

KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT INGENIEURSWETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
AFDELING ESAT-SCD: SISTA/COSIC/DOCARCH
Kasteelpark Arenberg 10, B-3001 Leuven

DESIGN AND EVALUATION OF NOISE REDUCTION TECHNIQUES FOR BINAURAL HEARING AIDS

Jury:
Prof. em. dr. ir. H. Hens, voorzitter
Prof. dr. ir. M. Moonen, promotor
Prof. dr. J. Wouters, promotor
Prof. dr. ir. P. Wambacq
Prof. dr. ir. S. Van Huffel
Prof. dr. ir. S. Doclo
(University of Oldenburg, Germany)
Prof. dr.-ing. P. Vary
(RWTH Aachen University, Germany)

Proefschrift voorgedragen
tot het behalen van de
graad van Doctor in de
Ingenieurswetenschappen

door

Bram CORNELIS

December 2011

© 2011 Katholieke Universiteit Leuven, Groep Wetenschap & Technologie, Arenberg Doctoraatsschool, W. de Croylaan 6, 3001 Heverlee, België

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotocopie, 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 or any other means without written permission from the publisher.

ISBN 978-94-6018-453-6
D/2011/7515/154

Voorwoord

't Is gebeurd, na een dikke vier jaar zwoegen ben ik (bijna) klaar met het doctoraat, enkel het voorwoord resteert nog. Het eindproduct van vier jaar doctoreren is uiteindelijk dit boekje van ongeveer 200 bladzijden. Zoals een zekere persoon[1] voor mij berekende, komt dit neer op een gemiddelde output van slechts 1 blz. per week, "zo impressionant is dat toch niet"? Wel, uiteraard is een doctoraat een verhaal van vallen en opstaan, met dieptepunten waarbij een experiment maar niet wil werken en hoogtepunten waarbij je plots toch die ene goede ingeving krijgt. En, zoals een niet-onbekende persoon[2] ooit opmerkte, *"If we knew what it was we were doing, it would not be called research, would it?"*. Het punt dat ik hier echter wil maken is dat dit boekje niet kan weergeven dat een doctoraat een heel "proces" is (het een "avontuur" noemen is er misschien wat over). En, nog belangijker, dat er vele mensen zijn die onderweg bijgedragen hebben tot dit werk. In dit voorwoord wil ik dus een aantal mensen van harte bedanken.

Vooreerst bedank ik de promotoren van dit werk, Marc Moonen en Jan Wouters. Initieel was Marc diegene die me overtuigde om een doctoraat te starten in dit onderzoekstopic. Ik dank hem oprecht voor zijn vertrouwen[3], voor de vrijheid die hij me schonk tijdens het onderzoek, en voor de zeer nauwgezette verbeteringen[4] van mijn papers. Ik bedank ook oprecht mijn andere promotor Jan, met wie ik steeds nauwer ben gaan samenwerken in de loop van het doctoraat. Hij heeft me zeer veel bijgebracht over de audiologische en fysiologische kant van het probleem, en ik apprecieer dit ten zeerste. Ik wil Marc en Jan ook bedanken voor het verstrekken van een toffe werkomgeving (ik voelde me thuis, zowel bij de ESAT-DSP groep als in het ExpORL lab) en ook (zeker niet onbelangrijk) voor het organiseren van de "sociale activiteiten": bedankt Marc voor de befaamde "Chez Hilaire" avonden[5], en bedankt Jan voor de "Holsbeek visit" waar ieder jaar zo naar

[1] Met dank aan broer Thomas voor de beenharde analyse!

[2] A. Einstein

[3] Ondanks zijn zichtbare teleurstelling dat ik geen muzikant bleek te zijn...

[4] Helaas waren het handgeschreven verbeteringen in een waar doktersgeschrift, en daarenboven vaak ook nog eens doorgefaxt. Het leek op den duur op de ontcijfering van de Rosetta steen...

[5] En sinds dit jaar zelfs een BBQ bij Marc thuis, dat we dat nog mogen meemaken!

wordt uitgekeken. Het feit dat ik twee "bazen" had, blijkt hier ook een voordeel te zijn, want ik kon er dus van profiteren en naar beide feesten gaan.

Ik wens hierbij ook de examencommissie te bedanken. *I would like to thank all members of the jury for their efforts to read the text, making valuable suggestions, and allocating their time to attend my private and public defenses*: Promotoren Prof. Marc Moonen en Prof. Jan Wouters, Prof. Patrick Wambacq, Prof. Sabine Van Huffel, Prof. Simon Doclo, Prof. Peter Vary en de voorzitter Prof. Hugo Hens.

Ik ben het I.W.T. erkentelijk voor de financiële ondersteuning van mijn onderzoek gedurende dit doctoraat.

Er zijn verder ook twee personen die ik in het bijzonder wil bedanken, aangezien ze cruciaal waren voor de goede afloop van dit doctoraat: ik heb veel te danken aan Simon Doclo en Tim Van den Bogaert, die ik als mijn "mentoren" beschouw. Simon ken ik sinds hij begeleider was van mijn Master's thesis, je kan stellen dat hij me heeft warmgemaakt voor dit onderzoeksdomein. De theoretische kennis en onderzoeksmethodieken die hij me heeft aangeleerd, zijn van onschatbare waarde gebleken. Ik ben ook zeer verheugd dat hij nu (als professor) in mijn jury zetelt, zodat de cirkel rond is. Tim heeft me de praktische kant van de zaken geleerd, de fijne kneepjes van het "slechthorenden testen" vak. In zijn eigen thesis heeft hij, samen met Simon, ook stevige fundamenten voor dit onderzoekstopic gelegd, een zeer belangrijke factor voor het succes van dit doctoraat. Ik weet niet wat ik zonder Simon en Tim had moeten beginnen. Bedankt voor de fijne samenwerking.

Dan zijn er ook mijn collega's van de DSP groep, waartussen ik het grootste deel van mijn werktijd doorbracht. Deze mensen komen uit alle uithoeken van de wereld, wat zorgde voor een unieke mix van verscheidene culturen. *Thank you, crazy DSP guys: Pepe, Joe, Alexander, Geert, Ann, Simon, Toon, Gert, Sam, Deepak, Rodrigo, Kim, Sylwek, Pascal, Bruno, Javier, Vincent, Jan, Beier, Amir, Romain, Prabin, Geert and Johnny.* I particularly enjoyed our lunchtime conversations about Formula one, football, and the latest developments in indie rock music. Also, there were some legendary parties, of which I better not reveal any details[6]. Another unique experience that I'll always remember is the wedding in Beijing[7], which unfortunately happened right at the time I was supposed to be finishing up the first draft of this thesis text (stress!).

Ik had ook het voorrecht om een korte tijd te werken bij de Exp ORL groep. Laten we zeggen dat er een hele andere "vibe" is dan wanneer je een hele dag tussen de ingenieurs zit, en dat ik tevreden terugdenk aan deze fijne periode. Ik ga hier niet alle namen exhaustief opnoemen, maar enkele mensen in het bijzonder bedanken. Vooreerst dank ik Tim, Heleen en Sofie, waarmee ik nauw samenwerkte bij perceptuele evaluaties. Daarnaast wil ik ook mijn reisgenoten in San Francisco

[6]Only that most of the "incidents" seem to involve our crazy Spanish guy Pepe.

[7]Thanks again for the invitation, Beier. I'm still recovering from some of the "Lost in translation" experiences that happened there.

en Lake Tahoe expliciet bedanken: Anneke, Sofie, Tom en Nilesh. Achteraf bekeken was de conferentie in Lake Tahoe wellicht mijn allerbeste conferentie, ik wou dat we het allemaal nog eens zouden kunnen overdoen.

En dan zijn we nu toegekomen tot het bedanken van de familie, vrienden en kennissen. Hoewel deze mensen misschien geen directe invloed hebben op het onderzoekswerk, zijn ze toch van onschatbare waarde: telkens het onderzoek in een dipje geraakte (en ja, dat is geregeld gebeurd), waren zij er om me weer op te beuren.

Vooreerst zijn er mijn mede "burgies", die vaak de eerste golf van frustratie te verwerken kregen als er weer eens een simulatie mislukte. Bij deze, sorry voor mijn gezaag, Alexander, Eline, Joris, Gerry, Lieboud, Bram, Karen, Fleur, William, Joram, Eleonor, Rogier, Cyrielle, Pieter en Roel. Dank jullie voor al de vriendschap doorheen de jaren, de toffe reizen en skivakanties, enzovoort. Ik wil ook mijn maat Alexander hier nog eens in het bijzonder vermelden, aangezien hij samen met mij het doctoraat startte bij Marc Moonen. We hadden bijna dagelijks discussies over het onderzoekswerk, maar ook over allerlei zaken daarbuiten. Daarenboven was Alex ook zowat mijn vaste reisgezel naar de verschillende conferenties. Het is heel zeldzaam om met iemand op zoveel vlakken goed overeen te komen, ik apprecieer je vriendschap dus ten zeerste.

Een andere groep vrienden die ik ook wil bedanken zijn de "boardgamegeeks" uit St.Amands (=waar ik ben opgegroeid, sommigen ken ik dus al redelijk lang). Ik ben content dat we nog steeds tijd vinden om af en toe een boardgame te spelen: het heeft me zeker behoed voor een ferme "sanity loss" en misschien ook enkele "fate points" opgeleverd[8].

Ten slotte, *last but not least*, wil ik mijn familie bedanken voor alle steun. Dank aan mijn broer Thomas, die me vaak weer met de voetjes op de grond brengt (cfr. zijn eerdere analyse aan het begin van dit voorwoord). Ook zijn vriendin Cilia (die ik nu toch al graag als familie beschouw), dank ik voor de steun. En dan zijn er natuurlijk mijn ouders. Waar kan ik beginnen om jullie voor alles te danken? Ik kan me voorstellen dat jullie bij momenten wat ongerust waren over mijn carrièrekeuze ("Wanneer ga je klaar zijn en een echte job beginnen?"). Desondanks kon ik altijd op jullie steun rekenen. Ik hoop dat deze verwezenlijking ook voor jullie een prachtige beloning betekent. Bedankt voor alles.

Ziezo, missie volbracht. Bedankt nog eens allemaal. Aan de moedigen die de hele tekst willen lezen (goed zo!), wens ik alvast veel leesplezier.

Bram Cornelis,
Leuven, Dec. 2011

[8] inside jokes voor de gamers, voor de rest is het wijselijk om dit te negeren.

Abstract

One of the main complaints of hearing aid users is their degraded speech understanding in noisy environments. Modern hearing aids therefore include noise reduction techniques. These techniques are typically designed for a monaural application, i.e. in a single device. However, the majority of hearing aid users currently have hearing aids at both ears in a so-called bilateral fitting, as it is widely accepted that this leads to a better speech understanding and user satisfaction. Unfortunately, the independent signal processing (in particular the noise reduction) in a bilateral fitting can destroy the so-called binaural cues, namely the interaural time and level differences (ITDs and ILDs) which are used to localize sound sources in the horizontal plane. A recent technological advance are so-called binaural hearing aids, where a wireless link allows for the exchange of data (or even microphone signals) between the two devices. An advantage is that the signal processing can then be coordinated so that the binaural cues are better preserved.

The goal of this thesis work is to develop a binaural noise reduction technique for binaural hearing aids. The binaural technique should achieve further speech intelligibility improvements (compared to a bilateral technique), while the binaural cues of both the speech and the residual noise should also be better preserved. For this purpose, the Speech-Distortion-Weighted Multi-channel Wiener Filter (SDW-MWF), which is a well-known multi-microphone noise reduction technique, is adopted in this thesis. It will be demonstrated that the SDW-MWF is particularly suited for binaural noise reduction, both in terms of noise reduction performance and in terms of binaural cue preservation.

The first part of the thesis gives a general introduction to the problem statement. A theoretical binaural noise reduction framework is then presented. Finally, the binaural SDW-MWF technique is introduced.

The second part of the thesis focuses on the noise reduction performance of the considered binaural SDW-MWF technique. A perceptual evaluation of a realistic adaptive implementation of the binaural SDW-MWF is presented first. It is demonstrated that the binaural SDW-MWF indeed allows for significant speech

intelligibility improvements (compared to bilateral techniques), both for normal hearing subjects and for hearing aid users, which motivates further development and optimizations of the technique. A theoretical analysis of new SDW-MWF formulations is then presented. It is demonstrated that implementations based on these expressions can lead to further improvements in noise reduction performance (compared to the previously evaluated implementation). It is also demonstrated that the conceptually interesting structure of these new SDW-MWF formulations can be exploited in a number of ways. First, it is demonstrated how the structure can be used to increase the robustness to certain estimation errors, which occur in practice. Second, it is demonstrated that the structure also allows for a numerically favorable implementation.

The third part of the thesis focuses on the binaural cue preservation of the considered binaural SDW-MWF technique. First, it is proven that a full-bandwidth binaural SDW-MWF technique (where all microphone signals can be exchanged) preserves the binaural cues of the target speech source, while extensions can be formulated so that the residual noise cues can also be preserved. Second, we focus on the (speech) cue preservation of reduced-bandwidth binaural or bilateral SDW-MWF techniques (where not all microphone signals can be exchanged). It is then shown that the speech ITD cues are still preserved, whereas the speech ILD cues are distorted. Novel reduced-bandwidth techniques, which better preserve the speech ILD cues, are also proposed.

The fourth part of the thesis presents a binaural Voice Activity Detection (VAD) algorithm, which is used to detect periods of speech activity. Many noise reduction techniques (including the SDW-MWF) rely on a VAD, hence a robust VAD algorithm is highly sought for. The approach here consists in utilizing the wireless link of the binaural hearing aids to combine local VAD decisions into a superior fused VAD decision. Two novel binaural fusion VADs, which indeed achieve an improved performance versus the local VAD algorithms, are proposed.

Korte Inhoud

Een van de belangrijkste klachten van hoorapparaatgebruikers is hun slechte spraakverstaan in rumoerige omgevingen. Moderne hoorapparaten bevatten daarom ruisonderdrukkingstechnieken. Deze technieken zijn meestal ontworpen voor een monaurale toepassing, d.w.z. voor één hoorapparaat. De meeste hoorapparaatgebruikers dragen tegenwoordig echter hoorapparaten aan beide oren in een zogenaamde bilaterale fitting, aangezien het nu algemeen aanvaard is dat dit leidt tot een beter spraakverstaan en grotere tevredenheid van de gebruikers. Helaas kan de afzonderlijke signaalverwerking (in het bijzonder de ruisonderdrukking) in een bilaterale fitting de zogenaamde binaurale cues verstoren, met name de interaurale tijd-en niveauverschillen (ITDs en ILDs), waarmee geluidsbronnen in het horizontale vlak kunnen gelokaliseerd worden. Een recente technologische ontwikkeling zijn zogenaamde binaurale hoorapparaten, waar een draadloze verbinding toelaat om gegevens (of zelfs microfoonsignalen) uit te wisselen tussen de twee apparaten. Een voordeel is dat de signaalverwerking dan kan worden gecoördineerd, zodat de binaurale cues beter worden bewaard.

Het doel van dit doctoraatsproefschrift is om een binaurale ruisonderdrukkingstechniek voor binaurale hoorapparaten te ontwikkelen. De binaurale techniek moet leiden tot een verdere verbetering van het spraakverstaan (in vergelijking met een bilaterale techniek), terwijl de binaurale cues van zowel de spraak als de residuele ruis ook beter moeten worden bewaard. Om dit te verwezenlijken wordt de Spraak-Distortie-gewogen Meerkanaals Wiener Filter (SDW-MWF), een bekende meermicrofoons ruisonderdrukkingtechniek, aangenomen in dit proefschrift. Het zal worden aangetoond dat de SDW-MWF bijzonder geschikt is voor binaurale ruisonderdrukking, zowel op het vlak van ruisonderdrukkingsperformantie als op het vlak van binaurale cue bewaring.

Het eerste deel van het proefschrift geeft een algemene inleiding op de probleemstelling. Een theoretisch binauraal ruisonderdrukkingsschema wordt vervolgens gepresenteerd. Ten slotte wordt de binaurale SDW-MWF techniek geïntroduceerd.

Het tweede deel van het proefschrift richt zich op de ruisonderdrukkingsperformantie van de beschouwde binaurale SDW-MWF techniek. Een perceptuele

evaluatie van een realistische adaptieve implementatie van de binaurale SDW-MWF wordt eerst gepresenteerd. Het wordt aangetoond dat de binaurale SDW-MWF inderdaad zorgt voor significante verbeteringen van het spraakverstaan (in vergelijking met bilaterale technieken), zowel voor normaal horende proefpersonen als voor hoorapparaatgebruikers. Deze bevindingen motiveren een verdere ontwikkeling en optimalisering van deze technieken. Een theoretische analyse van nieuwe SDW-MWF formuleringen wordt vervolgens gepresenteerd. Het wordt aangetoond dat implementaties op basis van deze formuleringen kunnen leiden tot een verdere verbetering van de ruisonderdrukkingsperformantie (in vergelijking met de eerder geëvalueerde implementatie). Het wordt daarnaast ook aangetoond dat de conceptueel interessante structuur van deze nieuwe SDW-MWF formuleringen kan worden benut op een aantal manieren. Ten eerste wordt aangetoond hoe de structuur kan worden gebruikt om de robuustheid tegen bepaalde schattingsfouten, die zich voordoen in de praktijk, te verhogen. Ten tweede wordt aangetoond dat de structuur ook een numeriek gunstige implementatie mogelijk maakt.

Het derde deel van het proefschrift richt zich op de binaurale cue bewaring in de beschouwde binaurale SDW-MWF techniek. Ten eerste wordt bewezen dat een binaurale SDW-MWF techniek de binaurale cues van de gewenste spraakbron bewaart, indien alle microfoonsignalen uitgewisseld kunnen worden tussen de apparaten. Er kunnen daarnaast extensies worden geformuleerd zodat de binaurale cues van de residuele ruis ook worden bewaard. Ten tweede bestuderen we de (spraak) binaurale cue bewaring in gereduceerde-bandbreedte binaurale of bilaterale SDW-MWF technieken (waar niet alle microfoonsignalen kunnen worden uitgewisseld). Het wordt hierbij aangetoond dat de spraak ITD cues nog steeds bewaard blijven, terwijl de spraak ILD cues worden verstoord. Nieuwe gereduceerde-bandbreedte technieken, die de spraak ILD cues beter bewaren, worden daarom ook voorgesteld.

Het vierde deel van het proefschrift presenteert een binauraal spraakdetectie (VAD) algoritme, dat gebruikt wordt om periodes van spraakactiviteit te detecteren. Vele ruisonderdrukkingstechnieken (inclusief de SDW-MWF) maken gebruik van een VAD, zodat een robuust VAD-algoritme zeer gewild is. De hier gevolgde aanpak maakt gebruik van de draadloze verbinding van de binaurale hoorapparaten om lokale VAD beslissingen samen te voegen tot een superieure gefusioneerde VAD beslissing. Twee nieuwe binaurale fusie VADs, die inderdaad tot een betere prestatie leiden ten opzichte van de lokale VAD algoritmes, worden voorgesteld.

Glossary

Mathematical notation

a	scalar a
$\mathbf{a}$	column vector $\mathbf{a}$
$\mathbf{A}$	matrix $\mathbf{A}$
$\mathbf{I}$	identity matrix
$\mathbf{0}$	zero matrix/vector
$\approx$	approximately equal to
$\cong$	is parallel to
$\ll$	much less than
$\gg$	much greater than
$\mathcal{E}\{.\}$	expected value operator
$\{.\}^*$	scalar complex conjugation
$\text{vec}\{.\}$	stack columns of a matrix into a column vector
$\{.\}^T$	matrix transpose
$\{.\}^H$	matrix Hermitian transpose
$\{.\}^{-1}$	matrix inverse
$\text{Tr}\{.\}$	trace of a matrix, i.e. sum of diagonal elements
$\lvert \cdot \rvert$	absolute value (real numbers) or modulus (complex numbers)
$\lVert \cdot \rVert$	Euclidian vector norm, L2 norm
$\angle\{.\}$	argument or phase of complex number
$< \mathbf{u}, \mathbf{v} >$	dotproduct of vectors $\mathbf{u}$ and $\mathbf{v}$
$\nabla J(\mathbf{x})$	gradient of function J with respect to $\mathbf{x}$
$\frac{\partial f}{\partial x}$	partial derivative of function f with respect to x
$\lim_{x \to a} f(x)$	limit of function f as x approaches a
$\int_a^b f(x)\ \mathrm{d}x$	definite integral of function f for x on interval $[a, b]$
$\min\{x, y\}$	minimum of scalars x and y
$\max\{x, y\}$	maximum of scalars x and y
$\min_x$	minimize over x

$\max_x$	maximize over x
argmin_x	values x for which a minimum is attained
argmax_x	values x for which a maximum is attained
$\sum^x$	sum of elements in a set or sequence
$\prod$	product of elements in a set or sequence
$x \leftarrow a$	value a is assigned to x
$\log_{10}(x)$	common/decadic logarithm
$\exp(x)$	exponential function
$\cos(x)$	cosine function
j	imaginary unit

Acronyms and Abbreviations

ACE	Advanced Combination Encoder
ADM	Adaptive Directional Microphone
ANC	Adaptive Noise Canceller
ANOVA	Analysis Of Variance
BSS	Blind Source Separation
BTE	Behind-The-Ear
CASA	Computational Auditory Scene Analysis
cfr.	*conferatur*: see, compare
CP	Common Postfilter
dB	decibel
dBA	decibel A-weighting
DFT	Discrete Fourier Transform
e.g.	*exempli gratia*: for example
et al.	*et alii*: and others
FB	Fixed Beamformer
GSC	Generalized Sidelobe Canceller
GUI	Graphical User Interface
HINT	Hearing-In-Noise Test
HL	Hearing Level
HRTF	Head-Related Transfer Function
Hz	Hertz
ICA	Independent Component Analysis
IBM	Ideal Binary Mask
i.e.	*id est*: that is
IDFT	Inverse Discrete Fourier Transform
ILD	Interaural Level Difference
ITD	Interaural Time Difference

ITF	Interaural Transfer Function
LCMV	Linearly Constrained Minimum Variance
LMS	Least Mean Squares
MMSE	Minimum-Mean-Squared-Error
MVDR	Minimum-Variance Distortionless Response
MWF	Multi-channel Wiener Filter
MWF-η	SDW-MWF with partial noise estimation
MWF-ITF	SDW-MWF with ITF extension
NAL	National Acoustic Laboratories (Sydney, Australia)
NAL-RP	NAL-Revised, Profound (prescription rule)
OLA	Overlap-Add
OLS	Overlap-Save
PSD	Power Spectral Density
QRD	QR Decomposition
R1-MWF	Rank-one SDW-MWF
RLS	Recursive Least Squares
ROC	Receiver Operating Characteristics
RT_{60}	Reverberation Time (60 dB decay)
s.t.	subject to
SD	Speech Distortion
SDW-MWF	Speech-Distortion-Weighted MWF
SI	Speech Intelligibility
SII	Speech Intelligibility Index
SNR	Signal-to-Noise Ratio
SOS	Second-Order Statistics
SPL	Sound Pressure Level
SP-MWF	Spatial Prediction SDW-MWF
SRT	Speech Reception Threshold
TDOA	Time Difference Of Arrival
TF-GSC	Transfer Function ratio GSC
VAD	Voice Activity Detection
VU	Vrije Universiteit (Amsterdam)
WOLA	Weighted Overlap-Add

Contents

Voorwoord i

Abstract v

Korte Inhoud vii

Glossary ix

Contents xiii

I Introduction 1

1 General introduction and thesis outline 3

1.1 Hearing impairment and hearing aids 4

1.2 Spatial hearing 6

1.2.1 Localization cues 6

1.2.2 Impact on speech intelligibility in noise 8

1.3 Binaural hearing aids 9

1.4 Noise reduction in hearing aids 10

1.4.1 Single-microphone noise reduction 11

1.4.2 Multi-microphone noise reduction 13

1.4.3 Binaural noise reduction 17
1.4.4 Voice Activity Detection 18
1.5 Outline of the thesis 20
1.5.1 Main research objectives and schematic overview 20
1.5.2 Chapter by chapter overview 22

2 Binaural noise reduction framework and SDW-MWF 25
2.1 Configuration and notation 25
2.1.1 Speech, noise and microphone signals 25
2.1.2 Binaural processing scheme 27
2.1.3 Second-order statistics estimation 29
2.1.4 Special case: single target speech source 30
2.1.5 Noise reduction performance measures 31
2.1.6 Binaural cue preservation measures 33
2.2 Binaural SDW-MWF 35
2.2.1 General cost function and optimal filters 35
2.2.2 Special case: single target speech source 36
2.2.3 Reduced-bandwidth binaural SDW-MWF 37
2.2.4 Implementation of binaural SDW-MWF 39
2.3 Conclusion 40

II Noise reduction performance of binaural SDW-MWF techniques 43

3 Perceptual evaluation of bilateral and binaural adaptive SDW-MWF algorithms 45
3.1 Problem statement and research questions 46
3.2 Evaluated noise reduction algorithms 47
3.2.1 Bilateral and binaural SDW-MWF 47

3.2.2 Bilateral fixed beamformer 48
3.3 Methods 49
3.4 Experimental results 56
3.4.1 S0 N45 scenario 56
3.4.2 S90 N270 scenario 57
3.4.3 S0 N90/180/270 scenario 58
3.4.4 Cafeteria scenario 60
3.5 Discussion 61
3.5.1 Differences between subject groups 61
3.5.2 Algorithm performance 62
3.6 Conclusion 66

4 Rank-one SDW-MWF techniques and performance analysis under second-order statistics estimation errors 67

4.1 Introduction 68
4.2 Rank-one SDW-MWF techniques 70
4.3 Impact of speech correlation matrix estimation errors: estimated filters 73
4.4 Impact of speech correlation matrix estimation errors: narrowband output SNR 76
4.5 Experimental results 79
4.5.1 Setup and stimuli 79
4.5.2 Narrowband SNR improvement 80
4.5.3 Broadband SNR improvement: effect of the single-channel postfilter 86
4.5.4 SI-weighted SNR improvement and distortion 88
4.5.5 Multiple target speech sources 90
4.6 Conclusion 94

