Post-multiplication of $\hat{\mathbf{R}}_{\mathbf{yy}}$ with $\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}$

   The EVD of the matrix product $\hat{\mathbf{R}}_{\mathbf{yy}}\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}$ (which is again not necessarily hermitian-symmetric) can also be considered for replacing the GEVD problem of (3.21):

   $$\hat{\mathbf{R}}_{\mathbf{yy}}\hat{\mathbf{R}}_{\mathbf{nn}}^{-1} = \hat{\mathbf{B}}\hat{\mathbf{\Xi}}\hat{\mathbf{B}}^{-1} \tag{3.33}$$

   where $\hat{\mathbf{\Xi}}$ is a diagonal matrix of the eigenvalues arranged in descending order, and $\hat{\mathbf{B}}$ contains the corresponding eigenvectors. Substitution of (3.24) in (3.33) shows that $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Xi}}$ and $\hat{\mathbf{Q}} = \hat{\mathbf{B}}$. Hence upon comparing to (3.27), the data-dependent RTF vector estimator is then given by:

   $$\hat{\mathbf{h}}_{\text{pom}} = \frac{\hat{\mathbf{B}}\mathbf{e}_1}{\mathbf{e}_1^T\hat{\mathbf{B}}\mathbf{e}_1} \tag{3.34}$$

   This is the preferred formulation in [200], where it is highlighted that unlike the previous two procedures, $\hat{\mathbf{R}}_{\mathbf{nn}}$ is not required to transform the principal eigenvectors in order to obtain an estimate for the data-dependent RTF vector. Nevertheless, this procedure still requires forming the product $\hat{\mathbf{R}}_{\mathbf{yy}}\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}$, which can also introduce numerical errors if $\hat{\mathbf{R}}_{\mathbf{nn}}$ is ill-conditioned and there is finite numerical precision.

A final comment to be made on the data-dependent RTF vector pertains how often it should be updated. Typically, it can be updated in each time frame and frequency bin where the target speaker is detected to be present, which was the approach taken throughout the experiments in the dissertation. However, in general, it would depend on the application or acoustic scenario. For instance, if it is expected that the target speaker will move quite regularly in space, then the data-dependent RTF vector should indeed be updated in each time frame and frequency bin where the target speaker is detected to be present. On the other hand, if the target speaker is not expected to move as frequently (such as when two persons are seated at a restaurant), it may be possible to update the data-dependent RTF vector less often, but still only when the the target speaker is detected to be present, thereby reducing some computational complexity.

This data-dependent RTF vector estimator can now be used for the MVDR beamformer defined in (3.18), and is given by:

$$\hat{\mathbf{w}} = \frac{\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\hat{\mathbf{h}}}{\hat{\mathbf{h}}^H\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\hat{\mathbf{h}}} \tag{3.35}$$

which will be referred to as the MVDR-EST. The speech estimate in the first microphone of the array is subsequently calculated as:

$$\hat{z}_1 = \hat{\mathbf{w}}^H\mathbf{y} \tag{3.36}$$

It is noted that the quantities denoted as $(\hat{\cdot})$ will be associated with estimated quantities.

While this MVDR beamformer that uses a data-dependent RTF vector estimate can be effective in many practical scenarios, it is important to acknowledge that if the data cannot yield an accurate RTF vector estimate, then the performance of the MVDR beamformer is subject to degradation. In particular for this subspace-based estimator, a critical requirement is to distinguish between the speech-plus-noise and noise-only periods for an accurate estimation of the respective correlation matrices.

3.5 LCMV with a priori and estimated constraints

Given the limitations of the previous MVDR beamformers, it is expected that each of them would only perform well in certain acoustic scenarios. Therefore a first attempt at obtaining good performance across a wider range of acoustic scenarios would be to consider an LCMV beamformer, where a distortionless response from both the a priori RTF vector, $\widetilde{\mathbf{h}}$, and the data-dependent RTF vector, $\hat{\mathbf{h}}$, can be maintained, i.e. multiple hard constraints can be imposed on the minimisation of the total noise power:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimise}} \quad & \mathbf{w}^H \hat{\mathbf{R}}_{\mathbf{nn}} \mathbf{w} \\ \text{subject to} \quad & \mathbf{H}^H \mathbf{w} = \mathbf{b} \end{aligned} \tag{3.37}$$

where

$$\mathbf{H} = [\widetilde{\mathbf{h}} \;\; \hat{\mathbf{h}}]; \;\; \mathbf{b} = [1 \;\; 1]^T \tag{3.38}$$

The corresponding solution to this equality constrained least squares problem can be found by defining the dual function with a Lagrange multiplier vector, $\boldsymbol{\nu}$, and employing the Karush-Kuhn-Tucker (KKT) conditions to solve the following linear set of equations:

$$\begin{bmatrix} \hat{\mathbf{R}}_{\mathbf{nn}} & \mathbf{H} \\ \mathbf{H}^H & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \boldsymbol{\nu} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{b} \end{bmatrix} \tag{3.39}$$

where the left-hand side coefficient matrix is known as the KKT matrix [123]. The result is the well-known equation:

$$\mathbf{w}_{\text{lcmv}} = \hat{\mathbf{R}}_{\mathbf{nn}}^{-1} \mathbf{H} \, \hat{\mathbf{V}}^{-1} \mathbf{b} \tag{3.40}$$

where

$$\hat{\mathbf{V}} = [\mathbf{H}^H \hat{\mathbf{R}}_{\mathbf{nn}}^{-1} \mathbf{H}] = \begin{bmatrix} k_{aa} & k_{ab} \\ k_{ba} & k_{bb} \end{bmatrix} \tag{3.41}$$

and

$$\begin{aligned} k_{aa} &= \widetilde{\mathbf{h}}^H \hat{\mathbf{R}}_{\mathbf{nn}}^{-1} \widetilde{\mathbf{h}}; \quad k_{bb} = \hat{\mathbf{h}}^H \hat{\mathbf{R}}_{\mathbf{nn}}^{-1} \hat{\mathbf{h}} \\ k_{ab} &= \widetilde{\mathbf{h}}^H \hat{\mathbf{R}}_{\mathbf{nn}}^{-1} \hat{\mathbf{h}}; \quad k_{ba} = \hat{\mathbf{h}}^H \hat{\mathbf{R}}_{\mathbf{nn}}^{-1} \widetilde{\mathbf{h}} \end{aligned} \tag{3.42}$$

Upon expansion of (3.40), it can also be observed that the $\mathbf{w}_{\text{lcmv}}$ can be expressed as a linear combination of the MVDR-APR, $\widetilde{\mathbf{w}}$, and the MVDR-EST, $\hat{\mathbf{w}}$:

$$\mathbf{w}_{\text{lcmv}} = \left[\frac{k_{aa}(k_{bb} - k_{ab})}{k_{aa}k_{bb} - k_{ab}k_{ba}} \right] \widetilde{\mathbf{w}} + \left[\frac{k_{bb}(k_{aa} - k_{ba})}{k_{aa}k_{bb} - k_{ab}k_{ba}} \right] \hat{\mathbf{w}} \tag{3.43}$$

where the denominator, $(k_{aa}k_{bb} - k_{ab}k_{ba})$ is the determinant of $\hat{\mathbf{V}}$, real-valued, and always ≥ 0 due to the Cauchy–Schwarz inequality. In the case of a single speaker, while this LCMV solution can potentially provide a distortionless response if $\widetilde{\mathbf{h}}$ or $\hat{\mathbf{h}}$ is a suitable approximation to the true RTF vector, less noise will be reduced as both constraints are always active.

Caution must however be taken with such an approach as the constraints in $\mathbf{H}$ can become redundant in scenarios when $\widetilde{\mathbf{h}}$ and $\hat{\mathbf{h}}$ are linearly dependent or even approximately so, i.e. $\hat{\mathbf{h}} \approx \delta\widetilde{\mathbf{h}}$, where δ is a scalar. This is possible for instance when $\hat{\mathbf{h}}$ is accurate and the desired speech source is in the a priori assumed direction. Even though such a case can be deemed as a "fortunate" circumstance of events, the consequence is that $\hat{\mathbf{V}}$ will become ill-conditioned. On closer observation of (3.39), it can be seen that $-\hat{\mathbf{V}}$ is the Schur complement of the $\mathbf{0}$-block in the KKT matrix [123] and the Lagrange multiplier vector is given as $\boldsymbol{\nu} = -\hat{\mathbf{V}}^{-1}\mathbf{b}$. Therefore, it can be concluded that when $\hat{\mathbf{h}} \approx \delta\widetilde{\mathbf{h}}$, the dual problem is not well defined.

As an extreme case, if $\hat{\mathbf{h}} = \widetilde{\mathbf{h}}$ ($\delta = 1$), the weights of $\widetilde{\mathbf{w}}$ and $\hat{\mathbf{w}}$ in (3.43) will be indeterminate forms of $\frac{0}{0}$. A distinct example is as follows: for $\widetilde{\mathbf{h}} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T$ and $\hat{\mathbf{h}} = \begin{bmatrix} 1 & 1 + a\cos\theta & 1 + a\sin\theta \end{bmatrix}^T$ (with a being some distance and $\theta \in [0,\ 2\pi]$), it is readily shown that the hard constraints $\mathbf{H}^T\mathbf{w} = \begin{bmatrix} 1 & 1 \end{bmatrix}^T$ are equivalent with $\begin{bmatrix} 1 & 1 & 1 \\ 0 & \cos\theta & \sin\theta \end{bmatrix} \mathbf{w} = \begin{bmatrix} 1 & 0 \end{bmatrix}^T$ where the a cancels out. As a result, even for infinitesimally small a (i.e. for $\hat{\mathbf{h}}$ arbitrarily close to $\widetilde{\mathbf{h}}$), the LCMV beamformer as well as its noise reduction performance will be independent of a but dependent on θ, which is indeed undesirable.

One solution to this problem would be the diagonal loading of $\hat{\mathbf{V}}$, which has been proven to avoid instabilities in the LCMV context [201]. The procedure consists of adding a scaled identity, $\gamma\mathbf{I}$, to $\hat{\mathbf{V}}$, resulting in a regularised version of (3.43):

$$\mathbf{w}_{\text{lcmv}}^{\gamma} = \left[\frac{k_{aa}(k_{bb}^{\gamma} - k_{ab})}{k_{aa}^{\gamma}k_{bb}^{\gamma} - k_{ab}k_{ba}}\right]\widetilde{\mathbf{w}} + \left[\frac{k_{bb}(k_{aa}^{\gamma} - k_{ba})}{k_{aa}^{\gamma}k_{bb}^{\gamma} - k_{ab}k_{ba}}\right]\hat{\mathbf{w}} \tag{3.44}$$

where $k_{aa}^{\gamma} = k_{aa} + \gamma$ and $k_{bb}^{\gamma} = k_{bb} + \gamma$. When $\hat{\mathbf{h}} = \delta\widetilde{\mathbf{h}}$ (3.44) will tend toward a scaled version of the MVDR-APR provided that $\gamma << k_{aa}$.

3.6 Integrated approach

3.6.1 Formulation

The fact that the MVDR-APR, the MVDR-EST, and LCMV beamformers impose hard constraints is the underlying reason why these approaches may still have a limited performance in dynamic acoustic scenarios. Consequently, an alternative approach may be considered such as that proposed in [172] where these hard constraints are softened. Such a beamformer can then merge the benefits of the MVDR-APR and the MVDR-EST beamformers in order to yield a more versatile one. For instance, if the speech source moves outside of the a priori assumed direction, more weight can be given to the data-dependent RTF vector to accommodate for the loss in performance that would otherwise result from using the a priori RTF vector alone. On the other hand, if the data-dependent RTF vector is inaccurate, less weight can be given to it and the system can revert to using the a priori RTF vector, which may have an improved performance if the speech source is indeed close to the a priori assumed direction.

Similar to [172], an integrated MVDR cost function that takes into consideration both $\widetilde{\mathbf{h}}$ and $\hat{\mathbf{h}}$ can be given as:

$$\underset{\mathbf{w}}{\text{minimise}} \quad \mathbf{w}^H\hat{\mathbf{R}}_{\mathbf{nn}}\mathbf{w} + \alpha\,|\mathbf{w}^H\widetilde{\mathbf{h}} - 1|^2 + \beta\,|\mathbf{w}^H\hat{\mathbf{h}} - 1|^2 \tag{3.45}$$

where $\alpha \in [0,\ \infty)$ and $\beta \in [0,\ \infty)$ are tuning parameters that control how much of the respective RTF vectors are imposed. This cost function is simply the combination of the MVDR-APR and MVDR-EST cost function, except that the constraints have been softened by α and β. The solution to (3.45) is then given by:

$$\mathbf{w}_{\text{int}} = (\hat{\mathbf{R}}_{\mathbf{nn}} + \alpha\widetilde{\mathbf{h}}\widetilde{\mathbf{h}}^H + \beta\hat{\mathbf{h}}\hat{\mathbf{h}}^H)^{-1}(\alpha\widetilde{\mathbf{h}}\widetilde{\mathbf{h}}^H + \beta\hat{\mathbf{h}}\hat{\mathbf{h}}^H)\mathbf{e}_1 \tag{3.46}$$

This will be referred to as the MVDR-INT, and can be re-written as:

$$\mathbf{w}_{\text{int}} = (\hat{\mathbf{R}}_{\mathbf{nn}} + \mathbf{HZH}^H)^{-1}(\mathbf{HZH}^H)\mathbf{e}_1 \tag{3.47}$$

where $\mathbf{Z} = \text{diag}\{\alpha, \beta\}$. Applying the matrix inversion lemma to the inverse in (3.47) yields:

$$(\hat{\mathbf{R}}_{\mathbf{nn}} + \mathbf{HZH}^H)^{-1} = \hat{\mathbf{R}}_{\mathbf{nn}}^{-1} - \hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\mathbf{H}\big[\mathbf{Z}^{-1} + (\mathbf{H}^H\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\mathbf{H})\big]^{-1}\mathbf{H}^H\hat{\mathbf{R}}_{\mathbf{nn}}^{-1} \tag{3.48}$$

Substitution of (3.48) into (3.47) eventually results in a simple expression for the MVDR-INT:

$$\mathbf{w}_{\text{int}} = \frac{\alpha k_{aa}[1 + \beta(k_{bb} - k_{ab})]}{\text{D}}\widetilde{\mathbf{w}} + \frac{\beta k_{bb}[1 + \alpha(k_{aa} - k_{ba})]}{\text{D}}\hat{\mathbf{w}} \tag{3.49}$$

where:

$$\text{D} = \alpha k_{aa} + \beta k_{bb} + \alpha\beta(k_{aa}k_{bb} - k_{ab}k_{ba}) + 1 \tag{3.50}$$

It can now be observed from (3.49) that the MVDR-INT beamformer is a linear combination of the MVDR-APR and the MVDR-EST beamformers, whose complex-valued weightings are defined by α, β, and the constants from (3.42).

In the limiting case where $\alpha \to \infty$ and $\beta \to \infty$, (3.45) is equivalent to (3.37) with the LCMV solution given by (3.43). As was discussed in Section 3.5, in scenarios where $\hat{\mathbf{h}} \approx \delta\widetilde{\mathbf{h}}$, the LCMV suffers from an ill-conditioning. Hence the limit as $(\alpha, \beta, \hat{\mathbf{h}}) \to (\infty, \infty, \delta\widetilde{\mathbf{h}})$ (or equivalently $(\frac{1}{\alpha}, \frac{1}{\beta}, \hat{\mathbf{h}}) \to (0, 0, \delta\widetilde{\mathbf{h}})$) will also result in an ill-conditiong problem in function of (3.49). For the LCMV, this problem was resolved with diagonal loading resulting in (3.44). Hence the ill-conditioning for the MVDR-INT can also be resolved by regularising (3.49) as follows:

$$\mathbf{w}_{\text{int}}^{\gamma} = \frac{\alpha k_{aa}[1 + \beta(k_{bb}^{\gamma} - k_{ab})]}{\text{D}^{\gamma}}\widetilde{\mathbf{w}} + \frac{\beta k_{bb}[1 + \alpha(k_{aa}^{\gamma} - k_{ba})]}{\text{D}^{\gamma}}\hat{\mathbf{w}} \tag{3.51}$$

where

$$\text{D}^{\gamma} = \alpha k_{aa} + \beta k_{bb} + \alpha\beta(k_{aa}^{\gamma}k_{bb}^{\gamma} - k_{ab}k_{ba}) + 1 \tag{3.52}$$

For $\alpha \to \infty$ and $\beta \to \infty$, (3.51) then indeed reduces to (3.44).

In order to realise practical values for α and β to analyse the MVDR-INT further, it is suggested to make the following normalisations:

$$\alpha = \frac{\bar{\alpha}}{k_{aa}}; \qquad \beta = \frac{\bar{\beta}}{k_{bb}} \tag{3.53}$$

where $\bar{\alpha}$ and $\bar{\beta}$ are still real-valued constants. These normalisations become particularly useful in Section 3.8.1 so that similar values of $\bar{\alpha}$ and $\bar{\beta}$ can be compared. Substituting (3.53) into (3.51) results in:

$$\bar{\mathbf{w}}_{\text{int}}^{\gamma} = \frac{\bar{\alpha}[1 + \bar{\beta}(\frac{k_{bb}^{\gamma} - k_{ab}}{k_{bb}})]}{\bar{\mathrm{D}}^{\gamma}} \widetilde{\mathbf{w}} + \frac{\bar{\beta}[1 + \bar{\alpha}(\frac{k_{aa}^{\gamma} - k_{ba}}{k_{aa}})]}{\bar{\mathrm{D}}^{\gamma}} \hat{\mathbf{w}} \tag{3.54}$$

where

$$\bar{\mathrm{D}}^{\gamma} = \bar{\alpha} + \bar{\beta} + \bar{\alpha}\bar{\beta}(\frac{k_{aa}^{\gamma} k_{bb}^{\gamma} - k_{ab}k_{ba}}{k_{aa}k_{bb}}) + 1 \tag{3.55}$$

While it is evident that the MVDR-INT is a linear combination of the MVDR-APR and the MVDR-EST beamformers, it should be highlighted that by the substitution of (3.44), (3.54) can also be expressed as:

$$\bar{\mathbf{w}}_{\text{int}}^{\gamma} = \frac{\bar{\alpha}}{\bar{\mathrm{D}}^{\gamma}} \widetilde{\mathbf{w}} + \frac{\bar{\beta}}{\bar{\mathrm{D}}^{\gamma}} \hat{\mathbf{w}} + \frac{\bar{\alpha}\bar{\beta}(\frac{k_{aa}^{\gamma} k_{bb}^{\gamma} - k_{ab}k_{ba}}{k_{aa}k_{bb}})}{\bar{\mathrm{D}}^{\gamma}} \mathbf{w}_{\text{lcmv}}^{\gamma} + \frac{1}{\bar{\mathrm{D}}^{\gamma}} \mathbf{w}_{\text{z}} \tag{3.56}$$

where $\mathbf{w}_{\text{z}}$ is an all-zero $(M \times 1)$ vector. (3.56) reveals that the MVDR-INT is a convex combination of the MVDR-APR beamformer, the MVDR-EST beamformer, the LCMV beamformer, and the all-zero vector. Furthermore, as opposed to (3.54) which consists of complex-valued coefficients, in (3.56), the coefficients are all positive and real-valued since $(k_{aa}^{\gamma} k_{bb}^{\gamma} - k_{ab}k_{ba}) > 0$ from the Cauchy–Schwarz inequality and the regularisation that was introduced. The MVDR-INT can therefore be truly regarded as a more global and versatile beamformer that encompasses wide range of filters bounded by the MVDR-APR, MVDR-EST, LCMV, and the all-zero vector. Consequently, it suggests that (3.56) can be tuned for an optimal performance depending on the acoustic environment. As such, some interesting cases of different tunings of (3.56) are discussed in the following.

3.6.2 Limiting cases of the tuning parameters

In terms of a contingency strategy as in [202], the fall back mechanism will be to revert to an MVDR-APR when it is believed that the MVDR-EST would result in poor performance due to an inaccurate RTF vector estimate. One means by which this can be achieved is to always keep the a priori constraint active, i.e. set $\bar{\alpha} \to \infty$ while only tuning $\bar{\beta}$. This additionally simplifies the practical implementation of the filter as only one tuning parameter needs to be

considered. Considering the limit of $\bar{\alpha} \to \infty$ in (3.56) results in:

$$\lim_{\bar{\alpha}\to\infty} \bar{\mathbf{w}}_{\text{int}}^{\gamma} = \frac{1}{1+\bar{\beta}(\frac{k_{aa}^{\gamma}k_{bb}^{\gamma}-k_{ab}k_{ba}}{k_{aa}k_{bb}})}\widetilde{\mathbf{w}} + \frac{\bar{\beta}(\frac{k_{aa}^{\gamma}k_{bb}^{\gamma}-k_{ab}k_{ba}}{k_{aa}k_{bb}})}{1+\bar{\beta}(\frac{k_{aa}^{\gamma}k_{bb}^{\gamma}-k_{ab}k_{ba}}{k_{aa}k_{bb}})}\mathbf{w}_{\text{lcmv}}^{\gamma} \tag{3.57}$$

which is a convex combination of the MVDR-APR and the LCMV. Clearly, if $\bar{\beta} = 0$, (3.57) reverts to (3.19), satisfying the notion of a contingency strategy. In scenarios where the a priori assumptions are not satisfied, increasing values of $\bar{\beta}$ in (3.57) will attempt to maintain a distortionless response and tend toward the LCMV. This however, may be at the cost of having less noise reduced as $\widetilde{\mathbf{h}}$ is always active. Nevertheless, this compromised performance will be an improvement as opposed to simply using only an MVDR-APR, provided that $\hat{\mathbf{h}}$ is accurate.

A contrary filter can also be derived, where alternatively all of the weight is placed on the data-dependent estimated RTF vector, i.e. $\bar{\beta} \to \infty$. Considering this limit in (3.56) results in:

$$\lim_{\bar{\beta}\to\infty} \bar{\mathbf{w}}_{\text{int}}^{\gamma} = \frac{1}{1+\bar{\alpha}(\frac{k_{aa}^{\gamma}k_{bb}^{\gamma}-k_{ab}k_{ba}}{k_{aa}k_{bb}})}\hat{\mathbf{w}} + \frac{\bar{\alpha}(\frac{k_{aa}^{\gamma}k_{bb}^{\gamma}-k_{ab}k_{ba}}{k_{aa}k_{bb}})}{1+\bar{\alpha}(\frac{k_{aa}^{\gamma}k_{bb}^{\gamma}-k_{ab}k_{ba}}{k_{aa}k_{bb}})}\mathbf{w}_{\text{lcmv}}^{\gamma} \tag{3.58}$$

which is now a convex combination of the MVDR-EST and the LCMV. If $\bar{\alpha} = 0$, (3.58) reverts to (3.35), indicating that this filter has the potential to perform better than the MVDR-APR regardless of the speech source location, provided that $\hat{\mathbf{h}}$ is accurate. If $\hat{\mathbf{h}}$ is not accurate as it will inevitably become in some scenarios, then $\bar{\alpha}$ can be increased to introduce more of the a priori assumed RTF vector and the MVDR-INT will tend toward the LCMV beamformer. If the speech source is such that it adheres to some a priori assumed conditions, for instance, if it is close to the direction defined by $\widetilde{\mathbf{h}}$, then increasing $\bar{\alpha}$ can result in a filter that has a better performance than that of the MVDR-EST when $\hat{\mathbf{h}}$ is inaccurate.

3.6.3 Equivalence to the Speech Distortion Weighted MWF

In the previous sections, it was demonstrated that the MVDR-INT encompasses a number of different spatial filters. It is worthwhile to highlight that the MVDR-INT can also incorporate a postfilter on top of these spatial filters for further noise suppression. This can be observed by considering the case where either $\bar{\alpha} = 0$ or $\bar{\beta} = 0$, which consequently results in an equivalence between the MVDR-INT and the speech-distortion-weighted MWF (SDW-MWF) [128]. In fact, the ability of the MVDR-INT to incorporate a postfilter is also evident

from (3.56) as the all-zero vector, $\mathbf{w}_{\mathrm{z}}$ is also included as one of the terms in the convex combination.

In [39], it was shown that the solution for the GEVD-based SDW-MWF is given by:

$$\mathbf{w}_{\mathrm{sdw-mwf}} = (\mathbf{R_{xx,r1}} + \mu\hat{\mathbf{R}}_{\mathbf{nn}})^{-1}\mathbf{R_{xx,r1}}\mathbf{e}_1 \tag{3.59}$$

where $\mu \in (0 \ \infty)$ is a parameter to trade-off the amount of speech distortion against the amount of noise reduction, and $\mathbf{R_{xx,r1}}$ is a rank-1 approximation to the speech-only correlation matrix, $\mathbf{R_{xx}}$. In the case of an a priori RTF vector, $\mathbf{R_{xx,r1}} = \widetilde{\mathbf{R}}_{\mathbf{xx,r1}} = \hat{\sigma}^2_{\mathrm{s}_1}\widetilde{\mathbf{h}}\widetilde{\mathbf{h}}^H$, whereas for the estimated RTF vector, $\mathbf{R_{xx,r1}} = \hat{\mathbf{R}}_{\mathbf{xx,r1}} = \hat{\sigma}^2_{\mathrm{s}_1}\hat{\mathbf{h}}\hat{\mathbf{h}}^H$, where $\hat{\sigma}^2_{\mathrm{s}_1}$ is an estimate of the speech PSD, $\sigma^2_{\mathrm{s}_1}$ in the first microphone. Considering either of these cases, it is well known that (3.59) can be decomposed into [53]:

$$\mathbf{w}_{\mathrm{sdw-mwf}} = \begin{cases} \dfrac{\hat{\sigma}^2_{\mathrm{s}_1}}{\hat{\sigma}^2_{\mathrm{s}_1} + \mu(\widetilde{\mathbf{h}}^H\hat{\mathbf{R}}^{-1}_{\mathbf{nn}}\widetilde{\mathbf{h}})^{-1}}\ \widetilde{\mathbf{w}} & \text{if } \mathbf{R_{xx,r1}} = \widetilde{\mathbf{R}}_{\mathbf{xx,r1}} \\ \dfrac{\hat{\sigma}^2_{\mathrm{s}_1}}{\hat{\sigma}^2_{\mathrm{s}_1} + \mu(\hat{\mathbf{h}}^H\hat{\mathbf{R}}^{-1}_{\mathbf{nn}}\hat{\mathbf{h}})^{-1}}\ \hat{\mathbf{w}} & \text{if } \mathbf{R_{xx,r1}} = \hat{\mathbf{R}}_{\mathbf{xx,r1}} \end{cases} \tag{3.60}$$

The conditions of $\mathbf{R_{xx,r1}} = \widetilde{\mathbf{R}}_{\mathbf{xx,r1}}$ and $\mathbf{R_{xx,r1}} = \hat{\mathbf{R}}_{\mathbf{xx,r1}}$ are equivalent to either $\bar{\beta} = 0$ or $\bar{\alpha} = 0$ respectively, hence for the MVDR-INT in (3.56), it can be deduced that:

$$\bar{\mathbf{w}}^{\gamma}_{\mathrm{int}} = \begin{cases} \dfrac{\bar{\alpha}}{\bar{\alpha} + k_{aa}(\widetilde{\mathbf{h}}^H\hat{\mathbf{R}}^{-1}_{\mathbf{nn}}\widetilde{\mathbf{h}})^{-1}}\ \widetilde{\mathbf{w}} & \text{if } \bar{\beta} = 0 \\ \dfrac{\bar{\beta}}{\bar{\beta} + k_{bb}(\hat{\mathbf{h}}^H\hat{\mathbf{R}}^{-1}_{\mathbf{nn}}\hat{\mathbf{h}})^{-1}}\ \hat{\mathbf{w}} & \text{if } \bar{\alpha} = 0 \end{cases} \tag{3.61}$$

Upon comparing (3.60) and (3.61), it is observed that both solutions are indeed equivalent provided that:

$$\bar{\alpha} = \frac{\hat{\sigma}^2_{\mathrm{s}_1}\, k_{aa}}{\mu} \quad \text{when } \bar{\beta} = 0 \tag{3.62}$$

$$\bar{\beta} = \frac{\hat{\sigma}^2_{\mathrm{s}_1}\, k_{bb}}{\mu} \quad \text{when } \bar{\alpha} = 0 \tag{3.63}$$

A summary of the various tunings for $\bar{\alpha}$ and $\bar{\beta}$ as applied to (3.56) and the corresponding speech enhancement strategies is presented in Table I.

Table 3.1: Table 1: Summary of speech enhancement strategies resulting from (3.56) for different tunings of $\bar{\alpha}$ and $\bar{\beta}$.

Tuning Parameters	Speech Enhancement Strategy	Equation
$\bar{\alpha} \to \infty$, $\bar{\beta} = 0$	MVDR-APR	(3.19)
$\bar{\alpha} = 0$, $\bar{\beta} \to \infty$	MVDR-EST	(3.35)
$\bar{\alpha} \to \infty$, $\bar{\beta} \to \infty$	LCMV	(3.44)
$\bar{\alpha} \to \infty$, $\bar{\beta} \in [0, \ \infty)$	MVDR-APR + LCMV	(3.57)
$\bar{\alpha} \in [0, \ \infty)$, $\bar{\beta} \to \infty$	MVDR-EST + LCMV	(3.58)
$\bar{\alpha} = 0$, $\bar{\beta} \in [0, \ \infty)$	SDW-MWF	(3.61)
$\bar{\beta} = 0$, $\bar{\alpha} \in [0, \ \infty)$	SDW-MWF	(3.61)

3.7 Tuning Strategy

Following the theoretical analysis of the MVDR-INT, the next issue to be addressed is what combinations of filters should be made in practice or how should the parameters $\bar{\alpha}$ and $\bar{\beta}$ be tuned in a practical situation? From a confidence-based perspective, the first step would be to establish a confidence metric that is indicative of the accuracy of the data-dependent RTF vector that can subsequently be used to assign more or less weight to $\bar{\alpha}$ and $\bar{\beta}$ according to some rule. In this section two potential tuning rules are proposed which make use of this metric of confidence as well as the relationship between the MVDR-INT and the SDW-MWF. The dependence of the quantities on time is reintroduced in this section to emphasize that $\bar{\alpha}$ and $\bar{\beta}$ are updated in each time frame (the dependence on frequency is still omitted as all frequencies are updated in the same manner).

3.7.1 Metric of confidence

One potential option for such a metric of confidence can be the posterior SNR obtained with the MVDR-EST beamformer, which under the assumption of a rank-1 speech-only correlation matrix, is equal to the the principal generalised eigenvalue, $\hat{\sigma}_1(l)$, from $\hat{\mathbf{\Sigma}}(l)$ in (3.21). Consequently, it is expected that larger values of $\hat{\sigma}_1(l)$ will correspond to a more accurate estimate of the RTF vector and hence be used to assign more weight to $\bar{\beta}$.

The posterior SNR obtained with the MVDR-EST beamformer is given as:

$$\mathrm{SNR}_{\mathrm{post}}(\mathrm{l}) = \frac{\hat{\mathbf{w}}^H(\mathrm{l})\hat{\mathbf{R}}_{\mathbf{yy}}(\mathrm{l})\hat{\mathbf{w}}(\mathrm{l})}{\hat{\mathbf{w}}^H(\mathrm{l})\hat{\mathbf{R}}_{\mathbf{nn}}(\mathrm{l})\hat{\mathbf{w}}(\mathrm{l})} \tag{3.64}$$

By substitution of (3.27) into (3.35), an alternate expression for the MVDR-EST beamformer is given by:

$$\hat{\mathbf{w}}(\mathrm{l}) = (\mathbf{e}_1^T\hat{\mathbf{Q}}^H(\mathrm{l})\mathbf{e}_1)\hat{\mathbf{Q}}^{-H}(\mathrm{l})\mathbf{e}_1 \tag{3.65}$$

Substitution of the joint diagonalisation expressions of $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$ from (3.24), as well as (3.65) into (3.64) then results in:

$$\mathrm{SNR}_{\mathrm{post}}(\mathrm{l}) = \mathbf{e}_1^T\hat{\mathbf{\Sigma}}(\mathrm{l})\mathbf{e}_1 = \hat{\sigma}_1(\mathrm{l}) \tag{3.66}$$

which shows that the posterior SNR obtained with the MVDR-EST beamformer is equal to the the principal generalised eigenvalue, $\hat{\sigma}_1(\mathrm{l})$.

The values of $\hat{\sigma}_1(\mathrm{l})$ can be considered for two extreme cases of a VAD: (i) a perfect VAD, where there is accurate classification of the speech and noise frames so that $\hat{\mathbf{R}}_{\mathbf{yy}}(\mathrm{l}) \approx \mathbf{R}_{\mathbf{yy}}(\mathrm{l})$ and $\hat{\mathbf{R}}_{\mathbf{nn}} \approx \mathbf{R}_{\mathbf{nn}}(\mathrm{l})$, i.e. the estimates of these correlation matrices are approximately equal to their true values and (ii) a worst case imperfect VAD where there is complete misclassification of the speech and noise frames so that $\hat{\mathbf{R}}_{\mathbf{yy}}(\mathrm{l}) \approx \mathbf{R}_{\mathbf{nn}}(\mathrm{l})$ and $\hat{\mathbf{R}}_{\mathbf{nn}} \approx \mathbf{R}_{\mathbf{yy}}(\mathrm{l})$.

The joint diagonalisation of the true correlation matrices $\mathbf{R}_{\mathbf{yy}}(\mathrm{l})$ and $\mathbf{R}_{\mathbf{nn}}(\mathrm{l})$ is firstly denoted respectively as $\mathbf{R}_{\mathbf{yy}}(\mathrm{l}) = \mathbf{Q}(\mathrm{l})\mathbf{\Sigma}(\mathrm{l})\mathbf{Q}^H(\mathrm{l})$ and $\mathbf{R}_{\mathbf{nn}}(\mathrm{l}) = \mathbf{Q}(\mathrm{l})\mathbf{I}_M\mathbf{Q}^H(\mathrm{l})$ where $\mathbf{Q}(\mathrm{l})$ and $\mathbf{\Sigma}(\mathrm{l})$ are the true values of $\hat{\mathbf{Q}}(\mathrm{l})$ and $\hat{\mathbf{\Sigma}}(\mathrm{l})$ respectively. In the case of the perfect VAD, replacing $\hat{\mathbf{R}}_{\mathbf{yy}}(\mathrm{l})$ and $\hat{\mathbf{R}}_{\mathbf{nn}}(\mathrm{l})$ respectively with $\mathbf{R}_{\mathbf{yy}}(\mathrm{l})$ and $\mathbf{R}_{\mathbf{nn}}(\mathrm{l})$ (and their joint diagonalisations) in (3.21) results in $\hat{\mathbf{\Sigma}}(\mathrm{l}) = \mathbf{\Sigma}(\mathrm{l})$. From (3.25), the relation can also be made that $\mathbf{\Sigma}(\mathrm{l}) = \mathbf{\Sigma}_{\mathbf{x}}(\mathrm{l}) + \mathbf{I}_M$, where $\mathbf{\Sigma}_{\mathbf{x}}(\mathrm{l})$ is the true rank-1 speech-only correlation matrix with the true principal eigenvalue, $\sigma_{\mathrm{x1}}(\mathrm{l})$. In the case of the imperfect VAD, replacing $\hat{\mathbf{R}}_{\mathbf{yy}}(\mathrm{l})$ and $\hat{\mathbf{R}}_{\mathbf{nn}}(\mathrm{l})$ respectively with $\mathbf{R}_{\mathbf{nn}}(\mathrm{l})$ and $\mathbf{R}_{\mathbf{yy}}(\mathrm{l})$ (and their joint diagonalisations) in (3.21) then results in $\hat{\mathbf{\Sigma}}(\mathrm{l}) = \mathbf{\Sigma}^{-1}(\mathrm{l}) = (\mathbf{\Sigma}_{\mathbf{x}}(\mathrm{l}) + \mathbf{I}_M)^{-1}$. Consequently for these two extreme cases of a VAD $\hat{\sigma}_1(\mathrm{l})$ will be given as:

$$\hat{\sigma}_1(\mathrm{l}) \approx \begin{cases} \sigma_{\mathrm{x1}}(\mathrm{l}) + 1 & \text{for perfect VAD} \\ \frac{1}{\sigma_{\mathrm{x1}}(\mathrm{l})+1} & \text{worst case imperfect VAD} \end{cases} \tag{3.67}$$

In the perfect VAD case, it can be observed that larger values of $\hat{\sigma}_1(\mathrm{l})$, i.e. $\hat{\sigma}_1(\mathrm{l}) \gg 1$ correspond to the situation where $\sigma_{\mathrm{x1}}(\mathrm{l}) \gg 1$, indicative that speech was present and correctly classified, which would presumably lead to an

accurately estimated RTF vector. On the other hand for a worst case imperfect VAD, since the speech is inaccurately classified, regardless the value of $\sigma_{x1}(l)$, $\hat{\sigma}_1(l) \leq 1$. It can also be observed that in general, when the speech is classified as noise, smaller values of $\hat{\sigma}_1(l)$ will be expected. In the case where the noise is classified as speech, this can result in larger values of $\hat{\sigma}_1(l)$ but is however, not as detrimental to overall performance in comparison to the case when speech is classified as noise [203].

A logisitic function can then be used to map $\hat{\sigma}_1(l)$ to a value, $F(l) \in [0 \;\; 1]$, where 0 indicates a low confidence in the accuracy of the estimated RTF and 1 a high confidence in the accuracy of the estimated RTF. The relationship between $F(l)$ and $\hat{\sigma}_1(l)$ is given by:

$$F(l) = \frac{1}{1 + e^{-\rho(\hat{\sigma}_1(l) - \sigma_t)}} \tag{3.68}$$

where ρ controls the gradient of the transition from 0 to 1, and σ_t is a threshold generalised eigenvalue, beyond which $F(l) \to 1$. Therefore, using higher values for σ_t would correspond to a more conservative approach where $F(l)$ would tend mostly to 0 and smaller values for σ_t would correspond to a more aggressive approach where $F(l)$ would tend mostly to 1. Figure 3.1 depicts examples of logistic functions that can be obtained by varying ρ and σ_t in (3.68).

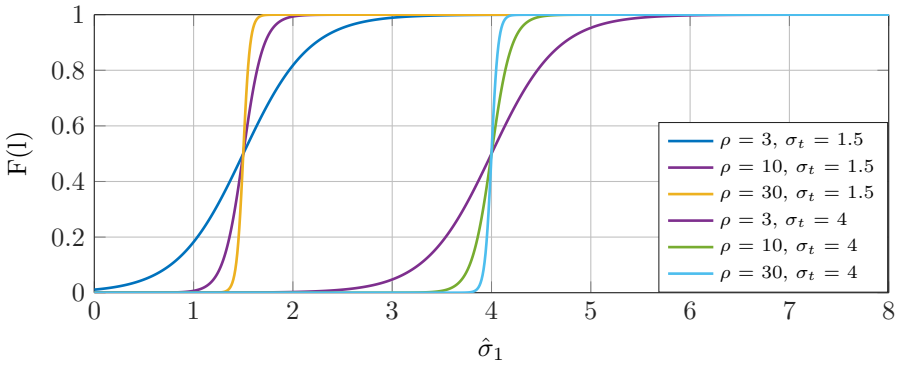


Figure 3.1: Examples of logistic functions that can be obtained by varying ρ and σ_t in (3.68) in order to define the confidence metric. ρ controls the slope of the functions, while σ_t controls the threshold principal generalised eigenvalue, beyond which $F(l) \to 1$.

3.7.2 Tuning rules

SDW tuning rule

The relationship between the MVDR-INT and the SDW-MWF from Section 3.6.3 suggests that (3.62) and (8.58) can be used for tuning $\bar{\alpha}(\mathrm{l})$ and $\bar{\beta}(\mathrm{l})$ respectively. However, in order to control the respective weights in accordance with how well the RTF vector has been estimated, the previously proposed metric of confidence can be incorporated such that $\bar{\alpha}(\mathrm{l})$ and $\bar{\beta}(\mathrm{l})$ are tuned as follows:

$$\bar{\alpha}^{\mathrm{sdw}}(\mathrm{l}) = (1 - \mathrm{F}(\mathrm{l}))\frac{\hat{\sigma}^2_{s_1}(\mathrm{l})k_{aa}(\mathrm{l})}{\mu} \tag{3.69}$$

$$\bar{\beta}^{\mathrm{sdw}}(\mathrm{l}) = \mathrm{F}(\mathrm{l})\frac{\hat{\sigma}^2_{s_1}(\mathrm{l})k_{bb}(\mathrm{l})}{\mu} \tag{3.70}$$

For this tuning, μ must be chosen (which is a familiar trade-off choice) and $\hat{\sigma}^2_{s_1}(\mathrm{l})$ must be computed, for instance as in [204]. Furthermore, it should be noted that if $\hat{\mathbf{h}} = \widetilde{\mathbf{h}}$, substitution of (3.69) and (3.70) into (3.45) would yield the cost function for the SDW-MWF for all values of $\mathrm{F}(\mathrm{l})$. Otherwise, there is a trade-off between the MVDR-APR and the MVDR-EST, with more emphasis placed on the respective beamformer as prescribed by $\mathrm{F}(\mathrm{l})$. With values computed for $\bar{\alpha}^{\mathrm{sdw}}(\mathrm{l})$ and $\bar{\beta}^{\mathrm{sdw}}(\mathrm{l})$, these parameters can then be substituted in (3.54) to compute the MVDR-INT beamformer. This tuning rule is summarised in Algorithm 1.

Algorithm 1 SDW tuning rule for the MVDR-INT at each time frame for a particular frequency.

Set value for μ, γ, ρ, and σ_t
for $l = 1$ to L **do**
(1) Estimate $\sigma^2_{s_1}(\mathrm{l})$ from the reference microphone.
(2) Compute $\widetilde{\mathbf{w}}(\mathrm{l})$ from (3.19) and $\hat{\mathbf{w}}(\mathrm{l})$ from (3.35)
(3) Compute F (l) from (3.68).
(4) Compute $\bar{\alpha}^{\mathrm{sdw}}(\mathrm{l})$ from (3.69) and $\bar{\beta}^{\mathrm{sdw}}(\mathrm{l})$ from (3.70).
(5) Compute $\bar{\mathbf{w}}^{\gamma}_{\mathrm{int}}(\mathrm{l})$ from (3.54)
end for

Contingency tuning rule

A second tuning rule may also be considered that is in line with the contingency noise reduction strategy of [202]. For this strategy, only $\bar{\beta}$ needs to be tuned,

which can be computed as $\bar{\beta}^{\text{sdw}}(\text{l})$ from (3.70). This can then be substituted along with $\widetilde{\mathbf{w}}$ from (3.19), $\hat{\mathbf{w}}(\text{l})$ from (3.35), and $\mathbf{w}_{\text{lcmv}}^{\gamma}$ from (3.44) into (3.57). In such a strategy, smaller values of $\bar{\beta}^{\text{sdw}}(\text{l})$ will tend to the MVDR-INT beamformer, while larger values of $\bar{\beta}^{\text{sdw}}(\text{l})$ will tend to the LCMV beamformer. This contingency tuning rule is summarised in Algorithm 2.

Algorithm 2 Contingency tuning rule for the MVDR-INT at each time frame for a particular frequency.

Set value for μ, γ, ρ, and σ_t
for $l = 1$ to L **do**
(1) Estimate $\sigma_{\text{s}_1}^2(\text{l})$ from the reference microphone.
(2) Compute $\widetilde{\mathbf{w}}(\text{l})$ from (3.19) and $\hat{\mathbf{w}}(\text{l})$ from (3.35)
(3) Compute $\mathbf{w}_{\text{lcmv}}^{\gamma}(\text{l})$ from (3.44)
(4) Compute F (l) from (3.68).
(5) Compute $\bar{\beta}^{\text{sdw}}(\text{l})$ from (3.70).
(6) Compute $\bar{\mathbf{w}}_{\text{int}}^{\gamma}(\text{l})$ from (3.57) (the contingency version)
end for

3.8 Evaluation and Discussion

In order to evaluate the MVDR-INT strategy, both simulated data as well as recorded data from a typical office room were used. In the simulated case, scenarios with accurately and inaccurately data-dependent RTF vectors were considered in order to understand how the tuning parameters, $\bar{\alpha}$ and $\bar{\beta}$ affect the MVDR-INT beamformer. For the recorded data, the tuning rules as described in Section 3.7 were applied to the MVDR-INT beamformer and the resulting performance was compared to that from the MVDR-APR, MVDR-EST, and LCMV beamformers.

For the processing of the algorithms, the weighted overlap-and-add (WOLA) method [101], with a discrete Fourier transform (DFT) size of 512, 50 % overlap, a square-root Hanning window, and a sampling frequency of 16 kHz was used. Depending on the scenario under consideration, either a perfect or imperfect means of voice activity detection was used to retrieve $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$.

The metrics used to evaluate the various experiments were the change in unweighted output SNR from the input at the reference microphone to the output and the short-time objective intelligibility (STOI) measure as described in 2.8. The reference signal for the STOI measure was the unprocessed speech-only time-domain signal.

3.8.1 Simulated Data

The simulation environment, depicted in Figure 3.2, consisted of a reverberant room with reverberation time (RT) 0.25 s and dimensions 4.3 m × 6.9 m × 2.6 m, a linear microphone array, and a single speech source within a spherically diffuse noise field. The array consisted of 4 omnidirectional microphones with an inter-element spacing of 4 cm. For the speech source signal, six sentences separated by silence from the English Hearing-In-Noise Test (HINT) database [205] were used. The diffuse noise field was generated using uncorrelated excerpts of multitalker babble noise from Audiotec [206] and the method outlined in [207]. Uncorrelated white noise was also added to each of the microphone signals such that the ratio of the speech signal power in the first microphone of the array to the uncorrelated white noise power was 30 dB. The room impulse responses were obtained using the randomised image method [208] and implemented from [209]. $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$ were then estimated accordingly via recursive averaging with an time constant of 3 s.

Two scenarios in which to observe the effect of the tuning parameters of the MVDR-INT were considered: (i) where the data-dependent RTF vector was accurately estimated and (ii) where the data-dependent RTF vector was not accurately estimated. For each scenario, the speech source was placed 1 m away from the centre of the microphone array and was swept through a series of angles from 0° to 180° (as indicated by the bold curved arrow in figure 3.2). For each speech source angle, the evaluation metrics previously discussed were then calculated for the MVDR-APR, MVDR-EST and MVDR-INT for several values of $\bar{\alpha}$ and $\bar{\beta}$ as a function of the angle of the speech source. The regularised version of the MVDR-INT from (3.54) was used for all evaluations with the regularisation parameter, $\gamma = 10^{-2}k_{aa}$.

For the MVDR-APR, the a priori RTF vector , $\widetilde{\mathbf{h}}$ was defined with respect to the endfire direction. A white noise signal of 20 s was played at the position of the speech source in the endfire direction in order to compute a rank-1 correlation matrix per frequency. The first column normalised with respect to a reference microphone was then used as the definition for $\widetilde{\mathbf{h}}$. For these simulations, the first microphone (m_1 in Fig. 3.2) was used as the reference microphone. For the MVDR-EST, $\hat{\mathbf{h}}$ was computed from (3.27).

Accurately estimated data-dependent RTF vector

In this scenario, the input SNR at the reference microphone was set to 4 dB, and a perfect VAD was used to estimate the correlation matrices, $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$. The use of a perfect VAD is idealistic of course, however, it was used to

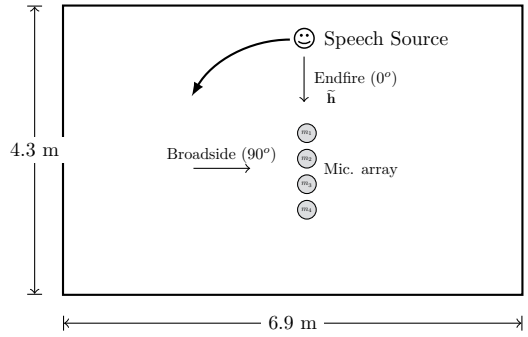


Figure 3.2: Plan view of simulation environment for an accurately estimated RTF vector with a reverberation time of 0.25 s. The arrow on the speech source indicates that simulations were done for different speech source angles, including that corresponding to $\widetilde{\mathbf{h}}$. Not shown is the simulated diffuse noise field. The room height was 2.6 m.

ensure that the data-dependent RTF vector would have indeed been accurately estimated. The VAD used was also applied to the time domain signal so that all frequency bins would have been treated as containing speech or not. In the experiments involving audio recordings, an SPP estimator is used, which allows for a frequency dependent VAD.

Fig. 3.3 displays the subsequent results from the simulations in this scenario. On the left column, Figs. 3.3 (a)-(b) show the performance metrics of the MVDR-APR, MVDR-EST and the MVDR-INT for the tuning parameters such that $\bar{\alpha} \geq \bar{\beta}$. On the right column, figures 3.3 (c)-(d) show the same performance metrics for the MVDR-APR and MVDR-EST, but with the MVDR-INT tuned such that $\bar{\beta} > \bar{\alpha}$. A global legend for all of the plots in this section is provided in Table II, indicating the exact values of $\bar{\alpha}$ and $\bar{\beta}$ used for the MVDR-INT.

From all of the plots, it can be observed that the MVDR-EST performs better than the MVDR-APR as expected with higher Δ SNR and Δ STOI for all speech source angles. The MVDR-APR on the other hand only has a performance similar to the MVDR-EST at a speech source angle of 0°, i.e. the direction from which $\widetilde{\mathbf{h}}$ was defined and a clear drop in performance or all other speech source angles. It is noted as well that for the MVDR-EST, there is a general reduction in performance closer to the broadside direction. This is in accordance with the fact that superdirective beamformers achieve higher gains at endfire directions [120] and is not surprising, considering the symmetry of the beam patterns created by the array.

Focusing on the left column of Fig. 3.3, the various plots of the MVDR-INT

Table 3.2: Global legend for the plots of the different noise reduction algorithms in Fig. 3.3 and Fig. 3.4.

Plot Style	Noise Reduction Strategy	$\bar{\alpha}$	$\bar{\beta}$
	MVDR-EST	-	-
	MVDR-APR	-	-
	MVDR-INT	100	0.1
	MVDR-INT	100	1
	MVDR-INT	100	5
	MVDR-INT	100	10
	MVDR-INT	100	100
	MVDR-INT	0.1	100
	MVDR-INT	1	100
	MVDR-INT	5	100
	MVDR-INT	10	100

are for $\bar{\alpha}$ fixed to 100 and $\bar{\beta} = \{0.1, 1, 5, 10, 100\}$, i.e. $\bar{\alpha} \geq \bar{\beta}$. Firstly, it can be seen that as $\bar{\alpha}$ is increasingly greater than $\bar{\beta}$, there is a convergence to the MVDR-APR as demonstrated by the extreme case where $\bar{\alpha} = 100$ and $\bar{\beta} = 0.1$. For non-zero values of $\bar{\beta}$, it is observed that there is an improved performance over the MVDR-APR beyond a source location of 30°.

On the right column of Fig. 3.3, the various plots of the MVDR-INT are now for $\bar{\beta}$ fixed to 100 and $\bar{\alpha} = \{0, 0.1, 1, 5, 10\}$, i.e. $\bar{\beta} > \bar{\alpha}$. It can be seen that as $\bar{\beta}$ is increasingly greater than $\bar{\alpha}$, there is a convergence to the MVDR-EST as demonstrated by the extreme case where $\bar{\beta} = 100$ and $\bar{\alpha} = 0.1$. As $\bar{\alpha}$ is increased, however, the Δ STOI seems to be maintained, with some minimal reduction in the Δ SNR. This reduction would have been due to the fact that there is extra noise from the direction defined by $\widetilde{\mathbf{h}}$ for greater values of $\bar{\alpha}$.

Inaccurately estimated data-dependent RTF vector

Fig. 3.4 displays simulation results in the second scenario where the RTF vector estimate has not been accurate. Here the noise was scaled such that the input SNR at the reference microphone was -3 dB. A time-domain imperfect VAD [210], using the Minimum Mean Square Error (MMSE) method and with a mean talkspurt length of 500 ms was used to compute $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$. Once again, for each speech source angle, the Δ SI-SNR and the Δ STOI were computed for the MVDR-APR, MVDR-EST and the regularised version of the MVDR-INT from (3.54) for several tunings of $\bar{\alpha}$ and $\bar{\beta}$.

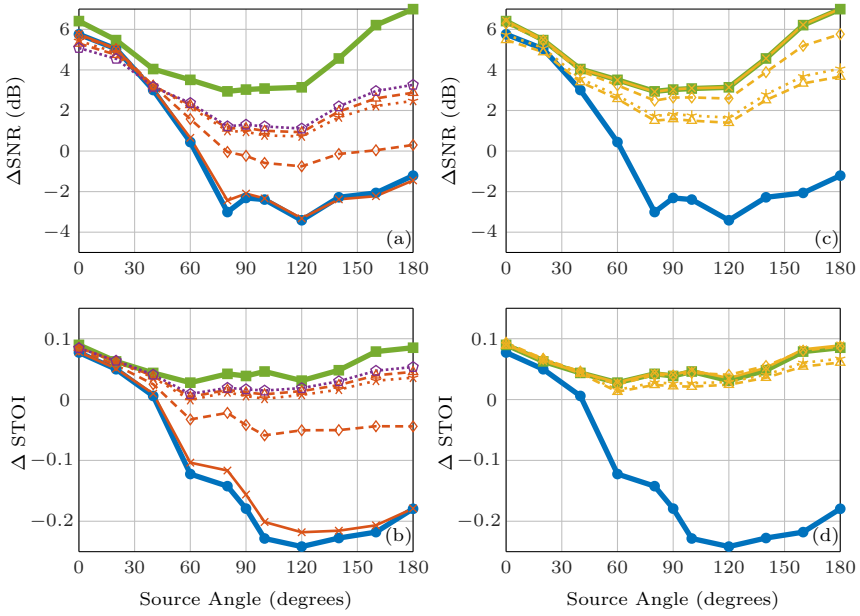


Figure 3.3: Performance of the MVDR-APR, MVDR-EST and MVDR-INT with different values of $\bar{\alpha}$ and $\bar{\beta}$, for a scenario where $\hat{\mathbf{h}}$ was accurately estimated. The input SNR was 4 dB. The left column, (a)-(b), displays the MVDR-INT with $\bar{\alpha} \geq \bar{\beta}$ and the right column, (c)-(d), displays the MVDR-INT with $\bar{\beta} > \bar{\alpha}$. The corresponding legend is shown in Table II.

Immediately it can be observed that the MVDR-APR maintains its typical performance characteristic as in Fig. 3.3 for Δ SNR and Δ STOI metrics, where there is a maximum improvement at the endfire direction that tapers off in all other source angles. The MVDR-EST on the other hand suffers from a uniform reduction in performance across all source angles in comparison to Fig. 3.3 when $\hat{\mathbf{h}}$ was accurately estimated. The loss is particularly apparent in the endfire direction with respect to the MVDR-APR, which implies that the RTF vector was poorly estimated. Although there seems to be an improvement over the MVDR-APR outside of the endfire direction, a robust performance is not expected in this region due to the unpredictability of the imperfect VAD.

With respect to the MVDR-INT in both the left and right columns of Fig. 3.4, it is seen again that in the extreme case of $\bar{\alpha} = 100$ and $\bar{\beta} = 0.1$ it converges to the MVDR-APR and for $\bar{\alpha} = 0.1$ and $\bar{\beta} = 100$, it converges to the MVDR-EST. In all other tunings of $\bar{\alpha}$ and $\bar{\beta}$ displayed (i.e. the same tunings as described for Fig. 3.3), a similar performance is achieved that is a compromise between

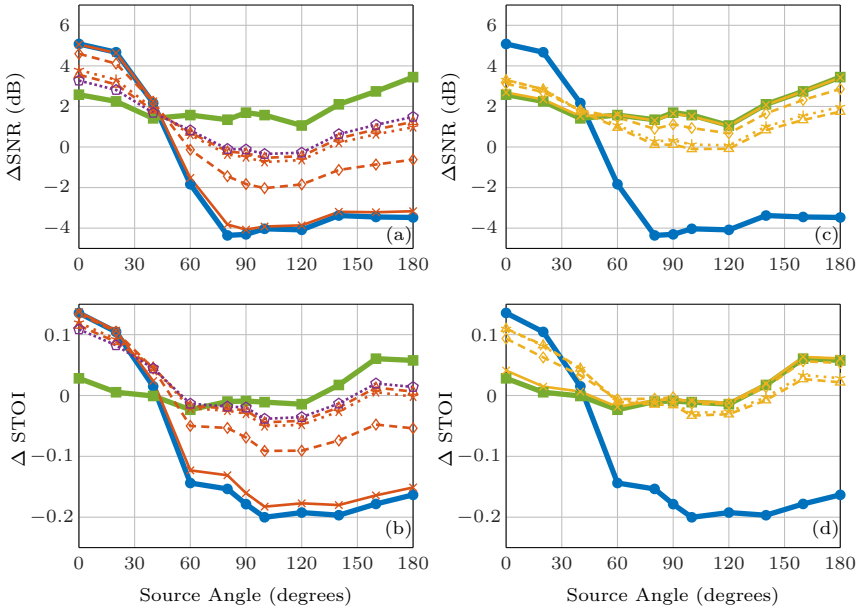


Figure 3.4: Performance of the MVDR-APR, MVDR-EST and MVDR-INT with different values of $\bar{\alpha}$ and $\bar{\beta}$, for a scenario where $\hat{\mathbf{h}}$ was inaccurately estimated. The input SNR was -3 dB. The left column, (a)-(b), displays the MVDR-INT with $\bar{\alpha} \geq \bar{\beta}$ and the right column, (c)-(d), displays the MVDR-INT with $\bar{\beta} > \bar{\alpha}$. The corresponding legend is shown in Table II.

the MVDR-APR and the MVDR-EST. A benefit is indeed achieved in the endfire region over the MVDR-EST as $\bar{\alpha}$ is increasingly greater than $\bar{\beta}$, which introduces more of the MVDR-APR beamformer.

These simulations have provided some insight into the question of how to set the $\bar{\alpha}$ and $\bar{\beta}$ for the MVDR-INT. In the following section, the MVDR-INT will be applied to recordings from a typical office room, which use the proposed tuning strategies from Section 3.7.

3.8.2 Recorded Data

Audio recordings were made in an office room of dimensions $5.4\,\mathrm{m} \times 3.5\,\mathrm{m} \times 2.5\,\mathrm{m}$ with an approximate reverberation time of 0.3 s. The audio signals (both speech and noise) were played through an RME UCX sound card, Samson SERVO 200 amplifier and a JBL CONTROL 1 PRO loudspeaker. These signals were then

captured by a linear microphone array consisting of four omni-directional AKG CK32 microphones with an inter-element spacing of 4 cm. The microphones were connected to a Behringer EURORACK MX 3242X mixer and acquired through Simulink, Matlab via a Speedgoat real-time acquisition system. The speech and noise were recorded separately and added afterwards to create a desired input SNR.

As in the simulations, this experiment served to evaluate the performance of the various algorithms on the basis of a speech source in different locations. Hence, the speech source, SS1, was instantaneously moved between locations (as depicted in Fig. 3.5) - (i) the endfire direction (0°, which would be set to the a priori direction), (ii) -45°, (iii) the broadside direction (-90°), and finally again to (iv) the endfire direction. For each of these locations 6 random sentences separated by silence from the HINT database were used, each lasting for a duration of approximately 20 s. A localised noise source of speech-shaped noise was also recorded at an angle of 60° relative to the endfire direction. In addition to recording the speech and the noise separately, a white noise signal played at the endfire position was also recorded and used to compute $\underset{\sim}{\mathbf{h}}$ in a similar manner to that of Section 3.8.1. The uppermost microphone in Fig. 3.5 was used as the reference microphone.

Using the individual speech and noise recordings, a noisy signal was created such that for the first 60 s (i.e. during the instantaneous movements from the endfire to the broadside direction), the input SNR at the reference microphone was approximately 4 dB and for the last 20 s (when the source was back in the endfire direction), the noise was increased so that the input SNR at the reference microphone in this segment was approximately -3 dB.

The upper plot of Fig. 3.6 illustrates this noisy input signal of the reference microphone and the lower plot of Fig. 3.6 depicts the corresponding probability if speech is present in the STFT domain after applying the SPP estimator from [192]. Using this result, periods for which the speech was active were extracted if the SPP ≥ 0.5, and periods of noise were extracted if the SPP < 0.5. $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$ were then computed accordingly via recursive averaging with a time constant of 1 s. On inspection of the lower plot of Fig. 3.6, it can be observed that in the final 20 s, where the input SNR is lower, particularly in the lower frequencies, the speech presence is not as distinct as compared for previous times, which is indicative of some misclassification.

Fig. 3.7 displays the resulting performance of the MVDR-APR, MVDR-EST, and the regularised LCMV from (3.44), along with the MVDR-INT, for the different tuning rules described in Section 3.7.2 when $\mu = 0.001$. $\hat{\sigma}^2_{s_1}(l)$ was computed using the method from [204] as implemented in [210] but with the noise estimation update computed as in [192]. The SDW tuning rule from

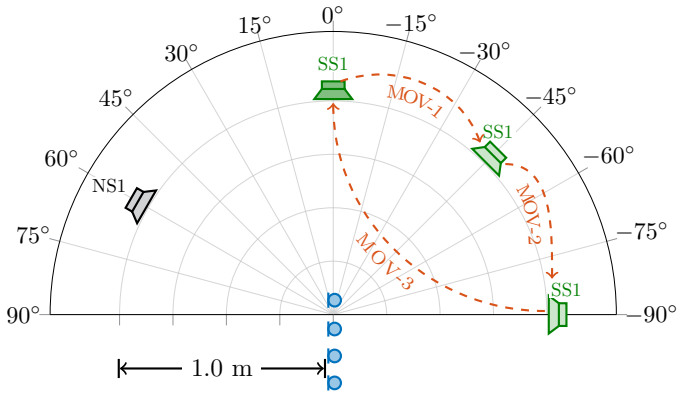


Figure 3.5: Scenario for recordings captured by a 4-element microphone array in an office room. The speech source was played through SS1, which was moved from 0°, -45°, -90°, and then back to 0°. The speech source remained in each position for approximately 20 s. Speech-shaped noise was played through the NS1 at 60°.

Algorithm 1 will be referred to as MVDR-INT-sdw and the contingency tuning rule from Algorithm 2 will be referred to as MVDR-INT-cnt. All algorithms were evaluated using the metrics mentioned at the beginning of this section and were computed over 5 s time frames with a 50 % overlap. For the logistic function of (3.68), the parameters were set such that $\rho = 10$, and $\sigma_t = 20$. For the regularisation parameter, $\gamma = 10^{-2} k_{aa}$.

Firstly, focusing on the MVDR-APR and the MVDR-EST, as expected, the MVDR-APR exhibits a robust performance when the speech is in the endfire direction (a priori assigned direction), i.e. during the time segments of the first 20 s and the final 20 s. For all other time segments, where the speech was outside of the endfire direction, the metrics are indicative of a reduced performance. The MVDR-EST on the other hand maintains its performance for changes in the speech source location during the first 60 s, but then suffers a reduction in performance in the final 20 s, where $\hat{\mathbf{h}}$ would have been an inaccurate estimate due to the lower input SNR. For the MVDR-EST, it should be noted that although the Δ STOI seems similar from 40 s to 80 s, the absolute STOI value of at the reference microphone is much lower in the final 20 s, and hence the MVDR-EST indeed exhibits a poor performance in these final 20 s. The LCMV is able to attain a performance that is in between that of the MVDR-APR and the MVDR-EST. However, as indicated by the Δ SNR, when the speech is outside of the endfire direction, less noise is reduced in comparison to the MVDR-EST , and when $\hat{\mathbf{h}}$ was inaccurate, less noise is reduced in comparison

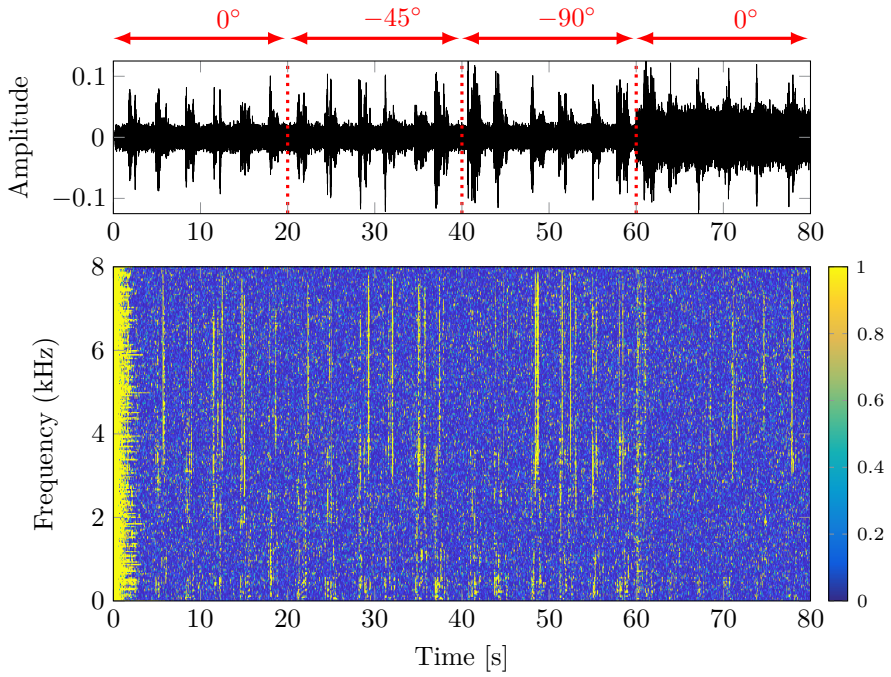


Figure 3.6: (Top) Input noisy signal of the reference microphone. The arrows above the plot indicate the 20 s intervals for which the speech source was at a particular angle in accordance with Fig. 3.5. (Bottom) Corresponding probability if speech is present in this signal after applying the SPP estimator from [192].

to the MVDR-APR.

For the first 60 s, it can be observed that the performance of the MVDR-INT-sdw is similar to that of the MVDR-EST in terms of both Δ SNR and Δ STOI. In the final 20 s, when $\hat{\mathbf{h}}$ was inaccurate, the MVDR-INT-sdw however moves away from the performance of the MVDR-EST and approaches that of the MVDR-APR, truly merging the benefits of both the MVDR-APR and MVDR-EST. The MVDR-INT-cnt, on the other hand, maintains the performance of the MVDR-APR when the speech source is in the endfire direction, but otherwise has a performance that is similar to the LCMV. Hence the MVDR-INT-cnt also reduces less noise when the speech is not in the a priori assumed location, but nevertheless merges the benefits of the MVDR-APR and the LCMV.

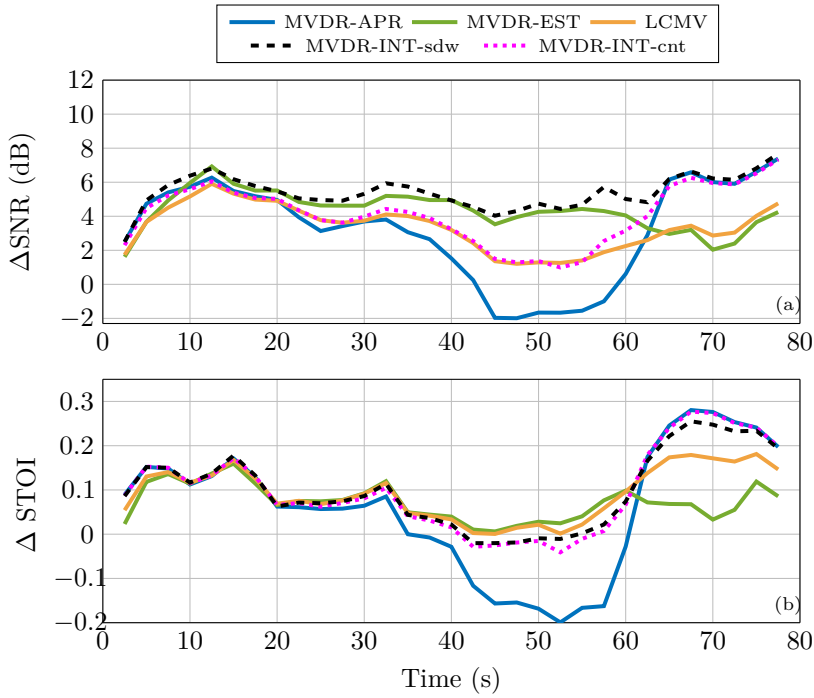


Figure 3.7: Performance of the MVDR-APR, MVDR-EST, LCMV, MVDR-INT-sdw, and MVDR-INT-cnt when $\mu = 0.001$ for the scenario of Fig. 3.5.

Fig. 3.8 displays the results for the same experiment as in Fig. 3.7, except that $\mu = 0.1$ for the MVDR-INT-sdw and MVDR-INT-cnt. This larger value of μ introduces a more aggressive post filter, which accounts for the increased Δ SNR for the MVDR-INT-sdw. The Δ STOI remains similar to that when $\mu = 0.001$, although some audible artifacts can now be heard, which is an expected result for more aggressive postfilters. The performance of the MVDR-INT-cnt remains relatively similar to that when $\mu = 0.001$.

The results of Fig. 3.7 and Fig. 3.8 can be further explained by observing the values of the metric of confidence, $\mathrm{F}(\mathrm{k},\mathrm{l})$ in Fig. 3.9. In general, there was more confidence for the higher frequencies and periods where the speech source was outside of the a priori assumed direction, resulting in larger values being assigned to $\bar{\beta}$ and hence imposing more of the MVDR-EST. In the final 20 s in particular, there was less confidence in the lower frequency region, resulting in lower values being assigned to $\bar{\beta}$, which would have imposed more of the

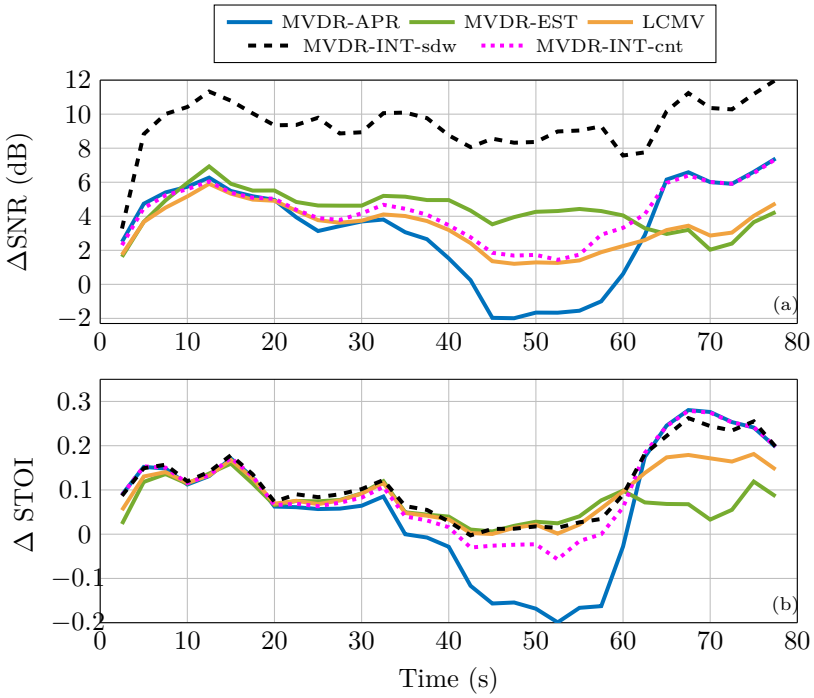


Figure 3.8: Performance of the MVDR-APR, MVDR-EST, LCMV, MVDR-INT-sdw, and MVDR-INT-cnt when $\mu = 0.1$ for the scenario of Fig. 3.5.

MVDR-APR. This behaviour is indeed consistent with the observations from the SPP of Fig. 3.6, where it was the lower frequencies in the final 20 s that suffered from some misclassification.

Both tuning rules have demonstrated that the MVDR-INT can indeed be a useful strategy as it has exhibited a performance that is more accomodating for changes in the acoustic environment as compared to using either the MVDR-APR, MVDR-EST, or LCMV only. The resulting audio signals from this experiment and a repeated version that uses a babble noise may be listened to for a subjective evaluation at [211].

3.9 Conclusion

This chapter has investigated a confidence-based approach to integrating a

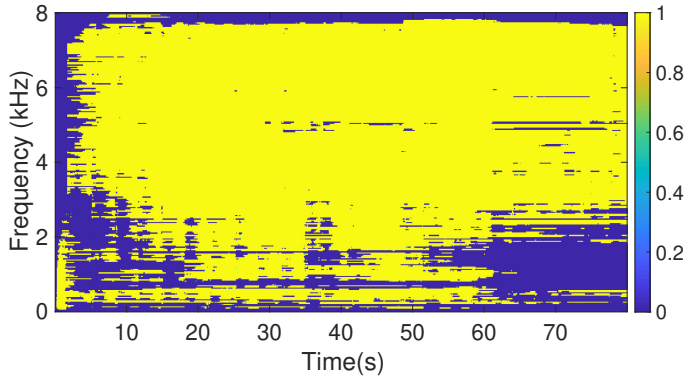


Figure 3.9: Metric of confidence, F(k, l) from the results of Fig. 3.7 and Fig. 3.8.

priori and data-dependent spatial information, in the form of a relative transfer function (RTF) vector for an MVDR beamformer. Whereas conventionally, the MVDR beamformer uses either an a priori RTF vector (MVDR-APR) or a data-dependent RTF vector (MVDR-EST) as a hard constraint, an alternative approach of using an integrated MVDR (MVDR-INT) beamformer was developed, where both the hard constraints from the two conventional approaches are softened to yield two tuning parameters, $\bar{\alpha}$ and $\bar{\beta}$.

It was found that the MVDR-INT could be expressed as a convex combination of the MVDR-APR beamformer, the MVDR-EST beamformer, a linearly constrained minimum variance (LCMV) beamformer, and an all-zero vector, with real, positive-valued coefficients. The effect of different tuning combinations of $\bar{\alpha}$ and $\bar{\beta}$ on the behaviour of the MVDR-INT beamformer was then explored theoretically and confirmed through simulations in cases where a data-dependent RTF vector estimate was accurate and where it was inaccurate, demonstrating that the MVDR-INT encompasses a wide range of speech enhancement filters.

It was then proposed that the posterior SNR resulting from the MVDR-EST beamformer could be used as a potential metric of confidence to indicate the accuracy of the data-dependent RTF vector. By making use of this confidence metric and the relation of the MVDR-INT to the speech-distortion-weighted Multi-channel Wiener Filter, two tuning rules were subsequently suggested to prescribe values for $\bar{\alpha}$ and $\bar{\beta}$ in practice. An evaluation with recorded data from an office room demonstrated that the MVDR-INT beamformer can indeed provide a more versatile beamformer by merging the benefits from the MVDR-APR and MVDR-EST beamformers. In particular it has shown a more

robust performance in terms of SNR and speech intelligibility for changes in the speech source location and inaccurately estimated correlation matrices.

Although it has been proposed to use the posterior SNR obtained with the MVDR-EST beamformer as the confidence metric, it is not to say that it is the only possible confidence metric that can be used. For instance, in [212], a confidence metric based on a computation of the input SNR has been used. Nevertheless, it raises the question as to what other potential confidence metrics can be used in practice and how well would they truly correlate to the accuracy of the data-dependent RTF vector. While this is an interesting avenue to explore, it is left as a separate body of work (see future work and perspectives in Chapter 9) as this dissertation mainly focuses on the development of the speech enhancement frameworks. Another point to consider, however, is that manual control of the weightings may also be given to the user themselves so that they have the ability to control their preferred sound. It may therefore be envisioned that in an assistive hearing device for instance, there may be two options: one where the confidence metric is used to automatically control optimal combinations, and another where the user has manual control over making such combinations, which can perhaps even be used to "train" the confidence metric.

On the relation to robust beamformers, it should be noted that the integrated MVDR beamformer is not intended to be better or worse than any of the particular robust beamformers. In the robust beamformers, one distortionless constraint is prescribed and the mismatch in relation to that single constraint is intended to be minimised. In the integrated MVDR beamformer on the other hand, two distortionless constraints are prescribed, one of which may be closer to the true distortionless constraint. The process of trading off between these two constraints is how robustness is intended to be maintained from the point of view of the integrated MVDR beamformer. Although these are two different approaches to address spatial robustness, it is not far-fetched to imagine that techniques from existing robust beamformers can be incorporated into the integrated MVDR beamformer and such an investigation should be explored in future work.

As a final remark, the integrated MVDR beamformer is not restricted to the use of the a priori RTF vector and the subspace-based data-dependent RTF vector. Alternatively, two differently estimated data-dependent RTF vectors may be used in the framework. This does not change the equations developed throughout the chapter, but rather the interpretation would be that the contingency option is a differently estimated data-dependent RTF vector as opposed to the a priori RTF vector.

Chapter 4

Extension of microphone array processing with an external microphone

4.1 Introduction

4.1.1 The use of external microphones

One of the applications that has largely benefited from the use of external microphones (XMs) or remote microphones as they are sometimes referred to, is that of speech intelligibility improvement in assistive hearing devices such as hearing aids (HAs) and cochlear implants (CIs). If an XM is placed near to a target speaker or other audio signal such as a television or radio, then listening directly to this higher SNR external signal (higher than those from the enhanced signal using only the microphones on the HA or CI) would presumably result in an improved speech intelligibility. Early research involving XMs was therefore focused on how to wirelessly transmit the XM signal to the HA or CI.

One of the earliest techniques of wirelessly transmitting external signals to HAs was via the use of a telecoil, a small magnetic metal core cylinder wrapped in copper wire that could pick up electro-magnetic signals [213]. Typically the signal from an XM placed on the target speaker (such as a teacher in a classroom) could be sent through an induction loop, creating an electromagnetic signal, which could then be received by the telecoil in the HA. Some of the

initial problems encountered with the telecoil was its optimal placement within the HA and ensuring that the magnetic field was sufficient throughout the looped area [214]. While such issues were being solved, radio frequency (RF) transmission methods then emerged[1] for wireless transmission of audio, with frequency modulation (FM) techniques being quite beneficial in a variety of listening conditions [214]. Some other advantages of the FM systems included the ability to transmit over long distances and independent operation on different channels. On the other hand, FM systems were susceptible to electromagnetic interference, could be high cost, and required separate hardware [213]. Another type of wireless transmission used was near-field magnetic induction (NFMI), which involved communication of a low power, non-propagating magnetic field among devices, i.e. communication in the near-field, which is in contrast to the far-field communication in the FM systems. The benefit with this near-field communication, was a reduced susceptibility to external noises and increasing security with signal encryption [213], but its range of communication was smaller. Even though the NFMI system has a higher power density than that of an FM system for small distances, the power density roll off for the NFMI is $-60\,\mathrm{dB}$ per decade with respect to distance, whereas the FM system has a $-20\,\mathrm{dB}$ per decade power density roll off [213]. More recently, Bluetooth has been considered as a suitable protocol that can be used for wireless transmission. While early Bluetooth systems suffered from excessive delays and heavy power consumption, the introduction of Bluetooth 4.0 and Bluetooth Low Energy, specifically targeted on low power consumption and low latency, are currently (at the time of this writing) being used on some modern devices.

With the successful development of wireless protocols, many HA and CI systems have been using an XM in a strategy where the XM signal is wirelessly transmitted to the hearing device and then the user may switch between listening to the sound of the XM signal or the sound of the enhanced signal using only the microphones on the hearing device. It has been shown that using an XM signal indeed provides a substantial benefit to the hearing impaired users [51, 214–216]. It was however noted in [215] that some subjects expressed concerns of persistent noise and particular annoyances[2] in very noisy environments as well as the problem of spatially localising the desired speaker. As opposed to switching between the use of the on board hearing device microphones and an XM, the natural question was then if it was possible to use all of the signals in conjunction for an even further improved speech intelligibility.

The signals from either a 'local' microphone array (LMA) (such as the

[1] Research into RF was actually triggered by the large number of hearing impaired children as a result of the rubella epidemic in the United States in 1964 and 1965 [214].

[2] Some interesting comments are noted in [215] such as someone not enjoying the sound of their husband chewing cereal.

microphones on the assistive hearing devices) or the XM only provide a localised spatial sampling of the acoustic environment and hence signal processing techniques with either of these signals are inherently limited. Using all of the microphone signals on the other hand, increases the spatial diversity as information would be available at both the locations of the LMA and at any XMs that are in the acoustic environment. Therefore, it would be expected that this additional spatial sampling would be quite beneficial when using the optimal filters described in Sec. 2.7.1. As devices embedded with microphones are becoming more ubiquitous, for instance, many persons have smartphones which are equipped with at least one microphone, it implies that there is a large spatial diversity to be exploited within the acoustic environment if all of these signals could be used in conjunction. Such a motivation has triggered significant research into the field of wireless acoustic networks (WASNs) [17, 45–47], which are are a specific class of wireless sensor networks [48].