5 VAD-robust rank-one SDW-MWF techniques 95

5.1 Introduction 96
5.1.1 Problem statement 96
5.1.2 Preliminaries 96
5.2 Impact of VAD errors 97
5.2.1 Impact of VAD errors under stationary noise 97
5.2.2 Impact of VAD errors under nonstationary noise 98
5.3 VAD-robust extension 98
5.4 Experimental results 100
5.4.1 Setup and stimuli 100
5.4.2 Results 100
5.5 Conclusion 103

6 QRD-RLS algorithms for rank-one SDW-MWF techniques 105

6.1 Introduction 106
6.1.1 Adaptive SDW-MWF algorithms 106
6.1.2 Preliminaries 107
6.2 QRD-RLS algorithms 107
6.2.1 QRD-RLS algorithm for R1-MWF 107
6.2.2 QRD-RLS algorithm for SP-MWF 110
6.2.3 Residual extraction 110
6.3 Experimental results 111
6.3.1 Setup 111
6.3.2 SI-weighted SNR improvement 112
6.3.3 Impact of μ: single-channel postfilter 113
6.3.4 Robustness: effect of reduced wordlength 113
6.4 Conclusion 115

III Binaural cue preservation in binaural SDW-MWF techniques 117

7 Binaural cue preservation in full-bandwidth binaural SDW-MWF techniques 119

7.1 Introduction . . . 120

7.1.1 Binaural cue preservation in binaural noise reduction techniques . . . 120

7.1.2 Preliminaries . . . 121

7.2 Binaural SDW-MWF . . . 122

7.3 Partial noise estimation extension (MWF-η) . . . 123

7.4 Interaural transfer function extension (MWF-ITF) . . . 126

7.5 Experimental results . . . 129

7.5.1 Data model . . . 129

7.5.2 Performance of MWF-η . . . 130

7.5.3 Performance of MWF-ITF . . . 133

7.5.4 Comparison of SDW-MWF, MWF-ITF and MWF-η in a reverberant room . . . 137

7.6 Conclusion . . . 138

8 Binaural cue preservation in reduced-bandwidth bilateral and binaural SDW-MWF techniques 141

8.1 Introduction . . . 142

8.1.1 Problem statement . . . 142

8.1.2 Preliminaries . . . 143

8.2 Speech binaural cue preservation by reduced bandwidth SDW-MWF and MWF-η . . . 145

8.2.1 Output speech binaural cues . . . 145

8.2.2 Influence of μ . . . 146

8.2.3 Influence of η . . . 146

8.3 Reduced-bandwidth SDW-MWF with common postfilter . . . 147

8.3.1 Introduction 147
8.3.2 Bilateral MWF with common postfilter 148
8.3.3 Binaural MWF-Front with common postfilter 150
8.4 Experimental results 151
8.4.1 ILD estimation 151
8.4.2 Setup and stimuli 152
8.4.3 Algorithms 153
8.4.4 Batch versus adaptive implementation 154
8.4.5 Speech ILD per critical band and normalized cross-correlation function 155
8.4.6 Average speech ILD error and broadband SNR improvement 159
8.5 Conclusion 163

IV Binaural VAD algorithms 165

9 Binaural VAD based on decision fusion 167

9.1 Introduction 168
9.2 Overview VAD algorithms 168
9.2.1 Single-microphone energy VAD 168
9.2.2 Two-microphone cross-correlation based VAD 169
9.3 Binaural fusion VADs 170
9.3.1 Decision fusion of energy VADs: Fusion-1 170
9.3.2 Decision fusion of energy and cross-correlation VADs: Fusion-2 172
9.3.3 Decision fusion of energy and cross-correlation VADs based on optimal AND rule: Fusion-opt 173
9.4 Experimental results 174
9.4.1 Setup and stimuli 174
9.4.2 VAD performance: ROC graphs 175

9.4.3 SDW-MWF performance 178
9.5 Conclusion 182

V Conclusion **183**

10 Conclusions and further research **185**
10.1 Conclusions and summary 185
10.2 Suggestions for further research 189

A Appendix to Chapter 2 **193**
A.1 Optimal SDW-MWF 193
A.2 Single target speech source SDW-MWF 193

B Appendix to Chapter 7 **195**
B.1 MWF-ITF optimal filters for extension with noise ITF ($\gamma = 0$) . . 195
B.2 MWF-ITF optimal filters for extension with speech and noise ITFs ($\gamma \neq 0$) 197

C Appendix to Chapter 8 **199**
C.1 Relative speech ILD error of SDW-MWF and MWF-η 199
C.2 Common postfilter for MWF-Front 200

Bibliography **203**

List of publications **219**

Curriculum Vitae **221**

Part I

Introduction

Chapter 1

General introduction and thesis outline

Since several decades hearing aids offer a technical solution to people suffering from hearing loss. Modern hearing aids are no longer simple sound amplifiers, but fully digital devices with complex functionalities. Most notably, current hearing aids include advanced noise reduction techniques to improve the speech intelligibility in noisy environments (one of the main complaints of hearing aid users). The majority of hearing aid users currently have a hearing aid at both ears (a so-called bilateral fitting), as it is now widely accepted that two devices indeed lead to a better speech intelligibility and user satisfaction. The general awareness has also grown that spatial hearing (i.e. sound localization using the two ears) is important to hearing aid users. In particular, spatial hearing can also improve speech intelligibility in noise, by the effect of binaural unmasking. However, as the design of signal processing techniques in bilateral hearing aids does not take spatial hearing in consideration, the localization abilities of bilateral hearing aids users are compromised. In particular, it has been demonstrated that the so-called binaural cues (the time and level difference of the received sound signal at both ears), are distorted by current noise reduction techniques in bilateral hearing aids.

Recently, so-called binaural hearing aids were introduced. These binaural hearing aids are connected by a wireless link, which allows the exchange of data between the devices. As a result, the signal processing can be better coordinated, so that the binaural cues are better preserved. It is expected that in future binaural hearing aids, it will be possible to exchange one (or even more) microphone signals between the devices. The aim of this thesis is to develop binaural noise reduction techniques for this type of binaural hearing aids, where microphone signals from both devices can be combined and filtered in each device. As a result, an improved

noise reduction performance can then be achieved, while the binaural cues can also be preserved.

This chapter gives a general introduction to the problem statement and presents the outline of the remainder of the thesis.

Section 1.1 gives an introduction to hearing impairment and hearing aid technology.

Section 1.2 presents an overview of the psychophysics of spatial hearing. First, the different localization cues, which enable spatial hearing, are discussed. Second, it is motivated why spatial hearing is also an important aspect of speech intelligibility in noisy environments.

Section 1.3 discusses current state of the art binaural hearing aids. It is illustrated how current devices improve spatial hearing and speech intelligibility, but it is also motivated how further improvements can be achieved.

Section 1.4 presents an overview of noise reduction techniques in hearing aids. Both single-microphone and multi-microphone techniques are discussed. Recently proposed binaural noise reduction techniques will also be reviewed. Finally, as several noise reduction techniques rely on a Voice Activity Detector (VAD), a discussion of VAD algorithms is also presented.

Section 1.5 presents the outline of the thesis.

1.1 Hearing impairment and hearing aids

Three types of hearing loss are typically distinguished, i.e. conductive hearing loss, sensorineural hearing loss, or a mixture of both types. Conductive hearing loss occurs when there is a problem conducting sound waves from the outer ear to the inner ear (cochlea). This type of hearing loss can for example be treated by surgery, inserting tubes in the eardrum, or by so-called bone-anchored hearing aids (BAHAs) which transmit sound through bone conduction. Sensorineural hearing loss, where the problem is situated in the cochlea or further in the auditory system, is the most common form of hearing loss. It is for example caused by aging (age-related hearing loss or presbycusis) or exposure to loud sounds. A 2006 study [155] indicates that one in six adult Europeans suffers from a sensorineural hearing loss that is bad enough to have a negative impact on their daily lives. Most probably, this number will increase in the future as the population is aging (e.g. 2004 and 2008 numbers for the U.S. population can be found in [104, 106]). Moreover, the younger generation may also be at a higher risk because of the exposure to loud sounds (e.g. listening to an MP3 player causes measurable short-term auditory effects [99], but long-term effects still have to be investigated). For the U.S., it is

indeed observed that over the last generation, the hearing loss population grew at the rate of 160% of the total population growth [106].

Sensorineural hearing loss can be (partially) treated by using a hearing aid or cochlear implant (the latter for profound hearing losses). To compensate for the hearing loss, hearing aids have to apply a frequency-dependent amplification (usually the higher frequencies are more degraded, a so-called sloping hearing loss). Unfortunately, there are many other problems associated with hearing loss, so that simple sound amplification is insufficient.

With the advent of fully digital hearing aids in the 1990s, several of these problems can be (partially) managed by advanced signal processing algorithms [47, 79]. For example, hearing impaired people usually have a reduced dynamic range, i.e. the difference between the hearing threshold and the uncomfortable level is smaller than for normal hearing people. Dynamic range compression techniques therefore apply an amplification that is not only frequency-dependent but also level-dependent. Another important problem is caused by the short distance between the microphones and the hearing aid receiver (loudspeaker), which can create an annoying acoustic feedback. This feedback problem is even more pronounced in open fitting hearing aids, which have become more popular in recent years. Modern hearing aids therefore include adaptive acoustic feedback cancellation techniques. A final example of the signal processing in hearing aids is noise reduction, which will be discussed in more detail in section 1.4. The largest variety of hearing aids (behind-the-ear hearing aids), which are especially required for people with moderate-to-profound hearing loss, are typically equipped with two or three microphones to allow for multi-microphone noise reduction. This is motivated by the fact that people suffering from sensorineural hearing loss particularly experience great difficulties in noisy environments, so that improved speech intelligibility in noise is highly sought for [102]. Finally, it is noted that modern hearing aids include several other signal processing techniques besides the techniques discussed here, overviews can be found in [47, 79].

The design of signal processing techniques for hearing aids poses interesting challenges. On the one hand, there are difficult technical constraints. For example, the power consumption has to be limited, as this will affect the battery life (typical battery capacities are between 5 and 575 mAh, with a required battery life between 1 and 14 days [47]), so that efficient algorithms are required. On the other hand, there are also psycho-acoustical constraints. For example, the total input-to-output delay introduced by the signal processing has to be limited. Hearing aid users with mild-to-moderate hearing losses can perceive their own voice through three different paths, namely through the open vent of the hearing aid, through bone conduction and finally, delayed and amplified through the hearing aid processing. It was demonstrated in [171] that hearing aid delays exceeding 20 ms are then likely to become disturbing (i.e. cause perceivable coloration effects). For moderate-to-severe hearing losses longer delays may be tolerable,

but then the synchronization of audio and visual cues (i.e. lip movements) may become disrupted. These observations imply that the signal processing should not introduce large delays, and efforts to further decrease the delays are still undertaken in recent research work [140].

1.2 Spatial hearing

This section discusses the different localization cues by which the auditory system is capable of localizing sound sources in three dimensions, also referred to as spatial hearing. Preserving the spatial hearing ability of hearing aid users is important, for example in traffic situations. Moreover, it will also be explained that spatial hearing allows for a better speech intelligibility in noisy environments, most notably in the so-called cocktail-party situation.

1.2.1 Localization cues

Interaural Time Differences (ITDs) and Interaural Level Differences (ILDs)

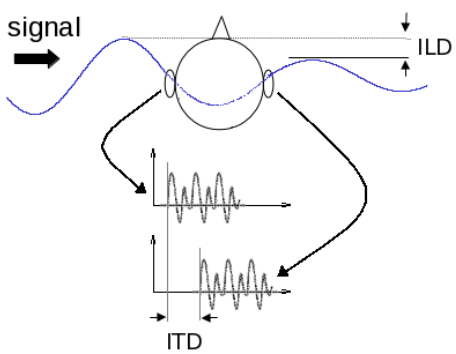


Figure 1.1: Sound wave impinging on the two ears, ITD and ILD cues.

The main cues responsible for localization in the horizontal (or azimuthal) plane are the so-called binaural cues, which are perceived using the two ears. When a sound wave arrives at the two ears from a certain direction, there will be a particular time and intensity difference between the two ears [14], cfr. Figure 1.1. These time and intensity differences are referred to as the Interaural Time Differences (ITDs) and Interaural Level Differences (ILDs). ITDs are within the range of -700 μs to +700 μs, with a time difference of 0 μs corresponding to a

source arriving from 0°, i.e. in front of the listener. Near the frontal direction (0°), listeners are sensitive to ITD differences as small as 13 μs, corresponding to an angle sensitivity of 1-2° [81]. The ILDs are highly frequency-dependent and can reach values up to 20-25 dB at higher frequencies. Listeners have shown sensitivity to ILD changes as small as 0.5 dB [81].

Over a century ago, Lord Rayleigh presented the duplex theory [147], which states that ITD cues are the dominant localization cues in the lower frequencies (< 1000-1500 Hz), while ILD cues are dominant in the higher frequencies. The fact that ITD cues are only usable in the lower frequencies, is explained by the observation that the wavelength of the sound signal becomes smaller than the diameter of the head in the 1000-1500 Hz range. Artificially introducing ITDs at higher frequencies does not lead to a spatial percept [81]. ILDs do occur over the entire frequency range (also in the lower frequencies), but normally only have a large magnitude in the higher frequency range. However, if ILDs are artificially introduced at lower frequencies, they can introduce a spatial percept [66, 81].

Other localization cues

The binaural cues allow for localization in the horizontal plane, but they do not contain information about the distance between the source and listener, the elevation angle (vertical localization), and finally, whether a source is in the front or the back hemisphere (front-back localization). A so-called cone of confusion can thus be defined, which contains all possible source positions for a particular value of the ITD and ILD cues. Other localization cues are therefore required to resolve this ambiguity. A recent survey [4] lists the following auditory localization cues (besides the binaural ITDs and ILDs):

- Monaural (spectral) cues depending on the shape of the outer ear (pinna), i.e. so-called pinna cues.
- Cues from torso reflection and diffraction.
- The ratio of direct to reverberant energy.
- Cue changes by voluntary head movements.
- Familiarity with the sound source.

Finally, it should be noted that visual cues and other non-auditory cues [7] can also influence spatial sound perception.

Pinna cues are less accessible to hearing aid users, as these only occur at very high frequencies (> 6000 Hz), or may not even be present (i.e. in behind-the-ear

hearing aids, which are considered here) [79]. As a result, the vertical and front-back localization ability of hearing aid users can be assumed to be limited. Front-back confusions can be reduced by using a directional microphone or acoustic beamforming (section 1.4) [98], as this typically introduces a higher sensitivity (amplification) towards the front hemisphere [79]. In this thesis, we will however only consider the binaural cues. Thus, when we refer to sound localization in this thesis, we implicitly mean horizontal or left-right localization.

Cue perception and hearing loss

Sensorineural hearing loss can have a negative effect on the horizontal localization ability, but this depends highly on the nature and the severity of the hearing loss. For people with mild hearing losses, it was demonstrated that the binaural cues can still be utilized [98]. However, people who have a severe low-frequency hearing loss have a degraded localization ability [98,137]. On the other hand, people with severe high-frequency hearing loss and mild low-frequency hearing loss (which is more typical) can still make use of the lower frequencies for localization [98,137]. They then especially utilize the low-frequency ITD cues, but not the ILD cues. The processing of the hearing aids should however still be controlled so that it does not introduce erroneous ILD cues. As noted before, artificially introduced ILD cues at lower frequencies can still create a spatial percept [66,81].

1.2.2 Impact on speech intelligibility in noise

Correct sound localization is an appealing and important property in its own right (consider for example traffic situations). As discussed here, it however also leads to a better speech intelligibility in noise. A well-known example is the so-called cocktail-party scenario [25,82], where a particular speaker of interest has to be understood in a mixture of competing speakers. The fact that two ears allow to localize the competing speakers in different directions, indeed helps accomplishing this task.

There are actually three distinct effects by which listening with two ears leads to improved speech intelligibility in noise. *Binaural summation* refers to the advantage of listening with two ears, when two identical mixtures of speech and noise (same input SNRs) are presented to the two ears. Informally, one could state that the brain gets "two looks" at the same sound, which makes it easier to identify the target speech component [47]. The equivalent SNR benefit by the binaural summation effect is in the order of 0.5 to 2 dB [19,47]. The *headshadow effect* is a purely physical effect which is due to the arrangement of the speech and noise sources around the head. In many scenarios a best-ear side (with high SNR) is created, e.g. when the speech source is closer to one ear than to the

other. Headshadow effects up to 10 dB have been measured [19]. Because of the headshadow effect, the listener will mainly use this best ear to understand the presented speech. However, adding a second (low SNR) ear can still lead to further speech intelligibility improvements. This is explained by the third effect, which is purely caused by the binaural spatial processing (with ITDs and ILDs), and denoted as *binaural unmasking* [19]. Other studies confirm this and report that due to binaural unmasking an intelligibility gain of 2-3 dB can be expected [194].

1.3 Binaural hearing aids

For normal hearing subjects, it has always been widely accepted that binaural hearing enhances speech intelligibility and spatial hearing, i.e. two ears are better than one [113]. In contrast, the question if two hearing-impaired ears coupled to two hearing aids (a bilateral fitting) are also better than one hearing aid (a unilateral fitting), has long been disputed [113]. Since the mid-1990s, the majority of hearing aid users do prefer bilateral fittings [107, 113]. Currently about four in five U.S. hearing aid users have a bilateral fitting [106, 113].

As the two devices in a bilateral fitting work independently, the hearing aid user is not able to benefit from the full binaural advantage. For example, uncoordinated fast compression schemes in the left and right device can lead to ILD distortions [110]. It was also demonstrated that the localization performance of bilateral hearing aid users is severely compromised when the multi-microphone noise reduction is switched on, versus the performance when it is switched off [178]. As stated in section 1.2.2, correct sound localization is an important property in its own right, but also affects speech intelligibility in noise. The fact that the uncoordinated signal processing in bilateral hearing aids compromises the localization performance may therefore be one of the reasons why only 29% of hearing aid users reported satisfaction with the performance in noise in a 2005 survey [103].

To improve the user satisfaction, the hearing aid industry sees great potential in the application of wireless technology in hearing aids [16, 56, 79, 105]. The industry hereby has two different applications of wireless technology in mind. In the first type of wireless application, the hearing aids are equipped with wireless receivers which pick up audio signals from remote external devices. A notable early example is the remote wireless microphone [61]. Other obvious audio sources for this application include radios, televisions or MP3 players. In the second type of wireless application, the hearing aids are not only able to receive data, but can also transmit data to the device at the other ear. These type of hearing aids, where the left and right device can exchange data over a wireless link, are referred to as binaural hearing aids. The benefit of binaural hearing aids is that coordination

of the signal processing is possible, so that the full binaural advantage can be achieved, in contrast to the previously discussed bilateral hearing aids.

The first commercial binaural hearing aid system was introduced by Siemens in 2004 (Siemens Acuris) [80,145]. The data transfer rate of the wireless link was 215 bits/s, which only allowed for the exchange of control data between the devices. An example application is the synchronization of the volume control: if the user changes the volume settings in one of the devices, the other device automatically adjusts its volume so that the ILDs are preserved [145]. Other possible applications include a binaural classification system [80], coordination of compression [110] and better identification of feedback [110]. More recently, Oticon introduced binaural hearing aids which can wirelessly exchange audio signals (at a rate of 120 kbit/s) with each other and other devices, through an external streamer device (Oticon Epoq Streamer) [115]. A similar product is the Widex Clear440, where one audio signal together with control data can be transmitted from one ear to the other (at a total rate of 212 kbit/s) [110]. A target application is the co-called Phone+ feature: during telephone use, the audio signal received at one device (closest to the telephone loudspeaker) is transmitted to the other device, whereby the other device mutes its own microphone signals. Because of binaural summation, the perceived phone signal is louder so that the conversation is easier to follow [110]. The Phonak Exelia line of hearing aids allow for real-time streaming of one audio signal between the devices, in a direction chosen by the user (ZoomControl) [138]. For example, if the user specifies that the signal of interest is to the right side of the head, a microphone signal of the right device is transmitted to the left device. The left device then uses this received signal instead of its own microphone signals, which are muted.

In summary, several commercial binaural hearing aid products have been introduced in recent years. In the current state of the art, one microphone signal can be transmitted from one device to the other, in half-duplex mode. It can be expected that in future devices it will also be possible to exchange one (or even more) microphone signals in full-duplex mode. These systems would allow for binaural multi-microphone noise reduction techniques, where microphone signals from both devices can be combined and filtered in each device. It is then expected that a superior noise reduction performance can be achieved, while the binaural cues can also be preserved. The aim of this thesis is to develop binaural noise reduction techniques for this type of binaural hearing aids.

1.4 Noise reduction in hearing aids

A major problem for hearing aid users is the degradation of their speech understanding in a noisy environment. Sensorineural hearing loss is most often accompanied by a loss of spectral and temporal resolution in the auditory

processing, which results in an SNR loss of about 4-10 dB [47, 79]. An SNR improvement of just 1 dB around the SRT level (the SNR level where 50% of the speech can be correctly identified) can already generate an increase in speech understanding of 15% in everyday conversations [144]. In a 2002 study [102], improved noise reduction performance is the most cited improvement requested by hearing aid users. Not surprisingly, noise reduction in hearing aids has therefore been an active area of research for several years. The different noise reduction techniques can be roughly classified into two categories, i.e. single-microphone noise reduction and multi-microphone noise reduction. Binaural noise reduction can be viewed as a special case of multi-microphone noise reduction (i.e. using microphone signals from both devices), but will be discussed in a separate section. Finally, as many noise reduction techniques rely on a Voice Activity Detector (VAD), which classifies audio frames as noise-only or speech+noise frames, an overview of VAD algorithms is also presented in this section.

1.4.1 Single-microphone noise reduction

Single-microphone noise reduction is concerned with improving the perceptual aspects of speech degraded by additive noise [118]. In literature, it is therefore also referred to as speech enhancement.

A common technique in hearing aids applies long-term attenuations on frequency bands with very low SNRs [26, 79]. A *modulation frequency* analysis hereby determines whether a frequency band is more likely to contain a speech or a noise signal. A disadvantage is that the speech and noise signals should be in different frequency ranges in order to achieve a good noise reduction performance [79].

Other techniques apply short-term attenuations so that a better noise reduction performance can be achieved if the speech and noise components are in the same frequency range [79]. A recent overview [118] divides the techniques in three classes:

- *Spectral-subtraction techniques*: spectral subtraction was initially proposed in [11, 15, 189]. In these techniques, an estimate of the noise spectrum is subtracted from the input microphone signal spectrum. The output signal is then constructed using the same phase as the original microphone signal, as it was observed that the phase does not affect speech intelligibility. To estimate the noise spectrum, it is assumed that the noise is additive and stationary. By using a VAD, the noise spectrum can then be updated in periods of speech absence.
- *Statistical-model-based techniques*: statistical-model-based techniques were introduced in [59, 114, 129]. These techniques pose the speech enhancement problem in a statistical estimation framework. The objective is then to find

an optimal linear or nonlinear estimator of the DFT coefficients of the target speech signal.

- *Subspace techniques*: subspace techniques were introduced in [45, 60]. The techniques decompose the vector space of the noisy microphone signal into a subspace that is occupied primarily by the target speech signal, and a subspace occupied primarily by the noise signal. An estimate of the target speech signal is then obtained by nulling the component residing in the noise subspace. The decomposition into signal and noise subspaces is obtained using well-known matrix factorization techniques, in particular the singular value decomposition [45] or eigenvalue decomposition [60].

Another related technique is the so-called *time-frequency masking* approach [187]. In this approach, the microphone signal is first transformed into a time-frequency representation (e.g. obtained by a short-time Fourier transform). A mask (i.e. gain function) is then applied on the time-frequency points, so that the resulting signal after synthesis is enhanced. In essence, the previously discussed speech enhancement techniques [118] can also be viewed as time-frequency masking approaches, but there are two distinct features in the techniques considered here [187]. First, the algorithms for computing the mask originate from the field of Computational Auditory Scene Analysis (CASA), which aims at segregating sound sources based on perceptual principles [17]. Second, the time-frequency masks in these techniques typically take a binary value (0 or 1), whereas the classical speech enhancement techniques apply continuous gains. The so-called Ideal Binary Mask (IBM) is currently viewed as the major computational goal of CASA techniques [112, 186], and encouraging perceptual results have indeed been obtained [111]. More recently however, it was reported that a practical IBM algorithm, where the IBM has to be estimated from the noisy microphone signal, is actually outperformed by algorithms which apply a continuous gain [93].

Although single-microphone noise reduction techniques may increase the SNR, this necessarily comes at the price of introducing speech distortion [24]. In addition, other implementation-specific artefacts like so-called musical noise [118] may occur in practice. Maybe as a consequence, current single-microphone noise reduction techniques are found not to be able to achieve speech intelligibility improvements in perceptual evaluations [87,119,123,173]. However, although intelligibility is not improved, hearing aid users do find single-microphone noise reduction useful, as it can attenuate the overall noise level and improve the listening comfort and listener preferences [123,173]. In particular, single-microphone noise reduction is especially suited for combatting diffuse noise scenarios or highly reverberant environments, where the performance of multi-microphone noise reduction (cfr. next section) is limited and a lot of residual noise remains [58]. In hearing aids, single-microphone noise reduction is therefore often included as a postfiltering stage, acting on the output of a multi-microphone noise reduction algorithm [157].

1.4.2 Multi-microphone noise reduction

Multi-microphone noise reduction techniques combine and filter different microphone signals in order to achieve an SNR improvement. The fact that the target speech source and interfering noise sources are in different locations is hereby exploited: constructive interference is created in the direction of the speech source, while destructive interference is created in others directions, cfr. Figure 1.2. This concept is also referred to as beamforming, and was originally applied in antenna transmission/reception and radar [181]. For hearing aid applications, it is widely accepted that multi-microphone noise reduction or beamforming can achieve significant speech intelligibility improvements [47,79,149,173]. Most state of the art hearing aids therefore contain two or three microphones, and perform some form of multi-microphone noise reduction. In the remainder of this section, different classes of multi-microphone noise reduction techniques are discussed.