In a WASN, each microphone or set of microphones is considered as a *node* within the network, which exchanges information accordingly to perform the audio DSP task. In the case of an HA for instance, the LMA on the HA device of the user can be considered as a node and each XM belonging to say the smart phone of a group of persons interacting with the HA user can be treated as separate nodes to constitute the WASN. All of the acquired signals may be broadcast to a fusion centre for processing in what is known as a *centralised* processing scheme, or the processing may be shared throughout the WASN in what is known as *distributed* processing. One such distributed processing algorithm is the distributed noise-specific signal estimation (DANSE) algorithm [45]. Several other distributed audio DSP algorithms have also been proposed in the recent years [217–221].

For the case of assistive hearing devices, the XM(s) have been proposed to be used in a number of ways. In [222], a robust version of the DANSE algorithm was applied for speech enhancement. Variants of the MWF (recall Sec. 2.7.1) have been used in [223–225] for preservation of binaural cues for HA users. In [226], the use of the XM as a noise reference for speech enhancement was analysed while taking into account, issues associated with the wireless transmission of the audio signal. For single microphone HAs, the procedure in [227] used the XM to design a post-filter in order to resolve a front-back ambiguity. A different approach altogether used an XM (typically worn on the desired speaker) to estimate the sound direction of arrival (DoA) and then applied the appropriate binaural cues onto the "clean" XM signal [228, 229].

4.1.2 Signal processing challenges with external microphones

The algorithms presented in this dissertation pertain mostly to an application layer level of an Open Systems Interconnection (OSI) model [230], but it is nevertheless important to recognise that there are still several challenges to overcome before algorithms with an LMA and one or more XMs can be effectively used in practice. Using an LMA in conjunction with one or more XMs can be considered as a wireless acoustic network (WASN) [45–47] and hence there several issues that are encountered as opposed to a wired context. In the following, some of the more important challenges are highlighted.

Communication bandwidth constraints. In order to use the optimal beamformers from Sec. 2.7.1, the microphone nodes require a certain amount of communication bandwidth to exchange their signals. For instance, in a scenario of centralised processing, as the number of microphones increases, so too does the required communication bandwidth and hence energy consumption. However, there is usually a constraint on the amount of communication bandwidth that is available and hence it must be used in an efficient manner.

Techniques for data compression are therefore quite important and should be considered on the various levels of the OSI model. In the application layer, *fusion* rules can be applied, which optimally combines several signals to yield one signal which is an improved estimate of a target signal. As opposed to transmitting multiple signals, only this "fused" signal can be transmitted for further processing. These types of fusion rules should, in fact, not be so alien to us as the optimal beamformers that have been discussed are all examples of fusion rules for compressing the data.

On the network level, data compression may take the form of source coding techniques to appropriately encode the transmitted data and decode it accordingly at the receiver, which will reduce the communication bandwidth as well as match specific rate constraints in wireless links [45]. Consequently there is a trade-off between the minimum bit rate that can be achieved and the amount of tolerable distortion. Theoretical optimal procedures in the context of wireless hearing aids have been proposed in [231] and practical investigation has been pursued in [232] [233].

Microphone subset selection is another technique for reducing communication bandwidth and has the additional benefit of conserving energy. In a WASN, not all signals are necessary for obtaining an acceptable level of performance. Depending on the location of sensors, they may only marginally contribute to the overall performance and in such a case should be put to sleep for conserving communication bandwidth and energy. Some approaches to this

problem involve computing a utility metric related to a mean squared error cost function as in [234] or minimising a transmission cost with a constraint on the noise reduction performance [235].

Clock synchronisation. Different nodes in a WASN operate with a separate clock, for instance with an LMA on a hearing aid will have a different clock from that of an XM in a mobile phone. As these clocks are subject to imperfections and offset, it is inevitable that they will be asynchronous at times, which can reduce the performance of audio signal processing algorithms [236]. Therefore synchronisation of the clocks within the WASN is an important issue to address [237]. Existing methods for clock synchronisation among various nodes in a WASN may involve broadcasting an RF reference signal to all devices [238] or blindly estimating the sample rate offsets [239] [240]. Once the sample rate offset is estimated, the signals can then be resampled at the same rate. More recent approaches have however avoided the need for resampling by considering partially synchronised nodes, i.e. where only subsets of nodes share the same clock and are synchronised, while others are not [241].

Communication delay. The delay in wirelessly streaming audio between nodes within a network is one of the more important criteria for applications involving assistive hearing devices. The optimal filters that were previously discussed essentially attempt to align the speech component in the various microphone signals that originate from the same source. However, such alignment is only possible if the STFT frame lengths used for computing the filter are at least twice the maximum time difference of arrival (TDOA) among the microphones [222]. This TDOA will include both the acoustic propagation delay from the desired source to the microphones as well as the delay of the communication link among the microphones. At a sampling frequency of 16 kHz, an STFT frame length of 128 samples corresponds to 8 ms. Therefore, the maximum permissible TDOA would have to be less than 4 ms. Furthermore, the delay should be also fixed as algorithms may be too slow to track stochastic delays. Currently, RF, Near-Field Magnetic Induction (NFMI), and Bluetooth systems are all used for for streaming wireless audio to HAs for instance from mobile phones [213].

Distributed processing. As opposed to the centralised scheme of processing, processing may be alternatively *distributed* across the entire WASN, i.e. processing may be done at different points in the network using a subset of nodes. This can result in reduction of communication bandwidth and energy savings as only certain nodes would be required to exchange their information. Distributed processing also comes with the additional benefit that algorithms

become scalable as the number of microphones in the network grows, i.e. as the network grows, there is minimal impact on the resources (such as communication bandwidth, energy savings, etc.) that are required.

In the DANSE algorithm, for example, individual nodes update their appropriate local fusion rules to transmit compressed data across the WASN [45]. Under conditions where the number of speakers is small, it has been shown that DANSE even converges to the centralised case. In other words, the enhanced signal at the particular node is equivalent to that which would have been obtained had all the signals been used in a centralised manner. Such a convergence should however, be fast enough such that the acoustic scenario does not change before convergence is obtained, a criterion which is not so easily controlled in practice.

Topology. The topology or organisation in which the various nodes share their respective information is another important consideration within a WASN. A typical consideration may be that nodes which are far away or are not in a particular line of sight with respect to another node may be excluded as it will not contribute much valuable information. Hence a topology employing a nearest neighbour communication may be a preferred option. Nevertheless, the topology should be optimised in general for energy conservation and delays. In relation to the topology, other intelligent decisions also need to be made for cases where there may be noisy communication links or even broken communication links. Various topologies for the DANSE algorithm have been analysed in [46], which also includes a topology independent strategy.

While there are still several challenges open for practical implementation, it is still worthwhile to understand quantitatively the potential benefits that can be achieved with an XM. In the remaining of this chapter, firstly a perspective on the notion of beamforming with an LMA and an XM is presented. In Sec. 4.4 some fundamental results on how an XM can improve the performance of some of the optimal filters from Sec. 2.7.1 are presented. Conclusions and the outline for the remaining chapters are then given in Sec. 4.5.

4.2 Beam patterns of a local microphone array with an external microphone

In Sec. 2.3, it was observed how beam patterns could be formed for a microphone array by analysing the frequency-wavenumber response function, parametrised by the incidence angle of the source signal in (2.32). Such a convenient expression could however only be derived by considering the source in the far field with

respect to the microphone array. With the introduction of one or more XMs, this far field assumption is no longer valid as it is quite likely that an XM is near to the source signal, for instance, if the target speaker is wearing the XM. Therefore, the frequency-wavenumber response function cannot be parametrised any more in terms of an incidence angle of the source signal, which obscures our previous picture of beam patterns.

Nevertheless, beamforming with XMs can be conceptualised by observing the frequency-wavenumber response as a function of space in (2.30), i.e. specifically as a function of the distance between the source position and the different microphone positions. A normalised version of the vector representation of the Green's function in the free-field from (2.29), with respect to the first microphone of the LMA can be firstly defined as follows:

$$\breve{\boldsymbol{f}}(\omega, \mathbf{r}_m^{\mathrm{s}}) = \begin{bmatrix} 1 \\ \frac{||\mathbf{r}_{\mathrm{s}}-\mathbf{r}_1||}{||\mathbf{r}_{\mathrm{s}}-\mathbf{r}_2||} e^{-j\frac{\omega(||\mathbf{r}_{\mathrm{s}}-\mathbf{r}_2||-||\mathbf{r}_{\mathrm{s}}-\mathbf{r}_1||)}{c}} \\ \vdots \\ \frac{||\mathbf{r}_{\mathrm{s}}-\mathbf{r}_1||}{||\mathbf{r}_{\mathrm{s}}-\mathbf{r}_{\mathrm{M_a}}||} e^{-j\frac{\omega(||\mathbf{r}_{\mathrm{s}}-\mathbf{r}_{\mathrm{M_a}}||-||\mathbf{r}_{\mathrm{s}}-\mathbf{r}_1||)}{c}} \\ \frac{||\mathbf{r}_{\mathrm{s}}-\mathbf{r}_1||}{||\mathbf{r}_{\mathrm{s}}-\mathbf{r}_{\mathrm{M_a}+1}||} e^{-j\frac{\omega(||\mathbf{r}_{\mathrm{s}}-\mathbf{r}_{\mathrm{M_a}+1}||-||\mathbf{r}_{\mathrm{s}}-\mathbf{r}_1||)}{c}} \\ \vdots \\ \frac{||\mathbf{r}_{\mathrm{s}}-\mathbf{r}_1||}{||\mathbf{r}_{\mathrm{s}}-\mathbf{r}_{\mathrm{M_a}+\mathrm{M_e}}||} e^{-j\frac{\omega(||\mathbf{r}_{\mathrm{s}}-\mathbf{r}_{\mathrm{M_a}+\mathrm{M_e}}||-||\mathbf{r}_{\mathrm{s}}-\mathbf{r}_1||)}{c}} \end{bmatrix} \quad \text{under free-field conditions} \tag{4.1}$$

where $\mathrm{M_a}$ is the number of microphones in the LMA and $\mathrm{M_e}$ is the number of XMs. As in (2.29), $\mathbf{r}_m$ is the position of the mth microphone but now for $m \in [1, 2, \ldots, \mathrm{M_a} + \mathrm{M_e}]$ and the index $\mathbf{r}_m^{\mathrm{s}}$ indicates that $\breve{\boldsymbol{f}}$ is a function of the vector $[||\mathbf{r}_{\mathrm{s}} - \mathbf{r}_1||\ ||\mathbf{r}_{\mathrm{s}} - \mathbf{r}_2||\ \ldots\ ||\mathbf{r}_{\mathrm{s}} - \mathbf{r}_{\mathrm{M_a}+\mathrm{M_e}}||]^T$, i.e. a vector of distances from the source position, $\mathbf{r}_{\mathrm{s}}$, to all microphone positions.

As there is no more parametrisation in terms of incidence angle of the source, in order to visualise the spatial response from a particular beamformer, the frequency-wavenumber response needs to be computed in a 2D or 3D cartesian space. In other words, there must be a parametrisation over a set of spatial positions that would correspond to a virtual source. Hence it is useful to define another steering vector, $\mathbf{v}(\omega, \mathbf{r}_m^{\mathrm{v}})$, that will correspond to these virtual sources and positions, $\mathbf{r}_{\mathrm{v}}$. This steering vector, $\mathbf{v}(\omega, \mathbf{r}_m^{\mathrm{v}})$ can be defined identically to (4.1) except that $\mathbf{r}_{\mathrm{s}}$ is replaced with the position of the virtual source, $\mathbf{r}_{\mathrm{v}}$. Using $\mathbf{v}(\omega, \mathbf{r}_m^{\mathrm{v}})$, the frequency-wavenumber response function from (2.30) is now

re-defined as:

$$\breve{\Psi}(\omega, \mathbf{r}^{\mathrm{s}}_m) = \boldsymbol{w}^H(\omega)\mathbf{v}(\omega, \mathbf{r}^{\mathrm{v}}_m) \tag{4.2}$$

where $\boldsymbol{w}(\omega) = [w_1(\omega)\ w_2(\omega)\ \ldots w_{\mathrm{M_a+M_e}}(\omega)]^T$ is the beamformer with the dimension of $\mathrm{M_a}+\mathrm{M_e}$. By defining an appropriate grid of positions with $\mathbf{v}(\omega, \mathbf{r}^{\mathrm{v}}_m)$, $|\breve{\Psi}(\omega, \mathbf{r}^{\mathrm{s}}_m)|$ can now be computed and visualised for a given set of microphone positions and a particular beamformer. When the position of the virtual source is that of the true source, then $\mathbf{r}^{\mathrm{v}}_m = \mathbf{r}^{\mathrm{s}}_m$, and hence $\mathbf{v}(\omega, \mathbf{r}^{\mathrm{v}}_m) = \breve{\boldsymbol{f}}(\omega, \mathbf{r}^{\mathrm{s}}_m)$.

In the following, the beam patterns are observed for three different scenarios as depicted in Fig. 4.1 for a 6 m × 6 m 2D cartesian space. The source position was always at the coordinates $\{3\,\mathrm{cm}, 4.5\,\mathrm{cm}\}$ and a linear LMA with $\mathrm{M_a} = 4$ and inter-element spacing, $d = 5\,\mathrm{cm}$, was placed such that the uppermost and reference microphone in the LMA was located at $\{3\,\mathrm{cm}, 3\,\mathrm{cm}\}$ in an end-fire direction in relation to the source. Hence the distance between the uppermost microphone of the LMA and the source was fixed to 1.5 m in all scenarios. The scenario in Fig. 4.1 (a) consisted of only the LMA, the scenario in Fig. 4.1 (b) consisted of the same LMA and one XM at $\{3\,\mathrm{cm}, 3.5\,\mathrm{cm}\}$, and the the scenario in Fig. 4.1 (b) consisted of the same LMA and one XM now at $\{3\,\mathrm{cm}, 4\,\mathrm{cm}\}$, i.e. closer to the source.

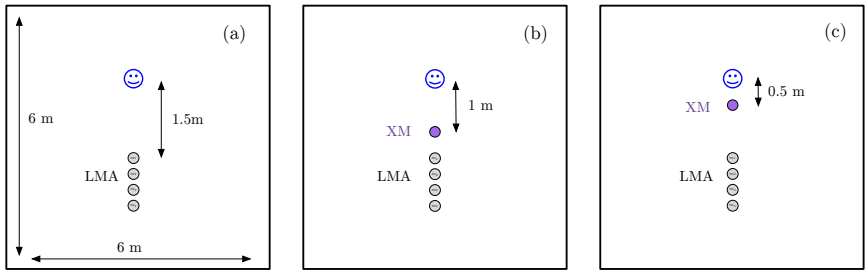


Figure 4.1: Three different scenarios used to observe the beam patters, $|\breve{\Psi}(\omega, \mathbf{r}^{\mathrm{s}}_m)|$ as defined in (4.2). Details of each scenario are given in the text.

Fig. 4.2 illustrates the resulting beam patterns, i.e. $|\breve{\Psi}(\omega, \mathbf{r}^{\mathrm{s}}_m)|$ for the scenarios depicted in Fig. 4.1 for the matched filter beamformer defined as:

$$\boldsymbol{w}_{mf}(\omega) = \frac{\breve{\boldsymbol{f}}(\omega, \mathbf{r}^{\mathrm{s}}_m)}{\breve{\boldsymbol{f}}(\omega, \mathbf{r}^{\mathrm{s}}_m)^H \breve{\boldsymbol{f}}(\omega, \mathbf{r}^{\mathrm{s}}_m)} \tag{4.3}$$

In Fig. 4.2, $|\breve{\Psi}(\omega, \mathbf{r}^{\mathrm{s}}_m)|$ was evaluated in 1 cm intervals within the 6 m × 6 m grid. Each row of Fig. 4.2 displays $|\breve{\Psi}(\omega, \mathbf{r}^{\mathrm{s}}_m)|$ for a different value of the ratio, $\frac{d}{\lambda}$. It should be noted that $|\breve{\Psi}(\omega, \mathbf{r}^{\mathrm{s}}_m)|$ is not plotted (set to zero) for regions

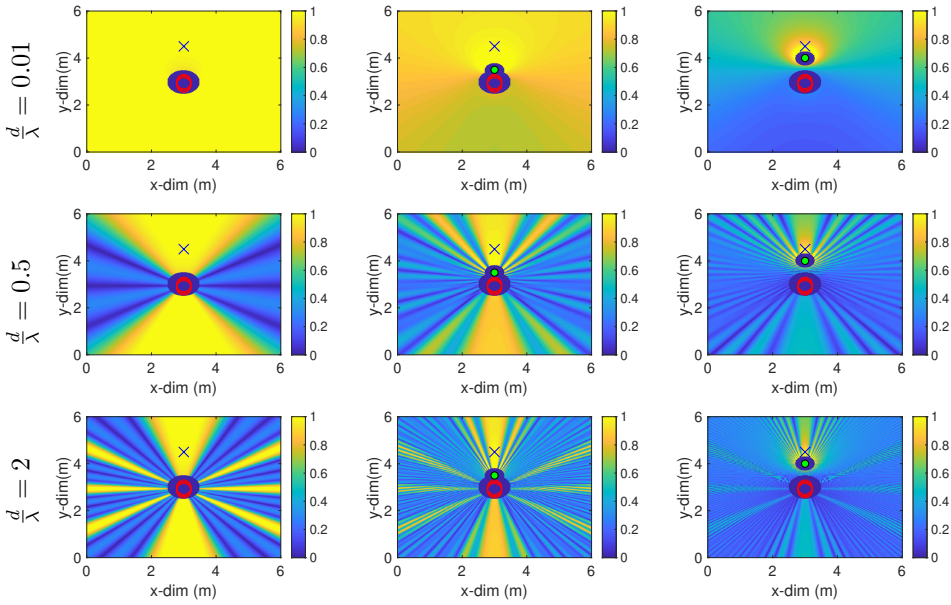


Figure 4.2: Beam patterns, $|\breve{\Psi}(\omega, \mathbf{r}_m^{\mathrm{s}})|$, for different microphone configurations involving an LMA and an XM for the matched filter, $\boldsymbol{w}_{mf}(\omega)$ at different ratios of $\frac{d}{\lambda} = \{0.01, 0.5, 2\}$ (Further description in the text). The values of $|\breve{\Psi}(\omega, \mathbf{r}_m^{\mathrm{s}})|$ are denoted by the colourbar. The x marker represents the speech source, the group of circles at the centre represents the LMA, and the highlighted circle in the middle and right columns represents the XMs.

very close to any of the microphones as $\mathbf{v}(\omega, \mathbf{r}_m^{\mathrm{v}})$ would be undefined (see the denominator of the elements from (4.1)).

Focusing firstly on the left column of Fig. 4.2, as expected, the same behaviour that was observed for the beam patterns in Sec. 2.3 is evident once again. In using the LMA alone, at lower ratios of $\frac{d}{\lambda}$, an omnidirectional type of pattern can be observed, with increasing directionality as $\frac{d}{\lambda}$ increases. At the highest ratio of $\frac{d}{\lambda}$, when $\frac{d}{\lambda} > 0.5$, spatial aliasing can also be observed. In the middle column of Fig. 4.2, when the XM is introduced, at $\frac{d}{\lambda} = 0.01$, it can be seen that the beam pattern is already more directional than using the LMA only. The directionality of the main beam continues to increase as $\frac{d}{\lambda}$ increases. With the introduction of the XM, the effective microphone spacing is now larger than d, which also implies that spatial aliasing will occur at lower ratios of $\frac{d}{\lambda}$ in

comparison to using the LMA only[3]. This is evident from the beam patterns at $\frac{d}{\lambda} = 0.5$ and $\frac{d}{\lambda} = 2$. Although aliasing occurs at lower ratios of $\frac{d}{\lambda}$, it can been seen that the sidelobes generated are of a lower amplitude with respect to the main beam. Finally, in the right column of Fig. 4.2, where the position of the XM is now even closer to the source position, the increased directionality is clearly visible across all ratios of $\frac{d}{\lambda}$. Once again, the spatial aliasing occurs at lower ratios of $\frac{d}{\lambda}$, but the sidelobes are generally of a lower amplitude, which is particularly evident by observing the last row of plots in Fig. 4.2. These plots are a clear indication that including an XM along with an LMA for spatial processing can prove to be quite advantageous. The beam patterns for various configurations of one or multiple XMs with the LMA can be observed from the code provided in [95].

4.3 Signal model with a local microphone array and external microphones

Through the remaining chapters, the microphone configuration will consist of an LMA with $\mathrm{M_a}$ microphones and $\mathrm{M_e}$ XMs. The same signal model in the STFT domain from (2.52) will be used, however with some minor changes to the notation so as to specifically refer to the LMA and XMs. This section simply serves to highlight those details and can be referred to throughout the rest of the dissertation. As in (2.52), the observable vector of ($\mathrm{M_a}$ + $\mathrm{M_e}$) microphone signals the STFT domain can be modelled as:

$$\mathbf{y}(\mathrm{k},\mathrm{l}) = \underbrace{\mathbf{h}(\mathrm{k},\mathrm{l})\mathrm{s_{a,1}}(\mathrm{k},\mathrm{l})}_{\mathbf{x}(\mathrm{k},\mathrm{l})} + \mathbf{n}(\mathrm{k},\mathrm{l}) \tag{4.4}$$

where now, with the inclusion of the XMs, the notation is introduced such that $\mathbf{y}(\mathrm{k},\mathrm{l}) = [\mathbf{y_a}(\mathrm{k},\mathrm{l})^T\ \mathbf{y_e}(\mathrm{k},\mathrm{l})^T]^T$, $\mathbf{y_a}(\mathrm{k},\mathrm{l}) = [\mathrm{y_{a,1}}(\mathrm{k},\mathrm{l})\ \mathrm{y_{a,2}}(\mathrm{k},\mathrm{l})\ \ldots \mathrm{y_{a,M_a}}(\mathrm{k},\mathrm{l})]^T$ is a vector of the LMA signals, $\mathbf{y_e}(\mathrm{k},\mathrm{l}) = [\mathrm{y_{e,1}}(\mathrm{k},\mathrm{l})\ \mathrm{y_{e,2}}(\mathrm{k},\mathrm{l})\ \ldots \mathrm{y_{e,M_e}}(\mathrm{k},\mathrm{l})]^T$ is a vector of the XM signals, $\mathbf{x}(\mathrm{k},\mathrm{l})$ is the speech contribution, represented by $\mathrm{s_{a,1}}(\mathrm{k},\mathrm{l})$, the speech signal in the first microphone of the LMA[4], filtered with $\mathbf{h}(\mathrm{k},\mathrm{l}) = [\mathbf{h_a}^T(\mathrm{k},\mathrm{l})\ \mathbf{h_e}(\mathrm{k},\mathrm{l})^T]^T$, $\mathbf{h_a}(\mathrm{k},\mathrm{l})$ is the RTF vector for the LMA (with the first microphone used as the reference, i.e. the first component of $\mathbf{h_a}$ equal to 1), and $\mathbf{h_e}(\mathrm{k},\mathrm{l})$ is the RTF vector for the XM signals. Finally,

[3]In fact, we may think of the entire microphone configuration as consisting of two nested arrays, one which is the XM and the reference microphone of the LMA, and the other as the remaining elements of the LMA.

[4]In the chapters that follow, one of the microphones in the LMA will always be used as a reference microphone. Any microphone may however in general be used as a reference microphone.

$\mathbf{n}(\mathrm{k},\mathrm{l}) = [\mathbf{n_a}(\mathrm{k},\mathrm{l})^T\ \mathbf{n_e}(\mathrm{k},\mathrm{l})^T]^T$ represents the noise contribution. Variables with the subscript "$\mathbf{a}$" refer to the LMA and variables with the subscript "$\mathbf{e}$" refer to the XMs. In the following, the dependency on k and l will be dropped again as they are assumed to be independently treated.

The $(\mathrm{M_a} + \mathrm{M_e}) \times (\mathrm{M_a} + \mathrm{M_e})$ speech-plus-noise, noise-only, and speech-only spatial correlation matrices are given respectively as:

$$\mathbf{R_{yy}}(\mathrm{k},\mathrm{l}) = \mathbb{E}\{\mathbf{y}(\mathrm{k},\mathrm{l})\mathbf{y}(\mathrm{k},\mathrm{l})^H\} \tag{4.5}$$

$$\mathbf{R_{nn}}(\mathrm{k},\mathrm{l}) = \mathbb{E}\{\mathbf{n}(\mathrm{k},\mathrm{l})\mathbf{n}(\mathrm{k},\mathrm{l})^H\} \tag{4.6}$$

$$\mathbf{R_{xx}}(\mathrm{k},\mathrm{l}) = \mathbb{E}\{\mathbf{x}(\mathrm{k},\mathrm{l})\mathbf{x}(\mathrm{k},\mathrm{l})^H\} \tag{4.7}$$

Similar correlation matrices can also be defined for only the LMA signals or only the XM signals. It is once again assumed that the speech signal is uncorrelated with the noise signal so that:

$$\mathbf{R_{yy}}(\mathrm{k},\mathrm{l}) = \mathbf{R_{xx}}(\mathrm{k},\mathrm{l}) + \mathbf{R_{nn}}(\mathrm{k},\mathrm{l}) \tag{4.8}$$

For a single target speaker, then $\mathbf{R_{xx}}$ can be modelled as a rank-1 matrix:

$$\mathbf{R_{xx}}(\mathrm{k},\mathrm{l}) = \sigma^2_{\mathrm{s_{a,1}}}(\mathrm{k},\mathrm{l})\mathbf{h}(\mathrm{k},\mathrm{l})\mathbf{h}(\mathrm{k},\mathrm{l})^H \tag{4.9}$$

where $\sigma^2_{\mathrm{s_{a,1}}}(\mathrm{k},\mathrm{l}) = \mathbb{E}\{\mathrm{s_{a,1}}(\mathrm{k},\mathrm{l})\mathrm{s_{a,1}}(\mathrm{k},\mathrm{l})^*\}$ is the speech power spectral density (PSD) in a reference microphone of the LMA.

Throughout the remaining chapters, it will always be assumed that all signal correlations can be estimated as if all signals were available in a centralised processor, i.e., a perfect communication link is assumed between the LMA and the XM signals with no bandwidth constraints and with synchronous sampling.

4.4 Optimal beamforming with a local microphone array and external microphones

In the STFT domain, any of the optimal beamformers from Sec. 2.7.1 can also be applied to a configuration of microphones consisting of an LMA with $\mathrm{M_a}$ microphones and $\mathrm{M_e}$ XMs. In this section, it will be demonstrated that the introduction of XMs can provide an improved performance to some of these optimal beamformers as opposed to using a system of microphones consisting of an LMA only.

4.4.1 Array gain for an MVDR beamformer

From Sec. 2.7.1, the array gain for an MVDR beamformer is given by (the dependency on k and l is dropped):

$$\mathrm{AG}^{\mathrm{MVDR}} = \sigma^2_{\mathrm{n_{a,1}}} \mathbf{h}^H \mathbf{R}^{-1}_{\mathbf{nn}} \mathbf{h} \tag{4.10}$$

with $\sigma^2_{\mathrm{n_{a,1}}}$ being the noise PSD in the first microphone of the LMA and $\mathbf{R_{nn}}$ can be written as a block matrix as follows:

$$\mathbf{R_{nn}} = \begin{bmatrix} \mathbf{R}_{\mathbf{n_a n_a}} & \mathbf{R}_{\mathbf{n_a n_e}} \\ \mathbf{R}^H_{\mathbf{n_a n_e}} & \mathbf{R}_{\mathbf{n_e n_e}} \end{bmatrix} \tag{4.11}$$

where $\mathbf{R}_{\mathbf{n_a n_a}} = \mathbb{E}\{\mathbf{n_a} \mathbf{n_a}^H\}$ is the $(\mathrm{M_a} \times \mathrm{M_a})$ noise-only correlation matrix for the LMA signals, $\mathbf{R}_{\mathbf{n_a n_e}} = \mathbb{E}\{\mathbf{n_a} \mathbf{n_e}^H\}$ is the $(\mathrm{M_a} \times \mathrm{M_e})$ cross-correlation matrix between the LMA signals and the XM signals and $\mathbf{R}_{\mathbf{n_e n_e}} = \mathbb{E}\{\mathbf{n_e} \mathbf{n_e}^H\}$ is the noise-only correlation matrix for the XM signals. Applying a blockwise matrix inversion [242] to (4.11) then results in:

$$\mathbf{R}^{-1}_{\mathbf{nn}} = \begin{bmatrix} \mathbf{R}^{-1}_{\mathbf{n_a n_a}} + \mathbf{R}^{-1}_{\mathbf{n_a n_a}} \mathbf{R}_{\mathbf{n_a n_e}} \mathbf{P}^{-1} \mathbf{R}^H_{\mathbf{n_a n_e}} \mathbf{R}^{-1}_{\mathbf{n_a n_a}} & -\mathbf{R}^{-1}_{\mathbf{n_a n_a}} \mathbf{R}_{\mathbf{n_a n_e}} \mathbf{P}^{-1} \\ -\mathbf{P}^{-1} \mathbf{R}^H_{\mathbf{n_a n_e}} \mathbf{R}^{-1}_{\mathbf{n_a n_a}} & \mathbf{P}^{-1} \end{bmatrix} \tag{4.12}$$

where $\mathbf{P} = \mathbf{R}_{\mathbf{n_e n_e}} - \mathbf{R}^H_{\mathbf{n_a n_e}} \mathbf{R}^{-1}_{\mathbf{n_a n_a}} \mathbf{R}_{\mathbf{n_a n_e}}$ is the Schur complement of $\mathbf{R}_{\mathbf{n_a n_a}}$. Decomposing $\mathbf{P}$ into its square-root matrices such that $\mathbf{P} = \mathbf{P}^{1/2} \mathbf{P}^{H/2}$ and substitution of (4.12) into (4.10) results in the array gain as:

$$\mathrm{AG}^{\mathrm{MVDR}}_{\mathrm{LMA+XM}} = \underbrace{\sigma^2_{\mathrm{n_{a,1}}} \mathbf{h}^H_{\mathbf{a}} \mathbf{R}^{-1}_{\mathbf{n_a n_a}} \mathbf{h}_{\mathbf{a}}}_{\mathrm{AG}^{\mathrm{MVDR}}_{\mathrm{LMA}}} + \underbrace{\sigma^2_{\mathrm{n_{a,1}}} ||\mathbf{P}^{-1/2} \mathbf{R}^H_{\mathbf{n_a n_e}} \mathbf{R}^{-1}_{\mathbf{n_a n_a}} \mathbf{h}_{\mathbf{a}} - \mathbf{P}^{-1/2} \mathbf{h}_{\mathbf{e}}||^2}_{\mathrm{AG}^{\mathrm{MVDR}}_{\mathrm{XM}}} \tag{4.13}$$

which consists of the array gain for the LMA only, $\mathrm{AG}^{\mathrm{MVDR}}_{\mathrm{LMA}}$ plus an additional term, $\mathrm{AG}^{\mathrm{MVDR}}_{\mathrm{XM}}$, that is due to the inclusion of the XMs. Due to the squared Euclidean norm in the $\mathrm{AG}^{\mathrm{MVDR}}_{\mathrm{XM}}$ term, it means that theoretically, the array gain of the system with an LMA and XMs is at least equal to or greater than the array gain for a system consisting only of an LMA.

In the case of $\mathrm{M_e} = 1$, the Schur complement of $\mathbf{R}_{\mathbf{n_a n_a}}$ becomes a scalar, ϱ, such that the array gain as demonstrated in [243] becomes:

$$\mathrm{AG}^{\mathrm{MVDR}}_{\mathrm{LMA+XM}} = \underbrace{\sigma^2_{\mathrm{n_{a,1}}} \mathbf{h}^H_{\mathbf{a}} \mathbf{R}^{-1}_{\mathbf{n_a n_a}} \mathbf{h}_{\mathbf{a}}}_{\mathrm{AG}^{\mathrm{MVDR}}_{\mathrm{LMA}}} + \frac{\sigma^2_{\mathrm{n_{a,1}}}}{\varrho} |\mathbf{r}^H_{\mathbf{n_{ae}}} \mathbf{R}^{-1}_{\mathbf{n_a n_a}} \mathbf{h}_{\mathbf{a}} - \mathrm{h_e}|^2 \quad \text{when } \mathrm{M_e} = 1 \tag{4.14}$$

where $\mathrm{h_e}$ is the RTF pertaining to the single XM, $\varrho = \sigma^2_{\mathrm{n_e}} - \mathbf{r}^H_{\mathbf{n_{ae}}}\mathbf{R}^{-1}_{\mathbf{n_a n_a}}\mathbf{r}_{\mathbf{n_{ae}}}$, $\mathbf{r}_{\mathbf{n_{ae}}} = \mathbb{E}\{\mathbf{n_a}\mathrm{n_e^*}\}$ is the $(\mathrm{M_a}\times 1)$ cross-correlation vector between the LMA signals and the XM, and $\sigma^2_{\mathrm{n_e}} = \mathbb{E}\{\mathrm{n_e n_e^*}\}$ is the noise power in the XM. Furthermore, if the noise field is homogeneous and the XM is sufficiently far away from the LMA, then it can be expected that the noise signals of the LMA will be uncorrelated with that of the XM signal and hence $\mathbf{r}_{\mathbf{n_{ae}}} = \mathbf{0}$. The array gain in this case then becomes:

$$\mathrm{AG}^{\mathrm{MVDR}}_{\mathrm{LMA+XM}} = \underbrace{\sigma^2_{\mathrm{n_{a,1}}}\mathbf{h}^H_{\mathbf{a}}\mathbf{R}^{-1}_{\mathbf{n_a n_a}}\mathbf{h}_{\mathbf{a}}}_{\mathrm{AG}^{\mathrm{MVDR}}_{\mathrm{LMA}}} + \sigma^2_{\mathrm{n_{a,1}}}\frac{|\mathrm{h_e}|^2}{\sigma^2_{\mathrm{n_e}}} \quad \text{when } \mathrm{M_e} = 1 \text{ and } \mathbf{r}_{\mathbf{n_{ae}}} = \mathbf{0} \tag{4.15}$$

which indicates that the $\mathrm{AG}^{\mathrm{MVDR}}_{\mathrm{LMA+XM}}$ increases as the noise power of the XM decreases and/or when $\mathrm{h_e}$ increases. If $\mathrm{h_e}$ is defined according to the free-field model (recall (2.31)), then a large $\mathrm{h_e}$ can be interpreted as being closer to the source signal.

4.4.2 Matched Filter and Superdirective Beamformers

As previously discussed in Sec. 2.7.1, there is a trade-off between the DI and the WNG for the MFB and the SDB. With the introduction of the XM, however, the nature of this trade-off is dependent on the position of the XM in relation to the source. Fig. 4.3 illustrates three different scenarios for which the DI and WNG were computed for the MFB and the SDB. The expressions for the DI and WNG are the same for both beamformers from Sec. 2.7.1, except that the dimensionality changes accordingly to accommodate the additional XM. In all cases, $\mathbf{h}$ was computed according to the free-field model in (4.1) and $\mathring{\mathbf{\Gamma}}_{\mathbf{nn}}$ from (2.62). The details of the various scenarios in Fig. 4.3 are as follows:

(a) A linear microphone array consisting of $\mathrm{M_a} = 4$ with a microphone spacing of 5 cm, which is equivalent to the scenario from Fig. 2.15.

(b) The same microphone array as in (a) but with an additional XM placed such that the new microphone configuration would be a linear microphone array of $\mathrm{M_a} = 5$ and a microphone spacing of 5 cm.

(c) The microphone array from (a), but with the XM now placed closer to the source (50 cm away from the source) along the end-fire direction.

Fig. 4.4 displays the resulting DI and WNG for the MFB and the SDB from each of the scenarios in Fig. 4.3. The scenario of Fig. 4.3 (a) is identical to that

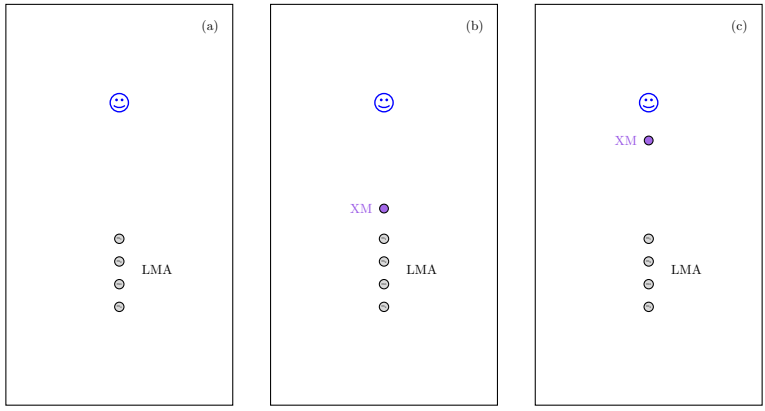


Figure 4.3: Different scenarios (a),(b),(c) outlined in the text for evaluation the MFB and SDB for using an LMA and an XM.

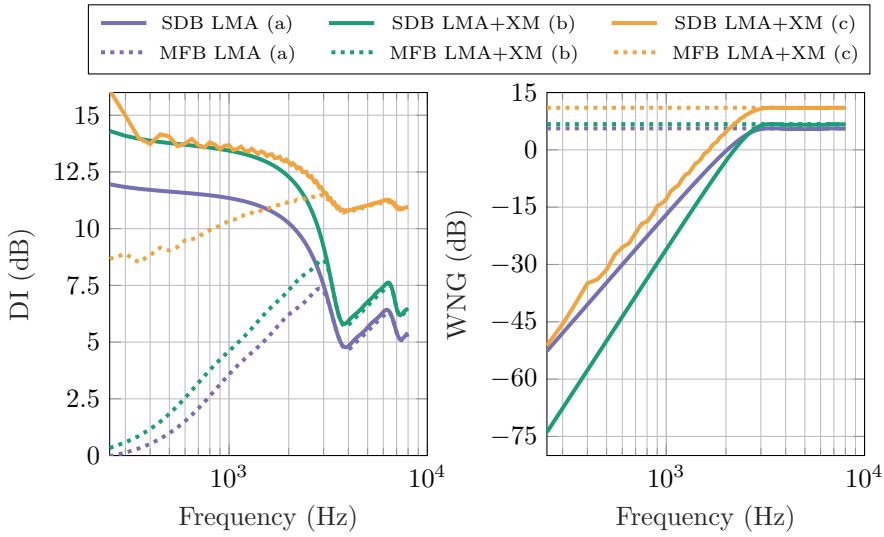


Figure 4.4: DI and WNG performance metrics for the MFB and SDB beamformers when using a microphone array and an external microphone.

from Fig. 2.15 and once again the trade-off between DI and WNG is observed for both beamformers. In the scenario of Fig. 4.3 (b), for the MFB, there is a small improvement in performance in comparison to the scenario of Fig. 4.3 (a) since the microphone configuration can be interpreted as an LMA with an additional microphone. For the SDB, however, while the DI improves, it can be

observed that the WNG significantly degrades for lower frequencies, which is even worse than the scenario of Fig. 4.3 (a). This is once again due to the rank deficient $\mathring{\mathbf{\Gamma}}_{\mathbf{nn}}$ which has an increased dimension with the XM as opposed to using the LMA only. For the scenario of Fig. 4.3 (c), the benefit of placing the XM close to the source is clearly observed. For the SDB, there is a significant increase in DI and WNG at higher frequencies as compared to the other two scenarios. Furthermore, for lower frequencies, the WNG is also better than the other two scenarios. Since the XM is closer to the source, and further from the LMA, the cross-correlation between the LMA signals and the XM is reduced, resulting in the $\mathring{\mathbf{\Gamma}}_{\mathbf{nn}}$ being less rank deficient than either of the two previous cases. This also impacts the MFB as there is also a significant improvement for its DI at lower frequencies.

While these results equip us with some understanding of how the XM provides an additional benefit to the MFB and SDB, it only applies for the case when the source is in the position for which $\mathbf{h}$ was defined. Hence, in order to gain some more intuition about the behaviour of the MFB and the SDB when an XM is used, a scenario was considered where the position of a source varied while the MFB and the SDB was designed to be fixed for a particular direction. Such a scenario will help us to understand the impact of the increased directionality from introducing an XM along with an LMA.

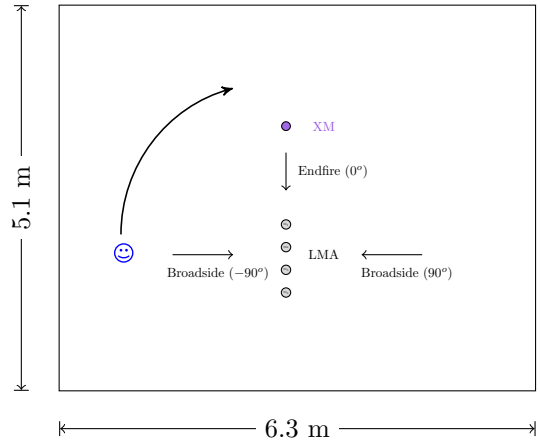


Figure 4.5: Plan view of simulation environment for observing the performance of the MFB and SDB as a function of source position. The arrow on the speech source indicates that simulations were done for different speech source angles. Not shown is the simulated diffuse noise field. The room height was 2.5 m.

The simulation scenario (plan view) is depicted in Fig. 4.5. In an anechoic room

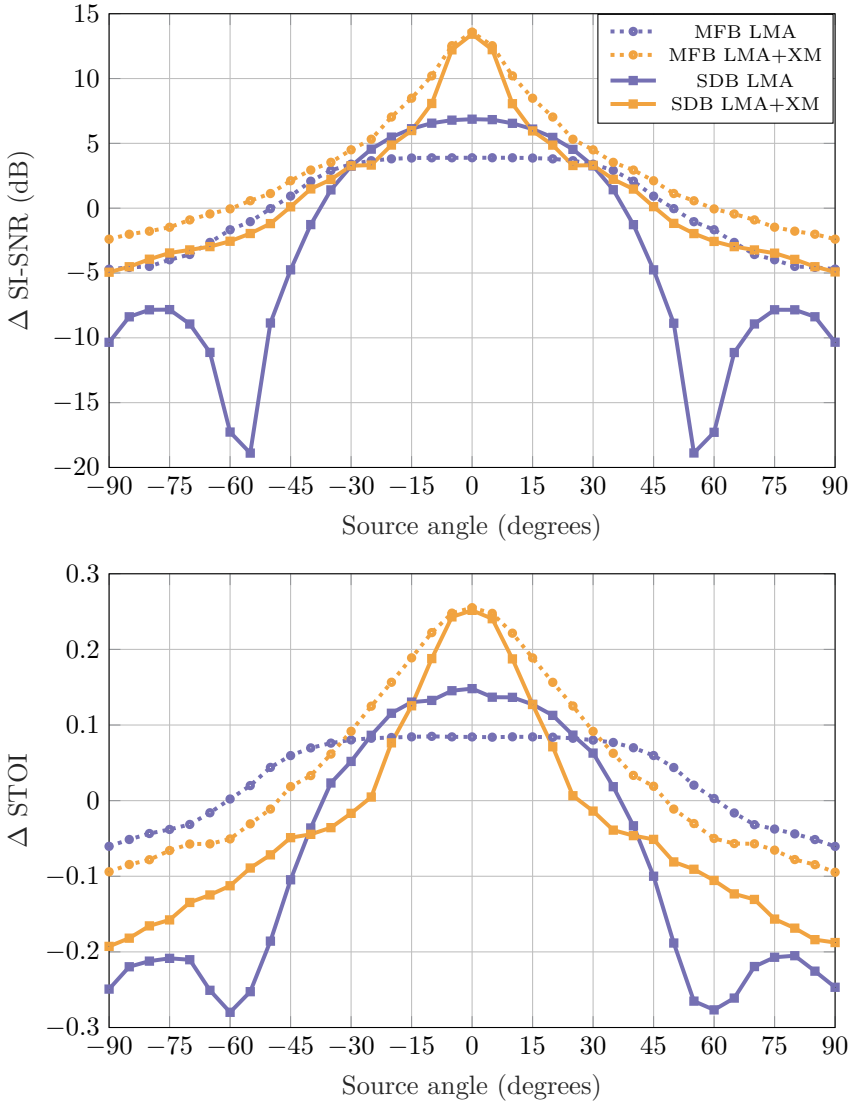


Figure 4.6: (Top) Δ SI-SNR for the MFB and SDB with the LMA only and with the LMA and XM as a function of source angle. The input SI-SNR at the first microphone was −3 dB. (Bottom) Δ STOI for the MFB and SDB with the LMA only and with the LMA and XM as a function of source angle. The input STOI at the first microphone was 0.68.

of dimensions $6.3\,\mathrm{m} \times 5.1\,\mathrm{m} \times 2.5\,\mathrm{m}$, an LMA with $\mathrm{M_a} = 4$ and inter-element spacing of 5 cm was placed at $\{3.6\,\mathrm{m}, 1.9\,\mathrm{m}, 1.4\,\mathrm{m}\}$, and an XM was placed at $\{3.6\,\mathrm{m}, 3.025\,\mathrm{m}, 1.4\,\mathrm{m}\}$. The source signal, 10 s excerpt from a male speaker taken from [244], was simulated at different angles from $-90°$ to $90°$ in $5°$ intervals, always at a distance of 1.5 m from the centre of the LMA (the height of the source was fixed to 1.4 m). At each of these angles, uncorrelated white noise was also added to each of the microphone signals such that the ratio of the speech signal power in the first (uppermost and reference) microphone of the array to the uncorrelated white noise power was 30 dB. In addition to this, diffuse noise generated from [207] using multitalker babble noise from Audiotec [206] was also added such that the input SNR at the first microphone was approximately 0 dB.

For a source located at $\{3.6\,\mathrm{m}, 3.325\,\mathrm{m}, 1.4\,\mathrm{m}\}$, i.e. at $0°$, 1.5 m from the centre of the LMA, and 30 cm from the XM, $\mathbf{h}$ was computed according to (4.1) using the known positions of the LMA and the XM. This was then subsequently used to compute the MFB and SDB beamformers as in (2.72) and (2.74). For the SDB, $\mathring{\mathbf{\Gamma}}_{\mathbf{nn}}$ was computed from (2.62) and a scaled identity matrix with a scaling of $1e^{-6}$ was added to it so as to avoid ill-conditioning of its inverse. All simulations were performed using the Weighted Overlap and Add (WOLA) method [101], with a Discrete Fourier Transform (DFT) size of 512, 50% overlap, sampling frequency of 16 kHz, and a square-root hanning window. The room impulse responses were obtained using the randomised image method [208] and implemented from [209].

Fig. 4.6 displays the performance of the MFB and SDB with the LMA only and with the LMA and XM in terms of Δ SI-SNR and Δ STOI. The changes in these metrics are with respect to that of the first microphone in the LMA. The input SI-SNR at the first microphone was -3 dB and the corresponding STOI was 0.68. In Fig. 4.6, both performance metrics exhibit very similar behaviour. Firstly, as expected, all beamformers perform their best at $0°$, i.e. the direction for which the MFB and SDB designed. More interestingly however, is the nature of the roll-off for the different beamformers as the source moves outside of the look-direction of the beamformers. In general, while the performance of the MFB and SDB that use the LMA and XM have a much increased performance at $0°$ in comparison to using the LMA only, their performance begins to fall quite rapidly as soon as the source moves outside of $0°$. The MFB and SDB for the LMA only, on the other hand, maintain a fairly broad response within the look direction before rolling off.

The performance of the various beamformers are further explained by observing Figs. 4.7 and 4.8 which display the unweighted third-octave band Δ SNRs for the different source locations. In Fig. 4.7, it can be observed that there is indeed a much improved SNR along the look-direction of $0°$ when using the LMA and

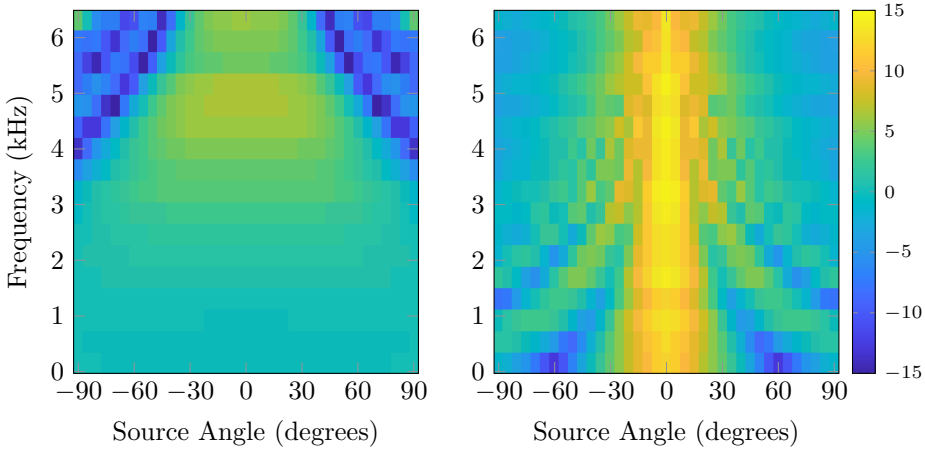


Figure 4.7: Unweighted third-octave band Δ SNRs (indicated by the colour) as a function of incident source angle for the MFB (Left) using an LMA only and (Right) using an LMA and an XM.

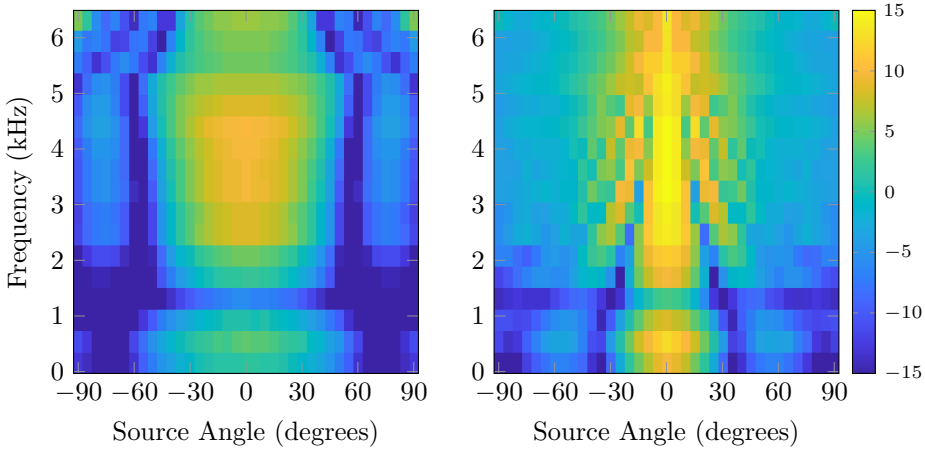


Figure 4.8: Unweighted third-octave band Δ SNRs (indicated by the colour) as a function of incident source angle for the SDB (Left) using an LMA only and (Right) using an LMA and an XM.

XM as compared to using the LMA only. The fairly omnidirectional response of the MFB for the LMA up to approximately 3 kHz explains the slow roll-off in performance as the source moves outside of the look-direction as observed in

Fig. 4.6.

In Fig. 4.8, it is also observed that there is a much improved SNR along the look-direction of 0° when using the LMA and XM as compared to using the LMA only and again accounts of the roll-off in performance that was observed in Fig. 4.6. For the SDB that uses the LMA only, there is a more directional response in the lower frequencies in comparison to the MFB that uses the LMA only, which also accounts for its performance as observed in Fig. 4.6. The results from Figs. 4.6, 4.7, and 4.8 can be regenerated from the codes provided in [95], and the resulting audio files can also be listened to.

4.5 Conclusion

From sections 4.2 and 4.4, it is clear that the introduction of an XM can indeed improve the performance of a beamformer as opposed to using an LMA only. The ultimate goal is to apply the integrated strategy from Chapter 3 to a system that involves XMs. However, before being able to investigate this, an approach for using a priori spatial knowledge in this type of microphone configuration needs to be addressed. This is the subject of the following chapter in the context of an MVDR beamformer. Chapters 6 and 7, then apply this approach to binaural speech enhancement and speech dereverberation respectively. Finally in chapter 8, the integrated approach similar to that of chapter 3, but for a local microphone array linked with external microphones will be analysed and evaluated.

Chapter 5

Estimation of spatial information for external microphones using a priori spatial knowledge from a local microphone array

5.1 Introduction

In this chapter[1], the design of a minimum variance distortionless response (MVDR) beamformer and its generalised sidelobe canceller (GSC) counterpart as applied to a microphone configuration consisting of a local microphone array (LMA) and multiple external microphones (XMs) will be investigated in

[1]This chapter is partially based on the following publications: (1) Ali R., Bernardi G., van Waterschoot T., Moonen M., "Methods of extending a generalised sidelobe canceller with external microphones", *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 27, no. 9, Sep. 2019, pp. 1349-1364; (2) Ali R., van Waterschoot T., Moonen M., "Completing the RTF vector for an MVDR beamformer as applied to a local microphone array and an external microphone", in *Proc. of the Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, Sep. 2018., and (3) Ali R., van Waterschoot T., Moonen M., "Generalised Sidelobe Canceller for Noise Reduction in Hearing Devices using an external microphone", in Proc. of the 2018 *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Calgary, Canada, Apr. 2018.

centralised processing scheme. While the benefits of extending a local LMA with an XM were evident in Chapter 4, such an improvement will only be achieved in practice for an MVDR beamformer when there are accurate estimates of the noise-only correlation matrix, $\mathbf{R_{nn}}$, and the relative transfer function (RTF) vector, $\mathbf{h}$, pertaining to all microphones. As discussed in chapter 3, for an LMA, given an estimate of $\mathbf{R_{nn}}$, two approaches can be followed to obtain an RTF vector: using either a priori or fully data-dependent spatial information. For the case of an LMA linked to XMs, however, a priori spatial information is not typically available for the XMs. For instance, the relative location of an XM with respect to the LMA and the desired speaker is unlikely to be known a priori[2], or the XMs may not even be calibrated with respect to the LMA. Consequently, when using an LMA with XMs, a fully data-dependent approach is favoured for estimating the RTF vector pertaining to all of the microphones [47, 217, 220, 222–225].

Although a priori spatial information is typically unavailable for the XMs, it is more likely to be available for the LMA. Hence, in opting for a fully data-dependent approach for estimating the RTF vector when using an LMA with XMs, any a priori spatial information that is available for the LMA is ignored. In chapter 3, it was shown that the use of such a priori spatial information can be beneficial and that relying solely on a data-dependent approach can be disadvantageous in some situations. Therefore, in this chapter, an alternative to the fully data-dependent approach for estimating the RTF vector when using an LMA and XMs is proposed; whereby any a priori spatial information available for the LMA is explicitly used.

For an MVDR beamformer, this approach corresponds to redefining the RTF vector into two distinct parts: (i) a known, a priori RTF vector pertaining to the LMA and (ii) an unknown, missing part of the RTF vector corresponding to the XMs that needs to be estimated from the data. In other words, as opposed estimating the entire RTF vector from the data, only a partial estimation of the RTF vector from the data is required when the a priori RTF vector is used. Such an approach can also be interpreted as a means of integrating a priori and data-dependent spatial information, although from a different perspective than the confidence-based one in Chapter 3. The nature of the integration in this case stems from a more *geometrically-based* perspective as it is related to the geometrical configuration of the microphones.

The advantage of this approach is that the XMs can be incorporated in a modular fashion or as “add-ons" for an improved performance to an existing LMA-based speech enhancement system (such as a hearing aid). With such

[2] A simple example is that of a microphone on a mobile device. With persons continuously using the mobile phones to write messages or make calls, its location, and hence location of the XM relative to the LMA and desired speaker is subject to continuous changes.

modularity, this approach has a built-in contingency option of reverting to the original performance of the LMA-based speech enhancement system in cases where data-dependent estimation becomes challenging. This is in contrast to the fully data-dependent approach for estimating the RTF vector, since in such cases if the data-dependent estimation is poor, there are no alternative options or decisions which can be taken to yield an acceptable performance.

Three methods of estimating the RTF vector by following this a priori approach will be discussed in this chapter, as well as the associated extensions to a practical GSC framework. Two of these methods involve a process of completing a blocking matrix similar to that of [245], while the third method uses the speech estimate from an LMA-based GSC directly with an orthogonalised version of the XM signals to obtain an improved speech estimate via a rank-1 generalised eigenvalue decomposition (GEVD). Following an evaluation of these methods, it will be shown that the third method offers the most improvement and has the advantage of not compromising an existing LMA-based GSC as both speech estimates, i.e. that from only using the existing LMA-based GSC, and that from using the LMA in co-operation with the XMs are independently available. This method is then further evaluated in a public environment and is also compared to the case when the entire RTF vector is estimated using a data-dependent approach.

Throughout this chapter, the signal model in the STFT domain from (4.4) is considered consisting of an LMA with M_a microphones and M_e XMs. It is to be recalled that variables with the subscript "$\mathbf{a}$" refer to the LMA and variables with the subscript "$\mathbf{e}$" refer to the XMs. The dependency on k and l will be dropped again as they are assumed to be independently treated. The rest of this chapter is organised as follows. In section 5.2, the LMA-based MVDR beamformer and GSC processing schemes are briefly reviewed. In section 5.3.1, the use of a priori spatial knowledge for the LMA is further motivated for the MVDR beamformer that includes multiple XMs. In section 5.4, the method of completing the blocking matrix for an extension of the LMA-based GSC for two different RTF estimation procedures involving multiple XMs is discussed. In section 5.5, an alternative approach to extending the LMA-based GSC is proposed, which involves an orthogonalisation of the XM signals and a rank-1 GEVD procedure. In section 5.6, the various algorithms are evaluated with recorded data taken in a typical office scenario. The third method is then evaluated in a public environment and also compared to the case of estimating the entire RTF vector from the data. In all evaluations the LMA was comprised of the microphones from a behind-the-ear (BTE) hearing aid (HAs). Conclusions are finally drawn in section 5.7.

5.2 Processing with a local microphone array using a priori spatial knowledge

5.2.1 LMA-based MVDR

Considering only the LMA, the constrained optimisation problem for the MVDR beamformer that takes into consideration a priori knowledge is:

$$\begin{aligned} \underset{\mathbf{w_a}}{\text{minimise}} \quad & \mathbf{w_a}^H \hat{\mathbf{R}}_{\mathbf{n_a n_a}} \mathbf{w_a} \\ \text{subject to} \quad & \mathbf{w_a}^H \widetilde{\mathbf{h}}_{\mathbf{a}} = 1 \end{aligned} \tag{5.1}$$

where $\hat{\mathbf{R}}_{\mathbf{n_a n_a}}$ is an estimate of the noise-only correlation matrix for the LMA, $\mathbf{R}_{\mathbf{n_a n_a}}$. The optimal MVDR beamformer given as:

$$\widetilde{\mathbf{w}}_{\mathbf{a}} = \frac{\hat{\mathbf{R}}^{-1}_{\mathbf{n_a n_a}} \widetilde{\mathbf{h}}_{\mathbf{a}}}{\widetilde{\mathbf{h}}_{\mathbf{a}}^H \hat{\mathbf{R}}^{-1}_{\mathbf{n_a n_a}} \widetilde{\mathbf{h}}_{\mathbf{a}}} \tag{5.2}$$

which will be referred to as the $\text{MVDR}^{\text{AP}}_{\text{LMA}}$. In (5.2), $\widetilde{\mathbf{h}}_{\mathbf{a}} = [\widetilde{\text{h}}_{\text{a},1}\ \widetilde{\text{h}}_{\text{a},2}\ \ldots \widetilde{\text{h}}_{\text{a,M}_\text{a}}]^T$ is the a priori RTF vector for the LMA that defines the constraint direction for which the speech is to be left undistorted. As in Chapter 3, $\widetilde{\mathbf{h}}_{\mathbf{a}}$ can be based on a priori assumptions regarding microphone characteristics, position, speaker location and room acoustics (e.g. no reverberation). The speech estimate, $\widetilde{\text{z}}_{\text{a},1}$, is then obtained through the linear filtering of the microphone signals with the complex-valued filter $\widetilde{\mathbf{w}}_{\mathbf{a}}$:

$$\widetilde{\text{z}}_{\text{a},1} = \widetilde{\mathbf{w}}_{\mathbf{a}}^H \mathbf{y}_{\mathbf{a}} \tag{5.3}$$

5.2.2 LMA-based GSC

As discussed in Chapter 2, the practical implementation of the LMA-based MVDR as proposed by Griffiths and Jim [124], involved the conversion of the constrained minimisation problem of (5.1) to an unconstrained one. The resulting beamformer, an LMA-based GSC, is displayed in Fig. 5.1. This LMA-based GSC that uses a priori spatial knowledge will be referred to as the $\text{GSC}^{\text{AP}}_{\text{LMA}}$. The top branch provides a speech reference by satisfying the constraint in (5.1) through the use of a fixed beamformer, $\tilde{\mathbf{f}}_{\mathbf{a}}$. The output of the top branch is then given by:

$$\text{y}_\text{f} = \tilde{\mathbf{f}}_{\mathbf{a}}^H \mathbf{y}_{\mathbf{a}} \tag{5.4}$$

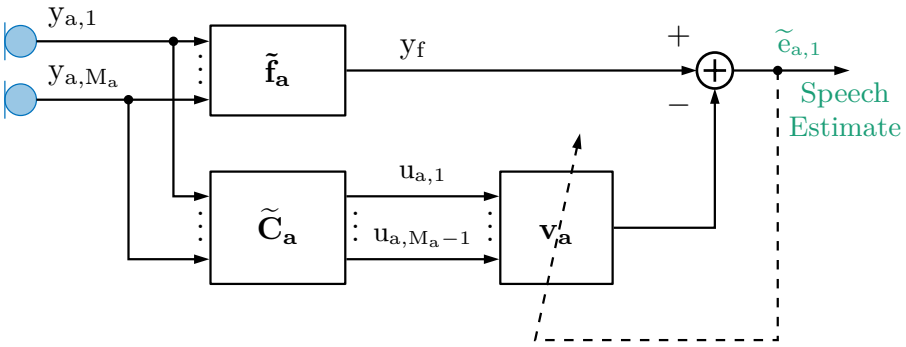


Figure 5.1: LMA-based Generalised Sidelobe Canceller using a priori spatial knowledge $\text{GSC}^{\text{AP}}_{\text{LMA}}$.

The bottom branch provides the noise reference signals $\mathbf{u_a} = \widetilde{\mathbf{C}}_\mathbf{a}^H \mathbf{y_a} = [\text{u}_{\text{a},1}\ \text{u}_{\text{a},2} \dots \text{u}_{\text{a},\text{M}_\text{a}-1}]^T$ through the $\text{M}_\text{a} \times (\text{M}_\text{a}-1)$ blocking matrix, $\widetilde{\mathbf{C}}_\mathbf{a}$, which is defined as being orthogonal to the corresponding RTFs such that $\widetilde{\mathbf{C}}_\mathbf{a}^H \widetilde{\mathbf{h}}_\mathbf{a} = \mathbf{0}$. Therefore, $\widetilde{\mathbf{C}}_\mathbf{a}$ can be defined as follows:

$$\widetilde{\mathbf{C}}_\mathbf{a} = \begin{bmatrix} -\widetilde{\text{h}}^*_{\text{a},2} & -\widetilde{\text{h}}^*_{\text{a},3} & \dots & -\widetilde{\text{h}}^*_{\text{a},\text{M}_\text{a}} \\ & \mathbf{I}_{\text{M}_\text{a}-1} & & \end{bmatrix} \tag{5.5}$$

where $\{.\}^*$ denotes the complex conjugate and $\mathbf{I}_{\text{M}_\text{a}-1}$ is the $(\text{M}_\text{a}-1) \times (\text{M}_\text{a}-1)$ identity matrix (in general $\mathbf{I}_\vartheta$ will denote the $\vartheta \times \vartheta$ identity matrix).

The adaptive noise cancelling (ANC) filter, $\mathbf{v_a}$, is then updated such as to reduce the residual noise in the speech reference at each time frame, l[1], by solving the following unconstrained optimisation problem:

$$\underset{\mathbf{v_a}(\text{l})}{\text{minimise}} \quad \mathbb{E}\{|\tilde{\mathbf{f}}_\mathbf{a}^H \mathbf{y_a}(\text{l}) - \mathbf{v}_\mathbf{a}^H(\text{l}) \widetilde{\mathbf{C}}_\mathbf{a}^H \mathbf{y_a}(\text{l})|^2\} \tag{5.6}$$

In order to avoid speech cancellation due to speech leakage into the noise reference, $\mathbf{v_a}$ is usually updated in frames where only noise is present. The optimal solution for $\mathbf{v_a}$ is given as:

$$\hat{\mathbf{v}}_\mathbf{a}(\text{l}) = (\widetilde{\mathbf{C}}_\mathbf{a}^H \hat{\mathbf{R}}_{\mathbf{n_a n_a}}(\text{l}) \widetilde{\mathbf{C}}_\mathbf{a})^{-1} \widetilde{\mathbf{C}}_\mathbf{a}^H \hat{\mathbf{R}}_{\mathbf{n_a n_a}}(\text{l}) \tilde{\mathbf{f}}_\mathbf{a} \tag{5.7}$$

from which the filter output representing a speech estimate follows as:

$$\widetilde{\text{e}}_{\text{a},1}(\text{l}) = \tilde{\mathbf{f}}_\mathbf{a}^H \mathbf{y_a}(\text{l}) - \hat{\mathbf{v}}_\mathbf{a}^H(\text{l}) \widetilde{\mathbf{C}}_\mathbf{a}^H \mathbf{y_a}(\text{l}) \tag{5.8}$$

[1]The dependency on l will be re-introduced to highlight the importance of the time dependence on some quantities. These quantities are still per frequency and the dependency on k will continue to be omitted for brevity.

In practice, the solution to (5.6) is often implemented with a Normalised Least Mean Squares (NLMS) approach [125].

5.3 MVDR beamformer with an LMA and XMs using a priori spatial knowledge

The following processing schemes which make using of only the LMA and a priori spatial knowledge serve as the benchmark for comparison with the proposed algorithms in this chapter as they are representative of a class of existing algorithms used in assistive hearing devices, e.g. the BEAM algorithm used in cochlear implants [33].

5.3.1 MVDR formulation

For the system of the LMA and XMs, the MVDR beamformer will consist of an extended noise-correlation matrix, $\mathbf{R_{nn}}$ of dimensions $(\mathrm{M_a}+\mathrm{M_e})\times(\mathrm{M_a}+\mathrm{M_e})$ and an RTF vector, $\mathbf{h}$ that consists of $\mathrm{M_a}$ components corresponding to the LMA, $\mathbf{h_a}$, and $\mathrm{M_e}$ components corresponding to the XM signals, $\mathbf{h_e}$ so that $\mathbf{h}=[\mathbf{h_a}^T\ \mathbf{h_e}^T]^T$. If there is a priori knowledge of $\mathbf{h_a}$, then an a priori RTF vector can be defined as follows:

$$\widetilde{\mathbf{h}}=[\ \widetilde{\mathbf{h}}_\mathbf{a}^T\ \mathbf{h}_\mathbf{e}^T\ ]^T \tag{5.9}$$

which consists partially of the a priori RTF vector pertaining to the LMA, $\widetilde{\mathbf{h}}_\mathbf{a}$, and partially of the RTF vector pertaining to the XM, $\mathbf{h_e}$, which is unknown and remains to be estimated. The estimate of $\mathbf{h_e}$ will therefore be explicitly based on the a priori spatial knowledge given by $\widetilde{\mathbf{h}}_\mathbf{a}$ and will be subsequently denoted as $\hat{\widetilde{\mathbf{h}}}_\mathbf{e}$ to emphasise that it is constrained by $\widetilde{\mathbf{h}}_\mathbf{a}$ but estimated from the observed microphone data.

The MVDR beamformer for a microphone configuration of an LMA with XMs that uses a priori spatial knowledge is then given by:

$$\widetilde{\mathbf{w}}=\frac{\hat{\mathbf{R}}_\mathbf{nn}^{-1}\widetilde{\mathbf{h}}}{\widetilde{\mathbf{h}}^H\hat{\mathbf{R}}_\mathbf{nn}^{-1}\widetilde{\mathbf{h}}} \tag{5.10}$$

where $\hat{\mathbf{R}}_\mathbf{nn}$ is an estimate of $\mathbf{R_{nn}}$ and $\hat{\widetilde{\mathbf{h}}}_\mathbf{e}$ is used in place of $\mathbf{h_e}$ in $\widetilde{\mathbf{h}}$[3] and the speech estimate, $\widetilde{\mathrm{z}}_1=\widetilde{\mathbf{w}}^H\mathbf{y}$. This will be referred to as the $\mathrm{MVDR}^{\mathrm{AP}}_{\text{LMA-XM}}$

[3]It is acknowledged that there is a slight abuse of notation here as $\widetilde{\mathbf{h}}$ should perhaps be denoted as $\hat{\widetilde{\mathbf{h}}}$ when $\hat{\widetilde{\mathbf{h}}}_\mathbf{e}$ is used in place of $\mathbf{h_e}$. However, in favour of legibility and to stress

beamformer in the following. In [246], various methods for computing $\hat{\tilde{\mathbf{h}}}_{\mathbf{e}}$ (in the case of $\mathrm{M_e} = 1$) have been presented. In the following sections, these methods are generalised for $\mathrm{M_e}$ XMs and corresponding implementation of the $\mathrm{MVDR}^{\mathrm{AP}}_{\text{LMA-XM}}$ in a GSC framework, referred to as the $\mathrm{GSC}^{\mathrm{AP}}_{\text{LMA-XM}}$, will be discussed.

5.3.2 Advantages of using partial a priori spatial knowledge

Before moving onto the various methods of extending the $\mathrm{GSC}^{\mathrm{AP}}_{\mathrm{LMA}}$, the approach of using a priori spatial knowledge is further justified by comparing it to the data-dependent approach in terms of the array gain. From Sec. 2.7.1, an expression for the array gain (2.70) was derived for the optimal MVDR beamformer. This expression is however only valid for a perfect estimation of $\mathbf{h}$ so that $\mathbf{w}^H_{\mathrm{MVDR}}\mathbf{h} = 1$. As we already know, since $\mathbf{h}$ needs to be estimated, the MVDR beamformer obtained in practice will not be equivalent to the "perfect" $\mathbf{w}_{\mathrm{MVDR}}$. Denoting $\bar{\mathbf{w}}$ as an MVDR beamformer based on an imperfect RTF vector, $\bar{\mathbf{h}} \neq \mathbf{h}$, then in general $\bar{\mathbf{w}}^H\mathbf{h} \neq 1$. For a perfectly estimated $\mathbf{R_{nn}}$, the array gain due to a mismatch in the RTF vector is then:

$$\mathrm{AG}^{\mathrm{MVDR}} = |\bar{\mathbf{w}}^H\mathbf{h}|^2\sigma^2_{\mathrm{n_r}}\ \bar{\mathbf{h}}^H\mathbf{R}^{-1}_{\mathbf{nn}}\bar{\mathbf{h}} \quad \text{for } \bar{\mathbf{h}} \neq \mathbf{h} \tag{5.11}$$

With $\bar{\mathbf{h}} = [\bar{\mathbf{h}}_{\mathbf{a}}^T\ \ \bar{\mathbf{h}}_{\mathbf{e}}^T]^T$, then the array gain expression with the inclusion of the XMs from (4.13) becomes:

$$\mathrm{AG}^{\mathrm{MVDR}}_{\mathrm{LMA+XM}} = \underbrace{|\bar{\mathbf{w}}^H\mathbf{h}|^2}_{\text{Mismatch}}\Bigg(\underbrace{\sigma^2_{\mathrm{n_{a,1}}}\bar{\mathbf{h}}_{\mathbf{a}}^H\mathbf{R}^{-1}_{\mathbf{n_a n_a}}\bar{\mathbf{h}}_{\mathbf{a}}}_{\mathrm{AG}^{\mathrm{MVDR}}_{\mathrm{LMA}}} + \underbrace{\sigma^2_{\mathrm{n_{a,1}}}||\mathbf{P}^{-1/2}\mathbf{R}^H_{\mathbf{n_a n_e}}\mathbf{R}^{-1}_{\mathbf{n_a n_a}}\bar{\mathbf{h}}_{\mathbf{a}} - \mathbf{P}^{-1/2}\bar{\mathbf{h}}_{\mathbf{e}}||^2}_{\mathrm{AG}^{\mathrm{MVDR}}_{\mathrm{XM}}}\Bigg) \tag{5.12}$$

which consists of three terms: (i) a mismatch term due to the fact that $\bar{\mathbf{w}}^H\mathbf{h} \neq 1$, (ii) the array gain due to the LMA, and (iii) the additional gain introduced by the XMs. The question here is then should the a priori RTF vector be used for $\bar{\mathbf{h}}_{\mathbf{a}}$, i.e. $\bar{\mathbf{h}}_{\mathbf{a}} = \widetilde{\mathbf{h}}_{\mathbf{a}}$ or should all components in the RTF vector be estimated from

that the estimation is done in accordance to the a priori assumptions set by $\widetilde{\mathbf{h}}_{\mathbf{a}}$ is why the notation is maintained as $\widetilde{\mathbf{h}}$.

noisy data? In fact, as seen in Chapter 3, both approaches have their benefits and it is entirely scenario dependent.

In order to have a better intuition of (5.12), the array gain was computed for two separate cases of $\bar{\mathbf{h}}$ which are representative of using a partial a priori RTF vector or not. Considering the case when $\mathrm{M_e} = 1$, the true RTF vector, $\mathbf{h}$ is denoted in terms of its magnitude and phase as follows:

$$\mathbf{h} = |\mathbf{h}| \angle \boldsymbol{\theta} = \begin{bmatrix} |\mathbf{h_a}| \, \angle \boldsymbol{\theta_a} \\ |\mathrm{h_e}| \, \angle \theta_\mathrm{e} \end{bmatrix} \tag{5.13}$$

where $|\mathbf{h}|$ is the vector of magnitudes of $\mathbf{h}$ consisting of $|\mathbf{h_a}|$, the $\mathrm{M_a} \times 1$ vector of magnitudes of $\mathbf{h_a}$ and $|\mathrm{h_e}|$ is the magnitude of $\mathrm{h_e}$, $\angle\boldsymbol{\theta}$ is the vector of phases of $\mathbf{h}$ consisting of $\angle\boldsymbol{\theta_a}$, the $\mathrm{M_a} \times 1$ vector of phases of $\mathbf{h_a}$ and $\angle\theta_\mathrm{e}$, the phase of $\mathrm{h_e}$. Two separate RTF vector perturbations can then be defined as:

$$\bar{\mathbf{h}}_\mathrm{ap} = \begin{bmatrix} (|\mathbf{h_a}| + \boldsymbol{\xi_a}) \, \angle(\boldsymbol{\theta_a} + \boldsymbol{\xi_a}) \\ (|\mathrm{h_e}| + \xi_\mathrm{e}) \, \angle(\theta_\mathrm{e} + \xi_\mathrm{e}) \end{bmatrix} \qquad \bar{\mathbf{h}}_\mathrm{dd} = (|\mathbf{h}| + \boldsymbol{\Delta}) \, \angle(\boldsymbol{\theta} + \boldsymbol{\Delta}) \tag{5.14}$$

where $\boldsymbol{\xi_a}$ and $\boldsymbol{\Delta}$ are gaussian random vectors and ξ_e is a gaussian random variable.

A scenario was then considered for $\mathbf{R_{nn}} = \mathbf{I}$ and $\mathbf{h} = [1 \; 1 \; 1 \; 1 \; 1.3]^T$, i.e. with $\mathrm{M_a} = 4$. The variance of $\boldsymbol{\xi_a}$ was fixed to 0.2 so that it represented a fixed deviation from $\mathbf{h_a}$, emulating a priori spatial information pertaining to the LMA that is not perfectly accurate. $\bar{\mathbf{h}}_\mathrm{ap}$ and $\bar{\mathbf{h}}_\mathrm{dd}$ were then computed from (5.14) for different realisations of ξ_e and $\boldsymbol{\Delta}$ whose variances ranged between 0 and 1. This was then followed by the computation of the corresponding array gain (in dB) from (5.11). The entire process was then repeated 1000 times, and the array gains were averaged. Figure 5.2 shows the resulting averaged array gain as a function of the variance of ξ_e and the variance of $\boldsymbol{\Delta}$. The two surfaces correspond to the array gain computed with $\bar{\mathbf{h}}_\mathrm{ap}$ and that computed with $\bar{\mathbf{h}}_\mathrm{dd}$.

For the surface of $\bar{\mathbf{h}}_\mathrm{dd}$, it can be seen that when there is no error in the RTF vector, then a maximum possible array gain is achieved. However, as errors are introduced, there is a fairly rapid decline in the amount of achievable array gain. For the $\bar{\mathbf{h}}_\mathrm{ap}$ surface on the other hand, when there is no error in the RTF component corresponding to the XM, the array gain is not as large as when there were no errors in the $\bar{\mathbf{h}}_\mathrm{dd}$ surface. This is due to the error introduced with the constant variance of $\xi_\mathbf{a}$ in the part of the RTF vector corresponding to the LMA. This is indeed indicative of the situation when a priori information is used for the LMA, since it is expected that the a priori information is not perfect and hence the maximum achievable array gain cannot be obtained. Nevertheless, the benefit is seen as errors are introduced to the RTF component corresponding to the XM. As errors are introduced, the array gain decreases, but not as rapidly

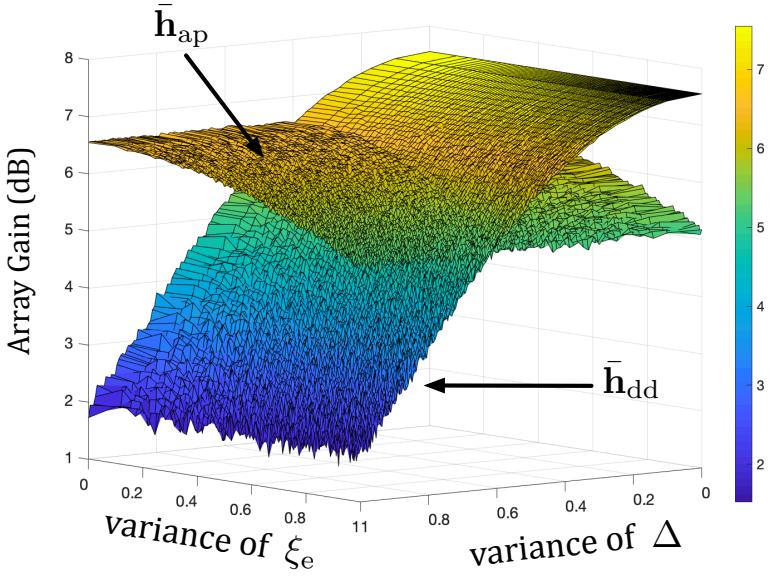


Figure 5.2: Array Gain as computed in (5.11) using two different RTF vector perturbation models from (5.14) illustrating the trade-off between using an a priori or a data-dependent approach. The vertical axis and the colourbar are both indicative of the array gain in dB. This plot may be re-created from the Matlab code provided in [95].

as that when $\bar{\mathbf{h}}_{\mathrm{dd}}$ was used. Hence using $\bar{\mathbf{h}}_{\mathrm{ap}}$ can result in a compromised, but more robust performance than that of the $\bar{\mathbf{h}}_{\mathrm{dd}}$. Consequently, it can be concluded that imposing spatial a priori information with respect to the LMA can indeed be beneficial. The optimal scenario however, would be to transition between the two surfaces, i.e. $\bar{\mathbf{h}}_{\mathrm{dd}}$ should be used when the estimation errors are small, and $\bar{\mathbf{h}}_{\mathrm{ap}}$ when estimation errors are larger. This trade-off should feel familiar as it was the basis of the integrated MVDR beamformer from Chapter 3. However, before revisiting the integrated MVDR beamformer in the context of the LMA with XMs, the algorithms for using a priori spatial information need to be firstly developed.

5.4 Completing the Blocking Matrix

One approach for implementing a GSC structure with an LMA and multiple XMs is to use the estimate, $\hat{\tilde{\mathbf{h}}}_{\mathbf{e}}$ to complete the additional columns of the blocking matrix. In [245], this was demonstrated (for $\mathrm{M_e} = 1$) using an estimate based on a cross-correlation method. In this section, this approach of completing the blocking matrix will be extended for $\mathrm{M_e} \geq 1$ with a further discussion of relevant implementation details. Two block schemes for a GSC will also be presented: (i) using the cross-correlation method to obtain $\hat{\tilde{\mathbf{h}}}_{\mathbf{e}}$, and (ii) using the EVD method to obtain $\hat{\tilde{\mathbf{h}}}_{\mathbf{e}}$ adopted from [246] (where $\mathrm{M_e} = 1$).