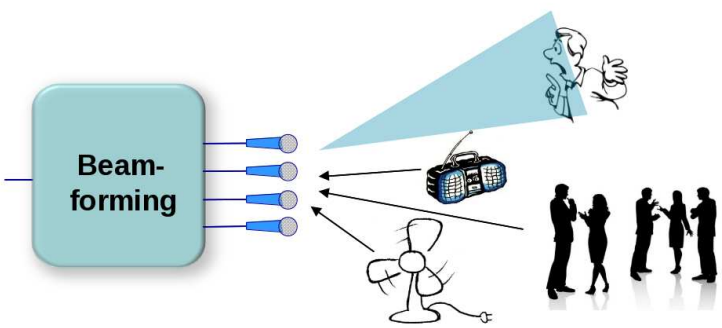


Figure 1.2: Acoustic beamforming with constructive interference for the direction of a target speech source, and destructive interference in other directions.

Fixed beamforming

Fixed beamformers use time-invariant filters and are hence data-independent techniques. Several techniques have been considered for hearing aids, e.g. directional microphones, delay-and-sum beamformers and superdirective beamformers (overviews are found in [47,57,79]). Notable examples are the above-the-ear endfire array [124] and the so-called hearing glasses, where multiple microphones are built into the frame of eyeglasses [159].

As fixed beamformers are data-independent techniques, an assumption has to be made about the target speech source location. Most often, it is assumed that

the target speech source is in the frontal (facing) direction. Unfortunately, this asssumption is not fulfilled in some commonly encountered scenarios (consider for example a conversation between a driver and passengers in a car, where the hearing aid user is driving and has to keep his eyes on the road). Some techniques therefore try to estimate the direction of the target speech source using source localization algorithms [72].

Another disadvantage of fixed beamformers is the sensitivity to "model imperfections" such as microphone mismatch, caused by aging effects or environmental influences [79]. Nevertheless, significant intelligibility improvements can indeed be demonstrated in certain conditions (see for example [149] for an evaluation of directional microphones, [124] for an evaluation of the above-the-ear endfire array and [132] for an evaluation of the hearing glasses).

Adaptive beamforming

In contrast to fixed beamformers, adaptive beamformers are data-dependent techniques that can adapt to changing noise scenarios. As a result, they can offer a better noise reduction performance in certain scenarios [79].

The most commonly used and computationally cheapest adaptive beamforming technique is the so-called adaptive directional microphone (ADM) [122, 150]. The ADM consists of two fixed directional microphones (where one has a forward-facing spatial pattern and the other has a backward-facing spatial pattern), followed by an adaptive part. The adaptive part, typically implemented as a one-tap adaptive scalar, hereby combines the two fixed spatial patterns so that a null is formed in the direction of the strongest noise interferer.

A more general class of techniques continuously minimizes the total output power of the beamformer under the constraint that the response towards one (or more) target speech directions is preserved. The so-called Minimum Variance Distortionless Response (MDVR) beamformer [21] constrains the response towards a single target direction (e.g. the facing direction of the hearing aid user), the so-called Linearly Constrained Minimum Variance (LCMV) beamformer [68, 181] extends this to a set of linear constraints.

A popular implementation of these beamformers is the Generalized Sidelobe Canceller (GSC) algorithm [76], which transforms the constrained optimization problem into an equivalent but easier unconstrained optimization problem. The GSC consists of a spatial preprocessing stage followed by a multi-channel adaptive noise canceller (ANC) [191]. The spatial preprocessing stage has two branches operating on orthogonal subspaces: a fixed beamfomer which generates a so-called speech reference, and a blocking matrix which steers nulls in the direction of the target speech source, generating so-called noise references. The initial GSC

algorithm [76] assumed freefield propagation, but this was extended into arbitrary transfer functions in the Transfer Function ratio GSC (TF-GSC) algorithm [69].

For hearing aid applications, the GSC algorithm may be viewed as the current state of the art, and it indeed leads to a significant benefit in certain scenarios [166,183]. However, similar to fixed beamformers it still relies on *a priori* knowledge or assumptions about the target speech location, room reverberation and microphone characteristics, in order to construct the spatial preprocessing stage. These assumptions are usually violated in practice, so that some of the speech signal leaks into the noise references. As a result, the performance may degrade significantly [164]. Several extensions were proposed to mitigate the leakage problem (see e.g. [49] for a discussion). A common solution is to make use of a VAD algorithm: the ANC is only activated in noise-only periods in order to avoid signal cancellation of the target speech signal [75,174].

Multi-channel Wiener Filtering (MWF)

A different class of multi-microphone noise reduction techniques is based on Multi-channel Wiener Filtering (MWF) [52,54,162], which is basically a generalization of single-microphone techniques [114, 118]. The MWF produces a minimum-mean-squared-error (MMSE) estimate of the speech component in a reference microphone signal, by exploiting speech and noise correlation matrices (second-order statistics). To provide an explicit trade-off between speech distortion and noise reduction, the Speech-Distortion-Weighted Multi-channel Wiener Filter (SDW-MWF) was also proposed in [52,54,162]. This extension is equivalent to applying additional single-microphone noise reduction to the output of a multi-microphone noise reduction algorithm, which, as previously mentioned, is generally considered useful by hearing aid users.

In contrast to the previously discussed adaptive beamformers, the SDW-MWF can adapt to both changing speech and changing noise scenarios. In principle, the SDW-MWF does not require *a priori* knowledge or assumptions about the target speech location and microphone characteristics, unlike the GSC. It is therefore expected to be more robust, which was indeed demonstrated in [164]. A frequency-domain adaptive implementation of the SDW-MWF was also proposed in [54]. This approach is computationally advantageous as every frequency bin can be processed separately. The SDW-MWF thus offers a robust and computationally efficient alternative to the GSC.

In this thesis, the SDW-MWF technique is adopted for binaural noise reduction, and it will be demonstrated that it is especially suited for this application. A more detailed discussion of the SDW-MWF and its binaural extension will therefore be presented in chapter 2.

Blind Source Separation (BSS)

Blind Source Separation (BSS) techniques consider the microphone signals as different observations of mixtures of audio signals, whereby each audio signal is a filtered version of a certain source signal. Under the assumption that the sources are statistically independent, the number of microphones is at least equal to the number of sources, and at most one source is Gaussian, the original source signals can be recovered (up to a scaling and permutation) using algorithms such as Independent Component Analysis (ICA) [29,90]. An interesting feature of the BSS techniques is that they do not require *a priori* knowledge or assumptions about the target speech and noise locations and microphone characteristics, similar to the SDW-MWF. Moreover, unlike the SDW-MWF, BSS techniques do not require a VAD. The BSS techniques for acoustic signal separation can be divided into two classes, i.e. frequency-domain techniques and time-domain techniques.

As audio signal mixtures (i.e. microphone signals) in a reverberant environment are convolutive mixtures, the first class of BSS techniques considers the problem in the frequency domain (e.g. [154]). Per frequency bin, the observed mixture can be approximated as an easier instantaneous mixture, so that the BSS problem is solvable by the standard ICA algorithm [90]. However, as each frequency bin is processed separately, the inherent BSS permutation and scaling ambiguities pose a problem, so that the different frequency bins have to be re-aligned in a postprocessing stage [154]. The second class of BSS techniques considers the original convolutive BSS problem in the time domain (e.g. [20]), so that permutation and scaling ambiguities are not an issue. However, a disadvantage is that the algorithms then involve computationally expensive operations on large matrices. A computationally cheaper variant using block-online adaptation and frequency-domain fast convolution techniques was therefore also proposed [1].

As the computational complexity of BSS algorithms is prohibitive for implementation in current hearing aids, relatively few studies have measured the perceptual speech intelligibility improvements obtained with these algorithms. A two-microphone implementation of the time-domain BSS technique [1, 20] was assessed in a multi-center evaluation [123], together with an implementation of the SDW-MWF and single-microphone algorithms. Good results were obtained for a scenario with a single target source and one interfering noise source (i.e. number of sources equal to number of microphones), although the results were only marginally better than the SDW-MWF results. In contrast, the BSS performance completely degraded in a scenario with three interfering noise sources (i.e. number of sources larger than number of microphones), as one of the BSS assumptions is violated. Even without considering the computational complexity, the applicability of current BSS techniques in hearing aids thus seems limited.

More recently, several BSS extensions have been proposed in order to retain the performance in an underdetermined BSS problem (i.e. more sources than

microphones) [125]. These approaches make use of the disjoint sparsity of speech signals, i.e. it is assumed that the involved signals rarely have energy concentrated at identical time-frequency points [116]. In another recent contribution [120], the time-domain BSS technique [20] is applied for sound source localization. The BSS-based method outperforms other known techniques, most notably in high-reverberant environments. Interestingly, the approach remains applicable in the underdetermined case, where there are more sources than microphones. In light of these advances, and assuming that the computational power of hearing aids will continue to increase, it is thus possible that BSS techniques will become applicable for noise reduction or sound localization in future hearing aids.

1.4.3 Binaural noise reduction

In recent years, several binaural noise reduction techniques have been proposed which are targeted towards future binaural hearing aids, where the full-duplex exchange of microphone signals between the two devices is feasible. The aim of binaural noise reduction techniques is to improve the SNR of the signal, while simultaneously preserving the binaural cues of both target speech and residual noise. The binaural noise reduction techniques can be roughly divided into two classes.

In a first class of binaural techniques, identical real-valued spectral weights are applied to one microphone signal on the left device and one microphone signal on the right device [78, 80, 97, 108, 121, 148, 151, 192], so that the binaural cues are indeed preserved. The same concept was also recently applied in a binaural dereverberation technique [94, 95]. Although the outputs of a beamformer can be utilized to derive the spectral weights (e.g. a superdirective beamformer [121,151]), in essence these techniques can be viewed as spectral filtering approaches, similar to techniques in single-microphone noise reduction. As previously discussed, single-microphone noise reduction usually introduces speech distortion and other artefacts, leading to limited or no speech intelligibility improvements [173]. The perceptual speech intelligibility improvements obtained with this class of binaural techniques may thus also be limited, or not competitive with true beamforming techniques.

A second class of binaural techniques combines microphone signals from both devices to perform a true beamforming. Some techniques first construct a monaural output and then apply a postprocessing stage to reconstruct the binaural signals with correct binaural cues (e.g. postprocessing for binaural BSS techniques [2, 80, 188]; postprocessing for a binaural superdirective beamformer [131]). Other techniques apply fixed or adaptive beamformers which produce a binaural ouput, whereby the beamformers are designed or constrained so that the binaural cues are also preserved [2, 46, 117, 133, 136, 190]. For example, in [46, 190]

(adaptive) beamforming is only applied in the higher frequencies, while for the lower frequencies the (low-pass filtered) original microphone signal, with correct binaural cues, is used. The binaural SDW-MWF technique adopted in this thesis also performs a true beamforming with the microphone signals in order to produce binaural output signals, hence it belongs to this second class of techniques. It will be demonstrated that the SDW-MWF cost function can be extended so that the binaural cues of both the target speech and the residual noise can be preserved.

In the second class of binaural techniques, there is necessarily a trade-off between noise reduction performance and binaural cue preservation, which is not the case for the first class of binaural techniques. However, as noted before, the first class of techniques may offer only limited actual speech intelligibility improvements. The performance comparison in [117] showed that a binaural MVDR beamformer (second class of techniques) introduced less speech distortion compared to the technique in [108] (first class of techniques), for a comparable (physical) noise reduction performance. More recently [126], the fixed beamformer with binaural postfilter of [151] (first class of techniques) was compared to binaural SDW-MWF techniques (second class of techniques), on a group of 10 normal hearing subjects. The fixed beamformer with binaural postfilter did achieve small SRT improvements, but was significantly outperformed by the SDW-MWF based approaches which perform true beamforming.

1.4.4 Voice Activity Detection

A VAD algorithm classifies audio frames as either noise-only or speech+noise frames. In Figure 1.3, a sample microphone signal (four sentences with periods of silence and additive noise) and the corresponding perfect VAD output are shown. The VAD makes a binary decision on a frame-by-frame basis, whereby frames are typically 20-40 ms. VAD algorithms are commonly used in both single- and multi-microphone noise reduction. In single-microphone noise reduction, where an estimate of the noise spectrum is often required, a VAD can be used so that the noise spectrum is updated during noise-only periods. For adaptive beamforming, it was noted that the ANC is usually only updated in noise-only periods to avoid speech cancellation, again requiring a VAD. Finally, the SDW-MWF technique, which is adopted in this thesis, requires a VAD to estimate the speech+noise and noise-only correlation matrices (second-order statistics) in periods of speech+noise and noise-only respectively.

As VADs are so widely used, many algorithms have been proposed in literature (cfr. [48,118] for a summary of some algorithms). In essence, VADs exploit the different properties and features of speech and noise, e.g. short-term energy, zero-crossing rate, speech and noise probability distributions, or combinations of properties. If more than one microphone is available, localization-based VADs (where speech is

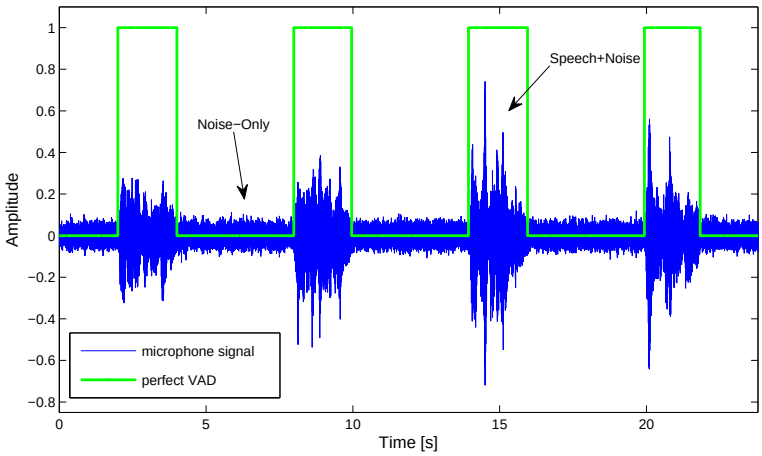


Figure 1.3: Microphone signal and output of perfect VAD.

detected if the estimated location is close to the assumed target speech location) are also possible (e.g. [84, 109]).

Many VAD algorithms are known to have difficulties if the noise is nonstationary, and the performance can be especially unsatisfactory at low input SNR levels [118]. For single-microphone noise reduction, some techniques have therefore been developed to obtain an estimate of the noise spectrum without using a VAD. The best-known techniques are probably the Minimum Statistics [127] and Minima-Controlled Recursive Averaging [27] techniques. Very recently, an extension for multi-microphone noise reduction was also proposed, where the single-microphone noise estimation techniques are extended to obtain an estimate of a noise correlation matrix [83]. These techniques could indeed provide good alternatives to VAD-based estimation, especially so for nonstationary noise, as the noise spectrum or noise correlation matrix can then be continuously updated and thus better tracked, compared to only updating during noise-only periods.

In this thesis, the second-order statistics will be estimated using the classic VAD-based approach (cfr. section 2.1.3 of chapter 2). When studying the binaural SDW-MWF performance, a perfect VAD as in figure 1.3 will usually be assumed in a first step. However, a realistic binaural VAD (proposed in chapter 9) will also be utilized in this work to evaluate the impact of VAD errors on the binaural SDW-MWF performance.

1.5 Outline of the thesis

1.5.1 Main research objectives and schematic overview

Figure 1.4 gives a schematic overview of thesis. The arrows in the figure indicate that a certain technique or result of one chapter is utilized in another chapter. Corresponding chapters are grouped together in larger parts, whereby each part represents a distinct research goal.

As the SDW-MWF will be studied throughout this thesis, a general introduction to the technique will be presented in the following chapter. The binaural noise reduction framework and objective performance measures are also formally defined. Three main research objectives are then identified and presented in Parts II-IV:

- **Part II** (chapters 3-6): This part focuses on the noise reduction performance of the considered binaural SDW-MWF techniques. In a first study, realistic adaptive bilateral and binaural SDW-MWF algorithms are compared in a perceptual evaluation. It is assessed whether a binaural SDW-MWF algorithm indeed allows for significant speech intelligibility improvements compared to bilateral algorithms, which is a main motivation for this thesis work. Theoretical contributions are then presented which aim at further improvements of the noise reduction performance, particularly for real-life situations where estimation errors in the second-order statistics occur and real VAD algorithms have to be utilized. Moreover, a numerically favorable algorithm is derived.
- **Part III** (chapters 7-8): This part focuses on the binaural cue preservation of the considered binaural SDW-MWF techniques. A first theoretical contribution demonstrates how the full-bandwidth binaural SDW-MWF (where all microphone signals can be exchanged) can be extended so that the binaural cues of both the target speech as the residual noise can be preserved. A second theoretical contribution considers reduced-bandwidth bilateral and binaural SDW-MWF techniques (where not all microphone signals can be exchanged). It is demonstrated how reduced-bandwidth techniques can distort the binaural cues. Novel extensions for the reduced-bandwidth techniques are also proposed, which allow for a better binaural cue preservation.
- **Part IV** (chapter 9): As the presented SDW-MWF techniques require a VAD algorithm, this part aims at deriving improved VAD algorithms. The main idea is to utilize the wireless link of the binaural hearing aids, so that a superior binaural VAD can be obtained.

The conclusions of the thesis, as well as suggestions for further research, are finally summarized in **Part V** (chapter 10).

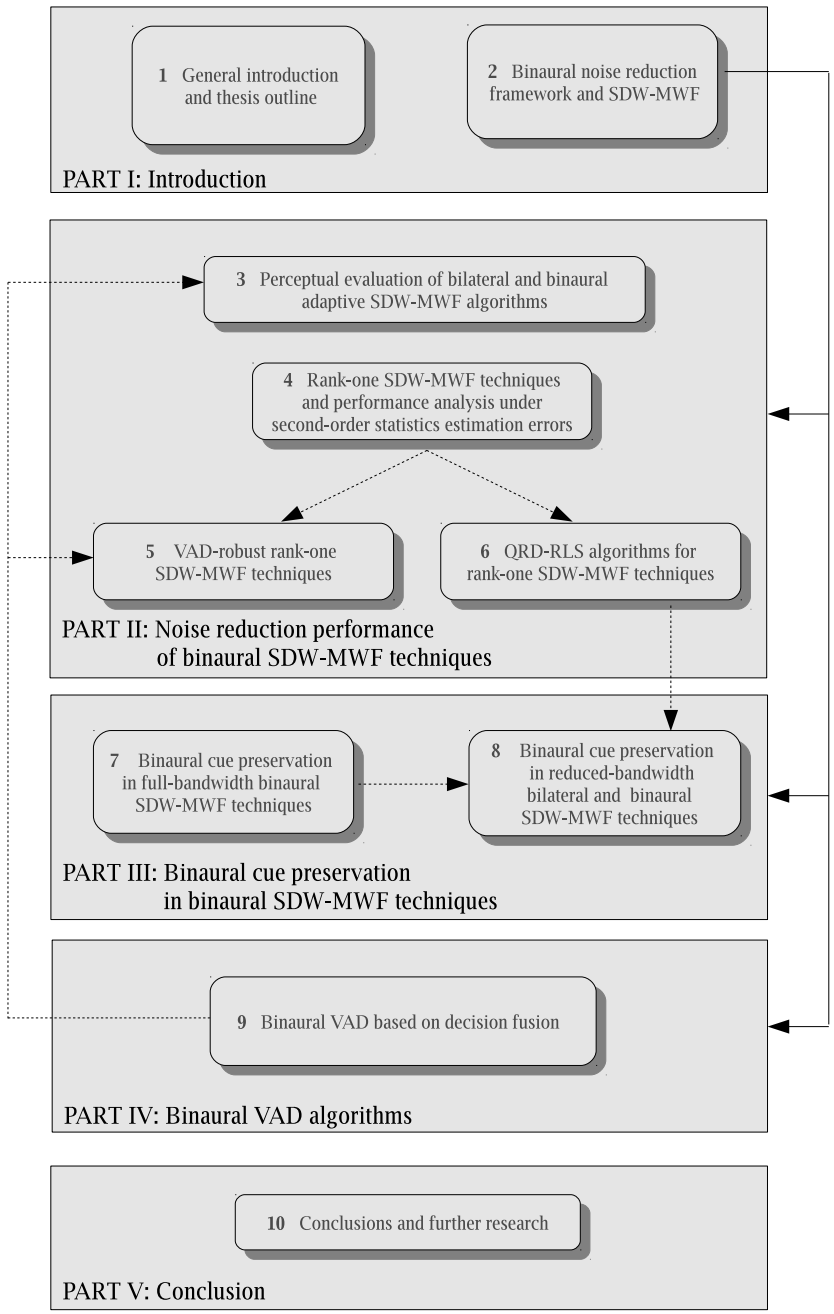


Figure 1.4: Schematic overview of the thesis.

1.5.2 Chapter by chapter overview

Chapter 2 continues the introductory part. The mathematical framework, which describes the binaural noise reduction problem, is presented. Different objective performance measures, which quantify both the noise reduction performance and the binaural cue preservation, are also defined. The binaural SDW-MWF technique, which is adopted throughout the thesis, is then reviewed. Both theoretical aspects and implementational aspects of the SDW-MWF are hereby discussed.

Part II: Noise reduction performance of binaural SDW-MWF techniques

Chapter 3 presents a perceptual speech intelligibility evaluation of adaptive bilateral and binaural SDW-MWF algorithms. It was previously demonstrated that the binaural SDW-MWF allows for significant performance improvements compared to bilateral algorithms, although the evaluation was conducted in idealized conditions. In contrast to the previous work, online adaptive SDW-MWF algorithms were implemented on a real-time platform. A realistic VAD algorithm is also included. 20 normal hearing and 8 hearing impaired subjects are tested in four different spatial scenarios, including a realistic cafeteria scenario with nonstationary noise. The main research question is then whether the binaural SDW-MWF algorithm retains its benefits over a bilateral algorithm, in these realistic conditions. The material presented in this chapter has also been submitted for publication in the Journal of the Acoustical Society of America [38]. A preliminary version of this work was presented as an invited talk at the International hearing aid research conference (IHCON), Lake Tahoe, California, August 2010.

Chapter 4 presents a theoretical contribution where the impact of estimation errors in the second-order statistics is assessed. In theoretical works, it is often assumed that there is a single target speech source. In this case, the SDW-MWF can also be decomposed in a conceptually interesting structure. However, it will be demonstrated that an algorithm based on the standard SDW-MWF formulation does not behave as predicted theoretically, when estimation errors in the (speech) second-order statistics occur. Second, two novel SDW-MWF formulations, namely the R1-MWF and SP-MWF, are also studied. In contrast to the standard SDW-MWF, these formulations are in an explicitly decomposed form. Moreover, algorithms based on these formulations behave close to the theoretical optimum, in the presence of estimation errors. The R1-MWF and SP-MWF thus offer more robust and conceptually interesting alternatives to the standard SDW-MWF formulation. The material presented in this chapter was published in the proceedings of ICASSP [32] and in the IEEE Trans. Audio, Speech, Lang. Process. journal [36].

Chapter 5 focuses on second-order statistics estimation errors which are specifically introduced by an erroneous VAD. Remarkably, the novel R1-MWF and SP-MWF show an inherent robustness to VAD errors, provided the noise is stationary. It is also investigated whether the performance can be improved for (spectrally) nonstationary noise. The material presented in this chapter was published in the proceedings of ICASSP [39].

Chapter 6 exploits the decomposed structures of the R1-MWF and SP-MWF to derive numerically favorable algorithms based on QR decomposition Recursive Least Squares (QRD-RLS). It will be demonstrated that the new algorithms outperform an algorithm based on the the standard SDW-MWF expression. Moreover, it is assessed whether the wordlength can be reduced without affecting the noise reduction performance. The material presented in this chapter was published in the proceedings of EUSIPCO [34].

Part III: Binaural cue preservation by binaural SDW-MWF techniques

Chapter 7 focuses on the binaural cue preservation of full-bandwidth SDW-MWF techniques, i.e. where all microphone signals can be exchanged between the devices. From previous work it was known that the full-bandwidth binaural SDW-MWF preserves the binaural cues of the target speech, but distorts the binaural cues of the residual noise. Two extensions were therefore proposed, namely the MWF-ITF and MWF-η. Previous perceptual validations indicated that these extensions indeed allow for a better preservation of the residual noise cues, but a theoretical analysis and comparison was not performed. This chapter therefore presents a theoretical analysis of the binaural SDW-MWF and its two extensions. The MWF-ITF and MWF-η are also directly compared by physical performance measures. The material presented in this chapter was published in the proceedings of EUSIPCO [30] and in the IEEE Trans. Audio, Speech, Lang. Process. journal [31].

Chapter 8 considers the binaural cue preservation of reduced-bandwidth bilateral and binaural SDW-MWF techniques, i.e. where not every microphone signal can be exchanged between the devices. It is then demonstrated that the SDW-MWF does not necessarily preserve the speech binaural cues: although the speech ITD cues are still preserved (even for the bilateral SDW-MWF), the speech ILD cues are distorted. Novel reduced-bandwidth bilateral and binaural SDW-MWF techniques are then proposed, which aim at better preserving the speech ILD cues. The novel techniques exploit the decomposed structures of the R1-MWF and SP-MWF. Namely, by using a common spectral postfilter in these approaches, ILD distortion can be eliminated. Experiments in a reverberant environment confirm the superior performances of the reduced-bandwidth SDW-MWF techniques with common postfilter. Part of the material presented in this chapter was published in the proceedings of IWAENC [33]. A journal paper version has been submitted

to IEEE Trans. Audio, Speech, Lang. Process. [37].

Part IV: Binaural VAD algorithms

Chapter 9 proposes two novel binaural VADs. As many noise reducion techniques (including the SDW-MWF) rely on a VAD, a robust VAD algorithm is highly sought for. The main idea in this work is to utilize the wireless link of the binaural hearing aids to combine local VAD decisions into a superior fused VAD decision. The local VADs are energy-based VADs, calculated in each device separately, and a two-microphone cross-correlation based VAD which uses microphone signals from both devices. The local VAD decisions are then combined using a fusion rule based on local SNR estimates, or by an optimal AND-rule. The performances of the VADs are assessed in terms of receiver operating characteristics and by evaluating the impact on the SDW-MWF performance. The material presented in this chapter was published in the proceedings of EUSIPCO [35].

Part V: Conclusions and further research

Chapter 10 concludes the thesis. First, conclusions are formulated for each of the identified research goals. Second, some suggestions for further research directions are presented.

Chapter 2

Binaural noise reduction framework and SDW-MWF

This chapter describes the binaural noise reduction framework which will be used in the remainder of the thesis. The binaural Speech-Distortion-Weighted Multi-channel Wiener Filter (SDW-MWF), which is the binaural multi-microphone noise reduction algorithm adopted in this thesis, is also reviewed.