The cost function of (5.6) is firstly extended to include the XMs:

$$\underset{\mathbf{g}}{\text{minimise}} \quad \mathbb{E}\{|\tilde{\mathbf{f}}^H(\mathrm{l})\mathbf{y}(\mathrm{l}) - \mathbf{g}^H(\mathrm{l})\widetilde{\mathbf{C}}^H(\mathrm{l})\mathbf{y}(\mathrm{l})|^2\} \tag{5.15}$$

where $\tilde{\mathbf{f}}$ is an $(\mathrm{M_a} + \mathrm{M_e}) \times 1$ fixed beamformer acting on both the LMA and XM signals, $\mathbf{g} = [\mathbf{g}_{\mathbf{a}}^T \; \mathbf{g}_{\mathbf{e}}^T]^T$ is the ANC filter to be designed, and the extended $(\mathrm{M_a} + \mathrm{M_e}) \times (\mathrm{M_a} + \mathrm{M_e} - 1)$ blocking matrix is now given as:

$$\widetilde{\mathbf{C}}(\mathrm{l}) = \left[\begin{array}{c|c} \widetilde{\mathbf{C}}_{\mathbf{a}} & \begin{array}{c} -\hat{\tilde{\mathbf{h}}}_{\mathbf{e}}^H(\mathrm{l}) \\ \mathbf{0}_{(\mathrm{M_a}-1)\times \mathrm{M_e}} \end{array} \\ \hline \mathbf{0}_{\mathrm{M_e}\times(\mathrm{M_a}-1)} & \mathbf{I}_{\mathrm{M_e}} \end{array}\right] \tag{5.16}$$

where $\widetilde{\mathbf{C}}_{\mathbf{a}}$ is defined from (5.5), the zero blocks are indicated with their dimensions.

The role of the fixed beamformer within the context of a GSC is to satisfy the distortionless constraint, which can be accomplished regardless of the XMs. Consequently, the fixed beamformer, $\tilde{\mathbf{f}}$, can be readily simplified by setting $\tilde{\mathbf{f}} = [\tilde{\mathbf{f}}_{\mathbf{a}}^T \; \mathbf{0}_{(\mathrm{M_e}\times 1)}^T]^T$, i.e. using the fixed beamformer from (5.4) for the LMA signals and an $(\mathrm{M_e} \times 1)$ vector of zeros for the XM signals, hence $\tilde{\mathbf{f}}^H\widetilde{\mathbf{h}} = \tilde{\mathbf{f}}_{\mathbf{a}}^H\widetilde{\mathbf{h}}_{\mathbf{a}}$. As a result, only $\hat{\tilde{\mathbf{h}}}_{\mathbf{e}}(\mathrm{l})$ will be required to complete the blocking matrix, $\widetilde{\mathbf{C}}(\mathrm{l})$, which requires an update for each time frame. The optimal solution for $\mathbf{g}(\mathrm{l})$ is also computed in noise-only periods in a similar manner to $\mathbf{v}_{\mathbf{a}}(\mathrm{l})$ for the $\mathrm{GSC}_{\mathrm{LMA}}^{\mathrm{AP}}$, and is given by:

$$\hat{\mathbf{g}}(\mathrm{l}) = (\widetilde{\mathbf{C}}^H(\mathrm{l})\hat{\mathbf{R}}_{\mathbf{nn}}(\mathrm{l})\widetilde{\mathbf{C}}(\mathrm{l}))^{-1}\widetilde{\mathbf{C}}^H(\mathrm{l})\hat{\mathbf{R}}_{\mathbf{nn}}(\mathrm{l})\tilde{\mathbf{f}} \tag{5.17}$$

On substitution of (5.16) into (5.15), and with $\tilde{\mathbf{f}} = [\tilde{\mathbf{f}}_{\mathbf{a}}^T \ \mathbf{0}_{(\mathrm{M_e}\times 1)}^T]^T$, the new speech estimate then follows as:

$$\widetilde{\mathrm{e}}_1(\mathrm{l}) = \underbrace{\tilde{\mathbf{f}}_{\mathbf{a}}^H \mathbf{y}_{\mathbf{a}}(\mathrm{l}) - \hat{\mathbf{g}}_{\mathbf{a}}^H(\mathrm{l}) \underbrace{\widetilde{\mathbf{C}}_{\mathbf{a}}^H \mathbf{y}_{\mathbf{a}}(\mathrm{l})}_{\mathbf{u_a}(\mathrm{l})}}_{\text{LMA contribution, } \widetilde{\varepsilon}_{\mathrm{a}}(\mathrm{l})} - \underbrace{\hat{\mathbf{g}}_{\mathbf{e}}^H(\mathrm{l}) \hat{\widetilde{\mathbf{C}}}_{\mathbf{e}}^H(\mathrm{l}) \begin{bmatrix} \mathrm{y}_1(\mathrm{l}) \\ \mathbf{y_e}(\mathrm{l}) \end{bmatrix}}_{\mathbf{u}_{\mathrm{e}}(\mathrm{l})}}_{\text{XM contribution, } \widetilde{\varepsilon}_{\mathrm{e}}(\mathrm{l})} \tag{5.18}$$

where $\hat{\widetilde{\mathbf{C}}}_{\mathbf{e}}(\mathrm{l})$ is defined as:

$$\hat{\widetilde{\mathbf{C}}}_{\mathbf{e}}(\mathrm{l}) = \begin{bmatrix} -\hat{\widetilde{\mathbf{h}}}_{\mathbf{e}}^H(\mathrm{l}) \\ \mathbf{I}_{\mathrm{M_e}} \end{bmatrix} \tag{5.19}$$

and $\mathbf{u_a}(\mathrm{l})$ and $\mathbf{u}_{\mathrm{e}}(\mathrm{l})$ are the noise reference signals corresponding to the LMA and the XM signals respectively. It is apparent that there are two sets of updates that are required - (i) an update for $\hat{\widetilde{\mathbf{h}}}_{\mathbf{e}}(\mathrm{l})$, which will subsequently be used to complete the blocking matrix $\widetilde{\mathbf{C}}(\mathrm{l})$, by defining $\hat{\widetilde{\mathbf{C}}}_{\mathbf{e}}(\mathrm{l})$, and (ii) an update for the ANC filter, $\hat{\mathbf{g}}(\mathrm{l})$.

It is also evident that the speech estimate in (5.18) consists of two distinct components, $\widetilde{\varepsilon}_{\mathrm{a}}$, as a result of the contribution from the LMA signals, and $\widetilde{\varepsilon}_{\mathrm{e}}$, from the contribution from the XM signals. It is clear that when $\hat{\mathbf{g}}_{\mathbf{e}} = \mathbf{0}$, the contribution from the XM signals is disabled and the error or speech estimate will be identical to that of the $\mathrm{GSC}_{\mathrm{LMA}}^{\mathrm{AP}}$ in (5.8), i.e. $\hat{\mathbf{g}}_{\mathbf{a}} = \hat{\mathbf{v}}_{\mathbf{a}}$, and hence $\widetilde{\varepsilon}_{\mathrm{a}} = \widetilde{\mathrm{e}}_{\mathrm{a},1}$. However, in general, $\widetilde{\varepsilon}_{\mathrm{a}} \neq \widetilde{\mathrm{e}}_{\mathrm{a},1}$ as two different errors are minimised from (5.6) and (5.15).

Whereas in practice an NLMS approach could be used for updating $\hat{\mathbf{v}}_{\mathbf{a}}$ in the $\mathrm{GSC}_{\mathrm{LMA}}^{\mathrm{AP}}$, care should be taken for the approach used for updating $\hat{\mathbf{g}}$. This is because the PSD of the noise references from the XMs could be quite different as opposed to the case of the LMA, where it would be expected that the PSD of noise references from the LMA would be similar. Consequently, it is suggested that a diagonal step size normalised by the respective noise references be used in an NLMS context, or that a recursive least squares (RLS) [125] algorithm be used for updating $\hat{\mathbf{g}}$. A further analysis of adaptive techniques and their respective trade-offs is outside the scope of this chapter.

5.4.1 Cross Correlation RTF vector estimation

In [245], using the cross-correlation method to obtain $\hat{\widetilde{\mathbf{h}}}_{\mathbf{e}}$ (for $\mathrm{M_e} = 1$), a GSC method as previously described was presented. The signal, $\widetilde{\varepsilon}_{\mathrm{a}}$ from (5.18), was

used as a speech reference in order to carry out a cross-correlation with the XM signal for computing the RTF estimate. As opposed to $\widetilde{\varepsilon}_{\mathrm{a}}$, an alternative speech reference may be the output from the fixed beamformer, i.e. $\mathrm{y_f} = \tilde{\mathbf{f}}_{\mathbf{a}}^{H}\mathbf{y_a}$. Although this signal would be more noisy than $\widetilde{\varepsilon}_{\mathrm{a}}$, it will still be preferred to $\widetilde{\varepsilon}_{\mathrm{a}}$ due to its stability, i.e., it would be fixed and not time-varying due to adaptation. It should be noted however, that in using such a speech reference, this estimator takes into consideration the a priori information of the LMA. Hence, for $\mathrm{M_e} \geq 1$, the update of the ith component of the corresponding RTF vector for the XMs, $\hat{\widetilde{\mathbf{h}}}_{\mathbf{e,cc}}$, follows as:

$$\hat{\widetilde{\mathrm{h}}}_{\mathrm{e},i,\mathrm{cc}}(\mathrm{l}) = \begin{cases} \frac{r_{\mathrm{ea},i}(\mathrm{l})}{r_{\mathrm{aa},i}(\mathrm{l})} & \text{speech frames} \\ \hat{\widetilde{\mathrm{h}}}_{\mathrm{e},i,\mathrm{cc}}(\mathrm{l}-1) & \text{otherwise} \end{cases} \tag{5.20}$$

where

$$r_{\mathrm{ea},i}(\mathrm{l}) = \alpha_{\mathrm{e},i}\, r_{\mathrm{ea},i}(\mathrm{l}-1) \;+\; (1-\alpha_{\mathrm{e},i})\, \mathrm{y_{e},}_{i}(\mathrm{l})\, \mathrm{y}_{\mathrm{f}}^{*}(\mathrm{l}) \tag{5.21}$$

$$r_{\mathrm{aa},i}(\mathrm{l}) = \alpha_{\mathrm{e},i}\, r_{\mathrm{aa},i}(\mathrm{l}-1) \;+\; (1-\alpha_{\mathrm{e},i})\, |\mathrm{y_f}(\mathrm{l})|^2 \tag{5.22}$$

are computed in frames where speech is present and $\alpha_{e,i} \in [0,1]$ is a forgetting factor for the ith XM component. Although this estimator is of low complexity, it is a biased estimator due to the presence of noise in $\mathrm{y_f}$.

This GSC that uses the cross-correlation RTF estimate for the XM signals is referred to here as the $\mathrm{GSC}^{\text{AP-CC}}_{\text{LMA-XM}}$ and can be encapsulated by the block diagram as shown in Fig. 5.3, similar to [245], except that $\mathrm{y_f}$ is used as opposed to $\widetilde{\varepsilon}_{\mathrm{a}}$ for updating $\hat{\widetilde{\mathrm{h}}}_{\mathrm{e},i,\mathrm{cc}}$. It is reiterated here that the top branch remains unchanged from the $\mathrm{GSC}^{\mathrm{AP}}_{\mathrm{LMA}}$, and hence only changes are made to the lower branch. The cross-correlation RTF estimation procedure is used to complete the blocking matrix, $\widetilde{\mathbf{C}}$ (i.e. define $\hat{\widetilde{\mathbf{C}}}_{\mathbf{e}}$) and generate the extended set of noise references, $\mathbf{u} = [\mathbf{u}_{\mathbf{a}}^{T}\ \mathbf{u}_{\mathrm{e}}^{T}]^{T}$. The block diagram also intuitively depicts the two separate components of (5.18), with the speech estimate denoted as $\widetilde{\mathrm{e}}_{1,\mathrm{cc}}$. A further advantage of such a block scheme is that it does not compromise the initial structure of the $\mathrm{GSC}^{\mathrm{AP}}_{\mathrm{LMA}}$ and can be interpreted as an "add-on" since it can easily be seen that if $\mathbf{g_e} = \mathbf{0}$, the $\mathrm{GSC}^{\text{AP-CC}}_{\text{LMA-XM}}$ is reduced to the $\mathrm{GSC}^{\mathrm{AP}}_{\mathrm{LMA}}$.

5.4.2 EVD RTF vector estimate

Recalling the assumption of $\mathbf{R_{yy}} = \mathbf{R_{xx}} + \mathbf{R_{nn}}$, and that $\mathbf{R_{xx}}$ is to be modelled as a rank-1 matrix, then given the estimates, $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$, a matrix approximation

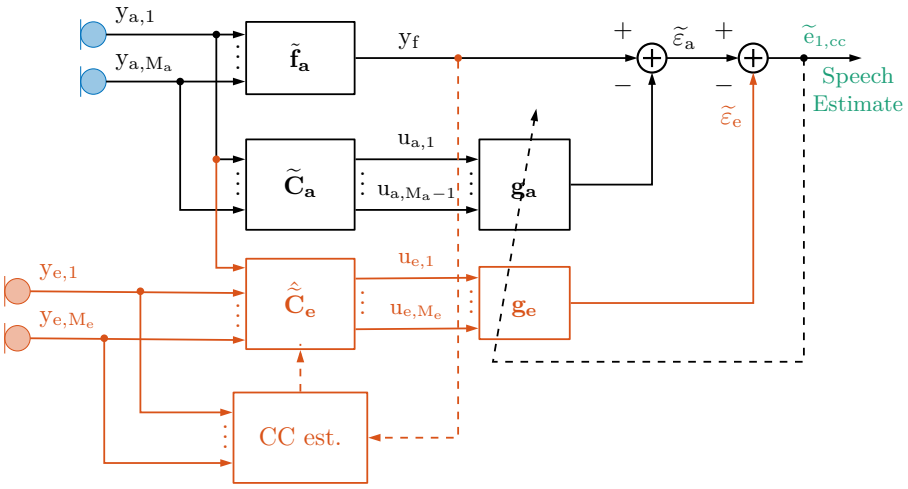


Figure 5.3: $\text{GSC}^{\text{AP-CC}}_{\text{LMA-XM}}$: The $\text{GSC}^{\text{AP}}_{\text{LMA}}$ extended with XMs and using the cross corrleation method of estimating the missing RTF component (CC est.) for the XMs from (5.20).

problem can be posed as follows [39]:

$$\begin{aligned} \underset{\mathbf{R_{xx}}}{\text{minimise}} \quad & ||(\hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{nn}}) - \mathbf{R_{xx}}||_F^2 \\ \text{subject to} \quad & \text{rank}(\mathbf{R_{xx}}) = 1 \end{aligned} \tag{5.23}$$

where $||.||_F$ is the Frobenius norm. From the Eckart–Young–Mirsky theorem [247], it is known that the solution to this problem can be found from the truncated EVD of the matrix $(\hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{nn}})$ using the principal eigenvalue and corresponding eigenvector. By substitution of (4.9) in (5.23), the problem can alternatively be expressed as:

$$\underset{\sigma^2_{s_{a,1}}, \mathbf{h}}{\text{minimise}} \quad ||(\hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{nn}}) - \sigma^2_{s_{a,1}} \mathbf{h}\mathbf{h}^H||_F^2 \tag{5.24}$$

where the minimisation is specifically over the RTF vector,$\mathbf{h}$ and the speech PSD in the reference microphone of the LMA, $\sigma^2_{s_{a,1}}$. At this point, the principal eigenvector form the EVD of $(\hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{nn}})$ can yield an estimate for the entire RTF vector for all microphones. However, given that there is an available a priori RTF vector for the LMA, $\tilde{\mathbf{h}}_{\mathbf{a}}$, then (5.24) can alternatively be posed

as [246]:

$$\underset{\sigma^2_{s_{a,1}},\,\mathbf{h_e}}{\text{minimise}} \quad ||(\hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{nn}}) - \sigma^2_{s_{a,1}} \begin{bmatrix} \widetilde{\mathbf{h}}_{\mathbf{a}} \\ \mathbf{h_e} \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{h}}^H_{\mathbf{a}} & \mathbf{h}^H_{\mathbf{e}} \end{bmatrix} ||^2_F \tag{5.25}$$

In [246], for $M_e = 1$, it has been demonstrated that by introducing a transform, (5.25) is further simplified and the estimate of $\mathbf{h_e}$ can be computed from a 2×2 correlation matrix. In the following, this procedure is generalised for the case of $M_e \geq 1$.

Proceeding to solve (5.25), an $M_a \times (M_a - 1)$ blocking matrix $\widetilde{\mathbf{B}}_{\mathbf{a}}$ and a specific $M_a \times 1$ fixed beamformer, $\widetilde{\mathbf{b}}_{\mathbf{a}}$ are defined such that:

$$\widetilde{\mathbf{B}}^H_{\mathbf{a}} \widetilde{\mathbf{h}}_{\mathbf{a}} = \mathbf{0}; \qquad \widetilde{\mathbf{b}}_{\mathbf{a}} = \frac{\widetilde{\mathbf{h}}_{\mathbf{a}}}{||\widetilde{\mathbf{h}}_{\mathbf{a}}||} \tag{5.26}$$

where $\widetilde{\mathbf{B}}^H_{\mathbf{a}} \widetilde{\mathbf{B}}_{\mathbf{a}} = \mathbf{I}_{(M_a-1)}$. In practice, $\widetilde{\mathbf{B}}_{\mathbf{a}}$ can be computed from a QR decomposition of $\widetilde{\mathbf{C}}_{\mathbf{a}}$. Using $\widetilde{\mathbf{B}}_{\mathbf{a}}$ and $\widetilde{\mathbf{b}}_{\mathbf{a}}$, an $(M_a + M_e) \times (M_a + M_e)$ unitary transformation matrix, $\widetilde{\mathbf{T}}$, can be subsequently defined:

$$\widetilde{\mathbf{T}} = \left[\begin{array}{c|c} \widetilde{\mathbf{T}}_{\mathbf{a}} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I}_{M_e} \end{array}\right] \tag{5.27}$$

where $\widetilde{\mathbf{T}}_{\mathbf{a}} = [\widetilde{\mathbf{B}}_{\mathbf{a}} \;\; \widetilde{\mathbf{b}}_{\mathbf{a}}]$, $\widetilde{\mathbf{T}}^H_{\mathbf{a}} \widetilde{\mathbf{T}}_{\mathbf{a}} = \mathbf{I}_{M_a}$, and hence $\widetilde{\mathbf{T}}^H \widetilde{\mathbf{T}} = \mathbf{I}_{(M_a+M_e)}$. As the Frobenius norm is invariant under a unitary transformation [247], (5.25) can be rewritten as:

$$\underset{\sigma^2_{s_{a,1}},\,\mathbf{h_e}}{\text{minimise}} \quad ||\widetilde{\mathbf{T}}^H((\hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{nn}}) - \sigma^2_{s_{a,1}} \begin{bmatrix} \widetilde{\mathbf{h}}_{\mathbf{a}} \\ \mathbf{h_e} \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{h}}^H_{\mathbf{a}} & \mathbf{h}^H_{\mathbf{e}} \end{bmatrix}) \widetilde{\mathbf{T}}||^2_F \tag{5.28}$$

By using (5.26) and (5.27), it can be seen that a transformed version of the RTF vector can be expressed as follows:

$$\widetilde{\mathbf{T}}^H \begin{bmatrix} \widetilde{\mathbf{h}}_{\mathbf{a}} \\ \mathbf{h_e} \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{B}}^H_{\mathbf{a}} \widetilde{\mathbf{h}}_{\mathbf{a}} \\ \widetilde{\mathbf{b}}^H_{\mathbf{a}} \widetilde{\mathbf{h}}_{\mathbf{a}} \\ \mathbf{h_e} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ ||\widetilde{\mathbf{h}}_{\mathbf{a}}|| \\ \mathbf{h_e} \end{bmatrix} \tag{5.29}$$

and hence the expansion of (5.28) becomes:

$$\min_{\sigma^2_{s_{a,1}},\,\mathbf{h_e}} \quad || \left[\begin{array}{c|c} \mathbf{K}_{\mathbf{a}_-} & \mathbf{K}^H_{\mathbf{c}} \\ \hline \mathbf{K_c} & \mathbf{K}_{\mathbf{e}_+} \end{array}\right] - \left[\begin{array}{c|c} \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{K}_{\mathbf{xx,r1}} \end{array}\right] ||^2_F \tag{5.30}$$

where $\mathbf{K}_{\mathbf{a}_-}$ is an $(M_a - 1) \times (M_a - 1)$ matrix, $\mathbf{K_c}$ an $(M_e + 1) \times (M_a - 1)$ matrix and $\mathbf{K}_{\mathbf{e}_+}$ and $\mathbf{K}_{\mathbf{xx,r1}}$ are $(M_e + 1) \times (M_e + 1)$ matrices realised as:

$$\mathbf{K}_{\mathbf{e}_+} = \left[\begin{array}{c|c} \widetilde{\mathbf{b}}_{\mathrm{a}}^H & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I}_{\mathrm{M_e}} \end{array}\right] (\hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{nn}}) \left[\begin{array}{c|c} \widetilde{\mathbf{b}}_{\mathbf{a}} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I}_{\mathrm{M_e}} \end{array}\right] \tag{5.31}$$

which is an estimate of $\mathbb{E}\left\{\begin{bmatrix}\widetilde{\mathbf{b}}_{\mathrm{a}}^H\mathbf{y_a} \\ \mathbf{y_e}\end{bmatrix}\begin{bmatrix}\mathbf{y_a}^H\widetilde{\mathbf{b}}_{\mathbf{a}} & \mathbf{y_e}^H\end{bmatrix}\right\} - \mathbb{E}\left\{\begin{bmatrix}\widetilde{\mathbf{b}}_{\mathrm{a}}^H\mathbf{n_a} \\ \mathbf{n_e}\end{bmatrix}\begin{bmatrix}\mathbf{n_a}^H\widetilde{\mathbf{b}}_{\mathbf{a}} & \mathbf{n_e}^H\end{bmatrix}\right\}$ and

$$\mathbf{K}_{\mathbf{xx,r1}} = \sigma^2_{\mathrm{s_{a,1}}}\begin{bmatrix}||\widetilde{\mathbf{h}}_{\mathbf{a}}|| \\ \mathbf{h_e}\end{bmatrix}\begin{bmatrix}||\widetilde{\mathbf{h}}_{\mathbf{a}}||\ \mathbf{h}_{\mathbf{e}}^H\end{bmatrix} \tag{5.32}$$

From (5.30), it can be seen that the additional a priori knowledge, $\widetilde{\mathbf{h}}_{\mathbf{a}}$, reduces the estimation problem further to:

$$\min_{\sigma^2_{\mathrm{s_{a,1}}},\,\mathbf{h_e}} \quad ||\mathbf{K}_{\mathbf{e}_+} - \mathbf{K}_{\mathbf{xx,r1}}||_F^2 \tag{5.33}$$

which is that of a rank-1 approximation of the $(\mathrm{M_e}+1)\times(\mathrm{M_e}+1)$ matrix, $\mathbf{K}_{\mathbf{e}_+}$. The estimation of $\mathbf{h_e}$ follows by initially extracting the principal eigenvector, $\mathbf{k_{max}} = [\mathrm{k_a}\ \ \mathbf{k}_{\mathbf{e}}^T]^T$, corresponding to the largest eigenvalue of $\mathbf{K}_{\mathbf{e}_+}$. Applying the appropriate scaling and normalisation of the elements in $\mathbf{k_{max}}$, the estimate of $\mathbf{h_e}$ that makes use of the a priori knowledge, $\widetilde{\mathbf{h}}_{\mathbf{a}}$, and an EVD procedure is then given by:

$$\hat{\widetilde{\mathbf{h}}}_{\mathbf{e,evd}} = \frac{||\widetilde{\mathbf{h}}_{\mathbf{a}}||\ \mathbf{k_e}}{\mathrm{k_a}} \tag{5.34}$$

This EVD-based RTF estimation method can easily be realised in a GSC scheme similar to that of the cross-correlation method as illustrated in Fig. 5.4, which will be referred to as the $\mathrm{GSC}^{\text{AP-EVD}}_{\text{LMA-XM}}$. In this case, however, a specific fixed beamformer of $\tilde{\mathbf{f}}_{\mathbf{a}} = \widetilde{\mathbf{b}}_{\mathbf{a}}$ is required. The output from the fixed beamformer, $\mathrm{y_f}$, and XM signals, $\mathbf{y_e}$, are then used to generate the correlation matrix $\mathbf{K}_{\mathbf{e}_+}$ from (5.31). The first term of (5.31) is updated when speech is active, and the second term updated in noise-only periods. $\hat{\widetilde{\mathbf{h}}}_{\mathbf{e,evd}}$ is computed accordingly and used to generate the extra noise reference, which completes the missing part of the blocking matrix, $\hat{\widetilde{\mathbf{C}}}_{\mathbf{e}}$. It is also noted that although another blocking matrix is defined in (5.26), this is only used for the derivation in computing $\hat{\widetilde{\mathbf{h}}}_{\mathbf{e,evd}}$. Consequently, the $\mathrm{GSC}^{\text{AP-EVD}}_{\text{LMA-XM}}$ scheme as depicted in Fig. 5.4 still uses $\widetilde{\mathbf{C}}_{\mathbf{a}}$ and $\hat{\widetilde{\mathbf{C}}}_{\mathbf{e}}$ as the blocking matrices, and the procedure of completing the blocking matrix follows as previously described, with the speech estimate, $\widetilde{\mathrm{e}}_{1,\mathrm{evd}}$.

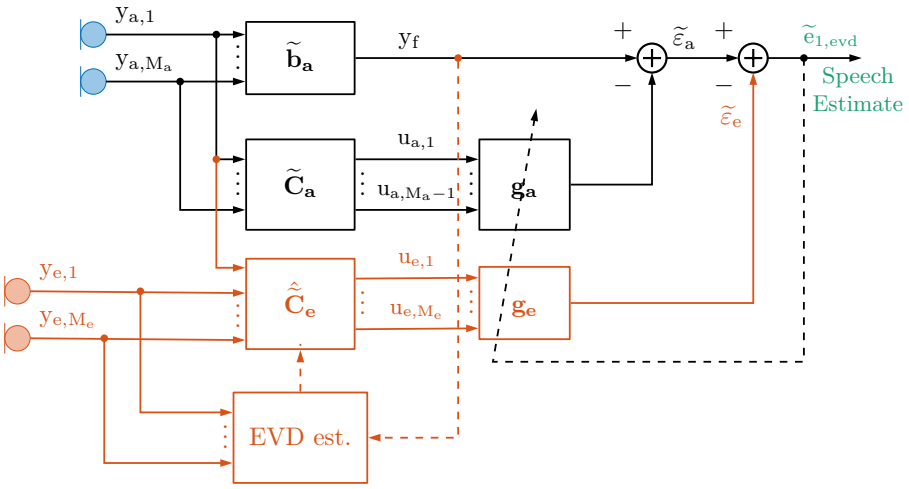


Figure 5.4: $\text{GSC}^{\text{AP-EVD}}_{\text{LMA-XM}}$: A $\text{GSC}^{\text{AP}}_{\text{LMA}}$ extended with XMs and using the EVD method of estimating the missing RTF (EVD est.) component for the XMs from (5.34).

5.5 Rank-1 GEVD Method

In [246], a method of estimating $\mathbf{h_e}$ (for $M_e = 1$) using covariance whitening, or equivalently, a GEVD was proposed. In this section, some modifications will be made to this method, as well as an extension for the general case of $M_e \geq 1$, which will lead to an alternative scheme compared to the previous section. This new scheme will still make use of the $\text{GSC}^{\text{AP}}_{\text{LMA}}$, and the inclusion of the XM will once again be used as an "add-on" to the speech enhancement system. As the mathematical derivations involved may detract from the conceptual aspect of this method, an overview of the resulting scheme and its utility is firstly presented in this section, followed by the relevant mathematical details.

5.5.1 Overview of the method

Fig. 5.5 reveals the resulting scheme, which will be referred to as the $\text{GSC}^{\text{AP-GEVD}}_{\text{LMA-XM}}$. Firstly, the $(M_a + M_e)$ signals will undergo the transformation from (5.41), which is simply the application of the fixed beamformer, $\tilde{\mathbf{f}}_\mathbf{a}$, and the blocking matrix, $\widetilde{\mathbf{C}}_\mathbf{a}$ on the LMA signals as is done in the $\text{GSC}^{\text{AP}}_{\text{LMA}}$, along with the unmodified XM signals.

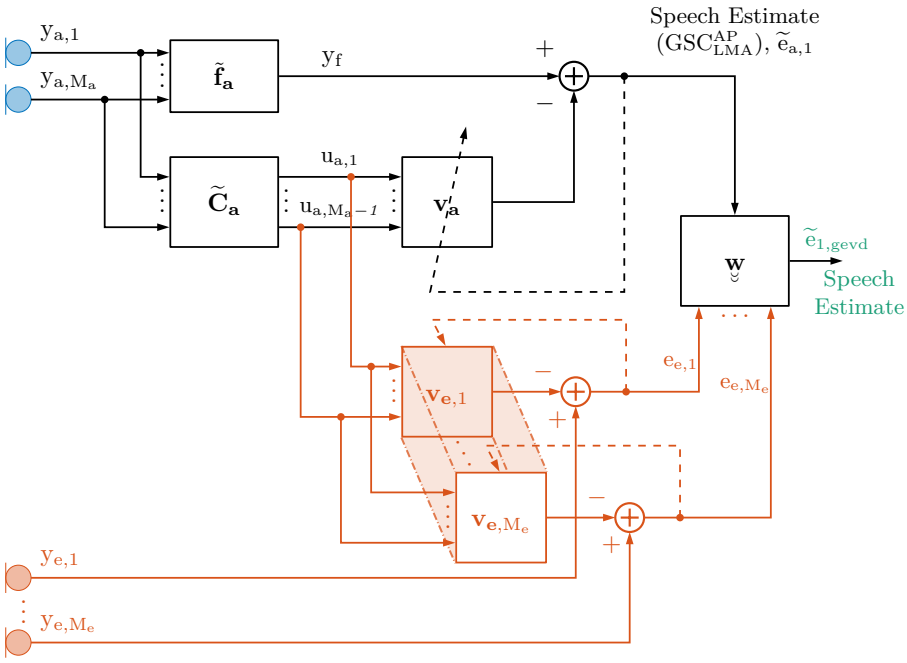


Figure 5.5: GSC$^{\text{AP-GEVD}}_{\text{LMA-XM}}$: A GSC$^{\text{AP}}_{\text{LMA}}$ extended with XMs involving a rank-1 GEVD-based RTF vector estimation procedure.

This is then followed by the orthogonalisation of the noise components of y_f and $\mathbf{y_e}$ onto the noise components of $\mathbf{u_a}$. Such an orthogonalisation can be performed in noise-only periods by using adaptive filters. The resulting (M_e+1) signals after this orthogonalisation procedure are then denoted as:

$$\underset{\sim}{\mathbf{y}}(l) = \left[\widetilde{e}_{a,1}(l) \;\; e_{e,1}(l) \;\; \dots \;\; e_{e,M_e}(l)\right]^T \tag{5.35}$$

consisting of the speech output from a GSC$^{\text{AP}}_{\text{LMA}}$, $\widetilde{e}_{a,1}(l)$, and the vector of XM signals who have had their noise components orthogonalised onto the noise components of $\mathbf{u_a}$, $\left[e_{e,1}(l) \;\; \dots \;\; e_{e,M_e}(l)\right]^T$. Since the orthogonalisation of the noise components of y_f onto the noise components of $\mathbf{u_a}$ is equivalent to considering the optimisation equation of (5.6) from the GSC$^{\text{AP}}_{\text{LMA}}$, the speech output from a GSC$^{\text{AP}}_{\text{LMA}}$, $\widetilde{e}_{a,1}$ corresponds to the first element in $\underset{\sim}{\mathbf{y}}$.

For the orthogonalisation involving the XM signals, a separate $(M_a-1)\times 1$ adaptive filter, $\mathbf{v}_{\mathbf{e},i}$ will have to be introduced for each of the XMs, such that it minimises the same equation as in (5.6), but with $y_{e,i}$ as the desired signal.

Therefore, the optimal filter for $\mathbf{v}_{\mathbf{e},i}$ can be computed in noise-only frames as:

$$\hat{\mathbf{v}}_{\mathbf{e},i}(\mathrm{l}) = (\widetilde{\mathbf{C}}_{\mathbf{a}}^{H}\hat{\mathbf{R}}_{\mathbf{n_a n_a}}(\mathrm{l})\widetilde{\mathbf{C}}_{\mathbf{a}})^{-1}\widetilde{\mathbf{C}}_{\mathbf{a}}^{H}\hat{\mathbf{r}}_{\mathbf{n_a}\mathrm{n}_{\mathrm{e},i}}(\mathrm{l}) \tag{5.36}$$

where $\hat{\mathbf{r}}_{\mathbf{n_a}\mathrm{n}_{\mathrm{e},i}}$ is an estimate of $\mathbf{r}_{\mathbf{n_a}\mathrm{n}_{\mathrm{e},i}} = \mathbb{E}\{\mathbf{n_a}\mathrm{n}_{\mathrm{e},i}^{*}\}$. The resulting error from this orthogonalisation step for the ith XM is then:

$$\mathrm{e}_{\mathrm{e},i}(\mathrm{l}) = \mathrm{y}_{\mathrm{e},i}(\mathrm{l}) - \hat{\mathbf{v}}_{\mathbf{e},i}^{H}(\mathrm{l})\widetilde{\mathbf{C}}_{\mathbf{a}}^{H}\mathbf{y_a}(\mathrm{l}) \tag{5.37}$$

Finally, the filter, $\underset{\sim}{\mathbf{w}}$, which involves a GEVD procedure (derived in the following section), can be used to filter the signals $\mathbf{y}$ in the corresponding time frame to yield the corresponding speech estimate, $\underset{\sim}{\widetilde{\mathrm{e}}}_{1,\mathrm{gevd}}$.

From Fig. 5.5, it can easily be observed that the XMs are truly incorporated in a modular fashion or as "add-ons" to an existing $\mathrm{GSC}_{\mathrm{LMA}}^{\mathrm{AP}}$. One advantage of this implementation over the previously described approach of completing the blocking matrix is that the speech estimate, $\widetilde{\mathrm{e}}_{\mathrm{a},1}$ is still available and could be reverted to in cases where using the XM signals may yield undesirable behaviour.

Furthermore, as will be demonstrated in (5.62), $\underset{\sim}{\mathbf{w}}$ is a low dimensional MVDR beamformer, which uses a rank-1 GEVD-based RTF vector estimate. Consequently, the $\mathrm{GSC}_{\mathrm{LMA\text{-}XM}}^{\mathrm{AP\text{-}GEVD}}$ represents a combination of a $\mathrm{GSC}_{\mathrm{LMA}}^{\mathrm{AP}}$ and an MVDR with a rank-1 GEVD-based RTF vector estimate, which encompasses a variety of filtering schemes. In particular, for $\mathrm{M_e} = 0$, the $\mathrm{GSC}_{\mathrm{LMA\text{-}XM}}^{\mathrm{AP\text{-}GEVD}}$ is equivalent to the $\mathrm{GSC}_{\mathrm{LMA}}^{\mathrm{AP}}$, and for $\mathrm{M_a} = 1, \mathrm{M_e} \neq 0$, the $\mathrm{GSC}_{\mathrm{LMA\text{-}XM}}^{\mathrm{AP\text{-}GEVD}}$ is a MVDR with a rank-1 GEVD-based RTF vector estimate, which uses the single local microphone as the reference signal.

In the sections that follow, a derivation is firstly given for estimating $\mathbf{h_e}$ with the a priori knowledge, $\widetilde{\mathbf{h}}_{\mathbf{a}}$, and the GEVD procedure, followed by details on how this is then incorporated into the $\mathrm{MVDR}_{\mathrm{LMA\text{-}XM}}$ from section 5.3.1 in order to yield the $\mathrm{GSC}_{\mathrm{LMA\text{-}XM}}^{\mathrm{AP\text{-}GEVD}}$ scheme depicted in Fig. 5.5.

5.5.2 GEVD-based RTF vector estimation

In order to extend the estimation problem of (5.25) to a GEVD, a spatial pre-whitening operation is firstly defined from $\hat{\mathbf{R}}_{\mathbf{nn}}$ using the Cholesky decomposition:

$$\hat{\mathbf{R}}_{\mathbf{nn}} = \hat{\mathbf{R}}_{\mathbf{nn}}^{1/2}\,\hat{\mathbf{R}}_{\mathbf{nn}}^{H/2} \tag{5.38}$$

where $\hat{\mathbf{R}}_{\mathbf{nn}}^{1/2}$ is a lower triangular matrix, and $\hat{\mathbf{R}}_{\mathbf{nn}}^{H/2}$ is its hermitian transpose. It is recalled from subsection 3.4.2 that spatial pre-whitening is then performed

by pre-multiplying the signal vector of interest by $\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2}$. For an autocorrelation matrix, spatial pre-whitening is performed by pre-multiplying it by $\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2}$ and post-multiplying it by $\hat{\mathbf{R}}_{\mathbf{nn}}^{-H/2}$. Firstly, a pre-whitened version of (5.24) is:

$$\underset{\sigma^2_{\mathrm{s_{a,1}}},\mathbf{h}}{\text{minimise}} \quad ||\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2}((\hat{\mathbf{R}}_{\mathbf{yy}}-\hat{\mathbf{R}}_{\mathbf{nn}})-\sigma^2_{\mathrm{s_{a,1}}}\mathbf{h}\mathbf{h}^H)\hat{\mathbf{R}}_{\mathbf{nn}}^{-H/2}||_F^2 \tag{5.39}$$

which is the optimisation problem considered when there is no available a priori RTF vector, i.e. the entire RTF vector for all microphones is directly estimated from the observed data. It is essentially a matrix approximation problem similar to (5.24), however it is now to find the best rank-1 matrix to approximate $\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2}((\hat{\mathbf{R}}_{\mathbf{yy}}-\hat{\mathbf{R}}_{\mathbf{nn}})\hat{\mathbf{R}}_{\mathbf{nn}}^{-H/2}$, and hence the solution follows from an EVD of this matrix. However, since $\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2}\hat{\mathbf{R}}_{\mathbf{nn}}\hat{\mathbf{R}}_{\mathbf{nn}}^{-H/2}=\mathbf{I}_{(\mathrm{M_a}+\mathrm{M_e})}$, an EVD can be performed on the matrix $\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2}\hat{\mathbf{R}}_{\mathbf{yy}}\hat{\mathbf{R}}_{\mathbf{nn}}^{-H/2}$, since the identity matrix would have only reduced all of the eigenvalues equally by one. Consequently, the eigenvectors remain the same for both decompositions and the relative ordering is the same for both. Therefore the data-dependent RTF vector for all microphones can be obtained from the appropriate rotation and scaling of the principal eigenvector from the EVD of $\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2}\hat{\mathbf{R}}_{\mathbf{yy}}\hat{\mathbf{R}}_{\mathbf{nn}}^{-H/2}$ as discussed in subsection 3.4.2, which also corresponds to a GEVD of the matrix pencil $(\hat{\mathbf{R}}_{\mathbf{yy}},\hat{\mathbf{R}}_{\mathbf{nn}})$.

Now for the case when there is an a priori RTF vector available, (5.39) becomes (the pre-whitened version of (5.25)):

$$\underset{\sigma^2_{\mathrm{s_{a,1}}},\mathbf{h_e}}{\text{minimise}}\ ||\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2}((\hat{\mathbf{R}}_{\mathbf{yy}}-\hat{\mathbf{R}}_{\mathbf{nn}})-\sigma^2_{\mathrm{s_{a,1}}}\begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}}\\ \mathbf{h_e}\end{bmatrix}\begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}}^H & \mathbf{h}_{\mathbf{e}}^H\end{bmatrix})\hat{\mathbf{R}}_{\mathbf{nn}}^{-H/2}||_F^2 \tag{5.40}$$

so that only $\mathbf{h_e}$ needs to be estimated as opposed to the entire RTF vector.

In [246], an appropriate transformation matrix has been defined and then (5.40) is solved (for $\mathrm{M_e}=1$) by performing an EVD on a 2×2 pre-whitened correlation matrix.

An alternative approach can however be taken to solve (5.40) that will yield a practical scheme and include the general case of $\mathrm{M_e}\geq1$. Firstly, using the blocking matrix[4]„ $\widetilde{\mathbf{C}}_{\mathbf{a}}$ from (5.5), and the fixed beamformer, $\tilde{\mathbf{f}}_{\mathbf{a}}$, the $(\mathrm{M_a}+\mathrm{M_e})\times(\mathrm{M_a}+\mathrm{M_e})$ transformation matrix, $\widetilde{\mathbf{\Upsilon}}$, can be defined:

$$\widetilde{\mathbf{\Upsilon}}=\left[\begin{array}{c|c}[\widetilde{\mathbf{C}}_{\mathbf{a}}\ \ \tilde{\mathbf{f}}_{\mathbf{a}}] & \mathbf{0}\\ \hline \mathbf{0} & \mathbf{I}_{\mathrm{M_e}}\end{array}\right] \tag{5.41}$$

[4]The blocking matrix, $\widetilde{\mathbf{C}}_{\mathbf{a}}$ is used here as opposed to $\widetilde{\mathbf{B}}_{\mathbf{a}}$ from (5.27) to construct the transformation matrix.

from which the transformed speech-plus-noise signals and the transformed noise-only signals are defined respectively as:

$$\widetilde{\boldsymbol{\Upsilon}}^H\mathbf{y} = \begin{bmatrix} \widetilde{\mathbf{C}}_\mathbf{a}^H\mathbf{y_a} \\ \tilde{\mathbf{f}}_\mathbf{a}^H\mathbf{y_a} \\ \mathbf{y_e} \end{bmatrix}; \quad \widetilde{\boldsymbol{\Upsilon}}^H\mathbf{n} = \begin{bmatrix} \widetilde{\mathbf{C}}_\mathbf{a}^H\mathbf{n_a} \\ \tilde{\mathbf{f}}_\mathbf{a}^H\mathbf{n_a} \\ \mathbf{n_e} \end{bmatrix} \tag{5.42}$$

consisting of the blocking matrix signals from the LMA, the fixed beamformer output signal, and the XM signals.

Another spatial pre-whitening operation can be subsequently defined from the noise-only correlation matrix of the transformed noise-only signals, similar to that of (5.38):

$$\mathbb{E}\{(\widetilde{\boldsymbol{\Upsilon}}^H\mathbf{n})(\widetilde{\boldsymbol{\Upsilon}}^H\mathbf{n})^H\} = \widetilde{\boldsymbol{\Upsilon}}^H\mathbf{R_{nn}}\widetilde{\boldsymbol{\Upsilon}} = \mathbf{L}\mathbf{L}^H \tag{5.43}$$

where $\mathbf{L}$ is an $(\mathrm{M_a}+\mathrm{M_e})\times(\mathrm{M_a}+\mathrm{M_e})$ lower triangular matrix. As $\mathbf{R_{nn}}$ will be replaced with an estimate, $\hat{\mathbf{R}}_\mathbf{nn}$, a corresponding estimate of $\mathbf{L}$ will be obtained, $\hat{\mathbf{L}}$, which can be represented as a block matrix:

$$\hat{\mathbf{L}} = \left[\begin{array}{c|c} \hat{\mathbf{L}}_{\mathbf{a}_-} & \mathbf{0} \\ \hline \hat{\mathbf{L}}_\mathbf{c} & \hat{\mathbf{L}}_{\mathbf{e}_+} \end{array}\right] \tag{5.44}$$

whose block dimensions are such that $\hat{\mathbf{L}}_{\mathbf{a}_-}$ is an $(\mathrm{M_a}-1)\times(\mathrm{M_a}-1)$ lower triangular matrix, $\hat{\mathbf{L}}_\mathbf{c}$ an $(\mathrm{M_e}+1)\times(\mathrm{M_a}-1)$, and $\hat{\mathbf{L}}_{\mathbf{e}_+}$ is an $(\mathrm{M_e}+1)\times(\mathrm{M_e}+1)$ lower triangular matrix. By computing the block inverse of $\hat{\mathbf{L}}$ in (5.44), $\hat{\mathbf{L}}^{-1}$ is then:

$$\begin{aligned} \hat{\mathbf{L}}^{-1} &= \left[\begin{array}{c|c} \hat{\mathbf{L}}_{\mathbf{a}_-}^{-1} & \mathbf{0} \\ \hline -\hat{\mathbf{L}}_{\mathbf{e}_+}^{-1}\hat{\mathbf{L}}_\mathbf{c}\hat{\mathbf{L}}_{\mathbf{a}_-}^{-1} & \hat{\mathbf{L}}_{\mathbf{e}_+}^{-1} \end{array}\right] \\ &= \underbrace{\left[\begin{array}{c|c} \hat{\mathbf{L}}_{\mathbf{a}_-}^{-1} & \mathbf{0} \\ \hline \mathbf{0} & \hat{\mathbf{L}}_{\mathbf{e}_+}^{-1} \end{array}\right]}_{\hat{\boldsymbol{\Theta}}_\mathbf{2}^H} \underbrace{\left[\begin{array}{c|c} \mathbf{I}_{\mathrm{M_a}-1} & \mathbf{0} \\ \hline -\hat{\mathbf{L}}_\mathbf{c}\hat{\mathbf{L}}_{\mathbf{a}_-}^{-1} & \mathbf{I}_{\mathrm{M_e}+1} \end{array}\right]}_{\hat{\boldsymbol{\Theta}}_\mathbf{1}^H} \end{aligned} \tag{5.45}$$

which is split into two separate orthogonalisations, $\hat{\boldsymbol{\Theta}}_\mathbf{1}^H$, which will orthogonalise the noise components from the last $(\mathrm{M_e}+1)$ signals onto the first $(\mathrm{M_a}-1)$ signals, and $\hat{\boldsymbol{\Theta}}_\mathbf{2}^H$, which will complete the entire orthogonalisation operation. A second transformation of the transformed speech-plus-noise and the noise-only signals from (5.42) can then be defined respectively such that:

$$\underset{\smile}{\mathbf{y}} = \hat{\boldsymbol{\Theta}}_\mathbf{1}^H\widetilde{\boldsymbol{\Upsilon}}^H\mathbf{y}; \quad \underset{\smile}{\mathbf{n}} = \hat{\boldsymbol{\Theta}}_\mathbf{1}^H\widetilde{\boldsymbol{\Upsilon}}^H\mathbf{n} \tag{5.46}$$

where the notation $\underset{\smile}{\{.\}}$ indicates that the transformed signals from (5.42) have had the noise components from the fixed beamformer output as well as the XM

signals orthogonalised onto the noise components of the blocking matrix signals from the LMA (i.e. orthogonalised onto $\mathbf{u_a}$).

With the relevant transformations defined, the estimation problem from (5.40) can be subsequently re-formulated as:

$$\begin{aligned}\underset{\sigma^2_{\mathrm{s_{a,1}}},\,\mathbf{h_e}}{\text{minimise}}||\hat{\mathbf{R}}_{\mathbf{nn}}^{^{-1/2}}\widetilde{\mathbf{\Upsilon}}^{-H}\hat{\mathbf{\Theta}}_{\mathbf{1}}^{-H}\hat{\mathbf{\Theta}}_{\mathbf{1}}^{H}\widetilde{\mathbf{\Upsilon}}^{H}((\hat{\mathbf{R}}_{\mathbf{yy}}-\hat{\mathbf{R}}_{\mathbf{nn}})\\ -\sigma^2_{\mathrm{s_{a,1}}}\begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}}\\ \mathbf{h_e}\end{bmatrix}\begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}}^{H}\ \mathbf{h_e}^{H}\end{bmatrix})\widetilde{\mathbf{\Upsilon}}\hat{\mathbf{\Theta}}_{\mathbf{1}}\hat{\mathbf{\Theta}}_{\mathbf{1}}^{-1}\widetilde{\mathbf{\Upsilon}}^{-1}\hat{\mathbf{R}}_{\mathbf{nn}}^{^{-H/2}}||_F^2\end{aligned} \tag{5.47}$$

which can be simplified by realising:

$$\begin{aligned}\hat{\mathbf{R}}_{\mathbf{nn}}^{^{-1/2}}\widetilde{\mathbf{\Upsilon}}^{-H}\hat{\mathbf{\Theta}}_{\mathbf{1}}^{-H}&=(\widetilde{\mathbf{\Upsilon}}^{H}\hat{\mathbf{R}}_{\mathbf{nn}}^{^{1/2}})^{-1}\hat{\mathbf{\Theta}}_{\mathbf{1}}^{-H}\\ &=(\hat{\mathbf{L}}\mathbf{U})^{-1}\hat{\mathbf{\Theta}}_{\mathbf{1}}^{-H}\\ &=\mathbf{U}^{H}\underbrace{\hat{\mathbf{L}}^{-1}\hat{\mathbf{\Theta}}_{\mathbf{1}}^{-H}}_{\hat{\mathbf{\Theta}}_{\mathbf{2}}^{H}}\end{aligned} \tag{5.48}$$

where $\mathbf{U}$ is some unitary matrix. Since the Frobenius norm is invariant under a unitary transformation, (5.47) can be re-written as:

$$\underset{\sigma^2_{\mathrm{s_{a,1}}},\,\mathbf{h_e}}{\text{minimise}}||\hat{\mathbf{\Theta}}_{\mathbf{2}}^{H}(\hat{\underset{\smile}{\mathbf{R}}}_{\mathbf{yy}}-\hat{\underset{\smile}{\mathbf{R}}}_{\mathbf{nn}})\hat{\mathbf{\Theta}}_{\mathbf{2}}-\hat{\mathbf{\Theta}}_{\mathbf{2}}^{H}\hat{\mathbf{\Theta}}_{\mathbf{1}}^{H}\widetilde{\mathbf{\Upsilon}}^{H}(\sigma^2_{\mathrm{s_{a,1}}}\begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}}\\ \mathbf{h_e}\end{bmatrix}\begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}}^{H}\ \mathbf{h_e}^{H}\end{bmatrix})\widetilde{\mathbf{\Upsilon}}\hat{\mathbf{\Theta}}_{\mathbf{1}}\hat{\mathbf{\Theta}}_{\mathbf{2}}||_F^2 \tag{5.49}$$

where:

$$\hat{\underset{\smile}{\mathbf{R}}}_{\mathbf{yy}}=\hat{\mathbf{\Theta}}_{\mathbf{1}}^{H}\widetilde{\mathbf{\Upsilon}}^{H}\hat{\mathbf{R}}_{\mathbf{yy}}\widetilde{\mathbf{\Upsilon}}\hat{\mathbf{\Theta}}_{\mathbf{1}} \tag{5.50}$$

$$\hat{\underset{\smile}{\mathbf{R}}}_{\mathbf{nn}}=\hat{\mathbf{\Theta}}_{\mathbf{1}}^{H}\widetilde{\mathbf{\Upsilon}}^{H}\hat{\mathbf{R}}_{\mathbf{nn}}\widetilde{\mathbf{\Upsilon}}\hat{\mathbf{\Theta}}_{\mathbf{1}}=\left[\begin{array}{c|c}\hat{\mathbf{L}}_{\mathbf{a}_-}\hat{\mathbf{L}}_{\mathbf{a}_-}^{H} & \mathbf{0}\\ \hline \mathbf{0} & \hat{\mathbf{L}}_{\mathbf{e}_+}\hat{\mathbf{L}}_{\mathbf{e}_+}^{H}\end{array}\right] \tag{5.51}$$

and also since $\widetilde{\mathbf{C}}_{\mathbf{a}}^{H}\widetilde{\mathbf{h}}_{\mathbf{a}}=\mathbf{0}$:

$$\hat{\mathbf{\Theta}}_{\mathbf{1}}^{H}\widetilde{\mathbf{\Upsilon}}^{H}\begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}}\\ \mathbf{h_e}\end{bmatrix}=\begin{bmatrix}\mathbf{0}\\ \mathrm{d_a}\\ \mathbf{h_e}\end{bmatrix} \tag{5.52}$$

which consists of a vector with $(\mathrm{M_a}-1)$ zeros, together with $\mathrm{d_a}=\tilde{\mathbf{f}}_{\mathbf{a}}^{H}\widetilde{\mathbf{h}}_{\mathbf{a}}$ and the RTF vector for the XM signals to be estimated. Upon substitution of (5.52) into (5.49), it can be observed that the dimensionality of (5.49) can be reduced to a lower order of $(\mathrm{M_e}+1)$, similarly to what was done in the EVD method of

section 5.4.2. The lower order $(\mathrm{M_e}+1)$ speech-plus-noise and noise-only signals can be defined as follows, denoted with the double underbreve:

$$\underbrace{\mathbf{y}} = \mathbf{J}^T\mathbf{y}; \quad \underbrace{\mathbf{n}} = \mathbf{J}^T\mathbf{n} \tag{5.53}$$

where $\mathbf{J} = [\ \mathbf{0}_{(\mathrm{M_e}+1)\times(\mathrm{M_a}-1)} \mid \mathbf{I}_{(\mathrm{M_e}+1)}\ ]^T$ is a selection matrix. Hence, $\underbrace{\mathbf{y}}$ and $\underbrace{\mathbf{n}}$ consist of the single fixed beamformer output signal and the $\mathrm{M_e}$ XM signals, both of which have had their noise components orthogonalised onto the blocking matrix signals. The reduced dimensionality of (5.49) then follows as:

$$\underset{\sigma^2_{\mathrm{s_{a,1}}},\,\mathbf{h_e}}{\text{minimise}} ||\hat{\mathbf{L}}^{-1}_{\mathbf{e}_+}((\hat{\underbrace{\mathbf{R}}}_{\mathbf{yy}} - \hat{\underbrace{\mathbf{R}}}_{\mathbf{nn}}) - (\sigma^2_{\mathrm{s_{a,1}}} \begin{bmatrix} \mathrm{d_a} \\ \mathbf{h_e} \end{bmatrix} \begin{bmatrix} \mathrm{d_a^*}\ \mathbf{h_e}^H \end{bmatrix}))\hat{\mathbf{L}}^{-H}_{\mathbf{e}_+}||^2_F \tag{5.54}$$

where the $(\mathrm{M_e}+1)\times(\mathrm{M_e}+1)$ matrices $\hat{\underbrace{\mathbf{R}}}_{\mathbf{yy}}$ and $\hat{\underbrace{\mathbf{R}}}_{\mathbf{nn}}$ are:

$$\hat{\underbrace{\mathbf{R}}}_{\mathbf{yy}} = \mathbf{J}^T\hat{\underbrace{\mathbf{R}}}_{\mathbf{yy}}\mathbf{J}; \quad \hat{\underbrace{\mathbf{R}}}_{\mathbf{nn}} = \mathbf{J}^T\hat{\underbrace{\mathbf{R}}}_{\mathbf{nn}}\mathbf{J} = \hat{\mathbf{L}}_{\mathbf{e}_+}\hat{\mathbf{L}}^H_{\mathbf{e}_+} \tag{5.55}$$

and hence $\hat{\mathbf{L}}^{-1}_{\mathbf{e}_+}\hat{\underbrace{\mathbf{R}}}_{\mathbf{nn}}\hat{\mathbf{L}}^{-H}_{\mathbf{e}_+} = \mathbf{I}_{\mathrm{M_e}+1}$.

The solution then follows from a GEVD of the matrix pencil $(\hat{\underbrace{\mathbf{R}}}_{\mathbf{yy}}, \hat{\underbrace{\mathbf{R}}}_{\mathbf{nn}})$, where a joint diagonalisation is done:

$$\hat{\underbrace{\mathbf{R}}}_{\mathbf{yy}} = \hat{\underbrace{\mathbf{Q}}}\hat{\underbrace{\mathbf{\Sigma}}}_{\mathbf{y}}\hat{\underbrace{\mathbf{Q}}}^H; \quad \hat{\underbrace{\mathbf{R}}}_{\mathbf{nn}} = \hat{\underbrace{\mathbf{Q}}}\hat{\underbrace{\mathbf{\Sigma}}}_{\mathbf{n}}\hat{\underbrace{\mathbf{Q}}}^H \tag{5.56}$$

where $\hat{\underbrace{\mathbf{Q}}}$ is a full-rank, $(\mathrm{M_e}+1)\times(\mathrm{M_e}+1)$ invertible matrix, $\hat{\underbrace{\mathbf{\Sigma}}}_{\mathbf{y}}$ and $\hat{\underbrace{\mathbf{\Sigma}}}_{\mathbf{n}}$ are real valued $(\mathrm{M_e}+1)\times(\mathrm{M_e}+1)$ diagonal matrices arranged in descending order according to the magnitude of the generalised eigenvalues. Using the principal eigenvector from the matrix, $(\hat{\underbrace{\mathbf{R}}}_{\mathbf{yy}} - \hat{\underbrace{\mathbf{R}}}_{\mathbf{nn}} = \hat{\underbrace{\mathbf{Q}}}(\hat{\underbrace{\mathbf{\Sigma}}}_{\mathbf{y}} - \hat{\underbrace{\mathbf{\Sigma}}}_{\mathbf{n}})\hat{\underbrace{\mathbf{Q}}}^H)$, the estimate of the lower dimensional RTF vector then follows as:

$$\begin{bmatrix} \mathrm{d_a} \\ \hat{\tilde{\mathbf{h}}}_{\mathbf{e,gevd}} \end{bmatrix} = \frac{\hat{\underbrace{\mathbf{Q}}}\,\mathbf{e_1}\ \mathrm{d_a}}{\mathbf{e_1}^T\hat{\underbrace{\mathbf{Q}}}\,\mathbf{e_1}} \tag{5.57}$$

where the $(\mathrm{M_e}+1)\times 1$ vector, $\mathbf{e_1} = [1\ \ 0\ldots 0]^T$, and $\hat{\tilde{\mathbf{h}}}_{\mathbf{e,gevd}}$ the estimate of $\mathbf{h_e}$ that makes use of the a priori knowledge, $\tilde{\mathbf{h}}_{\mathbf{a}}$, and the GEVD procedure. $\hat{\tilde{\mathbf{h}}}_{\mathbf{e,gevd}}$ could then, in fact, be used as a third option for completing the blocking matrix as was done in section 5.4. However, as will be demonstrated in the following section, a substitution of (5.57) into the $\mathrm{MVDR}^{\mathrm{AP}}_{\text{LMA-XM}}$ beamformer will reveal a convenient sequence of operations that leads to the alternative practical implementation depicted in Fig. 5.5 that does not affect the output of the $\mathrm{GSC}^{\mathrm{AP}}_{\mathrm{LMA}}$.

5.5.3 MVDR-LMA-XM beamformer

Using the definitions from (5.43), and the result of (5.57), the numerator of (5.10) can firstly be written as:

$$\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\widetilde{\mathbf{h}} = \widetilde{\boldsymbol{\Upsilon}}\hat{\mathbf{L}}^{-H}\hat{\mathbf{L}}^{-1}\widetilde{\boldsymbol{\Upsilon}}^{H}\begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}}\\ \hat{\underset{\sim}{\mathbf{h}}}_{\mathbf{e,gevd}}\end{bmatrix} \tag{5.58}$$

Using (5.44), (5.45), (5.55), and (5.57) eventually results in:

$$\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\widetilde{\mathbf{h}} = \widetilde{\boldsymbol{\Upsilon}}\begin{bmatrix}-\hat{\mathbf{L}}_{\mathbf{a}_-}^{-H}\hat{\mathbf{L}}_{\mathbf{c}}^{H}\\ \mathbf{I}_{\mathrm{M_e}+1}\end{bmatrix}\hat{\underset{\sim}{\mathbf{Q}}}^{-H}\hat{\underset{\sim}{\boldsymbol{\Sigma}}}_{\mathbf{n}}^{-1}\frac{\mathbf{e_1}\,\mathrm{d_a}}{\mathbf{e_1}^{T}\hat{\underset{\sim}{\mathbf{Q}}}\mathbf{e_1}} \tag{5.59}$$

Finally, making the relevant substitutions in the denominator of (5.10), the $\mathrm{MVDR}_{\text{LMA-XM}}^{\text{AP}}$ beamformer becomes:

$$\widetilde{\mathbf{w}} = \underbrace{\widetilde{\boldsymbol{\Upsilon}}}_{\text{Trans}}\underbrace{\begin{bmatrix}-\hat{\mathbf{L}}_{\mathbf{a}_-}^{-H}\hat{\mathbf{L}}_{\mathbf{c}}^{H}\\ \mathbf{I}_{\mathrm{M_e}+1}\end{bmatrix}}_{\text{Orthogonalisation}}\underbrace{\hat{\underset{\sim}{\mathbf{Q}}}^{-H}\mathbf{e_1}}_{\text{GEVD}}\underbrace{\frac{\mathbf{e_1}^{T}\hat{\underset{\sim}{\mathbf{Q}}}^{H}\mathbf{e_1}}{\mathrm{d_a}}}_{\text{scaling}} \tag{5.60}$$

which reveals the distinct operations that are required in order to implement the scheme of Fig. 5.5. Namely, it consists of the transformation operation from (5.41), the orthogonalisation operation onto the noise components of the blocking matrix signals from the LMA, the GEVD operation and the appropriate scaling.

Consequently, the resulting speech estimate, i.e. $\widetilde{\mathbf{w}}^{H}\mathbf{y}$, is then computed as:

$$\widetilde{\mathrm{e}}_{1,\mathrm{gevd}} = \underbrace{\frac{\mathbf{e_1}^{T}\hat{\underset{\sim}{\mathbf{Q}}}\mathbf{e_1}}{\mathrm{d_a}}\;\mathbf{e_1}^{T}\hat{\underset{\sim}{\mathbf{Q}}}^{-1}}_{\underset{\sim}{\mathbf{w}}^{H}}\;\underbrace{[-\hat{\mathbf{L}}_{\mathbf{c}}\hat{\mathbf{L}}_{\mathbf{a}_-}^{-1}\mid\mathbf{I}_{\mathrm{M_e}+1}]\begin{bmatrix}\widetilde{\mathbf{C}}_{\mathbf{a}}^{H}\mathbf{y_a}\\ \tilde{\mathbf{f}}_{\mathbf{a}}^{H}\mathbf{y_a}\\ \mathbf{y_e}\end{bmatrix}}_{\underset{\sim}{\mathbf{y}}} \tag{5.61}$$

which can be realised as a lower dimensional $(\mathrm{M_e}+1)$ tap beamformer, $\underset{\sim}{\mathbf{w}}$ that acts directly on the orthogonalised, transformed signals, $\underset{\sim}{\mathbf{y}}$ (see (5.35)). As opposed to directly computing the matrix products, $\underset{\sim}{\mathbf{y}}$ can also be obtained by using the adaptive filters as described in section 5.5.1((5.35) - (5.37)). It is also to be noted that $\underset{\sim}{\mathbf{w}}$ can be equivalently formulated in the familiar MVDR beamformer structure as:

$$\underset{\sim}{\mathbf{w}} = \frac{\hat{\underset{\sim}{\mathbf{R}}}_{\mathbf{nn}}^{-1}\widetilde{\underset{\sim}{\mathbf{h}}}}{\widetilde{\underset{\sim}{\mathbf{h}}}^{H}\hat{\underset{\sim}{\mathbf{R}}}_{\mathbf{nn}}^{-1}\widetilde{\underset{\sim}{\mathbf{h}}}} \tag{5.62}$$

where the lower dimensional RTF vector, $\widetilde{\underset{\sim}{\mathbf{h}}} = \begin{bmatrix}\mathrm{d_a}\\ \hat{\underset{\sim}{\mathbf{h}}}_{\mathbf{e,gevd}}\end{bmatrix}$.

5.6 Evaluation and Discussion

In this section, the various algorithms are evaluated on recorded audio data from dummy[5] behind-the-ear (BTE) hearing aid (HA) microphones as well as XMs. The BTE-HA microphones served as the LMA (either two or four were used depending on the experiment) and are shown in Fig. 5.6. In Sec. 5.6.1, the algorithms are evaluated in controlled scenario in an office room. In Sec. 5.6.2, the $\mathrm{GSC}^{\text{AP-GEVD}}_{\text{LMA-XM}}$ and the $\mathrm{GSC}^{\text{AP}}_{\text{LMA}}$ are evaluated in a more realistic scenario in a public environment. Finally, in Sec. 5.6.3, a comparison is made between the MVDR beamformer for an LMA and XMs that uses a priori spatial knowledge and one that relies on a fully data-dependent RTF vector estimate.

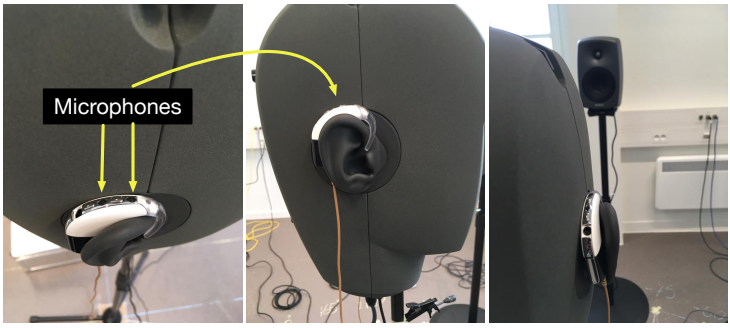


Figure 5.6: Dummy behind-the-ear (BTE) hearing aid (HA) on a Neumann KU-100 dummy head. The BTE-HA consisted of two microphones spaced approximately 1.3 cm apart as shown on the leftmost photo. In some of the experiments, the BTE-HA was used on either the dummy head as shown, or on a test subject.

5.6.1 Evaluation in an office room

The various algorithms were initially evaluated on audio recordings made in an office room of dimensions 5.4 m × 3.5 m × 2.5 m with an estimated broadband reverberation time of 0.3 s. The scenario under which audio recordings were made is depicted in Fig. 5.7. At the centre of the scenario, a test subject wore the BTE-HA hearing aid (HA), the two microphones of which served as the LMA. The test subject was instructed to always face towards the direction of 0°, i.e. in the direction of the speech source, which was placed 1 m away. Four XMs were distributed as shown, with XM3 being worn by the test subject,

[5]This means that microphone signals alone was captured, without any other pre-processing applied.

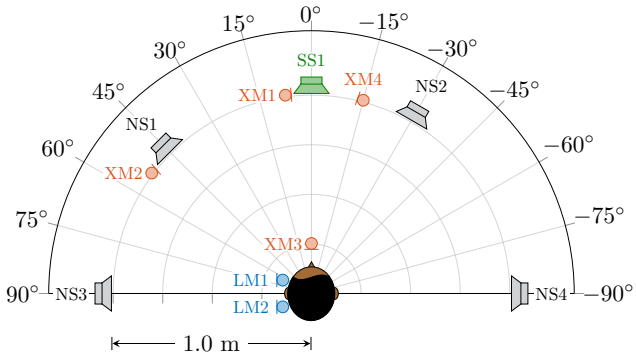


Figure 5.7: Acoustic scenario illustrating the spatial distribution of the speech source (SS1), the noise sources (NS1 - NS4), the LMA (LM1, LM2), and the XMs (XM1 - XM4).

clipped onto the chest. AKG-CK-97-O microphones were used for XM1 and XM3, and AKG-CK-32 microphones were used for XM2 and XM4. A Genelec 8020C loudspeaker was used to generate the speech signal (SS1), which was that of a male speaker from [244]. Several loudspeakers to serve as noise sources (NS) were also distributed as shown, also 1 m away from the test subject. NS1 and NS2 were JBL Control 1 Pro loudspeakers, and NS3 and NS4 were Harman-Kardon HK206 loudspeakers. The noise signals considered were two uncorrelated excerpts of multitalker babble noise (`bb1` and `bb2`) from [206], speech-shaped noise (`sn1`) constructed from a 12-coefficient linear predictive coder (LPC) using the male speaker from [244], and noise from the interior of a busy coffee shop (`cc1`) [248]. With the test subject facing toward 0°, each of these noises played through each of the NS positions was recorded separately. The speech signal from SS1 was also separately recorded. All recorded signals had a duration of 60 s. With this database of speech and noises, subsets of several scenarios stemming from that of Fig. 5.7 could then be constructed and analysed accordingly.

For the processing of the algorithms in the various acoustic scenarios, the Weighted Overlap and Add (WOLA) method [101], with a Discrete Fourier Transform (DFT) size of 256, 50 % overlap, a square-root Hanning window, and sampling frequency of 16 kHz was used. Separate experiments were done using either a perfect voice activity detector (VAD) or an imperfect VAD (using the minimum statistics method from [210]) to indicate the time frames where speech was present. All RTF estimates were performed in time frames where periods where the speech was present (as indicated by the respective VAD) and all the relevant correlation matrices were computed using a forgetting factor

corresponding to a time constant of 1 s. For all experiments, the correlation matrices were initialised such that all elements were set to zero. It should be noted that these VADs were applied on the time domain noisy signal of the microphone, LM1. Hence if it was detected that a certain time frame contained speech, all frequency bins were subsequently treated as such. As not all frequency bins will truly contain speech, the use of the word “perfect" in this context is somewhat of a misnomer. It is rather used to distinguish two instances of estimating the correlation matrices, i.e. one which is better (“perfect") and another that does not perform as well (“imperfect"). Further to this point, the “imperfect" VAD from [210] is just one option of a realistic VAD, and one may choose from other methods such as the speech presence probability [192]. The results that ensue are only meant to give an idea of the potential range in performance one can expect from using the various algorithms depending on how well the relevant correlation matrices are estimated.

The performance of the algorithms was also evaluated using two different procedures for defining the a priori RTF vector, $\widetilde{\mathbf{h}}_{\mathbf{a}}$, for the LMA. In the first procedure, a white noise signal of 60 s was played through SS1, which was subsequently used to compute a rank-1 correlation matrix per frequency. With the frontal microphone of the LMA, LM1, as the reference microphone, an EVD on the correlation matrix was performed and the resulting principal eigenvector was used to define $\widetilde{\mathbf{h}}_{\mathbf{a}}$. In the following, this definition of $\widetilde{\mathbf{h}}_{\mathbf{a}}$ will be referred to as $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$. In the second procedure, only the direct path of the RTF vector was used for defining $\widetilde{\mathbf{h}}_{\mathbf{a}}$ and will be referred to as $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$. Hence $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}} = [1\ e^{-j\omega\tau_2(\theta)}]^T$, where $\tau_2(\theta)$ is the relative time delay between the two microphones of the LMA, and θ is the a priori assumed location of the source with respect to the LMA, which was 0° in the experiments that follow.

The metrics used to evaluate the following experiments were the change in Speech Intelligibility-weighted SNR improvement [159] (Δ SI-SNR) from the input SI-SNR at LM1, and the change in short-time objective intelligibility (Δ STOI) [155] from the unprocessed speech only signal in LM1. In addition to the GSC algorithms, the metrics were also computed on the particular XMs used in the various acoustic scenarios.

Single XM, Single SS, Single NS

In the first experiment, only SS1 and NS1 were considered (i.e. a single correlated noise) along with the LMA and a single XM, hence 3 microphones in total. The single XM was chosen to be either XM1, XM2 or XM3 in order to observe the impact of the SNR of the XM signal on the performance of the algorithms. The input signal created was a combination of the speech signal (i.e. the male

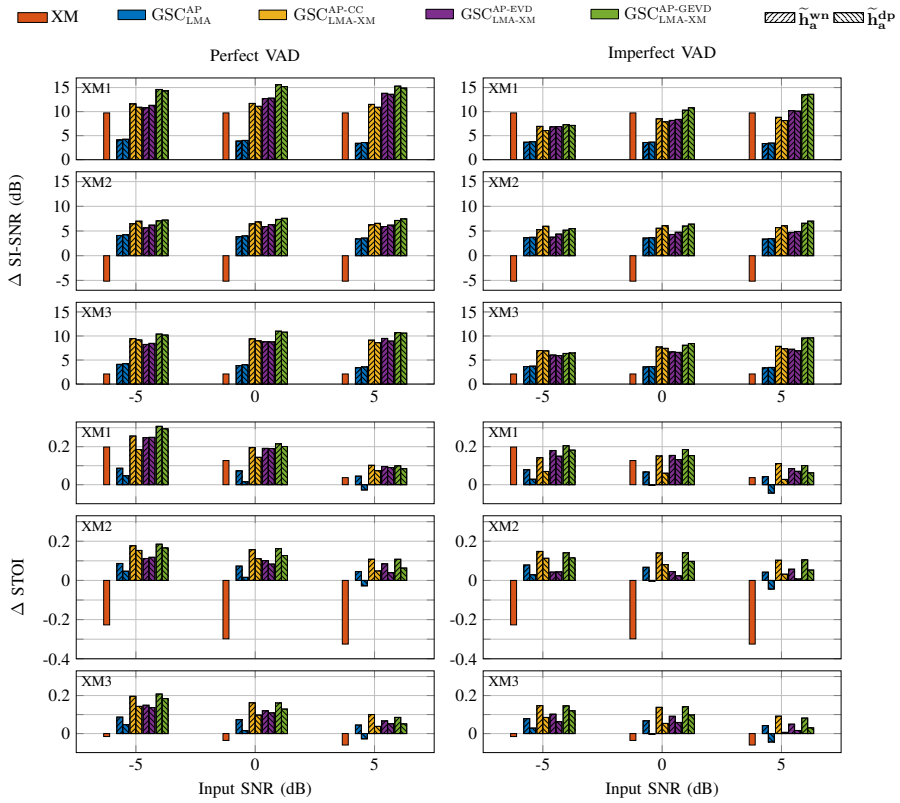


Figure 5.8: Objective metrics for the various GSC algorithms, as well as for the XMs as a function of input SNR at LM1 using a perfect VAD (left) and an imperfect VAD (right). The scenario considered one male speech source (SS1), one mutli-talker babble noise source (NS1), the LMA, and one XM. Each of the three sub-plots within the particular metric uses a different XM as indicated, which corresponds to the XMs from Fig. 5.7. Each group of 2 bars for a particular algorithm represents the processing as performed with either $\widetilde{\mathbf{h}}^{\mathbf{wn}}_{\mathbf{a}}$ or $\widetilde{\mathbf{h}}^{\mathbf{dp}}_{\mathbf{a}}$ as the a priori RTF vector for the LMA.

speaker from [244]) and the `bb1` noise. As these noises were recorded separately from the speech, each noise signal was also scaled such as to vary the unweighted input SNR at LM1 from −5 dB to 5 dB. This unweighted SNR (as opposed to an input SI-SNR) was computed simply from the ratio of the variance of the unprocessed speech signal to the variance of the unprocessed noise signal.

The algorithms evaluated were the GSC$^{AP}_{LMA}$ from section 5.2.2, GSC$^{AP-CC}_{LMA-XM}$

from section 5.4.1, $\text{GSC}_{\text{LMA-XM}}^{\text{AP-EVD}}$ from section 5.4.2, and $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ from section 5.5.3, as well as the individual XM signals. For each of the GSC methods, the optimal filters were used for the computation of the ANC filters, i.e. (5.7) for $\hat{\mathbf{v}}_{\mathbf{a}}$ (in the $\text{GSC}_{\text{LMA}}^{\text{AP}}$ and $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$), (5.17) for $\hat{\mathbf{g}}$ (in the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-CC}}$ and $\text{GSC}_{\text{LMA-XM}}^{\text{AP-EVD}}$) and (5.36) for $\hat{\mathbf{v}}_{\mathbf{e},i}$ (in the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$). For the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$, $\underset{\sim}{\mathbf{w}}$ was computed using (5.61) with $\underset{\sim}{\mathbf{R}}_{\mathbf{yy}}$ and $\underset{\sim}{\mathbf{R}}_{\mathbf{nn}}$ estimated per frame using the same forgetting factor of 1 s. Since the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-EVD}}$ is the only algorithm that requires a strict definition of its fixed beamformer, i.e. $\widetilde{\mathbf{b}}_{\mathbf{a}}$ (see (5.26)), the fixed beamformers of all other algorithms were also set to $\widetilde{\mathbf{b}}_{\mathbf{a}}$, i.e. $\tilde{\mathbf{f}}_{\mathbf{a}} = \widetilde{\mathbf{b}}_{\mathbf{a}}$ so as to provide an unambiguous comparison between the methods. Finally, despite the input signals having a duration of 60 s, only the latter 30 s of the processed signals were used for evaluation to avoid any transient behaviour that may have resulted from the convergence of the covariance estimates in computing the filters.

Figure 5.8 compiles the results from this experiment when a perfect VAD and an imperfect VAD was used. Each of the three sub-plots within the particular metric uses a different XM as indicated. Each group of 2 bars for a particular algorithm represents the processing as performed with either $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ or $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$ as the a priori RTF vector for the LMA. The x-axis indicates the unweighted input SNR of LM1 at which all of the algorithms were evaluated.

Focusing on the left-hand plots of Fig. 5.8 for the perfect VAD, in terms of Δ SI-SNR, it is clear that all of the algorithms which use the XM have an improved performance over that of the $\text{GSC}_{\text{LMA}}^{\text{AP}}$ as well as the XM for all input SNRs. The $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ also seems to offer a better performance than the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-CC}}$ or $\text{GSC}_{\text{LMA-XM}}^{\text{AP-EVD}}$ algorithms. In terms of Δ-STOI, it is once again evident that the algorithms which use the XM have an improved performance over that of the $\text{GSC}_{\text{LMA}}^{\text{AP}}$ as well as the XM for all input SNRs. Among these algorithms which use the XM, however, the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-CC}}$ and $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ have similar intelligibility improvements and are both mostly better than the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-EVD}}$. It is not surprising, however that the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-EVD}}$ exhibits a poorer performance, as it is known that estimating RTFs from a subtraction of correlation matrices is prone to error at low input SNRs [47] [39]. It is also noted at the highest input SNR that the Δ-STOI values are generally smaller as the input signal would have already been fairly intelligible. In general, it can also be observed that the best improvements in terms of both metrics is obtained when XM1 is used, which is close to the source and hence subject to a high SNR. Additionally, improvements are also evident even in cases where the SNR of the XM is quite low as in the case of XM2. In such a case the XMs would have the potential to act as a noise references and hence improve performance. Finally, using either $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ or $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$ did not demonstrate any considerable differences in performance, but using $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ was in most cases slightly preferred to $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$.

Focusing on the right-hand plots of Figure 5.8 where an imperfect VAD was now used, as expected, the absolute values of the metrics have decreased. However, for both metrics the algorithms using an XM still demonstrate an improvement over the $\text{GSC}_{\text{LMA}}^{\text{AP}}$ as well as the XM except for lower input SNRs when XM1 is used. At an input SNR of -5 dB for instance, the Δ SI-SNR is better than all the other algorithms, however the Δ-STOI indicates that the intelligibility is still on the order of using the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$. Among the algorithms that use the XM, in terms of Δ SI-SNR, the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ performs better than the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-CC}}$ or $\text{GSC}_{\text{LMA-XM}}^{\text{AP-EVD}}$ particularly at higher input SNRs and when the SNR of the XM is higher, for instance for the case of XM1 or XM3. Apart from such conditions, the differences between the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-CC}}$ and $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ are less pronounced and better than the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-EVD}}$. In terms of Δ STOI, the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-CC}}$ and $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ are again similar and better than the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-EVD}}$. The difference in performance from using either $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ or $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$ is now more evident, more so in terms of the intelligibility, where $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ would be preferred to $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$.

From these results of Fig. 5.8, the concluding point would be that the proposed algorithms that use the XM can indeed offer an improved intelligibility as opposed to using only a $\text{GSC}_{\text{LMA}}^{\text{AP}}$ or an XM alone despite imperfect estimation of the relevant correlation matrices. Furthermore, there is the most benefit to be gained from $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ algorithm as it performs either equally well or better than the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-CC}}$ or $\text{GSC}_{\text{LMA-XM}}^{\text{AP-EVD}}$ algorithms depending on the acoustic conditions.

The processed audio files from all of the algorithms, the reference, and the XM signals can be listened to for a personal subjective evaluation at [249].

Multiple XM combinations

In this section, the full scenario of Fig. 5.7 was now considered with the single speech source and the four noise sources active. The noise signals, `bb1`, `bb2`, `cc1`, and `sn1`, were used respectively for NS1, NS2, NS3, and NS4. The sum of these noises were scaled such that the unweighted input SNR at LM1 was 0 dB. Using the optimal filters, the $\text{GSC}_{\text{LMA}}^{\text{AP}}$, $\text{GSC}_{\text{LMA-XM}}^{\text{AP-CC}}$, $\text{GSC}_{\text{LMA-XM}}^{\text{AP-EVD}}$, and $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ algorithms were once again evaluated, however using all the possible permutations from the set of XM signals available. Hence, for the four XMs that were available, there were 15 possible XM combinations to choose from for use with the LMA.

Figure 5.9 displays the results of this experiment when using either $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ or $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$ as the a priori RTF vector for the LMA, as well as when using a perfect VAD. The x-axis indicates which set of XMs was used and is grouped by the number of XMs used in the respective algorithm. So for instance, the first four points

indicate that only one XM was used, while the second set of six points indicates that two XMs were used. The numbers indicated in these groupings correspond to the XM positions as depicted in Fig. 5.7. The corresponding Δ SI-SNR and Δ STOI metrics for the individual XMs used are also displayed in Table 5.1.

Table 5.1: Corresponding metrics for the individual XMs used for the algorithms corresponding to Fig. 5.9 and Fig. 5.10.

	XM1	XM2	XM3	XM4
Δ SI-SNR	7.1	−1.8	1.6	3.2
Δ STOI	0.12	−0.2	−0.02	0.05

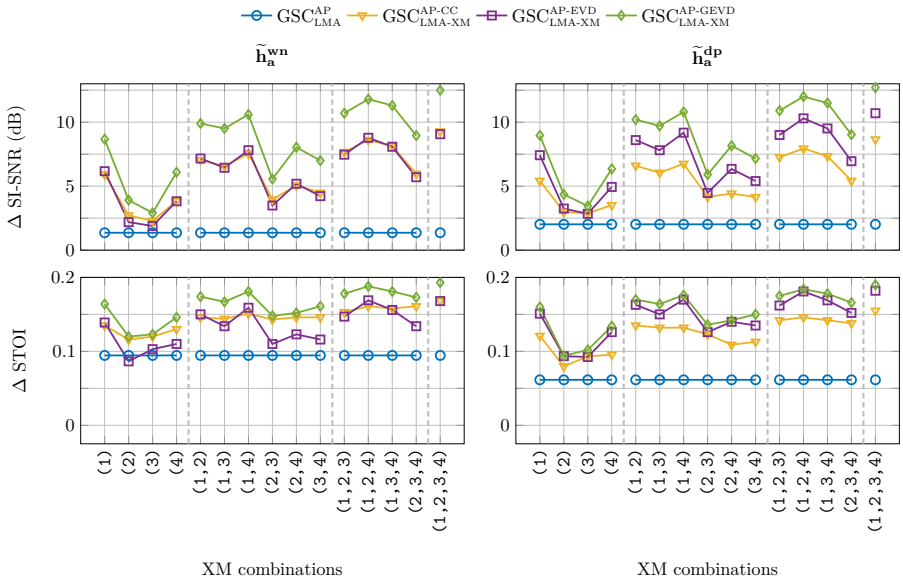


Figure 5.9: Objective metrics from the acoustic scenario of Fig. 5.7 with one speech source and four noise sources, as a function of various combinations of XMs when using a perfect VAD.

Focusing on the left-hand plot of Fig. 5.9 when $\widetilde{\mathbf{h}}^{\mathbf{wn}}_{\mathbf{a}}$ was used, it can be observed that the GSC$^{\text{AP-GEVD}}_{\text{LMA-XM}}$ outperforms the GSC$^{\text{AP}}_{\text{LMA}}$, GSC$^{\text{AP-CC}}_{\text{LMA-XM}}$, GSC$^{\text{AP-EVD}}_{\text{LMA-XM}}$ as well as any of the XMs (upon comparison with the metrics in Table 5.1). In general, a greater improvement can be achieved with the addition of more XMs. The relative position of the XM to the speech source is also observed to have an influence, where XMs closer to the speech source are more beneficial. This can

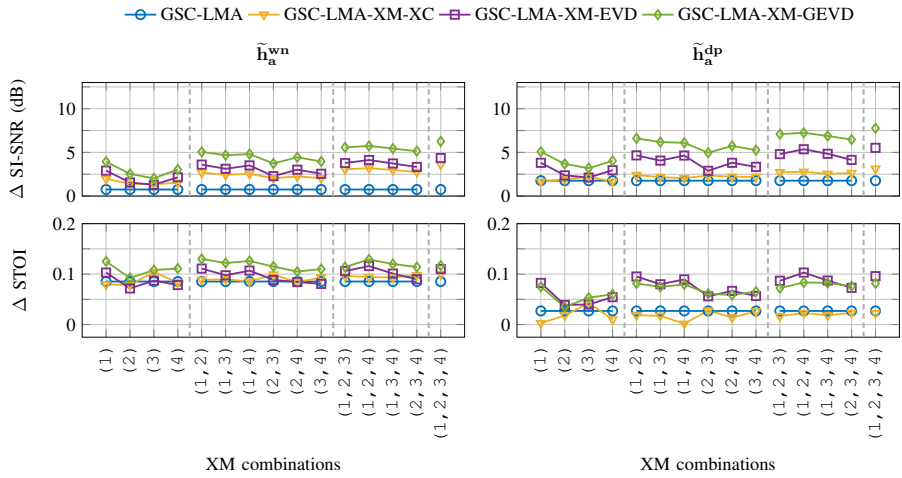


Figure 5.10: Objective metrics from the acoustic scenario of Fig. 5.7 with one speech source and four noise sources, as a function of various combinations of XMs when using an imperfect VAD.

be seen by the increase in the metrics whenever XM1 is included as one of the XMs in the respective algorithm. Such results are indeed consistent with previous findings such as in [222]. Focusing now on the right-hand plot of Fig. 5.9 when $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$ was used, the performance of the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ is quite similar to that when using $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ and the trend is maintianed in most cases where it outperforms the $\text{GSC}_{\text{LMA}}^{\text{AP}}$, $\text{GSC}_{\text{LMA-XM}}^{\text{AP-CC}}$, $\text{GSC}_{\text{LMA-XM}}^{\text{AP-EVD}}$ as well as any of the XMs. In terms of Δ STOI, however, the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ and $\text{GSC}_{\text{LMA-XM}}^{\text{AP-EVD}}$ demonstrate a similar performance. In general, the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-CC}}$ and $\text{GSC}_{\text{LMA-XM}}^{\text{AP-EVD}}$ algorithms appear to be more sensitive than the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ to the choice of either $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ or $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$ as the a priori RTF vector for the LMA.

Figure 5.10 now displays the results of the experiment when using either $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ or $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$ as the a priori RTF vector for the LMA, but when using an imperfect VAD. Focusing on the left-hand plot of Fig. 5.10 when $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ was used, the reduction in the absolute performance can immediately be seen due to the misclassification of frames where speech was present. Nevertheless, the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ algorithm maintains its trend of outperforming the $\text{GSC}_{\text{LMA}}^{\text{AP}}$, $\text{GSC}_{\text{LMA-XM}}^{\text{AP-CC}}$, $\text{GSC}_{\text{LMA-XM}}^{\text{AP-EVD}}$. In terms of Δ SI-SNR, however, XM1 now offers a better performance (in cases when it is used), but the intelligibility improvement as indicated by its Δ STOI is still on the order of that of the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ algorithm. On the right-hand plot of Fig. 5.10 when $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$ was used, it can be observed that the Δ SI-SNR maintains a

similar trend for all algorithms. In terms of intelligibility improvement however, all algorithms have a reduced performance, with $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ and $\text{GSC}_{\text{LMA-XM}}^{\text{AP-EVD}}$ demonstrating a similar improvement (similar to the right-hand plot of Fig. 5.9). Furthermore, in cases where XM1 is used, this XM may offer some intelligibility improvement over the other algorithms. However, the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-EVD}}$ and $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ still offer intelligibility improvements over the other XMs, for instance in cases when XM2 and XM3 or XM2, XM3, and XM4 are used.

The processed audio files from all of the algorithms, the reference, and the XM signals for this experiment can be listened to for a personal subjective evaluation at [249].

Switching XMs

In this experiment, the impact of instantaneously switching between different XMs over time was investigated. Such a scenario can occur for instance if an XM on a mobile device is being used and then one moves the mobile device from one location to the other. As demonstrated in the previous two sections, the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ has the potential for the best performance among all the algorithms that use the XMs. Hence, a comparison will be considered only for this algorithm will along with the $\text{GSC}_{\text{LMA}}^{\text{AP}}$ (using either $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ or $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$), as well as the XMs in this section.

The scenario as depicted in Fig. 5.11 was considered. The speech and noise signals used were identical to that of section 5.6.1, however only two XMs were used along with the LMA. Initially both XM1 and XM4 were used. After 20 s, XM4 was switched to XM3 as indicated by the arrow XM-SWITCH-1, and after 40 s, XM1 was switched to XM2 as indicated by the arrow XM-SWITCH-2.

For both the $\text{GSC}_{\text{LMA}}^{\text{AP}}$ and the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$, an NLMS procedure was used to compute the adaptive filters $\mathbf{v_a}$ and $\mathbf{v}_{\mathbf{e},i}$. As a comparison was not done with the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-EVD}}$, the fixed beamformer was now set such that $\tilde{\mathbf{f}}_{\mathbf{a}} = \widetilde{\mathbf{h}}_{\mathbf{a}}/||\widetilde{\mathbf{h}}_{\mathbf{a}}||^2$, which is generally more of a common choice as it is a matched filter. The difference between this definition and $\widetilde{\mathbf{b}}_{\mathbf{a}}$ is simply a scaling and does not affect the relative comparison of the various algorithms. With the input SNR at the LM1 scaled to 0 dB, the metrics of Δ SI-SNR and Δ STOI were computed over time in 3 s frames with a 50 % overlap.

Figure 5.12 displays the results of this experiment when using a perfect VAD (left) and an imperfect VAD (right). The uppermost plot displays the reference speech signal at LM1 with the respective VAD superimposed, while the bottom two plots display the Δ SI-SNR and Δ STOI metrics. Focusing on the left-hand plot of Fig. 5.12, i.e., with the perfect VAD, it can firstly be observed that

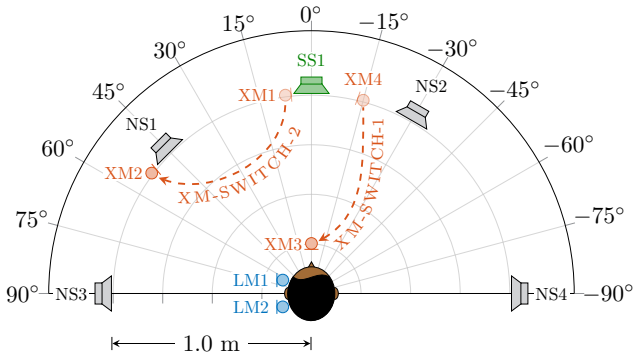


Figure 5.11: Acoustic scenario analagous to that of Fig 5.7, except with two XMs. After 20 s, XM4 was switched to XM3 as indicated by the arrow XM-SWITCH-1, and after 40 s, XM1 was switched to XM1 as indicated by the arrow XM-SWITCH-2.

at the points of switching, both of the XMs transition from having a higher SI-SNR and intelligibility improvement to a lower SI-SNR and intelligibility improvement. This transition obviously does not affect the $\text{GSC}_{\text{LMA}}^{\text{AP}}$, whose SNR and intelligibility improvement remain relatively constant over time. It is also clear that the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ results in the best performance regardless of which XMs are used for processing. Additionally, it can also be seen that using XM1 results in the most improvement (in line with the previous results) since at 40 s, when XM1 switches to XM3, the absolute values of the metrics for the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ are reduced. In terms of the different a priori RTF vectors for the LMA, there was no significant difference between using either $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ or $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$.

Focusing now on the right-hand plot of Fig. 5.12, i.e., with the imperfect VAD, the missclassification of the time periods for which speech was present can be observed from the uppermost plot. For the $\text{GSC}_{\text{LMA}}^{\text{AP}}$ and $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ that use $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$, it can be seen that the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ suffers a reduction in performance in both SI-SNR and STOI improvement, but still performs better than the $\text{GSC}_{\text{LMA}}^{\text{AP}}$ in most cases. During the first 40 s, the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ also exhibits a performance on the order of XM1, being slightly better or slightly worse at times. However, when any of the other XMs are used, the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ maintains a better performance. When using $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$, the relative trends among the algorithms and the XMs are maintained in comparison with using $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$, except that the absolute intelligibility has been reduced.

The processed audio files from all of the algorithms, the reference, and the XM signals for this experiment can be listened to at [249].

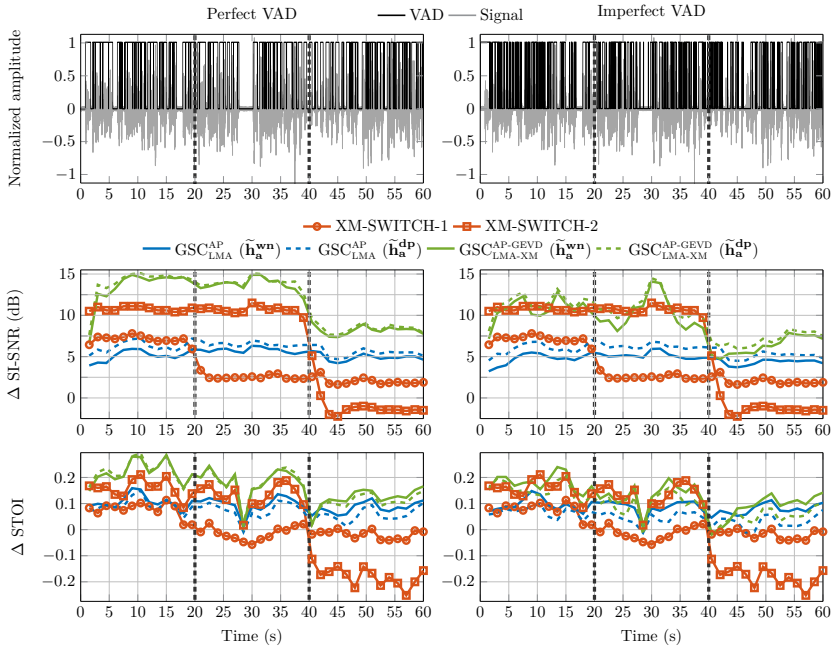


Figure 5.12: Performance of the algorithms when switching between different XMs over time using a perfect VAD (left) and an imperfect VAD (right). The uppermost plots display the clean speech reference signal in LM1 with the respective VAD superimposed. The middle plots show the Δ SI-SNR metric and the final row of plots show the Δ STOI metric. Both Δ SI-SNR and Δ STOI metrics were computed with 3 s frames with a 50 % overlap. The markers for the XMs correspond to the middle of these time frames. Markers were not used for the other plots for clarity. The vertical dotted line at 20 s indicate the switching of XMs from XM-SWITCH-1 and the vertical dotted line at 40 s indicate the switching of XMs from XM-SWITCH-2.

5.6.2 Performance in a public environment

For a more realistic evaluation of the $\text{GSC}^{\text{AP-GEVD}}_{\text{LMA-XM}}$, measurements were carried out in a public space, namely a cafe in the city centre of Leuven, Belgium, involving two persons engaging in conversation as shown in Fig. 5.13. One of the persons (listener) was equipped with the same BTE HAs as previously described (one for each ear) and the other person (target speaker) sat directly in front of the listener (0°) approximately 1 m away, which would be designated as the direction for assigning the a priori RTF vector for the BTE HAs. One XM, an omni-directional AKG CK32 microphone, was also used in the measurements and placed on a table closer to the target speaker. The BTE HA microphones (two for the left ear and two for the right ear) were connected to a proprietary pre-amplifier with an equivalent gain for all of the BTE HAs, after which they were passed to an RME-UCX interface connected to a laptop for signal acquisition. The XM was also connected directly to the RME-UCX amplifier for signal acquisition. Throughout the conversation, there were several types of noises, including persons talking at other tables and the noise of coffee machines, all of which were quite non-stationary. The recordings can be heard in [95].

Prior to recording the noisy signals in the cafe, however, it was necessary to obtain the a priori RTF vector for the BTE HAs. As opposed to the procedures described in the previous evaluation scenario, an impulse response measurement was performed in the KU Leuven ESAT-STADIUS audio laboratory. As depicted in Fig. 5.14 and Fig. 5.15, the same listener from Fig. 5.13 wore the BTE HAs and sat in the middle of the laboratory with a loudspeaker located at 0° (i.e. directly in front of the listener, corresponding to the scenario in the cafe from Fig. 5.13) and approximately and 1 m from the listener. The impulse responses from the loudspeaker to each of the BTE HA microphones were then computed from an exponential sine-sweep measurement [89].

Fig. 5.16(a) is a plot of the resulting impulse response corresponding to one of the BTE HA microphones at a sampling frequency of 16 kHz, demonstrating that the laboratory has a very long reverberation time (RT60 = 1.5 sec in fact). A Tukey window of 100 samples and 75 % taper [250] was applied to the first 100 samples of all of the impulse responses and then padded with zeros according to the length of the DFT window to be used in the STFT processing. This was done so as to extract an approximate representation of the direct component of the impulse responses. Fig. 5.16(b) is a plot of the corresponding truncated and smoothed impulse response from that of Fig. 5.16(a). A DFT was then applied to each of these impulse responses, which yielded the a priori ATF vectors per frequency, $\widetilde{\mathbf{a}}_{\mathbf{a}}(\mathrm{k}) = [\tilde{\mathrm{a}}_{\mathrm{a},1}(\mathrm{k})\ \tilde{\mathrm{a}}_{\mathrm{a},2}(\mathrm{k})\ \ldots \tilde{\mathrm{a}}_{\mathrm{a},\mathrm{M_a}}(\mathrm{k})]^T$. This ATF vector was then normalised with respect to the reference microphone to obtain the a priori RTF vector in each frequency bin, $\widetilde{\mathbf{h}}_{\mathbf{a}}(\mathrm{k}) = [1\ \frac{\tilde{\mathrm{a}}_{\mathrm{a},2}(\mathrm{k})}{\tilde{\mathrm{a}}_{\mathrm{a},1}(\mathrm{k})}\ \ldots \frac{\tilde{\mathrm{a}}_{\mathrm{a},\mathrm{M_a}}(\mathrm{k})}{\tilde{\mathrm{a}}_{\mathrm{a},1}(\mathrm{k})}]^T$. This



Figure 5.13: (Top) Listener wearing the BTE HAs. (Bottom) Conversation between the target speaker and the listener in a noisy cafe. The XM can also be seen on the table next to the target speaker.

a priori RTF vector was then kept the same for each time frame.

One consequence of this measurement was that the listener had to remove the BTE HAs after the laboratory measurement and then put them on again at

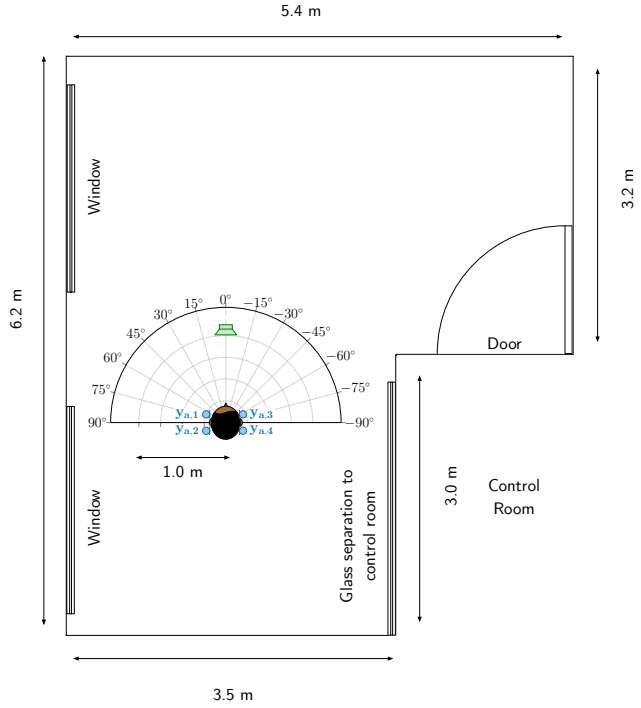


Figure 5.14: Plan view of the KU Leuven ESAT-STADIUS audio laboratory depicting the setup for performing an impulse response measurement from a source located at 0° (directly in front) with respect to the user equipped with the BTE HAs.

a later time in the cafe. This is a somewhat subtle but important conceptual detail of realism for the entire experiment because it means that there is some inherent mismatch between the measured a priori RTF vector for a target speaker at 0° and that of the true RTF vector for a target speaker at 0° while at the cafe.

In the following, only the left-ear BTE HA was used along with the XM, so that $M_a = 2$ and $M_e = 1$. Throughout the recording, the listener faced directly toward the target speaker as in Fig. 5.13 so that the scenario would correspond to a similar one under which the a priori RTF vector was computed. Midway through the recording, however, the target speaker moves the XM closer toward them.

The recorded signals were then processed using the $\mathrm{GSC}^{\mathrm{AP}}_{\mathrm{LMA}}$ and $\mathrm{GSC}^{\mathrm{AP\text{-}GEVD}}_{\mathrm{LMA\text{-}XM}}$



Figure 5.15: Measurement of the impulse response inside the KU Leuven ESAT-STADIUS audio laboratory corresponding to the depiction in Fig. 5.14. The BTE HAs can be seen on the listener's ear.

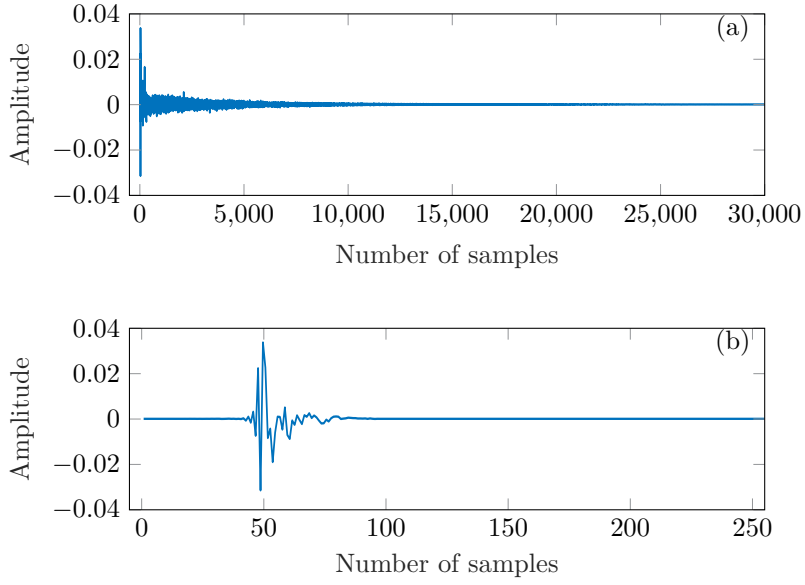


Figure 5.16: (a) Impulse response from the loudspeaker to one of the BTE HA microphones from the measurement scenario of Fig. 5.14. (b) The truncated and smoothed impulse response of 256 samples used for computing the corresponding component in the a priori RTF vector.

algorithms with the WOLA method, a Discrete Fourier Transform (DFT) size of 256 samples, 50 % overlap, a square-root hanning window, and a sampling frequency of 16 kHz. All RTF vector estimates were performed in frames where the speech source was active as indicated by the SPP in the XM. Frames for which the speech was active were chosen if the SPP > 0.85, otherwise the frames were classified as noise. All relevant correlation matrices were also estimated using a forgetting factor corresponding to a time constant of 300 ms in the respective speech plus noise or noise only frames. For the $\text{GSC}_{\text{LMA}}^{\text{AP}}$ and $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$, the adaptive filters, $\mathbf{v_a}$ and $\mathbf{v_{e},}_i$, were computed both with the optimal expressions, i.e. (5.7) for $\hat{\mathbf{v}}_{\mathbf{a}}$ and (5.36) for $\hat{\mathbf{v}}_{\mathbf{e},i}$, as well as with an NLMS procedure with a step size = 0.3. The fixed beamformer was also set such that $\tilde{\mathbf{f}}_{\mathbf{a}} = \widetilde{\mathbf{h}}_{\mathbf{a}}/||\widetilde{\mathbf{h}}_{\mathbf{a}}||^2$.

Fig. 5.17 show the resulting spectrograms for a 20 sec segment of the noisy signals. At the point of 10 sec is where the target speaker moves the XM closer to them, which can be observed from the spectrogram of the XM in Fig. 5.17(b). While the $\text{GSC}_{\text{LMA}}^{\text{AP}}$ in Figs. 5.17(c) and 5.17(d) offers some improvement upon the noisy BTE HA microphone, it is quite evident that a large performance improvement can be achieved by using the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$. When the XM is closer to the target speaker, it can also be observed that the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ has an improved noise reduction. The spectrograms also suggest that there is not too much difference between the optimal and NLMS computation of the adaptive filters. A video of the scenario in the cafe and the processed audio from various algorithms superimposed is available in [95], and is recommended to be viewed by the reader in order to have a subjective impression of the results.

Finally, it can perhaps be argued that the degree of noise reduction achieved by the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ could be too severe for hearing aid users as the spatial impression of interfering noises is fairly suppressed, although it could potentially be beneficial for cochlear implant users, but this is left to be further evaluated. Nevertheless, in order to retain some more spatial impression of the interfering sources with a compromise of reducing less noise, there are a few practical options. For instance, a convex combination can be made with the enhanced output of the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ and that of the $\text{GSC}_{\text{LMA}}^{\text{AP}}$, or even the noisy BTE HA as is sometimes done within the context of binaural hearing aid speech enhancement [251]. Another option to consider would be to use an alternative estimate for $\mathbf{R_{nn}}$ that is a fixed diffuse spatial coherence matrix scaled with some estimated noise PSD. In this case it will be diffuse noise that will be suppressed and not any interfering sources. In [252], an RTF vector estimate was in fact proposed by assuming a diffuse noise field. These are all worthwhile areas of exploration for the use of the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ in practice and would require a more detailed evaluation for the particular application at hand. Recent work in [253] has also investigated how the perception of an enhanced speech signal is

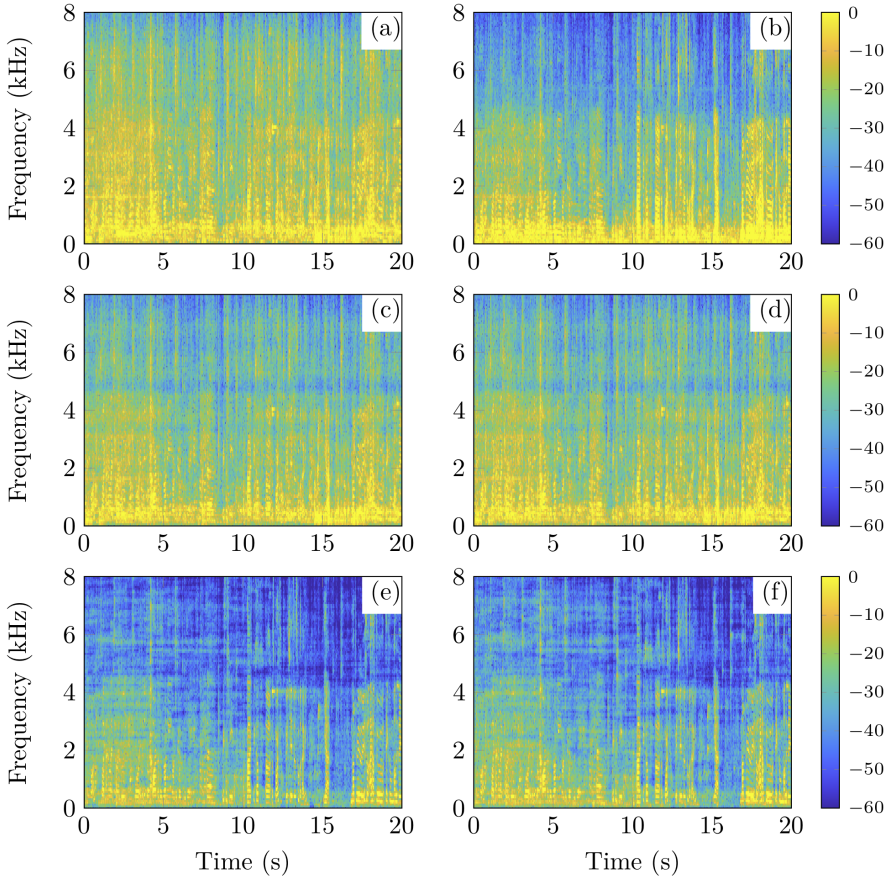


Figure 5.17: Spectrograms of the (a) front left BTE HA signal (reference microphone), (b) XM signal, (c) $\text{GSC}_{\text{LMA}}^{\text{AP}}$ with the optimal adaptive filters, (d) $\text{GSC}_{\text{LMA}}^{\text{AP}}$ with the NLMS computed adaptive filters, (e) $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ with the optimal adaptive filters, (f) $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ with the NLMS computed adaptive filters. At 10 sec the target speaker moves the XM closer to them. The colour bar indicates the intensity of the STFT coefficients (in dB), which has been constrained to the range as shown for visual clarity (since the intensity of the XM STFT coefficients were substantially greater than those on the reference signal).

affected when incorporating an XM, with the conclusion that there is a trade-off between the quality and spatial attributes of the perceived enhanced speech signal.

5.6.3 Comparison to a fully data-dependent RTF vector

Although the advantages of incorporating a priori spatial knowledge into an MVDR beamformer with an LMA and XMs have been demonstrated, it is still unclear as to how this compares to the case when a fully data-dependent method is used, i.e., when the RTF vector is estimated without the use of any a priori spatial knowledge. In this section, these two strategies are compared to better understand the benefits of each approach.

For this evaluation audio recordings of speech and noise were used from a dataset created at the KU Leuven ESAT-STADIUS audio laboratory for the setup depicted in Fig. 5.18. A Neumann KU-100 dummy head was placed in a central location of the room and equipped with the same two BTE HAs as described in the previous sections. The first microphone of the left ear HA was used as the reference microphone. Three omnidirectional XMs (XM1 was an AKG CK97O microphone, while XM2 and XM3 were AKG CK32 microphones) were placed at heights of 1 m from the floor and at varying distances from the dummy head as shown in Fig. 5.18. A Genelec 8030C loudspeaker, was placed at 1 m and different azimuth angles from the dummy head to generate a speech signal from male and female speakers [94]. The loudspeaker and the dummy head were placed at a height of approximately 1.3 m from the floor. For the noise, a cocktail party scenario was re-created. With the same configuration of the dummy head and external microphones from Fig. 5.18, persons were invited to stand outside of a 1 m circumference from the dummy head in a random manner (i.e. all participants were not confined to a particular corner in the room). Beverages in glasses as well as snacks were served while the participants engaged in conversation. There were nine male speakers and six female speakers present in the room. A recording of such a scenario was made for approximately one hour.

The noisy microphone signals used in the following consisted of an 8 s sample of the female speech mixed with a random sample of the cocktail party noise. For the STFT processing, the WOLA method, with a DFT size of 256 samples, 50 % overlap, a square-root hanning window, and a sampling frequency of 16 kHz were used. By using the SPP [192] computed on one of the microphone signals (will be specified in the experiments), frames were classified as containing speech if the SPP > 0.8, otherwise the frames were classified as noise only. All RTF vector estimates were performed in frames which were classified as containing speech. All the relevant correlation matrices were also estimated using a forgetting factor corresponding to a time constant of 500 ms. $\mathbf{R_{nn}}$ was only estimated when the SPP < 0.8.

The a priori RTF vector pertaining to the HAs was defined in accordance with a

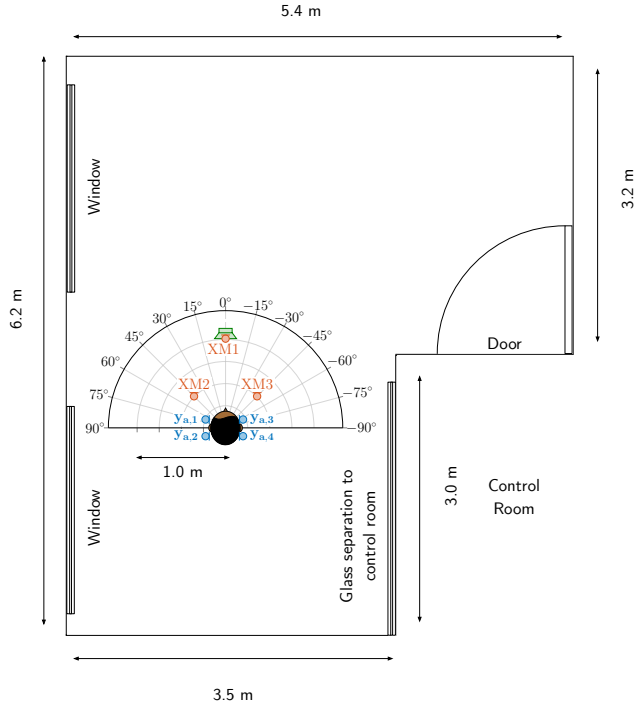


Figure 5.18: Plan view of the KU Leuven ESAT-STADIUS audio laboratory depicting the setup used in recording the data for comparing the use of an a priori RTF vector with a data-dependent RTF vector.

source located at 0° (the azimuth direction directly in front of the dummy head) and 1 m from the dummy head. Impulse response measurements were carried out as described in sec. 5.6.2 and used to obtain $\widetilde{\mathbf{h}}_{\mathbf{a}}$. Using a fixed beamformer where $\tilde{\mathbf{f}}_{\mathbf{a}} = \widetilde{\mathbf{h}}_{\mathbf{a}}/||\widetilde{\mathbf{h}}_{\mathbf{a}}||^2$, $\hat{\widetilde{\mathbf{h}}}_{\mathbf{e,gevd}}$ was computed as in (5.57) and used as the estimate of $\mathbf{h_e}$ in (5.9). This a priori RTF vector was then subsequently used to compute the $\text{MVDR}^{\text{AP}}_{\text{LMA-XM}}$ from (5.10) (which would then be equivalent to the $\text{GSC}^{\text{AP-GEVD}}_{\text{LMA-XM}}$ when $\hat{\mathbf{R}}_{\mathbf{nn}}$ is explicitly used as opposed to adaptive filters). For the fully data-dependent RTF vector, the procedure as outlined in subsection 3.4.2 from Chapter 3 was used, i.e. a GEVD of the matrix pencil $(\hat{\mathbf{R}}_{\mathbf{yy}}, \hat{\mathbf{R}}_{\mathbf{nn}})$, where these matrices are of size $(\text{M}_\text{a} + \text{M}_\text{e}) \times (\text{M}_\text{a} + \text{M}_\text{e})$, whereas in Chapter 3, they were of size $\text{M}_\text{a} \times \text{M}_\text{a}$. With this data-dependent RTF vector, then the MVDR beamformer was applied and will be referred to as $\text{MVDR}^{\text{DD}}_{\text{LMA-XM}}$. The $\text{MVDR}^{\text{AP}}_{\text{LMA}}$ from (5.2) was also included in the evaluation.

The $\text{MVDR}^{\text{AP}}_{\text{LMA-XM}}$ and $\text{MVDR}^{\text{DD}}_{\text{LMA-XM}}$ beamformers were updated accordingly

and applied to the noisy microphone signals at each frequency bin and time frame. The last 5 s of the resulting processed signal in the time domain was used in computing the evaluation metrics of Δ SI-SNR, Δ-STOI, and Δ-fwSNRseg, where the relative changes are with respect to the metrics as computed on the frontal-left-ear microphone, $y_{a,1}$ (see Fig. 5.18). One detail to be highlighted in the computation of the Δ-STOI and Δ-fwSNRseg metrics is the choice of a reference signal. Usually, the speech-only signal in the reference microphone can be used as the reference signal. However, as the laboratory where the recordings were made has a reverberation time of 1.5 s, such as signal was not suitable as a reference. This in fact reveals a fundamental problem on being able to compare the $\text{MVDR}^{\text{AP}}_{\text{LMA-XM}}$ and the $\text{MVDR}^{\text{DD}}_{\text{LMA-XM}}$ (with the STOI and fwSNRseg metrics) as it is dependent on a reference signal, which is subjective. Nevertheless, based on the discussion in Chapter 2 regarding the signal model interpretation, the reference signal used was chosen to be the unprocessed desired speech signal convolved with 256 samples (i.e, same length as the DFT size) of the pre-measured impulse response from the desired speech signal location to the reference microphone. This should be kept in mind when interpreting the results that follow as the relative performance of $\text{MVDR}^{\text{AP}}_{\text{LMA-XM}}$ and the $\text{MVDR}^{\text{DD}}_{\text{LMA-XM}}$ in terms of Δ-STOI and Δ-fwSNRseg are with respect to this reference. The Δ SI-SNR metric on the other hand does not depend on a reference and can be interpreted as is.

In the following, results of the evaluation are presented for the variation of four different criteria that can affect the performance of the algorithms: (i) the input SNR (unweighted) at the reference microphone, (ii) the microphone used in computing the SPP, (iii) the number of samples of communication delay from an XM signal, and (iv) the location of the desired source. Furthermore, two cases were considered: the first when $\text{M}_\text{a} = 2$ so that LMA consisted of only the two microphones on the left-ear, and also when $\text{M}_\text{a} = 4$ so that the LMA consisted of all of the BTE-HA microphones. Table 5.2 summarises the values used for the other criteria (control variables) when one of the performance affecting criteria is varied. For the SPP microphone, $y_{a,1}$ means that the noisy reference microphone was used, where as $x_{a,1}$ means that the speech-only component of the noisy microphone was used. Similarly for the input SNR criterion, the speech-only component of XM1 was used to compute the SPP. These speech-only microphone signals were used to reduce any performance reductions due to misclassification, and hence isolate the effect of that particular criterion. More details are given in the respective experiment. All of the codes and audio samples for this section are provided in [95].

Table 5.2: Summary of the experimental conditions used in evaluating the four different performance affecting criteria.

Criterion to vary	Input SNR (dB)	SPP mic.	XM sample delay	Source Location
Input SNR	Varied	XM1 (speech-only)	0	0°
SPP mic.	2	Varied	0	0°
XM sample delay	2	$x_{a,1}$	Varied	0°
Source Location	2	$y_{a,1}$	0	Varied

Input SNR

Fig. 5.19 displays the performance of the beamformers as function of the unweighted input SNR. In this experiment all three XMs as depicted in Fig. 5.18 were used in processing. It is immediately clear that the $\text{MVDR}^{\text{AP}}_{\text{LMA-XM}}$ consistently performs better than the $\text{MVDR}^{\text{AP}}_{\text{LMA}}$, highlighting once again the benefit of introducing the XMs to an existing LMA. Additionally, it can also be observed that the $\text{MVDR}^{\text{AP}}_{\text{LMA-XM}}$ performs better than the $\text{MVDR}^{\text{DD}}_{\text{LMA-XM}}$ in all cases, which supports the claim that there are advantages to explicitly using a priori spatial knowledge in adverse acoustic conditions. For $M_a = 2$, the $\text{MVDR}^{\text{DD}}_{\text{LMA-XM}}$ performs better than the $\text{MVDR}^{\text{AP}}_{\text{LMA}}$ in most cases, however, the distinction is not as clear for $M_a = 4$. Even in some cases, it can be observed that using the LMA alone with a priori spatial knowledge has a performance on the order of that obtained by relying on a fully-data-dependent $\text{MVDR}_{\text{LMA-XM}}$ beamformer.

SPP microphone

In this experiment, the performance of the beamformers was observed as a function of the microphone signal used for computing the SPP. Either the noisy signal of $y_{a,1}$ or XM1 was used for this purpose. As the SPP was used for accurately classifying periods of speech activity, it was needed for the estimation of $\mathbf{R_{nn}}$ as well as for estimating the missing part of the RTF vector for the $\text{MVDR}^{\text{AP}}_{\text{LMA-XM}}$, and the entire RTF vector for the $\text{MVDR}^{\text{DD}}_{\text{LMA-XM}}$. Therefore, it would be expected that different estimates of the SPP (resulting from the computation on different noisy signals) would affect the performance of the beamformers.

Fig. 5.20 displays the results for $M_a = 2$ and $M_a = 4$. In both cases, all three XMs were used in the processing. For both the $\text{MVDR}^{\text{AP}}_{\text{LMA-XM}}$ and the $\text{MVDR}^{\text{DD}}_{\text{LMA-XM}}$, the performance is observed to improve when XM1 was used as

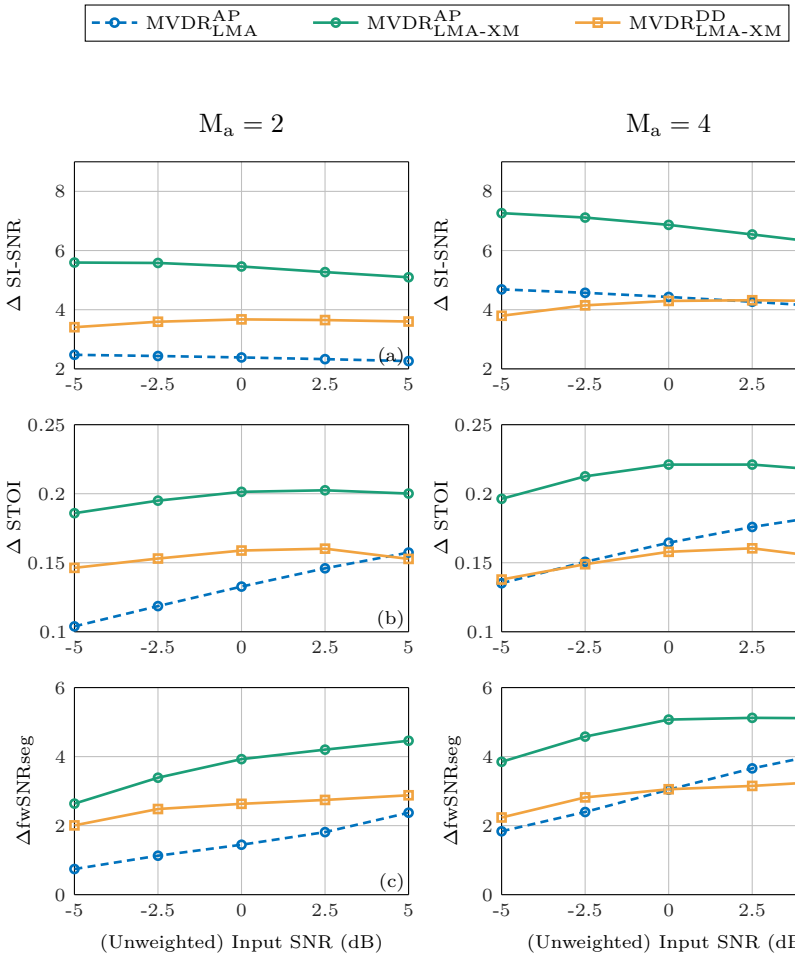


Figure 5.19: Effect of unweighted input SNR on the performance of the $\text{MVDR}^{\text{AP}}_{\text{LMA}}$, $\text{MVDR}^{\text{AP}}_{\text{LMA-XM}}$, and $\text{MVDR}^{\text{DD}}_{\text{LMA-XM}}$. The left column (a)-(c) corresponds to using an LMA with $M_a = 2$ and the right column (d)-(f) corresponds to using an LMA with $M_a = 4$.

opposed to $y_{a,1}$ for computing the SPP. This was expected as a better estimate of the SPP would have been obtained with XM1 as it was closer to the desired source than $y_{a,1}$. The $\text{MVDR}^{\text{AP}}_{\text{LMA-XM}}$ is also seen in general to have a better performance than the $\text{MVDR}^{\text{DD}}_{\text{LMA-XM}}$.

With respect to the $\text{MVDR}^{\text{AP}}_{\text{LMA}}$, its performance remains relatively unchanged

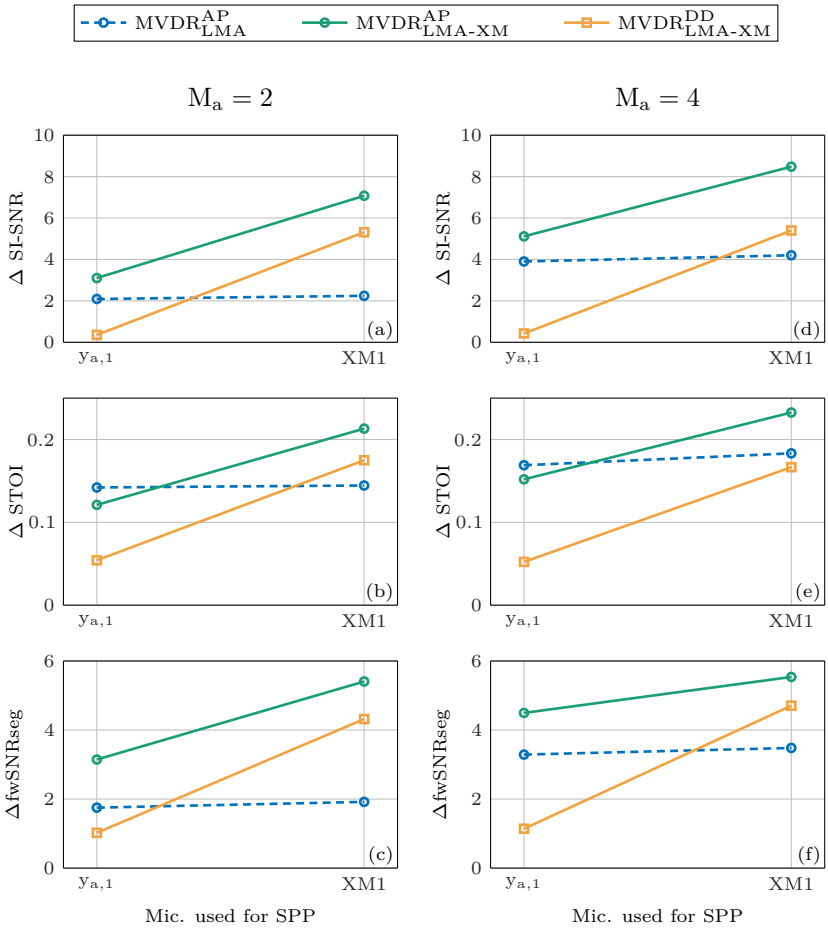


Figure 5.20: Effect of the microphone used to compute the SPP on the performance of the MVDR$^{AP}_{LMA}$, MVDR$^{AP}_{LMA-XM}$, and MVDR$^{DD}_{LMA-XM}$. The left column (a)-(c) corresponds to using an LMA with $M_a = 2$ and the right column (d)-(f) corresponds to using an LMA with $M_a = 4$.

regardless of the microphone used for computing the SPP as its RTF vector is completely known a priori and not dependent on the data. It is also observed that when $y_{a,1}$ is used to compute the SPP, the performance of the MVDR$^{AP}_{LMA}$ is better than that of the MVDR$^{DD}_{LMA-XM}$ and in terms of Δ-STOI slightly better than the MVDR$^{AP}_{LMA-XM}$.

This in fact highlights the problem that it becomes increasingly difficult to

accurately obtain the necessary statistics pertaining to the target speaker in adverse acoustic environments involving interfering speakers and excessive reverberation. Certainly, other methods of computing the SPP can be done which makes use of all of the microphones [5, 254] and may further improve the performance of the $\text{MVDR}^{\text{AP}}_{\text{LMA-XM}}$ and $\text{MVDR}^{\text{DD}}_{\text{LMA-XM}}$ beamformers. However, the point of these results was to demonstrate that the beamformers relying on partial or complete data-dependent RTF estimates can degrade in performance when there is misclassification of the speech and the noise time-frequency segments. In such a case therefore, the use of a priori spatial knowledge then becomes a contingency strategy. In other words, when the $\text{MVDR}^{\text{DD}}_{\text{LMA-XM}}$ beamformer has a poor performance, the $\text{MVDR}^{\text{AP}}_{\text{LMA-XM}}$ can be a better option, and furthermore, even when the $\text{MVDR}^{\text{AP}}_{\text{LMA-XM}}$ has poor performance, then the $\text{MVDR}^{\text{AP}}_{\text{LMA}}$ may be a better option.

Communication delay

As mentioned in Chapter 4, one of the practical issues encountered when using the XMs with an LMA is the communication delay. In this experiment, only XM1 was used in the processing. Each time the algorithms was executed, XM1 was delayed by an increasing number of samples. Therefore only the $\text{MVDR}^{\text{AP}}_{\text{LMA-XM}}$ and $\text{MVDR}^{\text{DD}}_{\text{LMA-XM}}$ beamformers would be affected. Fig. 5.21 shows the resulting performance of the beamformers as a function of this sample delay. In all cases, it can be seen that as the number of samples exceed 128 (half of the DFT window size of 256), the performance of the $\text{MVDR}^{\text{AP}}_{\text{LMA-XM}}$ and $\text{MVDR}^{\text{DD}}_{\text{LMA-XM}}$ beamformers begin to degrade, which is in line with that observed in [222]. For the $\text{MVDR}^{\text{AP}}_{\text{LMA-XM}}$ the lower bound of the performance is that of the $\text{MVDR}^{\text{AP}}_{\text{LMA}}$. It is also seen that the $\text{MVDR}^{\text{AP}}_{\text{LMA-XM}}$ performs better than the $\text{MVDR}^{\text{DD}}_{\text{LMA-XM}}$. For the case when $\text{M}_{\text{a}} = 4$, the $\text{MVDR}^{\text{AP}}_{\text{LMA}}$ is seen to do better than the $\text{MVDR}^{\text{DD}}_{\text{LMA-XM}}$, especially when performance of the $\text{MVDR}^{\text{DD}}_{\text{LMA-XM}}$ begins to degrade due to the sample delay.

Location of desired source

In all the previous experiments, the source location was fixed to 0°. In practice, it can be expected that the source location will move and in such a case a priori spatial knowledge of the source location will become invalid. In this experiment, all three XMs were used and the microphone used for the SPP computation was that of $\text{y}_{\text{a},1}$, i.e., the noisy reference microphone. This microphone was chosen so as to observe the performance under realistic and adverse circumstances. Fig. 5.22 displays the results for $\text{M}_{\text{a}} = 2$ and $\text{M}_{\text{a}} = 4$.

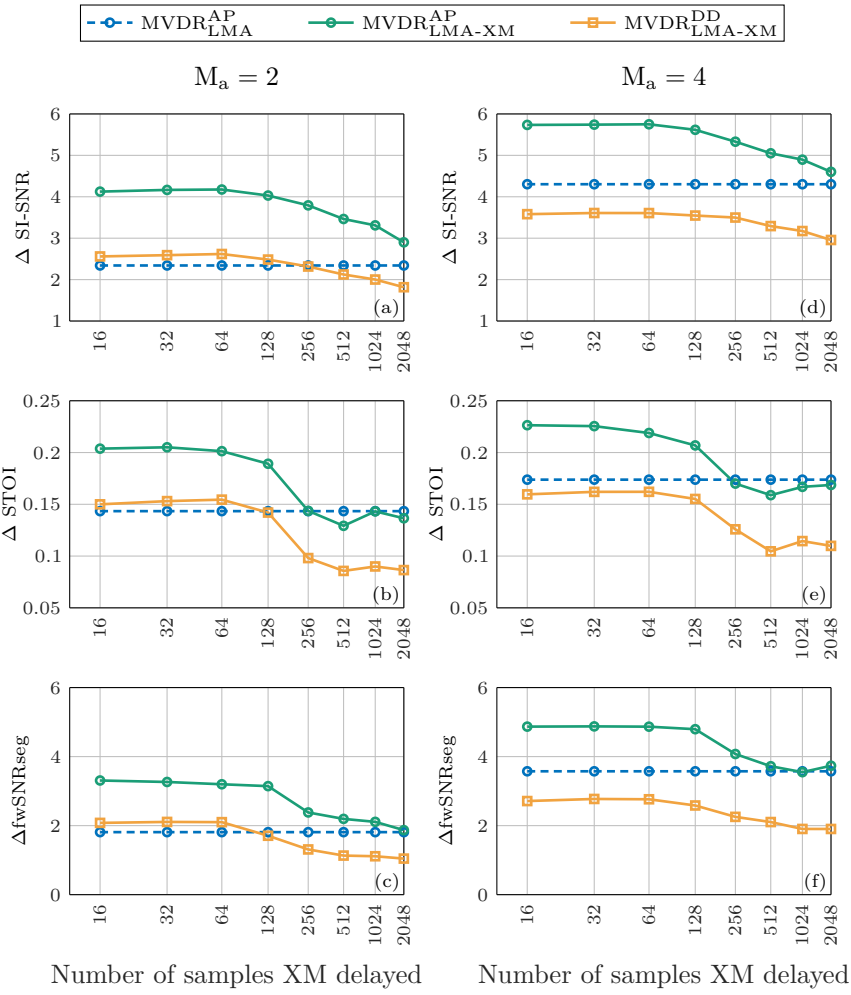


Figure 5.21: Effect of the communication delay of the XM on the performance of the $\text{MVDR}^{\text{AP}}_{\text{LMA}}$, $\text{MVDR}^{\text{AP}}_{\text{LMA-XM}}$, and $\text{MVDR}^{\text{DD}}_{\text{LMA-XM}}$. The left column (a)-(c) corresponds to using an LMA with $M_a = 2$ and the right column (d)-(f) corresponds to using an LMA with $M_a = 4$.

For $M_a = 2$, the $\text{MVDR}^{\text{DD}}_{\text{LMA-XM}}$ has a consistent performance across all angles, although not with as large increases compared to when the SPP was esitmated from XM1. The $\text{MVDR}^{\text{AP}}_{\text{LMA}}$ and $\text{MVDR}^{\text{AP}}_{\text{LMA-XM}}$ on the other hand, have a better performance within 0° and 30°, but then gradually decrease. The metrics also suggest that the performance of both the $\text{MVDR}^{\text{AP}}_{\text{LMA}}$ and $\text{MVDR}^{\text{AP}}_{\text{LMA-XM}}$

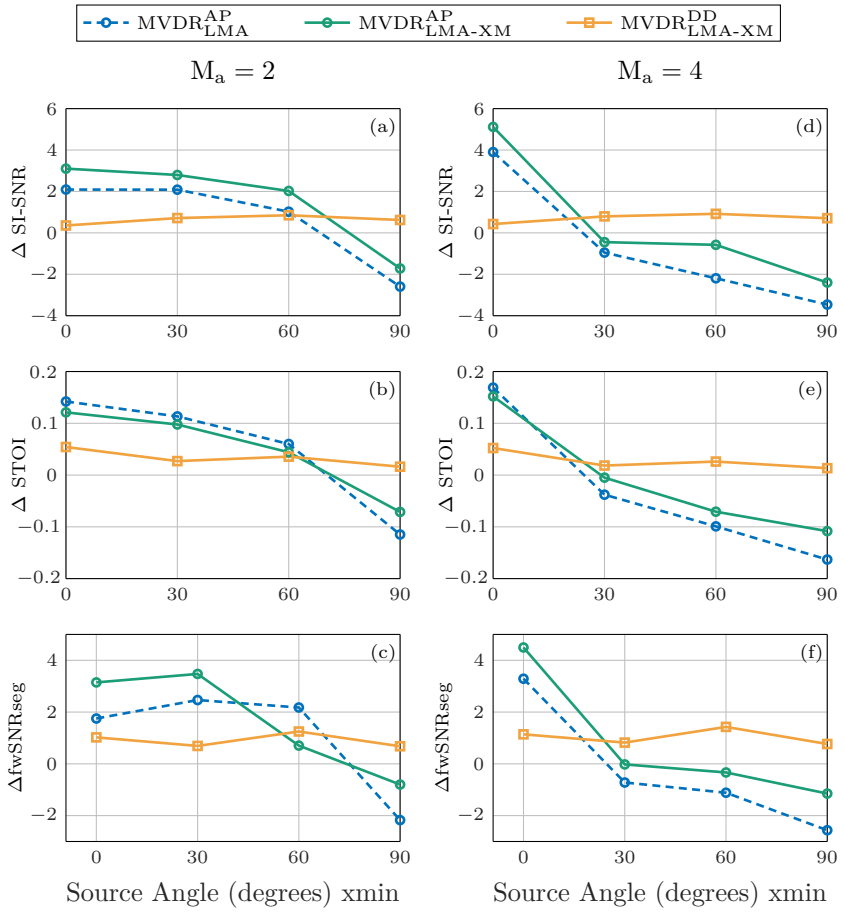


Figure 5.22: Effect of the location of the desired source on the performance of the MVDR$^{AP}_{LMA}$, MVDR$^{AP}_{LMA-XM}$, and MVDR$^{DD}_{LMA-XM}$. The left column (a)-(c) corresponds to using an LMA with $M_a = 2$ and the right column (d)-(f) corresponds to using an LMA with $M_a = 4$.

is fairly similar. For $M_a = 4$, however, the MVDR$^{AP}_{LMA}$ and MVDR$^{AP}_{LMA-XM}$ beamformers only demonstrate a better performance than the MVDR$^{DD}_{LMA-XM}$ beamformer at 0°. Beyond 0°, their performances rapidly degrade, while the MVDR$^{DD}_{LMA-XM}$ beamformer maintains a fairly consistent performance. Once again the MVDR$^{AP}_{LMA}$ and MVDR$^{AP}_{LMA-XM}$ beamformers seem to have a similar performance. The more rapid decline in performance for these beamformers when $M_a = 4$ as compared to $M_a = 2$ could be due to a more directional

response being created as the number of microphones increased. Although a more directional response would be responsible for the better performance at 0° from $\mathrm{M_a} = 4$ compared with $\mathrm{M_a} = 2$, it would also be the reason for an increased sensitivity to source locations outside of 0°. Moreover, these results demonstrate that the use of a priori spatial knowledge, while valuable, can also significantly degrade performance when reality does not match with the a priori spatial knowledge. In such a case, therefore, it is seen that relying on the data becomes beneficial as demonstrated by the $\mathrm{MVDR}^{\mathrm{DD}}_{\mathrm{LMA\text{-}XM}}$, hence the need to merge the benefits of the a priori and data-dependent spatial information.

5.7 Conclusion

This chapter has investigated an alternative to a fully data-dependent approach of designing a speech enhancement system for a microphone configuration involving an LMA and XMs, which involves the explicit use of any available a priori spatial information with respect to the LMA. In such an approach, it is only the relevant parameters related to the XMs that need to be estimated from the microphone data. As different classes of spatial information are used for the different sets of microphones, this approach can also be viewed as an integration of the a priori and data-dependent spatial information from a geometrically-based perspective, which is structurally different from the confidence-based type of integration considered in Chapter 3. This geometrically-based integration approach was developed within the context of a minimum variance distortionless (MVDR) beamformer and its generalised sidelobe canceller (GSC) counterpart, where the a priori spatial information is an a priori RTF vector for the LMA. Using this a priori RTF vector, it was then shown how the complementary RTF vector for the XMs could be estimated from the microphone data.

In particular, three methods of extending an LMA-based GSC with XMs have been proposed following this approach. While all of the methods demonstrated intelligibility improvements over the LMA-based GSC, it was found that the one which yielded the best performance, the $\mathrm{GSC}^{\mathrm{AP\text{-}GEVD}}_{\mathrm{LMA\text{-}XM}}$, involved a procedure where the speech estimate from an LMA-based GSC is directly used along with an orthogonalised version of the XM signals to obtain an improved speech estimate via a generalised eigenvalue decomposition (GEVD). One of the advantages of this method is that the speech estimate from the LMA-based GSC has not been compromised in any way (as illustrated in Fig 5.5) and if so desired, this signal is still available as an option to the listener. In extreme cases of poor data-dependent estimation with the XMs, the $\mathrm{GSC}^{\mathrm{AP\text{-}GEVD}}_{\mathrm{LMA\text{-}XM}}$ offers a contingency option of simply using the LMA-based solution (which was what would be used in the absence of XMs). In fact, a convex combination of the speech estimate

from the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ and the speech estimate from the $\text{GSC}_{\text{LMA}}^{\text{AP}}$ would be an appropriate strategy, where the weightings could be decided manually by the user or automatically with a confidence metric as was done in the confidence-based integration approach. An XM signal itself can also be an option for improved intelligibility, however, the $\text{GSC}_{\text{LMA-XM}}^{\text{AP-GEVD}}$ is still the preferred method as listening to an XM signal only would contribute to intelligibility improvements when it is close to the desired speaker.

A comparison between the proposed geometrically-based integration approach that uses the a priori spatial knowledge for the LMA and a fully data-dependent approach was also carried out in this chapter. It was observed that both methods were capable of delivering performance improvements in terms of SNR and speech intelligibility, but the amount of improvement that could be achieved depended on several factors such as the accuracy of the estimated spatial correlation matrices, the communication delay experienced from an XM, and the accuracy of the a priori spatial knowledge (such as an assumed source location). Depending on the accuracy of the a priori spatial knowledge in particular, a clear trade-off was evident between both approaches, similar to that observed in Chapter 3.

This chapter has therefore now provided a technique which can be used to revisit an integrated MVDR beamformer similar to that of Chapter 3, but for the microphone configuration involving an LMA linked with XMs. This would then correspond to merging both the confidence-based and geometric-based integration approaches into a more global integration of a priori and data-dependent spatial information. Before continuing to this however, in the following chapters, the geometric-based integration approach of using a priori spatial knowledge developed in this chapter is applied to two other speech enhancement tasks: binaural speech enhancement and speech dereverberation.

Chapter 6

Binaural speech enhancement using a priori spatial knowledge and external microphones

6.1 Introduction

In the previous chapter, while the speech enhancement strategy developed demonstrated improvements with respect to noise reduction and speech intelligibility, they did not preserve any spatial auditory perception of the acoustic environment. For instance, by listening to the enhanced audio signal alone from chapter 5, we would not be able to tell where the desired speaker or a particular noise was located in space. To accomplish this, we would alternatively need a binaurally enhanced signal, consisting of a signal to be presented to the left ear and another to the right ear, which encode relevant binaural cues related to the acoustic environment. This chapter[1] therefore investigates how the geometrically-based integration approach from chapter 5 can be extended to obtain a binaurally enhanced signal in order to preserve some aspects of the spatial auditory perception of the acoustic environment. The context where

[1]This chapter is partially based on the publication: Ali R., van Waterschoot T., Moonen M., "Using partial a priori knowledge of relative transfer functions to design an MVDR beamformer for a binaural hearing assistive device with external microphones", in *Proc. of the 23rd Int. Congress on Acoustics (ICA)*, Aachen, Germany, Sep. 2019, pp. 1-8.

this is applicable is for hearing impaired individuals who are equipped with a binaural assistive hearing device, which consists of one assistive hearing device for the left ear and another for the right ear.

Normal hearing individuals are able to localise sounds in a horizontal (azimuth) plane primarily by exploiting differences in the signals received at both ears. Two such differences which play an important role are the *interaural time differences* (ITDs) and the *interaural level differences* (ILDs) [4]. The ITD is defined as the time difference of arrival of a sound from one ear to the other. For instance, a sound generated from the right-hand side of a listener arrives to the right ear first before arriving to the left ear. In addition to this time delay, the sound on the right ear will be louder than that of the left ear, which defines the ILD, i.e. as the difference in the level (amplitude) of the sound from one ear to the other. At low frequencies, the wavelength of sound is larger in comparison to the dimensions of the human head, and hence an acoustic wave propagates unimpeded resulting in smaller ILDs at lower frequencies. At higher frequencies, however, as the wavelength becomes smaller, the human head casts an acoustic shadow on the ear contralateral to the sound source resulting in a lower level for this ear in comparison to the one ipsilateral to the source and hence a larger ILD. On the other hand, at these higher frequencies, the ITDs are less important since it is more likely that these shorter wavelengths arrive to both ears in phase (without any discernible delay). Therefore, the ITDs are more important in the lower frequencies. In general, however, both ITDs and ILDs are also dependent on the location of the source and the size of the listener's head (related to the physical spacing of both ears).

In addition to the ITDs and ILDs, the shape of the pinna and the ear canal also aid with localisation particularly in the elevation plane by acting as a system of acoustical resonators, where specific resonances are excited according to the direction and distance of a sound source [4]. Furthermore, the interaural coherence (IC) [4] is another important binaural cue which as its name implies is a measure of the coherence[2] between the two sounds that arrive to each ear. In a reverberant environment therefore, it would be the direct and early components of the signal that would have a higher coherence as compared to the late reverberant component (cfr. Chapter 2).

It is known that the binaural aspects of hearing benefit normal hearing individuals in adverse acoustic environments [4, 255]. The potential of restoring these aspects for hearing impaired individuals [251, 256, 257] has therefore

[2]Blauert [4] is clear on the definition of coherence of two signals here to mean if either both signals are identical or if they differ in one or more of the following attributes: (i) a level difference independent of freuqency, i.e., different amplitudes but the same signal, (ii) a phase delay independent of frequency, i.e., a pure delay of one of the signals, and (iii) a phase difference of 180°.

initiated the development of several binaural speech enhancement algorithms for binaural assistive hearing devices [14, 29, 34, 251, 258–263]. If the processing done on the binaural hearing assistive device is such that each device performs its own processing independently to yield an enhanced signal for the respective ear, then it is referred to as a bilateral processing scheme. Alternatively, if a communication link is available between both devices, then the two devices can be used in conjunction, i.e., all of the microphones from both devices can be used to do the processing to yield an enhanced signal for the respective ear. This is then referred to as binaural processing.

For a binaural assistive hearing device equipped with a separate local microphone array (LMA) for the left and right ear, the binaural minimum variance distortionless response beamformer (BMVDR) [259] is known to exhibit substantial noise reduction and to preserve the ITD and ILD of a target speaker[3]. One issue with this strategy, however, in a binaural processing scheme, is that the BMVDR beamformer preserves the binaural cues for the target speaker, but distorts the binaural cues for the noise. For instance, in a diffuse noise field, the spatial perception of the noise will be distorted such that it appears as coming from the direction of the target speaker. As several remedies [6, 223, 251, 260, 261] have been proposed to deal with this issue[4], this chapter is mostly concerned with the preservation of the binaural cues for the target speaker.

In recent work [223, 225], such a binaural assistive hearing device has also been supplemented with an external microphone (XM) and it was demonstrated that the XM could contribute to additional noise reduction and preserve the relevant binaural cues. For the successful operation of the BMVDR in this case, an estimate of the entire vector of transfer functions from the target signal at a left ear reference microphone to all the other microphones, i.e., the left ear relative transfer function (RTF) vector, and a corresponding right ear RTF vector is required. As discussed throughout the dissertation, a data-dependent estimation of such an entire RTF vector can be challenging in adverse acoustic environments.

In this chapter, therefore, the geometrically-based integration approach of explicitly using the a priori spatial knowledge as developed in chapter 5 will be explored. As opposed to considering only one XM as in [223, 225], the system will also be generalised to include multiple XMs. Whereas in chapter 5, the a priori spatial knowledge to be used corresponded to one LMA, in the binaural case, there are 2 LMAs to be considered, each with their own a priori spatial

[3] Herein after the ITDs and ILDs are referred to as the binaural cues and are the cues of interest for preservation.

[4] A simple remedy is that of reintroducing some the binaural unprocessed microphone signal to the enhanced binaural signal [251], which leads to a natural trade-off between noise reduction and the preservation of the spatial auditory perception of the acoustic environment.

knowledge. Specifically, the available a priori knowledge to be exploited will be a known RTF vector for the left ear LMA and a separate known RTF vector for the right ear LMA, which is not unrealistic to be available in practice [262]. As will be shown, upon incorporating such a priori knowledge, it is then only the estimation of an appropriate scaling between the left and right ear RTF vectors, and the missing part of these RTF vectors pertaining to the XMs that need to be estimated.

The remainder of this chapter is organised as follows. In Section 6.2, the data model pertaining specifically to a case of binaural assistive hearing devices and the associated notation are defined. It should be noted that there will be a slight abuse of notation in that some of the same symbols are used as in the previous chapter. The symbols are the same conceptually, however it is only the dimensions of the vectors and matrices that change accordingly. In Section 6.3, the state of the art procedure for estimating the entire RTF vector is reviewed. In Section 6.4, the proposed procedure that makes use of the partial a priori knowledge of the RTF vector is discussed Two strategies are proposed, one that implements a solution via binaural processing, and another which is related to a bilateral processing scheme. In Section 6.5, the various algorithms are evaluated using recorded audio data. Conclusions are drawn and future work is discussed in Section 6.6.

6.2 Data Model

The scenario as depicted in Figure 6.1 is considered, in which a user of a binaural assistive hearing device is listening to one target speaker of interest in a noisy, reverberant environment. The binaural assistive hearing device consists of an LMA with M_a microphones for the left ear and an LMA with M_a microphones for the right ear. Additionally, there are M_e XMs randomly placed within the room ($M_e = 2$ in Fig.6.1). In the short-time Fourier transform (STFT) domain,

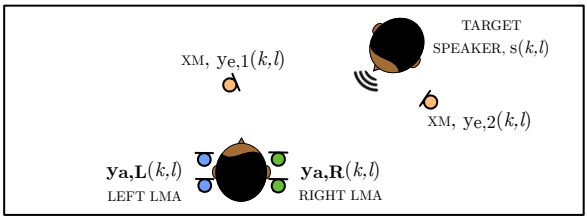


Figure 6.1: Scenario with a user of a binaural assistive hearing device having access to XMs, listening to the target speaker.

the microphone signals at frequency bin, k, and time frame, l, can be stacked into a vector and represented as follows:

$$\mathbf{y}(\mathrm{k},\mathrm{l}) = \underbrace{\mathbf{a}(\mathrm{k},\mathrm{l})\mathrm{s}(\mathrm{k},\mathrm{l})}_{\mathbf{x}(\mathrm{k},\mathrm{l})} + \mathbf{n}(\mathrm{k},\mathrm{l}) \tag{6.1}$$

$$\begin{bmatrix} \mathbf{y}_{\mathbf{a},\mathrm{L}}(\mathrm{k},\mathrm{l}) \\ \mathbf{y}_{\mathbf{a},\mathrm{R}}(\mathrm{k},\mathrm{l}) \\ \mathbf{y}_{\mathbf{e}}(\mathrm{k},\mathrm{l}) \end{bmatrix} = \begin{bmatrix} \mathbf{a}_{\mathbf{a},\mathrm{L}}(\mathrm{k},\mathrm{l}) \\ \mathbf{a}_{\mathbf{a},\mathrm{R}}(\mathrm{k},\mathrm{l}) \\ \mathbf{a}_{\mathbf{e}}(\mathrm{k},\mathrm{l}) \end{bmatrix} \mathrm{s}(\mathrm{k},\mathrm{l}) + \begin{bmatrix} \mathbf{n}_{\mathbf{a},\mathrm{L}}(\mathrm{k},\mathrm{l}) \\ \mathbf{n}_{\mathbf{a},\mathrm{R}}(\mathrm{k},\mathrm{l}) \\ \mathbf{n}_{\mathbf{e}}(\mathrm{k},\mathrm{l}) \end{bmatrix} \tag{6.2}$$

where[5] $\mathbf{a}_{\mathbf{a},\mathrm{L}} = [\mathrm{a}_{\mathrm{a1,L}},\ \mathrm{a}_{\mathrm{a2,L}},\ \ldots\ \mathrm{a}_{\mathrm{aM_a,L}}]^T$, $\mathbf{a}_{\mathbf{a},\mathrm{R}} = [\mathrm{a}_{\mathrm{a1,R}},\ \mathrm{a}_{\mathrm{a2,R}},\ \ldots\ \mathrm{a}_{\mathrm{aM_a,R}}]^T$, and $\mathbf{a}_{\mathbf{e}} = [\mathrm{a}_{\mathrm{e1}},\ \mathrm{a}_{\mathrm{e2}},\ \ldots\ \mathrm{a}_{\mathrm{eM_e}}]^T$ are the acoustic transfer function (ATF) vectors from the target speaker to the microphones on the left LMA, the right LMA, and the XMs respectively. Furthermore, s is the target speaker, and $\mathbf{n}_{\mathbf{a},\mathrm{L}}$, $\mathbf{n}_{\mathbf{a},\mathrm{R}}$, and $\mathbf{n}_{\mathbf{e}}$ are the noise contributions stacked in vectors similar to $\mathbf{a}_{\mathbf{a},\mathrm{L}}$, $\mathbf{a}_{\mathbf{a},\mathrm{R}}$, and $\mathbf{a}_{\mathbf{e}}$ respectively. Without loss of generality, the first microphone in each of the LMAs is also chosen as the reference microphone:

$$\mathrm{y}_{\mathrm{a1,L}} = \mathbf{e}_{\mathbf{L}}^T\mathbf{y} = \mathrm{s}_{\mathrm{a1,L}} + \mathrm{n}_{\mathrm{a1,L}} \qquad \mathrm{y}_{\mathrm{a1,R}} = \mathbf{e}_{\mathbf{R}}^T\mathbf{y} = \mathrm{s}_{\mathrm{a1,R}} + \mathrm{n}_{\mathrm{a1,R}} \tag{6.3}$$

where $\mathrm{s}_{\mathrm{a1,L}} = \mathrm{a}_{\mathrm{a1,L}}\mathrm{s}$, $\mathrm{s}_{\mathrm{a1,R}} = \mathrm{a}_{\mathrm{a1,R}}\mathrm{s}$, which are the speech components that need to be estimated, and $\mathbf{e}_{\mathbf{L}}$ and $\mathbf{e}_{\mathbf{R}}$ are all-zero vectors except for a one in the left and right LMA reference microphone position respectively. In order to perform the estimation, it is firstly convenient to re-define eq. (6.1) in terms of a relative transfer function (RTF) vector, as opposed to the ATF vector, $\mathbf{a}$. As the RTF vector is simply the ATF vector normalised to a reference microphone, in the binaural context, a separate RTF vector can be defined for the left ear and another for the right ear. Hence, eq. (6.1) can be expressed as follows:

$$\mathbf{y} = \mathbf{h}_{\mathrm{L}}\mathrm{s}_{\mathrm{a1,L}} + \mathbf{n} \qquad \mathbf{y} = \mathbf{h}_{\mathrm{R}}\mathrm{s}_{\mathrm{a1,R}} + \mathbf{n} \tag{6.4}$$

where $\mathbf{h}_{\mathrm{L}}$ and $\mathbf{h}_{\mathrm{R}}$ are the RTF vectors defined as:

$$\mathbf{h}_{\mathrm{L}} = \frac{1}{\mathrm{a}_{\mathrm{a1,L}}}\begin{bmatrix} \mathbf{a}_{\mathbf{a},\mathrm{L}} \\ \mathbf{a}_{\mathbf{a},\mathrm{R}} \\ \mathbf{a}_{\mathbf{e}} \end{bmatrix} = \begin{bmatrix} \mathbf{h}_{\mathbf{a},\mathrm{L}} \\ \varphi\mathbf{h}_{\mathbf{a},\mathrm{R}} \\ \mathbf{h}_{\mathbf{e},\mathrm{L}} \end{bmatrix} \qquad \mathbf{h}_{\mathrm{R}} = \frac{1}{\mathrm{a}_{\mathrm{a1,R}}}\begin{bmatrix} \mathbf{a}_{\mathbf{a},\mathrm{L}} \\ \mathbf{a}_{\mathbf{a},\mathrm{R}} \\ \mathbf{a}_{\mathbf{e}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\varphi}\mathbf{h}_{\mathbf{a},\mathrm{L}} \\ \mathbf{h}_{\mathbf{a},\mathrm{R}} \\ \frac{1}{\varphi}\mathbf{h}_{\mathbf{e},\mathrm{L}} \end{bmatrix} \tag{6.5}$$

where $\mathbf{h}_{\mathbf{a},\mathrm{L}} = \frac{\mathbf{a}_{\mathbf{a},\mathrm{L}}}{\mathrm{a}_{\mathrm{a1,L}}}$ and $\mathbf{h}_{\mathbf{a},\mathrm{R}} = \frac{\mathbf{a}_{\mathbf{a},\mathrm{R}}}{\mathrm{a}_{\mathrm{a1,R}}}$ are the individual RTF vectors corresponding to each of the left and right LMAs and the complex scaling, $\varphi = \frac{\mathrm{a}_{\mathrm{a1,R}}}{\mathrm{a}_{\mathrm{a1,L}}}$. It should be noted that the part of the RTF vector pertaining to the XMs in $\mathbf{h}_{\mathrm{R}}$ is a scaled version of that in $\mathbf{h}_{\mathrm{L}}$, where $\mathbf{h}_{\mathbf{e},\mathrm{L}} = \frac{\mathbf{a}_{\mathbf{e}}}{\mathrm{a}_{\mathrm{a1,L}}}$. In fact, it can be seen that $\mathbf{h}_{\mathrm{L}} = \varphi\mathbf{h}_{\mathrm{R}}$, which means that the RTF vectors are parallel.

[5] The dependence on (k, l) is dropped as each time-frequency bin is independently treated and for notational convenience as usual.

The speech-plus-noise spatial correlation matrix, $\mathbf{R_{yy}}$, the noise-only correlation matrix, $\mathbf{R_{nn}}$, and the speech-only correlation matrix, $\mathbf{R_{xx}}$, all $\in \mathbb{C}^{(2\mathrm{M_a}+\mathrm{M_e})\times(2\mathrm{M_a}+\mathrm{M_e})}$, are given respectively as:

$$\mathbf{R_{yy}} = \mathbb{E}\{\mathbf{y}\mathbf{y}^H\}; \;\; \mathbf{R_{nn}} = \mathbb{E}\{\mathbf{n}\mathbf{n}^H\}; \;\; \mathbf{R_{xx}} = \mathbb{E}\{\mathbf{x}\mathbf{x}^H\} \tag{6.6}$$

where $\mathbb{E}\{.\}$ is the expectation operator, ^{H} is the Hermitian transpose, and $\mathbf{R_{xx}}$ is a rank-1 correlation matrix:

$$\mathbf{R_{xx}} = \mathbb{E}\{\mathbf{x}\mathbf{x}^H\} = \sigma^2_{\mathrm{s_{a1,L}}}\mathbf{h}_\mathrm{L}\mathbf{h}_\mathrm{L}^H = \sigma^2_{\mathrm{s_{a1,R}}}\mathbf{h}_\mathrm{R}\mathbf{h}_\mathrm{R}^H \tag{6.7}$$

where $\sigma^2_{\mathrm{s_{a1,L}}} = \mathbb{E}\{|\mathrm{s_{a1,L}}|^2\}$ and $\sigma^2_{\mathrm{s_{a1,R}}} = \mathbb{E}\{|\mathrm{s_{a1,R}}|^2\}$ are the speech power spectral densities in the reference left and right microphone respectively. It is also assumed that the speech components are uncorrelated with the noise components, and hence $\mathbf{R_{yy}} = \mathbf{R_{xx}} + \mathbf{R_{nn}}$. A perfect communication link is additionally assumed among the left and right LMAs in the binaural assistive hearing device, and the XMs, with no bandwidth constraints and synchronous sampling.

The estimate of the speech component in the reference microphone of the left and right LMAs, i.e. the estimate of $\mathrm{s_{a1,L}}$ and $\mathrm{s_{a1,R}}$ is then obtained through the linear filtering of the microphone signals, with the complex-valued filters, $\mathbf{w}_\mathrm{L}$ and $\mathbf{w}_\mathrm{R}$ respectively:

$$\hat{\mathrm{s}}_{\mathrm{a1,L}} = \mathbf{w}_\mathrm{L}^H\mathbf{y} \qquad\qquad \hat{\mathrm{s}}_{\mathrm{a1,R}} = \mathbf{w}_\mathrm{R}^H\mathbf{y} \tag{6.8}$$

The BMVDR beamformer filters, $\mathbf{w}_\mathrm{L}$ and $\mathbf{w}_\mathrm{R}$, are then given by:

$$\mathbf{w}_\mathrm{L} = \frac{\mathbf{R_{nn}}^{-1}\mathbf{h}_\mathrm{L}}{\mathbf{h}_\mathrm{L}^H\mathbf{R_{nn}}^{-1}\mathbf{h}_\mathrm{L}} \qquad\qquad \mathbf{w}_\mathrm{R} = \frac{\mathbf{R_{nn}}^{-1}\mathbf{h}_\mathrm{R}}{\mathbf{h}_\mathrm{R}^H\mathbf{R_{nn}}^{-1}\mathbf{h}_\mathrm{R}} \tag{6.9}$$

which require estimates for $\mathbf{R_{nn}}$ and the left and right RTF vectors, $\mathbf{h}_\mathrm{L}$, and $\mathbf{h}_\mathrm{R}$. In the following, the focus is on the estimation of $\mathbf{h}_\mathrm{L}$ and $\mathbf{h}_\mathrm{R}$.