Section 2.1 presents the configuration and notation used in the binaural noise reduction problem. Performance measures which quantify both the noise reduction performance and the binaural cue preservation are also defined.

Section 2.2 presents a review of the binaural SDW-MWF algorithm. An overview of previously proposed reduced-bandwidth binaural algorithms, i.e. where not every microphone signal is exchanged between the two devices, is also given. Finally, a practical implementation of the binaural SDW-MWF, which will be used as reference implementation, is discussed.

Section 2.3 concludes the chapter.

2.1 Configuration and notation

2.1.1 Speech, noise and microphone signals

A speech-in-noise scenario as in Figure 2.1 is considered. Each hearing aid microphone m observes a signal $Y_m(\omega)$, which consists of a target speech component $X_m(\omega)$ and an (unwanted) noise component $V_m(\omega)$. The observed

mth microphone signal in the hearing aid $Y_m(\omega)$ can thus be specified as

$$Y_m(\omega) = X_m(\omega) + V_m(\omega), \quad m = 0 \dots M-1 \,, \tag{2.1}$$

where ω is the frequency-domain variable and M is the total number of microphones in the hearing aid.

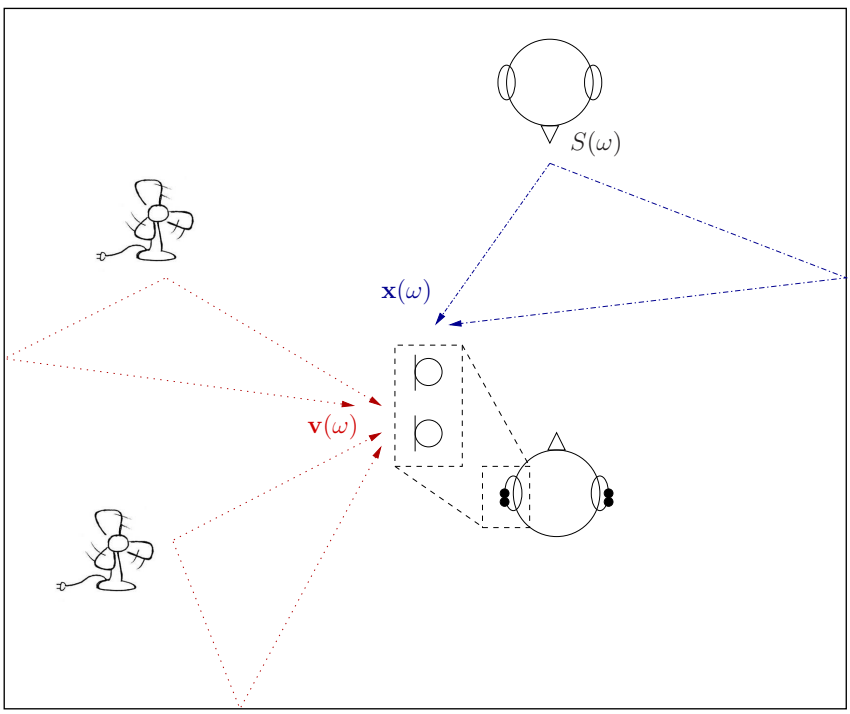


Figure 2.1: Signals impinging on the multi-microphone array of a hearing aid, in a speech-in-noise scenario.

As in Figure 2.1, the target speech component can be a filtered version of a speech signal $S(\omega)$ produced by a target speaker, so that

$$X_m(\omega) = A_m(\omega)S(\omega) \,. \tag{2.2}$$

$A_m(\omega)$ hereby contains the complete acoustic transfer function from the speech source to microphone m (including reflections against walls and objects, microphone characteristics and head shadow effect). However, a more general case will first be considered in the following, where $X_m(\omega)$ can be a superposition of multiple target speech sources which are all of interest to the hearing aid user.

The noise component $V_m(\omega)$ contains all unwanted signals which should be suppressed by the noise reduction. The noise can be a superposition of different localized noise sources (e.g. the two fans in Figure 2.1), but can also include diffuse noise (noise coming from all directions, e.g. in a car environment) or spatially uncorrelated noise (e.g. thermal sensor noise). The noise is not necessarily non-speech, i.e. undesired speech sources (e.g. in a cafeteria scenario) can also be included in the noise component.

All available microphone signals in the hearing aid can be stacked in the M-dimensional signal vector $\mathbf{y}(\omega)$. As in (2.1), the signal vector can also be decomposed into a speech component vector and noise component vector, i.e.[1]

$$\mathbf{y}(\omega) = \mathbf{x}(\omega) + \mathbf{v}(\omega) \, . \tag{2.3}$$

For conciseness, the frequency-domain variable ω is omitted from now on.

2.1.2 Binaural processing scheme

A binaural processing scheme is now considered, where the hearing aid user wears a left and a right device which are able to exchange one or more microphone signals over a wireless link. The scheme proposed in [33, 37, 55] is depicted in Figure 2.2. Both hearing aids have a microphone array consisting of M microphones[2]. As in (2.1), the mth microphone signal in the left hearing aid $Y_{0,m}$ can be specified as[3]

$$Y_{0,m} = X_{0,m} + V_{0,m}, \quad m = 0 \ldots M-1, \tag{2.4}$$

where $X_{0,m}$ represents the speech component and $V_{0,m}$ represents the noise component. Similarly, the mth microphone signal in the right hearing aid is equal to $Y_{1,m} = X_{1,m} + V_{1,m}$. We define the M-dimensional stacked microphone signal vectors $\mathbf{y}_0$ and $\mathbf{y}_1$ and the $2M$-dimensional signal vector $\mathbf{y}$ as

$$\mathbf{y}_0 = \begin{bmatrix} Y_{0,0} \\ \vdots \\ Y_{0,M-1} \end{bmatrix}, \quad \mathbf{y}_1 = \begin{bmatrix} Y_{1,0} \\ \vdots \\ Y_{1,M-1} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_0 \\ \mathbf{y}_1 \end{bmatrix} . \tag{2.5}$$

As in (2.3), the signal vector $\mathbf{y}$ can be written as $\mathbf{y} = \mathbf{x} + \mathbf{v}$, where $\mathbf{x}$ and $\mathbf{v}$ are defined similarly as $\mathbf{y}$.

[1] Although the signal vectors contain complex-valued frequency-domain variables, they are denoted with lower-case letters throughout the thesis to distinguish them from matrices.

[2] It is possible to use different array sizes as in [50], but here both arrays are assumed to have M microphones for the sake of simplicity.

[3] An index 0 is now added to explicitly refer to the microphones of the left device. Similarly, an index 1 is added to refer to the microphones of the right device.

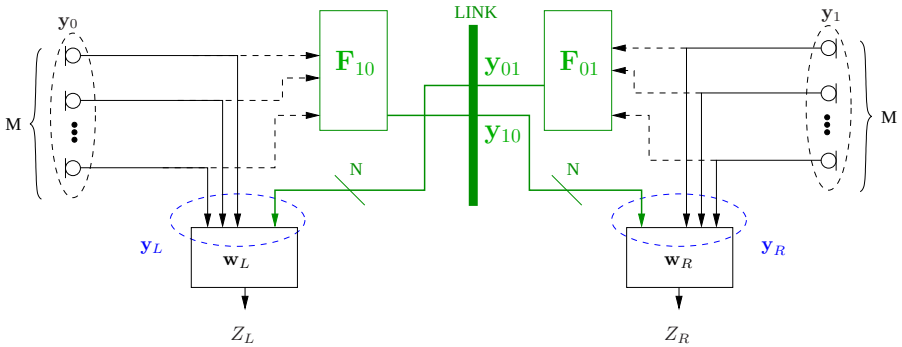


Figure 2.2: General binaural processing scheme

In the binaural processing scheme, signals are exchanged between the two devices over a wireless link. However, due to bandwidth limitations, it may not be feasible to exchange every microphone signal, so that each device does not have access to the full signal vector $\mathbf{y}$. In general, it is assumed that each device receives linear combinations of the contralateral microphone signals, i.e.

$$\mathbf{y}_{10} = \mathbf{F}_{10}^H \, \mathbf{y}_0 \, , \quad \mathbf{y}_{01} = \mathbf{F}_{01}^H \, \mathbf{y}_1 \, , \tag{2.6}$$

where $\{.\}^H$ denotes the Hermitian transpose operator and where $\mathbf{F}_{10}$ and $\mathbf{F}_{01}$ are $M \times N$ dimensional complex matrices with $1 \leq N < M$. The available signals at the left and right device[4] can then be defined as:

$$\mathbf{y}_L \;=\; \begin{bmatrix} \mathbf{y}_0 \\ \mathbf{y}_{01} \end{bmatrix} = \left[\begin{array}{c:c} \mathbf{I}_M & \mathbf{0}_{M\times M} \\ \hdashline \mathbf{0}_{N\times M} & \mathbf{F}_{01}^H \end{array}\right] \mathbf{y} = \mathbf{C}_L^H \mathbf{y} \, , \tag{2.7}$$

$$\mathbf{y}_R \;=\; \begin{bmatrix} \mathbf{y}_{10} \\ \mathbf{y}_1 \end{bmatrix} = \left[\begin{array}{c:c} \mathbf{F}_{10}^H & \mathbf{0}_{N\times M} \\ \hdashline \mathbf{0}_{M\times M} & \mathbf{I}_M \end{array}\right] \mathbf{y} = \mathbf{C}_R^H \mathbf{y} \, , \tag{2.8}$$

where $\mathbf{I}_M$ is the $M \times M$ identity matrix and where $\mathbf{0}_{M\times M}$ and $\mathbf{0}_{N\times M}$ are matrices whose entries are all equal to 0. $\mathbf{C}_L$ and $\mathbf{C}_R$ are $2M \times (M+N)$ matrices which compress the signal vector $\mathbf{y}$ into the lower-dimensional $\mathbf{y}_L$ and $\mathbf{y}_R$. The speech components $\mathbf{x}_L$ and $\mathbf{x}_R$ and the noise components $\mathbf{v}_L$ and $\mathbf{v}_R$ are similarly defined.

[4] In contrast to [31, 55], a distinction is made here between the microphone signal vectors $\mathbf{y}_0$ and $\mathbf{y}_1$, and the signal vectors $\mathbf{y}_L$ and $\mathbf{y}_R$ which represent the available signals at the left and right device.

For the special case of a bilateral setup (i.e. no signals are exchanged), the same framework can be used, where then $N = 0$ and so $\mathbf{C}_L$ and $\mathbf{C}_R$ are then $2M \times M$ matrices, defined as:

$$\mathbf{C}_L^H = \left[\begin{array}{c:c} \mathbf{I}_M & \mathbf{0}_{M\times M} \end{array} \right] , \tag{2.9}$$

$$\mathbf{C}_R^H = \left[\begin{array}{c:c} \mathbf{0}_{M\times M} & \mathbf{I}_M \end{array} \right] . \tag{2.10}$$

The full-bandwidth case, i.e. where all microphone signals can be exchanged between the devices, corresponds to the case $N = M$ and $\mathbf{F}_{10} = \mathbf{F}_{01} = \mathbf{I}_M$, so that $\mathbf{y}_L = \mathbf{y}_R = \mathbf{y}$.

The r_L-th signal of the left device and the r_R-th signal of the right device will be used as the so-called reference signals for the noise reduction algorithms. The reference signals at the left and the right device are denoted as $Y_{L,\text{ref}}$ and $Y_{R,\text{ref}}$, which are then equal to

$$Y_{L,\text{ref}} = \mathbf{e}_L^H \mathbf{y}_L, \quad Y_{R,\text{ref}} = \mathbf{e}_R^H \mathbf{y}_R , \tag{2.11}$$

where $\mathbf{e}_L$ and $\mathbf{e}_R$ are $(M+N)$-dimensional vectors with only one element equal to 1 and the other elements equal to 0, i.e. $\mathbf{e}_L(r_L) = 1$ and $\mathbf{e}_R(N + r_R) = 1$. Typically, the front microphone signals are used as reference signals, which corresponds to $r_L = r_R = 0$ (assuming zero-based indexing). The reference signals can also be written as $Y_{L,\text{ref}} = X_{L,\text{ref}} + V_{L,\text{ref}}$ and $Y_{R,\text{ref}} = X_{R,\text{ref}} + V_{R,\text{ref}}$.

The output signals Z_L and Z_R at the left and the right device are obtained by filtering and summing the left and right signal vectors, i.e.

$$Z_L = \mathbf{w}_L^H \mathbf{y}_L, \quad Z_R = \mathbf{w}_R^H \mathbf{y}_R , \tag{2.12}$$

where $\mathbf{w}_L$ and $\mathbf{w}_R$ are $(M + N)$-dimensional complex weight vectors. The output signal at the left hearing aid can also be written as

$$Z_L = Z_{xL} + Z_{vL} = \mathbf{w}_L^H \mathbf{x}_L + \mathbf{w}_L^H \mathbf{v}_L , \tag{2.13}$$

where Z_{xL} represents the speech component and Z_{vL} represents the noise component of the output signal. Similarly, the output signal at the right hearing aid can be written as $Z_R = Z_{xR} + Z_{vR} = \mathbf{w}_R^H \mathbf{x}_R + \mathbf{w}_R^H \mathbf{v}_R$.

2.1.3 Second-order statistics estimation

The noise reduction techniques presented in section 2.2 make use of the speech and noise second-order statistics (SOS) or correlation matrices. The $2M \times 2M$ correlation matrices, which contain the autocorrelation and crosscorrelation values of $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{v}$ over the different microphones, are defined as

$$\mathbf{R}_y = \mathcal{E}\{\mathbf{y}\mathbf{y}^H\}, \quad \mathbf{R}_x = \mathcal{E}\{\mathbf{x}\mathbf{x}^H\}, \quad \mathbf{R}_v = \mathcal{E}\{\mathbf{v}\mathbf{v}^H\} , \tag{2.14}$$

where $\mathcal{E}\{.\}$ denotes the expected value operator. Assuming that the speech and the noise components are uncorrelated, $\mathbf{R}_y = \mathbf{R}_x + \mathbf{R}_v$. By using definitions (2.7) and (2.8), the $(M+N)\times(M+N)$ left and right correlation matrices (i.e. the correlation matrices estimated at the left and right device) can be defined as

$$\mathbf{R}_{yL} = \mathcal{E}\{\mathbf{y}_L\mathbf{y}_L^H\} = \mathbf{C}_L^H\mathbf{R}_y\mathbf{C}_L, \quad \mathbf{R}_{yR} = \mathcal{E}\{\mathbf{y}_R\mathbf{y}_R^H\} = \mathbf{C}_R^H\mathbf{R}_y\mathbf{C}_R\ . \tag{2.15}$$

The speech correlation matrices $\mathbf{R}_{xL}$ and $\mathbf{R}_{xR}$ and the noise correlation matrices $\mathbf{R}_{vL}$ and $\mathbf{R}_{vR}$ can be similarly defined.

It is assumed that the noise is continuously active, while there are pauses in the target speech signal, during which only the noise is present. If a Voice Activity Detection (VAD) algorithm is available, which classifies frames as noise-only or speech+noise frames, the estimates $\hat{\mathbf{R}}_{vL}$, $\hat{\mathbf{R}}_{vR}$ and $\hat{\mathbf{R}}_{yL}$, $\hat{\mathbf{R}}_{yR}$ can thus be calculated by a recursive updating procedure, i.e.

- In noise-only frames:

$$\begin{aligned}\hat{\mathbf{R}}_{yL}[k] &= \hat{\mathbf{R}}_{yL}[k-1]\ ,\\ \hat{\mathbf{R}}_{vL}[k] &= \lambda_v\hat{\mathbf{R}}_{vL}[k-1] + (1-\lambda_v)\mathbf{v}_L[k]\mathbf{v}_L^H[k]\ ,\end{aligned} \tag{2.16}$$

- In speech+noise frames:

$$\begin{aligned}\hat{\mathbf{R}}_{yL}[k] &= \lambda_y\hat{\mathbf{R}}_{yL}[k-1] + (1-\lambda_y)\mathbf{y}_L[k]\mathbf{y}_L^H[k]\ ,\\ \hat{\mathbf{R}}_{vL}[k] &= \hat{\mathbf{R}}_{vL}[k-1]\ ,\end{aligned} \tag{2.17}$$

for the left device and similarly for the right device. k is the frame-index, λ_y and λ_v are exponential forgetting factors which are typically chosen close to one[5]. The speech correlation matrix estimates $\hat{\mathbf{R}}_{xL}$ and $\hat{\mathbf{R}}_{xR}$ cannot be directly computed because by definition, the speech component vectors $\mathbf{x}_L$ and $\mathbf{x}_R$ are unknown. However, assuming the speech and noise are uncorrelated and that the noise is sufficiently stationary, the speech correlation matrix estimates can be obtained as:

$$\hat{\mathbf{R}}_{xL} = \hat{\mathbf{R}}_{yL} - \hat{\mathbf{R}}_{vL}, \quad \hat{\mathbf{R}}_{xR} = \hat{\mathbf{R}}_{yR} - \hat{\mathbf{R}}_{vR}\ . \tag{2.18}$$

2.1.4 Special case: single target speech source

A frequently occuring scenario is the case where there is only a single target speech source S, as in Figure 2.1. Each microphone observes a filtered version of S as in (2.2), so that the speech signal vector can be modelled as

$$\mathbf{x} = \mathbf{a}S\ , \tag{2.19}$$

[5] An averaging time window of 2-3 seconds is typically required to achieve reliable estimates.

where the $2M$-dimensional steering vector $\mathbf{a}$ contains the acoustic transfer functions from the speech source to the microphones (including room acoustics, microphone characteristics and head shadow effect). The vector $\mathbf{a}$ is defined similarly as $\mathbf{y}$ in (2.5), i.e.

$$\mathbf{a}_0 = \begin{bmatrix} A_{0,0} \\ \vdots \\ A_{0,M-1} \end{bmatrix}, \quad \mathbf{a}_1 = \begin{bmatrix} A_{1,0} \\ \vdots \\ A_{1,M-1} \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} \mathbf{a}_0 \\ \mathbf{a}_1 \end{bmatrix}. \tag{2.20}$$

As in (2.7) and (2.8), the vectors $\mathbf{a}_L$ and $\mathbf{a}_R$ can be defined as:

$$\mathbf{a}_L = \mathbf{C}_L^H \mathbf{a}\ , \quad \mathbf{a}_R = \mathbf{C}_R^H \mathbf{a}\ . \tag{2.21}$$

The reference microphone signals at the left and the right hearing aid can be written as

$$Y_{L,\text{ref}} = A_{L,\text{ref}} S + V_{L,\text{ref}}, \quad Y_{R,\text{ref}} = A_{R,\text{ref}} S + V_{R,\text{ref}}\ , \tag{2.22}$$

with $A_{L,\text{ref}}$ the reference element of $\mathbf{a}_L$ (i.e. $A_{L,\text{ref}} = \mathbf{e}_L^H \mathbf{a}_L$) and $A_{R,\text{ref}}$ the reference element of $\mathbf{a}_R$ (i.e. $A_{R,\text{ref}} = \mathbf{e}_R^H \mathbf{a}_R$).

With assumption (2.19), the speech correlation matrix is a rank-1 matrix, i.e.

$$\mathbf{R}_x = P_s \mathbf{a}\mathbf{a}^H\ , \tag{2.23}$$

with $P_s = \mathcal{E}\{|S|^2\}$ the power of the speech signal. The speech correlation matrices $\mathbf{R}_{xL}$ and $\mathbf{R}_{xR}$ (i. e. the correlation matrices estimated at the left and right devices) can be similarly written as $\mathbf{R}_{xL} = P_s \mathbf{a}_L \mathbf{a}_L^H$ and $\mathbf{R}_{xR} = P_s \mathbf{a}_R \mathbf{a}_R^H$.

2.1.5 Noise reduction performance measures

In the theoretical analysis of the noise reduction techniques, performance measures based on the second-order statistics will be used in the thesis.

The *narrowband*[6] *input SNR* is defined as the power ratio of the speech and noise component in the reference microphone signal, i.e. for the left device:

$$\text{SNR}_L^{in} = \frac{\mathcal{E}\{|X_{L,\text{ref}}|^2\}}{\mathcal{E}\{|V_{L,\text{ref}}|^2\}} = \frac{\mathbf{e}_L^H \mathbf{R}_{xL} \mathbf{e}_L}{\mathbf{e}_L^H \mathbf{R}_{vL} \mathbf{e}_L}\ , \tag{2.24}$$

and similarly for the right device. The *narrowband output SNR* is defined as the power ratio of the speech and noise component in the output signal, i.e. for the

[6]It should be noted that the terms *narrowband SNR* and *broadband SNR* indicate that the SNR is calculated in either a single frequency bin or over the complete frequency range, and that they are not related to the bandwidth of the wireless link. In other publications, these terms are sometimes referred to as *local SNR* and *global SNR*, e.g. [160].

left device:

$$\mathrm{SNR}_L^{out} = \frac{\mathcal{E}\{|Z_{xL}|^2\}}{\mathcal{E}\{|Z_{vL}|^2\}} = \frac{\mathbf{w}_L^H \mathbf{R}_{xL} \mathbf{w}_L}{\mathbf{w}_L^H \mathbf{R}_{vL} \mathbf{w}_L} \,, \tag{2.25}$$

and similarly for the right device. The *narrowband SNR improvements* at the left and the right device are then defined as

$$\Delta\mathrm{SNR}_L = \frac{\mathrm{SNR}_L^{out}}{\mathrm{SNR}_L^{in}}, \quad \Delta\mathrm{SNR}_R = \frac{\mathrm{SNR}_R^{out}}{\mathrm{SNR}_R^{in}} \,. \tag{2.26}$$

To assess the practical performance of the noise reduction techniques, SNR and distortion measures are calculated on the input and output time-domain signals. The *broadband input SNR* (in dB) at the left device is defined as

$$\mathrm{SNR}_{\mathrm{broad},L}^{in} = 10 \log_{10}\left(\frac{\sum x_{L,\mathrm{ref}}^2[t]}{\sum v_{L,\mathrm{ref}}^2[t]}\right) , \tag{2.27}$$

where t is the discrete time-domain variable, and where $x_{L,\mathrm{ref}}$ and $v_{L,\mathrm{ref}}$ are the time-domain speech and noise component signals in the left reference microphone signal. The summations hereby only count the samples where speech and noise are both active. The *broadband output SNR* (in dB) obtained at the left device is defined as

$$\mathrm{SNR}_{\mathrm{broad},L}^{out} = 10 \log_{10}\left(\frac{\sum z_{xL}^2[t]}{\sum z_{vL}^2[t]}\right) , \tag{2.28}$$

where z_{xL} and z_{vL} are the time-domain speech and noise component signals in the left output signal. The *broadband SNR improvement* (in dB) at the left device is

$$\Delta\mathrm{SNR}_{\mathrm{broad},L} = \mathrm{SNR}_{\mathrm{broad},L}^{out} - \mathrm{SNR}_{\mathrm{broad},L}^{in} \,. \tag{2.29}$$

The broadband input SNR, output SNR and SNR improvement for the right device are similarly defined.

The speech intelligibility (SI) weighted SNR improvement, which correlates well with perceptual speech intelligibility improvements, will also be used in the thesis. As for the broadband SNR, the (time-domain) speech and noise components of the output signals and microphone signals are treated separately. The signals are filtered by one-third octave band bandpass filters, and the SNR is calculated per band. The *SI-weighted SNR improvement* (in dB) [74, 164] for the left device is then defined as:

$$\Delta\mathrm{SNR}_{\mathrm{SI},L} = \sum_i I_i \, (\mathrm{SNR}_{i,L}^{out} - \mathrm{SNR}_{i,L}^{in}) \,, \tag{2.30}$$

where the band importance function I_i expresses the importance of the ith one-third octave band with center frequency f_i^c for intelligibility, and where $\mathrm{SNR}_{i,L}^{out}$ and $\mathrm{SNR}_{i,L}^{in}$ are the output SNR and input SNR (in dB) in this band. The center frequencies f_i^c and the values I_i are defined in [5]. Again, the SI-weighted SNR

improvement for the right device is obtained similarly.
An *SI-weighted distortion* measure was proposed in [164], i.e. for the left device (and similarly for the right device):

$$\mathrm{SD}_{\mathrm{SI},L} = \sum_i I_i \ \mathrm{SD}_{i,L} \ , \tag{2.31}$$

where $\mathrm{SD}_{i,L}$ is the average spectral distortion in the ith one-third octave band, calculated as

$$\mathrm{SD}_{i,L} = \int_{2^{-1/6} f_i^c}^{2^{1/6} f_i^c} |10 \ \log_{10} \ G_L^s(f)| \, df \ , \tag{2.32}$$

with $G_L^s(f)$ the power transfer function of the speech component from the input to the output of the noise reduction. A distortion value of 0 dB corresponds to an undistorted signal, while larger distortion values correspond to more introduced speech distortion.

2.1.6 Binaural cue preservation measures

In this section, theoretical measures are defined by which the preservation of the binaural cues (ITD and ILD) by the noise reduction techniques can be quantified. As for section 2.1.5, the measures are based on the speech and noise second-order statistics.