6.3 Data-dependent estimate of the entire RTF vector

Given $\hat{\mathbf{R}}_\mathbf{yy}$ and $\hat{\mathbf{R}}_\mathbf{nn}$, which are estimates of $\mathbf{R_{yy}}$ and $\mathbf{R_{nn}}$ respectively, a generalised eigenvalue decomposition (GEVD) [39] or what is equivalently known as covariance whitening [225] can be used to estimate $\mathbf{h}_\mathrm{L}$ and $\mathbf{h}_\mathrm{R}$, as discussed in subsection 3.4.2. As in (5.38), a spatial pre-whitening operation is firstly defined from $\hat{\mathbf{R}}_\mathbf{nn}$ using the Cholesky decomposition:

$$\hat{\mathbf{R}}_\mathbf{nn} = \hat{\mathbf{R}}_\mathbf{nn}^{1/2}\,\hat{\mathbf{R}}_\mathbf{nn}^{H/2} \tag{6.10}$$

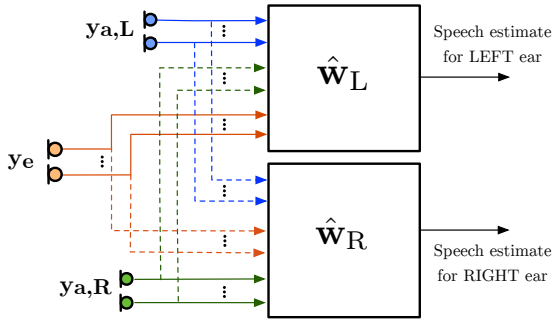


Figure 6.2: Block scheme for a BMVDR that uses partial a priori knowledge of the RTF vectors.

where $\hat{\mathbf{R}}_{\mathbf{nn}}^{1/2}$ is a lower triangular matrix. An optimisation problem analagous to (5.39) can then be considered to estimate $\mathbf{h}_{\mathrm{L}}$ (and $\mathbf{h}_{\mathrm{R}}$ by an appropriate scaling):

$$\underset{\sigma^2_{\mathrm{s_{a1,L}}},\,\mathbf{h}_{\mathrm{L}}}{\text{minimise}} \; ||\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2}((\hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{nn}}) - \sigma^2_{\mathrm{s_{a1,L}}}\mathbf{h}_{\mathrm{L}}\mathbf{h}_{\mathrm{L}}^H)\hat{\mathbf{R}}_{\mathbf{nn}}^{-H/2}||_F^2 \tag{6.11}$$

where $||.||_F$ is the frobenius norm. The solution to eq. (6.11) then follows from an eigenvalue decomposition (EVD) of $\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2}\hat{\mathbf{R}}_{\mathbf{yy}}\hat{\mathbf{R}}_{\mathbf{nn}}^{-H/2}$ or equivalently, GEVD of the matrix pencil $(\hat{\mathbf{R}}_{\mathbf{yy}}, \hat{\mathbf{R}}_{\mathbf{nn}})$. As discussed in subsection 3.4.2, this is equivalent to a joint diagonalisation of $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$ such that:

$$\hat{\mathbf{R}}_{\mathbf{yy}} = \hat{\mathbf{Q}}\hat{\mathbf{\Sigma}}\hat{\mathbf{Q}}^H \qquad \hat{\mathbf{R}}_{\mathbf{nn}} = \hat{\mathbf{Q}}\mathbf{I}_{(2\mathrm{M_a}+\mathrm{M_e})}\hat{\mathbf{Q}}^H \tag{6.12}$$

where $\hat{\mathbf{\Sigma}}$ is a diagonal matrix of the generalised eigenvalues arranged in descending order and $\hat{\mathbf{Q}}$ is an invertible matrix. From subsection 3.4.2, the estimates of $\mathbf{h}_{\mathrm{L}}$ and $\mathbf{h}_{\mathrm{R}}$ then follow with the appropriate scaling for the left and right ears as:

$$\hat{\mathbf{h}}_{\mathrm{L}} = \frac{\hat{\mathbf{Q}}\mathbf{e_1}}{\mathbf{e}_{\mathbf{L}}^T\hat{\mathbf{Q}}\mathbf{e_1}} \qquad \hat{\mathbf{h}}_{\mathrm{R}} = \frac{\hat{\mathbf{Q}}\mathbf{e_1}}{\mathbf{e}_{\mathbf{R}}^T\hat{\mathbf{Q}}\mathbf{e_1}} \tag{6.13}$$

where $\mathbf{e_1} \in \mathbb{C}^{2\mathrm{M_a}+\mathrm{M_e}}$ is an all-zero vector except for a one as the first element (and it is noted that $\mathbf{e_1} = \mathbf{e_L}$). Finally, a substitution of $\hat{\mathbf{R}}_{\mathbf{nn}}$ from eq. (6.12) and $\hat{\mathbf{h}}_{\mathrm{L}}$ and $\hat{\mathbf{h}}_{\mathrm{R}}$ from eq. (6.13) into eq. (6.9) results in the corresponding BMVDR filters:

$$\hat{\mathbf{w}}_{\mathrm{L}} = (\mathbf{e}_1^T\hat{\mathbf{Q}}^H\mathbf{e_L})\hat{\mathbf{Q}}^{-H}\mathbf{e_1} \qquad \hat{\mathbf{w}}_{\mathrm{R}} = (\mathbf{e}_1^T\hat{\mathbf{Q}}^H\mathbf{e_R})\hat{\mathbf{Q}}^{-H}\mathbf{e_1} \tag{6.14}$$

Figure 6.2 illustrates the block scheme for such a procedure. The two filters from (6.26) directly filter all of the microphone signals (as in (6.8)) to obtain the respective left ear and right ear speech estimates.

6.4 Using a priori knowledge of the left and right RTF vector

In this section, a binaural scheme of Fig. 6.3 is firstly discussed which makes use of the partial a priori knowledge of the left and right ear RTF vectors. With the development of this scheme, it will be evident that another scheme, related to (but not exactly) a bilateral processing scheme in Fig. 6.4 can also be realised within the same structure. The performance of these schemes, along with that from Sec. 6.3 will then be evaluated in Sec. 6.5.

6.4.1 Binaural processing

As opposed to estimating the entire RTF vectors, $\mathbf{h}_{\mathrm{L}}$, and $\mathbf{h}_{\mathrm{R}}$, an alternative procedure may be followed if there is a priori knowledge of the RTF vectors for the separate left and right LMA, i.e., if a suitable approximation to $\mathbf{h}_{\mathbf{a},\mathrm{L}}$ and $\mathbf{h}_{\mathbf{a},\mathrm{R}}$ is available. For instance, such an approximation may be the measured RTF vectors for the separate left and right LMA in an anechoic room or RTF vectors from an existing binaural noise reduction system that uses only the LMAs. Denoting $\widetilde{\mathbf{h}}_{\mathbf{a},\mathrm{L}}$ and $\widetilde{\mathbf{h}}_{\mathbf{a},\mathrm{R}}$ as the a priori RTF vectors for the left and right ears respectively, and recalling the definitions from eq. (6.5), an alternative optimisation problem to eq. (6.11) can be considered:

$$\underset{\sigma^2_{\mathrm{s_{a1,L}}},\,\varphi,\,\mathbf{h}_{\mathbf{e},\mathrm{L}}}{\text{minimise}} \; ||\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2}((\hat{\mathbf{R}}_{\mathbf{yy}}-\hat{\mathbf{R}}_{\mathbf{nn}})-\sigma^2_{\mathrm{s_{a1,L}}}\begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a},\mathrm{L}}\\ \varphi\widetilde{\mathbf{h}}_{\mathbf{a},\mathrm{R}}\\ \mathbf{h}_{\mathbf{e},\mathrm{L}}\end{bmatrix}\begin{bmatrix}\widetilde{\mathbf{h}}^H_{\mathbf{a},\mathrm{L}} & \varphi^*\widetilde{\mathbf{h}}^H_{\mathbf{a},\mathrm{R}} & \mathbf{h}^H_{\mathbf{e},\mathrm{L}}\end{bmatrix})\hat{\mathbf{R}}_{\mathbf{nn}}^{-H/2}||^2_F \tag{6.15}$$

where now it is only the scaling, φ, and the RTF vector for the XMs, $\mathbf{h}_{\mathbf{e},\mathrm{L}}$, which need to be found as opposed to the entire $\mathbf{h}_{\mathrm{L}}$ as in eq. (6.11). Furthermore, this equation only differs from (5.40) in Chapter 5 by the additional estimation of the scaling, φ. As will be discussed in the following, the solution can be realised in the block scheme of Figure 6.3, which consists of compressing the left and right LMA signals, an orthogonalisation operation, and finally a GEVD on a lower dimensional ($\mathbb{C}^{(\mathrm{M_e}+2)\times(\mathrm{M_e}+2)}$) matrix pencil. In order to solve eq. (6.15), the following blocking matrices, $\widetilde{\mathbf{C}}_{\mathbf{a}} \in \mathbb{C}^{2\mathrm{M_a}\times(2\mathrm{M_a}-2)}$, $\widetilde{\mathbf{C}}_{\mathbf{a},\mathrm{L}} \in \mathbb{C}^{\mathrm{M_a}\times(\mathrm{M_a}-1)}$, $\widetilde{\mathbf{C}}_{\mathbf{a},\mathrm{R}} \in \mathbb{C}^{\mathrm{M_a}\times(\mathrm{M_a}-1)}$, fixed beamformers, $\widetilde{\mathbf{F}}_{\mathbf{a}} \in \mathbb{C}^{2\mathrm{M_a}\times 2}$, $\widetilde{\mathbf{f}}_{\mathbf{a},\mathrm{L}} \in \mathbb{C}^{\mathrm{M_a}}$, $\widetilde{\mathbf{f}}_{\mathbf{a},\mathrm{R}} \in \mathbb{C}^{\mathrm{M_a}}$

are firstly defined:

$$\widetilde{\mathbf{C}}_{\mathbf{a}} = \begin{bmatrix} \widetilde{\mathbf{C}}_{\mathbf{a},\mathrm{L}} & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{C}}_{\mathbf{a},\mathrm{R}} \end{bmatrix} \qquad \widetilde{\mathbf{F}}_{\mathbf{a}} = \begin{bmatrix} \widetilde{\mathbf{f}}_{\mathbf{a},\mathrm{L}} & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{f}}_{\mathbf{a},\mathrm{R}} \end{bmatrix} \tag{6.16}$$

$$\widetilde{\mathbf{C}}_{\mathbf{a},\mathrm{L}}^{H}\widetilde{\mathbf{h}}_{\mathbf{a},\mathrm{L}} = \mathbf{0};\ \widetilde{\mathbf{C}}_{\mathbf{a},\mathrm{R}}^{H}\widetilde{\mathbf{h}}_{\mathbf{a},\mathrm{R}} = \mathbf{0} \qquad \widetilde{\mathbf{f}}_{\mathbf{a},\mathrm{L}}^{H}\widetilde{\mathbf{h}}_{\mathbf{a},\mathrm{L}} = 1;\ \widetilde{\mathbf{f}}_{\mathbf{a},\mathrm{R}}^{H}\widetilde{\mathbf{h}}_{\mathbf{a},\mathrm{R}} = 1$$

followed by the and transformation matrix, $\widetilde{\mathbf{T}} \in \mathbb{C}^{(2\mathrm{M_a}+\mathrm{M_e})\times(2\mathrm{M_a}+\mathrm{M_e})}$:

$$\widetilde{\mathbf{T}} = \left[\begin{array}{c|c|c} \widetilde{\mathbf{C}}_{\mathbf{a}} & \widetilde{\mathbf{F}}_{\mathbf{a}} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{I}_{\mathrm{M_e}} \end{array}\right] \tag{6.17}$$

where $\mathbf{I}_{\mathrm{M_e}} \in \mathbb{C}^{\mathrm{M_e}\times\mathrm{M_e}}$ is an identity matrix. The first two blocks in Fig. 6.3 apply the transformation, $\widetilde{\mathbf{T}}^{H}$, to $\mathbf{y}$ to yield a set of blocking matrix signals, $\widetilde{\mathbf{C}}_{\mathbf{a}}^{H}\mathbf{y}_{\mathbf{a}} \in \mathbb{C}^{2\mathrm{M_a}-2}$, two compressed signals, $\widetilde{\mathbf{f}}_{\mathbf{a},\mathrm{L}}^{H}\mathbf{y}_{\mathbf{a},\mathrm{L}}$ and $\widetilde{\mathbf{f}}_{\mathbf{a},\mathrm{R}}^{H}\mathbf{y}_{\mathbf{a},\mathrm{R}}$ resulting from the left and right fixed beamformers respectively, and the unaltered set of XM signals, $\mathbf{y}_{\mathbf{e}}$. An alternative spatial pre-whitening operation can then be defined by applying the transformation to $\hat{\mathbf{R}}_{\mathbf{nn}}$:

$$\widetilde{\mathbf{T}}^{H}\hat{\mathbf{R}}_{\mathbf{nn}}\widetilde{\mathbf{T}}^{H} = \hat{\mathbf{L}}\hat{\mathbf{L}}^{H} \tag{6.18}$$

where $\hat{\mathbf{L}}$ is a lower triangular matrix. Following a similar procedure from Sec. 5.5.2 in chapter 5, it can then be shown that eq. (6.15) can be equivalently re-written as:

$$\underset{\sigma^2_{\mathrm{s_{a1,L}}},\,\varphi,\,\mathbf{h}_{\mathbf{e},\mathrm{L}}}{\text{minimise}} \; ||\hat{\mathbf{L}}^{-1}\widetilde{\mathbf{T}}^{H}((\hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{nn}}) - \sigma^2_{\mathrm{s_{a1,L}}}\begin{bmatrix} \widetilde{\mathbf{h}}_{\mathbf{a},\mathrm{L}} \\ \varphi\widetilde{\mathbf{h}}_{\mathbf{a},\mathrm{R}} \\ \mathbf{h}_{\mathbf{e},\mathrm{L}} \end{bmatrix}\left[\widetilde{\mathbf{h}}_{\mathbf{a},\mathrm{L}}^{H}\ \varphi^{*}\widetilde{\mathbf{h}}_{\mathbf{a},\mathrm{R}}^{H}\ \mathbf{h}_{\mathbf{e},\mathrm{L}}^{H}\right])\widetilde{\mathbf{T}}\,\hat{\mathbf{L}}^{-H}||_F^2 \tag{6.19}$$

from which the solution follows eventually from a GEVD of a matrix pencil consisting of the lower dimensional $\mathbb{C}^{(\mathrm{M_e}+2)\times(\mathrm{M_e}+2)}$ correlation matrices [264] [265] corresponding to the $\mathrm{M_e}+2$ compressed signals, $\underset{\sim}{\mathbf{y}} = [\underset{\sim}{\mathrm{y}}_{\mathrm{L}}\ \underset{\sim}{\mathrm{y}}_{\mathrm{R}}\ \underset{\sim}{\mathrm{y}}_{\mathrm{e},1}\cdots\underset{\sim}{\mathrm{y}}_{\mathrm{e,Me}}]^{T} \in \mathbb{C}^{\mathrm{M_e}+2}$. These compressed signals can be computed by performing an orthogonalisation involving the previously transformed signals, or equivalently, with a GSC beamformer, $\mathbf{A}$:

$$\underset{\sim}{\mathbf{y}} = \mathbf{A}^{H}\mathbf{y} = \widetilde{\mathbf{F}}^{H}\mathbf{y} - \hat{\mathbf{G}}^{H}\widetilde{\mathbf{C}}_{\mathbf{a}}^{H}\mathbf{E}^{T}\mathbf{y} \tag{6.20}$$

with $\mathbf{A} = \widetilde{\mathbf{F}} - \mathbf{E}\widetilde{\mathbf{C}}_{\mathbf{a}}\hat{\mathbf{G}}$, $\widetilde{\mathbf{F}} \in \mathbb{C}^{(2\mathrm{M_a}+\mathrm{M_e})\times(\mathrm{M_e}+2)}$ defined as $\widetilde{\mathbf{F}} = \begin{bmatrix} \widetilde{\mathbf{F}}_{\mathbf{a}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{\mathrm{M_e}} \end{bmatrix}$, the selection matrix, $\mathbf{E} = [\mathbf{I}_{2\mathrm{M_a}}\,|\,\mathbf{0}_{(2\mathrm{M_a}\times\mathrm{M_e})}]^{T}$, and $\hat{\mathbf{G}} \in \mathbb{C}^{(2\mathrm{M_a}-2)\times(\mathrm{M_e}+2)}$, which can be computed as follows:

$$\hat{\mathbf{G}} = (\widetilde{\mathbf{C}}_{\mathbf{a}}^{H}\mathbf{E}^{T}\hat{\mathbf{R}}_{\mathbf{nn}}\mathbf{E}\widetilde{\mathbf{C}}_{\mathbf{a}})^{-1}\widetilde{\mathbf{C}}_{\mathbf{a}}^{H}\mathbf{E}^{T}\hat{\mathbf{R}}_{\mathbf{nn}}\widetilde{\mathbf{F}} = [\hat{\mathbf{g}}_{\mathbf{a},\mathrm{L}}\ \hat{\mathbf{g}}_{\mathbf{a},\mathrm{R}}\ \hat{\mathbf{g}}_{\mathbf{e},1}\ \cdots\ \hat{\mathbf{g}}_{\mathbf{e},\mathrm{M_e}}] \tag{6.21}$$

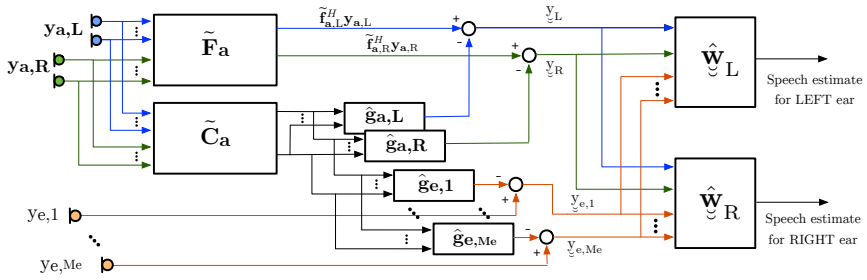


Figure 6.3: Block scheme for a BMVDR that uses partial a priori knowledge of the RTF vectors.

is comprised of the $\mathbb{C}^{M_e+2}$ orthogonalisation filters in each column. As depicted in Fig. 6.3, $\hat{\mathbf{g}}_{\mathbf{a},L}$ and $\hat{\mathbf{g}}_{\mathbf{a},R}$ will orthogonalise the noise components of the signals $\widetilde{\mathbf{f}}^H_{\mathbf{a},L}\mathbf{y}_{\mathbf{a},L}$ and $\widetilde{\mathbf{f}}^H_{\mathbf{a},R}\mathbf{y}_{\mathbf{a},R}$ onto the noise components of the blocking matrix signals, and $[\hat{\mathbf{g}}_{\mathbf{e},1} \ \dots \ \hat{\mathbf{g}}_{\mathbf{e},M_e}]$ will orthogonalise the noise components of each of the respective XMs onto the noise components of the blocking matrix signals.

A new signal model that resembles eq. (6.4) can in fact then be realised by substituting eq. (6.4) into eq. (6.20):

$$\underset{\sim}{\mathbf{y}} = \underset{\sim}{\mathbf{h}}_L s_{a1,L} + \underset{\sim}{\mathbf{n}} \qquad \underset{\sim}{\mathbf{y}} = \underset{\sim}{\mathbf{h}}_R s_{a1,R} + \underset{\sim}{\mathbf{n}} \tag{6.22}$$

where $\underset{\sim}{\mathbf{h}}_L = \mathbf{A}^H\mathbf{h}_L = \left[1 \ \varphi \ \mathbf{h}^T_{e,L}\right]^T$, $\underset{\sim}{\mathbf{h}}_R = \mathbf{A}^H\mathbf{h}_R = \left[\frac{1}{\varphi} \ 1 \ \frac{1}{\varphi}\mathbf{h}^T_{e,L}\right]^T$, $\underset{\sim}{\mathbf{n}} = \mathbf{A}^H\mathbf{n}$, and the associated correlation matrices, $\underset{\sim}{\mathbf{R}}_{\mathbf{yy}} = \mathbb{E}\{\underset{\sim}{\mathbf{y}}\underset{\sim}{\mathbf{y}}^H\}$ and $\underset{\sim}{\mathbf{R}}_{\mathbf{nn}} = \mathbb{E}\{\underset{\sim}{\mathbf{n}}\underset{\sim}{\mathbf{n}}^H\}$. These compressed signals, $\underset{\sim}{\mathbf{y}}$, can now be used to design the BMVDR as opposed to $\mathbf{y}$. Estimates of $s_{a1,L}$ and $s_{a1,R}$ can be obtained by directly filtering $\underset{\sim}{\mathbf{y}}$ with the BMVDR filters, $\underset{\sim}{\mathbf{w}}_L \in \mathbb{C}^{M_e+2}$ and $\underset{\sim}{\mathbf{w}}_R \in \mathbb{C}^{M_e+2}$ respectively (as in eq. (6.8) and depicted in Fig. 6.3):

$$\underset{\sim}{\mathbf{w}}_L = \frac{\underset{\sim}{\mathbf{R}}^{-1}_{\mathbf{nn}}\underset{\sim}{\mathbf{h}}_L}{\underset{\sim}{\mathbf{h}}^H_L\underset{\sim}{\mathbf{R}}^{-1}_{\mathbf{nn}}\underset{\sim}{\mathbf{h}}_L} \qquad \underset{\sim}{\mathbf{w}}_R = \frac{\underset{\sim}{\mathbf{R}}^{-1}_{\mathbf{nn}}\underset{\sim}{\mathbf{h}}_R}{\underset{\sim}{\mathbf{h}}^H_R\underset{\sim}{\mathbf{R}}^{-1}_{\mathbf{nn}}\underset{\sim}{\mathbf{h}}_R} \tag{6.23}$$

Hence, similar to eq. (6.9), estimates are now required for $\underset{\sim}{\mathbf{R}}_{\mathbf{nn}}$, $\underset{\sim}{\mathbf{h}}_L$, and $\underset{\sim}{\mathbf{h}}_R$ in order to compute these filters. With the estimates of the correlation matrices, $\hat{\underset{\sim}{\mathbf{R}}}_{\mathbf{yy}} = \mathbf{A}^H\hat{\mathbf{R}}_{\mathbf{yy}}\mathbf{A}$ and $\hat{\underset{\sim}{\mathbf{R}}}_{\mathbf{nn}} = \mathbf{A}^H\hat{\mathbf{R}}_{\mathbf{nn}}\mathbf{A}$, the GEVD of the matrix pencil $(\hat{\underset{\sim}{\mathbf{R}}}_{\mathbf{yy}}, \hat{\underset{\sim}{\mathbf{R}}}_{\mathbf{nn}})$ can be used to estimate $\underset{\sim}{\mathbf{h}}_L$ and $\underset{\sim}{\mathbf{h}}_R$. The joint diagonalisation of $\hat{\underset{\sim}{\mathbf{R}}}_{\mathbf{yy}}$ and $\hat{\underset{\sim}{\mathbf{R}}}_{\mathbf{nn}}$ follows as:

$$\hat{\underset{\sim}{\mathbf{R}}}_{\mathbf{yy}} = \hat{\underset{\sim}{\mathbf{Q}}}\,\hat{\underset{\sim}{\mathbf{\Sigma}}}\,\hat{\underset{\sim}{\mathbf{Q}}}^H; \qquad \hat{\underset{\sim}{\mathbf{R}}}_{\mathbf{nn}} = \hat{\underset{\sim}{\mathbf{Q}}}\,\mathbf{I}_{(M_e+2)}\,\hat{\underset{\sim}{\mathbf{Q}}}^H \tag{6.24}$$

where $\hat{\underset{\sim}{\mathbf{\Sigma}}}$ is a diagonal matrix of the generalised eigenvalues arranged in descending order and $\hat{\underset{\sim}{\mathbf{Q}}}$ is an invertible matrix. The estimates for $\hat{\underset{\sim}{\mathbf{h}}}_{\mathrm{L}}$ and $\hat{\underset{\sim}{\mathbf{h}}}_{\mathrm{R}}$ then follow as:

$$\hat{\underset{\sim}{\mathbf{h}}}_{\mathrm{L}} = \frac{\hat{\underset{\sim}{\mathbf{Q}}}\,\underset{\sim}{\mathbf{e}}_{\mathbf{1}}}{\underset{\sim}{\mathbf{e}}_{\mathbf{L}}^{T}\hat{\underset{\sim}{\mathbf{Q}}}\,\underset{\sim}{\mathbf{e}}_{\mathbf{1}}} \qquad \hat{\underset{\sim}{\mathbf{h}}}_{\mathrm{R}} = \frac{\hat{\underset{\sim}{\mathbf{Q}}}\,\underset{\sim}{\mathbf{e}}_{\mathbf{1}}}{\underset{\sim}{\mathbf{e}}_{\mathbf{R}}^{T}\hat{\underset{\sim}{\mathbf{Q}}}\,\underset{\sim}{\mathbf{e}}_{\mathbf{1}}} \tag{6.25}$$

where the selection vectors, $\underset{\sim}{\mathbf{e}}_{\mathbf{1}} = \underset{\sim}{\mathbf{e}}_{\mathbf{L}} = [1,\, 0,\, \mathbf{0}_{(1\times \mathrm{M_e})}]^T$ and $\underset{\sim}{\mathbf{e}}_{\mathbf{R}} = [0,\, 1,\, \mathbf{0}_{(1\times \mathrm{M_e})}]^T$. Finally, the substitution of $\hat{\underset{\sim}{\mathbf{R}}}_{\mathbf{nn}}$ from eq. (6.24) and $\hat{\underset{\sim}{\mathbf{h}}}_{\mathrm{L}}$ and $\hat{\underset{\sim}{\mathbf{h}}}_{\mathrm{R}}$ from eq. (6.25) into eq. (6.23) results in the corresponding BMVDR filters:

$$\hat{\underset{\sim}{\mathbf{w}}}_{\mathrm{L}} = (\underset{\sim}{\mathbf{e}}_{\mathbf{1}}^{T}\hat{\underset{\sim}{\mathbf{Q}}}^{H}\underset{\sim}{\mathbf{e}}_{\mathbf{L}})\hat{\underset{\sim}{\mathbf{Q}}}^{-H}\underset{\sim}{\mathbf{e}}_{\mathbf{1}} \qquad \hat{\underset{\sim}{\mathbf{w}}}_{\mathrm{R}} = (\underset{\sim}{\mathbf{e}}_{\mathbf{1}}^{T}\hat{\underset{\sim}{\mathbf{Q}}}^{H}\underset{\sim}{\mathbf{e}}_{\mathbf{R}})\hat{\underset{\sim}{\mathbf{Q}}}^{-H}\underset{\sim}{\mathbf{e}}_{\mathbf{1}} \tag{6.26}$$

6.4.2 Semi-bilateral processing

In the binaural processing schemes previously described, all available signals were used to design the corresponding filters. In a true bilateral processing system (involving only the head-mounted microphones), however, only signals from the left ear LMA, i.e. $\mathbf{y}_{\mathbf{a},\mathrm{L}}$ are used to design the left ear filter, and only signals from the right ear LMA, i.e. $\mathbf{y}_{\mathbf{a},\mathrm{R}}$ are used to design the right ear filter. For instance, a bilateral scheme would produce (i) a left ear enhanced signal by filtering $\mathbf{y}_{\mathbf{a},\mathrm{L}}$ with a GSC that uses $\widetilde{\mathbf{f}}_{\mathbf{a},\mathrm{L}}$ and $\widetilde{\mathbf{C}}_{\mathbf{a},\mathrm{L}}$ as the fixed beamformer and blocking matrix respectively and (ii) a right ear enhanced signal by filtering $\mathbf{y}_{\mathbf{a},\mathrm{R}}$ with a GSC that uses $\widetilde{\mathbf{f}}_{\mathbf{a},\mathrm{R}}$ and $\widetilde{\mathbf{C}}_{\mathbf{a},\mathrm{R}}$ as the fixed beamformer and blocking matrix respectively.

In Fig. 6.3, the signals, $[\underset{\sim}{\mathrm{y}}_{\mathrm{L}}\ \underset{\sim}{\mathrm{y}}_{\mathrm{R}}]^T$ are related to such a bilateral scheme when considering only the head-mounted microphones. The difference to the bilateral scheme, however, is that the blocking matrix is the same in each GSC used to produce the left and right enhanced signal. Theoretically, for such a scheme, the binaural cues for the target speaker should remain undistorted, but there will still be some distortion of the binaural cues for the noise. However, the difference as opposed to the binaural scheme is that the RTF vectors that will be used are not parallel anymore and it is possible that some of the binaural cues for the noise may not be as heavily distorted. So even though this semi-bilateral processing may seem suboptimal (fewer degrees of freedom), it is nevertheless investigated to observe the effect on the binaural cues for the target speaker and noise.

Furthermore, upon incorporating the XMs, another bilateral type scheme can be considered as in Fig. 6.4, where the signals $[\underset{\sim}{y}_{L}\ \underset{\sim}{y}_{e,1}\cdots\underset{\sim}{y}_{e,Me}]^{T}$ are filtered by $\hat{\mathbf{w}}_{\underset{\sim}{L}}^{biL}\in\mathbb{C}^{M_e+1}$ to produce the left ear enhanced signal and $[\underset{\sim}{y}_{R}\ \underset{\sim}{y}_{e,1}\cdots\underset{\sim}{y}_{e,Me}]^{T}$ are filtered by $\hat{\mathbf{w}}_{\underset{\sim}{R}}^{biL}\in\mathbb{C}^{M_e+1}$ to produce the right ear enhanced signal. This scheme will be referred to as semi-bilateral in the following. The filters $\hat{\mathbf{w}}_{\underset{\sim}{L}}^{biL}$ and $\hat{\mathbf{w}}_{\underset{\sim}{R}}^{biL}$ are computed identically to those of $\hat{\mathbf{w}}_{\underset{\sim}{L}}$ and $\hat{\mathbf{w}}_{\underset{\sim}{R}}$ except with the corresponding lower dimension. The corresponding RTF vectors to be estimated for $\hat{\mathbf{w}}_{\underset{\sim}{L}}^{biL}$ and $\hat{\mathbf{w}}_{\underset{\sim}{R}}^{biL}$ respectively would now be $\left[1\ \mathbf{h}_{\mathbf{e},L}^{T}\right]^{T}$ and $\left[1\ \frac{1}{\varphi}\mathbf{h}_{\mathbf{e},L}^{T}\right]^{T}$.

As this semi-bilateral processing scheme with or without the XMs was a by product of the binaural scheme from Sec. 6.4.1, it was decided that it should be evaluated along with the binaural schemes in order to have some insight if such a scheme could be useful in practice.

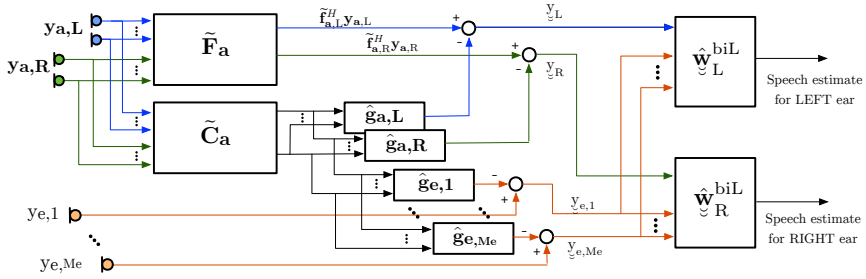


Figure 6.4: Block scheme for a BMVDR that uses partial a priori knowledge of the RTF vectors.

6.5 Evaluation with BTE-HAs

In order to evaluate these algorithms, audio recordings were made in the KU Leuven ESAT-STADIUS audio laboratory for the scenario depicted in Fig. 6.5. A Neumann KU-100 dummy head was placed in a central location of the room and equipped with the two behind-the-ear (BTE) hearing aids (HAs) used in Chapter 5, each consisting of an LMA with two microphones spaced approximately 1.3 cm apart. Two XMs (AKG-CK-97-O) were positioned such that XM1 was at $-60°$ and 1 m away from the dummy head, and XM2 was at 20 cm below the dummy head. A Genelec 8030C loudspeaker, 1 m from the dummy head at $-45°$, was used to generate the speech signal from a target male speaker [244]. Four of the same loudspeakers, facing away from the dummy head toward opposite corners (depicted in grey in Fig. 6.5), were used to generate

uncorrelated excerpts of babble noise [206] to approximate a diffuse noise field. The speech and noises were recorded separately, but mixed together at an input signal-to-noise ratio (SNR) of 0 dB at the reference (frontal) microphone on the right LMA, $y_{a1,R}$.

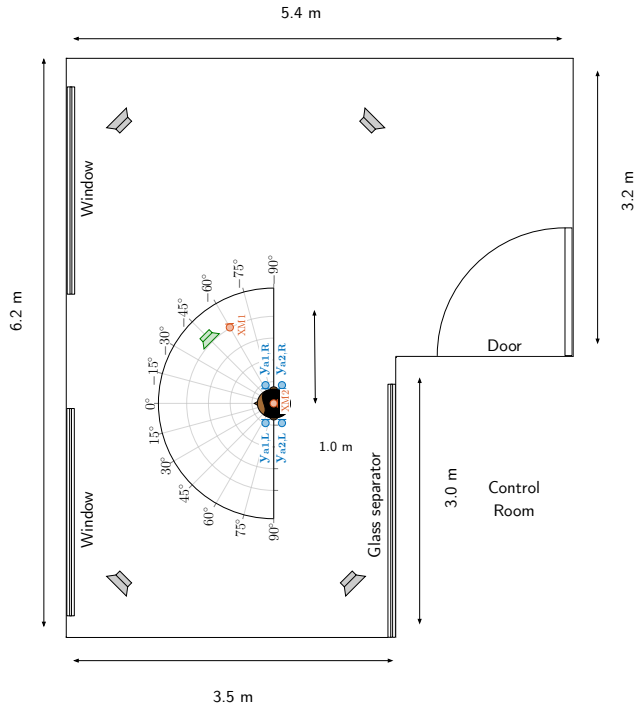


Figure 6.5: Acoustic scenario in the KU Leuven ESAT-STADIUS audio laboratory used to evaluate the binaural processing schemes.

The three algorithms: (i) the binaural estimation procedure from Sec. 6.3, (ii) the binaural procedure using a priori spatial knowledge from Sec. 6.4.1, and (iii) the semi-bilateral using a priori spatial knowledge procedure from Sec. 6.4.2, were then evaluated with and without the XMs. Table 6.1 summarises the details of the six cases and assigns a name for which each will be referred to in the following discussion.

For the processing of the algorithms, the Weighted Overlap and Add (WOLA) method [101], with a Discrete Fourier Transform (DFT) size of 128 samples, 50 % overlap, a square-root hanning window, and a sampling frequency of 16 kHz was used. Using the speech presence probability (SPP) [192], periods for which the speech was active were extracted if the SPP > 0.6, and not active if the SPP

Table 6.1: The various algorithms to be evaluated.

Algorithm Name	Description
HABIN-DD	Processing scheme of Fig. 6.2 but without the XMs.
HABIN-AP	Processing scheme of Fig. 6.3 but without the XMs.
HABiL-AP	$[\underset{\smile}{\mathrm{y}}_{\mathrm{L}}\ \underset{\smile}{\mathrm{y}}_{\mathrm{R}}]^T$ used as speech estimates for left and right ear respectively
HABIN-XM-DD	Processing scheme of Fig. 6.2.
HABIN-XM-AP	Processing scheme of Fig. 6.3.
HABiL-XM-AP	Processing scheme of Fig. 6.4

< 0.4. The estimated correlation matrices, $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$, were then computed accordingly via recursive averaging with a time constant of 1 s. The a priori RTF vectors, $\widetilde{\mathbf{h}}_{\mathbf{a},\mathrm{L}}$ and $\widetilde{\mathbf{h}}_{\mathbf{a},\mathrm{R}}$, were then computed from the DFT of 128 samples of the pre-measured impulse responses (representative of the direct component) from the target speaker position depicted in Fig. 6.5 to each of the left and right LMAs using the procedure previously discussed in Sec. 5.6.2 of Chapter 5. The metrics used to evaluate the following experiments were the change in speech intelligibility-weighted SNR improvement [159] (ΔSI-SNR) from the input SI-SNR at $\mathrm{y}_{\mathrm{a1,R}}$, a change in a binaural version of the STOI metric (ΔMB-STOI) [266] from that of the MB-STOI of the input binaural signal, $[\mathrm{y}_{\mathrm{a1,L}}\ \mathrm{y}_{\mathrm{a1,R}}]^T$, and change in binaural cue errors, ΔITD and ΔILD for both the target speech and the noise from the respective reference binaural signals, using the auditory model from [267]. Once again, due to the long reverberation time of 1.5 s in the KU Leuven ESAT-STADIUS audio laboratory, the speech component of the input binaural signal were very reverberant and was not an appropriate reference signal for computing the MB-STOI, the speech ITD, and the speech ILD. Consequently, it was decided that the reference binaural signal for computing these metrics would be the speech signal of the target speaker convolved with the time domain impulse responses (direct component representation) used to define $\widetilde{\mathbf{h}}_{\mathbf{a},\mathrm{L}}$ and $\widetilde{\mathbf{h}}_{\mathbf{a},\mathrm{R}}$. The reference binaural signal for computing the ITD and ILD for the noise however, consisted of the noise components of the input binaural signal, $[\mathrm{n}_{\mathrm{a1,L}}\ \mathrm{n}_{\mathrm{a1,R}}]^T$. All metrics were then computed in 2 s time frames with a 50 % overlap. The total length of the noisy signal to be processed was 13 s.

Figure 6.6 displays the resulting performance of the BMVDR algorithms outlined in Table 6.1 when the SPP was computed using the speech-only component from the reference microphone signal, $\mathrm{y}_{\mathrm{a1,R}}$. This is representative of a case when $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$ are fairly accurate estimates and can be considered as a best-case scenario under which the algorithms will perform.

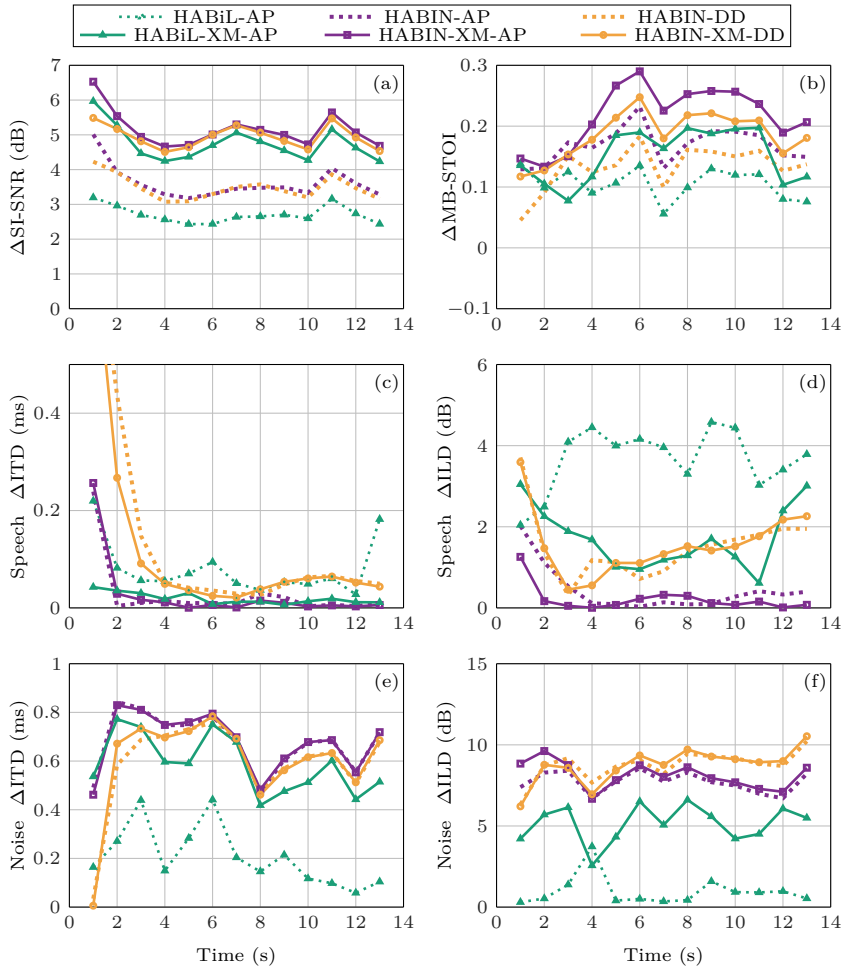


Figure 6.6: Performance of the various BMVDR algorithms outlined in Table 6.1 when the correlation matrices have been estimated fairly accurately.

Firstly, in terms of ΔSI-SNR, all of the algorithms that use the LMAs along with the XMs demonstrate an improvement over those that use the left and right ear LMA only. Additionally, the ΔSI-SNR of the HABIN-XM-AP and the HABIN-XM-DD indicate that both algorithms have a similar performance, which would be expected for the case of fairly accurately estimated correlation matrices. In terms of the ΔMB-STOI, after about 4 sec, once again the algorithms that use the LMAs along with the XMs demonstrate an improvement over their counterparts that use the LMA only. The HABIN-XM-AP is

indicative of an improvement over the HABiL-XM-AP and the HABIN-XM-DD, although informal listening suggested that they were quite similar.

Furthermore, for the HABIN-XM-AP and the HABIN-XM-DD algorithms, the ΔITD and ΔILD for both the target speech and the noise remain relatively unchanged between their LMA counterparts, i.e. HABIN-AP and HABIN-DD respectively. For these algorithms, it therefore suggests that the inclusion of the XMs does not contribute to any additional distortion in the binaural cues. The semi-bilateral algorithms on the other hand appear to be more sensitive to the XMs being used or not. In comparison to the HABiL-AP, when XMs are used in HABiL-XM-AP, ΔITD and ΔILD for the speech is seen to decrease, while the ΔITD and ΔILD for the noise is seen to increase. As expected, the ΔITD and ΔILD for the noise in the HABiL-AP are generally lower than all of the other algorithms, but this comes at the expense of a reduced intelligibility, and an increased speech ΔILD.

Figure 6.7 displays the resulting performance of the BMVDR algorithms outlined in Table 6.1 but now for when the SPP was computed using the actual noisy signal, $y_{a1,R}$. This is representative of a more realistic scenario since in this case the estimates, $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$, would have been less accurate.

Firstly, the impact of the less accurate $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$ is evident from the ΔSI-SNR and ΔMB-STOI metrics as the corresponding absolute values from all algorithms now indicate a lower performance in comparison to Fig. 6.6. In particular, it is the HABIN-DD and HABIN-XM-DD algorithms that have suffered degradation as they are completely dependent on the accuracy of the estimates $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$. On the other hand, the HABIN-XM-AP and HABiL-XM-AP maintain a fairly robust performance in terms of intelligibility.

It can also be observed that for the HABIN-XM-AP and the HABIN-XM-DD algorithms, the ΔITD and ΔILD for both the target speech and the noise remain relatively unchanged between their LMA counterparts, i.e. HABIN-AP and HABIN-DD respectively. The HABIN-AP and HABIN-XM-AP however exhibit lower ΔITD and ΔILD for the target speech as compared with the HABIN-DD and HABIN-XM-DD algorithms (although upon subjective listening, both appear to have the same spatial impression for the target speech). In terms of the binaural cues, HABiL-XM-AP and the HABiL-AP exhibit a similar relative performance as in Fig. 6.6 when the correlation matrices were more accurately estimated. For a subjective evaluation, the audio files from these experiments may be listened to at [95].

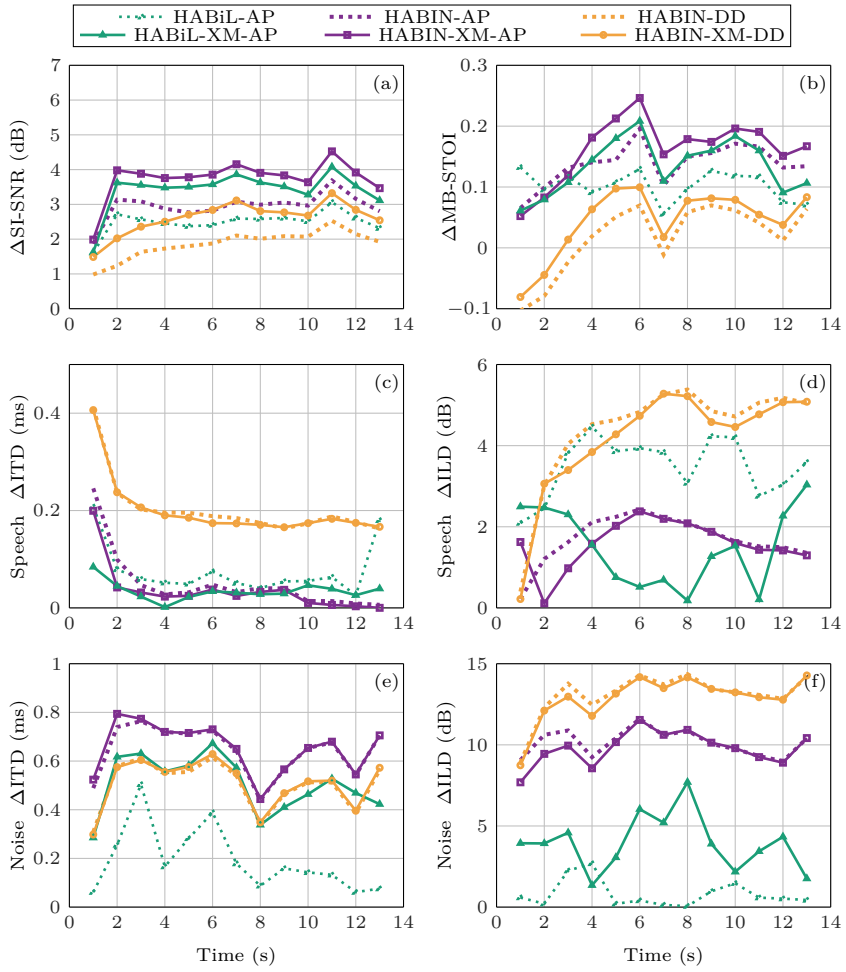


Figure 6.7: Performance of the various BMVDR algorithms outlined in Table 6.1 when the correlation matrices have been estimated less accurately than in Fig. 6.6.

6.6 Conclusion

In this chapter, the technique of incorporating partial spatial a priori knowledge of RTF vectors into a speech enhancement system involving XMs from chapter 5 has been extended to the case where there is available a priori knowledge from more than one LMA. Specifically, this has been considered within the context of designing a BMVDR beamformer for a binaural assistive hearing device that

consists of a separate LMA for the left ear and another LMA for the right ear, and has access to multiple XMs. As opposed to the conventional approach of estimating entire RTF vectors from the data, it was demonstrated that with the introduction of separate a partial a priori knowledge of each of the left and right ear LMAs, that only the missing part of the RTF vector corresponding to the XMs as well as an additional scalar needed to be estimated from the data.

Using the geometrically-based integration approach of using a priori spatial knowledge as developed in chapter 5, both a semi-bilateral and a binaural scheme was proposed and compared to the binaural scheme that uses a fully data-dependent RTF vector for all microphones. An evaluation scenario with a dummy head, two behind-the-ear dummy hearing aids, and XMs suggested that for the binaural processing schemes, the introduction of XMs improves intelligibility and does not further distort the binaural cues already preserved from using only the left and right ear LMAs. Additionally, when there is inaccurate estimation of the relevant spatial correlation matrices, more intelligibility improvement can be achieved when using a priori spatial knowledge for the left and right ear LMAs as opposed to relying on a fully data-dependent approach. For the semi-bilateral scheme that uses this a priori knowledge, the intelligibility also improves when XMs are introduced, however the binaural cues are different in comparison to using only the left and right hearing devices. Particularly a reduction in the speech ITD and ILD was observed, while there was an increase in the noise ITD and ILD. Informal listening has suggested that the binaural cues related to the speech source have been preserved in all algorithms.

The binaural cues related to the noise however remained distorted, which is a direct consequence of the use of a BMVDR beamformer. Nevertheless, it is known that reintroducing some of the signal from the reference microphones can assist in preservation of the binaural cues related to the noise, with the compromise of a lower achievable SNR output [251]. In fact, in [243] it has been shown that for a homogeneous noise field, when XMs are introduced into a system involving only LMAs, a better compromise between preservation of spatial cues related to the noise and achievable SNR can be attained.

As an extension to the work presented in this chapter, the scenario of an XM in motion should also be considered. In such a case, it is still unclear as to how well the binaural cues would be preserved and any associated trade-off with intelligibility. A preliminary investigation into this question has suggested that the binaural cues related to the speech can be preserved when an XM is in motion, but the resulting intelligibility is dependent on how fast the correlation matrices are updated in relation to the velocity of the XM. Audio and video samples from these preliminary experiments can be found in [95].

Moreover, while the results in this chapter are indicative of improvements

in performance, it is still not conclusive as there has not been an evaluation with hearing impaired users. In practice there are several factors which can hinder the performance of binaural processing schemes and degrade localisation such as differing characteristics (such as different gain settings) on each of the assistive hearing devices, the modification of spectral cues provided by the pinna due to the presence of the physical device, or effects of acclimatisation [257]. Consequently, future work should also be dedicated toward a subjective evaluation with binaural assistive hearing devices users.

Chapter 7

MWF-based speech dereverberation using a priori spatial knowledge and an external microphone

7.1 Introduction

Our focus shifts in this chapter[1] toward addressing the speech enhancement task of dereverberation. As was discussed in chapter 2, reverberation can be described as the persistence of a sound and is known to be detrimental to speech intelligibility [10, 268–270]. The task of speech dereverberation is therefore the process of removing the effects of the reverberation from a reverberant speech signal so as to restore an acceptable degree[2] of speech intelligibility and quality. Research into addressing this task has subsequently led to the development of an extensive range of multi-microphone speech dereverberation strategies [13], which make varying assumptions on signal modelling and the availability of a priori information.

[1] This chapter is partially based on the publication: Ali R., van Waterschoot T., Moonen M., "MWF-based speech dereverberation with a local microphone array and an external microphone", in *Proc. of the 27th European Signal Processing Conference (EUSIPCO)*, A Coruña, Spain, Sep. 2019.

[2] The degree to which a speech signal is acceptably dereverberated would be defined by a particular application.

For instance, when the room impulse response (RIR) is known between the desired source and the multiple microphones, i.e., when the acoustic channel is known, an inverse filter or channel equalisation can be applied to mitigate reverberation. If the RIRs do not share common zeros, then the multiple-input multiple-output (MINT) theorem [271] can provide perfect dereverberation if the RIRs are exactly known. Due to the sensitivity of the MINT method to RIR mismatch, however, its performance can degrade in practice as the RIRs are typically unknown and have to be estimated. A number of MINT-based methods have since been developed which intend to address this issue of robustness [272–275]. As opposed to relying on some knowledge of the RIRs, the strategy of Multi-Channel Linear Prediction (MCLP) can blindly estimate the dereverberated speech [276–279]. In MCLP, the reverberant component is modelled as a filtered version of the delayed microphone signals. After obtaining an estimate for this filter, then the modelled reverberant component can be subtracted from the signal of a reference microphone to yield the dereverberated speech estimate. It should come to no surprise to us that beamforming is another popular strategy for speech dereverberation. Since spatial filters, such as the MVDR beamformer "listens" in a specific direction and rejects interference from all other directions, it can be used for dereverberation albeit there is an inherent trade-off between the amount of noise reduction and dereverberation that can be achieved [280]. Nevertheless, a combined approach, along with a spectral filter can provide an improved performance in a Multi-Channel Wiener Filter (cfr. Chapter 2) framework [26, 281–284]. A final speech dereverberation strategy to be mentioned is that which makes use of a state-space framework [285, 286]. In these strategies, the RIRs are assumed to be time varying and unknown. By formulating an appropriate multichannel state-space representation for the system and using an expectation-maximisation scheme, an estimate for the dereverberated speech can be obtained.

Throughout this chapter, an MWF will be used as the speech dereverberation strategy, where the reverberant component is modelled as an additive distortion. Many of the speech dereverberation methods that make use of the MWF have considered a microphone system typically consisting of a local microphone array (LMA). As it has become more evident in previous chapters that an XM can be beneficial for improving speech intelligbility, the intent of this chapter is to investigate the performance of the MWF for speech dereverberation in cases when an XM[3] is available in addition to the LMA. In particular, the approach of making use of a priori spatial knowledge pertaining to the LMA as developed in Chapter 5 will be explored.

[3]Only one XM is considered in this chapter due to an assumption that will be made on the diffuse field coherence matrix in the following. Further suggestions on how to extend the proposed procedure for multiple XMs are discussed in the conclusion and left for future work.

When using an LMA, the MWF has shown to be effective in suppressing the reverberant component, provided that there are estimates of the relevant parameters, namely the speech and reverberant power spectral densities (PSDs), and the relative transfer function (RTF) vector pertaining to all of the microphones [26, 282–284]. In some speech dereverberation methods, namely blocking-based methods [26, 281, 283], a priori spatial knowledge of an RTF vector for the LMA has in fact been exploited. However, such methods cannot directly extend to the microphone configuration of an LMA and an XM due to the typically unknown position of the XM in relation to that of the LMA. It would then be necessary to obtain an estimate for the RTF component corresponding to the XM in order to complete the entire RTF vector. Therefore it can already be seen that the geometrically-based integration approach using a priori spatial knowledge as in Chapter 5 is directly applicable.

In [282], the MWF parameters were estimated by considering pre-whitened correlation matrices with a dimension equal to the number of microphones in the system. In this chapter, however, it will be demonstrated that by considering signals in a pre-whitened-transformed (PWT) domain, the relevant MWF parameters can be alternatively estimated by considering the eigenvalue decomposition (EVD) of a 2×2 matrix. This procedure makes direct use of the available a priori knowledge of the RTF vector for the LMA, and the assumption that the XM is sufficiently far from the LMA so that the reverberant component of the LMA signals is uncorrelated with the reverberant component of the XM signal. The latter of these assumptions was proposed in [225, 252] and used for speech enhancement tasks in diffuse noise fields involving an LMA and an XM. Whereas in previous chapters, the focus was on the estimation of the missing component of the RTF vector pertaining to the XM, this chapter also focuses on the estimation of the speech and reverberant PSDs that can be obtained when making use of the partial a priori knowledge.

The remainder of this chapter is organised as follows. In Section 7.2, the data model pertaining specifically to a case of speech dereverberation and the associated notation are defined, as well as a brief review of the MWF and the parameters to be estimated. In Section 7.3, a state of the art MWF-based speech dereverberation procedure is discussed. In Section 7.4, the proposed procedure that makes use of the a priori spatial knowledge of the RTF vector for the LMA is discussed, and a simple approximation is also suggested. In section 7.5, the various algorithms are evaluated using simulated as well as recorded audio data. Conclusions are drawn and future work is discussed in Section 7.6.

7.2 Data model

A reverberant environment consisting of an LMA of M_a microphones, one additional XM, and one target speaker is considered as in Figure 7.1. In the short-time Fourier transform (STFT) domain, the stacked vector of microphone signals at frequency bin k and time frame l, is modelled as:

$$\mathbf{y}(\mathrm{k},\mathrm{l}) = \underbrace{\mathbf{d}(\mathrm{k},\mathrm{l})\mathrm{s}_{1\mathrm{d}}(\mathrm{k},\mathrm{l})}_{\mathbf{x_d}(\mathrm{k},\mathrm{l})} + \mathbf{x_r}(\mathrm{k},\mathrm{l}) \tag{7.1}$$

where $\mathbf{y}(\mathrm{k},\mathrm{l}) = [\mathbf{y_a}^T(\mathrm{k},\mathrm{l})\ \mathrm{y_e}(\mathrm{k},\mathrm{l})]^T$, consists of the stacked LMA signals, $\mathbf{y_a}(\mathrm{k},\mathrm{l}) = [\mathrm{y}_1(\mathrm{k},\mathrm{l})\ \mathrm{y}_2(\mathrm{k},\mathrm{l})\ \ldots \mathrm{y}_{\mathrm{M_a}}(\mathrm{k},\mathrm{l})]^T$, and the XM signal, $\mathrm{y_e}(\mathrm{k},\mathrm{l})$. $\mathbf{x_d}(\mathrm{k},\mathrm{l})$ is the contribution from the direct component of the speech, represented by $\mathrm{s}_{1\mathrm{d}}(\mathrm{k},\mathrm{l})$, the direct component of the speech signal in the first microphone of the LMA (i.e. the target signal of interest), filtered with the direct RTF vector, $\mathbf{d}(\mathrm{k},\mathrm{l}) = [\mathbf{d_a}^T(\mathrm{k},\mathrm{l})\ \mathrm{d_e}(\mathrm{k},\mathrm{l})]^T$, consisting of the direct RTF vector for the LMA, $\mathbf{d_a}(\mathrm{k},\mathrm{l})$ (with the first microphone used as the reference, i.e. the first component of $\mathbf{d_a}(\mathrm{k},\mathrm{l})$ equal to 1), and the direct RTF component for the XM, $\mathrm{d_e}(\mathrm{k},\mathrm{l})$. Finally, $\mathbf{x_r}(\mathrm{k},\mathrm{l})$ is the reverberant component.

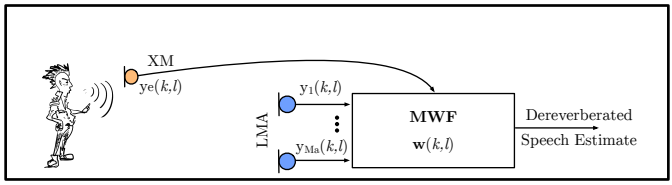


Figure 7.1: Acoustic scenario consisting of a target speaker, an LMA, and an XM.

It should be noted that the data model described in (7.1) is slightly different as opposed to that which was outlined in Chapter 2. In particular, the speech component only consists of the direct component upon which a diffuse field (instead of noise) is added. This model is in fact that of a reverberant suppression model and has been used to design successful speech dereverberation strategies [6]. From Chapter 2, however, we saw that the true acoustic field will also consist of early reflections of the speech component within the acoustic environment. Hence (7.1) is indeed more of an approximation to the true acoustic field and will be inherently limited. The benefit however, is that with such a simplified signal model, it will allow us to take advantage of spatial a priori knowledge, which can lead to a low cost solution with acceptable performance. In Sec. 7.5, the impact of the assumptions on the signal model will become evident as the proposed dereverberation procedure in this chapter will be evaluated

using signals that contain all direct, early, and reverberant components. On continuing with the signal model, it is once again assumed that all frequency bins can be treated independently and only the dependence on time will be retained for the parameters that are time-varying in the following derivations.

With the consideration of a single speaker in a fixed position, and modelling the reverberant field as spatially homogeneous (recall Sec. 2.7.1), the corresponding $(\mathrm{M_a}+1)\times(\mathrm{M_a}+1)$ correlation matrices for the microphone signals, $\mathbf{\Phi_y}(\mathrm{l})$, direct component of the speech, $\mathbf{\Phi_{x_d}}(\mathrm{l})$, and the reverberant component, $\mathbf{\Phi_{x_r}}(\mathrm{l})$, can be given respectively as:

$$\mathbf{\Phi_y}(\mathrm{l}) = \mathbb{E}\{\mathbf{y}(\mathrm{l})\mathbf{y}^H(\mathrm{l})\} \tag{7.2}$$

$$\mathbf{\Phi_{x_d}}(\mathrm{l}) = \mathbb{E}\{\mathbf{x_d}(\mathrm{l})\mathbf{x_d}^H(\mathrm{l})\} = \Phi_{\mathrm{s}}(\mathrm{l})\mathbf{d}\mathbf{d}^H \tag{7.3}$$

$$\mathbf{\Phi_{x_r}}(\mathrm{l}) = \mathbb{E}\{\mathbf{x_r}(\mathrm{l})\mathbf{x_r}^H(\mathrm{l})\} = \Phi_{\mathrm{r}}(\mathrm{l})\mathbf{\Gamma} \tag{7.4}$$

$\mathbf{\Phi_{x_d}}(\mathrm{l})$ is a rank-1 matrix, $\Phi_{\mathrm{s}}(\mathrm{l}) = \mathbb{E}\{|\mathrm{s_{1d}}(\mathrm{l})|^2\}$ is the time-varying power spectral density (PSD) of the target signal, $\Phi_{\mathrm{r}}(\mathrm{l})$ is the time-varying PSD of the reverberant component, and $\mathbf{\Gamma}$ is a time-invariant spatial coherence matrix. $\mathbf{d}$ and particularly, $\mathrm{d_e}$ is assumed time-invariant, however, the position of the XM still remains unknown with respect to the LMA. Assuming that the direct component of the speech is uncorrelated with the reverberant component, $\mathbf{\Phi_y}(\mathrm{l})$ can be expressed as:

$$\mathbf{\Phi_y}(\mathrm{l}) = \mathbf{\Phi_{x_d}}(\mathrm{l}) + \mathbf{\Phi_{x_r}}(\mathrm{l}) = \Phi_{\mathrm{s}}(\mathrm{l})\mathbf{d}\mathbf{d}^H + \Phi_{\mathrm{r}}(\mathrm{l})\mathbf{\Gamma} \tag{7.5}$$

It will also be assumed that there is a perfect communication link between the LMA and XM with no bandwidth constraints and synchronous sampling, so that the signal correlations can be estimated as if all signals were available in a centralised processor. The estimate of the target signal, $\hat{\mathrm{s}}_{1d}(\mathrm{l})$, i.e. the direct component of the speech in the first microphone of the LMA, is then obtained through the linear filtering of the microphone signals, such that $\hat{\mathrm{s}}_{1d}(\mathrm{l}) = \mathbf{w}^H(\mathrm{l})\mathbf{y}(\mathrm{l})$, where $\mathbf{w}(\mathrm{l}) = [\mathbf{w_a}^T(\mathrm{l})\ \mathrm{w_e}(\mathrm{l})]^T$. As discussed, an MWF will be used, which consists of an MVDR beamformer, followed by a single-channel post-filter:

$$\mathbf{w}(\mathrm{l}) = \underbrace{\frac{\mathbf{\Gamma}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{\Gamma}^{-1}\mathbf{d}}}_{\text{MVDR}}\ \underbrace{\frac{\Phi_{\mathrm{s}}(\mathrm{l})}{\Phi_{\mathrm{s}}(\mathrm{l}) + \Phi_{\mathrm{r}}(\mathrm{l})(\mathbf{d}^H\mathbf{\Gamma}^{-1}\mathbf{d})^{-1}}}_{\text{Single-Channel Post-Filter}} \tag{7.6}$$

Consequently, estimates are required for the parameters $\mathbf{d}$, $\mathbf{\Gamma}$, $\Phi_{\mathrm{s}}(\mathrm{l})$, and $\Phi_{\mathrm{r}}(\mathrm{l})$ in order to compute the MWF filter.

7.3 Data-dependent estimation of the MWF parameters

This section summarises the state-of-the-art methods for estimating the MWF parameters, $\mathbf{d}$, $\mathbf{\Gamma}$, $\Phi_{\mathrm{s}}(\mathrm{l})$, and $\Phi_{\mathrm{r}}(\mathrm{l})$. As such methods have only considered an LMA, they are also extended to include an XM.

Firstly, $\mathbf{\Gamma}$ can be modelled as a spherically diffuse coherence matrix, so that each element, $\gamma_{p,q}$, in the matrix can be computed as $\gamma_{p,q} = \mathrm{sinc}\,(\omega\, r_{pq}/c)$ [207], where ω is the angular frequency (rad/s), c is the speed of sound (m/s), and r_{pq} is the distance (m) between the p-th and q-th microphone (recall Sec. 2.7.1). Although the distance between the microphones in the LMA and the XM are unknown in practice, it can be assumed that the XM is far enough away from the LMA so that the reverberant component of the XM signal is uncorrelated with the reverberant component of the LMA signal [252]. An estimate for $\mathbf{\Gamma}$ in block matrix representation can then be given as [252]:

$$\hat{\mathbf{\Gamma}} = \left[\begin{array}{c|c} \hat{\mathbf{\Gamma}}_{\mathbf{a}} & \mathbf{0}_{\mathrm{M_a}\times 1} \\ \hline \mathbf{0}_{1\times \mathrm{M_a}} & 1 \end{array}\right] \tag{7.7}$$

where $\hat{\mathbf{\Gamma}}_{\mathbf{a}}$ is the $(\mathrm{M_a} \times \mathrm{M_a})$ diffuse field coherence matrix for the LMA, whose elements can be computed as the inter-microphone distances in an LMA are typically known. A spatial pre-whitening operation can then be defined by using the Cholesky decomposition:

$$\hat{\mathbf{\Gamma}} = \hat{\mathbf{\Gamma}}^{1/2}\, \hat{\mathbf{\Gamma}}^{H/2} \tag{7.8}$$

where $\hat{\mathbf{\Gamma}}^{1/2}$ is a lower triangular matrix. The MWF parameters can be estimated by using the pre-whitened correlation matrices in the optimization problem:

$$\underset{\Phi_{\mathrm{r}}(\mathrm{l}),\Phi_{\mathrm{s}}(\mathrm{l}),\,\mathbf{d}}{\text{minimise}}\; ||\hat{\mathbf{\Gamma}}^{-1/2}(\hat{\mathbf{\Phi}}_{\mathbf{y}}(\mathrm{l}) - \Phi_{\mathrm{r}}(\mathrm{l})\hat{\mathbf{\Gamma}} - \Phi_{\mathrm{s}}(\mathrm{l})\mathbf{d}\mathbf{d}^{H})\hat{\mathbf{\Gamma}}^{-H/2}||_F^2 \tag{7.9}$$

where $||.||_F$ is the Frobenius norm and $\hat{\mathbf{\Phi}}_{\mathbf{y}}(\mathrm{l})$ is the estimate of $\mathbf{\Phi}_{\mathbf{y}}(\mathrm{l})$ (for instance with recursive averaging [287]). Performing an eigenvalue decomposition (EVD) on the pre-whitened microphone signal PSD matrix, results in:

$$\hat{\mathbf{\Gamma}}^{-1/2}\hat{\mathbf{\Phi}}_{\mathbf{y}}(\mathrm{l})\hat{\mathbf{\Gamma}}^{-H/2} = \hat{\mathbf{D}}\mathbf{\Lambda}(\mathrm{l})\hat{\mathbf{D}}^{H} \tag{7.10}$$

where $\mathbf{\Lambda}(\mathrm{l}) = \mathrm{diag}\{\lambda_1(\mathrm{l}), \lambda_2(\mathrm{l}), \ldots, \lambda_{\mathrm{M_a}+1}(\mathrm{l})\}$ is a diagonal matrix of eigenvalues arranged in descending order and $\hat{\mathbf{D}}$ is a unitary matrix of the corresponding eigenvectors. As the Frobenius norm is invariant under a unitary transformation

[247], substituting (7.10) in (7.9) followed by a pre-multiplication with $\hat{\mathbf{D}}^H$, and a post-multiplication with $\hat{\mathbf{D}}$, results in:

$$\underset{\Phi_{\mathrm{r}}(\mathrm{l}),\Phi_{\mathrm{s}}(\mathrm{l}),\,\mathbf{d}}{\text{minimise}} ||\mathbf{\Lambda}(\mathrm{l}) - \Phi_{\mathrm{r}}(\mathrm{l})\mathbf{I}_{\mathrm{M_a}+1} - \hat{\mathbf{D}}^H\hat{\mathbf{\Gamma}}^{-1/2}\Phi_{\mathrm{s}}(\mathrm{l})\mathbf{d}\mathbf{d}^H\hat{\mathbf{\Gamma}}^{-H/2}\hat{\mathbf{D}}||_F^2 \tag{7.11}$$

where $\mathbf{I}_{\mathrm{M_a}+1}$ is the $(\mathrm{M_a}+1)\times(\mathrm{M_a}+1)$ identity matrix. The solution to (7.11) can be interpreted as the best approximation of $\mathbf{\Lambda}(\mathrm{l})$ by means of a sum of a scaled identity matrix and a rank-1 $(\mathrm{M_a}+1)\times(\mathrm{M_a}+1)$ matrix[4]. Firstly an estimate for $\mathbf{d}$ can be computed from the principal eigenvector of the matrix in (7.10), as contained in $\hat{\mathbf{D}}$ [25]:

$$\boxed{\hat{\mathbf{d}}^{\mathrm{pw}} = \frac{1}{\rho}\hat{\mathbf{\Gamma}}^{1/2}\hat{\mathbf{D}}\mathbf{e_1}} \tag{7.12}$$

where $\rho = \mathbf{e_1}^T\hat{\mathbf{\Gamma}}^{1/2}\hat{\mathbf{D}}\mathbf{e_1}$, and the $(\mathrm{M_a}+1)$ selection vector, $\mathbf{e_1} = [1\,0\ldots0]^T$. On replacing $\mathbf{d}$ with $\hat{\mathbf{d}}^{\mathrm{pw}}$ in (7.11) then gives:

$$\underset{\Phi_{\mathrm{r}}(\mathrm{l}),\Phi_{\mathrm{s}}(\mathrm{l})}{\text{minimise}} ||\mathbf{\Lambda}(\mathrm{l}) - \Phi_{\mathrm{r}}(\mathrm{l})\mathbf{I}_{\mathrm{M_a}+1} - \Phi_{\mathrm{s}}(\mathrm{l})\mathbf{\Lambda_x}||_F^2 \tag{7.13}$$

where $\mathbf{\Lambda_x} = \mathrm{diag}\{\frac{1}{|\rho|^2},\ 0,\ldots0\}$. An estimate for $\Phi_{\mathrm{r}}(\mathrm{l})$ follows by averaging the last $\mathrm{M_a}$ eigenvalues of $\mathbf{\Lambda}(\mathrm{l})$ [282][5]:

$$\boxed{\hat{\Phi}_{\mathrm{r}}^{\mathrm{pw}}(\mathrm{l}) = \frac{1}{\mathrm{M_a}}\big(\mathrm{trace}\{\mathbf{\Lambda}(\mathrm{l})\} - \lambda_1(\mathrm{l})\big)} \tag{7.14}$$

Finally, on replacing $\Phi_{\mathrm{r}}(\mathrm{l})$ with $\hat{\Phi}_{\mathrm{r}}^{\mathrm{pw}}(\mathrm{l})$ in (7.13) an estimate for $\Phi_{\mathrm{s}}(\mathrm{l})$ can be computed as [287]:

$$\boxed{\hat{\Phi}_{\mathrm{s}}^{\mathrm{pw}}(\mathrm{l}) = \big(\lambda_1(\mathrm{l}) - \hat{\Phi}_{\mathrm{r}}^{\mathrm{pw}}(\mathrm{l})\big)\,|\rho|^2} \tag{7.15}$$

7.4 Estimation of MWF parameters using a priori knowledge of the RTF vector for the LMA

In this section, as was done in previous chapters, the intention is to incorporate available spatial a priori knowledge of the RTF vector pertaining to the LMA

[4]It is worthwhile to acknowledge here the difference in the estimation problem of (7.11) and that of say (5.40) from Chapter 5. In (5.40), the entire noise correlation matrix, $\mathbf{R_{nn}}$, was pre-whitened to be an identity matrix, whereas in (7.11), only $\hat{\mathbf{\Gamma}}$ has been pre-whitened to be an identity matrix, leaving the time-varying reverberant PSD (or noise PSD if it were a diffuse noise field) to be estimated in addition to the other parameters.

[5]This estimate is in fact equivalent to a maximum-likelihood based estimate from [26] when the true MWF parameters, $\mathbf{d}$, $\mathbf{\Gamma}$, and $\Phi_{\mathrm{r}}(\mathrm{l})$ are known.

for the estimation of the relevant MWF parameters. In the case of speech dereverberation, it would be the direct RTF vector for the LMA that could be practically available[6].

The available a priori direct speech component and the associated correlation matrix is firstly re-defined respectively as:

$$\widetilde{\mathbf{d}} = [\widetilde{\mathbf{d}}_{\mathbf{a}}^{T} \;\; \mathrm{d_e}]^T; \quad \widetilde{\mathbf{\Phi}}_{\mathbf{x_d}}(\mathrm{l}) = \Phi_{\mathrm{s}}(\mathrm{l})\widetilde{\mathbf{d}}\widetilde{\mathbf{d}}^H \tag{7.16}$$

where the $\widetilde{\mathbf{d}}_{\mathbf{a}}$ is the known direct RTF vector for the LMA. Therefore only an estimate is required for $\mathrm{d_e}$ as opposed to the entire RTF vector as in Section 7.3. Following from the approach outlined in Chapter 5, a transformation matrix can be defined such that:

$$\widetilde{\mathbf{\Upsilon}} = \left[\begin{array}{c|c} [\widetilde{\mathbf{C}}_{\mathbf{a}} \;\; \tilde{\mathbf{f}}_{\mathbf{a}}] & \mathbf{0}_{\mathrm{M_a}\times 1} \\ \hline \mathbf{0}_{1\times \mathrm{M_a}} & 1 \end{array}\right] \tag{7.17}$$

where the $\mathrm{M_a} \times (\mathrm{M_a} - 1)$ blocking matrix, $\widetilde{\mathbf{C}}_{\mathbf{a}}$, and $\mathrm{M_a} \times 1$ fixed beamformer, $\tilde{\mathbf{f}}_{\mathbf{a}}$, are defined such that:

$$\widetilde{\mathbf{C}}_{\mathbf{a}}^H\widetilde{\mathbf{d}}_{\mathbf{a}} = \mathbf{0}_{(\mathrm{M_a}-1)\times 1}; \qquad \tilde{\mathbf{f}}_{\mathbf{a}}^H\widetilde{\mathbf{d}}_{\mathbf{a}} = 1 \tag{7.18}$$

A transformed version of the microphone signals is therefore:

$$\widetilde{\mathbf{\Upsilon}}^H\mathbf{y}(\mathrm{l}) = \left[(\widetilde{\mathbf{C}}_{\mathbf{a}}^H\mathbf{y}_{\mathbf{a}}(\mathrm{l}))^T \;\; \tilde{\mathbf{f}}_{\mathbf{a}}^H\mathbf{y}_{\mathbf{a}}(\mathrm{l}) \;\; \mathrm{y_e}(\mathrm{l})\right]^T \tag{7.19}$$

consisting of the blocking matrix signals from the LMA, the fixed beamformer output signal, and the XM signal. A new spatial pre-whitening operator, $\hat{\mathbf{L}}$, can then be defined from the transformed spatial coherence matrix:

$$\widetilde{\mathbf{\Upsilon}}^H\hat{\mathbf{\Gamma}}\widetilde{\mathbf{\Upsilon}} = \hat{\mathbf{L}}\hat{\mathbf{L}}^H \tag{7.20}$$

where $\hat{\mathbf{L}}$ is lower triangular and can be factorised as:

$$\hat{\mathbf{L}} = \widetilde{\mathbf{\Upsilon}}^H\hat{\mathbf{\Gamma}}^{1/2}\Theta \tag{7.21}$$

for some unitary matrix, Θ. In fact, since the reverberant component of the XM signal is assumed to be uncorrelated with the reverberant component of the LMA signals, the last row of $\hat{\mathbf{L}}$ consists of only zeros except for a one in

[6]This is precisely the reason for using the signal model of (7.1), i.e. modelling the speech component as only due to the direct RTF vector. The signal model could have certainly contained an additional component for the early reflections, however it would be less unlikely that a priori knowledge for the early reflections would be available in practice as this would be dependent on the acoustic environment and relative positions of the target speaker and microphones.

the last entry. After some rearranging and following of similar steps from Sec. 5.5.3, (7.9) can eventually be re-written as:

$$\underset{\Phi_{\mathrm{r}}(\mathrm{l}),\Phi_{\mathrm{s}}(\mathrm{l}),\,\mathrm{d_e}}{\text{minimise}} \; ||\boldsymbol{\Omega}^H\hat{\boldsymbol{\Phi}}_{\mathbf{y}}(\mathrm{l})\boldsymbol{\Omega} - \Phi_{\mathrm{r}}(\mathrm{l})\mathbf{I}_{\mathrm{M_a}+1} - \boldsymbol{\Omega}^H\Phi_{\mathrm{s}}(\mathrm{l})\widetilde{\mathbf{d}}\widetilde{\mathbf{d}}^H\boldsymbol{\Omega}||_F^2 \tag{7.22}$$

where $\boldsymbol{\Omega}^H = \hat{\mathbf{L}}^{-1}\widetilde{\boldsymbol{\Upsilon}}^H$. Signals consequently multiplied by $\boldsymbol{\Omega}^H$ are considered to be in a pre-whitened-transformed (PWT) domain. As a consequence of this operation, the last term in (7.22) is all zeros except for the bottom-right 2×2 block. Hence the EVD of the 2×2 matrix is considered:

$$\mathbf{J}^T\boldsymbol{\Omega}^H\hat{\boldsymbol{\Phi}}_{\mathbf{y}}(\mathrm{l})\boldsymbol{\Omega}\mathbf{J} = \underline{\underline{\mathbf{D}}}\,\underline{\underline{\boldsymbol{\Lambda}}}(\mathrm{l})\underline{\underline{\mathbf{D}}}^H \tag{7.23}$$

where $\mathbf{J} = [\,\mathbf{0}_{2\times(\mathrm{M_a}-1)} \mid \mathbf{I}_2\,]^T$, $\underline{\underline{\mathbf{D}}}$ is a 2×2 unitary matrix of eigenvectors and $\underline{\underline{\boldsymbol{\Lambda}}}(\mathrm{l}) = \mathrm{diag}\{\underline{\underline{\lambda}}_1(\mathrm{l}), \underline{\underline{\lambda}}_2(\mathrm{l})\}$ is a diagonal matrix of eigenvalues arranged in descending order. The notation of the double underbar, i.e. $\{\underline{\underline{.}}\}$ indicates that the signals are of a lower dimension and are in the PWT domain.

Applying a unitary transform to (7.22) with the block diagonal matrix, $\mathbf{G} = \mathrm{blkdiag}\{\mathbf{I}_{\mathrm{M_a}-1}, \underline{\underline{\mathbf{D}}}\}$, (blkdiag{.} is an operator that creates a block diagonal matrix from its arguments), then results in:

$$\underset{\Phi_{\mathrm{r}}(\mathrm{l}),\Phi_{\mathrm{s}}(\mathrm{l}),\,\mathrm{d_e}}{\text{minimise}} \; || \left[\begin{array}{c|c} \mathbf{P_{11}} & \mathbf{P_{12}} \\ \hline \mathbf{P_{21}} & \underline{\underline{\boldsymbol{\Lambda}}}(\mathrm{l}) \end{array}\right] - \Phi_{\mathrm{r}}(\mathrm{l}) \left[\begin{array}{c|c} \mathbf{I}_{\mathrm{M_a}-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I}_2 \end{array}\right] - \mathbf{G}^H\boldsymbol{\Omega}^H\Phi_{\mathrm{s}}(\mathrm{l})\widetilde{\mathbf{d}}\widetilde{\mathbf{d}}^H\boldsymbol{\Omega}\mathbf{G}||_F^2 \tag{7.24}$$

where only the bottom 2×2 block is diagonalised from the first term in (7.22) and $\mathbf{P_{11}}, \mathbf{P_{12}}$, and $\mathbf{P_{21}}$ are the residual matrices. The solution to (7.24) is now the best approximation of the first term in (7.24) by means of a sum of a scaled identity matrix and a rank-1 2×2 matrix, which is in contrast to the rank-1 $(\mathrm{M_a}+1) \times (\mathrm{M_a}+1)$ matrix required to solve (7.11). From (7.23), an estimate for $\mathrm{d_e}$ then follows as the last element from:

$$\hat{\widetilde{\mathbf{d}}} = \begin{bmatrix} \widetilde{\mathbf{d}}_{\mathbf{a}}^T \\ \hat{\mathrm{d}}_{\mathrm{e}}^{\mathrm{pwt}} \end{bmatrix} = \frac{1}{\zeta}\,\boldsymbol{\Omega}^{-H}\mathbf{J}\,\underline{\underline{\mathbf{D}}}\,\underline{\underline{\mathbf{e}}}_{\mathbf{1}} \tag{7.25}$$

where $\zeta = \mathbf{e_1}^T\boldsymbol{\Omega}^{-H}\mathbf{J}\,\underline{\underline{\mathbf{D}}}\,\underline{\underline{\mathbf{e}}}_{\mathbf{1}}$, and the 2-element selection vector, $\underline{\underline{\mathbf{e}}}_{\mathbf{1}} = [1,\, 0]^T$. Substitution of (7.25) for $\widetilde{\mathbf{d}}$ in (7.24) eventually results in:

$$\underset{\Phi_{\mathrm{r}}(\mathrm{l}),\Phi_{\mathrm{s}}(\mathrm{l})}{\text{minimise}} \; || \left[\begin{array}{c|c} \mathbf{P_{11}} & \mathbf{P_{12}} \\ \hline \mathbf{P_{21}} & \underline{\underline{\boldsymbol{\Lambda}}}(\mathrm{l}) \end{array}\right] - \Phi_{\mathrm{r}}(\mathrm{l}) \left[\begin{array}{c|c} \mathbf{I}_{\mathrm{M_a}-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I}_2 \end{array}\right] - \Phi_{\mathrm{s}}(\mathrm{l}) \left[\begin{array}{c|c} \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \underline{\underline{\boldsymbol{\Lambda}}}_{\mathbf{x}} \end{array}\right] ||_F^2 \tag{7.26}$$

where $\underline{\underline{\boldsymbol{\Lambda}}}_{\mathbf{x}} = \mathrm{diag}\{\frac{1}{|\zeta|^2}, 0\}$. Similar to (7.13), it is once again the diagonal elements which contribute to the solution of (7.26). Estimates for $\Phi_{\mathrm{r}}(\mathrm{l})$ and

$\Phi_{\mathrm{s}}(\mathrm{l})$ then follow similarly to those in Sec. 7.3:

$$\boxed{\hat{\Phi}_{\mathrm{r}}^{\mathrm{pwt}}(\mathrm{l}) = \frac{1}{\mathrm{M_a}}\left(\mathrm{trace}\{\hat{\mathbf{P}}(\mathrm{l})\} - \underline{\underline{\lambda}}_1(\mathrm{l})\right)} \tag{7.28}$$

$$\boxed{\hat{\Phi}_{\mathrm{s}}^{\mathrm{pwt}}(\mathrm{l}) = (\underline{\underline{\lambda}}_1(\mathrm{l}) - \hat{\Phi}_{\mathrm{r}}^{\mathrm{pwt}}(\mathrm{l}))\,|\zeta|^2} \tag{7.29}$$

where $\hat{\mathbf{P}}(\mathrm{l}) = \mathbf{G}^H\mathbf{\Omega}^H\hat{\mathbf{\Phi}}_{\mathbf{y}}(\mathrm{l})\mathbf{\Omega}\mathbf{G}$, i.e., the first term of (7.26).