The input and output *interaural transfer function (ITF)* of the speech and noise components are defined as the ratio of the components at the left and right hearing aid, i.e.

$$\mathrm{ITF}_x^{in} = \frac{X_{L,\mathrm{ref}}}{X_{R,\mathrm{ref}}} = \frac{\mathbf{e}_L^H \mathbf{x}_L}{\mathbf{e}_R^H \mathbf{x}_R}, \ \ \mathrm{ITF}_x^{out} = \frac{Z_{xL}}{Z_{xR}} = \frac{\mathbf{w}_L^H \mathbf{x}_L}{\mathbf{w}_R^H \mathbf{x}_R} \ , \tag{2.33}$$

$$\mathrm{ITF}_v^{in} = \frac{V_{L,\mathrm{ref}}}{V_{R,\mathrm{ref}}} = \frac{\mathbf{e}_L^H \mathbf{v}_L}{\mathbf{e}_R^H \mathbf{v}_R}, \ \ \mathrm{ITF}_v^{out} = \frac{Z_{vL}}{Z_{vR}} = \frac{\mathbf{w}_L^H \mathbf{v}_L}{\mathbf{w}_R^H \mathbf{v}_R} \ . \tag{2.34}$$

In the case of a single target speech source (section 2.1.4), the ITF_x^{in} and ITF_x^{out} are independent of the actual input signal S, namely

$$\mathrm{ITF}_x^{in} = \frac{\mathbf{e}_L^H \mathbf{a}_L}{\mathbf{e}_R^H \mathbf{a}_R} = \frac{A_{L,\mathrm{ref}}}{A_{R,\mathrm{ref}}}, \quad \mathrm{ITF}_x^{out} = \frac{\mathbf{w}_L^H \mathbf{a}_L}{\mathbf{w}_R^H \mathbf{a}_R} \ . \tag{2.35}$$

Hence it is also possible to use alternative formulas for these ITFs, namely

$$\mathrm{ITF}_x^{in} = \frac{\mathbf{e}_L^H \mathbf{R}_{xLR} \mathbf{e}_R}{\mathbf{e}_R^H \mathbf{R}_{xR} \mathbf{e}_R} = \frac{\mathbf{e}_L^H \mathbf{R}_{xL} \mathbf{e}_L}{\mathbf{e}_R^H \mathbf{R}_{xRL} \mathbf{e}_L} \ , \tag{2.36}$$

$$\mathrm{ITF}_x^{out} = \frac{\mathbf{w}_L^H \mathbf{R}_{xLR} \mathbf{w}_R}{\mathbf{w}_R^H \mathbf{R}_{xR} \mathbf{w}_R} = \frac{\mathbf{w}_L^H \mathbf{R}_{xL} \mathbf{w}_L}{\mathbf{w}_R^H \mathbf{R}_{xRL} \mathbf{w}_L} \ , \tag{2.37}$$

where $\mathbf{R}_{xLR} = \mathbf{C}_L^H \mathbf{R}_x \mathbf{C}_R$ and $\mathbf{R}_{xRL} = \mathbf{R}_{xLR}^H = \mathbf{C}_R^H \mathbf{R}_x \mathbf{C}_L$.

These input and output ITFs are complex valued scalars, of which the amplitude and phase can be defined as the (square root of the) *interaural level differences (ILDs)* and *interaural time differences (ITDs)*, namely

$$\mathrm{ILD}_x^{in} = \left[\left(\frac{\mathbf{e}_L^H \mathbf{R}_{xLR} \mathbf{e}_R}{\mathbf{e}_R^H \mathbf{R}_{xR} \mathbf{e}_R} \right) \left(\frac{\mathbf{e}_L^H \mathbf{R}_{xL} \mathbf{e}_L}{\mathbf{e}_R^H \mathbf{R}_{xRL} \mathbf{e}_L} \right)^H \right] = \frac{\mathbf{e}_L^H \mathbf{R}_{xL} \mathbf{e}_L}{\mathbf{e}_R^H \mathbf{R}_{xR} \mathbf{e}_R} \ , \tag{2.38}$$

$$\mathrm{ILD}_x^{out} = \frac{\mathbf{w}_L^H \mathbf{R}_{xL} \mathbf{w}_L}{\mathbf{w}_R^H \mathbf{R}_{xR} \mathbf{w}_R} \ , \tag{2.39}$$

and

$$\mathrm{ITD}_x^{in} = \angle \left(\frac{\mathbf{e}_L^H \mathbf{R}_{xLR} \mathbf{e}_R}{\mathbf{e}_R^H \mathbf{R}_{xR} \mathbf{e}_R} \right) = \angle \left(\mathbf{e}_L^H \mathbf{R}_{xLR} \mathbf{e}_R \right) \ , \tag{2.40}$$

$$\mathrm{ITD}_x^{out} = \angle \left(\mathbf{w}_L^H \mathbf{R}_{xLR} \mathbf{w}_R \right) \ . \tag{2.41}$$

Formulas (2.36)-(2.41) will be used as ITF, ILD and ITD definitions, also in the more general case with more than one speech source. The input and output ITF, ILD and ITD for the noise component are then also similarly defined:

$$\mathrm{ITF}_v^{in} = \frac{\mathbf{e}_L^H \mathbf{R}_{vLR} \mathbf{e}_R}{\mathbf{e}_R^H \mathbf{R}_{vR} \mathbf{e}_R} = \frac{\mathbf{e}_L^H \mathbf{R}_{vL} \mathbf{e}_L}{\mathbf{e}_R^H \mathbf{R}_{vRL} \mathbf{e}_L} \ , \tag{2.42}$$

$$\mathrm{ITF}_v^{out} = \frac{\mathbf{w}_L^H \mathbf{R}_{vLR} \mathbf{w}_R}{\mathbf{w}_R^H \mathbf{R}_{vR} \mathbf{w}_R} = \frac{\mathbf{w}_L^H \mathbf{R}_{vL} \mathbf{w}_L}{\mathbf{w}_R^H \mathbf{R}_{vRL} \mathbf{w}_L} \ , \tag{2.43}$$

$$\mathrm{ILD}_v^{in} = \frac{\mathbf{e}_L^H \mathbf{R}_{vL} \mathbf{e}_L}{\mathbf{e}_R^H \mathbf{R}_{vR} \mathbf{e}_R} \ , \tag{2.44}$$

$$\mathrm{ILD}_v^{out} = \frac{\mathbf{w}_L^H \mathbf{R}_{vL} \mathbf{w}_L}{\mathbf{w}_R^H \mathbf{R}_{vR} \mathbf{w}_R} \ , \tag{2.45}$$

$$\mathrm{ITD}_v^{in} = \angle \left(\mathbf{e}_L^H \mathbf{R}_{vLR} \mathbf{e}_R \right) \ , \tag{2.46}$$

$$\mathrm{ITD}_v^{out} = \angle \left(\mathbf{w}_L^H \mathbf{R}_{vLR} \mathbf{w}_R \right) \ , \tag{2.47}$$

where $\mathbf{R}_{vLR} = \mathbf{C}_L^H \mathbf{R}_v \mathbf{C}_R$ and $\mathbf{R}_{vRL} = \mathbf{R}_{vLR}^H = \mathbf{C}_R^H \mathbf{R}_v \mathbf{C}_L$.

The *ILD errors* are calculated as

$$\Delta\mathrm{ILD}_x = 10 \log_{10}(\mathrm{ILD}_x^{out}) - 10 \log_{10}(\mathrm{ILD}_x^{in}), \tag{2.48}$$

$$\Delta\mathrm{ILD}_v = 10 \log_{10}(\mathrm{ILD}_v^{out}) - 10 \log_{10}(\mathrm{ILD}_v^{in}) \ . \tag{2.49}$$

The *ITD errors* are calculated as

$$\Delta\text{ITD}_x = \frac{|\text{ITD}_x^{out} - \text{ITD}_x^{in}|}{\pi}, \tag{2.50}$$

$$\Delta\text{ITD}_v = \frac{|\text{ITD}_v^{out} - \text{ITD}_v^{in}|}{\pi}\,. \tag{2.51}$$

By this definition, ΔITD_x and ΔITD_v are relative errors which always lie between 0 (no error) and 1 (maximum error, phase shift of π).

2.2 Binaural SDW-MWF

2.2.1 General cost function and optimal filters

The binaural Multi-channel Wiener Filter (MWF) produces a minimum-mean-squared-error (MMSE) estimate of the speech component in the reference microphone of each device, hence simultaneously reducing noise and limiting speech distortion [101]. To provide a more explicit trade-off between speech distortion and noise reduction, the Speech-Distortion-Weighted Multi-channel Wiener Filter (SDW-MWF) has been proposed, which minimizes a weighted sum of the mean-squared residual noise and speech distortion [52, 54]. The binaural SDW-MWF[7] cost function for the filter $\mathbf{w}_L$ estimating the speech component $X_{L,\text{ref}}$ in the reference microphone of the left device is equal to

$$J_{\text{MWF}}(\mathbf{w}_L) = \mathcal{E}\left\{\left|X_{L,\text{ref}} - \mathbf{w}_L^H\mathbf{x}_L\right|^2 + \mu\left|\mathbf{w}_L^H\mathbf{v}_L\right|^2\right\}\,, \tag{2.52}$$

where the speech distortion parameter μ provides a trade-off between noise reduction and speech distortion. The SDW-MWF cost function for the right filter $\mathbf{w}_R$ can be similarly defined. The optimal SDW-MWF filters for the left and right device, which minimize the cost function, are then equal to (cfr. Appendix A.1):

$$\boxed{\begin{array}{rcl}\mathbf{w}_L &=& (\mathbf{R}_{xL} + \mu\mathbf{R}_{vL})^{-1}\mathbf{R}_{xL}\mathbf{e}_L \\ \mathbf{w}_R &=& (\mathbf{R}_{xR} + \mu\mathbf{R}_{vR})^{-1}\mathbf{R}_{xR}\mathbf{e}_R\end{array}} \tag{2.53}$$

The SDW-MWF is a good choice for binaural noise reduction: if it is assumed that all microphone signals can be exchanged between the devices and that there is a single target speech source, the SDW-MWF preserves the speech binaural cues (cfr. [31, 50] and chapter 7) in addition to obtaining a noise reduction performance which is superior to the performance obtained in a bilateral configuration. The

[7] For conciseness, *SDW-MWF* is abbreviated to *MWF* in subsequent formulas.

binaural cues of the residual noise can also be preserved by including extensions to the SDW-MWF cost function (cfr. [31,101] and chapter 7). Perceptual evaluations of an idealized binaural SDW-MWF algorithm have also illustrated that both a significant speech intelligibility improvement [177] and a good localization performance [176] can be achieved.

2.2.2 Special case: single target speech source

As shown in [50] and Appendix A.2, in the case of a single target speech source (section 2.1.4), the filters (2.53) are equal to

$$\boxed{\begin{aligned} \mathbf{w}_L &= \mathbf{R}_{vL}^{-1}\mathbf{a}_L \frac{P_s}{\mu+\rho_L} A_{L,\text{ref}}^* \\ \mathbf{w}_R &= \mathbf{R}_{vR}^{-1}\mathbf{a}_R \frac{P_s}{\mu+\rho_R} A_{R,\text{ref}}^* \end{aligned}} \tag{2.54}$$

with

$$\rho_L = P_s \mathbf{a}_L^H \mathbf{R}_{vL}^{-1} \mathbf{a}_L \,, \tag{2.55}$$

$$\rho_R = P_s \mathbf{a}_R^H \mathbf{R}_{vR}^{-1} \mathbf{a}_R \,. \tag{2.56}$$

Using definition (2.25), the (narrowband) output SNRs obtained at the left and right device are also equal to ρ_L and ρ_R, i.e.

$$\text{SNR}_L^{out} = P_s \mathbf{a}_L^H \mathbf{R}_{vL}^{-1} \mathbf{a}_L = \rho_L \,, \tag{2.57}$$

$$\text{SNR}_R^{out} = P_s \mathbf{a}_R^H \mathbf{R}_{vR}^{-1} \mathbf{a}_R = \rho_R \,. \tag{2.58}$$

As shown in [157, 161], the filter expressions (2.54) can be decomposed into a spatial filter, which is equivalent to the so-called Minimum Variance Distortionless Response (MVDR) beamformer [8, 21], followed by a (single-channel) spectral postfilter, i.e.

$$\mathbf{w}_L = \underbrace{\mathbf{R}_{vL}^{-1}\mathbf{a}_L \frac{A_{L,\text{ref}}^*}{\mathbf{a}_L^H \mathbf{R}_{vL}^{-1}\mathbf{a}_L}}_{\text{MVDR}} \cdot \underbrace{\frac{\rho_L}{\mu+\rho_L}}_{\text{postfilter}} \,, \tag{2.59}$$

and similarly for $\mathbf{w}_R$. The speech distortion parameter μ only appears in the spectral postfilter and fulfills the same role as in single-channel constrained Wiener filters [60, 118] or spectral oversubtraction [11, 118]. As the filters obtained for different values of μ merely have different postfilters, the narrowband output SNRs (2.57) and (2.58) are independent of μ. However, larger values of μ allow further attenuation of the residual noise, thus leading to a better listening comfort, at the cost of a higher speech distortion.

2.2.3 Reduced-bandwidth binaural SDW-MWF

Suboptimal reduced-bandwidth SDW-MWF

In [55, 169] several reduced-bandwidth binaural SDW-MWF techniques, which differ in the choice of $\mathbf{F}_{01}$ and $\mathbf{F}_{10}$ in (2.6), have been investigated. In these techniques, only one audio signal is exchanged between the devices (in full-duplex), corresponding to $N = 1$. The considered reduced-bandwidth binaural SDW-MWF techniques are:

- **MWF-Superd**: $\mathbf{F}_{01}$ and $\mathbf{F}_{10}$ are M-microphone monaural fixed superdirective beamformers. A disadvantage is that the acoustic scenario (e.g. the location of the target speech source) has to be known *a priori*.
- **MWF-Front**: The front microphone signals are exchanged between the devices, corresponding to $\mathbf{F}_{01} = \mathbf{F}_{10} = [1\ 0 \dots 0]^T$.
- **MWF-Contra**: $\mathbf{F}_{01}$ and $\mathbf{F}_{10}$ are found by minimizing a local SDW-MWF cost function, which estimates the contralateral speech component using only the M contralateral microphone signals.
- **MWF-Contra-Noise**: Instead of estimating the contralateral speech component as in the previous technique, it is also possible to estimate the noise component, so that a reference of the noise signal is transmitted.

These techniques are suboptimal in the sense that in most scenarios, they do not obtain the same (physical) SNR improvements as a technique which has access to all microphone signals (denoted as **MWF-All**) [49, 55, 177]. In a perceptual evaluation [177], it was however observed that no significant differences in perceptual speech intelligibility occurred between MWF-Front ($M = 2,\ N = 1$) and MWF-All ($M = N = 2$), while both techniques offered significant improvements compared to a bilateral MWF ($M = 2,\ N = 0$). The suboptimal technique thus offer sufficient SNR improvements versus a bilateral technique, hence justifying (reduced-bandwidth) binaural noise reduction. An optimal (full-bandwidth) SDW-MWF however does not seem to offer sufficient further SNR improvements, which also correspond to noticeable intelligibility improvements.

When comparing the different reduced-bandwidth techniques, the MWF-Superd is observed to give the worst performance [55], especially so when the acoustic scenario does not match the conditions for which the fixed beamformer was designed. The other techniques (MWF-Front, MWF-Contra and MWF-Contra-Noise) were studied from an information-theoretical point-of-view using rate-distortion theory in [169]. Intuitively, a transmitted signal can improve the MWF performance if it provides new information, i.e. information which is not available at the receiving device. Which type of information contributes the most, i.e.

information about the speech (MWF-Contra), noise (MWF-Contra-Noise) or both (MWF-Front), then depends on the acoustic scenario and input SNR [169]. For example, in a scenario with only spatially uncorrelated noise (e.g. thermal sensor noise) where the SDW-MWF has to achieve constructive interference for the target speech signal, the MWF-Contra obtains the best SNR improvement (be it only a marginal improvement over the MWF-Front). The MWF-Front achieves a good performance in most scenarios, while its computational complexity is also the lowest of all considered binaural techniques.

In conclusion, it is suggested in [169] that the MWF-Front is the best candidate for binaural reduced-bandwidth noise reduction, when considering both performance and computational complexity. In this thesis, the MWF-Front will therefore be selected in practical evaluations, unless explicitly stated otherwise. However, all of the theoretical derivations, which consider the general binaural processing scheme of section 2.1.2 and make abstraction of the particular choice of $\mathbf{F}_{01}$ and $\mathbf{F}_{10}$, are also applicable to the other binaural reduced-bandwidth techniques.

Optimal reduced-bandwidth SDW-MWF

It is shown in [55] that under the assumption of a single speech source (section 2.1.4), the MWF-All solution can be obtained using the same bandwidth (i.e. $N = 1$) as the previous reduced-bandwidth algorithms. In the proposed distributed MWF algorithm (**DB-MWF**), the filters $\mathbf{F}_{01}$ and $\mathbf{F}_{10}$ are again found by minimizing a local SDW-MWF cost function which estimates the contralateral speech component. However, in contrast to the MWF-Contra, the signal received from the other device is also included in the optimization. Assuming a single speech source (so that (2.23) holds), it can then be shown [55] that by applying this optimization iteratively at both devices, the procedure converges to the optimal MWF-All solution.

A problem with the DB-MWF is that for each new block of data, a few iterations are required before the procedure converges. As each iteration also requires re-transmission of data, the required bandwidth will then actually be higher than the MWF-All algorithm where all microphone signals are exchanged, while the computational complexity is also higher. Therefore, it is assumed that the iterations can be spread over time, i.e. that each new iteration is performed for a new block of data, so that the required bandwidth is again the same as for the suboptimal reduced-bandwidth algorithms. Convergence to the optimum can still be achieved under the assumption of short-term stationarity of the speech and noise signals, so that the second-order statistics stay constant in the considered time window.

Pilot experiments with a more realistic implementation (compared to [55]) of the DB-MWF indicate that the optimal performance is not reached in practice. Most

probably, this is due to the fact that the short-term stationarity assumption is violated, so that the procedure cannot converge. It was also observed that in some scenarios, the performance is even lower than the performance of the (more straightforward) MWF-Front. A final complication is that, as the iterations are spread over time, the tracking performance (for moving speech and noise sources) of the DB-MWF may also be lower. In conclusion, again the MWF-Front seems the best candidate for binaural reduced-bandwidth noise reduction. The DB-MWF algorithm will not be explored further in this thesis, although most of the presented theoretical results can also be applied to the DB-MWF. Namely, as the DB-MWF can be considered to be equivalent to the MWF-All after convergence, the theoretical results for the MWF-All also apply for the DB-MWF. It should also be noted that an extension of the DB-MWF was successfully applied in a wireless acoustic sensor network, which can be viewed as a promising future application[8] [12]. This could imply that as the number of available signals and devices increases, the DB-MWF will become a more efficient and better performing algorithm compared to the straightforward algorithms, while for the (simpler) binaural case the straightforward approaches are advantageous.

2.2.4 Implementation of binaural SDW-MWF

For a monaural (single-device) SDW-MWF, different adaptive algorithms have been proposed [49]. Computationally expensive matrix-decomposition-based algorithms for a time-domain SDW-MWF were presented in [52, 152]. The computational complexity was mitigated by the stochastic gradient algorithms proposed in [165]. However, the latter make use of large circular buffers which have a prohibitive memory requirement. In [54], a stochastic gradient algorithm based on a frequency-domain SDW-MWF cost function was proposed. This implementation reduces both the memory requirement and the computational complexity, and is thus also the best candidate for the implementation of the binaural SDW-MWF.

In the algorithms in [54] a fixed spatial preprocessing stage is included, similar to the classic GSC structure [76] (cfr. Figure 2.3). This preprocessing stage generates a speech reference and noise references, so that the stochastic gradient method only cancels out the noise while preserving the speech signal. It is however proposed here to omit the fixed preprocessing stage in the binaural extension of [54]. A first motivation is that, as the SDW-MWF takes speech distortion explicitly into account [49, 54], the preprocessing stage is less crucial to its performance, compared to the impact on the performance of a GSC. Furthermore, in the context of (binaural) hearing aids there are also some problems associated with fixed

[8] A wireless acoustic sensor network can be viewed as a more general case of the binaural hearing aid application, where other external devices equipped with microphones and a processing unit (e.g. cell-phones or other hearing aids) also exchange signals with the binaural hearing aids.

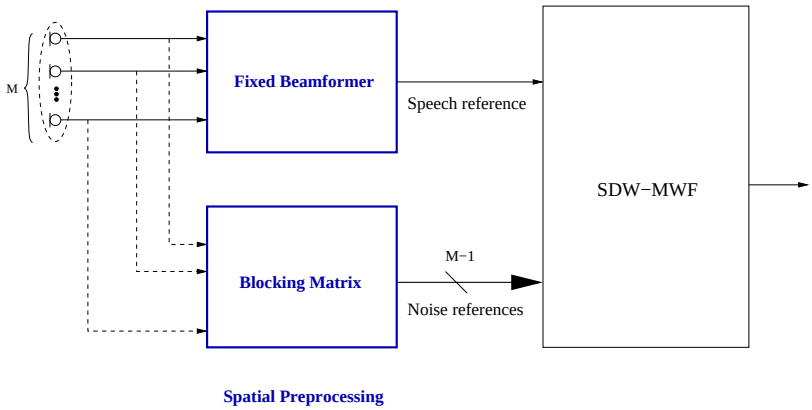


Figure 2.3: Structure of the monaural adaptive SDW-MWF algorithms with spatial preprocessing stage [54].

preprocessing. First, as a fixed preprocessing stage assumes the acoustic scenario is known *a priori*, the performance is degraded when the actual scenario does not match the scenario for which this preprocessing was designed (e.g. when the target speech source is not in the assumed frontal direction [177]). Second, designing a binaural fixed preprocessing (also using received signals from the other device) is not so straightforward, as even small changes in the microphone placement (e.g. due to improper placement of the devices) may have a large impact on the performance. Third, it was demonstrated that fixed preprocessing can distort the binaural cues so that the localization performance is degraded [178, 179].

The binaural extension of the adaptive frequency-domain algorithm of [54], without preprocessing, will thus be used as reference algorithm. However, novel SDW-MWF formulations and adaptive algorithms will also be proposed in this thesis. Compared to the reference algorithm, these are also based on frequency-domain cost functions, have a similar computational complexity and memory requirements, and also do not require a fixed preprocessing stage, while it will be demonstrated that they allow for a better noise reduction and localization performance.

2.3 Conclusion

In this chapter, a description of the binaural noise reduction framework and binaural SDW-MWF technique was given. The presented framework and definitions will be used in the remainder of the thesis.

In section 2.1.1, the speech, noise and microphone signals, which impinge on a hearing aid in a typical speech-in-noise scenario, were defined. The input and output signals in a general binaural processing scheme, where two hearing aids exchange signals over a wireless link, were discussed in section 2.1.2. The definitions and practical estimation of the second-order statistics (SOS), which are crucial to the considered noise reduction techniques, were discussed in section 2.1.3. A frequently occurring scenario is the scenario where there is only a single target speech source. This scenario will often be assumed in the theoretical derivations in this thesis, and was therefore discussed in-depth in section 2.1.4. Finally, performance measures which quantify the noise reduction performance and the binaural cue preservation, were defined in sections 2.1.5 and 2.1.6 respectively.

A review of the binaural SDW-MWF cost function and optimal filters was also presented, both for the general case with an arbitrary number of target speech sources (section 2.2.1) as for the special case where there is only a single target speech source (section 2.2.2). If it is not feasible to exchange all microphone signals between the devices, a reduced-bandwidth binaural SDW-MWF technique is required. Previously proposed reduced-bandwidth binaural SDW-MWF techniques were therefore discussed in section 2.2.3. It was motivated that the MWF-Front is the most appealing reduced-bandwidth technique, so that it is adopted as reference reduced-bandwidth binaural technique in this thesis. Finally, a practical implementation of the binaural SDW-MWF, which will be used as reference algorithm, was discussed in section 2.2.4.

Part II

Noise reduction performance of binaural SDW-MWF techniques

Chapter 3

Perceptual evaluation of bilateral and binaural adaptive SDW-MWF algorithms

This chapter presents a perceptual evaluation of bilateral and binaural SDW-MWF algorithms, which were introduced in chapter 2. In previous work [177], it was demonstrated that off-line batch implementations of the SDW-MWF can significantly improve speech intelligibility, especially so for the binaural SDW-MWF. In this study, more realistic online adaptive implementations of the bilateral and binaural SDW-MWF and a competing bilateral fixed beamformer are evaluated. As the SDW-MWF relies on a VAD, a realistic binaural VAD is also included. The test subjects (both normal hearing subjects and hearing aid users) are tested by an adaptive speech reception threshold (SRT) test in different spatial scenarios, including a realistic cafeteria scenario with nonstationary noise. This realistic performance evaluation, in challenging acoustic scenarios, forms a motivation for the development of binaural SDW-MWF based noise reduction techniques presented in further chapters.

Section 3.1 discusses relevant perceptual evaluations in the literature and defines the research questions addressed in this study.

Section 3.2 specifies the evaluated bilateral and binaural noise reduction algorithms, i.e. the bilateral and binaural SDW-MWF and bilateral fixed beamformer.

Section 3.3 specifies the test procedure, setup, stimuli, acoustic enviromnent, spatial scenarios and parameter settings used in the study.

Section 3.4 presents the performances measured in the perceptual evaluation.

Section 3.5 is a discussion of the obtained results. Answers are formulated for the defined research questions.

Section 3.6 finally gives an overall conclusion of the chapter.

The material presented in this chapter has also been submitted for publication in the Journal of the Acoustical Society of America [38]. A preliminary version of this work was presented as an invited talk at the International hearing aid research conference (IHCON), Lake Tahoe, California, August 2010.

3.1 Problem statement and research questions

Well-known multi-microphone noise reduction techniques are based on fixed and adaptive beamforming [181]. The Generalized Sidelobe Canceller (GSC) [76], which belongs to the class of adaptive beamformers, can be considered as the current state of the art solution adopted in hearing aids. A two-microphone implementation was indeed shown to be able to achieve a speech intelligibility improvement [166, 183]. The SDW-MWF can be viewed as a more recent adaptive beamforming technique. A benefit over the GSC is that in principle, no *a priori* knowledge about the acoustic environment, target speech location or microphone characteristics is required. As a result, the SDW-MWF is also less sensitive to imperfections such as microphone mismatch [164]. In a previous study *Luts et al.* [123], a three-microphone SDW-MWF implementation was evaluated at different test sites, together with other single and multi-microphone noise reduction algorithms. Overall, the SDW-MWF achieved the largest speech intelligibility improvements, even in highly reverberant environments. Even though in laboratory evaluations, it was demonstrated that current adaptive beamformers can achieve large speech intelligibility improvements compared to fixed beamformers [79], there has remained some skepticism about the real-world benefit [193]. That is, relatively few situations are encountered in everyday life where current adaptive beamformers offer significant benefits over fixed beamformers. Still, it should be noted that only monaural implementations in a single device (with a small microphone array) were assessed in [193] and similar studies.

In this chapter, a binaural noise reduction technique, which has access to microphone signals from both hearing aids, is evaluated. As a larger microphone array is available to the noise reduction procedure, an improved performance can be expected compared to bilateral techniques. As in a binaural noise reduction technique two separate microphone arrays are combined, small variations of intermicrophone distances (between microphones of the two different devices) can

occur, for example if one of the devices is not properly placed. An adaptive beamformer such as the SDW-MWF, which does not require *a priori* knowledge and is not sensitive to such variations, is therefore suited for binaural noise reduction. It is noted that also from the perspective of binaural cue preservation, the SDW-MWF offers a valuable approach to binaural noise reduction, as will be demonstrated in chapters 7 and 8.