Alternative estimates for $\Phi_{\mathrm{r}}(\mathrm{l})$ and $\Phi_{\mathrm{s}}(\mathrm{l})$ may also be considered by approximating (7.26) with its lower 2×2 blocks only, i.e. by solving the following problem:

$$\underset{\Phi_{\mathrm{r}}(\mathrm{l}),\Phi_{\mathrm{s}}(\mathrm{l})}{\text{minimise}}\ ||\underline{\underline{\mathbf{\Lambda}}}(\mathrm{l}) - \Phi_{\mathrm{r}}(\mathrm{l})\mathbf{I}_2 - \Phi_{\mathrm{s}}(\mathrm{l})\underline{\underline{\mathbf{\Lambda}}}_{\mathbf{x}}||_F^2 \tag{7.30}$$

Estimates for $\Phi_{\mathrm{r}}(\mathrm{l})$ and $\Phi_{\mathrm{s}}(\mathrm{l})$ would then follow as:

$$\boxed{\hat{\Phi}_{\mathrm{r}}^{\mathrm{pwt,22}}(\mathrm{l}) = \mathrm{trace}\{\underline{\underline{\mathbf{\Lambda}}}(\mathrm{l})\} - \underline{\underline{\lambda}}_1(\mathrm{l}) = \underline{\underline{\lambda}}_2(\mathrm{l})} \tag{7.31}$$

$$\boxed{\hat{\Phi}_{\mathrm{s}}^{\mathrm{pwt,22}}(\mathrm{l}) = (\underline{\underline{\lambda}}_1(\mathrm{l}) - \hat{\Phi}_{\mathrm{r}}^{\mathrm{pwt,22}}(\mathrm{l}))\,|\zeta|^2} \tag{7.32}$$

where the estimate for $\Phi_{\mathrm{r}}(\mathrm{l})$ is not anymore an average of the diagonal elements of $\hat{\mathbf{P}}(\mathrm{l})$. The advantage here is that it is not necessary to compute $\hat{\mathbf{P}}(\mathrm{l})$, but then it is only an approximation to the original problem of (7.22).

In terms of complexity of this approach, an EVD is performed on a 2×2 matrix as opposed to a $(\mathrm{M_a}+1) \times (\mathrm{M_a}+1)$ matrix, but the operation, $\mathbf{\Omega}^H$ still remains to be computed. However, as $\hat{\mathbf{L}}$, $\widetilde{\mathbf{\Upsilon}}$, and $\hat{\mathbf{\Gamma}}$ are all data-independent, $\mathbf{\Omega}^H$ can be pre-computed and multiplied with the microphone signal vector as a pre-processing stage. It is then the last two elements of this pre-processed vector which can be used to construct the 2×2 matrix on the left hand side of (7.23).

7.5 Evaluation

7.5.1 Simulated Data

The dereverberation algorithms were firstly evaluated in a simulated acoustic scenario as depicted in Fig. 7.2. It consisted of a linear LMA with five omnidirectional microphones separated by 8 cm and an end-fire positioned speech source 1.8 m from the centre microphone of the LMA in a room of

dimensions 6.3 m × 5.1 m × 2.5 m with a reverberation time of 600 ms. Also depicted in Fig. 7.2 is what will be referred to as the critical distance boundary, which is the circumference around the speech source, whose radius is the critical distance. Recalling (2.27) from Sec. 2.2.2 in Chapter 2, the critical distance for this acoustic scenario for an omnidirectional source is 0.66 m. In addition to the LMA and the speech source, XMs were also placed along a line from the LMA toward the speech source in a 5° trajectory, some of which within the critical distance. The speech source consisted of seven sentences from the hearing in noise test (HINT) database [205]. The room impulse responses from the speech source to each of the microphones were obtained using the randomised image method [208] and implemented from [209]. These impulse responses were consequently convolved with the speech source in order to simulate the respective microphone signal.

For the evaluation of the algorithms, only one of the XMs depicted in Fig. 7.2 was used along with the LMA, so that a total of $M_a + 1$ microphones were used. The effect of the XM placement with respect to the speech and the LMA would then be able to be observed. This would also implicitly be indicative of how the direct-to-reverberant ratio (DRR) at a particular XM position would affect the performance of the algorithms. All simulations were performed using the weighted overlap and add (WOLA) method [101], with a discrete Fourier transform (DFT) size of 512, a square-root hanning window, 50% overlap, and sampling frequency of 16 kHz. $\hat{\mathbf{\Phi}}_{\mathbf{y}}$ was estimated using recursive averaging with a time constant of 500 ms. Following the procedure as described in Sec. 5.6.2 of Chapter 5, the a priori direct-path RTF vector, $\widetilde{\mathbf{d}}_{\mathbf{a}}$, was computed from the DFT of 128 samples of the room impulse responses from the target speech to each of microphones in the LMA. Using this definition of $\widetilde{\mathbf{d}}_{\mathbf{a}}$, $\widetilde{\mathbf{C}}_{\mathbf{a}}$, and $\tilde{\mathbf{f}}_{\mathbf{a}}$ were defined accordingly from (7.18).

Table 7.1 summarises the list of dereverberation algorithms evaluated and the estimates used for the direct RTF vector, $\Phi_s(l)$, and $\Phi_r(l)$ for the MWF filter. LMA-XM-DD uses the procedure from Section 7.3, LMA-XM-AP uses the procedure involving the 2×2 matrices from Section 7.4, and LMA-XM-AP-22 is the approximation to LMA-XM-AP considering only the diagonalised matrices from (7.26). Processing with the LMA only and the unprocessed XM signal were also included as benchmarks against which processing with both LMA signals and the XM signal together could be compared.

Figure 7.3 displays the results of the various algorithms outlined in Table 7.1 as function of which XM was used. The XM position number on the x-axis corresponds to the XM number as indicated on Fig. 7.2. The metrics used for evaluation were the STOI [155], Cepstral Distance (CD) [288], and the fwSNRseg [288] with p=1. The reference signal used for computation of these metrics was the speech signal of the target speaker convolved with the time

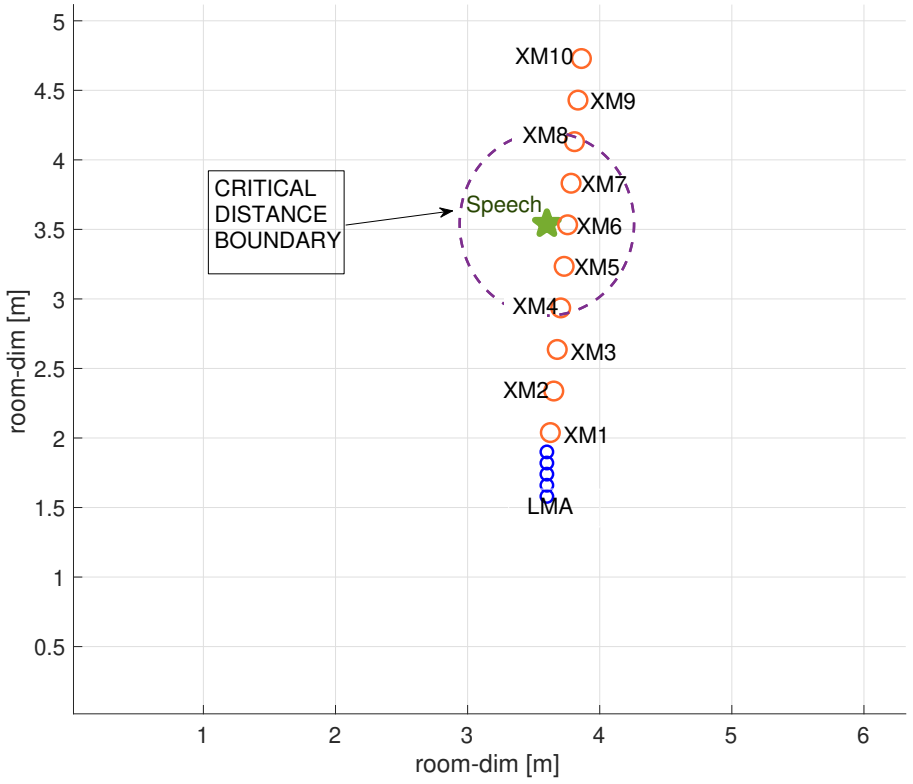


Figure 7.2: Simulated acoustic scenario with an LMA, a speech source, and several XMs placed at varying distances from the LMA and the speech source. The dereverberation algorithms were evaluated using the LMA and one of these XMs. The critical distance for the scenario was 0.66 m and defines the radius of the critical distance boundary (circumference) which is also depicted.

domain direct part of the room impulse response used to define the first element of $\tilde{\mathbf{d}}_{\mathbf{a}}$, i.e. corresponding direct component of the speech signal in the uppermost microphone of the LMA.

For the unprocessed, uppermost microphone (reference microphone) of the LMA, the values of the metrics were as follows: STOI = 0.62, CD = 4.6, and fwSNRseg = 15 dB. In Fig. 7.3, the relative change (Δ) of the metrics from these values as computed on the reference microphone of the LMA is reported. It should be recalled that higher values of Δ-STOI, and Δ-fwSNRseg indicate an improvement, whereas lower values for Δ-CD indicate an improvement.

Table 7.1: Dereverberation algorithms evaluated with a microphone array and an XM.

Algorithm	Signals used	RTF vector	$\Phi_{\mathrm{r}}(\mathrm{l})$	$\Phi_{\mathrm{s}}(\mathrm{l})$
XM	XM	-	-	-
LMA-AP	LMA	$\widetilde{\mathbf{d}}_{\mathbf{a}}$	$\hat{\Phi}_{\mathrm{r}}^{\mathrm{pw}(1)}$	$\hat{\Phi}_{\mathrm{s}}^{\mathrm{pw}(1)}$
LMA-XM-DD	LMA+XM	$\hat{\mathbf{d}}^{\mathrm{pw}}$	$\hat{\Phi}_{\mathrm{r}}^{\mathrm{pw}}$	$\hat{\Phi}_{\mathrm{s}}^{\mathrm{pw}}$
LMA-XM-AP	LMA+XM	$[\widetilde{\mathbf{d}}_{\mathbf{a}}^{T} \;\; \hat{\mathrm{d}}_{\mathrm{e}}^{\mathrm{pwt}}]^{T}$	$\hat{\Phi}_{\mathrm{r}}^{\mathrm{pwt}}$	$\hat{\Phi}_{\mathrm{s}}^{\mathrm{pwt}}$
LMA-XM-AP-22	LMA+XM	$[\widetilde{\mathbf{d}}_{\mathbf{a}}^{T} \;\; \hat{\mathrm{d}}_{\mathrm{e}}^{\mathrm{pwt}}]^{T}$	$\hat{\Phi}_{\mathrm{r}}^{\mathrm{pwt,22}}$	$\hat{\Phi}_{\mathrm{s}}^{\mathrm{pwt,22}}$

[1] $\hat{\Phi}_{\mathrm{r}}^{\mathrm{pw}}$and $\hat{\Phi}_{\mathrm{s}}^{\mathrm{pw}}$for the LMA algorithm were modified accordingly using only the LMA signals (i.e. as per [282] and [287]).

The first observation from Fig. 7.3, as expected, is that the performance of using the LMA-AP is independent of the XM, and already exhibits an improvement over the reference microphone. In cases where the XM is further away from the speech source, the use of the LMA-AP is seen to even perform better than listening to the XM alone. The benefit of using an XM for processing is however clearly evident as it gets within the critical distance of the target speaker, i.e for XM4 to XM8 (recall Fig. 7.2). As indicated by all metrics, the LMA-XM-AP and LMA-XM-AP-22 algorithms always perform better than using the LMA alone, with a maximum improvement achieved at XM position 6. The Δ-STOI and the Δ-fwSNRseg metrics indicate that these two algorithms also perform consistently better than using the XM alone, even when the XM is close to the target speaker. Furthermore, it can be seen that the performance of the LMA-XM-AP-22 and the LMA-XM-AP are quite similar, suggesting that the LMA-XM-AP-22 may not be such a bad approximation.

The LMA-XM-DD also follows a similar pattern of performance, in that the performance improves when the XM is within the critical distance. The Δ-STOI and the Δ-fwSNRseg metrics indicate that the LMA-XM-AP and LMA-XM-AP-22 perform better, whereas the Δ-CD suggests the opposite. However, care must be taken when comparing LMA-XM-DD to the LMA-XM-AP or LMA-XM-AP-22 as they are in fact making estimates of slightly different parameters. For the LMA-XM-AP, the a priori information used specifically corresponds to the direct component of the target speech, and the metrics are computed with a reference signal also as a direct component of the target speech in the reference microphone. The LMA-XM-DD, on the other hand will compute an RTF vector from (7.12) that will correspond to a speech estimate that inevitably contains some of the early components of the speech. Therefore, it can be argued that the LMA-XM-DD is being unfairly compared as the reference signal being used

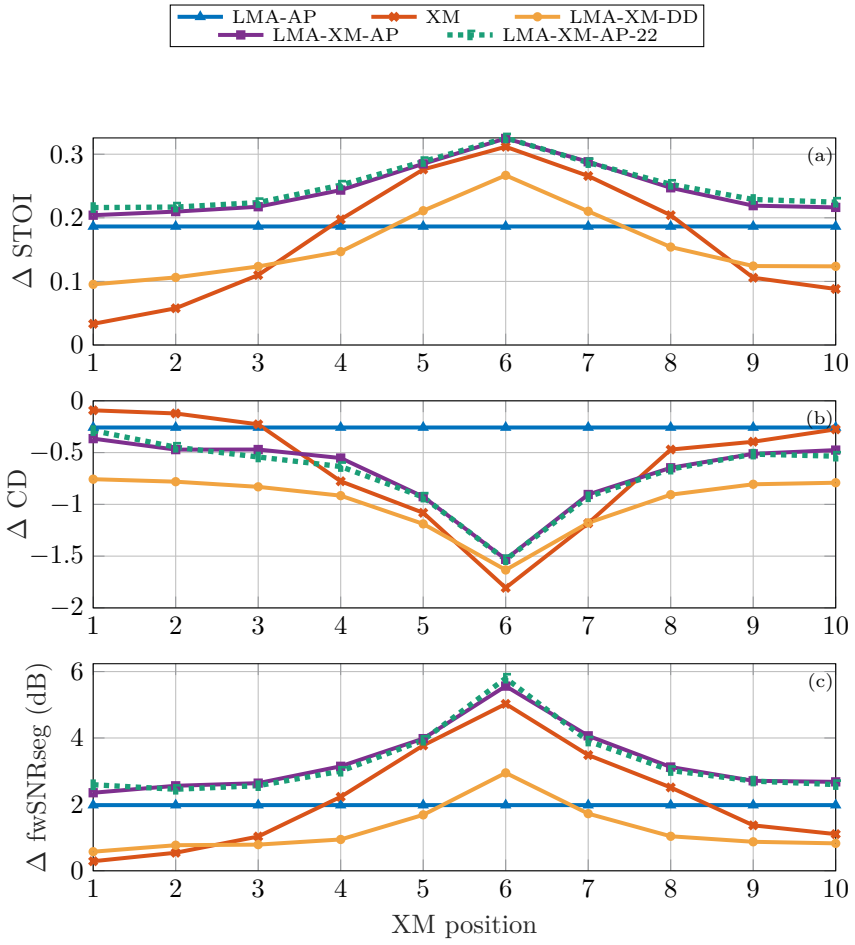


Figure 7.3: Performance of the various dereverberation algorithms outlined in Table 7.1. The XM position number on the x-axis corresponds to the XM number as indicated on Fig. 7.2.

is one that contains a direct component of the target speech. For this reason, it cannot be objectively said whether the LMA-XM-AP or LMA-XM-DD would yield better performance. Rather, it is important to understand that they are estimates of different parameters and each may be better suited depending on the application at hand.

Fig. 7.4 displays the resulting spectrograms for a section of the audio where two of the sentences were spoken for the case when XM-5 was used from Fig.

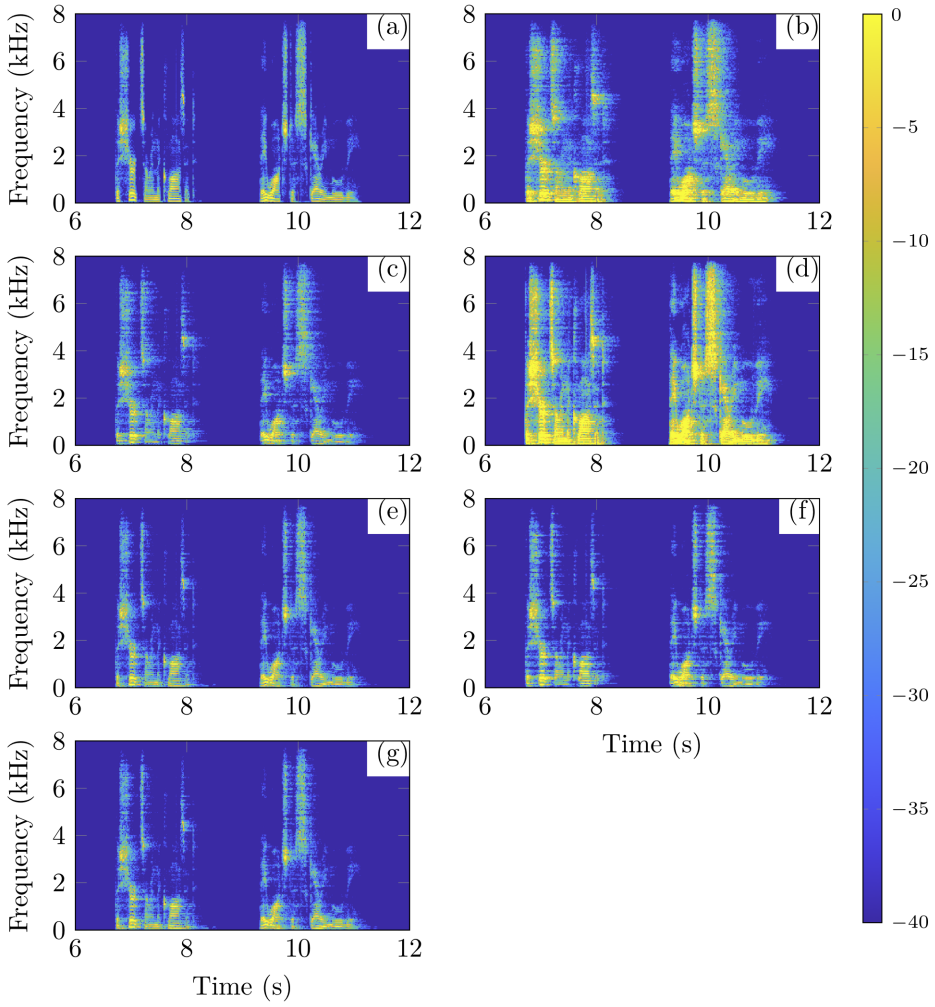


Figure 7.4: Spectrograms from two sentences using XM-5 that is within the critical distance. (a) Direct component of the target speech, (b) Unprocessed signal from the uppermost microphone of the LMA, (c) LMA-AP processed signal, (d) XM signal, (e) LMA-XM-AP processed signal, (f) LMA-XM-DD processed signal, (g) LMA-XM-AP-22 processed signal. The colour bar indicates the intensity of the STFT coefficients (in dB), which has been constrained to the range as shown for visual clarity (since the intensity of the XM STFT coefficients were substantially greater than those on the reference signal).

7.2. The reverberant nature of the unprocessed signal in Fig. 7.4 (b) can be clearly observed in comparison to that of the direct component from Fig. 7.4 (a). In Fig. 7.4 (d), it can also be seen that the XM signal exhibits some degree of reverberation, even though it is close to the target speaker. The improvement of the LMA-XM-AP and LMA-XM-AP-22 from Figs. 7.4 (e) and (g) respectively over LMA-AP in Fig. 7.4 (c) can also bee seen. Furthermore, it is evident from Fig. 7.4 (f) that the LMA-XM-DD appears to include some degree of reverberation or early components as compared to LMA-XM-AP or LMA-XM-AP-22. To have a better understanding of the differences between the LMA-XM-AP, LMA-XM-AP-22, and the LMA-XM-DD, audio samples from this simulation are provided in [95].

7.5.2 Recorded data from BTE hearing aid devices

The algorithms were also evaluated using recorded audio data from the KU Leuven ESAT-STADIUS audio laboratory, where the LMA was now a dummy behind-the-ear (BTE) hearing aid device with two microphones spaced approximately 1.3 cm apart. Recall that this room has a reverberation time of 1.5 s, and hence was quite appropriate for the evaluation of the dereverberation algorithms. The scenario for the recording is depicted in Fig. 7.5. A Neumann KU-100 dummy head was placed in a central location of the room and equipped with one BTE-HA on the left ear. A Genelec 8030C loudspeaker, was placed at 0° (the azimuth direction directly in front of the dummy head) and 1.2 m from the dummy head, and used to generate 5 s of a speech signal from a target female speaker [244]. The loudspeakers and the dummy head were placed at a height of approximately 1.25 m from the floor. Four XMs (AKG-CK-97-O) were placed at heights of 1.15 m from the floor and at increments of 10 cm away from the loudspeaker. In the evaluation of the algorithms, however, only one of these XMs was used. As in the previous simulation, the purpose of this evaluation would be to observe the effect of the position of the XM in relation to the source and the LMA (and hence also DRR) on the performance of the various algorithms. For this room, for an omnidirectional source, the critical distance for the audio lab is 0.45 m.

Once again, the procedure as described in Sec. 5.6.2 of Chapter 5, was used to obtain the a priori direct-path RTF vector, $\widetilde{\mathbf{d}}_{\mathbf{a}}$. Using this definition of $\widetilde{\mathbf{d}}_{\mathbf{a}}$, $\widetilde{\mathbf{C}}_{\mathbf{a}}$, and $\tilde{\mathbf{f}}_{\mathbf{a}}$ were defined accordingly from (7.18). All the processing of the algorithms was done using the WOLA method [101], with a Discrete Fourier Transform (DFT) size of 256 samples, a square-root hanning window, 50 % overlap, and a sampling frequency of 16 kHz. The metrics used to evaluate the following experiments were the same as described in the previous section. The reference signal used for computation of these metrics was the speech signal

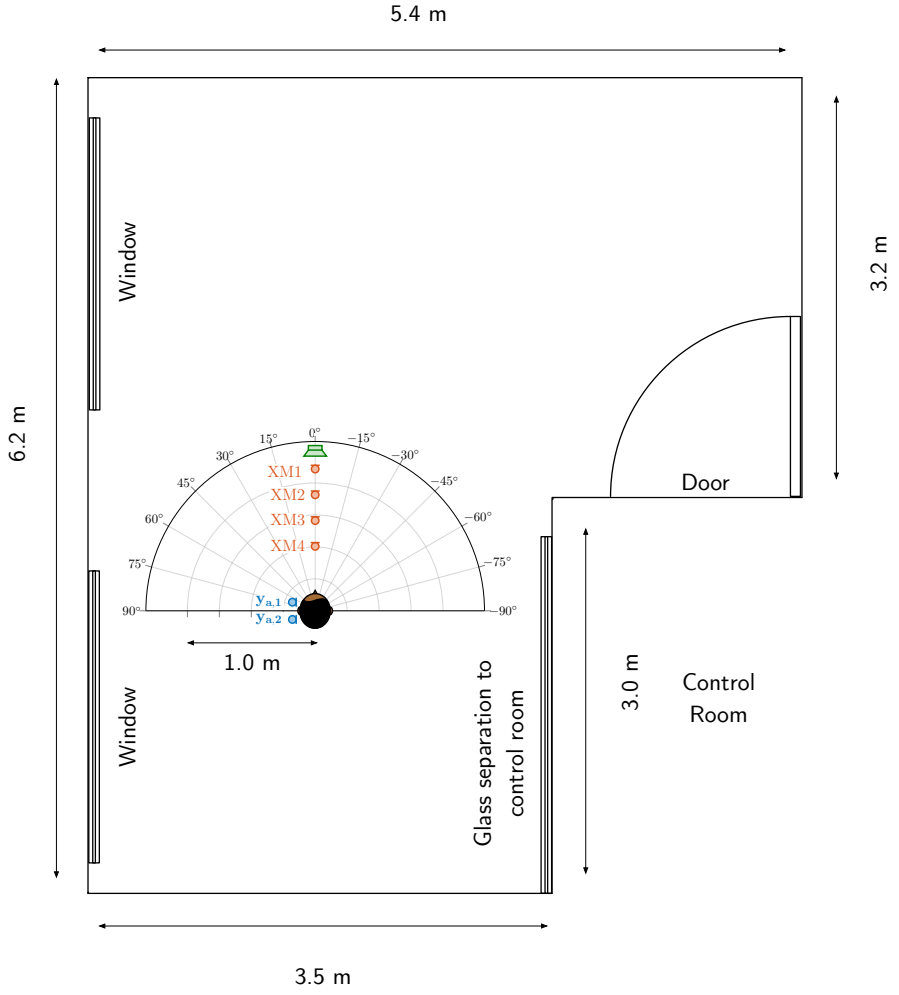


Figure 7.5: Acoustic scenario in the KU Leuven ESAT-STADIUS audio laboratory used to evaluate the dereverberation processing schemes.

of the target speaker convolved with the time domain direct part of the room impulse response used to define the first element of $\widetilde{\mathbf{d}}_{\mathbf{a}}$, i.e. corresponding direct component of the speech signal in the left-frontal microphone of the BTE-HA.

Table 7.2 summarises the list of dereverberation algorithms evaluated with the BTE HAs and the respective estimates used for the direct RTF vector, $\Phi_{\mathrm{s}}(l)$, and $\Phi_{\mathrm{r}}(l)$ for the MWF filter. HAL-XM-DD uses the procedure from Section

7.3, HAL-XM-AP uses the procedure involving the 2×2 matrices from Section 7.4, and HAL-XM-AP-22 is the approximation to HAL-XM-AP considering only the diagonalised matrices from (7.26).

Table 7.2: Dereverberation algorithms evaluated with BTE hearing aids and an XM.

Algorithm	Signals used	RTF vector	$\Phi_{\mathrm{r}}(\mathrm{l})$	$\Phi_{\mathrm{s}}(\mathrm{l})$
XM	XM	-	-	-
HAL-AP	Left ear mics + XM	$\widetilde{\mathbf{d}}_{\mathbf{a}}$	$\hat{\Phi}_{\mathrm{r}}^{\mathrm{pw}(1)}$	$\hat{\Phi}_{\mathrm{s}}^{\mathrm{pw}(1)}$
HAL-XM-DD	Left ear mics + XM	$\hat{\mathbf{d}}^{\mathrm{pw}}$	$\hat{\Phi}_{\mathrm{r}}^{\mathrm{pw}}$	$\hat{\Phi}_{\mathrm{s}}^{\mathrm{pw}}$
HAL-XM-AP	Left ear mics + XM	$[\widetilde{\mathbf{d}}_{\mathbf{a}}^{T} \;\; \hat{\mathrm{d}}_{\mathrm{e}}^{\mathrm{pwt}}]^{T}$	$\hat{\Phi}_{\mathrm{r}}^{\mathrm{pwt}}$	$\hat{\Phi}_{\mathrm{s}}^{\mathrm{pwt}}$
HAL-XM-AP-22	Left ear mics + XM	$[\widetilde{\mathbf{d}}_{\mathbf{a}}^{T} \;\; \hat{\mathrm{d}}_{\mathrm{e}}^{\mathrm{pwt}}]^{T}$	$\hat{\Phi}_{\mathrm{r}}^{\mathrm{pwt,22}}$	$\hat{\Phi}_{\mathrm{s}}^{\mathrm{pwt,22}}$

[1] $\hat{\Phi}_{\mathrm{r}}^{\mathrm{pw}}$and $\hat{\Phi}_{\mathrm{s}}^{\mathrm{pw}}$for the HAL-AP algorithm were modified accordingly using only the LMA signals (i.e. as per [282] and [287]).

Figure 7.6 displays the results of the various algorithms outlined in Table 7.2 as function of which XM from Fig. 7.2 was used. For the left-frontal microphone of the BTE-HA (reference microphone), the values of the metrics were as follows: STOI = 0.63, CD = 6.1, and fwSNRseg = 6.5 dB. In Fig. 7.6, the relative change (Δ) of the metrics from these reference values is reported.

The HAL-XM-AP and the HAL-XM-DD algorithms are firstly seen to have a similar performance across all metrics and exhibit a similar trend with respect to the XM position from the source. The HAL-XM-AP-22 also exhibits a quite similar performance to that of HAL-XM-AP, which is further evidence to suggest that the HAL-XM-AP-22 may serve as an appropriate low-cost derverberation procedure with the XM. For all of these algorithms, the best performance is achieved when the XM is closest to the source, but however drops off rapidly as the distance of the XM from the source increases, tending toward the performance of the HAL-AP. Although the simulation in the previous section suggested that such a drop off occurs at the critical distance, these results suggest that in practice, the drop off can occur even earlier. This may be due to a number of reasons including the imperfect modelling of the reverberant field as well as the noise present in the microphones. In terms of the Δ-STOI and Δ-CD metrics, the XM demonstrates an improvement for all positions over all the algorithms. However, the Δ-fwSNRseg indicates that it is worse than the all of the algorithms that use the XM. Certainly, the XM can offer an improvement in intelligibility and is an alternative option to any of the dereverberation algorithms. However, it should be recalled that listening to the XM only may not be desirable if the preservation of spatial cues are also important. Furthermore, in using all of the microphones as opposed to the single XM, additional noise reduction can be achieved as was observed in Chapter 5.

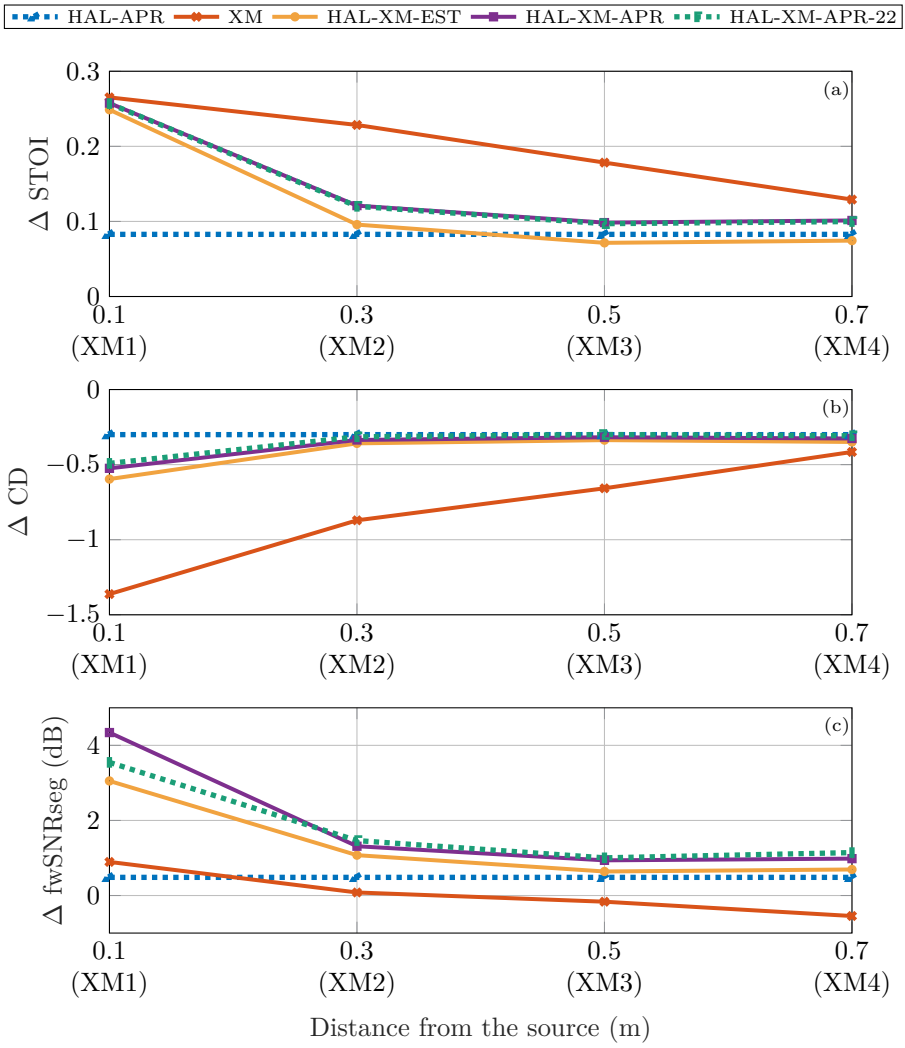


Figure 7.6: Performance of the various dereverberation algorithms outlined in Table 7.2 using the BTE HAs. The XM position number on the x-axis corresponds to the XM number as indicated on Fig. 7.5.

Fig. 7.7 displays the spectrograms for the reference speech signal (a), the reverberant microphone signal from the frontal-left microphone of the BTE-HA (b), and the various algorithms (d-g) for the case when XM1 (c) was used from Fig. 7.5. The excessive reverberation in the room can be clearly observed

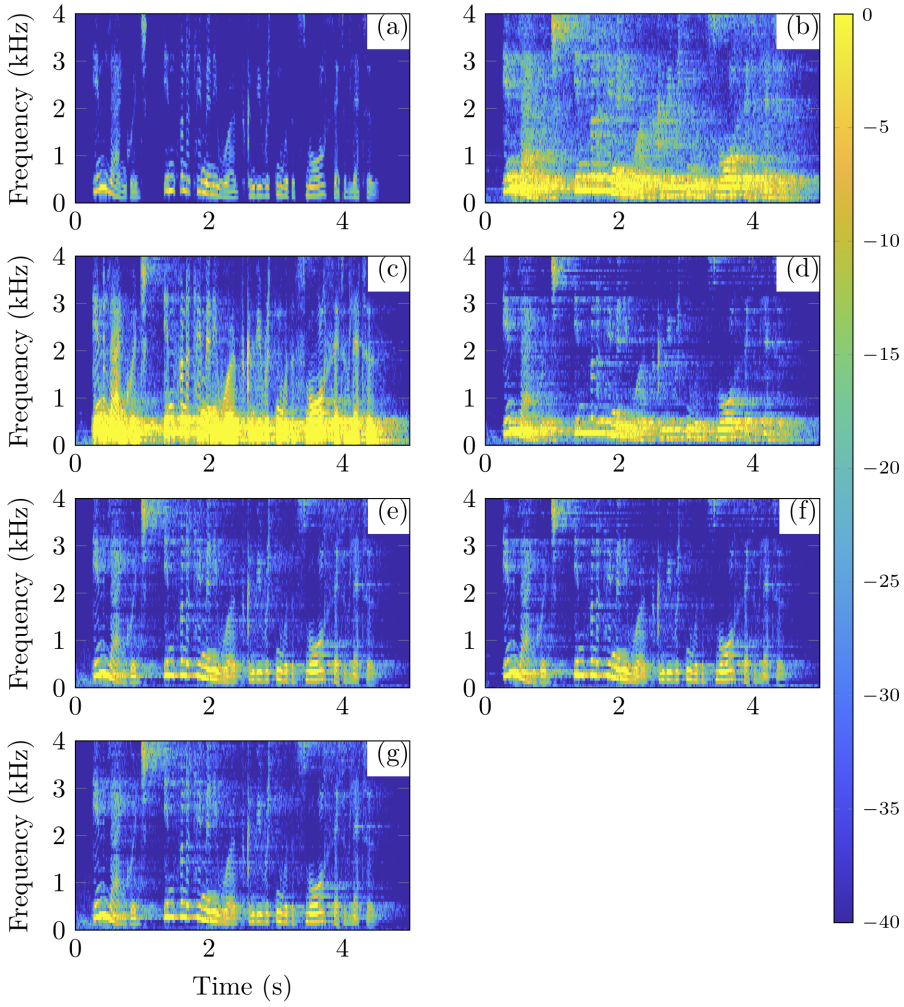


Figure 7.7: Spectrograms from a 4 second sample of the unprocessed and processed audio with the dereverberation algorithms when using XM1. (a) Direct component of the target speech, (b) Unprocessed signal from the frontal left-ear BTE HA, (c) XM1 (d) HAL-AP, (e) HAL-XM-AP, (f) HAL-XM-DD, (g) HAL-XM-AP22. The colour bar indicates the intensity of the STFT coefficients (in dB), which has been constrained to the range as shown for visual clarity (since the intensity of the XM STFT coefficients were substantially greater than those on the reference signal).

in the frontal left ear BTE HA in Fig. 7.7 (b) in comparison to that of the reference speech signal. Although XM1 has a very strong direct component, some reverberant components are also evident from Fig. 7.7 (c). The improvement of the HAL-XM-AP, HAL-XM-DD, HAL-XM-AP22 algorithms in Fig. 7.7 (e-g) over that of the HAL-AP (d) is also clearly observed. The HAL-XM-AP, HAL-XM-AP-22, and HAL-XM-DD are all seen to be quite similar. For a further intuition of these results, the audio samples are also provided in [95] for a subjective evaluation.

7.6 Conclusion

In this chapter, the technique of incorporating a priori spatial knowledge of an RTF vector for an LMA has been applied to the speech enhancement task of dereverberation within an MWF framework for a system of microphones consisting of an LMA and a single XM. In addition to estimating the missing part of the RTF vector corresponding to the XM as was done in previous chapters, estimates were also proposed for the corresponding speech and reverberant PSDs required for the MWF. It was demonstrated that by considering signals in a pre-whitened-transformed domain (PWT), all of the relevant parameters for the MWF could be obtained from a procedure involving an EVD on a 2×2 matrix as opposed to a $(\mathrm{M_a}+1)\times(\mathrm{M_a}+1)$ matrix. As the operator for obtaining signals in the PWT domain was data independent, it could also be pre-computed and multiplied with the microphone signal vector as a pre-processing stage, from which its last two elements can be used to form the 2×2 matrix. An evaluation using simulated data as well as recorded data involving BTE hearing aid devices has indicated that the proposed procedure can improve intelligibility compared to that achieved with using an LMA only, especially for the case when the XM is within the critical distance of the desired source.

In comparison to the previous chapters, this chapter has only considered one XM. This was directly due to the assumption made on the diffuse field coherence matrix in (7.7), i.e., the assumption regarding reverberant components in the XM and the LMA being uncorrelated. If more XMs are included in the microphone configuration and if the reverberant components of each of these signals as well as that of the LMA are uncorrelated, then the diffuse field coherence matrix can be trivially extended and the speech dereverberation procedure is the same apart from the increased dimensionality. In cases where this will not be a valid assumption, then the speech dereverberation procedure with additional XMs will be contingent on obtaining an appropriate model for the diffuse field coherence matrix.

It should be acknowledged that the price being paid for such a straightforward and practical speech dereverberation procedure is that there will be an upper bound on the best performance that the MWF can achieve due to the assumptions that have been placed on the signal model. In comparison to previous chapters, the signal model used in this chapter for speech dereverberation was somewhat more specific as it consisted of the sum of a direct-path speech component and a diffuse reverberant field. Firstly, this model does not take into account additional noise, however, if the noise was also diffuse, then this would not pose a problem. In such a case, as opposed to obtaining an estimate for the reverberant power, an estimate will be obtained for the sum of the reverberant power as well as the diffuse noise power, which can then be used in the MWF instead of the reverberant power alone. Secondly, and perhaps more critical, is that the signal model blatantly ignores the early reflections in the acoustic field. This means that there would be some correlation between the direct speech component and the reverberant component in the signal model used, which then puts a limit on how accurate the speech and reverberant and/or diffuse noise powers can be estimated. One potential modification that can be done to the signal model is to model the speech component as consisting of direct and early reflections. In this case however, it would then be necessary to include a priori information on the early reflections, which is not typically available in practice. Consequently, this modelling aspect of speech dereverberation is still an open issue and one for which further investigation should be dedicated towards.

Chapter 8

An integrated MVDR beamformer for a local microphone array linked with external microphones

8.1 Introduction

Having developed an approach for using available a priori spatial information for a local microphone array (LMA) that is linked with external microphones (XMs) in Chapter 5, an integrated version of the minimum variance distortionless response (MVDR) beamformer from Chapter 3 can finally be revisited for this microphone configuration. In this chapter[1], the integrated MVDR beamformer is formulated from an alternative perspective, namely that of a quadratically constrained quadratic program (QCQP), with two constraints, one of which is related to a maximum tolerable speech distortion for the imposition of an a priori relative transfer function (RTF) vector and the other related to a maximum tolerable speech distortion for the imposition of a fully data-dependent RTF vector. Unlike chapter 3 for the LMA only, when XMs are also used, the a priori RTF vector is based on partial a priori knowledge pertaining to the

[1] This chapter is partially based on the publication: Ali R., van Waterschoot T., Moonen M., "An integrated MVDR beamformer for speech enhancement using a local microphone array and external microphones", Internal Report 20-76, ESAT-STADIUS, KU Leuven, *submitted for publication*, 2020.

LMA and partial data-dependent estimation. Hence, an integrated MVDR beamformer that uses LMA and XMs will merge an a priori RTF vector that is based on partial a priori knowledge and a fully data-dependent RTF vector. In this regard, the integrated MVDR beamformer developed in this chapter can be viewed as combining both the confidence-based integration approach from Chapter 3 and the geometrically-based integration approach from Chapter 5 into a more global integration of the a priori and data-dependent spatial information.

The procedures for obtaining an a priori RTF vector and a data-dependent RTF vector will be briefly re-derived for a microphone configuration involving an LMA and XMs in a pre-whitened-transformed (PWT) domain, which consists of an initial transformation of the microphone signals through a blocking matrix and a fixed beamformer, followed by a pre-whitening operation. Whereas in Chapter 7, this pre-whitening operation was defined using only a spatial coherence matrix, an estimate of the noise-only correlation matrix is used in this chapter. With the a priori and data-dependent RTF vectors for the LMA and XMs estimated, it will become evident that the optimal filter from the integrated MVDR beamformer, formulated as a QCQP, is identical to that which was derived from Chapter 3, where the Lagrangian multipliers associated with the QCQP are equivalent to the tuning parameters that were considered in Chapter 3.

Through an analysis of how the maximum tolerable speech distortions affect the behaviour of the QCQP, it is revealed that the parameter space spanned by these distortions can be divided into four distinct regions for which a particular set of constraints are active or inactive. It will also be shown how this parameter space relates to the speech distortion weighted multi-channel Wiener filter (SDW-MWF) and that it also provides framework for optimally combining an SDW-MWF that relies on an a priori RTF vector with an SDW-MWF that relies on a data-dependent RTF vector. The advantage of the integrated MVDR beamformer over the SDW-MWF is therefore that it encapsulates a wider range of speech enhancement filters (including those of Chapter 3 or Chapter 5) that may be more accommodating for various acoustic scenarios.

By making use of this parameter space, a general tuning framework will be discussed, where values are prescribed for the the maximum tolerable speech distortions as opposed to making combinations of filters as was done in Chapter 3. The tuning framework, however still relies on a confidence metric which is indicative of the accuracy of the data-dependent RTF vector, for which a posterior SNR will once again be used. Some practical tuning strategies are also suggested.

An evaluation with experimental data involving a single speech source within

a re-created cocktail-party scenario for an LMA of hearing aid microphones and XMs is also discussed. The results support those of Chapter 3 in that the integrated MVDR beamformer can indeed be a compromise between relying only on either an a priori or data-dependent RTF vector, leading to a potentially more versatile performance in practice. More concretely, the integrated MVDR beamformer is observed to yield different enhanced speech signals depending on the tuning, which may be more suitable for consistently improving speech intelligibility despite changes in the source position and imperfectly estimated spatial correlation matrices.

The remainder of this chapter is organised as follows. In section 8.2, the data model and the PWT domain are defined. In section 8.3, the procedures for obtaining the a priori RTF vector based on partial a priori knowledge and the data-dependent RTF vector in the PWT domain are presented. Section 8.4 reformulates the integrated MVDR beamformer as a QCQP and provides an analysis on the effect of the maximum tolerable speech distortions due to the imposition of the a priori RTF vector and the data-dependent RTF vector. In section 8.5 a general tuning framework is presented, as well as some suggested tuning strategies. In section 8.6, the integrated MVDR approach is analysed and evaluated with both simulated data as well as experimental data involving the use of behind-the-ear hearing aid microphones and three XMs. Conclusions are drawn in section 8.7.

8.2 Data Model

8.2.1 Unprocessed Signals

The signal model in the STFT domain from section 4.3 of Chapter 4 is restated once again for convenience:

$$\mathbf{y}(\mathrm{k},\mathrm{l}) = \underbrace{\mathbf{h}(\mathrm{k},\mathrm{l})\mathrm{s}_{\mathrm{a},1}(\mathrm{k},\mathrm{l})}_{\mathbf{x}(\mathrm{k},\mathrm{l})} + \mathbf{n}(\mathrm{k},\mathrm{l}) \tag{8.1}$$

where (dropping the dependency on k and l for brevity) $\mathbf{y} = [\mathbf{y}_\mathbf{a}^T\ \mathbf{y}_\mathbf{e}^T]^T$, $\mathbf{y}_\mathbf{a} = [\mathrm{y}_{\mathrm{a},1}\ \mathrm{y}_{\mathrm{a},2}\ \cdots \mathrm{y}_{\mathrm{a},\mathrm{M_a}}]^T$ is a vector of the LMA signals, $\mathbf{y}_\mathbf{e} = [\mathrm{y}_{\mathrm{e},1}\ \mathrm{y}_{\mathrm{e},2}\ \cdots \mathrm{y}_{\mathrm{e},\mathrm{M_e}}]^T$ is a vector of the XM signals, $\mathbf{x}$ is the speech contribution, represented by $\mathrm{s}_{\mathrm{a},1}$, the desired speech signal in the first (reference) microphone of the LMA, filtered with $\mathbf{h} = [\mathbf{h}_\mathbf{a}^T\ \mathbf{h}_\mathbf{e}^T]^T$, $\mathbf{h}_\mathbf{a}$ is the RTF vector for the LMA (where the first component of $\mathbf{h}_\mathbf{a}$ is equal to 1 since the first microphone is used as the reference), and $\mathbf{h}_\mathbf{e}$ is the RTF vector for the XM signals. Finally, $\mathbf{n} = [\mathbf{n}_\mathbf{a}^T\ \mathbf{n}_\mathbf{e}^T]^T$ represents the noise contribution.

The $(M_a + M_e) \times (M_a + M_e)$ spatial correlation matrix for the speech-plus-noise, noise-only, and speech-only signals is defined respectively as:

$$\mathbf{R_{yy}} = \mathbb{E}\{\mathbf{y}\mathbf{y}^H\}; \;\; \mathbf{R_{nn}} = \mathbb{E}\{\mathbf{n}\mathbf{n}^H\}; \;\; \mathbf{R_{xx}} = \mathbb{E}\{\mathbf{x}\mathbf{x}^H\} \tag{8.2}$$

with the assumption that the speech signal is uncorrelated with the noise signal, and hence $\mathbf{R_{yy}} = \mathbf{R_{xx}} + \mathbf{R_{nn}}$. The estimate of the speech component in the first microphone of the LMA, z_1, is then obtained through the linear filtering of the microphone signals, such that:

$$z_1 = \mathbf{w}^H \mathbf{y} \tag{8.3}$$

where $\mathbf{w} = [\mathbf{w_a}^T \; \mathbf{w_e}^T]^T$ is a complex-valued filter.

8.2.2 Pre-whitened-transformed (PWT) domain

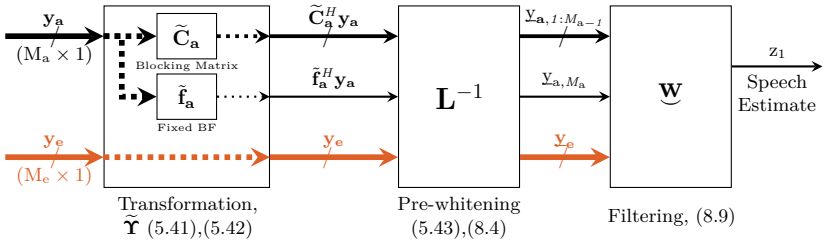


Figure 8.1: Signal processing flow for obtaining a speech estimate in the PWT domain.

As a pre-processing stage, the unprocessed microphone signals can be firstly transformed in accordance with an available a priori RTF vector for the LMA signals as in (5.41) and (5.42) from Chapter 5. A spatial pre-whitening operation can be defined from the transformed noise-only correlation matrix as in (5.43), which can subsequently be used to spatially pre-whiten the transformed signals, yielding a vector of pre-whitened-transformed (PWT) microphone signals[2]:

$$\underline{\mathbf{y}}(k,l) = \mathbf{L}^{-1}(k,l)\widetilde{\mathbf{\Upsilon}}^H(k,l)\mathbf{y}(k,l) \tag{8.4}$$

$$= \underbrace{\underline{\mathbf{h}}(k,l)s_{a,1}(k,l)}_{\underline{\mathbf{x}}(k,l)} + \underline{\mathbf{n}}(k,l) \tag{8.5}$$

[2]The dependence on the frequency bin and time frame was included here as a reminder and for completeness in the signal model. It will be subsequently dropped unless explicitly required.

where $\underline{\mathbf{y}}$ consists of the PWT LMA and XM signals, i.e. $\underline{\mathbf{y}} = \left[\underline{\mathbf{y}}_{\mathbf{a}}^T \; \underline{\mathbf{y}}_{\mathbf{e}}^T\right]^T$, the PWT RTF vector $\underline{\mathbf{h}} = \mathbf{L}^{-1}\widetilde{\boldsymbol{\Upsilon}}^H\mathbf{h}$ and the respective correlation matrices are:

$$\underline{\mathbf{R}}_{\mathbf{yy}} = \mathbb{E}\{\underline{\mathbf{y}}\,\underline{\mathbf{y}}^H\} = \mathbf{L}^{-1}\widetilde{\boldsymbol{\Upsilon}}^H\mathbf{R}_{\mathbf{yy}}\widetilde{\boldsymbol{\Upsilon}}\mathbf{L}^{-H} \tag{8.6}$$

$$\underline{\mathbf{R}}_{\mathbf{nn}} = \mathbb{E}\{\underline{\mathbf{n}}\,\underline{\mathbf{n}}^H\} = \mathbf{L}^{-1}\widetilde{\boldsymbol{\Upsilon}}^H\mathbf{R}_{\mathbf{nn}}\widetilde{\boldsymbol{\Upsilon}}\mathbf{L}^{-H} = \mathbf{I}_{(\mathrm{M_a}+\mathrm{M_e})} \tag{8.7}$$

$$\underline{\mathbf{R}}_{\mathbf{xx}} = \mathbb{E}\{\underline{\mathbf{x}}\,\underline{\mathbf{x}}^H\} = \sigma^2_{\mathrm{s_{a,1}}}\underline{\mathbf{h}}\,\underline{\mathbf{h}}^H \tag{8.8}$$

where the expression for $\underline{\mathbf{R}}_{\mathbf{nn}}$ is a direct consequence of (5.43). With the assumption of the desired speech source and noise being uncorrelated, it also holds that $\underline{\mathbf{R}}_{\mathbf{yy}} = \underline{\mathbf{R}}_{\mathbf{xx}} + \underline{\mathbf{R}}_{\mathbf{nn}}$. In the PWT domain, the estimate of the speech component in the first microphone of the LMA, z_1, which is equivalent to (8.3), is then obtained through the linear filtering of the PWT microphone signals, such that:

$$\mathrm{z}_1 = \underset{\sim}{\mathbf{w}}^H\underline{\mathbf{y}} \tag{8.9}$$

where $\underset{\sim}{\mathbf{w}} = \mathbf{L}^H\widetilde{\boldsymbol{\Upsilon}}^{-1}\mathbf{w}$ is a complex-valued filter[3]. Consequently, the flow of signal processing operations to obtain the speech estimate can be summarised as in Fig. 8.1.

8.3 MVDR with an LMA and XMs in the PWT domain

For the unprocessed microphone signals, the MVDR beamformer problem is recalled as:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimise}} \quad & \mathbf{w}^H\mathbf{R}_{\mathbf{nn}}\mathbf{w} \\ \text{subject to} \quad & \mathbf{w}^H\mathbf{h} = 1 \end{aligned} \tag{8.10}$$

with the solution:

$$\mathbf{w}_{\mathrm{mvdr}} = \frac{\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{h}}{\mathbf{h}^H\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{h}} \tag{8.11}$$

and the desired speech signal estimate, $\mathrm{z}_1 = \mathbf{w}_{\mathrm{mvdr}}^H\mathbf{y}$. As discussed in previous chapters, estimates are required for both $\mathbf{R}_{\mathbf{nn}}$ and $\mathbf{h}$. Firstly on replacing $\mathbf{R}_{\mathbf{nn}}$ with some estimate, $\hat{\mathbf{R}}_{\mathbf{nn}}$, in the PWT domain, it can be seen that using $\hat{\mathbf{R}}_{\mathbf{nn}}$

[3]Since the sequence of operations from $\mathbf{w}$ to $\underset{\sim}{\mathbf{w}}$ is not exactly that of a PWT signal vector, a slightly different notation is used for this quantity.

in (5.43) will result in an estimate for the pre-whitening operator as $\hat{\mathbf{L}}$. Hence from (8.7), $\hat{\mathbf{R}}_{\mathbf{nn}}$ can be expressed as:

$$\hat{\mathbf{R}}_{\mathbf{nn}} = \widetilde{\boldsymbol{\Upsilon}}^{-H}\hat{\mathbf{L}}\hat{\mathbf{L}}^{H}\widetilde{\boldsymbol{\Upsilon}}^{-1} \tag{8.12}$$

Replacing $\mathbf{R}_{\mathbf{nn}}$ in (8.10) with $\hat{\mathbf{R}}_{\mathbf{nn}}$ in (8.12) then results in the MVDR beamformer problem formulated in the PWT domain as:

$$\begin{aligned} \underset{\underset{\sim}{\mathbf{w}}}{\text{minimise}} \quad & \underset{\sim}{\mathbf{w}}^{H}\underset{\sim}{\mathbf{w}} \\ \text{subject to} \quad & \underset{\sim}{\mathbf{w}}^{H}\underline{\mathbf{h}} = 1 \end{aligned} \tag{8.13}$$

where $\underset{\sim}{\mathbf{w}}$ is redefined as $\underset{\sim}{\mathbf{w}} = \hat{\mathbf{L}}^{H}\widetilde{\boldsymbol{\Upsilon}}^{-1}\mathbf{w}$ and $\underline{\mathbf{h}}$ is redefined as $\underline{\mathbf{h}} = \hat{\mathbf{L}}^{-1}\widetilde{\boldsymbol{\Upsilon}}^{H}\mathbf{h}$. The solution to (8.13) then yields the optimal filter in the PWT domain:

$$\underset{\sim}{\mathbf{w}}_{\text{mvdr}} = \frac{\underline{\mathbf{h}}}{\underline{\mathbf{h}}^{H}\underline{\mathbf{h}}} \tag{8.14}$$

with the desired speech signal estimate, $\mathrm{z}_1 = \underset{\sim}{\mathbf{w}}_{\text{mvdr}}^{H}\underline{\mathbf{y}}$. As $\mathbf{h}$ is still unknown, however, it means that $\underline{\mathbf{h}}$ is also unknown and an estimate for this component is still required. Using the same $\hat{\mathbf{R}}_{\mathbf{nn}}$, two general approaches for the estimation of $\underline{\mathbf{h}}$ can be considered, either making use of an available a priori RTF vector pertaining to the LMA or making use of only the observable microphone data, i.e. a fully data-dependent estimate. The remainder of this section elaborates on these procedures.

8.3.1 Using an a priori RTF vector

As discussed in Chapter 5, an a priori RTF vector for a microphone configuration of an LMA linked with XMs can be defined as in (5.9), repeated here for convenience:

$$\widetilde{\mathbf{h}} = [\ \widetilde{\mathbf{h}}_{\mathbf{a}}^{T}\ \mathbf{h}_{\mathbf{e}}^{T}\ ]^{T} \tag{8.15}$$

which consists partially of the a priori RTF vector pertaining to the LMA, $\widetilde{\mathbf{h}}_{\mathbf{a}}$, and partially of the RTF vector pertaining to the XM, $\mathbf{h}_{\mathbf{e}}$, which is unknown and remains to be estimated. The estimate of $\mathbf{h}_{\mathbf{e}}$ will be denoted as $\hat{\widetilde{\mathbf{h}}}_{\mathbf{e}}$ to emphasise that it is constrained by the a priori knowledge set by $\widetilde{\mathbf{h}}_{\mathbf{a}}$ but estimated from the observed microphone data. In the following, the GEVD-based procedure for obtaining $\hat{\widetilde{\mathbf{h}}}_{\mathbf{e}}$ from section 5.5.2 of Chapter 5 is re-framed in the PWT domain.

In the PWT domain, using (8.6)-(8.8), a rank-1 matrix approximation problem can be firstly formulated to estimate the entire RTF vector:

$$\underset{\sigma^2_{\mathrm{s_{a,1}}},\,\mathbf{h}}{\text{minimise}} \quad ||(\underline{\hat{\mathbf{R}}}_{\mathbf{yy}} - \underline{\hat{\mathbf{R}}}_{\mathbf{nn}}) - \sigma^2_{\mathrm{s_{a,1}}}\hat{\mathbf{L}}^{-1}\widetilde{\boldsymbol{\Upsilon}}^H\mathbf{h}\,\mathbf{h}^H\widetilde{\boldsymbol{\Upsilon}}\hat{\mathbf{L}}^{-H}||^2_F \tag{8.16}$$

where $||.||_F$ is the Frobenius norm, $\underline{\hat{\mathbf{R}}}_{\mathbf{yy}} = \hat{\mathbf{L}}^{-1}\widetilde{\boldsymbol{\Upsilon}}^H\hat{\mathbf{R}}_{\mathbf{yy}}\widetilde{\boldsymbol{\Upsilon}}\hat{\mathbf{L}}^{-H}$, $\hat{\mathbf{R}}_{\mathbf{yy}}$ is the data-dependent estimate of $\mathbf{R}_{\mathbf{yy}}$, and $\underline{\hat{\mathbf{R}}}_{\mathbf{nn}} = \hat{\mathbf{L}}^{-1}\widetilde{\boldsymbol{\Upsilon}}^H\hat{\mathbf{R}}_{\mathbf{nn}}\widetilde{\boldsymbol{\Upsilon}}\hat{\mathbf{L}}^{-H} = \mathbf{I}_{(\mathrm{M_a}+\mathrm{M_e})}$. From (8.15), an a priori RTF vector in the PWT domain can be defined as follows:

$$\underline{\widetilde{\mathbf{h}}} = \hat{\mathbf{L}}^{-1}\widetilde{\boldsymbol{\Upsilon}}^H[\ \widetilde{\mathbf{h}}_{\mathbf{a}}^T\ \mathbf{h}_{\mathbf{e}}^T\ ]^T = \hat{\mathbf{L}}^{-1}\left[\mathbf{0}^T\ \ 1\ \ \mathbf{h}_{\mathbf{e}}^T\right]^T \tag{8.17}$$

where $\mathbf{0}$ is a vector of $(\mathrm{M_a}-1)$ zeros. Replacing $\mathbf{h}$ with the a priori RTF vector from (8.15) then results in:

$$\underset{\sigma^2_{\mathrm{s_{a,1}}},\,\mathbf{h_e}}{\text{minimise}} \quad ||(\underline{\hat{\mathbf{R}}}_{\mathbf{yy}} - \underline{\hat{\mathbf{R}}}_{\mathbf{nn}}) - \sigma^2_{\mathrm{s_{a,1}}}\underline{\widetilde{\mathbf{h}}}\,\underline{\widetilde{\mathbf{h}}}^H||^2_F \tag{8.18}$$

where now only an estimate is required for $\mathbf{h_e}$, which in turn will define the a priori RTF vector. As discussed in section 5.5.2 of Chapter 5, it is only the lower $(\mathrm{M_e}+1)\times(\mathrm{M_e}+1)$ blocks of $\underline{\hat{\mathbf{R}}}_{\mathbf{yy}}$ and $\underline{\hat{\mathbf{R}}}_{\mathbf{nn}}$ that are required for estimating $\mathbf{h_e}$. Hence (8.18) can be reduced to:

$$\underset{\sigma^2_{\mathrm{s_{a,1}}},\mathbf{h_e}}{\text{minimise}} \quad ||(\underline{\underline{\hat{\mathbf{R}}}}_{\mathbf{yy}} - \underline{\underline{\hat{\mathbf{R}}}}_{\mathbf{nn}}) - \sigma^2_{\mathrm{s_{a,1}}}\mathbf{J}^T\underline{\widetilde{\mathbf{h}}}\,\underline{\widetilde{\mathbf{h}}}^H\mathbf{J}||^2_F \tag{8.19}$$

where $\mathbf{J} = [\ \mathbf{0}_{(\mathrm{M_e}+1)\times(\mathrm{M_a}-1)}\ |\ \mathbf{I}_{(\mathrm{M_e}+1)}\ ]^T$ is a selection matrix, $\underline{\underline{\hat{\mathbf{R}}}}_{\mathbf{yy}} = \mathbf{J}^T\underline{\hat{\mathbf{R}}}_{\mathbf{yy}}\mathbf{J}$, and $\underline{\underline{\hat{\mathbf{R}}}}_{\mathbf{nn}} = \mathbf{J}^T\underline{\hat{\mathbf{R}}}_{\mathbf{nn}}\mathbf{J} = \mathbf{I}_{\mathrm{M_e}+1}$. The solution of (8.19) then follows from a generalised eigenvalue decomposition (GEVD) of the matrix pencil $(\underline{\underline{\hat{\mathbf{R}}}}_{\mathbf{yy}}, \underline{\underline{\hat{\mathbf{R}}}}_{\mathbf{nn}})$, which, due to the formulation in the PWT domain, is equivalent to an eigenvalue decomposition (EVD) of $\underline{\underline{\hat{\mathbf{R}}}}_{\mathbf{yy}}$:

$$\underline{\underline{\hat{\mathbf{R}}}}_{\mathbf{yy}} = \hat{\mathbf{V}}\hat{\boldsymbol{\Xi}}\hat{\mathbf{V}}^H \tag{8.20}$$

where $\hat{\mathbf{V}}$ is a $(\mathrm{M_e}+1)\times(\mathrm{M_e}+1)$ unitary matrix of eigenvectors and $\hat{\boldsymbol{\Xi}}$ is a diagonal matrix with the associated eigenvalues in descending order. The estimate of $\mathbf{h_e}$ then follows from the appropriate scaling of the principal eigenvector, $\hat{\mathbf{v}}_{\mathbf{max}}$:

$$\begin{bmatrix}\mathbf{0}\\ 1\\ \hat{\underline{\mathbf{h}}}_{\mathbf{e}}\end{bmatrix} = \frac{\hat{\mathbf{L}}\mathbf{J}\hat{\mathbf{v}}_{\mathbf{max}}}{\mathbf{e}^T_{\mathrm{M_a}}\hat{\mathbf{L}}\mathbf{J}\hat{\mathbf{v}}_{\mathbf{max}}} = \frac{\hat{\mathbf{L}}\mathbf{J}\hat{\mathbf{v}}_{\mathbf{max}}}{\hat{l}_{M_\mathrm{a}}\hat{\mathrm{v}}_1} \tag{8.21}$$

where $\mathbf{e}_{\mathrm{M_a}}$ is an $(\mathrm{M_a}+\mathrm{M_e})$ selection vector consisting of all zeros except for a one in the $\mathrm{M_a}$-th position, $\hat{\mathrm{v}}_1$ is the first element of $\hat{\mathbf{v}}_{\mathbf{max}}$, and $\hat{l}_{M_\mathrm{a}}$ is the

real-valued $(\mathrm{M_a}, \mathrm{M_a})$-th element in $\hat{\mathbf{L}}$. Substitution of this expression into (8.17) finally yields the a priori RTF vector in the PWT domain as[4]:

$$\widetilde{\underline{\mathbf{h}}} = \frac{1}{\hat{l}_{M_\mathrm{a}}\ \hat{\mathrm{v}}_1} \begin{bmatrix} \mathbf{0} & \hat{\mathbf{v}}_{\mathbf{max}} \end{bmatrix}^T \tag{8.22}$$

Finally, replacing $\underline{\mathbf{h}}$ in (8.14) with $\widetilde{\underline{\mathbf{h}}}$ from (8.22) results in the MVDR beamformer based on a priori knowledge pertaining to the LMA:

$$\widetilde{\underline{\mathbf{w}}}_{\mathrm{mvdr}} = \hat{l}_{M_\mathrm{a}}\ \hat{\mathrm{v}}_1^*\begin{bmatrix} \mathbf{0} & \hat{\mathbf{v}}_{\mathbf{max}} \end{bmatrix}^T \tag{8.23}$$

which will be referred to as MVDR-AP. The corresponding speech estimate is then computed using (8.9):

$$\widetilde{\mathrm{z}}_1 = \hat{l}_{M_\mathrm{a}}\ \hat{\mathrm{v}}_1\, \hat{\mathbf{v}}_{\mathbf{max}}^H \begin{bmatrix} \underline{\mathrm{y}}_{\mathrm{a},M_\mathrm{a}} \\ \underline{\mathbf{y}}_{\mathbf{e}} \end{bmatrix} \tag{8.24}$$

As a consequence of incorporating the a priori information into the rank-1 speech model, it can be seen that it is only necessary to filter the last $(\mathrm{M_e}+1)$ elements of $\underline{\mathbf{y}}$, i.e. $\underline{\mathrm{y}}_{\mathrm{a},M_\mathrm{a}}$ and $\underline{\mathbf{y}}_{\mathbf{e}}$, with the lower order, $(\mathrm{M_e}+1)$ filter defined by $\hat{l}_{M_\mathrm{a}}\ \hat{\mathrm{v}}_1^*\, \hat{\mathbf{v}}_{\mathbf{max}}$. The flow of signal processing operations is depicted in Fig. 8.2.

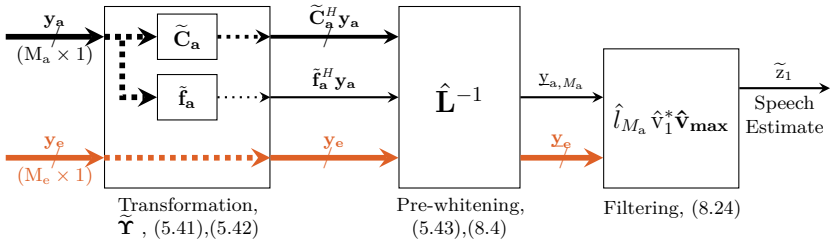


Figure 8.2: Signal processing flow for obtaining a speech estimate in the PWT domain using an a priori RTF vector.

8.3.2 Using a data-dependent RTF vector

In the case where no a priori knowledge is available with respect to the LMA, then an entire RTF vector for both the LMA and XM signals must be estimated from the microphone data. In the PWT domain, it is therefore (8.16) that will need to be directly solved.

[4]Once again, it is acknowledged that there is a slight abuse of notation here as the estimate for $\widetilde{\underline{\mathbf{h}}}$ should perhaps be denoted as $\widetilde{\widehat{\underline{\mathbf{h}}}}$. However, in favour of legibility and to stress that the estimation is done in accordance to the a priori assumptions set by $\widetilde{\mathbf{h}}_\mathbf{a}$ is why the notation is maintained as $\widetilde{\underline{\mathbf{h}}}$.

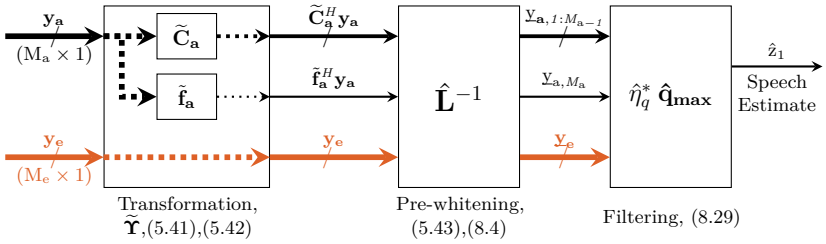


Figure 8.3: Signal processing flow for obtaining a speech estimate in the PWT domain using a data-dependent RTF vector.

The solution to (8.16) follows from a GEVD of the matrix pencil $(\hat{\underline{\mathbf{R}}}_{\mathbf{yy}}, \hat{\underline{\mathbf{R}}}_{\mathbf{nn}})$, which, due to the formulation in the PWT domain, is equivalent to an EVD of $\hat{\underline{\mathbf{R}}}_{\mathbf{yy}}$

$$\hat{\underline{\mathbf{R}}}_{\mathbf{yy}} = \hat{\mathbf{Q}}\hat{\boldsymbol{\Lambda}}\hat{\mathbf{Q}}^H \tag{8.25}$$

where $\hat{\mathbf{Q}}$ is an $(M_a + M_e) \times (M_a + M_e)$ unitary matrix of eigenvectors and $\hat{\boldsymbol{\Lambda}}$ is a diagonal matrix with the associated eigenvalues in descending order. The estimated RTF vector is then given by the principal (first in this case) eigenvector, $\hat{\mathbf{q}}_\mathbf{max}$:

$$\hat{\mathbf{h}} = \frac{\widetilde{\boldsymbol{\Upsilon}}^{-H}\hat{\mathbf{L}}\,\hat{\mathbf{q}}_\mathbf{max}}{\hat{\eta}_q} \tag{8.26}$$

where $\hat{\eta}_q = \mathbf{e}_1^T\widetilde{\boldsymbol{\Upsilon}}^{-H}\hat{\mathbf{L}}\,\hat{\mathbf{q}}_\mathbf{max}$ and $\mathbf{e}_1$ is an $(M_a + M_e)$ selection vector with a one as the first element and zeros everywhere else. In the PWT domain, this data-dependent RTF vector then becomes:

$$\hat{\underline{\mathbf{h}}} = \hat{\mathbf{L}}^{-1}\widetilde{\boldsymbol{\Upsilon}}^H\hat{\mathbf{h}} = \frac{\hat{\mathbf{q}}_\mathbf{max}}{\hat{\eta}_q} \tag{8.27}$$

Replacing $\underline{\mathbf{h}}$ in (8.14) with $\hat{\underline{\mathbf{h}}}$ from (8.27) results in the MVDR beamformer that makes use of a data-dependent RTF vector:

$$\hat{\underline{\mathbf{w}}}_{\text{mvdr}} = \hat{\eta}_q^*\,\hat{\mathbf{q}}_\mathbf{max} \tag{8.28}$$

which will be referred to as MVDR-DD. The corresponding speech estimate is then computed using (8.9):

$$\hat{z}_1 = \hat{\eta}_q\,\hat{\mathbf{q}}_\mathbf{max}^H\,\underline{\mathbf{y}} \tag{8.29}$$

where now all $(M_a + M_e)$ signals need to be filtered as opposed to only $(M_e + 1)$ signals in (8.24) when an a priori RTF vector is used. The flow of signal processing operations is depicted in Fig. 8.3.

8.4 Integrated MVDR beamformer

8.4.1 Quadratically Constrained Quadratic Program

As opposed to relying on only an a priori RTF vector or a data-dependent RTF vector, the merging or integration of both RTF vectors into a single approach can be framed as a quadratically constrained quadratic program (QCQP), firstly with respect to the unprocessed microphone signals:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimise}} \quad & \mathbf{w}^H \hat{\mathbf{R}}_{\mathbf{nn}} \mathbf{w} \\ \text{subject to} \quad & |\mathbf{w}^H \widetilde{\mathbf{h}} - 1|^2 \leq \widetilde{\epsilon}^2 \\ & |\mathbf{w}^H \hat{\mathbf{h}} - 1|^2 \leq \hat{\epsilon}^2 \end{aligned} \tag{8.30}$$

where $\widetilde{\epsilon}^2$ and $\hat{\epsilon}^2$ are maximum-tolerated squared deviations from a distortionless response due to $\widetilde{\mathbf{h}}$ or $\hat{\mathbf{h}}$ respectively. In the PWT domain, (8.30) is equivalently:

$$\begin{aligned} \underset{\underset{\sim}{\mathbf{w}}}{\text{minimise}} \quad & \underset{\sim}{\mathbf{w}}^H \underset{\sim}{\mathbf{w}} \\ \text{subject to} \quad & |\underset{\sim}{\mathbf{w}}^H \underline{\widetilde{\mathbf{h}}} - 1|^2 \leq \widetilde{\epsilon}^2 \\ & |\underset{\sim}{\mathbf{w}}^H \underline{\hat{\mathbf{h}}} - 1|^2 \leq \hat{\epsilon}^2 \end{aligned} \tag{8.31}$$

where $\underline{\widetilde{\mathbf{h}}}$ and $\underline{\hat{\mathbf{h}}}$ are given in (8.22) and (8.27) respectively. Whereas in (8.13), the hard constraint of $\underline{\mathbf{h}}$ is replaced by either $\underline{\widetilde{\mathbf{h}}}$ or $\underline{\hat{\mathbf{h}}}$, (8.31) can be interpreted as the relaxation of the hard constraints imposed by $\underline{\widetilde{\mathbf{h}}}$ or $\underline{\hat{\mathbf{h}}}$ by the specified deviations $\widetilde{\epsilon}^2$ and $\hat{\epsilon}^2$ respectively. In the following, the quantities $|\underset{\sim}{\mathbf{w}}^H \underline{\widetilde{\mathbf{h}}} - 1|^2$ and $|\underset{\sim}{\mathbf{w}}^H \underline{\hat{\mathbf{h}}} - 1|^2$ are referred to as speech distortions and $\widetilde{\epsilon}^2$ and $\widetilde{\epsilon}^2$ are the respective maximum tolerable speech distortions. Furthermore, the first inequality constraint in (8.31) will be referred to as the a priori constraint (APC), and the second inequality constraint will be referred to as the data-dependent constraint (DDC).

The QCQP of (8.30) is in fact a subset of the more general QCQP considered in [34, 189] and as well as an extension to the parametrised Multi-Channel Wiener Filter (PMWF) [289]. In [34, 189], the inequality constraints considered are a set of a priori measured RTF vectors, and in [289], only one inequality constraint is considered. The difference in (8.30) from both of these approaches is that two inequality constraints are considered, one that relies on a priori knowledge, and the other which is fully estimated from the data.

The Lagrangian of (8.31) is given by:

$$\mathcal{L}(\underset{\sim}{\mathbf{w}}, \alpha, \beta) = \underset{\sim}{\mathbf{w}}^H \underset{\sim}{\mathbf{w}} + \alpha(|\underset{\sim}{\mathbf{w}}^H \underline{\widetilde{\mathbf{h}}} - 1|^2 - \widetilde{\epsilon}^2) + \beta(|\underset{\sim}{\mathbf{w}}^H \underline{\hat{\mathbf{h}}} - 1|^2 - \hat{\epsilon}^2) \tag{8.32}$$

where α and β are Lagrangian multipliers. Taking the partial derivative of (8.32) with respect to $\underset{\sim}{\mathbf{w}}$ and setting to zero results in the integrated MVDR beamformer, MVDR-INT:

$$\underset{\sim}{\mathbf{w}}_{\text{int}} = (\mathbf{I}_{(\mathrm{M_a}+\mathrm{M_e})} + \alpha \underline{\widetilde{\mathbf{h}}} \underline{\widetilde{\mathbf{h}}}^H + \beta \underline{\hat{\mathbf{h}}} \underline{\hat{\mathbf{h}}}^H)^{-1} (\alpha \underline{\widetilde{\mathbf{h}}} \underline{\widetilde{\mathbf{h}}}^H + \beta \underline{\hat{\mathbf{h}}} \underline{\hat{\mathbf{h}}}^H) \mathbf{e}_1 \tag{8.33}$$

where the actual values of α and β depend on the prescribed maximum tolerable speech distortions $\widetilde{\epsilon}^2$ and $\widetilde{\epsilon}^2$. Unfortunately, there is no closed form solution to relate α and β to $\widetilde{\epsilon}^2$ and $\widetilde{\epsilon}^2$, hence an iterative procedure is usually required to obtain a solution (one of which is shortly discussed).

Upon further inspection of (8.33), it can be observed that it is identical to the integrated MVDR beamformer in (3.46) from Chapter 3, except that it is formulated in the PWT domain and the filter is now an $(\mathrm{M_a} + \mathrm{M_e}) \times 1$ vector due to the inclusion of the XMs. Consequently, $\underset{\sim}{\mathbf{w}}_{\text{int}}$ can also be written as a linear combination of $\underset{\sim}{\widetilde{\mathbf{w}}}_{\text{mvdr}}$ and $\underset{\sim}{\hat{\mathbf{w}}}_{\text{mvdr}}$ with complex weightings as in (3.49):

$$\underset{\sim}{\mathbf{w}}_{\text{int}} = g_{ap}\,(\alpha, \beta)\, \underset{\sim}{\widetilde{\mathbf{w}}}_{\text{mvdr}} + g_{dd}\,(\alpha, \beta)\, \underset{\sim}{\hat{\mathbf{w}}}_{\text{mvdr}} \tag{8.34}$$

where $\underset{\sim}{\widetilde{\mathbf{w}}}_{\text{mvdr}}$ and $\underset{\sim}{\hat{\mathbf{w}}}_{\text{mvdr}}$ are given in (8.23) and (8.28) respectively, and the complex weightings are given by:

$$g_{ap}\,(\alpha, \beta) = \left[\frac{\alpha k_{hh}[1 + \beta(k_{qq} - k_{hq})]}{\alpha k_{hh} + \beta k_{qq} + \alpha\beta(k_{hh}k_{qq} - k_{hq}k_{qh}) + 1} \right] \tag{8.35}$$

$$g_{dd}\,(\alpha, \beta) = \left[\frac{\beta k_{qq}[1 + \alpha(k_{hh} - k_{qh})]}{\alpha k_{hh} + \beta k_{qq} + \alpha\beta(k_{hh}k_{qq} - k_{hq}k_{qh}) + 1} \right] \tag{8.36}$$

and

$$k_{hh} = \underline{\widetilde{\mathbf{h}}}^H \underline{\widetilde{\mathbf{h}}};\quad k_{qq} = \underline{\hat{\mathbf{h}}}^H \underline{\hat{\mathbf{h}}};\quad k_{hq} = \underline{\widetilde{\mathbf{h}}}^H \underline{\hat{\mathbf{h}}};\quad k_{qh} = \underline{\hat{\mathbf{h}}}^H \underline{\widetilde{\mathbf{h}}}. \tag{8.37}$$

By substituting the expressions for $\underset{\sim}{\widetilde{\mathbf{w}}}_{\text{mvdr}}$ and $\underset{\sim}{\hat{\mathbf{w}}}_{\text{mvdr}}$ from (8.23) and (8.28) respectively, the resulting speech estimate from the MVDR-INT is then finally:

$$\hat{\mathrm{z}}_{\text{int}} = g_{ap}^*\,(\alpha, \beta)\, \widetilde{\mathrm{z}}_1 + g_{dd}^*\,(\alpha, \beta)\, \hat{\mathrm{z}}_1 \tag{8.38}$$

where $\widetilde{\mathrm{z}}_1$ and $\hat{\mathrm{z}}_1$ are defined in (8.24) and (8.29) respectively. Hence the integrated beamformer output is simply a linear combination of the two speech

estimates which relied on either a priori information or not. A block scheme for obtaining the integrated speech estimate is depicted in Fig. 8.4.