In a previous study [177], normal hearing subjects were tested by adaptive speech reception threshold (SRT) tests to evaluate the performances of bilateral and binaural SDW-MWF techniques. The binaural SDW-MWF indeed achieved significant SRT improvements compared to the bilateral SDW-MWF, even when only one microphone signal was transmitted to the contralateral device. However, the considered SDW-MWF implementations were idealized in two ways. First, the noise reduction filters of the SDW-MWF were estimated in a batch procedure using the complete microphone signals, whereas in practice, the filters should be adaptively estimated in an online procedure. Second, perfect knowledge of the target speech activity was assumed. In practice, a VAD algorithm is required, which can indicate the periods where the target speech source is active. It should be noted that a VAD is commonly required in adaptive beamformers such as the GSC. It was demonstrated by physical evaluations that the SDW-MWF outperforms the GSC for moderate VAD errors [163]. In [22], it was similarly demonstrated by a physical evaluation that the binaural SDW-MWF retains much of its benefit when using a real VAD, although again only an off-line implementation was assessed. In this study, a performance evaluation of an online adaptive binaural SDW-MWF implementation, using a real VAD, will therefore be conducted, both with normal hearing subjects and with hearing aid users.

The following research questions are addressed and discussed in this chapter: 1) Does an adaptive implementation of a bilateral and binaural SDW-MWF still achieve SRT improvements, even in challenging scenarios? 2) Does the binaural SDW-MWF algorithm retain its benefit over bilateral algorithms? 3) Does using a real VAD result in significant performance degradations for the bilateral and binaural SDW-MWF? 4) Do the adaptive SDW-MWF algorithms yield significant improvements compared to a fixed beamformer? 5) Do the noise reduction algorithms yield different outcomes for the different subject groups?

3.2 Evaluated noise reduction algorithms

3.2.1 Bilateral and binaural SDW-MWF

In this study, both a bilateral and a binaural SDW-MWF technique, which were introduced in chapter 2, are evaluated. For the binaural SDW-MWF, only the

front microphone signals are exchanged between the devices. This technique was denoted as *MWF-Front* in section 2.2.3. In [177], it was demonstrated that this binaural SDW-MWF achieved significant intelligibility improvements over the bilateral SDW-MWF. However, both evaluated SDW-MWF implementations were "idealized" in two ways:

- The SDW-MWF filters were calculated *a priori* in an off-line (batch) procedure, using the complete microphone signals. However, in practice the filters have to be adapted in an online procedure, so that the filters can adapt to any scenario without prior knowledge and so that a (moderately) changing spatial scenario can also be tracked.
- It was assumed that a perfect VAD was available so that the correlation matrices are perfectly estimated. However, in practice a realistic VAD will introduce estimation errors so that the noise reduction performance possibly degrades.

In this study, a realistic implemention of the bilateral and binaural SDW-MWF is therefore evaluated:

- The adaptive RLS-type SDW-MWF algorithm of [54] is extended for the binaural configuration and used as an online adaptive procedure.
- The impact of using a realistic (binaural) VAD algorithm (cfr. Chapter 9) is assessed.

3.2.2 Bilateral fixed beamformer

A bilateral Fixed Beamformer (FB) is also considered in this study as a competing noise reduction technique to the SDW-MWF. The FB calculates output signals in each device as in (2.12), where the filters $\mathbf{w}_L$ and $\mathbf{w}_R$ are calibrated *a priori*, and where $\mathbf{y}_L$ and $\mathbf{y}_R$ only contain the ipsi-lateral microphone signals, i.e. $N = 0$.

A major disadvantage of the FB is that an assumption has to be made about the target speech source location. Most often, it is assumed that the target speech source location is in the frontal (facing) direction. Another FB disadvantage is the sensitivity to "model imperfections" such as microphone mismatch, caused by aging effects or environmental influences [79]. In addition to a degraded noise reduction performance, the model imperfections can also cause distortion of the binaural cues [179]. Adaptive beamformers (such as the SDW-MWF) can offer a better noise reduction performance for coherent noise compared to a FB [79], while the SDW-MWF in particular also offers increased robustness to model imperfections [164]. As the SDW-MWF and other adaptive beamformers rely

on a VAD, they can however be outperformed by a FB in very adverse scenarios (such as in low SNR or nonstationary noise scenarios), where the VAD fails to discriminate between speech+noise and noise-only frames.

The directivity patterns of the considered bilateral FB, measured in an anechoic room on a CORTEX MK2 manikin, are illustrated in Figure 3.1. The directivity from the specified angle is defined as the power ratio of the output signal (in dB) for sound arriving for a certain azimuthal angle (with 0° in front of the head, 90° to the right of the head) and for the diffuse case where sound arrives from all directions. In order to relate the performance to speech intelligibility, the directivity is averaged across frequencies with a speech intelligibility (SI) weighting function as in [5]. The FB is designed so that a notch is formed at 180° while maximum directivity is obtained at 0°. It is however apparent from Figure 3.1 that the direction of maximum directivity is shifted towards the side of the head in practice, when the devices are worn behind the ear, as is also illustrated in [79].

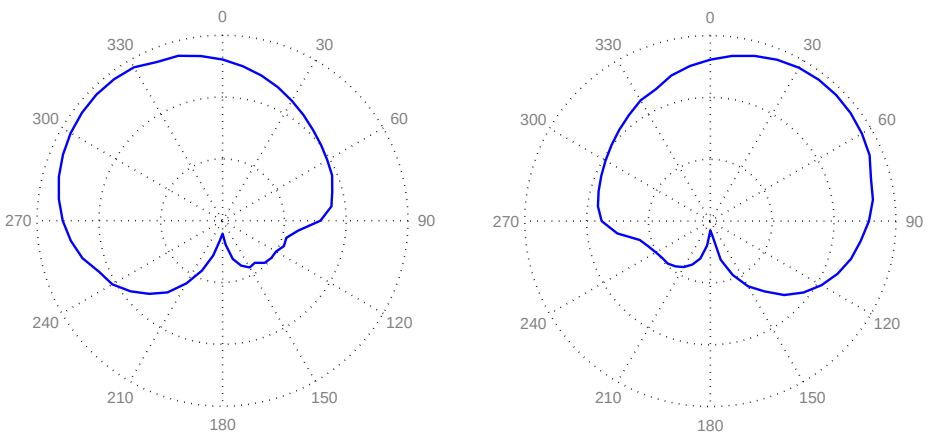


Figure 3.1: SI-weighted directivity pattern (10 dB grid) of bilateral FB at left and right device, for different azimuthal angles.

3.3 Methods

General procedure

The speech reception threshold (SRT) is measured by the adaptive test procedure of [144]. In this procedure the level of the target speech signal is adjusted in steps of 2 dB depending on the response of the test subject (-2dB after a correct sentence recognition, +2dB after an incorrect sentence recognition), while the noise level is

fixed. The 50% SRT, i.e. the SNR level at which 50% of the presented speech is correctly identified by the listener, is then obtained as the average input SNR of the last 10 presentations, including a virtual SNR level based on the response to the last sentence. The input SNRs are measured at the center point of the setup, in absence of the subject.

Each subject is tested in either one or three different spatial scenarios (cfr. subsections *Spatial scenarios* and *Subjects*), with six noise reduction settings (including a reference condition with noise reduction switched off) per scenario (cfr. subsection *Overview evaluated algorithms*). The resulting six or eighteen different conditions are evaluated in randomized order and conducted twice (test and retest) in one (normal hearing subjects) or two sessions (hearing aid users).

Setup

The test setup is depicted in Figure 3.2. Eight Fostex 6310B loudspeakers are placed in a lattice with an interspacing of 1 m, at a height of 1.3 m. The test subject is seated in the middle of this lattice, and is asked to look to the front loudspeaker during the test, whereby minor head movements are allowed. The experimenter and most of the equipment are placed in a separate control room. By means of a surveillance camera the experimenter is able to check if the test subject remains in the correct position during the test.

The adaptive SRT procedure is controlled by the APEX 3 psycho-acoustical test platform [67], which generates the speech and noise stimuli. The generated stimuli are then presented through the loudspeakers using an RME Hammerfall DSP II soundcard.

The test subjects wear behind-the-ear hearing aids (Phonak Versata) in both ears, whereby each device has two omnidirectional microphones ($M = 2$). The hearing aids do not contain a signal processor, but allow access to the microphone signals. These microphone signals are pre-amplified and routed to two hardware platforms (whereby each platform represents one hearing aid), which perform real-time signal processing. The software was developed in the xPC Target real-time software environment [128], which compiles the complete signal processing scheme into C-code and transmits it to the two hardware platforms (the so-called Target machines).

The experimenter can change the parameter settings of the Target machines during the test (at runtime) by means of the so-called Host PC, which is connected to the Target machines via an Ethernet connection. A graphical user interface (GUI), running on the Host PC, was created so that the experimenter can easily change the audio volume, hearing loss thresholds, and noise reduction algorithm. As the subject's own voice can erroneously influence the filter adaptation during the test,

a pushbutton was also implemented by which the updates of all internal variables (speech/noise statistics, noise floor thresholds used in the VAD,...) can be frozen. After each presented sentence, the button is pushed by the experimenter so that the filter adaptation is switched off during the subject's response, and pushed again after the subject's response so that the filter adaptation is again switched on.

Each Target machine has a total of four input audio signals: the two microphone signals of the associated hearing aid, the front microphone signal of the contralateral device (for binaural noise reduction) and a copy of the clean speech signal which is generated by APEX and is also sent to the appropriate loudspeaker. The copy of the clean speech signal is used for constructing a perfect VAD reference signal for some of the tested conditions. The (left and right) processed output signals are redirected to the receivers of the two hearing aids. Instead of earmolds, disposable foam earplugs with tubing are used, which prevent the direct sound from entering the ear canal.

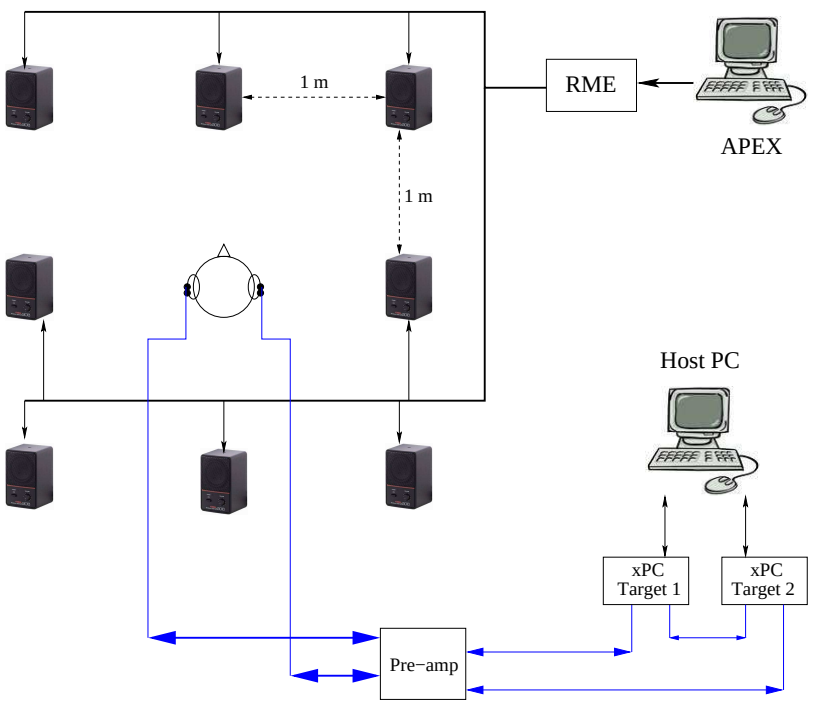


Figure 3.2: Test Setup

Room characteristics

Tests are performed in a soundproof room with dimensions 6m $\times$ 5.5m $\times$ 2.6 m (where 6 m corresponds to the horizontal direction in Figure 3.2 and where 2.6 m is the height of the room). Subjects are seated in the right half of the room, at 1.7 m from the right wall and 3.1 m from the front wall. The reverberation time was measured according to the [91] standard, whereby the equipment belonging to the setup was present, but in absence of a test subject. The room has an SI-weighted (averaged across frequencies with SI-weighting according to [5]) reverberation time of 0.62 s.

Spatial scenarios

Two single noise source scenarios (S0 N45 and S90 N270) and a pseudo-diffuse (three noise sources) scenario (S0 N90/180/270) are considered, which are similar to the scenarios tested in [177]. The scenarios are denoted as Sx Ny, with x the azimuthal angle-of-arrival of the target speech (S) signal and y the azimuthal angle-of-arrival of the noise (N) signal(s). 0° is located in front of the head, 90° to the right of the head. In these scenarios, the loudspeakers are pointed towards the subject, seated at the center of the lattice.

A more challenging realistic scenario, representative of a cafeteria situation, is also considered and shown in Figure 3.3. For this scenario the loudspeakers are not pointed towards the center of the lattice but are pointed in a vertical direction, as illustrated in the figure. This particular setup therefore mimics concurrent (interfering) conversations which are taking place around the hearing aid user, as in a cafeteria.

Stimuli

Sentences spoken by a Dutch male talker [184] are used as speech material. The database consists of 39 lists with 13 sentences per list. As six noise reduction algorithms (cfr. subsection *Overview evaluated algorithms*) are tested (in a test/retest setup), whereby every algorithm is tested in each scenario, this implies that at most three different spatial scenarios can be tested per subject (as this results in a total of 36 test conditions).

In the single noise source and pseudo-diffuse (three noise sources) scenarios, multi-talker babble noise [6] is used as noise signal(s). The signals are scaled so that the total input noise sound pressure level (SPL) at the center point of the lattice, measured in absence of the subject, is equal to 65 dBA.

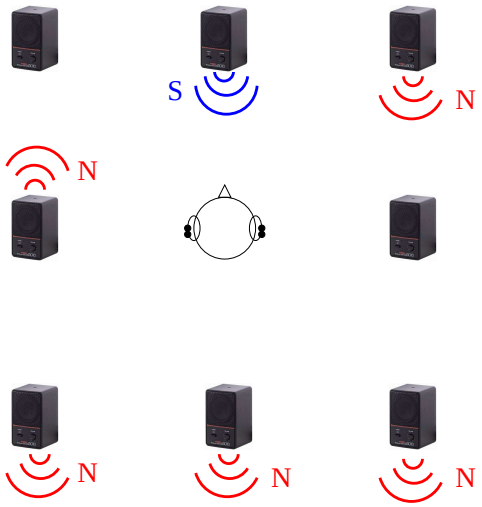


Figure 3.3: Cafeteria scenario

For the cafeteria scenario, sentences of the English hearing-in-noise test (HINT) [135] are used as noise signals. The different noise signals were time-shifted so that they are not all active at the same time, while at least one noise signal is active at any time instant. The signals are scaled so that the average SPL per noise signal is 60 dBA, measured at 1 m of the loudspeaker. For this calibration, speech-weighted noise with a spectrum that matches the average HINT spectrum, is used as calibration stimulus.

Subjects and fitting

Two groups of normal hearing subjects participated in the tests. The first group of ten subjects were tested in three spatial scenarios (S0 N45, S90 N270 and cafeteria). After these tests it was decided to test a second group of ten normal hearing subjects in one additional spatial scenario (S0 N90/180/270). All normal hearing subjects had average hearing thresholds better than or equal to 20 dB hearing level (HL) for octave frequencies between 250 and 8000 Hz.

Eight hearing aid users were tested in three spatial scenarios (cafeteria, S0 N45 and S0 N90/180/270). The subjects were experienced bilateral hearing aid users and had a moderate symmetrical sensorineural hearing loss. The average hearing loss is depicted in Figure 3.4.

A frequency-dependent gain is fitted based on the individual audiograms, in

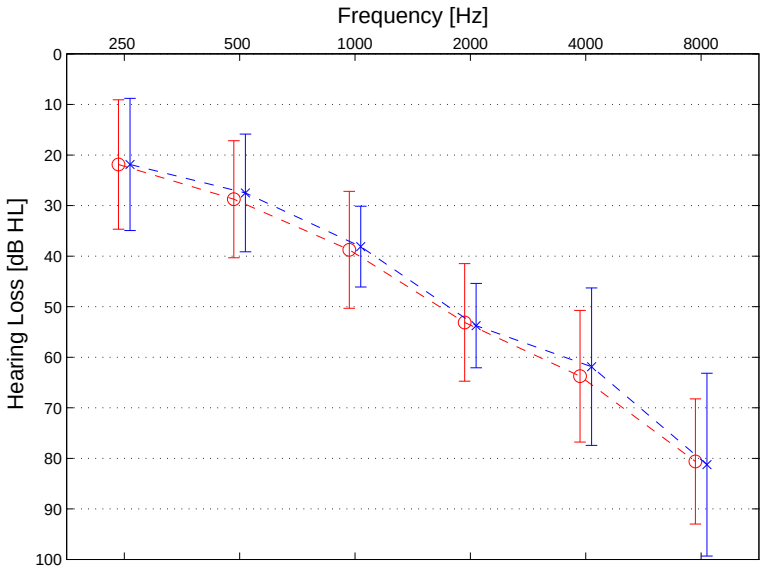


Figure 3.4: Average left (-x) and right (-o) Hearing Loss (HL) per octave frequency of 8 hearing aid users. Standard deviations are indicated by errorbars.

accordance to the NAL-RP prescription rule [47]. Compression is not included, only a limiter set at 100 dB SPL per frequency band. Prior to each SRT test, a pilot test is conducted in order to check if the overall gain is experienced as too loud or too quiet for the hearing aid user. If so, the overall audio volume is adjusted to a comfortable level.

SDW-MWF implementation

In [54], several monaural frequency-domain adaptive SDW-MWF algorithms were proposed, which perform frequency-domain filtering with an overlap-save (OLS) filterbank. The bilateral and binaural adaptive SDW-MWF algorithms considered here are based on the unconstrained algorithm with block-structured correlation matrices, denoted as Algo 1 in [54]. Two major changes are applied:

- The fixed spatial preprocessing stage is omitted, as was motivated in section 2.2.4. The considered adaptive algorithms are therefore fully matched to the batch algorithms evaluated in [177].
- A Weighted Overlap-Add (WOLA) filterbank (cfr. subsection *Support algorithms*) is used instead of an OLS filterbank.

As in [177], the speech distortion parameter μ is chosen as $\mu = 5$ for all test conditions, as this setting offers a good compromise between noise reduction and speech distortion.

Support algorithms

Besides the noise reduction algorithms, some support algorithms are also implemented in the real-time software environment:

- Filterbank: a Weighted Overlap-Add (WOLA) filterbank is implemented, as this is a flexible framework suitable for hearing aid applications [40,18]. The microphone signals are sampled at 20480 Hz, the DFT size is equal to 128, and the frames overlap by 75 %. For these settings, the WOLA filterbank introduces an input-output delay of only 4.6 ms, which is well below the 20 ms limit reported in [171], where the hearing aid delay likely becomes disturbing to the hearing aid user.
- Voice Activity Detection (VAD): the VAD algorithm, denoted as fusion-2 VAD and developed in [35] and chapter 9, is used in this evaluation. The fusion-2 VAD makes a decision fusion of log-energy based VAD algorithms operated at both sides of the head, together with a cross-correlation based VAD (which assumes the target speech signal azimuthal angle-of-arrival is 0°), whereby the decision fusion rule is based on local SNR estimates. More details on the VAD algorithm can be found in chapter 9.
- Frequency-dependent gain and subband power limiter: the gain and limiter use the same WOLA filterbank as the noise reduction algorithms, so that no additional delay is introduced.

Overview evaluated algorithms

The following six noise reduction settings are evaluated:

- Unprocessed (*REF*): Bilateral reference condition where no noise reduction is applied. The front microphone signals are selected as output signals.
- Bilateral FB (*BIL-FB*): Bilateral 2-microphone FB (devices work independently; $M = 2$, $N = 0$).
- Bilateral MWF perfect VAD (*BIL-MWF-P*): Bilateral 2-microphone SDW-MWF (devices work independently; $M = 2$, $N = 0$), which makes use of an ideal VAD reference.

- Bilateral MWF real VAD (*BIL-MWF-R*): Bilateral 2-microphone SDW-MWF (devices work independently; $M = 2$, $N = 0$), which uses a realistic VAD (fusion-2 VAD).
- Binaural MWF perfect VAD (*BIN-MWF-P*): Binaural 3-microphone SDW-MWF (2 ipsilateral microphone signals + the contralateral front microphone signal, denoted as *MWF-Front* in section 2.2.3; $M = 2$, $N = 1$), which makes use of an ideal VAD reference.
- Binaural MWF real VAD (*BIN-MWF-R*): Binaural 3-microphone SDW-MWF (2 ipsilateral microphone signals + the contralateral front microphone signal, denoted as *MWF-Front* in section 2.2.3; $M = 2$, $N = 1$), which uses a realistic VAD (fusion-2 VAD).

3.4 Experimental results

The average SRT scores obtained in the four spatial scenarios are provided in Figures 3.5-3.8. Statistical analyses were carried out on the SRT data using the SPSS 16.0 software. As it was illustrated in [177] that the performance of the algorithms is dependent on the position of the speech and noise source(s), every spatial scenario is analyzed separately. To analyze the SRT data, a mixed-model factorial repeated-measures analysis of variance (abbreviated to "ANOVA") is carried out. The mixed-model ANOVA includes two within-subjects factors, i.e. *algorithm* (6 levels) and *test/retest* (2 levels), and one between-subjects factor, i.e. *subject group* (normal hearing subject or hearing aid user). Pairwise comparisons between all algorithmic conditions are carried out as *Post Hoc* tests. To guard against Type I errors, Bonferroni corrections for multiple comparisons are applied. A p-value of $p = 0.05$ is used as the threshold for significance in all tests.

3.4.1 S0 N45 scenario

The SRT data of the normal hearing subjects and hearing aid users is provided in Figure 3.5. It can be observed that the hearing aid users consistently reach their SRT at higher SNR levels than the normal hearing subjects. The mixed-model ANOVA however indicates that the between-subjects variable *subject group* does not interact with the within-subjects factors, i.e. p-values of $p = 0.118$ and $p = 0.267$ are measured for the interactions *subject group* - *test/retest* and *subject group* - *algorithm* respectively. This finding motivates a single analysis for both subject groups.

A main effect is found for the factor *algorithm* ($p < 0.001$), but no significant effect of *test/retest* ($p = 0.866$) and no interaction effect between the two factors

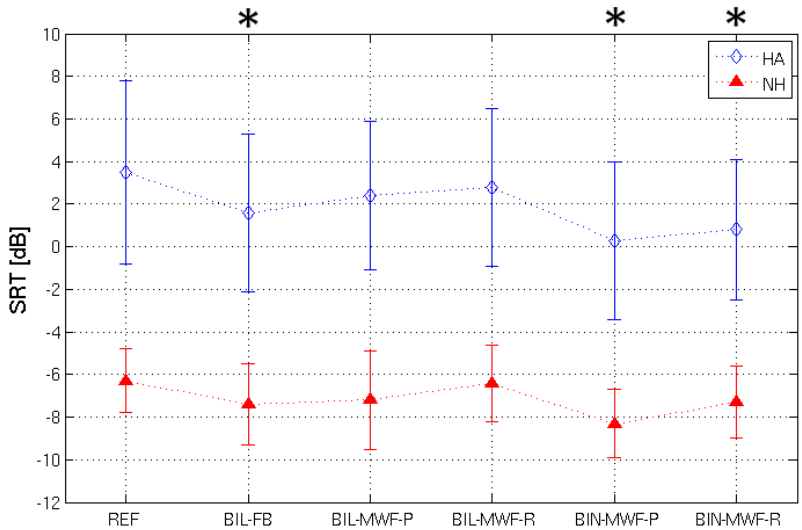


Figure 3.5: Average SRT results of 10 normal hearing subjects (NH) and 8 hearing aid users (HA), for the S0 N45 scenario. Standard deviations are indicated by errorbars. Algorithms which significantly improve the SRT for both groups, compared to the unprocessed (REF) condition, are marked by a "$*$" above the graph.

($p = 0.072$). *Post Hoc* tests show that three algorithms obtain a significant SRT improvement over the unprocessed condition: the largest SRT improvement of 2.6 ± 1.9 dB ($p < 0.001$) is obtained by BIN-MWF-P, followed by an improvement of 1.8 ± 1.9 dB ($p = 0.007$) for BIN-MWF-R, and finally an improvement of 1.5 ± 1.8 dB ($p = 0.035$) for the BIL-FB. The binaural SDW-MWF algorithms are significantly better than their bilateral counterparts ($p = 0.012$ for BIN-MWF-P vs. BIL-MWF-P and $p = 0.037$ for BIN-MWF-R vs. BIL-MWF-R), but not significantly better than the FB ($p = 0.189$ and $p = 1.000$ for BIL-FB vs. BIN-MWF-P and BIN-MWF-R respectively). BIN-MWF-R obtains a 0.8 dB lower SRT improvement than BIN-MWF-P, but this difference is statistically non-significant ($p = 0.911$).

3.4.2 S90 N270 scenario

The average SRT data of ten normal hearing subjects is provided in Figure 3.6. The ANOVA indicates a main effect for *algorithm* ($p = 0.002$), no significant effect for *test/retest* ($p = 0.526$) and no interaction effect ($p = 0.233$). As is

apparent from the Figure, none of the algorithms is however able to improve the intelligibility compared to the unprocessed (REF) condition. In fact, the *Post Hoc* tests reveal that BIL-MWF-R significantly degrades the performance, namely by 1.6 ± 1.3 dB ($p = 0.049$). For the other algorithms, no significant difference with the unprocessed (REF) condition can be demonstrated. Therefore, this scenario was not included in the evaluations with hearing aid users.