It is also worth reminding ourselves here of (3.56), revealing that the MVDR-INT of (8.34) may also be expressed as:

$$\underline{\mathbf{w}}_{\text{int}} = \psi_1 \, \widetilde{\underline{\mathbf{w}}}_{\text{mvdr}} + \psi_2 \, \hat{\underline{\mathbf{w}}}_{\text{mvdr}} + \psi_3 \, \underline{\mathbf{w}}_{\text{lcmv}} + \psi_4 \, \mathbf{w}_{\text{z}} \tag{8.39}$$

where ψ_1, ψ_2, ψ_3, and ψ_4 are real-valued scalars ≥ 0, and $\sum_{i=1}^{4} \psi_i = 1$. Hence the MVDR-INT is a convex combination of the MVDR-AP beamformer, the MVDR-DD beamformer, an LCMV beamformer, $\underline{\mathbf{w}}_{\text{lcmv}}$, and an all-zero vector, $\mathbf{w}_{\text{z}}$. The LCMV beamformer, $\underline{\mathbf{w}}_{\text{lcmv}}$, is defined similarly to (3.43), except in the PWT domain. From (8.37) and (8.38), defining $Đ = \alpha k_{hh} + \beta k_{qq} + \alpha\beta(k_{hh}k_{qq} - k_{hq}k_{qh}) + 1$, it can be shown that the scalars are given by:

$$\psi_1 = \frac{\alpha k_{hh}}{Đ}; \quad \psi_2 = \frac{\beta k_{qq}}{Đ}; \quad \psi_3 = \frac{\alpha\beta(k_{hh}k_{qq} - k_{hq}k_{qh})}{Đ}; \quad \psi_4 = \frac{1}{Đ} \tag{8.40}$$

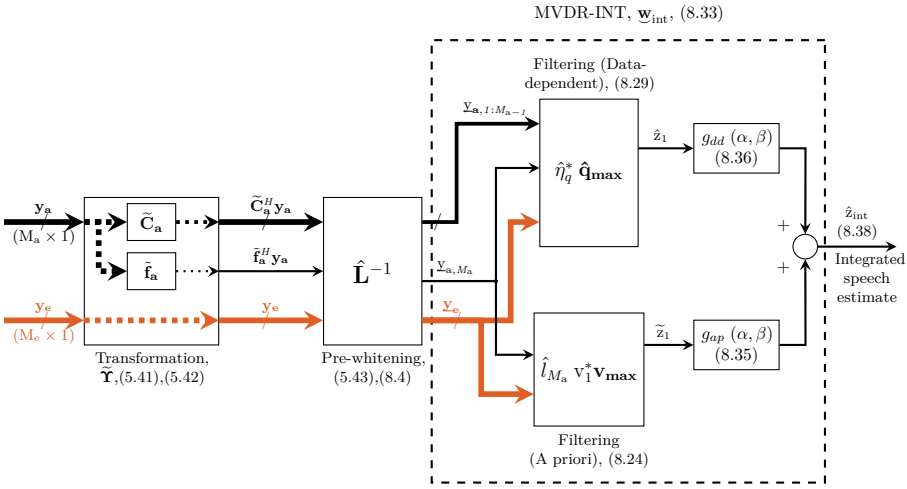


Figure 8.4: Signal processing flow for obtaining an integrated speech estimate in the PWT domain using both an a priori RTF vector and a data-dependent RTF vector.

In Chapter 3, α and β were tuned in order to obtain expressions for $g_{ap}\,(\alpha,\beta)$ and $g_{dd}\,(\alpha,\beta)$, which was then subsequently substituted into (8.38) to obtain the integrated speech estimate, $\hat{z}_{\text{int}}$. It is however evident from the QCQP formulation that $\underline{\mathbf{w}}_{\text{int}}$ (and hence the $\hat{z}_{\text{int}}$) can also be obtained by specifying the maximum tolerable speech distortions for the APC and DDC, which has

more of a tangible meaning as opposed to directly tuning the the Lagrangian multipliers. This will be discussed in more detail in the following section.

Once appropriate values are chosen for $\widetilde{\epsilon}^2$ and $\hat{\epsilon}^2$, then a package for specifying and solving convex programs such as CVX [290, 291] can be used for solving (8.31). Alternatively, more computationally efficient methods may be applied such as those proposed in [34, 189], one of which is highlighted in Algorithm 1. Here, a gradient ascent method [292] for solving (8.31) is described, which is based on solving the dual problem:

$$\begin{aligned} &\underset{(\alpha,\beta)}{\text{maximise}} \quad \mathscr{D}(\alpha,\beta) \\ &\text{subject to} \;\; \alpha \geq 0; \;\; \beta \geq 0 \end{aligned} \tag{8.41}$$

where $\mathscr{D}(\alpha,\beta) = \inf \mathcal{L}(\underset{\sim}{\mathbf{w}}_{\text{int}}, \alpha, \beta)$ is the dual function. As the dual function, $\mathscr{D}(\alpha,\beta)$, is concave [123], a gradient ascent procedure can be used to update the values of α and β using the gradients, $\frac{\partial \mathscr{D}(\alpha,\beta)}{\partial \alpha} = |\underset{\sim}{\mathbf{w}}_{\text{int}}^H \underline{\widetilde{\mathbf{h}}} - 1|^2 - \widetilde{\epsilon}^2$ and $\frac{\partial \mathscr{D}(\alpha,\beta)}{\partial \beta} = |\underset{\sim}{\mathbf{w}}_{\text{int}}^H \underline{\hat{\mathbf{h}}} - 1|^2 - \hat{\epsilon}^2$, i.e. the gradients of the dual function with respect to the particular Lagrange multiplier is the respective constraint. This then gives rise to Algorithm 1 [189], which makes use of the simplified expressions for $\underset{\sim}{\mathbf{w}}_{\text{int}}$ with the complex-valued weightings as opposed to computing (8.33) directly. The Lagrangian multipliers, α and β, are then updated via the gradient ascent procedure with the step size γ, whose value can be controlled using a backtracking method [185]. The algorithm continues until the respective gradients are within some specified tolerance, δ.

Algorithm 3 Gradient ascent method for solving the QCQP of (8.31)

Initialise α, β, $\underset{\sim}{\mathbf{w}}_{\text{int}}$. Set tolerance, δ. $n = 0$
while $(|\underset{\sim}{\mathbf{w}}_{\text{int}}^H(n)\underline{\widetilde{\mathbf{h}}} - 1|^2 - \widetilde{\epsilon}^2) > \delta$ OR $(|\underset{\sim}{\mathbf{w}}_{\text{int}}^H(n)\underline{\hat{\mathbf{h}}} - 1|^2 - \hat{\epsilon}^2) > \delta$ **do**
  $g_{ap}(n) = g_{ap}\,(\alpha(n-1), \beta(n-1))$ from (8.35)
  $g_{dd}(n) = g_{dd}\,(\alpha(n-1), \beta(n-1))$ from (8.36)
  $\underset{\sim}{\mathbf{w}}_{\text{int}}(n) = g_{ap}(n)\underset{\sim}{\widetilde{\mathbf{w}}}_{\text{mvdr}} + g_{dd}(n)\underset{\sim}{\hat{\mathbf{w}}}_{\text{mvdr}}$ from (8.34)
  Set γ according to a backtracking method.
  $\alpha(n) = \max\{\alpha(n-1) + \gamma(|\underset{\sim}{\mathbf{w}}_{\text{int}}^H(n)\underline{\widetilde{\mathbf{h}}} - 1|^2 - \widetilde{\epsilon}^2), 0\}$
  $\beta(n) = \max\{\beta(n-1) + \gamma(|\underset{\sim}{\mathbf{w}}_{\text{int}}^H(n)\underline{\hat{\mathbf{h}}} - 1|^2 - \hat{\epsilon}^2), 0\}$
  $n = n + 1$
end while

8.4.2 Effect of $\widetilde{\epsilon}$ and $\hat{\epsilon}$

As the QCQP of (8.31) in principle is to be solved for every time frame and frequency bin, it can therefore lead to quite a versatile beamformer as the parameters, $\widetilde{\epsilon}^2$ and $\hat{\epsilon}^2$ can be set independently for each frequency in every time frame in order to define the inequality constraints. However, although (8.31) is a well known QCQP for which there are several methods available to find the solution, it still remains unclear as to what would be a reasonable strategy for setting or tuning these maximum tolerable speech distortions in practice. As opposed to Chapter 3, where tuning rules were developed for the Lagrangian multipliers, here a strategy is outlined for tuning $\widetilde{\epsilon}^2$ and $\hat{\epsilon}^2$, which will in turn compute the corresponding Lagrangian multipliers and $\mathbf{w}_{\text{int}}$ (for instance as outlined in Algorithm 1), as this is believed to be a more insightful procedure.

In order to develop a strategy for tuning $\widetilde{\epsilon}$ and $\hat{\epsilon}$, it will be useful to observe the constraints of (8.5) in more detail. The derivations that follow will reveal that the space spanned by $\widetilde{\epsilon}$ and $\hat{\epsilon}$ can be divided into four distinct regions as illustrated in Fig. 8.5, where each of these regions corresponds to a particular set of constraints being active. This is one of the more important pictures in this dissertation and explicitly depicts how the integrated MVDR beamformer is able to encapsulate a wide range of speech enhancement filters through the integration of a priori and data-dependent spatial information.

Firstly, substitution of $\underline{\mathbf{w}}_{\text{int}} = \mathbf{0}$ into the APC and DDC from (8.31) shows that when $\widetilde{\epsilon} > 1$ and $\hat{\epsilon} > 1$, both the APC and the DDC are inactive. This condition therefore defines the upper-right region (region I) of Fig. 8.5 and indeed corresponds to a complete attenuation of the microphone signals, i.e. a zero output signal.

For the case when $\hat{\epsilon} \to \infty$,, i.e. when the DDC is inactive, then $\beta \to 0$. If the APC is still active however, it becomes[5]:

$$|\underline{\mathbf{w}}_{\text{int}}^H \underline{\widetilde{\mathbf{h}}} - 1| \leq \widetilde{\epsilon} \tag{8.42}$$

Furthermore, if $0 \leq \widetilde{\epsilon} \leq 1$, then it can be deduced that:

$$\lim_{\substack{\hat{\epsilon}\to\infty \\ 0\leq\widetilde{\epsilon}\leq 1}} \underline{\mathbf{w}}_{\text{int}} = (1 - \widetilde{\epsilon})\underline{\widetilde{\mathbf{w}}}_{\text{mvdr}} \tag{8.43}$$

Substitution of (8.43) into (8.42) readily makes this evident, recalling that $\underline{\widetilde{\mathbf{w}}}_{\text{mvdr}}^H \underline{\widetilde{\mathbf{h}}} = 1$. It is worthwhile to also note that by using (8.35), the relationship

[5]The square root has been taken on both sides of the inequality from (8.31) in order to simplify the derivations that follow.

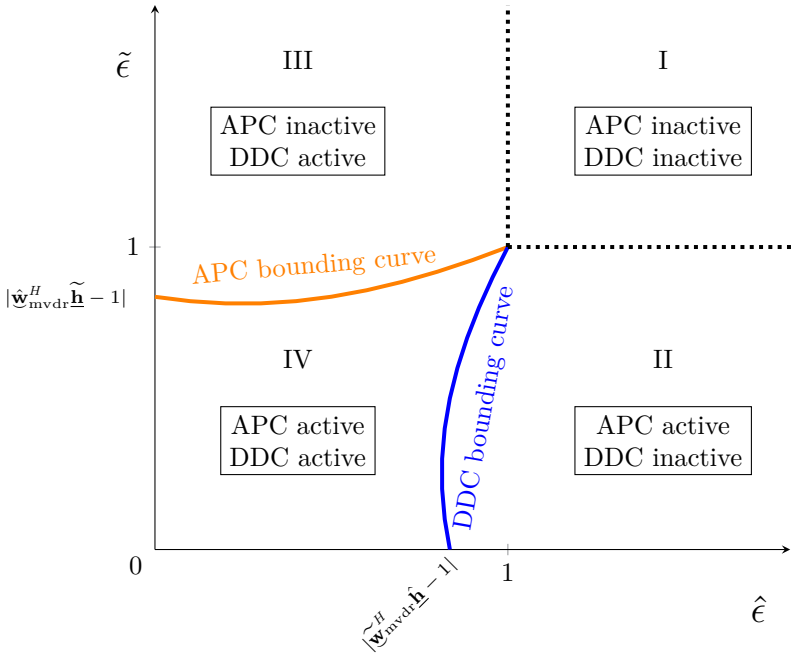


Figure 8.5: Depiction of the four regions for which the APC and DDC may be active or inactive within the space spanned by the maximum tolerable speech distortion parameters, $\widetilde{\epsilon}$ and $\hat{\epsilon}$. Region I corresponds to complete suppression of the microphone signals since both constraints are inactive and hence $\underset{\sim}{\mathbf{w}}_{\text{int}} = \mathbf{0}$. Region II corresponds to using only a priori spatial information as the APC is active and the DDC inactive. Region III corresponds to using only data-dependent spatial information as the APC is inactive and the DDC is active. Region IV corresponds to the merging of both a priori and data-dependent spatial information as both the APC and DDC are active. The curve dividing the regions II and IV is the DDC bounding curve defined when the equality is satisfied from (8.45). The curve dividing the regions III and IV is the APC bounding curve defined when the equality is satisfied from (8.49).

between α and $\widetilde{\epsilon}$ for $0 \leq \widetilde{\epsilon} \leq 1$ is then given as:

$$\lim_{\substack{\hat{\epsilon}\to\infty \\ 0\leq\widetilde{\epsilon}\leq 1}} \alpha = \frac{1}{k_{hh}} \frac{(1-\widetilde{\epsilon})}{\widetilde{\epsilon}} \tag{8.44}$$

In regards to the DDC, as $\hat{\epsilon}$ is decreased (from $\hat{\epsilon} \to \infty$), it remains inactive until $|\underset{\sim}{\mathbf{w}}^{H}\hat{\underline{\mathbf{h}}} - 1| = \hat{\epsilon}$. By substitution of (8.43) into the DDC of (8.31), the value of $\hat{\epsilon}$

at which the DDC becomes active, $\hat{\epsilon}_o$, is given by:

$$\hat{\epsilon}_o = |\underset{\sim}{\widetilde{\mathbf{w}}}^H_{\text{mvdr}}\hat{\underline{\mathbf{h}}}(1-\widetilde{\epsilon}) - 1| \tag{8.45}$$

In the limits of $\widetilde{\epsilon}$, when $\widetilde{\epsilon} = 1$, then $\hat{\epsilon}_o = 1$, and when $\widetilde{\epsilon} = 0$, then $\hat{\epsilon}_o = |\underset{\sim}{\widetilde{\mathbf{w}}}^H_{\text{mvdr}}\hat{\underline{\mathbf{h}}} - 1|$, where depending on $\hat{\underline{\mathbf{h}}}$, $|\underset{\sim}{\widetilde{\mathbf{w}}}^H_{\text{mvdr}}\hat{\underline{\mathbf{h}}} - 1| \leq 1$ or $|\underset{\sim}{\widetilde{\mathbf{w}}}^H_{\text{mvdr}}\hat{\underline{\mathbf{h}}} - 1| \geq 1$. The range of values obtained for $\hat{\epsilon}_o$ from (8.45) within the domain $0 \leq \widetilde{\epsilon} \leq 1$ define what will be referred to as the DDC bounding curve as depicted in Fig. 8.5. Hence region II in Fig. 8.5 is enclosed by the DDC bounding curve, $\widetilde{\epsilon} = 0$ and, $\widetilde{\epsilon} = 1$, representing the space where the APC is active and the DDC is inactive.

A similar analysis can be followed starting from the case when $\widetilde{\epsilon} \to \infty$, i.e. when the APC is inactive and hence $\alpha \to 0$. If the DDC is still active however, it becomes:

$$|\underset{\sim}{\mathbf{w}}^H_{\text{int}}\hat{\underline{\mathbf{h}}} - 1| \leq \hat{\epsilon} \tag{8.46}$$

When $0 \leq \hat{\epsilon} \leq 1$, then the following relationships can be deduced:

$$\lim_{\substack{\widetilde{\epsilon}\to\infty \\ 0\leq\hat{\epsilon}\leq 1}} \underset{\sim}{\mathbf{w}}_{\text{int}} = (1-\hat{\epsilon})\underset{\sim}{\hat{\mathbf{w}}}_{\text{mvdr}} \tag{8.47}$$

$$\lim_{\substack{\widetilde{\epsilon}\to\infty \\ 0\leq\hat{\epsilon}\leq 1}} \beta = \frac{1}{k_{qq}}\frac{(1-\hat{\epsilon})}{\hat{\epsilon}} \tag{8.48}$$

Finally, for the APC, as $\widetilde{\epsilon}$ is decreased (from initially $\widetilde{\epsilon} \to \infty$), the value, $\widetilde{\epsilon}_o$, at which this constraint becomes active is given by:

$$\widetilde{\epsilon}_o = |\underset{\sim}{\hat{\mathbf{w}}}^H_{\text{mvdr}}\widetilde{\underline{\mathbf{h}}}(1-\hat{\epsilon}) - 1| \tag{8.49}$$

In the limits of $\hat{\epsilon}$, when $\hat{\epsilon} = 1$, then $\widetilde{\epsilon}_o = 1$, and when $\hat{\epsilon} = 0$, then $\widetilde{\epsilon}_o = |\underset{\sim}{\hat{\mathbf{w}}}^H_{\text{mvdr}}\widetilde{\underline{\mathbf{h}}} - 1|$, where depending on $\widetilde{\underline{\mathbf{h}}}$, $|\underset{\sim}{\hat{\mathbf{w}}}^H_{\text{mvdr}}\widetilde{\underline{\mathbf{h}}} - 1| \leq 1$ or $|\underset{\sim}{\hat{\mathbf{w}}}^H_{\text{mvdr}}\widetilde{\underline{\mathbf{h}}} - 1| \geq 1$. The range of values obtained for $\widetilde{\epsilon}_o$ from (8.49) within the domain $0 \leq \hat{\epsilon} \leq 1$ define what will be referred to as the APC bounding curve as depicted in Fig. 8.5. Hence region III in Fig. 8.5 is enclosed by the APC bounding curve, $\hat{\epsilon} = 0$ and, $\hat{\epsilon} = 1$, representing the space where the APC is inactive and the DDC is active.

Finally, in the lower-left region, region IV, both the APC and the DDC become active within the area enclosed by the APC and DDC bounding curve. It should be kept in mind that Fig. 8.5 is only an illustration and that the shape of the area for which the APC and DDC are both active can change depending on the RTF vectors, $\widetilde{\underline{\mathbf{h}}}$ and $\hat{\underline{\mathbf{h}}}$. For instance, Fig. 8.5 shows $|\underset{\sim}{\hat{\mathbf{w}}}^H_{\text{mvdr}}\widetilde{\underline{\mathbf{h}}} - 1| < 1$ and $|\underset{\sim}{\widetilde{\mathbf{w}}}^H_{\text{mvdr}}\hat{\underline{\mathbf{h}}} - 1| < 1$ (points on the axes), whereas it is possible that either of these points may be greater than or equal to one.

8.5 Confidence Metric and Tuning

8.5.1 Confidence Metric

One of the ingredients towards developing a tuning strategy for setting appropriate values for $\widetilde{\epsilon}$ and $\hat{\epsilon}$ is that of a confidence metric, which is indicative of the confidence in the accuracy of the data-dependent RTF vector. In section 3.7 of Chapter 3, it was proposed that a principal generalised eigenvalue resulting from the data-dependent estimation procedure be used as such a confidence metric. In the following, it is proposed again to use such a metric, however, due to the formulation in the PWT domain, the principal eigenvalue, $\hat{\lambda}_1$ from the EVD in (8.25) will be used. For the MVDR-DD in (8.28), the resulting posterior SNR is given by:

$$\widehat{\mathrm{SNR}}_{\mathrm{DD}} = \frac{\underset{\sim}{\hat{\mathbf{w}}}^H_{\mathrm{mvdr}} \underline{\hat{\mathbf{R}}}_{\mathbf{yy}} \underset{\sim}{\hat{\mathbf{w}}}_{\mathrm{mvdr}}}{\underset{\sim}{\hat{\mathbf{w}}}^H_{\mathrm{mvdr}} \underline{\hat{\mathbf{R}}}_{\mathbf{nn}} \underset{\sim}{\hat{\mathbf{w}}}_{\mathrm{mvdr}}} \tag{8.50}$$

where it is recalled that $\underline{\hat{\mathbf{R}}}_{\mathbf{nn}} = \mathbf{I}_{(\mathrm{M_a}+\mathrm{M_e})}$. Substitution of (8.25) and (8.28)[6] into (8.50) results in $\widehat{\mathrm{SNR}}_{\mathrm{DD}} = \hat{\lambda}_1$, i.e. using $\hat{\lambda}_1$ as a confidence metric is equivalent to making a decision based on the resulting posterior SNR from the MVDR-DD.

Similar to (3.68), $\hat{\lambda}_1$ can then be used in a logisitic function to define the confidence metric, $\mathrm{F(l)}$[7]:

$$\mathrm{F(l)} = \frac{1}{1 + e^{-\rho(10\log_{10}(\hat{\lambda}_1(\mathrm{l})) - \lambda_{\mathrm{t}})}} \tag{8.51}$$

where $\mathrm{F(l)} \in [0,\ \ 1]$, ρ controls the gradient of the transition from 0 to 1, and λ_{t} is a threshold (in dB), beyond which $\mathrm{F(l)} \to 1$. Hence, as $\mathrm{F(l)} \to 0$, it will be indicative of a low confidence in the accuracy of the data-dependent RTF vector and as $\mathrm{F(l)} \to 1$, it will be indicative of a high confidence in the accuracy of the data-dependent RTF vector.

8.5.2 Tuning strategy

With the depiction of the space spanned by $\widetilde{\epsilon}$ and $\hat{\epsilon}$ from Fig. 8.5 in mind, a general two-step procedure can be followed to establish a particular tuning strategy:

[6] Recall that $\underset{\sim}{\hat{\mathbf{w}}}_{\mathrm{mvdr}}$ can be equivalently expressed as $\underset{\sim}{\hat{\mathbf{w}}}_{\mathrm{mvdr}} = \hat{\eta}_q^* \, \hat{\mathbf{Q}}\mathbf{e}_1$.

[7] The time index is reintroduced here to reinforce that these quantities are to be computed in each time frame. All frequencies are still treated equivalently.

1. Choose two points on the $\{\hat{\epsilon}, \widetilde{\epsilon}\}$ plane: ϵ_{AP} and ϵ_{DD}. The coordinates of ϵ_{AP}, $\{\hat{\epsilon}_{AP}, \widetilde{\epsilon}_{AP}\}$, will specify the maximum tolerable speech distortions for the case when there is no confidence in the accuracy of the data-dependent RTF vector. The coordinates of ϵ_{DD}, $\{\hat{\epsilon}_{DD}, \widetilde{\epsilon}_{DD}\}$, on the other hand, will specify the maximum tolerable speech distortions for the case when there is complete confidence in the accuracy of the data-dependent RTF vector.

2. Define an appropriate path in order to connect ϵ_{AP} and ϵ_{DD}, where the variation along this path would be a function of the confidence metric, F(l). As F(l) changes in each time-frequency segment, different values of $\hat{\epsilon}$ and $\widetilde{\epsilon}$ will be chosen along this path and subsequently used in the QCQP from (8.31).

Figure 8.6 depicts three examples of how such a general tuning strategy can be interpreted in the $\{\hat{\epsilon}, \widetilde{\epsilon}\}$ plane, where a linear path has been used to connect the points, ϵ_{AP} and ϵ_{DD}. Before further elaborating on Fig. 8.6, however, one possible tuning strategy will be briefly outlined. In this strategy, ϵ_{AP} and ϵ_{DD}, are chosen by making use of the relationship between the MVDR-INT and SDW-MWF [53,128] (recall the discussion on the SDW-MWF from section 2.7.1 of Chapter 2). Although ϵ_{AP} and ϵ_{DD} can in general be chosen without making use of this relation, it is done to highlight how the speech distortion parameter, μ, from the SDW-MWF is related to the maximum tolerable speech distortion parameters of the MVDR-INT, especially as this μ is a well-established trade-off parameter. For the path connecting ϵ_{AP} and ϵ_{DD}, a linear path will be defined using the confidence metric, F(l).

In the PWT domain, the cost function for the SDW-MWF is given by:

$$\underset{\underbrace{\mathbf{w}}}{\text{minimise}} \quad \mu\, \underbrace{\mathbf{w}}^H \underbrace{\mathbf{w}} + \sigma^2_{\mathrm{s_{a,1}}} |\underbrace{\mathbf{w}}^H \underline{\mathbf{h}} - 1|^2 \tag{8.52}$$

which consists of two terms, the first corresponding to the noise PSD after filtering and the second corresponding to the speech distortion. The speech distortion parameter, $\mu \in (0,\ \infty)$ is used to trade-off between the amount of noise reduction and speech distortion, where larger values of μ put more emphasis on reducing the noise and smaller values put more emphasis on reducing the speech distortion. Two separate SDW-MWF formulations can then be considered for $\widetilde{\underline{\mathbf{h}}}$ and $\hat{\underline{\mathbf{h}}}$ respectively:

$$\underset{\underbrace{\mathbf{w}}}{\text{minimise}} \quad \widetilde{\mu}\, \underbrace{\mathbf{w}}^H \underbrace{\mathbf{w}} + \sigma^2_{\mathrm{s_{a,1}}} |\underbrace{\mathbf{w}}^H \widetilde{\underline{\mathbf{h}}} - 1|^2 \tag{8.53}$$

$$\underset{\underbrace{\mathbf{w}}}{\text{minimise}} \quad \hat{\mu}\, \underbrace{\mathbf{w}}^H \underbrace{\mathbf{w}} + \sigma^2_{\mathrm{s_{a,1}}} |\underbrace{\mathbf{w}}^H \hat{\underline{\mathbf{h}}} - 1|^2 \tag{8.54}$$

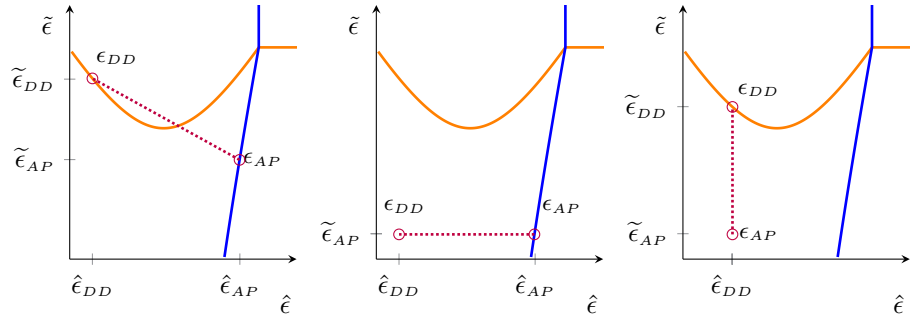


Figure 8.6: Depiction of three different tuning strategies (a) trading off the maximum tolerable speech distortions between the APC and DDC, (b) fixed maximum tolerable speech distortion for the APC but variable maximum tolerable speech distortion for the DDC, and (c) fixed maximum tolerable speech distortion for the DDC but variable maximum tolerable speech distortion for the APC.

where $\widetilde{\mu} \in (0, \ \infty)$ and $\hat{\mu} \in (0, \ \infty)$ are the separate speech distortion trade-off parameters for each cost function. The solutions to (8.53) and (8.54) are then given respectively by:

$$\widetilde{\mathbf{w}}_{\text{sdw}-\text{mwf}} = (\widetilde{\mu}\mathbf{I}_{(\text{M}_\text{a}+\text{M}_\text{e})} + \hat{\sigma}^2_{\text{s}_{\text{a},1}}\underline{\widetilde{\mathbf{h}}}\underline{\widetilde{\mathbf{h}}}^H)^{-1}\hat{\sigma}^2_{\text{s}_{\text{a},1}}\underline{\widetilde{\mathbf{h}}}\underline{\widetilde{\mathbf{h}}}^H\mathbf{e}_1 \tag{8.55}$$

$$\hat{\mathbf{w}}_{\text{sdw}-\text{mwf}} = (\hat{\mu}\mathbf{I}_{(\text{M}_\text{a}+\text{M}_\text{e})} + \hat{\sigma}^2_{\text{s}_{\text{a},1}}\underline{\hat{\mathbf{h}}}\underline{\hat{\mathbf{h}}}^H)^{-1}\hat{\sigma}^2_{\text{s}_{\text{a},1}}\underline{\hat{\mathbf{h}}}\underline{\hat{\mathbf{h}}}^H\mathbf{e}_1 \tag{8.56}$$

where $\hat{\sigma}^2_{\text{s}_{\text{a},1}}$ is an estimate of $\sigma^2_{\text{s}_{\text{a},1}}$. On comparing the $\underset{\sim}{\mathbf{w}}_{\text{int}}$ in (8.33) to (8.55) and (8.56), it can be observed that there is a relationship between the integrated MVDR beamformer and the SDW-MWF. By considering the expressions written as an MVDR beamformer followed by a single-channel post filter as was done in section 3.6.3 of Chapter 3, it can be deduced that:

$$\alpha = \frac{\hat{\sigma}^2_{\text{s}_{\text{a},1}}}{\widetilde{\mu}} \ \text{ when } \ \beta = 0 \tag{8.57}$$

$$\beta = \frac{\hat{\sigma}^2_{\text{s}_{\text{a},1}}}{\hat{\mu}} \ \text{ when } \ \alpha = 0 \tag{8.58}$$

Proceeding to define the coordinates of ϵ_{AP}, (8.57) is firstly substituted into (8.44) to obtain a value for $\widetilde{\epsilon}$ as:

$$\widetilde{\epsilon}_{AP} = \frac{\widetilde{\mu}}{\widetilde{\mu} + \hat{\sigma}^2_{\mathrm{s_{a,1}}} k_{hh}} \tag{8.59}$$

Hence the range of values for $\widetilde{\mu}$ are essentially compressed into a range of values for $\widetilde{\epsilon}_{AP}$ such that $0 \leq \widetilde{\epsilon}_{AP} \leq 1$. This means that $\widetilde{\epsilon}_{AP}$ can be chosen to be within this range without having to specify $\widetilde{\mu}$. However, (8.59) serves to clarify how the choice of $\widetilde{\epsilon}_{AP}$ is related to the cost function of (8.53).

Using the value of $\widetilde{\epsilon}_{AP}$ in (8.45) then yields a range of choices for $\hat{\epsilon}_{AP}$ such that $\hat{\epsilon}_{AP} \leq \hat{\epsilon}_o$:

$$\hat{\epsilon}_{AP} \leq |\widetilde{\underline{\mathbf{w}}}^H_{\mathrm{mvdr}} \hat{\underline{\mathbf{h}}}(1 - \widetilde{\epsilon}_{AP}) - 1| \tag{8.60}$$

If $\hat{\epsilon}_{AP} = |\widetilde{\underline{\mathbf{w}}}^H_{\mathrm{mvdr}} \hat{\underline{\mathbf{h}}}(1 - \widetilde{\epsilon}_{AP}) - 1|$, then ϵ_{AP} lies on the DDC curve of Fig. 8.5. For all values of $\hat{\epsilon}$ such that $\hat{\epsilon} > |\widetilde{\underline{\mathbf{w}}}^H_{\mathrm{mvdr}} \hat{\underline{\mathbf{h}}}(1 - \widetilde{\epsilon}_{AP}) - 1|$, the DDC remains inactive and hence setting a value of $\hat{\epsilon}$ within this region will always result in the same achievable[8] speech distortion defined by $|\widetilde{\underline{\mathbf{w}}}^H_{\mathrm{mvdr}} \hat{\underline{\mathbf{h}}}(1 - \widetilde{\epsilon}_{AP}) - 1|$. Furthermore, when the DDC is inactive, then (8.57) holds, so that values of $\widetilde{\epsilon}$ and $\hat{\epsilon}$ in region II from Fig. 8.5 would result in the SDW-MWF from (8.55).

Similarly, by firstly substituting (8.58) in (8.48) and making use of (8.49), the coordinates $\{\hat{\epsilon}_{DD}, \widetilde{\epsilon}_{DD}\}$ of ϵ_{DD} can be defined as:

$$\hat{\epsilon}_{DD} = \frac{\hat{\mu}}{\hat{\mu} + \hat{\sigma}^2_{\mathrm{s_{a,1}}} k_{qq}} \tag{8.61}$$

$$\widetilde{\epsilon}_{DD} \leq |\hat{\underline{\mathbf{w}}}^H_{\mathrm{mvdr}} \widetilde{\underline{\mathbf{h}}}(1 - \hat{\epsilon}_{DD}) - 1| \tag{8.62}$$

Now if $\widetilde{\epsilon}_{DD} = |\hat{\underline{\mathbf{w}}}^H_{\mathrm{mvdr}} \widetilde{\underline{\mathbf{h}}}(1 - \hat{\epsilon}_{DD}) - 1|$, then ϵ_{DD} lies on the APC bounding curve of Fig. 8.5. Additionally, for all values of $\widetilde{\epsilon}$ such that $\widetilde{\epsilon} > |\hat{\underline{\mathbf{w}}}^H_{\mathrm{mvdr}} \widetilde{\underline{\mathbf{h}}}(1 - \hat{\epsilon}_{DD}) - 1|$, the APC remains inactive and hence setting a value of $\widetilde{\epsilon}$ within this region will always result in the same achievable speech distortion defined by $|\hat{\underline{\mathbf{w}}}^H_{\mathrm{mvdr}} \widetilde{\underline{\mathbf{h}}}(1 - \hat{\epsilon}_{DD}) - 1|$. Furthermore, when the APC is inactive, then (8.58) holds, so that values of $\widetilde{\epsilon}$ and $\hat{\epsilon}$ in region III from Fig. 8.5 would result in the SDW-MWF from (8.56).

With ϵ_{AP} and ϵ_{DD} now defined, the second step of the general tuning procedure is to now define an appropriate path to connect them as a function of the confidence metric, F(l). Here, it is proposed to use F(l) to perform a linear

[8] Achievable here is meant to differentiate between the actual speech distortion that is obtained and the maximum tolerable value that was specified.

interpolation between ϵ_{AP} and ϵ_{DD} to yield the values for $\widetilde{\epsilon}$ and $\hat{\epsilon}$ respectively as:

$$\hat{\epsilon} = (1 - \mathrm{F(l)})\ \hat{\epsilon}_{AP} + \mathrm{F(l)}\ \hat{\epsilon}_{DD} \tag{8.63}$$

$$\widetilde{\epsilon} = (1 - \mathrm{F(l)})\ \widetilde{\epsilon}_{AP} + \mathrm{F(l)}\ \widetilde{\epsilon}_{DD} \tag{8.64}$$

which are subsequently squared to be used in the QCQP from (8.31). Consequently as the confidence in the accuracy of the data-dependent RTF vector increases, the maximum tolerable speech distortions will be specified by values tending towards $\{\hat{\epsilon}_{DD}, \widetilde{\epsilon}_{DD}\}$. On the contrary, as this confidence decreases, maximum tolerable speech distortions will be specified by values tending towards $\{\hat{\epsilon}_{AP}, \widetilde{\epsilon}_{AP}\}$.

The insight of Fig. 8.5 and additional value of the MVDR-INT as compared to the SDW-MWF is now apparent. In Fig. 8.5, SDW-MWF with the a priori RTF vector, (8.55), corresponds to regions I and II, and the SDW-MWF with the data-dependent RTF vector, (8.56), corresponds to region I and III. Conventionally, either of these SDW-MWFs is used in practice and hence the solution space is constrained within regions I and II or regions I and III. The MVDR-INT, on the other hand, introduces an additional region, namely region IV, where it is observed that specific values of $\widetilde{\epsilon}$ and $\hat{\epsilon}$ can be set so as to optimally (with respect to the QCQP) merge or integrate the a priori and data-dependent spatial information. Region IV also serves as a "bridge" for connecting regions II and III, which then allows for tuning strategies to be developed such as in Fig. 8.6 (a), where optimal combinations of the SDW-MWFs from (8.55) and (8.56) can be made by specifying $\widetilde{\epsilon}$ and $\hat{\epsilon}$. Region IV can in fact be related to the convex combination of filters in (8.39) as it can be deduced that each of the corners of region IV depicted in Fig. 8.5 corresponds to one of the filters (8.39) (top-right: $\mathbf{w}_{\mathrm{z}}$, bottom-right: $\underset{\sim}{\widetilde{\mathbf{w}}}_{\mathrm{mvdr}}$, bottom-left: $\underset{\sim}{\mathbf{w}}_{\mathrm{lcmv}}$, and top-left: $\underset{\sim}{\hat{\mathbf{w}}}_{\mathrm{mvdr}}$). Moreover, as the solution space with the MVDR-INT now covers regions I, II, III, and IV, it encapsulates a wider range of speech enhancement filters, which can be advantageous in accommodating various acoustic scenarios.

Returning focus to Fig. 8.6, the three examples of a tuning strategy can now be understood. A particular realisation of the APC and the DDC bounding curves has been plotted and the intersecting point of both curves corresponds to the $\{1, 1\}$ coordinate (recall Fig. 8.5). In the tuning of Fig. 8.6 (a), as F(l) increases, the path along the dotted line is taken from ϵ_{AP} to arrive at ϵ_{DD} which gradually sets a larger value of $\widetilde{\epsilon}$ for the APC and a smaller value of $\hat{\epsilon}$ for the DDC. Depending on the particular realisation of the APC and DDC bounding curves, it may be that such a path can entirely lie within the area enclosed by these curves or part of it may lie outside as shown in Fig. 8.6 (a). The latter is in fact a fortunate circumstance because the achieved speech distortion

corresponding to the inactive constraint will actually be lower than what was prescribed by the tuning. In the case of Fig. 8.6 (a) for instance, when the linear path is above the APC bounding curve, it means that $\widetilde{\epsilon} > |\hat{\underset{\sim}{\mathbf{w}}}^H_{\text{mvdr}}\widetilde{\underline{\mathbf{h}}}(1-\hat{\epsilon})-1|$ (recall (8.49)). Since beyond $|\hat{\underset{\sim}{\mathbf{w}}}^H_{\text{mvdr}}\widetilde{\underline{\mathbf{h}}}(1-\hat{\epsilon})-1|$ the APC continues to be inactive, the actual speech distortion that would be achieved in relation to this constraint would correspond to $|\hat{\underset{\sim}{\mathbf{w}}}^H_{\text{mvdr}}\widetilde{\underline{\mathbf{h}}}(1-\hat{\epsilon})-1|$, which is by definition less than $\widetilde{\epsilon}$. Hence, although there is a linear path from ϵ_{AP} to ϵ_{DD}, at the point where this linear path intersects with the APC bounding curve, the actual speech distortions that would be achieved are those that continue along the APC bounding curve in order to arrive at ϵ_{DD}.

The tunings depicted in Fig. 8.6 (b) and (c) are representative of strategies where the maximum tolerable speech distortion is fixed for one of the constraints, and only the maximum tolerable speech distortion for the other constraint is tuned. In Fig. 8.6 (b), ϵ_{DD} is defined by setting $\widetilde{\epsilon}_{DD} = \widetilde{\epsilon}_{AP}$, so that the maximum tolerable speech distortion for the APC is fixed. $\hat{\epsilon}$ is then tuned according to (8.63). This is representative of a case where the APC is always active and the DDC is only included if there is confidence in the accuracy of the data-dependent RTF vector. Fig. 8.6 (c) depicts an opposite strategy, where now ϵ_{AP} is set by setting $\hat{\epsilon}_{AP} = \hat{\epsilon}_{DD}$, so that the maximum tolerable speech distortion for the DDC is fixed.

8.6 Evaluation and discussion

In order to gain further insight into the behaviour of the integrated MVDR beamformer using the QCQP formulation, a simulation was firstly considered involving only an LMA without XMs. As will be demonstrated, observing such a scenario facilitates the visualisation of the theoretical beam patterns that would be generated under different tuning strategies. Following this simulation, recorded data from an acoustic scenario involving behind-the-ear (BTE) dummy[9] hearing aids (HAs) along with XMs in a cocktail party scenario was then analysed and evaluated.

8.6.1 Beam patterns for a linear microphone array

Although in chapter 4 beam patterns were developed for the microphone configuration of an LMA with XMs, in the following the beam patterns are

[9]This means that only the microphone signals alone, without any processing are captured.

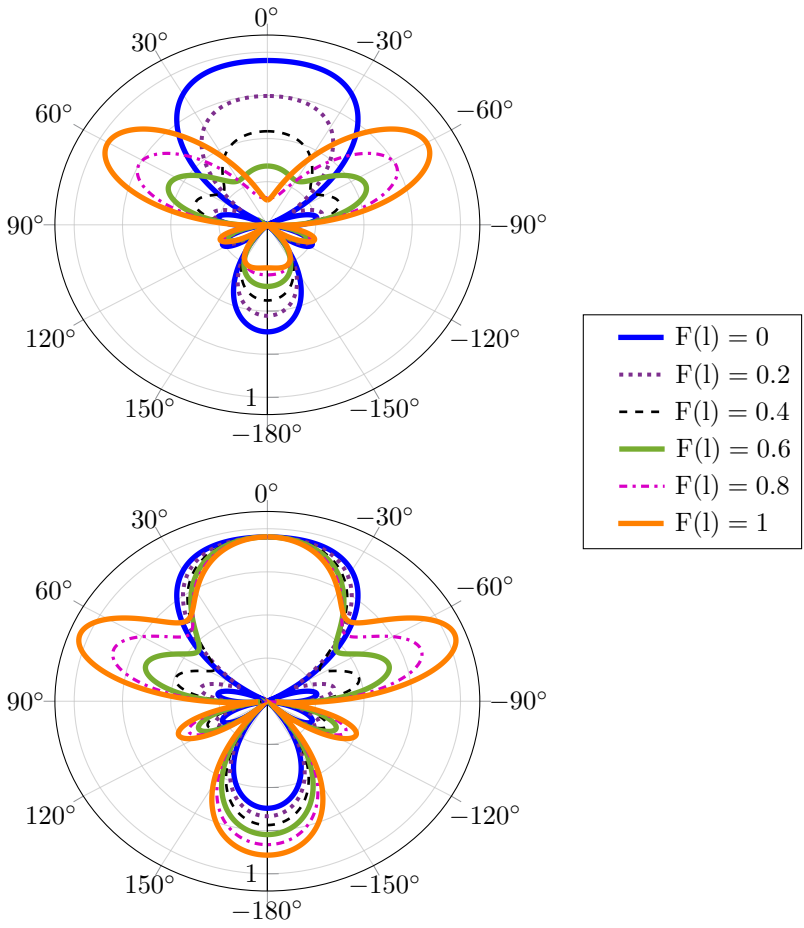


Figure 8.7: Beam patterns as a function of the confidence metric, F(l), for different tunings of the MVDR-INT as applied to a microphone configuration consisting of an LMA only. (Top) A tuning strategy similar to that depicted in Fig. 8.6 (a) and (Bottom) A tuning strategy similar to that depicted in Fig. 8.6 (b). F(l) = 0 corresponds to the position ϵ_{AP} and F(l) = 1 corresponds to the position ϵ_{DD} from Figs. 8.6. As F(l) increases, the path from ϵ_{AP} to ϵ_{DD} is followed resulting in the depicted beam patterns.

considered for an LMA only as it provides a better intuition of the behaviour of the MVDR-INT.

For visualising the beam patterns, a linear LMA consisting of 4 microphones and

5 cm spacing was considered. Two anechoic RTF vectors, simulating an a priori RTF vector, $\widetilde{\mathbf{h}}_{\mathbf{a}}$, and a data-dependent RTF vector, $\hat{\mathbf{h}}_{\mathbf{a}}$, were computed according to a far-field approximation, i.e. $[1\ e^{-j2\pi f\tau_2(\theta)}\ e^{-j2\pi f\tau_3(\theta)}\ e^{-j2\pi f\tau_4(\theta)}]^T$, where f is the frequency (Hz) which was set to 3 kHz, $\tau_m(\theta) = \frac{(m-1)0.05\cos(\theta)}{c}$ is the relative time delay between the mth microphone and the reference microphone (the microphone closest to the desired speech source) of the LMA, θ is the angle of the desired speech source, and $c = 345\,\mathrm{m}\cdot\mathrm{s}^{-1}$ is the speed of sound. For $\widetilde{\mathbf{h}}_{\mathbf{a}}$, $\theta = 0°$ and for $\hat{\mathbf{h}}_{\mathbf{a}}$, $\theta = 60°$. Using this definition of $\widetilde{\mathbf{h}}_{\mathbf{a}}$, $\widetilde{\mathbf{C}}_{\mathbf{a}}$ was defined as in (5.5) and $\tilde{\mathbf{f}}_{\mathbf{a}} = \frac{\widetilde{\mathbf{h}}_{\mathbf{a}}}{||\widetilde{\mathbf{h}}_{\mathbf{a}}||^2}$ and $\widetilde{\boldsymbol{\Upsilon}}_{\mathbf{a}} = [\widetilde{\mathbf{C}}_{\mathbf{a}}\ \tilde{\mathbf{f}}_{\mathbf{a}}]$. With $\mathbf{R}_{\mathbf{n_a n_a}} = \mathbf{I}_{\mathrm{M_a}}$, the pre-whitening operation from (5.43) was then computed but with $\widetilde{\boldsymbol{\Upsilon}}_{\mathbf{a}}$ instead of $\widetilde{\boldsymbol{\Upsilon}}$, and hence denoted as $\mathbf{L}_{\mathbf{a}}$. In the PWT domain, the respective RTF vectors are given by $\underline{\widetilde{\mathbf{h}}}_{\mathbf{a}} = \mathbf{L}_{\mathbf{a}}^{-1}\widetilde{\boldsymbol{\Upsilon}}_{\mathbf{a}}^{H}\widetilde{\mathbf{h}}_{\mathbf{a}}$ and $\underline{\hat{\mathbf{h}}}_{\mathbf{a}} = \mathbf{L}_{\mathbf{a}}^{-1}\widetilde{\boldsymbol{\Upsilon}}_{\mathbf{a}}^{H}\hat{\mathbf{h}}_{\mathbf{a}}$. The optimal PWT domain filters, $\underline{\widetilde{\mathbf{w}}}_{\mathbf{a}}$, and $\underline{\hat{\mathbf{w}}}_{\mathbf{a}}$ were then computed as in (8.14), but using either $\underline{\widetilde{\mathbf{h}}}_{\mathbf{a}}$ or $\underline{\hat{\mathbf{h}}}_{\mathbf{a}}$. Finally, (8.63) and (8.64) were used to $\widetilde{\epsilon}$ and $\hat{\epsilon}$, after which (8.31) was then solved using CVX [290, 291] to yield the integrated MVDR beamformer for the LMA only, denoted as $\underline{\mathbf{w}}_{\mathbf{a},\mathrm{int}}$. The beam patterns were computed as $|\underline{\mathbf{w}}_{\mathbf{a},\mathrm{int}}^{H}\underline{\mathbf{h}}(\theta)|$, where $\underline{\mathbf{h}}(\theta)$ is the PWT domain RTF vector corresponding to an angle, θ.

Figure 8.7 illustrates the resulting beam patterns for two tuning strategies for different values of F(l) (in this case l = 1). The left-hand plot of Fig. 8.7 corresponds to a tuning strategy similar to that depicted in Fig. 8.6 (a), where there is a trade-off between the two constraints. For this strategy, $\widetilde{\mu} = \hat{\mu} = 0.2$, which means that ϵ_{AP} and ϵ_{DD} were fairly close to the x-axis and y-axis respectively. As F(l) increases, the beam pattern is clearly seen to evolve from focusing on the a priori direction of 0° to eventually that of the data-dependent direction of 60°. As a linear path is followed, at the midpoint, both $\widetilde{\epsilon}$ and $\hat{\epsilon}$ are of a similarly larger values, which explains the nature of the lower magnitude in the beam pattern during the transition.

The right-hand plot of Fig. 8.7 corresponds to a tuning strategy as depicted in Fig. 8.6 (b), i.e. when the APC is always active. As F(l) increases, it can be observed that the beam in the a priori direction of 0° is maintained, while more gain is attributed to the data-dependent direction of 60°. In this particular case, however, it is noted that although the response at 60° is in accordance with the maximum tolerable speech distortion prescribed, there is a slight tilt of the beam towards 68° as compared to if only the DDC was active. Nevertheless, this can still be a useful tuning strategy for cases when a high confidence is placed on the a priori RTF vector.

8.6.2 Effect of $\widetilde{\epsilon}$ and $\hat{\epsilon}$

In this section the effect of $\widetilde{\epsilon}$ and $\hat{\epsilon}$ on the behaviour of the integrated MVDR beamformer for the case of an LMA and XMs is further investigated using recorded audio data. A batch processing framework will be applied so as to observe an average performance of individual frequencies. In the following section, the processing will be done using a Weighted Overlap and Add (WOLA) framework [101] and a broadband performance will be assessed.

The database of speech and noise recordings as described in section 5.6.3 of Chapter 5 was used for the evaluation in this section as well as the following section. The scenario is depicted in Fig. 8.8, with the Neumann KU-100 dummy head in a central location of the room and equipped with the two (i.e. left and right) BTE HAs. Hence, in the following, the LMA is considered as having a total of four microphones, i.e. the stacked left ear and right ear microphones. The first microphone of the left ear HA, $y_{a,1}$,was used as the reference microphone. Three omnidirectional XMs (XM1 was an AKG CK97O microphone; XM2 and XM3 were AKG CK32 microphones) in static positions as shown were used in conjunction with the BTE HAs. The speech signal was that from a male speaker [94] and was considered to be either at 0° or 60°. The noise was that of the re-created cocktail party noise described in section 5.6.3 of Chapter 5.

As opposed to a free-field a priori RTF vector, a more suitable a priori RTF vector for the BTE-HAs was obtained from a pre-measured impulse response in the scenario as depicted in Fig. 8.8, specifically for when a source was located at 0° (the azimuth direction directly in front of the dummy head) and 1 m from the dummy head. This was obtained in an identical manner as described in section 5.6.2 of Chapter 5, which is denoted as $\widetilde{\mathbf{h}}_{\mathbf{a}}(\mathrm{k})$, and was kept the same for each time frame.

A scenario was firstly considered for the desired speech source located at 0° in Fig. 8.8, i.e. the location where the a priori RTF vector was defined. A 4 s sample of the desired speech signal was mixed with a random sample of the cocktail party noise at a broadband input SNR of 0 dB. For the batch processing framework with a DFT size of 256 samples, $\mathbf{R}_{\mathbf{yy}}$ and $\mathbf{R}_{\mathbf{nn}}$ were estimated by time averaging across the entire length of the signal in the respective speech-plus-noise or noise-only frames. Using the SPP [192] from the first microphone of the left ear HA, frames for which the speech was active were chosen if the resulting SPP > 0.85. The RTF vectors, $\underline{\widetilde{\mathbf{h}}}$ and $\underline{\hat{\mathbf{h}}}$, were computed according to the procedures described in sections 8.3.1 and 8.3.2. Using CVX [290, 291], the MVDR-INT from (8.31) was then evaluated for a range of $0 < \widetilde{\epsilon} < 1.5$ and $0 < \hat{\epsilon} < 1.5$ at a frequency of 2 kHz.

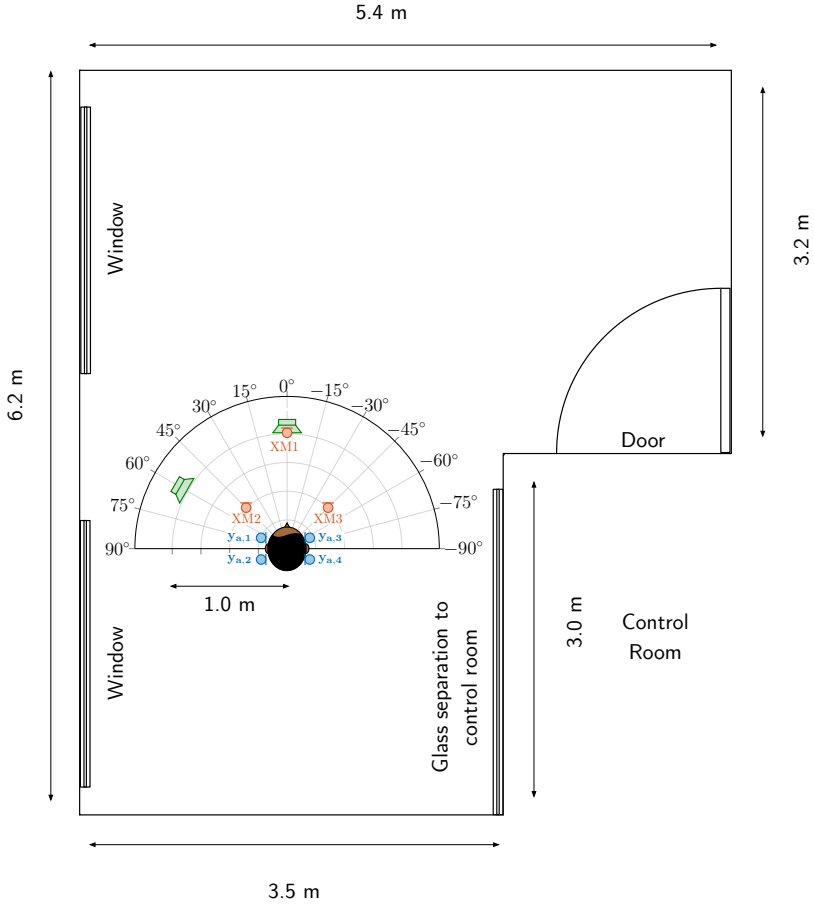


Figure 8.8: Spatial scenario for the audio recordings. Separate recordings were made of speech signals from the loudspeakers positioned at 0° and 60°. These were then mixed with a re-created cocktail party type noise as explained in section 5.6.3 of Chapter 5.

Figures 8.9 (a) and (b) display the resulting (base-10 log) values of the Lagrangian multipliers α and β respectively as a function of $\widetilde{\epsilon}$ and $\hat{\epsilon}$, along with the APC and DDC bounding curves. These plots support the theoretical analysis of the space spanned by $\widetilde{\epsilon}$ and $\hat{\epsilon}$ from Fig. 8.5. In Fig. 8.9 (a), it is clearly observed that as the value of $\widetilde{\epsilon}$ exceeds the APC bounding curve, then $\alpha \to 0$ so that the APC is inactive while the DDC remains active. Similarly, in Fig. 8.9 (b), as the value of $\hat{\epsilon}$ exceeds the DDC bounding curve, then $\beta \to 0$ so

that the APC remains active and the DDC is inactive. The regions where both constraints are active, and when neither are active can also be observed.

Figures 8.9 (c) and (d) are plots of the corresponding change in SNR (ΔSNR) from the reference microphone as well as the speech distortion which was computed as follows:

$$\Delta\text{SNR} = 10\log_{10}\left(\frac{|\underset{\sim}{\mathbf{w}}_{\text{int}}^H\underline{\mathbf{h}}|^2}{\underset{\sim}{\mathbf{w}}_{\text{int}}^H\underset{\sim}{\mathbf{w}}_{\text{int}}}\right) - 10\log_{10}\left(\frac{1}{\mathbf{e}_1^T\hat{\mathbf{R}}_{\mathbf{nn}}\mathbf{e}_1}\right) \tag{8.65}$$

$$\text{SD} = |\underset{\sim}{\mathbf{w}}_{\text{int}}^H\underline{\mathbf{h}} - 1|^2 \tag{8.66}$$

where the first term of the ΔSNR is the output SNR and the second term is the input SNR at the unprocessed reference microphone[10] and in this scenario $\underline{\mathbf{h}} = \widetilde{\underline{\mathbf{h}}}$. The true value of $\underline{\mathbf{h}}$ is unknown, hence the results of Figs. 8.9 (c) and (d) are suggestive for the case when the true RTF vector corresponds to that of the a priori assumed RTF vector. In Fig. 8.9 (c), since $\underset{\sim}{\mathbf{w}} \to \mathbf{0}$ in the region where $\hat{\epsilon} \geq 1$ and $\widetilde{\epsilon} \geq 1$, it is purposefully hatched so as to indicate that in this region an output SNR is undefined.

As expected, it can be observed that the best ΔSNR is achieved for the region where the DDC is inactive and the APC is active, with a compromise within the region where the two constraints are active. An interesting observation here is the poor ΔSNR in the region where $\widetilde{\epsilon} \to 0$ and $\hat{\epsilon} \to 0$. Even though the maximum tolerable speech distortions have been specified to be quite small, in this case $\widetilde{\underline{\mathbf{h}}}$ and $\hat{\underline{\mathbf{h}}}$ can be parallel, which can lead to redundant constraints and an ill-conditioning problem as discussed in section 3.5 pf Chapter 3. In terms of the SD, fairly low distortions are achieved when either of the constraints are active or when both are active. As both $\widetilde{\epsilon} \to 1$ and $\hat{\epsilon} \to 1$, the speech distortion increases, which is expected from (8.59) and (8.61), i.e. the SDW-MWF parameters, $\widetilde{\mu}$ and $\hat{\mu}$. As $\widetilde{\mu} \to \infty$, $\widetilde{\epsilon} \to 1$, and as $\hat{\mu} \to \infty$, $\hat{\epsilon} \to 1$, which accounts for the increasing speech distortion from Fig. 8.9 (d). Another point to highlight in Fig. 8.9 (d) is that a low speech distortion is also achieved in the region where the APC bounding curve is a minimum, regardless the value of $\widetilde{\epsilon}$. As discussed in section 8.5.2, for a value of $\widetilde{\epsilon} > \widetilde{\epsilon}_o$ (where $\widetilde{\epsilon}_o$ is the value of $\widetilde{\epsilon}$ on the APC bounding curve from (8.49)), the achievable distortion would in fact correspond to $\widetilde{\epsilon}_o$ on the APC bounding curve, which is quite low in this minimum region.

Figure 8.10 now displays a similar set of results, however for the case when the desired speech source was located at 60° as depicted in Fig. 8.8. As the a priori RTF vector was based on a speaker located at 0°, this scenario represented a

[10]The numerator of this term is 1 since the first component of the RTF vector for the unprocessed microphone signals is 1.

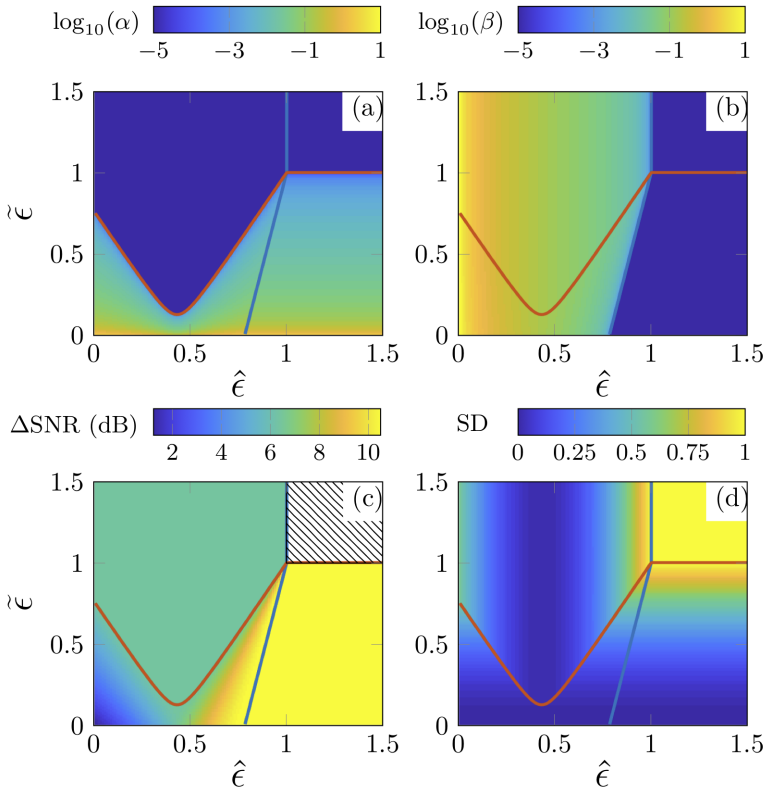


Figure 8.9: Behaviour of the integrated $\text{MVDR}_{\text{LMA-XM}}$ beamformer as a function of $\tilde{\epsilon}$ and $\hat{\epsilon}$ for the case when the desired speech source is at 0°, i.e., in the direction of the a priori constraint. (a) Lagrangian multiplier, $\log_{10}(\alpha)$, (b) Lagrangian multiplier, $\log_{10}(\beta)$, (c) ΔSNR, (d) Speech Distortion (SD). The APC and DDC bounding curves analogous to those from Fig. 8.5 are also shown.

mismatch between the a priori RTF vector and the true RTF vector. The same procedure as previously described was also followed to obtain the MVDR-INT filters.

Figures 8.10 (a) and (b) display the resulting values of the (base-10 log) Lagrangian multipliers α and β respectively as a function of $\tilde{\epsilon}$ and $\hat{\epsilon}$, along with the APC and DDC bounding curves. The nature of these plots is quite similar to that of Figs. 8.9 (a) and (b) in terms of how α and β vary with respect to the bounding curves. In comparison to Figs. 8.9 (a) and (b), Figs. 8.10 (a) and

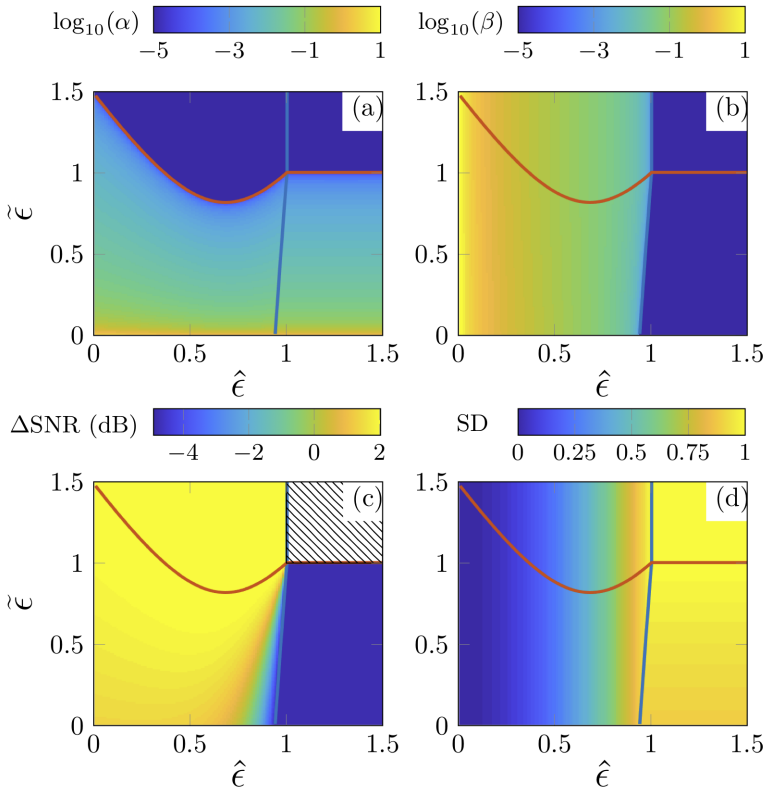


Figure 8.10: Behaviour of the integrated $\text{MVDR}_{\text{LMA-XM}}$ beamformer as a function of $\widetilde{\epsilon}$ and $\hat{\epsilon}$ for the case when the source is at 60°, i.e., not in the direction of the a priori constraint. (a) Lagrangian multiplier, $\log_{10}(\alpha)$, (b) Lagrangian multiplier, $\log_{10}(\beta)$, (c) ΔSNR, (d) Speech Distortion (SD). The APC and DDC bounding curves analogous to those from Fig. 8.5 are also shown.

(b) also highlight the fact that these bounding curves can have quite different appearances.

Figures 8.10 (c) and (d) display the corresponding ΔSNR and SD respectively, however with $\underline{\mathbf{h}} = \hat{\underline{\mathbf{h}}}$ in (8.65) and hence the results are suggestive for the case when the true RTF vector corresponds to that of the data dependent RTF vector. Now it can be observed that the best ΔSNR is achieved for the region where the APC is inactive and the DDC is active, with a compromise within the region where the two constraints are active. For the SD, fairly low speech distortions are achieved for small values of $\hat{\epsilon}$ as expected. For small values of $\widetilde{\epsilon}$

and large values of $\hat{\epsilon}$, i.e. toward the region where only the APC is active, it can be observed that the speech distortion increases, which is a direct result of the speech source not being in the a priori defined direction of 0°. Once again, it can also be seen that the speech distortion generally increases as both $\widetilde{\epsilon} \to 1$ and $\hat{\epsilon} \to 1$.

The results of Figs. 8.9 and 8.10 provide some more insight into the behaviour of the MVDR-INT and demonstrate that in some scenarios a better performance can be achieved when either only the APC or only the DDC is active. Furthermore, it was observed that there were transition regions where a compromise could be achieved between these limits of performance when either only the APC or only the DDC is active. Therefore it suggests that tuning strategies such as those depicted in Fig. 8.6 would indeed be appropriate means of obtaining an optimal filter as opposed to relying on only an APC or DDC.

8.6.3 Performance of tuning strategies

The audio recordings as previously described for the scenario depicted in Fig. 8.8 were also used to observe the performance of the tuning strategies. A desired speech signal was created where the desired speech source was initially located at 0° for a duration of 5 s, and then instantaneously moved to 60° for another 6 s. This was then mixed with a random sample of the cocktail party noise at a broadband input SNR of 2 dB. The same a priori RTF vector pertaining to the HAs, $\widetilde{\mathbf{h}}_{\mathbf{a}}(\mathrm{k})$, as previously described was used, i.e., $\widetilde{\mathbf{h}}_{\mathbf{a}}(\mathrm{k})$ was computed for a source located at 0° and 1 m from the dummy head.

For the STFT processing, the WOLA method, with a DFT size of 256 samples, 50 % overlap, a square-root hanning window, and a sampling frequency of 16 kHz were used. By using the SPP [192] computed on a particular microphone, frames were classified as containing speech if the SPP > 0.8, otherwise the frames were classified as noise-only. All RTF vector estimates were performed in frames which were classified as containing speech. All the relevant correlation matrices were also estimated using a forgetting factor corresponding to a time constant of 300 ms. $\mathbf{R}_{\mathbf{nn}}$ was only estimated when the SPP < 0.8.

For the MVDR-INT, two tuning strategies were considered - (i) the tradeoff between the maximum tolerable speech distortions for the APC and DDC, corresponding to Fig. 8.6 (a), which will be referred to as MVDR-INT-a and (ii) where the maximum tolerable speech distortion for the APC is constant, but the maximum tolerable speech distortion for the DDC varies, corresponding to Fig. 8.6 (b), and which will be referred to as MVDR-INT-b. For both tunings, $\widetilde{\mu} = \hat{\mu} = 0.001$, and $\hat{\sigma}^2_{\mathrm{s_{a,1}}}$ was computed using the method from [204] as implemented in [210] but with the noise estimation update computed as in [192].

A different setting was used for the confidence metric, F(l) in (3.68) for each of the tunings such that for the MVDR-INT-a, $\rho = 1$ and $\lambda_t = 5\,\text{dB}$, and for MVDR-INT-b, $\rho = 1$ and $\lambda_t = 10\,\text{dB}$, i.e. a higher thresholding was used for the MVDR-AP tuning. With all parameters assigned, the QCQP problem from (8.31) was solved using the gradient ascent procedure as described in Algorithm 1.

The metrics used to evaluate the following experiments were the change in Speech Intelligibility-weighted SNR improvement [159] (Δ SI-SNR) from the input SI-SNR at $y_{a,1}$, the change in short-time objective intelligibility (Δ STOI) [155] from a reference speech-only component in $y_{a,1}$, and the normalised speech-to-reverberation modulation energy ratio for cochlear implants, (SRMR-CI) [163]. For the STOI metric, the reference signal used was the unprocessed desired speech source convolved with 256 samples (i.e, same length as the DFT size) of the (pre-measured) impulse response from the desired speech signal location to the reference microphone. As the room was quite reverberant, however, a true reference signal is somewhat ambiguous to define, and hence the non-intrusive metric, SRMR-CI, suitable for hearing instruments, in particular cochlear implants was also used.

Computing SPP using XM2

Figure 8.11 displays the performance of the various algorithms, where all the metrics have been computed in 2 s time frames with a 25 % overlap for the case when the SPP was computed on XM2. The relative improvements of the SI-SNR and the STOI metrics in relation to the reference microphone have been plotted. The metrics for XM1 and XM2 from Fig. 8.8 is also plotted. In order to contextualise the values of the SRMR-CI metric, an additional plot of the performance for the reference signal (that which was used for the STOI metric) is displayed. From all the metrics, as expected the MVDR-AP performs better than the MVDR-DD in the first 5 s as the speech source was at 0°, i.e. the a priori direction. However, in the latter 6 s, when the speech source was at 60°, the MVDR-DD achieves a better performance.

With respect to the XMs, it can also be seen that the performance of XM1 decreases after 5 s as the source moves to the location of 60°, while XM2 has more of a consistent performance across the different speech locations. In terms of the Δ SI-SNR, the performance of all of the other algorithms are better than either of the XMs, which demonstrates that simply listening to the XM only would not always immediately yield satisfactory performance.

Within the first 5 s, the MVDR-INT-a is able to find a compromise between the MVDR-AP and MVDR-DD in terms of all metrics. In the final 6 s, although

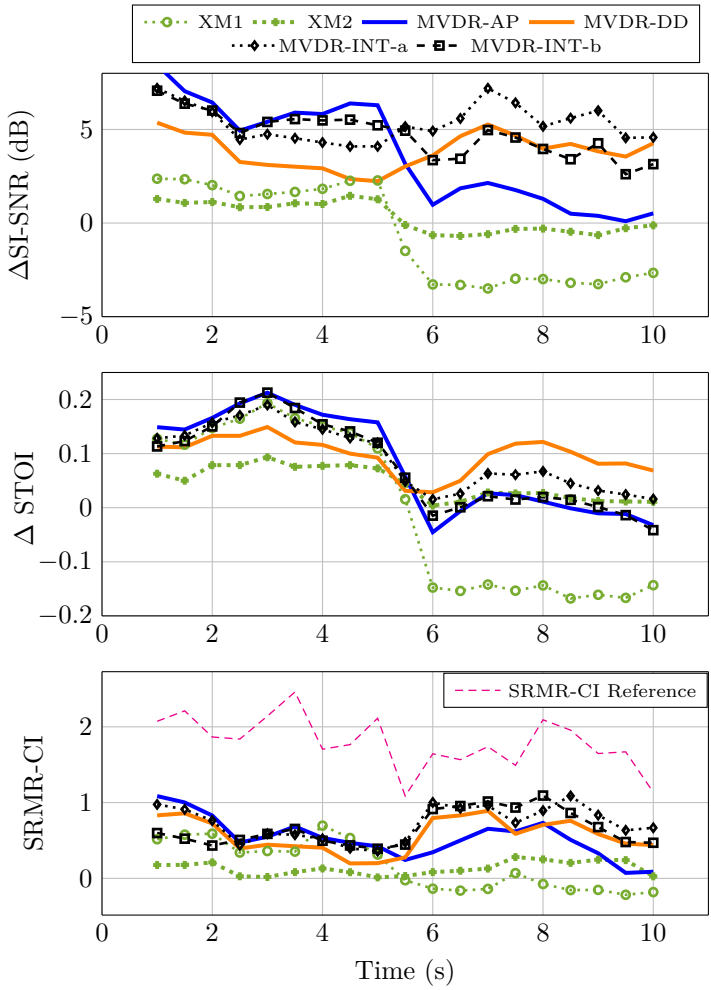


Figure 8.11: Performance of the MVDR-AP, MVDR-DD, and two tunings of the integrated MVDR beamformer, MVDR-INT-a and MVDR-INT-b along with XM1 and XM2 from Fig. 8.8 when SPP computed on XM2.

the Δ STOI is once again in between the MVDR-AP and MVDR-DD, the performance in terms of Δ SI-SNR and SRMR-CI is in fact better than either of the MVDR-AP or the MVDR-DD. This is a direct consequence of the nature of the MVDR-INT beamformer as different linear combinations of the MVDR-AP and the MVDR-DD are effectively applied to different time-frequency segments, yielding a broadband SI-SNR that could be better than either the MVDR-AP

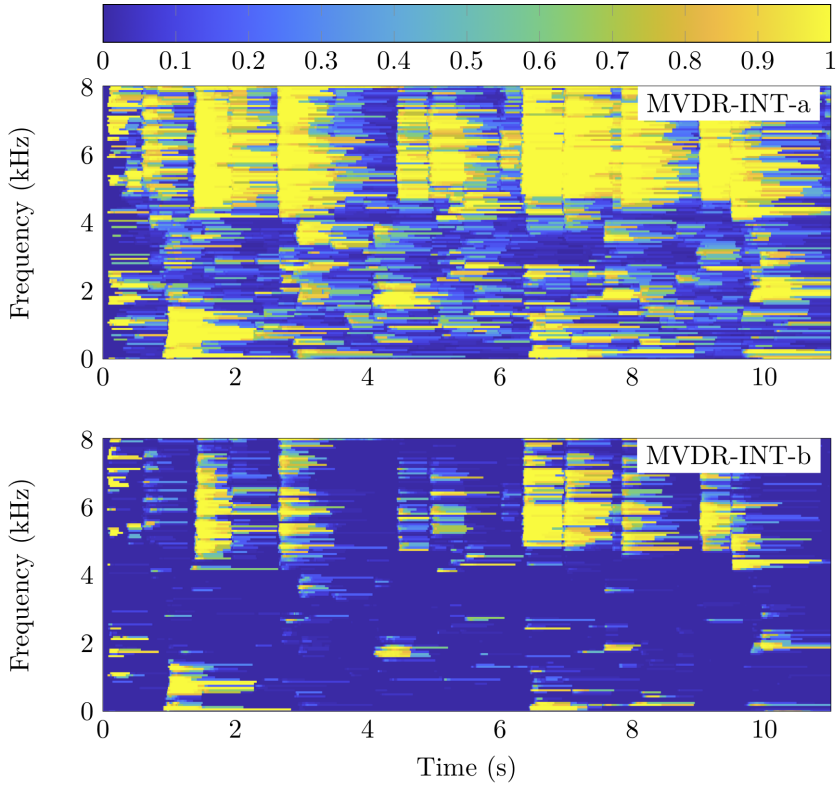


Figure 8.12: Confidence metrics of the evaluation performed in Fig. 8.11 for (Top) MVDR-INT-a and (Bottom) MVDR-INT-b.

or MVDR-DD.

For the MVDR-INT-b, within the first 5 s, the performance in terms of all the metrics is closer to that of the MVDR-AP which is expected as the APC is kept active at all times. In the following 6 s, the STOI metric indicates that the speech intelligibility has not changed from that of the MVDR-AP. However, an improvement can be observed in both Δ SI-SNR and SRMR-CI metrics as some frequency bins would have also had the DDC active.

The corresponding confidence metrics across all time frames and frequencies for the MVDR-INT-a and the MVDR-INT-b are displayed in Fig. 8.12. The upper plot corresponds to the confidence metric of MVDR-INT-a, and reveals that much of the confidence has been placed on the higher frequencies, presumably because there was less noise in this region. Therefore a smaller value of $\hat{\epsilon}$ and a

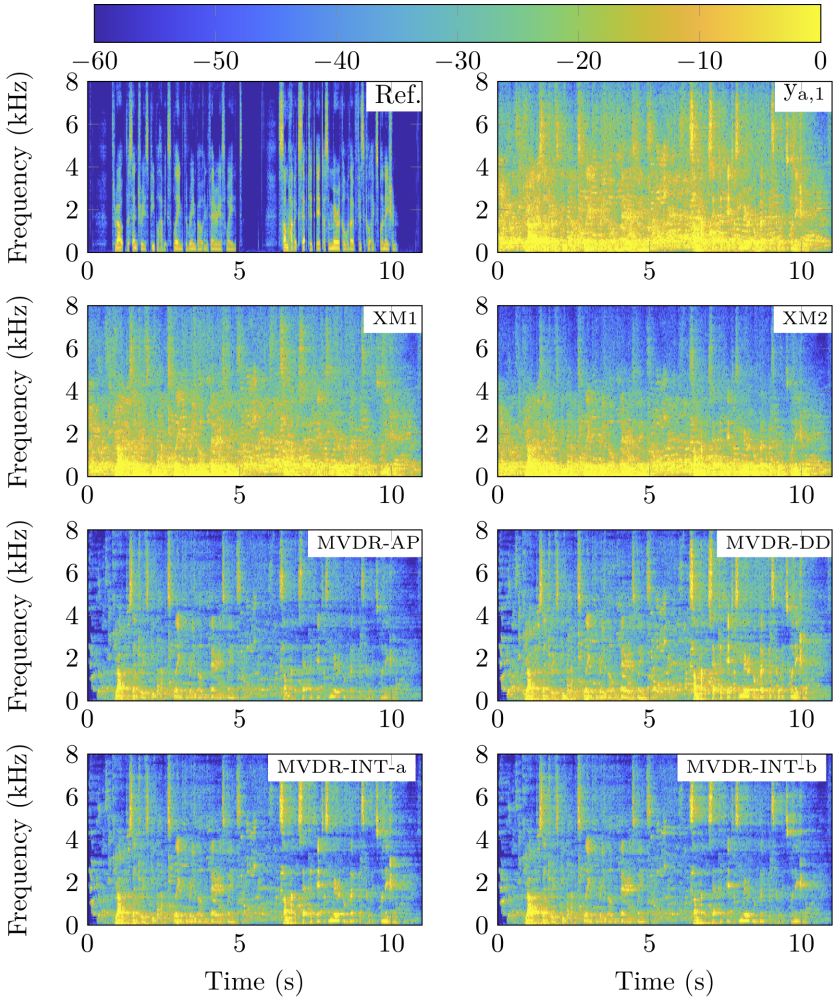


Figure 8.13: Spectrograms corresponding to the signals evaluated from Fig. 8.11. Also included is the the reference speech-only signal used in computing Δ STOI (Ref) and the noisy reference microphone signal, $y_{a,1}$. The colour bar indicates the intensity of the STFT coefficients (in dB), which has been constrained to the range as shown for visual clarity (since the intensity of the XM STFT coefficients were substantially greater than those on the reference signal).

larger value of $\widetilde{\epsilon}$ would have been assigned to the DDC and APC respectively,

i.e. the MVDR-INT-a in this region would have tended toward the MVDR-DD. Several regions of uncertainty are also observed where the MVDR-INT-a would then find a compromise between the MVDR-AP and the MVDR-DD. In the lower plot of Fig. 8.12, the confidence metric for the MVDR-INT-b shows a much more conservative behaviour due to the larger threshold of λ_t. It is observed that there are now many regions where there is little confidence, and hence a larger value of $\hat{\epsilon}$ and a smaller value of $\widetilde{\epsilon}$ would have been assigned to the DDC and APC respectively, i.e. the MVDR-INT-b in these regions would have tended toward the MVDR-AP. More confidence is now only placed in the higher frequency region and there are still some regions of uncertainty so that a compromise can be achieved.

Figure 8.13 displays the corresponding spectrograms for the reference speech-only signal used in the computation of the Δ STOI metric, the noisy reference microphone signal, $y_{a,1}$, two of the external microphone signals (XM1 and XM2), and the corresponding processed signals from the various beamformers. It can be observed that all input microphone signals are quite noisy. The differences in the responses of the different microphones used for XM1 and XM2 are also evident. With respect to the processed signals, it can be seen that within the first 5 s the MVDR-AP results in a less noisy signal that the MVDR-DD, whereas in the final 6 s, the MVDR-DD retains much more of the speech signal in comparison to the MVDR-AP, which is supportive of the results from Fig. 8.13. For the MVDR-INT-a and MVDR-INT-b, the integration of both the MVDR-AP and MVDR-DD is more evident as opposed to only observing the metrics. It can be seen that the resulting speech signal is directly related to the confidence metrics from Fig. 8.12.

Computing SPP using $y_{a,1}$

Figure 8.14 now shows the performance for the case when the SPP was computed on $y_{a,1}$, while all other conditions as in the previous experiment remained the same. As $y_{a,1}$ would have been more noisy in comparison to XM2, it was expected that there would have been more errors in the estimated correlation matrices as opposed to when the SPP was computed on XM2. This is evident from the decrease in the values of all the metrics for the MVDR-AP and the MVDR-DD as compared with Fig. 8.11. The general trend of the MVDR-AP and the MVDR-DD however remain the same in the MVDR-AP performs better for the desired speech source at 0° and the MVDR-DD better when the desired speech source is at 60°. Both integrated MVDR beamformers demonstrate a similar behaviour in comparison to that of Fig. 8.11.

Within the first 5 s, the MVDR-INT-a and the MVDR-INT-b finds a compromise

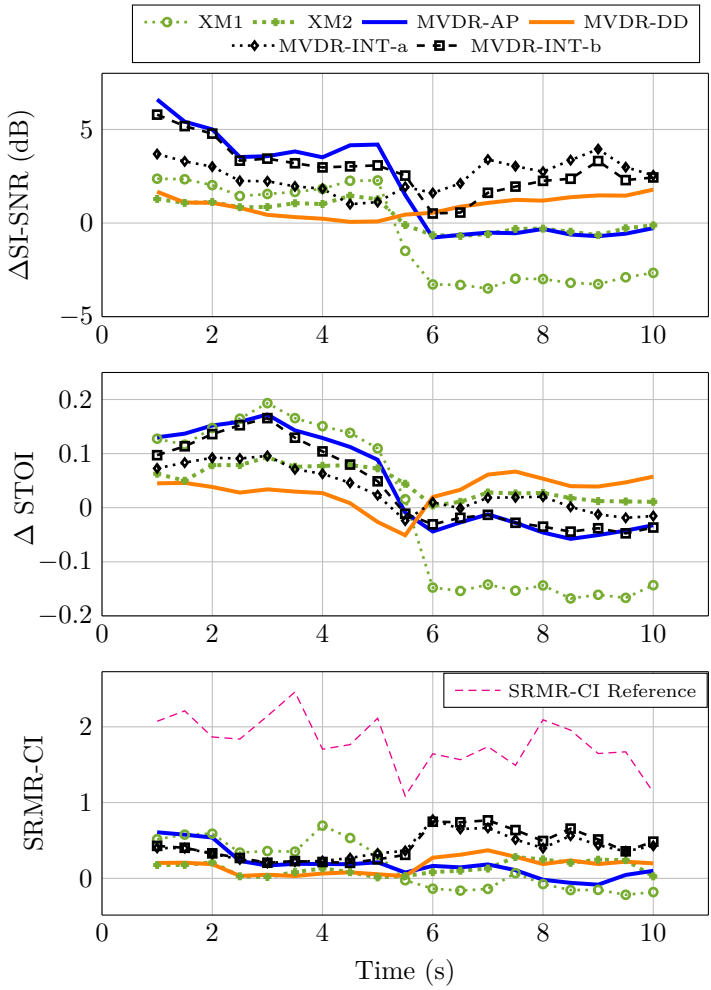


Figure 8.14: Performance of the MVDR-AP, MVDR-DD, and two tunings of the integrated MVDR beamformer, MVDR-INT-a and MVDR-INT-b along with XM1 and XM2 from Fig. 8.8 when SPP computed on $y_{a,1}$.

between MVDR-AP and MVDR-DD in terms of Δ SI-SNR and Δ STOI, with the MVDR-INT-b adhering more toward the performance of the MVDR-AP. In terms of the SRMR-CI, both the MVDR-INT-a and MVDR-INT-b are of similar performance to the MVDR-AP. In the last 6 s, in terms of Δ SI-SNR and SRMR-CI, both the MVDR-INT-a and the MVDR-INT-b demonstrate a performance that is slightly better than either the MVDR-AP and MVDR-DD.

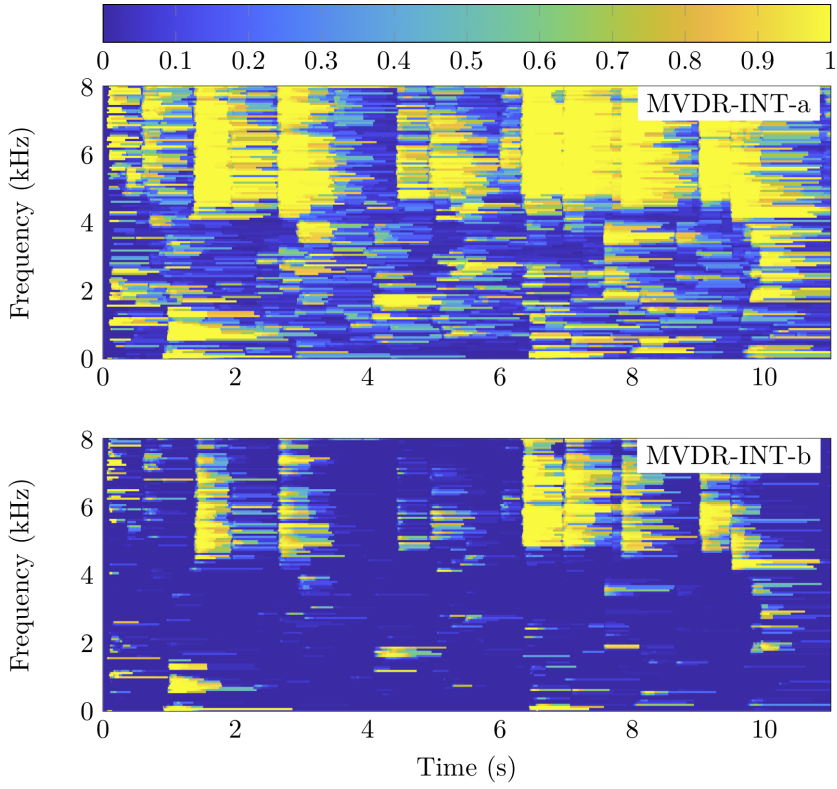


Figure 8.15: Confidence metric of the evaluation performed in Fig. 8.14 for (Top) MVDR-INT-a and (Bottom) MVDR-INT-b.

In terms of Δ STOI however, the MVDR-INT-a finds a compromise between the MVDR-AP and MVDR-DD while the MVDR-INT-b is of similar performance to the MVDR-AP, which is worse than the MVDR-DD. The corresponding confidence metrics across all time frames and frequencies for the MVDR-INT-a and the MVDR-INT-b from this experiment are displayed in Fig. 8.15, which suggests that there was generally more certainty in the accuracy of the data-dependent RTF vector at higher frequencies.

The corresponding spectrograms are also shown in Fig. 8.16. The differences between the MVDR-AP and MVDR-DD are more pronounced, with the MVDR-DD resulting in a particularly noisy signal in the first 5 s. Once again the integrated strategy is evident as it can be seen that the MVDR-INT-a and MVDR-INT-b result in signals which make use of both the MVDR-AP and MVDR-DD. Overall, these results suggest that appropriate tunings can

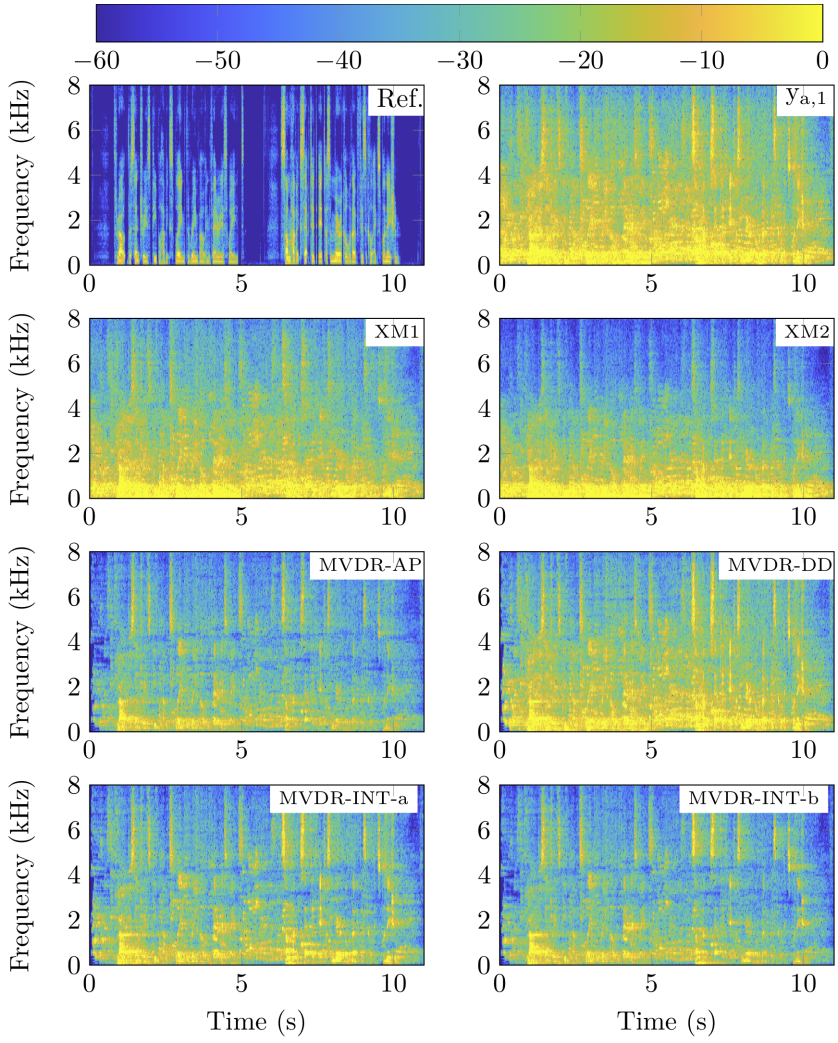


Figure 8.16: Spectrograms corresponding to the signals evaluated from Fig. 8.14. Also included is the the reference speech-only signal used in computing Δ STOI (Ref) and the noisy reference microphone signal, $y_{a,1}$. The colour bar indicates the intensity of the STFT coefficients (in dB), which has been constrained to the range as shown for visual clarity (since the intensity of the XM STFT coefficients were substantially greater than those on the reference signal).

give rise to different enhanced speech signals, which may be more suitable for consistently improving speech intelligibility despite changes in the source position and imperfectly estimated spatial correlation matrices.

The resulting audio signals from the experiments in this section may also be listened to for a subjective evaluation at [95].

8.7 Conclusion

This chapter has combined both the confidence-based integration approach from Chapter 3 and geometrically-based integration approach from Chapter 5 into a more global integration of the a priori and data-dependent spatial information for a microphone configuration consisting of an LMA linked with XMs. In particular, an integrated MVDR beamformer (similar to Chapter 3) has been developed which merges benefits from using an a priori RTF vector (as obtained from Chapter 5) and a data-dependent RTF vector. The framework has been presented in a pre-whitened-transformed (PWT) domain, which consists of an initial transformation of the microphone signals through a blocking matrix and a fixed beamformer, followed by a pre-whitening operation, facilitating convenient processing operations. In the PWT domain, procedures for obtaining an a priori RTF vector and data-dependent RTF vector have also been derived, where the a priori RTF vector is based on an a priori RTF vector pertaining to the LMA only.

With the two RTF vectors, an integrated MVDR beamformer was proposed by formulating a quadratically constrained quadratic program (QCQP), with two constraints, one of which is related to the maximum tolerable speech distortion for the imposition of the a priori RTF vector and the other related to the maximum tolerable speech distortion for the imposition of the data-dependent RTF vector. Such a formulation can be interpreted as a relaxation of the hard constraints that are usually imposed for conventional MVDR beamformers. The advantage of this approach is that by specifying the maximum tolerable speech distortions for each constraint in each time frame and frequency bin, it is possible to obtain a number of different enhanced signals, which can be more suitable for variable conditions within the acoustic environment.

A further insight into these maximum tolerable speech distortions was also provided where it was shown how the space spanned by each of the maximum tolerable speech distortions could be divided into four separate regions, each of which corresponded to a particular set of constraints being active or inactive. This insight then facilitated the development of a general tuning framework where the maximum tolerable speech distortions could be chosen in accordance

with the confidence in the accuracy of the data-dependent RTF vector. A particular set of tuning rules was also suggested, which made use of a relationship to the SDW-MWF. Furthermore, the advantage of using the integrated MVDR beamformer over the SDW-MWF was also observed in that it also provides framework for optimally combining an SDW-MWF that relies on an a priori RTF vector with an SDW-MWF that relies on a data-dependent RTF vector.

The theoretical behaviour of the QCQP as a function of the maximum tolerable speech distortion parameters was supported with audio recordings which involved a single speaker in a cocktail party noise scenario, and where the microphone configuration was an LMA of behind-the-ear hearing aid microphones and three XMs. A broadband evaluation was also performed when the speaker was in different locations. The results indicate that the integrated MVDR beamformer can indeed find a compromise between relying only on either an a priori or data-dependent RTF vector, resulting in enhanced signals which may be more suitable for consistently improving speech intelligibility despite changes in the source position and imperfectly estimated spatial correlation matrices.

As mentioned in Chapter 3, although the integrated MVDR beamformer has been developed using both an a priori RTF vector and a data-dependent RTF vector, this is not a strict requirement. In principle, any two RTF vector estimates can be used within the framework. Finally, it should be mentioned that the computational complexity of the algorithms developed has not been specifically addressed in this chapter. It is acknowledged that there several aspects which would have contributed to additional computational complexity such as the pre-whitening operations, the eigenvalue decompositions, and the optimal tuning of the step-size parameter for the gradient ascent procedure. A more complete analysis of the impact of these aspects on the computational complexity as well as the development of procedures for reducing this complexity is however left for a separate body of future work.

Chapter 9

Conclusion

The aim of this dissertation was to develop multi-microphone speech enhancement algorithms that could deliver robust performance in situations when parameters pertaining to spatial information in an acoustic environment, such as the location of a desired speaker, are inaccurately estimated. As opposed to a speech enhancement algorithm where such spatial parameters are based only on a priori knowledge or estimated only from observable microphone data, the speech enhancement algorithms in this dissertation have been designed by explicitly using *both* a priori and data-dependent spatial information in an integrated manner. Such a design has been applied to conventional speech enhancement systems that use a microphone array, as well as newer embodiments of speech enhancement systems which consist of a microphone array linked with external microphones. Depending on the microphone configuration, it was also shown that the integration of a priori and data-dependent spatial information can be accomplished by following a confidence-based approach, a geometrically-based approach, or a combination of these two approaches.

By following the integrated strategy in general, it has been argued that a priori spatial knowledge does not have to be fully discarded when transitioning to a more data-dependent speech enhancement system. Moreover, it was observed that a priori spatial knowledge can be advantageous in that it serves as a contingency framework when it is believed that the observable microphone data will result in an inaccurate estimation of the relevant spatial information. Theoretical and experimental results from this research have also demonstrated that through this strategy of integrating a priori and data-dependent spatial information, spatially robust speech enhancement algorithms can be designed. In particular, experiments involving hearing aid devices

and external microphones have indicated promising performance (signal-to-noise ratio and speech intelligibility improvements) for cases of inaccurately estimated correlation matrices and changes in speech source location for the speech enhancement tasks of noise reduction, binaural speech enhancement, and speech dereverberation.

The concept of integrating a priori and data-dependent spatial information as applied to a microphone array was initially investigated in **Chapter 3** from a confidence-based perspective. A variation to the cost function of a minimum variance distortionless response (MVDR) beamformer was analysed, where spatial constraints derived from a priori knowledge and a data-dependent method were treated as soft constraints within an unconstrained convex optimisation problem. It was shown that this integrated MVDR beamformer could also be expressed as a convex combination (with real-valued positive constants) of an MVDR beamformer that uses only a priori spatial information, one that uses only data-dependent spatial information, a linearly constrained minimum variance (LCMV) beamformer that uses both a priori and data-dependent spatial information, as well as an all-zero vector. This framework allows for microphone signals in a particular time-frequency bin to be processed differently, according to some level of confidence in the accuracy of the data-dependent spatial constraint. For instance, if it is believed that the data-dependent spatial constraint is accurate, then that is used in the processing. On the other hand, if there is low confidence in the accuracy, then the a priori spatial constraint can be used, which why it is perceived as a contingency option. The framework additionally facilitates in between decisions where different combinations of the a priori and data-dependent spatial constraints can be used. The resulting speech enhancement strategy in this chapter was in general found to be quite versatile and hence robust as it was able to produce a wide variety of enhanced speech signals.

A new microphone configuration was introduced in **Chapter 4**, which consisted of a microphone array linked with an external microphone. Through an insightful view of the beam patterns generated with this configuration and a theoretical performance analysis with classical beamformers, it was evident that there were substantial benefits to be obtained by using an external microphone. In particular, an increased directionality, signal to noise ratio, and speech intelligibility could all be obtained, especially for the case when the external microphone was close to the desired speaker.

Chapter 5 proceeded to discuss an approach of using a priori spatial knowledge for the microphone array linked with multiple external microphones. This was achieved by splitting the spatial constraint of the MVDR beamformer into two parts: a known part derived from a priori spatial knowledge pertaining to the microphone array and an unknown part pertaining to the external microphones,

which would have to be estimated from the microphone data. Such an approach can therefore also be interpreted as integration of a priori and data-dependent spatial information, but with a geometrically-based approach as opposed to the confidence-based approach taken in chapter 3. By following this approach, it was firstly found that optimal estimators for the unknown part of the spatial constraint could be estimated from a compressed set of the microphone data of dimension equal to the number of external microphones plus one. This could therefore lead to considerable computational complexity reductions especially for the case when there are a few external microphones but many microphones in the array. Secondly, using the best performing estimator, an extension to the conventional generalised sidelobe canceller (GSC) scheme was developed where there was independent access to two enhanced speech signals: (i) a speech estimate from a GSC using only the microphone array with its priori knowledge and (ii) a speech estimate that combines this first speech estimate with the additional data-dependent spatial information estimated using the external microphones. The access to two enhanced speech signals facilitates a robust approach as decisions can be taken in accordance with the level of confidence one has in the accuracy of the data-dependent spatial parameters, which now draws a parallel to integration using a confidence-based approach as discussed in chapter 3. For instance, in cases of inaccurately estimated spatial parameters from the microphone data, a contingency option of simply using the GSC with the microphone array using its a priori knowledge is available and is what would have been used in the absence of external microphones. These algorithms were also evaluated in a number of scenarios involving a microphone array on a hearing aid and randomly positioned external microphones. The results have indicated notable speech intelligibility improvements in using these proposed algorithms with the external microphones as opposed to using the microphone array alone with a priori knowledge or even in some cases a fully data-dependent approach. It can certainly be argued that one can simply listen to an external microphone if it is close to a desired speaker for an enhanced speech signal. However, (i) it should not be expected that an external microphone is always close to a desired speaker (such as in the case of a mobile phone), (ii) listening to an external microphone alone results in a loss of spatial cues, and (iii) results from several experiments have demonstrated that a more intelligible speech signal can be obtained when using all of the microphones. A final takeaway from this chapter is that when transitioning from a microphone array to a configuration involving external microphones, it is not necessary to discard any a priori spatial knowledge that may be available and completely transition to a fully data-dependent approach as is commonly the case. The algorithms developed in this chapter have presented an alternative approach in using microphone arrays along with external microphones, which integrates partial a priori and partial data-dependent spatial information, with the option of always

reverting to the original speech enhancement framework without the inclusion of the external microphones.

In **Chapter 6**, the geometrically-based integration approach of using a priori knowledge from Chapter 5 was then extended to a binaural speech enhancement task where the microphone configuration comprised of a left-ear hearing aid microphone array, a right-ear hearing aid microphone array, and external microphones. In the context of a binaural MVDR beamformer, it was demonstrated that by using separate a priori knowledge for each of the left and right ear microphone arrays, only a data-dependent estimate of the spatial constraint pertaining to the external microphones as well as an additional scaling between the two left and right ear microphone arrays was required. It was also shown that this data-dependent estimate could be done using a compressed set of microphone data of dimension equal to the number of external microphones plus two. This is in contrast to a fully data-dependent approach, which would have required microphone data of dimension equal to the total number of microphones. Furthermore, upon comparison to a fully data-dependent approach in adverse acoustic conditions, it was observed that using the a priori knowledge can yield a more intelligible speech estimate. In comparison to a binaural MVDR that uses only the left and right ear microphone arrays, it was also seen that the binaural cues with respect to the desired speaker are not degraded when external microphones are used in the processing. The experiments also highlight that listening to an external microphone in this context would be inadequate as the binaural cues with respect to the desired speaker will be severely distorted.

The geometrically-based integration approach of using a priori spatial knowledge was also extended to a speech dereverberation task using a microphone array and a single external microphone in **Chapter 7**. By making use of a data-independent operator for transforming and pre-whitening signals, it was shown that an eigenvalue decomposition of a 2×2 matrix was required for subsequently estimating the missing part of a spatial constraint pertaining to the external microphone as well as the speech and reverberant power spectral densities for a multi-channel Wiener filter. This is in contrast to the data-dependent approach where an eigenvalue decomposition of a matrix with dimension equal to the total number of microphones is required. Evaluations using both simulated and recorded data involving a hearing aid microphone array have demonstrated that both the a priori and fully data-dependent approaches can achieve similar a similar performance in terms of speech intelligibility improvement. Furthermore, it was also observed that the most speech intelligibility improvement over using the microphone array alone is obtained when the location of the external microphone is within the critical distance of the desired speaker. Listening directly to the XM was also found to offer an improvement in speech intelligibility in reverberant environments when close to the desired speaker, although the

loss of spatial cues would still persist.

Chapters 5-7 served to highlight the advantages of using a priori knowledge for speech enhancement tasks using a microphone array with external microphones. However that is not to say there is a complete negation of the advantages to be achieved by relying on fully data-dependent spatial information estimated using all the observable microphone data. In **Chapter 8** the confidence-based integration approach from chapter 3 and the geometrically-based integration approach from chapter 5 were combined to obtain a final conception of a more global integrated approach. The cost function for an MVDR beamformer was alternatively formulated as a quadratically constrained quadratic program with two constraints, one of which was related to a maximum tolerable speech distortion for the imposition of a priori spatial information (obtained from the approach in chapter 5) and the other related to a maximum tolerable speech distortion for the imposition of fully data-dependent spatial information. It was found that the parameter space spanned by these maximum tolerable speech distortions could be divided into four distinct regions for which a particular set of constraints are active or inactive, which provided deeper insight into the means by which a priori and data-dependent spatial information could be integrated. These results were then used to guide the proposition of a general tuning framework, from which some practical tuning strategies were suggested. Relations to the speech distortion weighted multi-channel Wiener filter (SDW-MWFs) were also discussed and it was further shown how two such SDW-MWFs (one based only on a priori spatial information and the other based only on data-dependent spatial information) could be optimally combined. The resulting integrated MVDR beamformer in this chapter is therefore considered to be the most overarching one throughout the dissertation as it encapsulates the widest range of speech enhancement filters. For instance, specific tunings can yield an enhanced signal that relies on a priori spatial knowledge from chapter 5, an enhanced signal that relies on a fully data-dependent method, or another linear combination of these two signals. Furthermore, by removal of the external microphones, the exact integrated MVDR beamformer from chapter 3 can be obtained where once again a trade-off can be made between the a priori and data-dependent spatial information. This facilitation of being able to transition from data-dependent to a priori approaches as well as from using external microphones to using the microphone array alone in fact serves as the contingency framework when it is believed that estimation using the observable microphone data will be inaccurate. The more globally integrated speech enhancement strategy developed in this chapter was also evaluated with hearing aid devices and external microphones in a recreated cocktail-party acoustic scenario, where the results have shown an improved spatial robustness in terms of signal-to-noise ratio and speech intelligibility.

Future research and perspectives

Throughout the discourse on multi-microphone speech enhancement, it has become evident that there is no one golden algorithm that can satiate all of the diverse needs from various end users. It is in response to this that the algorithms throughout this dissertation have been designed, i.e., with the provision of several possibilities of enhanced signals in one framework. This is however only a partial response as there are still several questions to be addressed in future research.

Tuning and confidence metrics. One of the main advantages of the integrating a priori and data-dependent spatial information is that a wide range of enhanced signals could be generated by specifying different tunings. In the dissertation, some specific tunings have been suggested and have performed well, but they have not been rigorously tested. As these tunings depended on some metric of confidence in the accuracy of the spatial information estimated from a data-dependent method, there is the need for two subsequent studies. The first is related to the question of what metrics of confidence can be developed that will have a high correlation to the accuracy of estimated spatial information. With such metrics available, the second question is then what would be an appropriate tuning strategy? The answers to these questions will inevitably involve the analysis of significant quantities of data pertaining to relevant scenarios which a speech enhancement system will encounter. To add to this complication, such a metric of confidence and tuning could differ for a hearing-impaired individual as opposed to a normal-hearing individual. This all suggests that the end user of the speech enhancement system should be directly involved in the tuning process. For instance, one research direction could be to investigate how machine learning techniques can be applied so that the end user is able to train or adapt the tuning for specific scenarios and hence decide on what is their preferred sound.

Robust beamforming. The integrated MVDR beamformers in this dissertation have used two constraints: one that is based on a priori knowledge and the other that is data-dependent. In other state-of-the-art robust beamforming techniques, however, one constraint is considered with a region of uncertainty [34, 183]. It is therefore hypothesised that an even more robust beamformer can be obtained by merging the integrated beamformer described in this work with that of the state-of-the-art robust beamformers. For instance, two regions of uncertainties can be prescribed, one pertaining to the a priori spatial constraint and the other for the data-dependent constraint. Such a formulation may result in a beamformer of high computational complexity, but remains to be investigated.

A statistical framework. The problems posed in this research consisted of deterministic, but unknown parameters which needed to be estimated. An alternative perspective which can be explored is that of a Bayesian perspective [130], where all the parameters are treated as unknown random variables. Extending the integration concept to a Bayesian framework could be an interesting research direction as the Bayesian framework allows for the inclusion of a priori distributions of certain parameters. Hence, in addition to the a priori spatial knowledge that was included in the algorithms developed in this work, a priori knowledge pertaining to a specific speaker or the acoustic environment for instance, can also be taken into consideration.

Communication with the external microphone and distributed processing. Many of the experiments throughout this dissertation have focused on assistive hearing devices. It was always assumed that there was a perfect communication link with no bandwidth constraints as well as synchronous sampling rates between the microphones on the assistive hearing device and the external microphones (as well as among the external microphones themselves for more than one external microphone). However this is an assumption that is inevitably broken in practice depending on the nature of the communication protocol used for a particular external microphone(s). Hence, although promising performance of several algorithms have been demonstrated throughout the dissertation, performance in practice needs to be evaluated in consideration of communication-related issues between the XMs and the microphone array. There have been suggestions in the literature to cope with bandwidth constraints [231–233] and sample rate offsets [239–241], however the communication delay remains to be one of the major limiting factors in using external microphones with assistive hearing devices. Currently, RF, Near-Field Magnetic Induction (NFMI), and Bluetooth systems are all used for streaming wireless audio to hearing aids for instance from mobile phones [213]. It still remains an open question if transmission is fast enough for speech enhancement algorithms such as the ones discussed in this dissertation. Additionally, a centralised processing scheme has been considered in all of the algorithms proposed, i.e, all of the microphone signals are broadcast to a fusion centre for processing. As discussed in Chapter 4, an alternative would be to pursue a distributed processing scheme, where the processing is shared among the different microphone nodes. Investigation into such schemes are in fact already underway [265].

A shift in evaluation methods. All of the experiments in this dissertation have been evaluated with objective metrics. These provide useful initial insights into how an algorithm will perform, but are inherently limited as several objective metrics depend on a reference signal which may not be available. Furthermore,

the results from such an evaluation are not as instructive for hearing-impaired individuals. The solution therefore would be to embark upon subjective testing in order to make some more concrete evaluations. However the problem with this method is that it can be quite time consuming and hence not as conducive for rapid algorithm development. Therefore it is apparent that some alternative is required where the best of both approaches can be achieved. In the case of hearing aid evaluations, both virtual sound environments [293–295] and the use of ecological momentary assessment [296–298] are promising methods.

Security and privacy Throughout the dissertation, it has been shown how noisy signals as captured by microphones can be "cleaned up" in order to obtain some desired speech signal. The more ethical concern of what is actually done with that speech signal (and even the noisy signals in general) also needs to be addressed. Although the intention has been primarily for assistive hearing devices and a general improvement in speech intelligibility for their end users, it would be naive to think that the captured and processed signals could not be used for other predatory interests. As we are transitioning to a society that is being increasingly monitored (laptops and phones all have microphones which can listen and cameras that can view), it is clear that we have entered into an age of surveillance capitalism [299] and there are legitimate threats on what can be done with our data, specifically speech data in this case [300]. Therefore, it is important that we also address issues of security and privacy within the role of technology developers. The recent Europen Union (EU) EU General Data Protection Regulation [301, 302] has been a major step towards increasing biometric data protection.

Bibliography

[1] W. T. Fitch, "The evolution of language: A comparative review," *Biology and Philosophy*, vol. 20, no. 2-3, pp. 193–203, 2005.

[2] P. Lieberman, *The biology and evolution of language*. Cambridge, MA: Harvard university press, 1983.

[3] N. Wiener, "Speech, Language, and Learning," *J. Acoust. Soc. Amer.*, vol. 22, no. 6, pp. 696–697, 1950.

[4] J. Blauert and J. Allen, *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, 1997.

[5] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multi-microphone speech enhancement and source separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 4, pp. 692–730, April 2017.

[6] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*. Wiley, Aug. 2018.

[7] N. Antonello. (2020). [Online]. Available: https://nantonel.github.io//animation/

[8] E. Cherry, "Some Experiments on the Recognition of Speech with One and with Two Ears," *J. Acoust. Soc. Amer.*, vol. 25, no. 5, pp. 975–979, 1953.

[9] J. H. McDermott, "The cocktail party problem," *Current Biology*, vol. 19, no. 22, 2009.

[10] P. Assmann and Q. Summerfield, "The perception of speech under adverse conditions," in *Speech Processing in the Auditory System*. New York, NY: Springer New York, 2004, pp. 231–308.

[11] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. USA: CRC Press, Inc., 2013, ch. 4, pp. 67–87.

[12] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*. Hoboken, NJ, USA: John Wiley and Sons, Inc., 2006.

[13] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*, 1st ed. Springer Publishing Company, Incorporated, 2010.

[14] S. Doclo, W. Kellermann, S. Makino, and S. Nordholm, "Multichannel Signal Enhancement Algorithms for Assisted Listening Devices," *IEEE Signal Process. Mag.*, pp. 18–30, February 2015.

[15] T. van Waterschoot, "Design and evaluation of digital signal processing algorithms for acoustic feedback and echo cancellation," Ph.D. dissertation, KU Leuven, Leuven, Belgium, 2009.

[16] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing, Techniques and Applications*. New York: Springer, 2001.

[17] A. Bertrand, S. Doclo, S. Gannot, N. Ono, and T. van Waterschoot, "Special issue on wireless acoustic sensor networks and ad hoc microphone arrays," *Signal Processing*, vol. 107, pp. 1–3, 2015.

[18] L. Cohen, *Time-frequency Analysis: Theory and Applications*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1995.

[19] 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, April 1988.

[20] H. L. V. Trees, *Optimum Array Processing*. John Wiley and Sons, Ltd, 2002, ch. 2, pp. 17–89.

[21] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art*. Morgan & Claypool Publishers, 2013.

[22] R. C. Hendriks and T. Gerkmann, "Noise Correlation Matrix Estimation for Multi-Microphone Speech Enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 1, pp. 223–233, 2012.

[23] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, 2003.

[24] J. K. Nielsen, M. S. Kavalekalam, M. G. Christensen, and J. Boldt, "Model-based noise psd estimation from speech in non-stationary noise," in *Proc. 2018 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '18)*, 2018, pp. 5424–5428.

[25] 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. 2015 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '15)*, Brisbane, Australia, April 2015, pp. 544–548.

[26] A. Kuklasinski, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE Trans. Audio Speech Lang. Process.*, vol. 24, no. 9, pp. 1595–1608, 2016.

[27] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 107–115, 2014.

[28] S. Srinivasan, "Knowledge-based speech enhancement," Ph.D. dissertation, KTH - Royal Institute of Technology, 2005.

[29] M. S. Kavalekalam, J. K. Nielsen, J. B. Boldt, and M. G. Christensen, "Model-based Speech Enhancement for Intelligibility Improvement in Binaural Hearing Aids," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 1, pp. 99–113, 2019.

[30] M. Kompis and N. Dillier, "Performance of an adaptive beamforming noise reduction scheme for hearing aid applications. I. Prediction of the signal-to-noise-ratio improvement." *J. Acoust. Soc. Amer.*, vol. 109, no. 3, pp. 1123–1133, 2001.

[31] J. Greenberg and P. Zurek, "Evaluation of an adaptive beamforming method for hearing aids," *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1662–1676, 1992.

[32] J. M. Kates and M. R. Weiss, "A comparison of hearing-aid array-processing techniques," *J. Acoust. Soc. Amer.*, vol. 99, no. 5, pp. 3138–3148, 1996.

[33] A. Spriet, L. Van Deun, K. Eftaxiadis, J. Laneau, M. Moonen, B. van Dijk, A. van Wieringen, and J. Wouters, "Speech understanding in background noise with the two-microphone adaptive beamformer BEAM in the Nucleus Freedom Cochlear Implant System." *Ear and Hearing*, vol. 28, no. 1, pp. 62–72, 2007.

[34] W. C. Liao, Z. Q. Luo, I. Merks, and T. Zhang, "An effective low complexity binaural beamforming algorithm for hearing AIDS," *Proc. 2015 IEEE Workshop Appls. Signal Process. Audio Acoust. (WASPAA '15)*, no. 2, pp. 1–5, 2015.

[35] A. H. Moore, J. M. de Haan, M. S. Pedersen, P. A. Naylor, M. Brookes, and J. Jensen, "Personalized signal-independent beamforming for binaural hearing aids," *J. Acoust. Soc. Amer.*, vol. 145, no. 5, pp. 2971–2981, 2019.

[36] W.-C. Liao, Z.-Q. Luo, I. Merks, M. Hong, and T. Zhang, "Hearing assistance device with beamformer optimized using a priori spatial information," U.S. Patent US9 949 041B2, Aug. 17, 2018.

[37] O. Valentin, S. Vij, and J. Voix, "Development of a real-time eog-based acoustical beamformer algorithm for binaural hearing devices," *Canadian Acoustics*, vol. 47, no. 2, pp. 41–48, Jul. 2019.

[38] T. R. Jennings and G. Kidd, "A visually guided beamformer to aid listening in complex acoustic environments," *Proceedings of Meetings on Acoustics*, vol. 33, no. 050005, 2019.

[39] 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 Trans. Audio Speech Lang. Process.*, vol. 22, no. 4, pp. 785–799, 2014.

[40] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Speech distortion weighted multichannel wiener filtering techniques for noise reduction," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 199–228.

[41] 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 *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2499–2503.

[42] P. Pertilä, A. Brutti, P. Svaizer, and M. Omologo, "Multichannel source activity detection, localization, and tracking," in *Audio Source Separation and Speech Enhancement*. John Wiley and Sons, Ltd, 2018, ch. 4, pp. 47–64.

[43] D. B. Ward, R. A. Kennedy, and R. C. Williamson, "Constant directivity beamforming," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 3–17.

[44] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, ser. Springer Topics in Signal Processing. Springer Berlin Heidelberg, 2008.

[45] A. Bertrand, "Signal processing algorithms for wireless acoustic sensor networks," Ph.D. dissertation, KU Leuven, Leuven, Belgium, 2011.

[46] J. Szurley, "Distributed signal processing algorithms for acoustic sensor networks," Ph.D. dissertation, KU Leuven, Leuven, Belgium, 2015.

[47] A. Hassani, "Distributed signal processing algorithms for multi-task wireless acoustic sensor networks," Ph.D. dissertation, KU Leuven, Leuven, Belgium, 2017.

[48] S. Haykin and K. R. Liu, *Handbook on Array Processing and Sensor Networks*. Wiley-IEEE Press, 2010.

[49] C. T. Jespersen and B. Kirkwood, "Benefit of the ReSound Multi Mic," *ReSound white paper*, 2016.

[50] J. Mecklenburger and T. Groth, "Wireless technologies and hearing aid connectivity," in *Hearing Aids*, G. R. Popelka, B. C. J. Moore, R. R. Fay, and A. N. Popper, Eds. Cham: Springer International Publishing, 2016, pp. 131–149.

[51] E. M. Fitzpatrick, C. Séguin, D. R. Schramm, S. Armstrong, and J. Chénier, "The benefits of remote microphone technology for adults with cochlear implants." *Ear and hearing*, vol. 30, pp. 590–599, 2009.

[52] WHO. (2020, March). [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[53] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. J. Ray Liu, Eds. John Wiley & Sons, Inc., 2010, ch. 10, pp. 269–302.

[54] S. Launer, J. A. Zakis, and B. C. J. Moore, "Hearing aid signal processing," in *Hearing Aids*, G. R. Popelka, B. C. J. Moore, R. R. Fay, and A. N. Popper, Eds. Cham: Springer International Publishing, 2016, pp. 93–130.

[55] T. Zhang, F. Mustiere, and C. Micheyl, "Intelligent hearing aids: The next revolution," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2016-Octob, pp. 72–76, 2016.

[56] P. C. Loizou, "Mimicking the human ear," *IEEE Signal Process. Mag.*, vol. 15, no. 5, pp. 101–130, 1998.

[57] Y. Hu and P. C. Loizou, "A new sound coding strategy for suppressing noise in cochlear implants." *The Journal of the Acoustical Society of America*, vol. 124, no. August 2007, pp. 498–509, 2008.

[58] A. Hersbach, D. B. Grayden, J. B. Fallon, and H. J. McDermott, "A beamformer post-filter for cochlear implant noise reduction." *J. Acoust. Soc. Amer.*, vol. 133, no. 4, pp. 2412–20, 2013.

[59] R. L. Goldsworthy, L. A. Delhorne, J. G. Desloge, and L. D. Braida, "Two-microphone spatial filtering provides speech reception benefits for cochlear implant users in difficult acoustic environments," *J. Acoust. Soc. Amer.*, vol. 136, no. 2, pp. 867–876, 2014.

[60] D. Berkley and O. Mitchell, "Seeking the ideal in hands-free telephony," *Bell Laboratories Record*, vol. 52, pp. 318–325, 1974.

[61] E. Hänsler, "The hands-free telephone problem: an annotated bibliography update," *Annales Des Télécommunications*, vol. 49, no. 7, pp. 360–367, 1994.

[62] "Convention on Road Traffic, Vienna, 8 November 1968," *United Nations Conference on Road Traffic Final Act, 36 Signatories 82 Parties*.

[63] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding And Error Concealment*. Hoboken, NJ, USA: John Wiley and Sons, Inc., 2006.

[64] I. Cohen, J. Benesty, and S. Gannot, *Speech Processing in Modern Communication: Challenges and Perspectives*. Springer Publishing Company, Incorporated, 2010.

[65] S. Watanabe, M. Delcroix, F. Metze, and J. R. Hershey, "Preliminaries," in *New Era for Robust Speech Recognition: Exploiting Deep Learning*, S. Watanabe, M. Delcroix, F. Metze, and J. R. Hershey, Eds. Cham: Springer International Publishing, 2017, pp. 3–17.

[66] R. M. Stern, A. Acero, F.-H. Liu, and Y. Ohshima, "Signal processing for robust speech recognition," in *Automatic Speech and Speaker Recognition: Advanced Topics*, C.-H. Lee, F. K. Soong, and K. K. Paliwal, Eds. Boston, MA: Springer US, 1996, pp. 357–384.

[67] J. P. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The CHiME challenges: Robust speech recognition in everyday environments," in

New Era for Robust Speech Recognition: Exploiting Deep Learning, S. Watanabe, M. Delcroix, F. Metze, and J. R. Hershey, Eds. Cham: Springer International Publishing, 2017, pp. 327–344.

[68] M. Seltzer, "Microphone array processing for robust speech recognition," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA 15213, 2003.

[69] K. Hermus, P. Wambacq, and H. Van hamme, "A review of signal subspace speech enhancement and its application to noise robust speech recognition," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, p. 045821, 2006.

[70] A. Spriet, S. Doclo, M. Moonen, and J. Wouters, "A Unification of Adaptive Multi-Microphone Noise Reduction Systems," *Proc. 2006 Int. Workshop Acoustic Echo Noise Control (IWAENC '06)*, Sept. 2006.

[71] R. Eklund, "Pulmonic ingressive phonation: Diachronic and synchronic characteristics, distribution and function in animal and human sound production and in human speech," *Journal of the International Phonetic Association*, vol. 38, no. 3, p. 235–324, 2008.

[72] T. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, 1st ed. USA: Prentice Hall Press, 2001.

[73] T. Chiba and M. Kajiyama, *The Vowel-its Nature and Structure.* Tokyo: Kaiseikan, 1941.

[74] G. Fant, *Acoustic Theory of Speech Production.* The Hague: Mouton, 1960.

[75] J. L. R. d'Alembert, "Investigation of the curve formed by a vibrating string (1747)," in *Acoustics: Historical and Philosophical Development*, R. B. Lindsay, Ed. Berlin, Heidelberg: Stroudsburg: Dowden, Hutchinson & Ross, 1973, p. 119–123.

[76] L. Euler, "De vibratione chordarum exercitatio," *Nova acta eruditorum*, 1749. [Online]. Available: http://eulerarchive.maa.org/pages/E119.html

[77] J. Lagrange, "Research on the nature and propagation of sound," *reprinted in Oeuvres de Lagrange*, vol. 1, pp. 39–148, 1759.

[78] P. Morse, *Vibration and Sound.* American Inst. of Physics, 1981.

[79] S. Garrett, *Understanding Acoustics: An Experimentalist's View of Acoustics and Vibration*, ser. Graduate Texts in Physics. Springer International Publishing, 2017.

[80] A. D. Pierce, *Acoustics: An introduction to its physical principles and applications*. McGraw-Hill Book Co New York, 1981.

[81] B. Kinsman, *Wind Waves: Their Generation and Propagation on the Ocean Surface*, ser. Dover Phoenix Editions. Dover, 2002.

[82] F. Jacobsen and P. Juhl, *Fundamentals of General Linear Acoustics*. Wiley, 2013.

[83] P. Morse and H. Feshbach, *Methods of Theoretical Physics*, ser. International series in pure and applied physics. McGraw-Hill, 1953, no. pt. 1.

[84] H. Kuttruff, *Room Acoustics, Fifth Edition*. Taylor & Francis, 2009.

[85] M. R. Schroeder, "Statistical Parameters of the Frequency Response Curves of Large Rooms," *J. Audio Eng. Soc.*, vol. 35, no. 5, pp. 299–306, 1987.

[86] W. Sabine, *Collected Papers on Acoustics*. Harvard University Press, 1922.

[87] D. Swanson, *Signal Processing for Intelligent Sensor Systems with MATLAB*. CRC Press, 2011.

[88] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2009.

[89] M. Holters, T. Corbach, and U. Zölzer, "Impulse response measurement techniques and their applicability in the real world," *Proceedings of the 12th International Conference on Digital Audio Effects, DAFx 2009*, 2009.

[90] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms." *J. Acoust. Soc. Amer.*, vol. 113, no. 6, pp. 3233–3244, 2003.

[91] I. Arweiler and J. M. Buchholz, "The influence of spectral characteristics of early reflections on speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 130, no. 2, pp. 996–1005, 2011.

[92] A. A. Kressner, A. Westermann, and J. M. Buchholz, "The impact of reverberation on speech intelligibility in cochlear implant recipients," *J. Acoust. Soc. Amer.*, vol. 144, no. 2, pp. 1113–1122, 2018.

[93] J. Polack, "Modifying Chambers to play Billiards: the Foundations of Reverberation Theory," *Acta Acustica united with Acustica*, vol. 76, no. 6, pp. 256–272, 1992.

[94] C. Veaux, J. Yamagishi, and K. MacDonald. (2016) CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit. [Online]. Available: http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html

[95] R.Ali. (2020). [Online]. Available: ftp://ftp.esat.kuleuven.be/pub/SISTA/rali/Thesis_data

[96] J. Mourjopoulos and M. Paraskevas, "Pole and zero modeling of room transfer functions," *J. Sound Vib.*, vol. 146, no. 2, pp. 281–302, 1991.

[97] Y. Haneda, A. Member, S. Makino, and Y. Kaneda, "Common Acoustical Pole and Zero Modeling of Room Transfer Functions," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 320–328, 1994.

[98] G. Vairetti, "Efficient parametric modeling, identification and equalization of room acoustics," Ph.D. dissertation, KU Leuven, 2018.

[99] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Amer.*, vol. 8, no. 3, pp. 185–190, 1937.

[100] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1, pp. 103 – 138, 1990.

[101] R. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/Synthesis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 1, pp. 99–102, 1980.

[102] J. J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Process. Mag.*, vol. 9, no. 1, pp. 14–37, Jan 1992.

[103] Anti-Everything, "Children of a Globalised World," *Children of a Globalised World (Musical Album)*, ISRC:TT-A10-11-00013, Boatshrimp Records, Port-of-Spain, 2011.

[104] J. O. Smith. Spectral audio processing. Online book, 2011 edition. [Online]. Available: http://ccrma.stanford.edu/~jos/sasp/

[105] A. Rosenblueth and N. Wiener, "The Role of Models in Science," *Philosophy of Science*, vol. 12, no. 4, pp. 316–321, 1945.

[106] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: analysis, experiments, and application to acoustic echo cancellation," *IEEE Trans. Signal Process.*, vol. 40, no. 8, pp. 1862–1875, Aug 1992.

[107] Y. Avargel and I. Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, 2007.

[108] R. Talmon, I. Cohen, and S. Gannot, "Relative Transfer Function Identification Using Convolutive Transfer Function Approximation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 4, pp. 546–555, 2009.

[109] S. Gannot, D. Burshtein, and E. Weinstein, "Signal Enhancement Using Beamforming and Nonstationarity with Applications to Speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, 2001.

[110] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 19–38.

[111] H. Marsh, "Correlation in wave fields," *U.S. Navy Underwater Sound Laboratory Quarterly Report for period ending 31 March, 1950*, pp. 63–68, 1950.

[112] J. Faran and R. Hills, "The application of correlation techniques to acoustic receiving systems," *Tech. Memo No. 28, Harvard University, Cambridge*, 1952.

[113] B. F. Cron and C. H. Sherman, "Spatial-Correlation Functions for Various Noise Models," *J. Acoust. Soc. Amer.*, vol. 34, no. 11, pp. 1732–1736, 1962.

[114] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 10, pp. 1365–1376, 1987.

[115] G. Strang, *Introduction to Linear Algebra*, 4th ed. Wellesley, MA: Wellesley-Cambridge Press, 2009.

[116] S. Darlington, "Linear least-squares smoothing and prediction, with applications," *Bell System Technical Journal*, vol. 37, no. 5, pp. 1221–1294, 1958.

[117] P. E. Green, E. J. Kelly, and M. J. Levin, "A Comparison of Seismic Array Processing Methods," *Geophysical Journal of the Royal Astronomical Society*, vol. 11, no. 1, pp. 67–84, 1966.

[118] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[119] H. Cox, "Resolving power and sensitivity to mismatch of optimum array processors," *J. Acoust. Soc. Amer.*, vol. 54, no. 3, pp. 771–785, 1973.

[120] G. W. Elko, "Superdirectional microphone arrays," in *Acoustic Signal Processing for Telecommunication*, S. L. Gay and J. Benesty, Eds. Boston, MA: Springer US, 2000, pp. 181–237.

[121] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug 1972.

[122] M. H. Er and A. Cantoni, "Derivative Constraints for Broad-band Element Space Antenna Array Processors," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-31, no. 6, pp. 1378–1393, 1983.

[123] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[124] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, 1982.

[125] S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 2013.

[126] B. R. Breed and J. Strauss, "A Short Proof of the Equivalence of LCMV and GSC Beamforming," *IEEE Signal Process. Lett.*, vol. 9, no. 6, pp. 168–169, 2002.

[127] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," in *Proc. 1993 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '93)*, vol. 2, April 1993, pp. 355–358 vol.2.

[128] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, 2004.

[129] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer Berlin Heidelberg, 2001, pp. 39–60.

[130] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, 1997.

[131] M. W. Callahan, "Acoustic signal processing based on the short-time spectrum," Ph.D. dissertation, University of Utah, 1976.

[132] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. MIT Press, 1949.

[133] Yi Hu and P. C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 59–67, 2004.

[134] S. Gustafsson, P. Jax, and P. Vary, "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics," *Proc. 1998 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '98)*, vol. 1, pp. 397–400, 1998.

[135] Y. Hu and P. C. Loizou, "A perceptually motivated subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 457–465, 2003.

[136] S. Gelfand, *Hearing: An Introduction to Psychological and Physiological Acoustics, Fourth Edition.* Taylor & Francis, 2004.

[137] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, 1988.

[138] B. Defraene, K. Ngo, T. van Waterschoot, M. Diehl, and M. Moonen, "A psychoacoustically motivated speech distortion weighted multi-channel Wiener filter for noise reduction," *Proc. 2012 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '12)*, pp. 4637–4640, 2012.

[139] J. Mcaulay, Robert and M. L. Malpass, "Speech Enhancement Using a Soft-Decision Noise," *IEEE Trans. Acoust., Speech, Signal Process.*, no. 2, pp. 137–145, 1980.

[140] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, 1984.

[141] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, April 1985.

[142] P. J. Wolfe and S. J. Godsill, "Towards a perceptually optimal spectral amplitude estimator for audio signal enhancement," *Proc. 2000 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '00)*, vol. 2, pp. 821–824, 2000.

[143] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 857–869, 2005.

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

[145] P. J. Wolfe and S. J. Godsill, "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement," *IEEE Workshop on Statistical Sig. Process. Proc.*, pp. 496–499, 2001.

[146] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 7, pp. 1110–1126, 2005.

[147] J. E. Porter and S. F. Boll, "Optimal Estimators for Spectral Restoration of Noisy Speech." *Proc. 1982 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '84)*, vol. 2, no. 1, pp. 53–56, 1984.

[148] Y. Ephraim and I. Cohen, "Recent advancements in speech enhancement," in *The Electrical Engineering Handbook, Circuits, Signals, and Speech and Image Processing, Third Edition*, R. C. Dorf, Ed. Boca Raton, FL: CRC Press, 2006, ch. 15, pp. 15–12 – 15–26.

[149] J. S. Erkelens, R. C. Hendriks, and R. Heusdens, "On the estimation of complex speech DFT coefficients without assuming independent real and imaginary parts," *IEEE Signal Process. Lett.*, vol. 15, pp. 213–216, 2008.

[150] R. C. Hendriks, J. S. Erkelens, and R. Heusdens, "Comparison of complex-DFT estimators with and without the independence assumption of real and imaginary parts," *Proc. 2008 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '08)*, no. 1, pp. 4033–4036, 2008.

[151] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Comm.*, vol. 53, no. 4, pp. 465–494, 2011.

[152] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, 2015.

[153] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Amer.*, vol. 125, no. 5, p. 3387, 2009.

[154] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices," *IEEE Signal Process. Mag.*, pp. 114–124, 2015.

[155] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time – Frequency Weighted Noisy Speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[156] S. Van Kuyk, W. Bastiaan Kleijn, and R. C. Hendriks, "An Evaluation of Intrusive Instrumental Intelligibility Metrics," *IEEE Trans. Audio Speech Lang. Process.*, vol. 26, no. 11, pp. 2153–2166, 2018.

[157] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. USA: CRC Press, Inc., 2013.

[158] J. Kreiman, D. Van Lancker Sidtis, and B. Gerratt, "Defining and measuring voice quality," in *Voice Quality: Functions, Analysis and Synthesis (VOQUAL'03), ISCA Tutorial and Research Workshop, Geneva, Switzerland*, 2003.

[159] J. E. Greenberg, P. M. Peterson, and P. M. Zurek, "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance." *J. Acoust. Soc. Amer.*, vol. 94, no. 5, pp. 3009–3010, 1993.

[160] ANSI-S3.5-1997, "American National Standard Methods for calculation of the speech intelligibility index," *J. Acoust. Soc. Amer.*, June 1997.

[161] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective Quality Evaluation for Low-Bit-Rate Speech Coding Systems," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 242–248, 1988.

[162] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality measure of dereverberated speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 7, pp. 4–7, 2008.

[163] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *Proc. 2014 Int. Workshop Acoustic Signal Enhancement (IWAENC '14)*, 2014, pp. 55–59.

[164] F. Chen, O. Hazrati, and P. Loizou, "Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure," *Biomedical signal processing and control*, vol. 8, pp. 311–314, 05 2013.

[165] A. H. Andersen, J. M. De Haan, Z. H. Tan, and J. Jensen, "A non-intrusive Short-Time Objective Intelligibility measure," *Proc. 2017 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '17)*, pp. 5085–5089, 2017.

[166] I. Cohen, "Relative Transfer Function Identification Using Speech Signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, 2004.

[167] S. Winter, H. Sawada, S. Araki, and S. Makino, "Overcomplete BSS for convolutive mixtures based on hierarchical clustering," in *Independent Component Analysis and Blind Signal Separation*, C. G. Puntonet and A. Prieto, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 652–660.

[168] M. Taseska and E. A. P. Habets, "Relative Transfer Function Estimation Exploiting Instantaneous Signals and the Signal Subspace," *Proc. 23rd European Signal Process. Conf. (EUSIPCO '15)*, pp. 404–408, 2015.

[169] Z. Wang and D. Wang, "Mask weighted STFT ratios for relative transfer function estimation and its application to robust ASR," in *Proc. 2018 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '18)*, 2018, pp. 5619–5623.

[170] C. Pan, J. Chen, and J. Benesty, "Performance Study of the MVDR Beamformer as a Function of the Source Incidence Angle," *IEEE Trans. Audio Speech Lang. Process.*, vol. 22, no. 1, pp. 67–79, 2014.

[171] K. Reindl, S. Markovich-Golan, H. Barfuss, S. Gannot, and W. Kellermann, "Geometrically Constrained TRINICON-based relative transfer function estimation in underdetermined scenarios," *Proc. 2013 IEEE Workshop Appls. Signal Process. Audio Acoust. (WASPAA '13)*, no. 2, pp. 1–4, 2013.

[172] A. Spriet, S. Doclo, M. Moonen, and J. Wouters, "A unification of adaptive multi-microphone noise reduction systems," in *Proc. 2006 Int. Workshop Acoustic Echo Noise Control (IWAENC '06)*, Paris, France, Sept. 2006.

[173] J. Li and P. Stoica, *Robust Adaptive Beamforming*, ser. Wiley Series in Telecommunications and Signal Processing. Wiley, 2005.

[174] S. A. Vorobyov, "Principles of minimum variance robust adaptive beamforming design," *Signal Processing*, vol. 93, no. 12, pp. 3264–3277, 2013.

[175] D. Feldman and L. Griffiths, "A Projection Approach for Robust Adaptive Beamforming David," *IEEE Trans. Signal Process.*, vol. 42, no. 4, pp. 867–876, 1994.

[176] Y. I. Abramovich, "Controlled Method for Adaptive Optimization of Filters using the criterion of maximum signal-to-noise ratio," *Radio Eng. Electron Phys.*, vol. V 25, no. N 3, pp. 87–95, 1981.

[177] D. Carlson, Blair, "Covariance matrix estimation errors and diagonal loading in adaptive arrays," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 24, no. 4, pp. 397–401, 1988.

[178] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-posed problems*. W.H. Winston, 1977.

[179] A. B. Gershman, U. Nickel, and J. F. Böhme, "Adaptive beamforming algorithms with robustness against jammer motion," *IEEE Trans. Signal Process.*, vol. 45, no. 7, pp. 1878–1885, 1997.

[180] R. J. Mailloux, "Covariance matrix augmentation to produce adaptive array pattern troughs," *Electronics Lett.*, vol. 31, no. 10, pp. 771–772, 1995.

[181] J. R. Guerci, "Theory and application of covariance matrix tapers for robust adaptive beamforming," *IEEE Trans. Signal Process.*, vol. 47, no. 4, pp. 977–985, 1999.

[182] M. Zatman, "Comments on "Theory and Application of Covariance Matrix Tapers for Robust Adaptive Beamforming"," *IEEE Trans. Signal Process.*, vol. 48, no. 6, pp. 1796–1800, 2000.

[183] S. A. Vorobyov, A. B. Gershman, and Z. Q. Luo, "Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem," *IEEE Trans. Signal Process.*, vol. 51, no. 2, pp. 313–324, 2003.

[184] A. Nemirovskii and Y. Nesterov, *Interior-point polynomial algorithms in convex programming*. Society for Industrial and Applied Mathematics, 1994.

[185] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York, NY, USA: Springer, 2006.

[186] S. A. Vorobyov, A. B. Gershman, Zhi-Quan Luo, and Ning Ma, "Adaptive beamforming with joint robustness against mismatched signal steering vector and interference nonstationarity," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 108–111, Feb 2004.

[187] R. Lorenz and S. Boyd, "Robust minimum variance beamforming," *IEEE Trans. Signal Process.*, vol. 53, no. 5, pp. 1684–1696, 2005.

[188] S. Shahbazpanahi, A. B. Gershman, Z. Q. Luo, and K. M. Wong, "Robust adaptive beamforming for general-rank signal models using worst-case performance optimization," *IEEE Trans. Signal Process.*, vol. 51, no. 9, pp. 2257–2269, 2003.

[189] W. C. Liao, M. Hong, I. Merks, T. Zhang, and Z. Q. Luo, "Incorporating spatial information in binaural beamforming for noise suppression in

hearing aids," in *Proc. 2015 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '15)*, 2015, pp. 5733–5737.

[190] C. A. Anderson, P. D. Teal, and M. A. Poletti, "Spatially Robust Far-field Beamforming Using the von Mises(-Fisher) Distribution," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 12, pp. 2189–2197, 2015.

[191] S. Van Gerven and F. Xie, "A comparative study of speech detection methods," *Fifth European Conference on Speech*, pp. 1095–1098, 1997.

[192] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *Proc. 2011 IEEE Workshop Appls. Signal Process. Audio Acoust. (WASPAA '11)*, Oct 2011, pp. 145–148.

[193] M. M. T. Dietzen, and and T. van Waterschoot, "Instantaneous psd estimation for speech enhancement based on generalized principal components," in *Proc. 28th European Signal Process. Conf. (EUSIPCO '20)*, Jan. 2021, pp. 1–5.

[194] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Communication*, vol. 49, no. 7-8, pp. 636–656, 2007.

[195] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, 2001.

[196] G. H. Golub and C. F. Van Loan, *Matrix Computations (3rd Ed.)*. Baltimore, MD, USA: Johns Hopkins University Press, 1996.

[197] B. N. Parlett, *The Symmetric Eigenvalue Problem*. Society for Industrial and Applied Mathematics, 1998.

[198] J. Wilkinson, *The algebraic eigenvalue problem*, ser. Monographs on numerical analysis. Oxford, 1965.

[199] A. Krueger, E. Warsitz, and R. Haeb-Umbach, "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation," vol. 19, no. 1, pp. 206–219, 2011.

[200] P. Thüne and G. Enzner, "Maximum-Likelihood and Maximum-A-Posteriori Perspectives for Blind Channel Identification on Acoustic Sensor Network Data," *Proc. of ITG Symp. Speech Comm.*, no. 2, pp. 241–245, 2018.

[201] S. Chakrabarty and E. A. P. Habets, "On the Numerical Instability of an LCMV Beamformer for a Uniform Linear Array," *IEEE Signal Process. Lett.*, vol. 23, no. 2, pp. 272–276, 2016.

[202] R. Ali and M. Moonen, "A contingency multi-microphone noise reduction strategy based on linearly constrained multi-channel wiener filtering," in *Proc. 2016 Int. Workshop Acoustic Signal Enhancement (IWAENC '16)*, Xi'an, China, Sept 2016, pp. 1–4.

[203] A. Spriet, M. Moonen, and J. Wouters, "The impact of speech detection errors on the noise reduction performance of multi-channel Wiener filtering and Generalized Sidelobe Cancellation," *Signal Processing*, vol. 85, no. 6, pp. 1073–1088, 2005.

[204] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, April 1985.

[205] M. Nilsson, S. D. Soli, and J. Sullivan, "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise." *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1085–1099, 1994.

[206] Auditec, "Auditory Tests (Revised), Compact Disc, Auditec, St. Louis," St. Louis, 1997.

[207] E. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint." *J. Acoust. Soc. Amer.*, vol. 124, pp. 2911–2917, 2008.

[208] E. De Sena, N. Antonello, M. Moonen, and T. van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 4, pp. 774–786, April 2015.

[209] N. Antonello. (2016) Room impulse response generator with the randomized image method. [Online]. Available: https://github.com/nantonel/RIM.jl/tree/master/src/MATLAB

[210] M. Brookes *et al.* (1997) Voicebox: Speech processing toolbox for matlab. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox

[211] R.Ali. (2019). [Online]. Available: ftp://ftp.esat.kuleuven.be/pub/SISTA/rali/Reports/INT_MVDR_LMA/Audio_Data

[212] A. A. Hersbach, S. J. Mauger, J. M. Heasman, and P. W. Dawson, "Sound processing based on a confidence measure," U.S. Patent US9 589 580B2, March 7, 2017.

[213] J. Mecklenburger and T. Groth, "Wireless technologies and hearing aid connectivity," in *Hearing Aids*, G. R. Popelka, B. C. J. Moore, R. R. Fay, and A. N. Popper, Eds. Cham: Springer International Publishing, 2016, pp. 131–149.

[214] M. Ross, "Keynote Address FM Systems : A Little History and Some Personal Reflections," *ACCESS: Achieving Clear Communication Employing Sound Solutions*, pp. 17–28, 1992.

[215] A. Boothroyd, "Hearing aid accessories for adults: the remote FM microphone." *Ear and Hearing*, vol. 25, no. 1, pp. 22–33, 2004.

[216] E. C. Schafer, K. Sanders, D. Bryant, K. Keeney, and N. Baldus, "Effects of Voice Priority in FM Systems for Children with Hearing Aids," *J. Educ. Audiol.*, vol. 19, pp. 12–24, 2013.

[217] S. Markovich-Golan, S. Gannot, and I. Cohen, "Distributed multiple constraints generalized sidelobe canceler for fully connected wireless acoustic sensor networks," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 2, pp. 343–356, 2013.

[218] M. Ferrer, M. De Diego, G. Piñero, and A. Gonzalez, "Active noise control over adaptive distributed networks," *Signal Process.*, vol. 107, pp. 82–95, 2015.

[219] A. Hassani, A. Bertrand, and M. Moonen, "Cooperative integrated noise reduction and node-specific direction-of-arrival estimation in a fully connected wireless acoustic sensor network," *Signal Process.*, vol. 107, pp. 68–81, 2015.

[220] 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 Trans. Audio Speech Lang. Process.*, vol. 26, no. 8, pp. 1434–1448, 2018.

[221] R. M. Corey, M. D. Skarha, and A. C. Singer, "Cooperative audio source separation and enhancement using distributed microphone arrays and wearable devices," in *Proceedings of IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, Guadeloupe, West Indies, 2019.

[222] A. Bertrand and M. Moonen, "Robust distributed noise reduction in hearing aids with external acoustic sensor nodes," *EURASIP J. Adv. Signal Process.*, vol. 2009, 2009.

[223] 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 Trans. Audio Speech Lang. Process.*, vol. 24, no. 5, pp. 952–966, 2016.

[224] 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, August 2017, pp. 101–106.

[225] 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. 1–5.

[226] N. Cvijanovic, O. Sadiq, and S. Srinivasan, "Speech enhancement using a remote wireless microphone," *IEEE Trans. on Consumer Electronics*, vol. 59, no. 1, pp. 167–174, February 2013.

[227] D. Yee, H. Kamkar-Parsi, R. Martin, and H. Puder, "A Noise Reduction Post-Filter for Binaurally-linked Single-Microphone Hearing Aids Utilizing a Nearby External Microphone," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 1, pp. 5–18, 2017.

[228] 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 Trans. Audio Speech Lang. Process.*, vol. 25, no. 3, pp. 611–623, 2017.

[229] G. Courtois, "Spatial hearing rendering in wireless microphone systems for binaural hearing aids," Ph.D. dissertation, École polytechnique fédérale de Lausanne (EPFL), Lausanne, 2016.

[230] "ITU-T Recommendation X.200: Information technology - Open Systems Interconnection - Basic Reference Model: The basic model," International Telecommunication Union (ITU), Standard, 1994.

[231] O. Roy and M. Vetterli, "Rate-constrained collaborative noise reduction for wireless hearing aids," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 645–657, 2009.

[232] S. Srinivasan and A. C. den Brinker, "Rate-constrained beamforming in binaural hearing aids," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, p. 257197, Aug 2009.

[233] T. C. Lawin-Ore and S. Doclo, "Analysis of rate constraints for MWF-based noise reduction in acoustic sensor networks," *Proc. 2011 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '11)*, no. 3, pp. 269–272, 2011.

[234] A. Bertrand, "Utility metrics for assessment and subset selection of input variables for linear estimation [tips tricks]," *IEEE Signal Process. Mag.*, vol. 35, no. 6, pp. 93–99, Nov 2018.

[235] J. Zhang, S. P. Chepuri, R. C. Hendriks, and R. Heusdens, "Microphone Subset Selection for MVDR Beamformer Based Noise Reduction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 26, no. 3, pp. 550–563, 2018.

[236] R. Lienhart, I. Kozintsev, S. Wehr, and M. Yeung, "On the importance of exact synchronization for distributed audio signal processing," *Proc. 2003 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '03)*, vol. 4, pp. 840–843, 2003.

[237] Y.-C. Wu, Q. Chaudhari, and E. Serpedin, "Clock Synchronization of Wireless Sensor Networks," *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 124–138, 2011.

[238] 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," *Proc. IEEE Sixth International Symposium on Multimedia Software Engineering*, pp. 18–25, 2004.

[239] S. Miyabe, N. Ono, and S. Makino, "Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation," *Signal Process.*, vol. 107, pp. 185 – 196, 2015.

[240] M. H. Bahari, A. Bertrand, and M. Moonen, "Blind sampling rate offset estimation based on coherence drift in wireless acoustic sensor networks," *IEEE Trans. Audio Speech Lang. Process.*, vol. 25, pp. 674–686, 2017.

[241] R. M. Corey and A. C. Singer, "Speech separation using partially asynchronous microphone arrays without resampling," *Proc. 2018 Int. Workshop Acoustic Signal Enhancement (IWAENC '18)*, no. 1, pp. 111–115, 2018.

[242] D. S. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas (Second Edition)*. Princeton University Press, 2009. [Online]. Available: http://www.jstor.org/stable/j.ctt7t833

[243] 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. 5th Joint Workshop Hands-Free Speech Commun. Microphone Arrays (HSCMA '17)*, San Francisco, U.S.A., 2017, pp. 1–5.

[244] Bang and Olufsen, "Music for Archimedes," CD B&O 101, 1992.

[245] R.Ali, T. van Waterschoot, and M. Moonen, "Generalised sidelobe canceller for noise reduction in hearing devices using an external microphone," in *Proc. 2018 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '18)*, Calgary, AB, Canada, April 2018.

[246] ——, "Completing the RTF vector for an MVDR beamformer as applied to a local microphone array and an external microphone," in *Proc. 2018 Int. Workshop Acoustic Signal Enhancement (IWAENC '18)*, Tokyo, Japan, Sept 2018, pp. 1–4.

[247] I. Markovsky, *Low Rank Approximation: Algorithms, Implementation, Applications*. Springer, 2012.

[248] BBC Sound Effects data for Research and Education Space, 16,000 sound effects and field recordings, 2018. [Online]. Available: http://bbcsfx.acropolis.org.uk/assets/07070195.wav

[249] R.Ali. (2018). [Online]. Available: http://homes.esat.kuleuven.be/~ali/meth_extend_gsc_xms.html

[250] F. J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proc. IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[251] T. Van den Bogaert, "Preserving binaural cues in noise reduction algorithms for hearing aids," Ph.D. dissertation, KU Leuven, 2008.

[252] N. Gößling and S. Doclo, "Relative transfer function estimation exploiting spatially separated microphones in a diffuse noise field," in *Proc. 2018 Int. Workshop Acoustic Signal Enhancement (IWAENC '18)*, Tokyo, Japan, Sep. 2018, pp. 146–150.

[253] J. M. Kates, K. H. Arehart, and L. O. Harvey, "Integrating a remote microphone with hearing-aid processing," *J. Acoust. Soc. Amer.*, vol. 145, no. 6, pp. 3551–3566, 2019.

[254] Y. Zhao, J. K. Nielsen, J. Chen, and M. G. Christensen, "Model-based distributed node clustering and multi-speaker speech presence probability estimation in wireless acoustic sensor networks," *J. Acoust. Soc. Amer.*, vol. 147, no. 6, pp. 4189–4201, 2020.

[255] A. W. Bronkhorst and R. Plomp, "The Effect of Head-Induced Interaural Time and Level Differences on Speech-Intelligibility in Noise," *J. Acoust. Soc. Amer.*, vol. 83, no. 4, pp. 1508–1516, 1988.

[256] G. Keidser, K. Rohrseitz, H. Dillon, V. Hamacher, L. Carter, U. Rass, and E. Convery, "The effect of multi-channel wide dynamic range compression, noise reduction, and the directional microphone on horizontal localization performance in hearing aid wearers," *International Journal of Audiology*, vol. 45, no. 10, pp. 563–579, 2006.

[257] M. A. Akeroyd and W. M. Whitmer, "Spatial hearing and hearing aids," in *Hearing Aids*, G. R. Popelka, B. C. J. Moore, R. R. Fay, and A. N. Popper, Eds. Cham: Springer International Publishing, 2016, ch. 7, pp. 181–215.

[258] B. Cornelis, S. Doclo, T. Van den Bogaert, M. Moonen, and J. Wouters, "Theoretical Analysis of Binaural Multimicrophone Noise Reduction Techniques," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 2, 2010.

[259] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Binaural multichannel wiener filter with directional interference rejection," in *Proc. 2015 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '15)*, Brisbane, Australia, April 2015, pp. 644–648.

[260] D. Marquardt, "Development and evaluation of psychoacoustically motivated binaural noise reduction and cue preservation techniques," Ph.D. dissertation, Carl von Ossietzky Universität Oldenburg, 2015.

[261] 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 J. Adv. Signal Process.*, vol. 2016, no. 1, p. 12, 2016.

[262] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, J. Jensen, and M. Guo, "Binaural beamforming using pre-determined relative acoustic transfer functions," in *Proc. 25th European Signal Process. Conf. (EUSIPCO '17)*, Aug 2017, pp. 1–5.

[263] T. Piechowiak, C. Ma, R. de Vries, and A. Dittberner, "A binaural auditory steering strategy based hearing-aid algorithm design," *J. Acoust. Soc. Amer.*, vol. 143, no. 6, pp. EL490–EL495, 2018.

[264] R. Ali, G. Bernardi, T. van Waterschoot, and M. Moonen, "Methods of extending a generalized sidelobe canceller with external microphones," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 9, pp. 1349–1364, Sep. 2019.

[265] R. Van Rompaey and M. Moonen, "Distributed adaptive node-specific signal estimation in a wireless sensor network with partial prior knowledge of the desired source steering vector," in *Proc. 27th European Signal Process. Conf. (EUSIPCO '19)*, Sep. 2019, pp. 1–5.

[266] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Comm.*, vol. 102, pp. 1 – 13, 2018.

[267] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Comm.*, vol. 53, no. 5, pp. 592–605, 2011.

[268] Y. Takata and A. K. Nábělek, "English consonant recognition in noise and in reverberation by Japanese and American listeners," *J. Acoust. Soc. Amer.*, vol. 88, no. 2, pp. 663–666, 1990.

[269] K. L. Payton, R. M. Uchanski, and L. D. Braida, "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing," *J. Acoust. Soc. Amer.*, vol. 95, no. 3, pp. 1581–1592, 1994.

[270] M. Lavandier and J. F. Culling, "Speech segregation in rooms: Monaural, binaural, and interacting effects of reverberation on target and interferer," *J. Acoust. Soc. Amer.*, vol. 123, no. 4, pp. 2237–2248, 2008.

[271] M. Miyoshi and Y. Kaneda, "Inverse Filtering of Room Acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, 1988.

[272] M. Kallinger and A. Mertins, "Multi-channel room impulse response shaping - A study," *Proc. 2006 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '06)*, vol. 5, no. 2, pp. 101–104, 2006.

[273] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," *EURASIP J. Adv. Signal Process.*, vol. 2007, 2007.

[274] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 9, pp. 1879–1890, 2013.

[275] F. Lim, "Robust multichannel equalization for blind speech dereverberation," Ph.D. dissertation, Imperial College London, 2016.

[276] M. Triki and D. T. M. Slock, "Blind Dereverberation of a Single Source Based on Multichannel Linear Prediction," *Proc. 2005 Int. Workshop Acoustic Echo Noise Control (IWAENC '05)*, pp. 1–4, 2005.

[277] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *Proc. 2008 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '08)*, 2008, pp. 85–88.

[278] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-Channel Linear Prediction-Based Speech Dereverberation With Sparse Priors," *IEEE Trans. Audio Speech Lang. Process.*, vol. 23, no. 9, pp. 1509–1520, 2015.

[279] T. Dietzen, "Spatio-temporal speech enhancement in adverse acoustic conditions," Ph.D. dissertation, KU Leuven, 2019.

[280] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the MVDR beamformer in room acoustics," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 1, pp. 158–170, 2010.

[281] S. Braun and E. A. P. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proc. 21st European Signal Process. Conf. (EUSIPCO '13)*, Sept 2013.

[282] I. Kodrasi and S. Doclo, "Analysis of Eigenvalue Decomposition-Based Late Reverberation Power Spectral Density Estimation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 6, pp. 1102–1114, 2018.

[283] O. Schwartz, S. Gannot, and E. A. Habets, "Joint maximum likelihood estimation of late reverberant and speech power spectral density in noisy environments," *Proc. 2016 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '16)*, vol. 2016-May, pp. 151–155, 2016.

[284] S. Braun, A. Kuklasiński, O. Schwartz, O. Thiergart, E. A. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and Comparison of Late Reverberation Power Spectral Density Estimators," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 6, pp. 1052–1067, 2018.

[285] D. Schmid, G. Enzner, S. Malik, D. Kolossa, and R. Martin, "Variational Bayesian inference for multichannel dereverberation and noise reduction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 22, no. 8, pp. 1320–1335, 2014.

[286] B. Schwartz, S. Gannot, and E. A. Habets, "Online Speech Dereverberation Using Kalman Filter and EM Algorithm," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 2, pp. 394–406, 2015.

[287] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Joint multi-microphone speech dereverberation and noise reduction using integrated sidelobe cancellation and linear prediction," in *Proc. 2018 Int. Workshop Acoustic Signal Enhancement (IWAENC '18)*, Tokyo, Japan, Sept 2018.

[288] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Amer.*, vol. 125, no. 5, p. 3387, 2009.

[289] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 2, pp. 260–276, 2010.

[290] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," Mar. 2014.

[291] ——, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110.

[292] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, p. 1–122, Jan. 2011.

[293] J. Cubick and T. Dau, "Validation of a virtual sound environment system for testing hearing aids," *Acta Acustica united with Acustica*, vol. 102, no. 3, 2016.

[294] N. Mansour, M. Marschall, A. Westermann, T. May, and T. Dau, "Speech intelligibility in a realistic virtual sound environment," *Proceedings of the 23rd International Congress on Acoustics, Aachen, Germany*, no. September, pp. 7623–7630, 2019.

[295] A. Devesse, A. van Wieringen, and J. Wouters, "AVATAR Assesses Speech Understanding and Multitask Costs in Ecologically Relevant Listening Situations," *Ear and Hearing*, vol. 41, no. 3, 2020.

[296] I. Holube, P. V. Gablenz, U. Kowalk, M. Markus, and J. Bitzer, "Hearing aid benefit in everyday environments," *Proceedings of the 23rd International Congress on Acoustics, Aachen, Germany*, vol. 12, no. 2, p. 3869, 2019.

[297] B. H. Timmer, L. Hickson, and S. Launer, "The use of ecological momentary assessment in hearing research and future clinical applications," *Hearing Research*, vol. 369, pp. 24–28, 2018.

[298] G. Galvez, M. B. Turbin, E. J. Thielman, J. A. Istvan, J. A. Andrews, and J. A. Henry, "Feasibility of ecological momentary assessment of hearing difficulties encountered by hearing aid users," *Ear and hearing*, vol. 33, no. 4, pp. 497–507, Jul-Aug 2012.

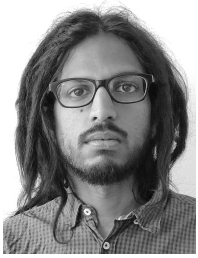
[299] S. Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs, 2019.

[300] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa, M. A. Abdelraheem, A. Abad, F. Teixeira, D. Matrouf, M. Gomez-Barrero, D. Petrovska-Delacrétaz, G. Chollet, N. Evans, T. Schneider, J.-F. Bonastre, B. Raj, I. Trancoso, and C. Busch, "Preserving privacy in speaker and speech characterisation," *Computer Speech and Language*, vol. 58, pp. 441 – 480, 2019.

[301] European Parliament and Council, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/ EC (General Data Protection Regulation)," 2016.

[302] ——, "Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data." 2016.

Biographical sketch



Randall Ali received a B.Sc. degree in electrical and computer engineering from the University of the West Indies in 2007 and an M.S. degree in acoustics from the Pennsylvania State University in 2013. Since 2016, he has been a Ph.D. candidate at KU Leuven in the research group of STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics, where his research has focused on the development of acoustic signal processing technologies for improving assistive hearing devices. Prior to KU Leuven, he has also been active in teaching and research in acoustics, including the fields of musical acoustics and thermoacoustics. In addition to his academic work, he is also an artist involved in music production.

List of publications

International journal publications

1. Ali R., Bernardi G., van Waterschoot T., Moonen M., "Methods of extending a generalised sidelobe canceller with external microphones", *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 27, no. 9, Sep. 2019, pp. 1349-1364

2. Ali R., van Waterschoot T., Moonen M., "Integration of a priori and estimated constraints into an MVDR beamformer for speech enhancement", *IEEE/ACM Trans. on Audio, Speech and Lang. Process.*, vol. 27, no. 12, Dec. 2019, pp. 2288-2300

International conference proceedings

1. Ali R., van Waterschoot T., Moonen M., "Using partial a priori knowledge of relative transfer functions to design an MVDR beamformer for a binaural hearing assistive device with external microphones", in *Proc. of the 23rd Int. Congress on Acoustics (ICA)*, Aachen, Germany, Sep. 2019, pp. 1-8.

2. Ali R., van Waterschoot T., Moonen M., "MWF-based speech dereverberation with a local microphone array and an external microphone", in *Proc. of the 27th European Signal Processing Conference (EUSIPCO)*, A Coruña, Spain, Sep. 2019.

3. Elvander F., Ali R., Jakobsson A., van Waterschoot T., "Offline Noise Reduction Using Optimal Mass Transport Induced Covariance Interpolation", in *Proc. 27th European Signal Process. Conf.*, A Coruña, Spain, Sep. 2019.

4. Ali R., van Waterschoot T., Moonen M., "Completing the RTF vector for an MVDR beamformer as applied to a local microphone array and an

external microphone", in *Proc. of the Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, Sep. 2018.

5. Ali R., van Waterschoot T., Moonen M., "Generalised Sidelobe Canceller for Noise Reduction in Hearing Devices using an external microphone", in Proc. of the 2018 *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Calgary, Canada, Apr. 2018.

6. Ali R. Moonen M., "A contingency multi-microphone noise reduction strategy based on linearly constrained multi-channel wiener filtering", in *Proc. of the 2016 IEEE Int. Workshop Acoustic Echo Noise Control (IWAENC 2016)*, Xi'an, China, Sept. 2016.

Internal reports

1. Ali R., van Waterschoot T., Moonen M., "An integrated MVDR beamformer for speech enhancement using a local microphone array and external microphones", Internal Report 20-76, ESAT-STADIUS, KU Leuven, *submitted for publication*, 2020.

Patent applications

1. Ali R., van Waterschoot T., Moonen M. "Integrated noise reduction", WO2020044166A1, *Application filed by Cochlear Limited*, 2019.



FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING
STADIUS CENTER FOR DYNAMICAL SYSTEMS, SIGNAL PROCESSING, AND DATA ANALYTICS
Kasteelpark Arenberg 10, bus 2446
B-3001 Leuven
randall.ali@esat.kuleuven.be
http://esat.kuleuven.be/stadius/