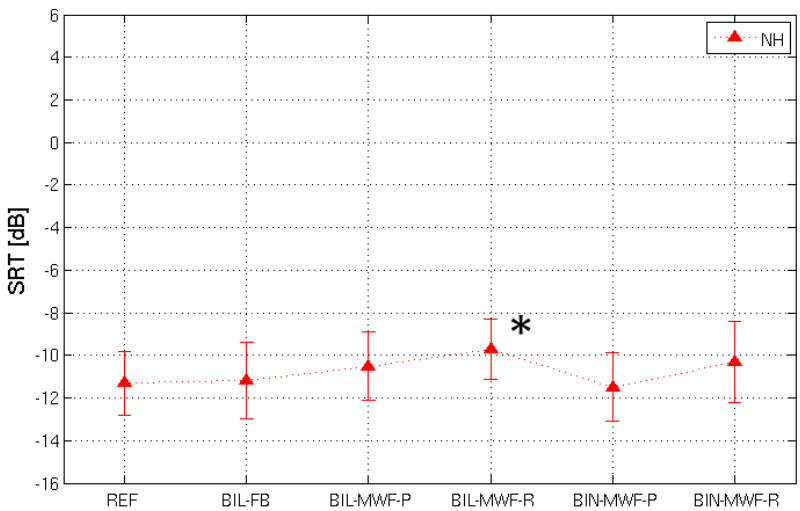


Figure 3.6: Average SRT results of 10 normal hearing subjects (NH), for the S90 N270 scenario. Standard deviations are indicated by errorbars. Significant degradations compared to the unprocessed (REF) condition are marked by a "$*$".

3.4.3 S0 N90/180/270 scenario

The SRT data of the normal hearing subjects and hearing aid users is provided in Figure 3.7. It can again be observed that the hearing aid users consistently reach their SRT at higher SNR levels than the normal hearing subjects. The mixed-model ANOVA again indicates that the between-subjects variable *subject group* does not interact with the within-subjects factors, i.e. p-values of $p = 0.132$ and $p = 0.218$ are measured for the interactions *subject group* - *test/retest* and *subject group* - *algorithm* respectively, so that a single analysis can be made for both subject groups.

A main effect is found for the factors *algorithm* ($p < 0.001$) and *test/retest* ($p = 0.026$). The significant effect for *test/retest* can most likely be attributed to

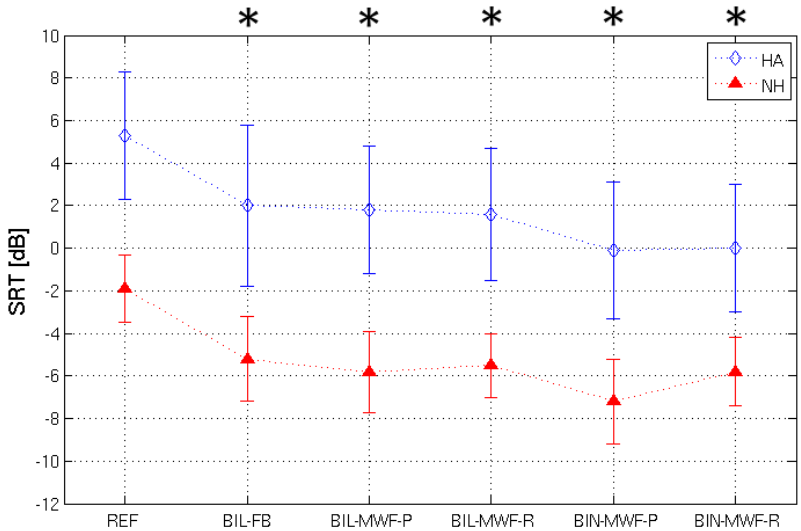


Figure 3.7: Average SRT results of 10 normal hearing subjects (NH) and 8 hearing aid users (HA), for the S0 N90/180/270 scenario. Standard deviations are indicated by errorbars. Algorithms which significantly improve the SRT for both groups, compared to the unprocessed (REF) condition, are marked by a "$*$" above the graph.

the results of the hearing aid users (tested in two separate sessions), whose retest results are indeed 1.0 dB better on average, compared to the test results. However, since no interaction between *algorithm* and *test/retest* is found ($p = 0.822$), the pairwise comparisons between algorithms are valid for both the test and retest session. All algorithms significantly improve the SRT over the unprocessed (REF) condition ($p < 0.001$). The largest improvements are found for the binaural SDW-MWF algorithms (5.4 ± 1.5 dB and 4.5 ± 1.6 dB for BIN-MWF-P and BIN-MWF-R respectively), followed by the bilateral SDW-MWF algorithms (3.7 ± 1.1 dB and 3.7 ± 1.4 dB for BIL-MWF-P and BIL-MWF-R respectively), and finally the smallest improvement for BIL-FB (3.3 ± 1.2 dB). Although BIN-MWF-P significantly outperforms all bilateral algorithms ($p \leq 0.004$), no significant improvement over the bilateral algorithms can be demonstrated ($p \geq 0.133$) when a real VAD is used. BIN-MWF-R obtains a 0.9 dB lower SRT improvement than BIN-MWF-P, but this difference is statistically non-significant ($p = 0.331$).

3.4.4 Cafeteria scenario

The SRT data of the normal hearing subjects and hearing aid users is provided in Figure 3.8. Again, the hearing aid users consistently reach their SRT at higher SNR levels than the normal hearing subjects. The mixed-model ANOVA indicates that there is a significant interaction effect between the factors *subject group* and *algorithm* ($p < 0.001$). It is also clearly visible in Figure 3.8 that the SDW-MWF algorithms using a real VAD seem to lead to different SRT improvements in the two subject groups. A separate ANOVA is therefore carried out for each subject group.

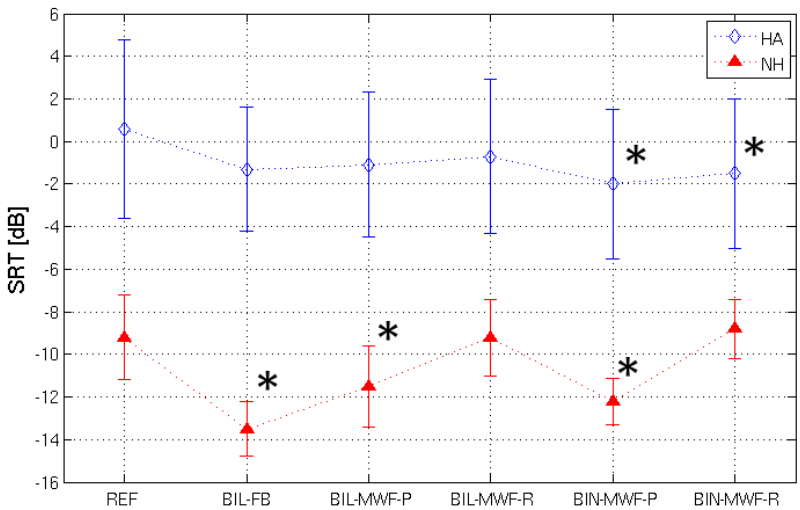


Figure 3.8: Average SRT results of 10 normal hearing subjects (NH) and 8 hearing aid users (HA), for the Cafeteria scenario. Standard deviations are indicated by errorbars. Significant SRT improvements compared to the unprocessed (REF) condition are marked by a "$*$".

For the normal hearing subjects, a main effect for the factor *algorithm* is found ($p < 0.001$), but no significant effect of *test/retest* ($p = 0.112$) and no interaction effect ($p = 0.411$). *Post Hoc* tests reveal that three algorithms obtain a significant SRT improvement over the unprocessed (REF) condition: the BIL-FB obtains the largest SRT improvement of 4.2 ± 1.9 dB ($p = 0.002$), followed by the BIN-MWF-P with 3.0 ± 1.7 dB ($p = 0.006$) and the BIL-MWF-P with 2.2 ± 1.7 dB ($p = 0.033$). The SDW-MWF algorithms using a real VAD do not obtain SRT improvements, and are significantly outperformed by the other algorithms.

For the hearing aid users, a main effect for the factor *algorithm* is found ($p = 0.002$),

but no significant effect of *test/retest* ($p = 0.437$) and no interaction effect ($p = 0.607$). Unlike for the normal hearing subjects, now only the binaural SDW-MWF obtains a significant SRT improvement which moreover does not seem affected by using a real VAD: the SRT improvements are 2.6 ± 1.1 dB ($p = 0.004$) for BIN-MWF-P and 2.2 ± 1.2 dB ($p = 0.025$) for BIN-MWF-R. No significant SRT improvements are found for the BIL-FB ($p = 0.529$) or BIL-MWF-P and BIL-MWF-R ($p = 0.435$ and $p = 1.000$ respectively), and no significant differences can be demonstrated for any other comparisons between algorithms.

3.5 Discussion

3.5.1 Differences between subject groups

The absolute SRT scores show differences between subject groups. As commonly observed in perceptual evaluations, the hearing aid users have higher (=worse) SRT scores than the normal hearing subjects. This worse performance can be attributed to a combination of effects [47, 143], but more particularly to supra-threshold auditory problems, such as worse spectral and temporal resolution.

As the SRT scores are determined at very different SNR levels for the normal hearing subjects and hearing aid users, it is possible that the algorithms performed differently for both subject groups. More in particular, a prior physical evaluation (not reported in this chapter) indicated that the SNR improvement obtained by the SDW-MWF (using a perfect VAD) decreases if the input SNR is lower, while the speech distortion increases. Moreover, the performance likely degrades further when using a real VAD, as the VAD performs worse at low SNRs and thus introduces estimation errors.

The statistical analysis indicates that for the S0 N45 and S0 N90/180/270 scenarios, the relative SRT improvements are not significantly different for both subject groups. This result suggests that on average, even for the normal hearing subjects, the SNR levels are sufficiently high for a good performance of both the SDW-MWF and the VAD. For the cafeteria scenario on the other hand, significantly different SRT improvements are observed for the two groups. First, as the normal hearing subjects are evaluated at very low SNR levels (also lower than for the S0 N45 and S0 N90/180/270 scenarios), the VAD cannot perform well. As a result, the SDW-MWF algorithms using a real VAD are indeed not able to improve the speech intelligibility. Second, as this is actually a speech-in-speech scenario (thus with highly nonstationary noise, which violates the noise-stationarity assumption, cfr. section 2.1.3), the SDW-MWF performance likely degrades, especially at low SNR levels. This may explain why the FB outperforms the SDW-MWF algorithms for the normal hearing subjects. At higher SNR levels, the VAD is less affected, and

the noise component has less weight in the SDW-MWF cost function so that the SDW-MWF performance is also less affected. For the hearing aid users, who obtain their SRTs at higher SNR levels, the results indeed indicate that the binaural SDW-MWF (both BIN-MWF-P and BIN-MWF-R) achieves the best performance.

3.5.2 Algorithm performance

Adaptive SDW-MWF performance

One of the general goals of this evaluation is to assess whether an adaptive bilateral or binaural SDW-MWF algorithm is able to improve speech intelligibility in a realistic listening environment. In previous work, a three-microphone adaptive bilateral SDW-MWF algorithm based on [54] was evaluated at multiple test sites and shown to indeed offer significant intelligibility improvements [123]. Although the SDW-MWF algorithms in this chapter are also based on [54], they are different in a few crucial points, most notably the omission of a fixed preprocessing stage and the use of a low-delay WOLA filterbank (cfr. section 3.3). The evaluated adaptive bilateral and binaural SDW-MWF algorithms are therefore better matched to the batch algorithms evaluated in [177]. The fixed preprocessing stage is omitted in order not to compromise the localization performance (an important aspect of binaural hearing aids). Moreover, if the target speech signal does not arrive from the assumed frontal direction, the performance can also degrade when using a fixed preprocessing stage.

The results indicate that the adaptive SDW-MWF algorithms retain their speech intelligibility improvements for scenarios with a single noise source (S0 N45) and with multiple noise sources (S0 N90/180/270), which are comparable to scenarios tested in [177]. For the more realistic cafeteria scenario, it was anticipated that the nonstationarity of the noise would limit the SDW-MWF performance. However, for hearing aid users who obtain their SRTs at high SNR levels, the binaural SDW-MWF is still able to obtain significant SRT improvements.

As in [177], ten normal hearing subjects were also tested in a scenario with speech and noise arriving from different sides of the head (S90 N270). Because of the headshadow effect, there is a large SNR difference between the two ears. In accordance with [177], prior physical evaluations indicate that a large SNR improvement is obtained in the worst SNR ear, while only a small SNR improvement is obtained in the best SNR ear (which is mainly used for speech understanding). Moreover, as no effort is made here to preserve the residual noise cues, the listener cannot benefit from the binaural unmasking effect [194]. As a result, no SRT improvements are found for any of the tested SDW-MWF algorithms, which is in accordance with the results in [177]. It should however be noted that it is possible to include so-called partial noise estimation, where

part of the unprocessed microphone signal is mixed into the SDW-MWF output signal (cfr. [31, 101] and chapter 7). As the listener can then potentially benefit from the binaural unmasking effect, the SRT can be improved for this particular scenario [177]. Due to limited testing time and limited sentence material, this algorithm is not included in the current evaluation.

Binaural versus bilateral SDW-MWF performance

The transmission of microphone signals between hearing aids clearly comes at a cost: both the computational complexity and the power consumption increase. A major research question is therefore whether the binaural benefit (in terms of SRT improvements) justifies the extra overhead. In array processing literature, it is well-known that if the size of an array increases and if more microphones are used, narrower beampatterns can be created. This helps in scenarios with a closely-spaced speech and noise source, as the noise can then be reduced without affecting the speech. Additional microphones also offer more degrees of freedom in the design, which particularly results in benefits in multiple noise sources scenarios. An extreme example based on these principles are the so-called hearing glasses where multiple microphones are built into the frame of eyeglasses [159]. The same effects are also observed for the binaural SDW-MWF [168, 177], although the analysis is only made for an idealized algorithm (with perfect knowledge of the speech and noise signal second-order statistics).

The results of this work indicate that the benefit of using a binaural SDW-MWF is retained in realistic circumstances (adaptive algorithm, real VAD, challenging listening environment). In the S0 N45 scenario, where a narrow beamwidth is required, the binaural SDW-MWF (both BIN-MWF-P and BIN-MWF-R) indeed obtains the largest SRT improvement, while the bilateral SDW-MWF (BIL-MWF-P and BIL-MWF-R) does not obtain a significant improvement. The bilateral FB (BIL-FB) is able to achieve a small SRT improvement. Although this SRT improvement is smaller than the improvement obtained with the binaural SDW-MWF, no significant difference can be demonstrated. In the multiple noise sources scenario (S0 N90/180/270), all algorithms are able to improve the SRT, but again the binaural SDW-MWF (both BIN-MWF-P and BIN-MWF-R) obtains the highest SRT improvement. When using a perfect VAD, a significant difference between the binaural SDW-MWF and all bilateral algorithms could also be demonstrated. The cafeteria scenario again presents a listening environment with multiple noise sources, which are however not all active at the same time (so that the noise directions are constantly changing). Nevertheless, for the hearing aid users, again the binaural SDW-MWF (BIN-MWF-P and BIN-MWF-R) offers significant SRT improvements, while no significant improvements can be demonstrated for any of the bilateral algorithms. Finally, for the S90 N270 scenario, none of the algorithms is able to improve the SRT. As previously

noted, it was demonstrated that including partial noise estimation can offer SRT improvements [177], but then only in combination with a binaural SDW-MWF algorithm.

In conclusion, it can be stated that in the tested scenarios, either only the binaural SDW-MWF achieves a significant SRT improvement, or otherwise the binaural SDW-MWF significantly outperforms the bilateral algorithms (with an exception in the S0 N45 scenario where the improvement over the FB is non-significant). As it can be assumed that the number of microphone signals that can be exchanged between the devices is limited, only the contralateral front microphone signals are exchanged in the evaluated binaural SDW-MWF algorithm. Although exchanging additional microphone signals could result in a further improvement, it was observed (for batch algorithms) that this does not have a significant impact on the performance [177]. This setup is not tested in the current study, but the same effect would most likely be observed for the adaptive algorithms.

VAD performance

In the S0 N45 and S0 N90/180/270 scenarios, the SRT improvement obtained with the binaural SDW-MWF degraded by 0.8-0.9 dB when using BIN-MWF-R instead of BIN-MWF-P, although this difference was non-significant. It should however be noted that the degradation is larger for the normal hearing subjects than for the hearing impaired subjects, as is visible in Figures 3.5 and 3.7. Although statistically no differences can be demonstrated between both groups, the degradation of 0.9 dB may thus be an overestimation. Nevertheless, BIN-MWF-R still outperforms the other algorithms. For the bilateral SDW-MWF, no significant differences were observed between the performance of BIL-MWF-P and BIL-MWF-R.

As previously mentioned, significant differences between subjects groups are observed for the cafeteria scenario, in contrast to the other scenarios. The VAD does not perform well at the low SNR levels where the normal hearing subjects are measured. However, for the hearing aid users, who obtain their SRTs at higher SNR levels, a non-significant difference of only 0.4 dB between BIN-MWF-P and BIN-MWF-R indicates that the VAD does not have a major impact. This example illustrates that for some scenarios, it is indeed necessary to also test hearing aid users, who are evidently the actual target users.

For the S90 N270 scenario, where only normal hearing subjects were tested, a very low SRT is achieved, as in this scenario the headshadow and binaural unmasking effects can be fully exploited. The VAD does not perform well which even results in a significant SRT degradation for the bilateral SDW-MWF. Again, as hearing aid users would obtain their SRT at higher SNR levels, the obtained degradation

would likely be smaller. For the binaural SDW-MWF algorithms, no significant degradation is observed.

In conclusion, it was demonstrated that the SDW-MWF retains its SRT improvements when using a real VAD, at the SNR levels where the hearing aid users would actually experience intelligibility improvements (around the SRT). Moreover, the binaural SDW-MWF retains its benefit over bilateral SDW-MWF algorithms, so that it is indeed a valid candidate for binaural noise reduction.

SDW-MWF versus fixed beamformer performance

A bilateral FB is tested here as competing noise reduction technique to the bilateral and binaural SDW-MWF. The FB has the advantage that it does not rely on a VAD and is not affected in nonstationary noise conditions, so that it was not clear *a priori* whether the SDW-MWF could outperform the FB in these scenarios.

The FB obtains a significant SRT improvement in the S0 N45 scenario, in contrast to the bilateral SDW-MWF. For the S0 N90/180/270, both algorithms achieve comparable SRT improvements. The FB outperforms the bilateral SDW-MWF for the normal hearing subjects in the S90 N270 and cafeteria scenarios, where very low SNRs are encountered. For hearing aid users, the algorithms obtain a similar performance. These results indicate that in these scenarios no SRT benefit can be demonstrated for the bilateral SDW-MWF, compared to the FB. It should however be noted that the bilateral SDW-MWF could achieve higher benefits when it is implemented as in [123], where a fixed preprocessing stage is included in the scheme.

The binaural SDW-MWF is shown to outperform the bilateral FB, as explained in the previous sections, and is thus certainly a good choice for binaural noise reduction. In contrast, designing a binaural FB is not so straightforward, as even small changes in the microphone placement (due to improper placement of the devices for example) could have a large impact on the performance (similarly to the sensitivity to model imperfections as explained in section 3.2.2).

Finally, we note that the evaluated bilateral FB is optimally calibrated and so does not suffer from the imperfections due to the wear and tear of daily use [79]. The results obtained with the FB should thus be viewed as a performance upper bound, while real-life performance could be significantly worse. For the same reason, some of the advantages of the SDW-MWF algorithms, which are not sensitive to these imperfections [164], can thus not be fully appreciated by this evaluation.

3.6 Conclusion

In [177], it was demonstrated that a binaural SDW-MWF achieves significant SRT improvements over a bilateral SDW-MWF or adaptive directional microphone. The evaluated algorithms were however idealized in a number of ways. The bilateral and binaural SDW-MWF filter coefficients were calculated in an off-line batch procedure, where the complete microphone signals are used. Moreover, perfect knowledge of the target speech activity, which is necessary to estimate the second-order statistics, was assumed to be available. Finally, it should also be noted that the stimuli were presented through headphones, and only normal hearing subjects participated in the evaluation.

In this chapter, online adaptive bilateral and binaural SDW-MWF algorithms were evaluated. The algorithms were implemented on a real-time platform to allow for a realistic evaluation. The performance of the SDW-MWF algorithms, a bilateral fixed beamformer and a reference unprocessed condition were assessed by an adaptive SRT procedure for a total of four spatial scenarios, including a challenging cafeteria scenario with highly nonstationary noise (i.e. interfering speakers). A VAD algorithm was also implemented on the real-time platform so that the impact of using a real VAD instead of a perfect VAD could also be assessed. Finally, basic hearing aid functionality was implemented so that hearing aid users could also be evaluated, in addition to normal hearing subjects.

The conclusions are as follows. First, it was observed that a realistic online adaptive implementation of the binaural SDW-MWF retains its benefit over bilateral algorithms. With only one exception, it can be stated that for the hearing aid users, either only the binaural SDW-MWF algorithm achieved significant SRT improvements, or otherwise the binaural SDW-MWF algorithm achieved a significant improvement over the bilateral algorithms. Second, no significant performance degradations were observed when using a real VAD instead of a perfect VAD, at the SNR levels where the hearing aid users reached their SRT. Third, it was observed that the bilateral adaptive SDW-MWF algorithm did not perform better than the bilateral FB. It was however noted that no model imperfections (such as microphone mismatch) were present. If these occur, it is known that the FB performance can degrade, while the SDW-MWF is robust against such effects. Fourth, a significant performance difference between the two subject groups was only observed for the cafeteria scenario. For the normal hearing subjects, who reached their SRT at relatively low SNR levels, both the SDW-MWF and the VAD performance were degraded so that the FB achieved the best improvement. For the hearing aid users, who reached their SRT at higher SNR levels, the SDW-MWF and VAD were not degraded. Significant improvements could only be demonstrated for the binaural SDW-MWF. This example has also illustrated that for some scenarios, evaluating hearing aid users, who are obviously the actual target subjects, indeed has great value.

Chapter 4

Rank-one SDW-MWF techniques and performance analysis under second-order statistics estimation errors

In theoretical studies of the SDW-MWF, it is usually assumed that there is a single target speech source. In this way, a closed form expression for the optimal filter is found, which allows for a thorough performance analysis. Moreover, the filter can be decomposed into a spatial filter (related to other known techniques) and a single-channel postfilter, which is a conceptually interesting structure. Unfortunately, it is not straightforward to make a robust practical algorithm based on this decomposition. Instead, an algorithm based on the standard SDW-MWF formulation, which only requires a (relatively easy) estimation of speech and noise correlation matrices, is mostly used in practice. This algorithm has been evaluated in chapter 3, where it was demonstrated that significant intelligibility improvements could be achieved, even in challenging scenarios. Even so, a higher robustness against estimation errors in the second-order statistics (SOS) is still very valuable.

This chapter presents a theoretical study and experimental validation on a binaural hearing aid setup of the standard SDW-MWF formulation, where the effect of estimation errors in the second-order statistics is analyzed. In this case, and for a single target speech source, an algorithm based on the standard SDW-MWF formulation is found not to behave as predicted theoretically. Second, two alternative formulations, namely the rank-one SDW-MWF (*R1-MWF*) and

the spatial prediction SDW-MWF (*SP-MWF*), are also studied in the presence of estimation errors in the second-order statistics. The R1-MWF and SP-MWF implicitly assume a single target speech source, but still only rely on the speech and noise correlation matrices. It is proven theoretically and illustrated through experiments that algorithms based on these alternative formulations behave close to the theoretical optimum, and hence outperform an algorithm based on the standard SDW-MWF formulation. Moreover, as they have a conceptually interesting form, they are better suited for an extension with other useful features, as will be demonstrated in subsequent chapters.

Section 4.1 introduces the problem and gives an overview of the relevant literature.

Section 4.2 describes the alternative SDW-MWF formulations (R1-MWF and SP-MWF), which implicitly assume a single target speech source.

Section 4.3 presents theoretical expressions for the standard SDW-MWF and the two novel formulations, under SOS estimation errors.

Section 4.4 presents theoretical expressions for the narrowband output SNRs obtained by the standard SDW-MWF and the two novel formulations, under SOS estimation errors.

Section 4.5 validates the theoretical results by experiments, both for a single target speech source scenario as well as for a more general scenario.

Section 4.6 presents an overall conclusion of the chapter.

The material presented in this chapter was published in the proceedings of ICASSP [32] and in the IEEE Trans. Audio, Speech, Lang. Process. journal [36].

4.1 Introduction

The SDW-MWF has been thoroughly studied in previous theoretical work (e.g. [24, 50, 77, 89, 160, 161]). In the analysis, it is often assumed that there is a single target speech source. As a consequence, the frequency-domain speech correlation matrices are rank-one matrices (2.23). A closed-form expression for the SDW-MWF can then be found, which explicitly depends on the speech power and steering vector (cfr. (2.54), Appendix A.2 and [50]). Using this closed-form expression, the SDW-MWF can be related to the MVDR [21], i.e. the SDW-MWF is equivalent to the MVDR followed by a single-channel spectral postfilter (cfr. (2.59) and [157, 161]). Recent theoretical contributions provide alternative formulations (still assuming a single target source) for the MVDR [77] or more generally for the SDW-MWF [160], that are also structured as a spatial filter followed by a single-channel postfilter. These formulations do not depend explicitly

on the speech power and steering vector, but only make use of the speech and noise second-order statistics. Using these formulations, the trade-off between speech distortion and noise reduction was quantified in [160], in analogy to the single-channel case [24].

The decomposition into a spatial filter and spectral postfilter is conceptually interesting, as for example, the spatial filter and postfilter can then be updated at different rates, or extended independently with other features, as will be demonstrated in subsequent chapters. Unfortunately the closed-form expression proposed in [50] does not allow for a practical algorithm as the speech power and steering vector would have to be calibrated or somehow estimated. Alternatively, the SDW-MWF could be implemented as a TF-GSC followed by a single-channel Wiener postfilter [157], but such algorithms would then suffer from the robustness issues of the GSC [164]. The standard SDW-MWF algorithm is based on the standard SDW-MWF formulation (cfr. (2.53) and [52, 54, 162]). As it does not assume a single target speech source, the filter is not in a decomposed structure. As a result, the algorithm only requires estimation of the speech and noise correlation matrices. However, as the closed-form formulations in [160] also only depend on the speech and noise correlation matrices, a robust algorithm can be derived which is similar to the standard SDW-MWF algorithm, but then explicitly decomposed into a spatial filter part and a spectral postfilter part.

In this chapter, the performances of the standard SDW-MWF and two novel formulations are studied theoretically and through experiments on a binaural hearing aid setup. The presented theoretical results are however relevant for any type of microphone array, especially so if the number of microphones in the array is large. The effect of SOS estimation errors is included in the theoretical analysis as well as in the experiments. It is shown that algorithms based on the different formulations behave differently in the presence of these estimation errors.

For a single target source, it is proven that an algorithm based on the standard SDW-MWF does not behave as predicted theoretically. In particular, when estimation errors are present in the speech correlation matrix, the narrowband SNR improvement (obtained by spatial filtering) is shown to be dependent on the speech distortion parameter, in contrast to the theoretical performance. Moreover, the intelligibility improvement is smaller than expected, especially so when a small speech distortion parameter value is chosen.

The performance of algorithms based on the two alternative SDW-MWF formulations, which assume a single target speech source and are structured as a spatial filter followed by a single-channel postfilter, is also studied. The alternative SDW-MWF formulations are referred to as the rank one SDW-MWF (*R1-MWF*), which is based on the filter in [160], and the Spatial Prediction SDW-MWF (*SP-MWF*), which is an extension of the filter in [8, 32]. It will be shown that algorithms based on these formulations perform close to the optimal (theoretical) performance,

and moreover, that they outperform an algorithm based on the standard SDW-MWF. In particular, in order to obtain a similar amount of noise reduction, the standard SDW-MWF introduces more speech distortion than the R1-MWF and SP-MWF. A simulation for a scenario with more than one target source, where the rank-one assumption is violated, is also performed. The R1-MWF and SP-MWF still obtain a large SNR improvement compared to the standard SDW-MWF, but they introduce different distortions. It is shown that the R1-MWF attenuates the dominant sources, while the SP-MWF tries to preserve the dominant sources. The SP-MWF therefore introduces the least overall distortion.

4.2 Rank-one SDW-MWF techniques

Configuration and standard SDW-MWF

The theoretical results in this chapter are applicable to any kind of microphone array, not just for the special case of a binaural hearing aid setup. To simplify the notation, a single generic microphone array, consisting of M microphones, will therefore be considered. As in section 2.1.1, the microphone signals Y_m, X_m and V_m are stacked in the M-dimensional vectors $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{v}$, with $\mathbf{y} = \mathbf{x} + \mathbf{v}$. The reference microphone signal is denoted as Y_{ref} and is then equal to $Y_{\text{ref}} = \mathbf{e}_{\text{ref}}^H \mathbf{y}$, where $\mathbf{e}_{\text{ref}} = [0 \dots 0 \quad 1 \quad 0 \dots 0]^T$ is an M-dimensional vector where the entry corresponding to the reference microphone is equal to one. The correlation matrix $\mathbf{R}_y$, the speech correlation matrix $\mathbf{R}_x$ and the noise correlation matrix $\mathbf{R}_v$ are defined as

$$\mathbf{R}_y = \mathcal{E}\{\mathbf{y}\mathbf{y}^H\}, \quad \mathbf{R}_x = \mathcal{E}\{\mathbf{x}\mathbf{x}^H\}, \quad \mathbf{R}_v = \mathcal{E}\{\mathbf{v}\mathbf{v}^H\}\,, \tag{4.1}$$

Assuming that the speech and the noise components are uncorrelated, $\mathbf{R}_y = \mathbf{R}_x + \mathbf{R}_v$. The output signal Z is obtained as $Z = \mathbf{w}^H \mathbf{y}$.

The standard SDW-MWF was given by (2.53), i.e.

$$\boxed{\mathbf{w}_{\text{MWF}} = \left(\mathbf{R}_x + \mu \mathbf{R}_v\right)^{-1} \mathbf{R}_x \mathbf{e}_{\text{ref}}\,.} \tag{4.2}$$

In the case of a single target speech source, this reduces to the following optimal filter (cfr. Appendix A.2):

$$\boxed{\mathbf{w}_{\text{opt.}} \;=\; \mathbf{R}_v^{-1}\mathbf{a}\;.\;\frac{P_s\; A_{\text{ref}}^*}{\mu + \rho}\,,} \tag{4.3}$$

with $\mathbf{a}$ the M-dimensional steering vector, $A_{\text{ref}}^* = \mathbf{a}^H \mathbf{e}_{\text{ref}}$ and $\rho = P_s \mathbf{a}^H \mathbf{R}_v^{-1} \mathbf{a}$. Using definition (2.25), the narrowband output SNR is also equal to ρ, i.e.

$$\text{SNR}_{\text{opt.}}^{out} = \rho = P_s \mathbf{a}^H \mathbf{R}_v^{-1} \mathbf{a}\,. \tag{4.4}$$

The speech distortion parameter μ only appears in the single-channel postfilter and fulfills the same role as in single-channel constrained Wiener filters [60,118] or spectral oversubtraction [11,118]. As the filters (4.3), obtained for different values of μ, are only related by a scalar factor, the narrowband output SNR (4.4) is indeed independent of μ. However, larger values of μ allow further attenuation of the residual noise, thus leading to a better listening comfort, at the cost of a higher speech distortion. The fact that the spatial filter is independent of μ is a desirable property. However, it is shown in the following section that this property is lost in an algorithm based on (4.2), with estimation errors in the speech correlation matrix $\mathbf{R}_x$.

Rank one SDW-MWF (R1-MWF)

Formula (4.3) incorporates prior knowledge (single target speech source) to obtain an alternative for the standard SDW-MWF formulation (4.2). It requires explicit estimation (or prior knowledge) of the steering vector $\mathbf{a}$ and the speech power P_s. It is however possible to derive an alternative formulation which only uses the speech and noise second-order statistics [77,160], similar to the standard formulation (4.2).

By rewriting ρ as

$$\rho = P_s\ \mathbf{a}^H \mathbf{R}_v^{-1} \mathbf{a}\ , \tag{4.5}$$

$$= P_s\ \mathrm{Tr}\{\mathbf{R}_v^{-1} \mathbf{a}\mathbf{a}^H\}\ , \tag{4.6}$$

$$= \mathrm{Tr}\{\mathbf{R}_v^{-1} \mathbf{R}_x\}\ , \tag{4.7}$$

where $\mathrm{Tr}\{.\}$ is the trace operator, (4.3) is equivalent to the following rank-one SDW-MWF (R1-MWF):

$$\boxed{\mathbf{w}_{\mathrm{R1-MWF}} = \mathbf{R}_v^{-1} \mathbf{R}_x \mathbf{e}_{\mathrm{ref}}\ .\ \frac{1}{\mu + \mathrm{Tr}\{\mathbf{R}_v^{-1} \mathbf{R}_x\}}} \tag{4.8}$$

Although (4.8) is derived for the special case of a single target speech source, it can also be used when this assumption is not fulfilled (as it only makes use of the speech and noise second-order statistics). Otherwise, for a single target speech source case, it is completely equivalent to (4.3).

Spatial Prediction SDW-MWF (SP-MWF)

The minimum distortion Spatial Prediction MWF (SP-MWF) was discussed in [32], and was originally proposed in [8] under the name Distortionless Multichannel

Wiener Filter. It can be viewed as a frequency-domain version of the spatial-temporal prediction approach [23, 89]. For a single target speech source this filter is theoretically equivalent to the MVDR [21], or the R1-MWF (4.8) where $\mu = 0$.

It is assumed that the M speech components can be related to the speech component in the reference microphone signal, i.e. $X_m = \mathcal{H}_{m,\text{ref}}\ X_{\text{ref}}$,for $m = 1...M$, so that

$$\mathbf{x} = \begin{bmatrix} \mathcal{H}_{1,\text{ref}} \\ \vdots \\ \mathcal{H}_{M,\text{ref}} \end{bmatrix} X_{\text{ref}} = \mathbf{h}\ X_{\text{ref}}\ , \tag{4.9}$$

For a single target speech source, $\mathbf{h}$ is then equal to[1]

$$\mathbf{h}_{\text{opt.}} = \mathbf{a}\frac{1}{A_{\text{ref}}}\ . \tag{4.10}$$

We only make use of the spatial correlations between the speech components, hence only a spatial prediction is performed. The spatial prediction vector $\mathbf{h}$ can be found in the Wiener sense, i.e. by minimizing

$$\min_{\mathbf{h}}\ \mathcal{E}\left\{(\mathbf{x} - \mathbf{h}\ X_{\text{ref}})^H\ (\mathbf{x} - \mathbf{h}\ X_{\text{ref}})\right\} \tag{4.11}$$

which leads to

$$\mathbf{h} = \mathbf{R}_x\mathbf{e}_{\text{ref}}\ \frac{1}{\mathbf{e}_{\text{ref}}^H\mathbf{R}_x\mathbf{e}_{\text{ref}}} \tag{4.12}$$

i.e. one column of the speech correlation matrix is selected and divided by the speech component power in the reference microphone. We can now impose the speech distortion to be zero, which leads to the following constrained optimization problem [32]:

$$\min_{\mathbf{w}} \qquad \mathbf{w}^H\mathbf{R}_v\mathbf{w} \tag{4.13}$$

$$\text{s.t.} \qquad \mathbf{w}^H\mathbf{h} = 1\ . \tag{4.14}$$

It is easily shown that the optimal filter is equal to

$$\mathbf{w} = \mathbf{R}_v^{-1}\mathbf{h}\ .\ \frac{1}{\mathbf{h}^H\mathbf{R}_v^{-1}\mathbf{h}}\ , \tag{4.15}$$

or by plugging (4.12) into (4.15),

$$\mathbf{w} = \mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{e}_{\text{ref}}\ .\ \frac{\mathbf{e}_{\text{ref}}^H\mathbf{R}_x\mathbf{e}_{\text{ref}}}{\text{Tr}\{\mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{e}_{\text{ref}}\mathbf{e}_{\text{ref}}^H\mathbf{R}_x\}} \tag{4.16}$$

[1]The denominator of this formula (and also of subsequent formula's) might decay to 0 at certain frequencies due to speech absence [160]. The denominator values are therefore kept above certain thresholds in a practical implementation.

Compared with the R1-MWF (4.8), expression (4.16) has the same spatial filter, but the single-channel postfilter is different. It is also possible to incorporate a speech distortion parameter μ into (4.16), thereby relaxing the minimum distortion hard constraint. This filter will be referred to as the SP-MWF. By enforcing that the postfilters of the SP-MWF and R1-MWF are equal for a single target source, the speech distortion weighted SP-MWF is obtained as

$$\mathbf{w}_{\mathrm{SP-MWF}} = \mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{e}_{\mathrm{ref}} \cdot \frac{\mathbf{e}_{\mathrm{ref}}^H\mathbf{R}_x\mathbf{e}_{\mathrm{ref}}}{\mu\ \mathbf{e}_{\mathrm{ref}}^H\mathbf{R}_x\mathbf{e}_{\mathrm{ref}} + \mathrm{Tr}\{\mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{e}_{\mathrm{ref}}\mathbf{e}_{\mathrm{ref}}^H\mathbf{R}_x\}} \tag{4.17}$$

For a single target speech source, (4.17) is thus again equivalent to (4.3).

4.3 Impact of speech correlation matrix estimation errors: estimated filters

In this section, the impact of estimation errors in the speech correlation matrix $\mathbf{R}_x$ in algorithms based on (4.2), (4.8) and (4.17), is investigated for a scenario with a single target speech source.

In practice, the correlation matrices $\hat{\mathbf{R}}_v$ and $\hat{\mathbf{R}}_y$ are estimated in a recursive procedure as in (2.16) and (2.17). The speech correlation matrix estimate is found as $\hat{\mathbf{R}}_x = \hat{\mathbf{R}}_y - \hat{\mathbf{R}}_v$. In order to obtain practical algorithms, the estimated correlation matrices $\hat{\mathbf{R}}_v$ and $\hat{\mathbf{R}}_x$ are then plugged into (4.2), (4.8) and (4.17) instead of the theoretical $\mathbf{R}_v$ and $\mathbf{R}_x$.

In the case of a single target speech source, the speech correlation matrix $\mathbf{R}_x$ is a rank-one matrix and given by (2.23), and the SDW-MWF is given by (4.3). However, the estimated speech correlation matrix $\hat{\mathbf{R}}_x$ will be

$$\hat{\mathbf{R}}_x = P_s\mathbf{a}\mathbf{a}^H + \mathbf{\Delta}\ , \tag{4.18}$$

i.e. $\hat{\mathbf{R}}_x$ will be equal to the theoretical rank-one matrix plus a full rank (Hermitian) error matrix $\mathbf{\Delta}$. Formula (4.18) will be plugged into formula's (4.2), (4.8) and (4.17) to analyze the impact of speech correlation matrix estimation errors. Inaccurate estimation of the speech statistics occurs because of several reasons [164]. The speech and noise may be nonstationary, while $\hat{\mathbf{R}}_v$ and $\hat{\mathbf{R}}_y$ are estimated at different moments in time. Speech detection errors made by the VAD will also introduce estimation errors in both the speech and the noise correlation matrices. By using the error model (4.18), the impact of errors can be studied, regardless of the origin of the errors.

It is noted that the impact of estimation errors in the noise correlation matrix can be investigated in a similar fashion. However, simulations indicate that

the narrowband SNR performance degradation caused by these errors is the same for the different filters, and independent of μ, in contrast to the impact of speech correlation matrix estimation errors. Therefore, and also to avoid overly complicated expressions, only speech correlation matrix estimation errors are included in the analysis.

SDW-MWF

By plugging (4.18) into (4.2) and applying the matrix inversion lemma (A.6), the following formula is obtained:

$$\hat{\mathbf{w}} = \left(P_s\mathbf{a}\mathbf{a}^H + \mathbf{\Delta} + \mu\mathbf{R}_v\right)^{-1}\left(P_s\mathbf{a}\mathbf{a}^H + \mathbf{\Delta}\right)\mathbf{e}_{\text{ref}}\ , \tag{4.19}$$

$$= \frac{P_s}{1+\tilde{\rho}}(\mathbf{\Delta} + \mu\mathbf{R}_v)^{-1}\mathbf{a}\ A_{\text{ref}}^* + \left[\mathbf{I} - \frac{P_s}{1+\tilde{\rho}}(\mathbf{\Delta} + \mu\mathbf{R}_v)^{-1}\mathbf{a}\mathbf{a}^H\right](\mathbf{\Delta} + \mu\mathbf{R}_v)^{-1}\mathbf{\Delta}\ \mathbf{e}_{\text{ref}}\ , \tag{4.20}$$

with

$$\tilde{\rho} = P_s\mathbf{a}^H(\mathbf{\Delta} + \mu\mathbf{R}_v)^{-1}\mathbf{a}\ . \tag{4.21}$$

Two limit cases are now considered.

- $\mathbf{\Delta} >> \mu\mathbf{R}_v \implies (\mathbf{\Delta} + \mu\mathbf{R}_v)^{-1} \approx \mathbf{\Delta}^{-1}$
  In this case (4.20) reduces to

  $$\hat{\mathbf{w}} = \mathbf{e}_{\text{ref}}\ . \tag{4.22}$$

  This means that if a small μ parameter is chosen or if the input SNR is high, the estimated SDW-MWF reduces to the trivial filter $\mathbf{e}_{\text{ref}}$ (i.e. pass the reference microphone signal, no noise reduction).

- $\mathbf{\Delta} << \mu\mathbf{R}_v \implies (\mathbf{\Delta} + \mu\mathbf{R}_v)^{-1} \approx \frac{1}{\mu}\mathbf{R}_v^{-1}$ and $\tilde{\rho} \approx \frac{1}{\mu}\rho$
  In this case (4.20) reduces to

  $$\hat{\mathbf{w}} = \mathbf{w}_{\text{opt.}} + \frac{1}{\mu}\left(\mathbf{I} - \frac{P_s}{\mu+\rho}\mathbf{R}_v^{-1}\mathbf{a}\mathbf{a}^H\right)\mathbf{R}_v^{-1}\mathbf{\Delta}\ \mathbf{e}_{\text{ref}}\ , \tag{4.23}$$

  where $\mathbf{w}_{\text{opt.}}$ is given by formula (4.3). Hence, for large values of μ or for a low input SNR, the estimated SDW-MWF will be a combination of the optimal theoretical filter and an extra bias term, which causes performance degradation. A larger μ value (which will result in more speech distortion) puts less weight on the bias term, so that it can be expected that the performance degradation will be smaller for larger μ values. In section 4.4, the output SNR obtained with (4.23) will be calculated.

R1-MWF

By plugging (4.18) into (4.8) and applying the matrix inversion lemma (A.6), the following formula is obtained:

$$\hat{\mathbf{w}} = \frac{\mathbf{R}_v^{-1}\left(P_s\mathbf{a}\mathbf{a}^H + \mathbf{\Delta}\right)\mathbf{e}_{\text{ref}}}{\mu + \text{Tr}\{\mathbf{R}_v^{-1}P_s\mathbf{a}\mathbf{a}^H + \mathbf{R}_v^{-1}\mathbf{\Delta}\}}\ , \tag{4.24}$$

$$= \frac{\mu+\rho}{\mu+\rho+\text{Tr}\{\mathbf{R}_v^{-1}\mathbf{\Delta}\}}\ .\ \mathbf{w}_{\text{opt.}} + \frac{1}{\mu+\rho+\text{Tr}\{\mathbf{R}_v^{-1}\mathbf{\Delta}\}}\ .\ \mathbf{R}_v^{-1}\mathbf{\Delta}\mathbf{e}_{\text{ref}}\ , \tag{4.25}$$

$$= \left(P_s\mathbf{R}_v^{-1}\mathbf{a}A_{\text{ref}}^* + \mathbf{R}_v^{-1}\mathbf{\Delta}\mathbf{e}_{\text{ref}}\right)\frac{1}{\mu+\rho+\text{Tr}\{\mathbf{R}_v^{-1}\mathbf{\Delta}\}}\ . \tag{4.26}$$

Equation (4.25) shows that the estimated SDW-MWF can still be written as a combination of the optimal filter (4.3) and a bias term. Equation (4.26) shows that the estimated filter can also be written as a spatial filter followed by a single-channel postfilter, where the parameter μ only occurs in the single-channel postfilter. Therefore, μ will not influence the obtained narrowband (single-frequency) output SNR, as in the optimal case (4.4).

SP-MWF

By plugging (4.18) into (4.12) and (4.17), the following formulas are obtained:

$$\hat{\mathbf{h}} = \frac{1}{\tilde{P}_s}\ .\ \left(P_s\mathbf{a}A_{\text{ref}}^* + \mathbf{\Delta}\mathbf{e}_{\text{ref}}\right) \tag{4.27}$$

and

$$\hat{\mathbf{w}} = \left(P_s\mathbf{R}_v^{-1}\mathbf{a}A_{\text{ref}}^* + \mathbf{R}_v^{-1}\mathbf{\Delta}\mathbf{e}_{\text{ref}}\right) \times\frac{\tilde{P}_s}{\mu\tilde{P}_s + \rho P_s|A_{\text{ref}}|^2 + \rho' P_s A_{\text{ref}} + (\rho')^* P_s A_{\text{ref}}^* + \rho''}\ , \tag{4.28}$$

where

$$\tilde{P}_s = P_s|A_{\text{ref}}|^2 + \mathbf{e}_{\text{ref}}^H\mathbf{\Delta}\mathbf{e}_{\text{ref}}\ , \tag{4.29}$$

$$\rho' = \mathbf{a}^H\mathbf{R}_v^{-1}\mathbf{\Delta}\mathbf{e}_{\text{ref}}\ , \tag{4.30}$$

$$\rho'' = \mathbf{e}_{\text{ref}}^H\mathbf{\Delta}^H\mathbf{R}_v^{-1}\mathbf{\Delta}\mathbf{e}_{\text{ref}}\ . \tag{4.31}$$

The estimated SP-MWF (4.28) can thus be written as a spatial filter followed by a single-channel postfilter. The spatial filter in (4.28) is equal to the spatial filter in (4.26), so that the estimated SP-MWF will obtain the same narrowband (single-frequency) output SNR as the estimated R1-MWF.

4.4 Impact of speech correlation matrix estimation errors: narrowband output SNR

In this section, the impact of speech correlation matrix estimation errors on the obtained output SNR will be analyzed. The estimated filters, derived in the previous section, are plugged into the narrowband output SNR definition (2.25). The rank-one model (2.23) is again used for the speech correlation matrix.

SDW-MWF

Again, the two limit cases of the previous section are considered.

For the case $\mathbf{\Delta} >> \mu\mathbf{R}_v$, the trivial filter $\mathbf{e}_{\text{ref}}$ is obtained, so that $\widehat{\text{SNR}}^{out} = \text{SNR}^{in}$.

For the case $\mathbf{\Delta} << \mu\mathbf{R}_v$, formula (4.20) can be plugged into the definition of the narrowband output SNR (2.25) to obtain:

$$\widehat{\text{SNR}}^{out}$$

$$= \frac{\frac{1}{(\mu+\rho)^2}\left(P_s|A_{\text{ref}}|^2\rho^2 + P_s A_{\text{ref}}(\rho^{'})\rho + P_s A_{\text{ref}}^*(\rho^{'})^*\rho - \frac{2\mu+\rho}{\mu^2}P_s|\rho^{'}|^2\rho\right) + \frac{1}{\mu^2}P_s|\rho^{'}|^2}{\frac{1}{(\mu+\rho)^2}\left(P_s|A_{\text{ref}}|^2\rho + P_s A_{\text{ref}}(\rho^{'}) + P_s A_{\text{ref}}^*(\rho^{'})^* - \frac{2\mu+\rho}{\mu^2}P_s|\rho^{'}|^2\right) + \frac{1}{\mu^2}\rho^{''}},$$

$$= \rho - \frac{\rho\rho^{''} - P_s|\rho^{'}|^2}{\frac{\mu^2}{(\mu+\rho)^2}\left(P_s|A_{\text{ref}}|^2\rho + P_s A_{\text{ref}}(\rho^{'}) + P_s A_{\text{ref}}^*(\rho^{'})^*\right) - \frac{2\mu+\rho}{(\mu+\rho)^2}P_s|\rho^{'}|^2 + \rho^{''}}, \quad (4.32)$$

$$= \rho -$$

$$\frac{\rho\rho^{''} - P_s|\rho^{'}|^2}{\frac{\mu^2}{(\mu+\rho)^2}\left[P_s|A_{\text{ref}}|^2\rho + P_s A_{\text{ref}}(\rho^{'}) + P_s A_{\text{ref}}^*(\rho^{'})^* + \rho^{''} + \frac{2\mu+\rho}{\mu^2}(\rho\rho^{''} - P_s|\rho^{'}|^2)\right]}, \quad (4.33)$$

where $\rho^{'}$ and $\rho^{''}$ were defined in (4.30) and (4.31). Formula (4.33) shows that the obtained narrowband output SNR is equal to the optimal narrowband output SNR (4.4) minus a bias term. By defining matrix $\tilde{\mathbf{V}}$ such that $\tilde{\mathbf{V}}\tilde{\mathbf{V}}^H = \mathbf{R}_v^{-1}$, the bias term can also be written as

$$\frac{\frac{(\mu+\rho)^2}{\mu^2}(\rho\rho^{''} - P_s|\rho^{'}|^2)}{||P_s A_{\text{ref}}\mathbf{a}^H\tilde{\mathbf{V}} + \mathbf{e}_{\text{ref}}^H\mathbf{\Delta}^H\tilde{\mathbf{V}}||^2 + \frac{2\mu+\rho}{\mu^2}(\rho\rho^{''} - P_s|\rho^{'}|^2)}, \quad (4.34)$$

which clearly shows that the bias term is always positive if $\left(\rho\rho^{''} - P_s|\rho^{'}|^2\right)$ is positive. By using the Cauchy-Schwarz inequality, it can indeed be shown that $\left(\rho\rho^{''} - P_s|\rho^{'}|^2\right)$ is positive[2]:

$$\begin{aligned}\rho\rho^{''} &= P_s\left(\mathbf{a}^H\mathbf{R}_v^{-1}\mathbf{a}\right)\left(\mathbf{e}_{\text{ref}}^H\mathbf{\Delta}^H\mathbf{R}_v^{-1}\mathbf{\Delta}\mathbf{e}_{\text{ref}}\right) , \\ &\geq P_s|\mathbf{a}^H\mathbf{R}_v^{-1}\mathbf{\Delta}\mathbf{e}_{\text{ref}}|^2 = P_s|\rho^{'}|^2 . \end{aligned} \tag{4.35}$$

These results indicate that there will always be a performance degradation, which is moreover dependent on μ. As can be seen from (4.32), the denominator of the bias term strictly increases when μ increases:

$$\begin{aligned}&\frac{\partial}{\partial\mu}\left(\frac{\mu^2}{(\mu+\rho)^2}\left[P_s|A_{\text{ref}}|^2\rho + P_sA_{\text{ref}}(\rho^{'}) + P_sA_{\text{ref}}^*(\rho^{'})^*\right] - \frac{2\mu+\rho}{(\mu+\rho)^2}P_s|\rho^{'}|^2 + \rho^{''}\right) \\ = &\quad \frac{2\mu\rho}{(\mu+\rho)^3}\left[P_s|A_{\text{ref}}|^2\rho + P_sA_{\text{ref}}(\rho^{'}) + P_sA_{\text{ref}}^*(\rho^{'})^*\right] + \frac{2\mu}{(\mu+\rho)^3}P_s|\rho^{'}|^2 \\ = &\quad \frac{2\mu}{(\mu+\rho)^3}P_s|\rho A_{\text{ref}} + (\rho^{'})^*|^2 \ > 0 . \end{aligned} \tag{4.36}$$

Therefore, the obtained narrowband output SNR strictly increases as μ increases. By calculating the limit for $\mu \to \infty$, i.e.

$$\lim_{\mu\to\infty}\widehat{\text{SNR}}^{out} = \rho - \frac{\rho\rho^{''} - P_s|\rho^{'}|^2}{P_s|A_{\text{ref}}|^2\rho + P_sA_{\text{ref}}(\rho^{'}) + P_sA_{\text{ref}}^*(\rho^{'})^* + \rho^{''}} , \tag{4.37}$$

it is apparent that there is a maximum obtainable output SNR, which is equal to the theoretical narrowband output SNR minus a fixed bias term. Thus, to achieve this maximum narrowband output SNR with the SDW-MWF, a large μ value should be used. This can however introduce too much speech distortion. Using a small μ value (low distortion) can also be unsatisfactory, as this will cause SNR performance degradation, especially so when $\mathbf{\Delta} >> \mu\mathbf{R}_v$.

R1-MWF and SP-MWF

As was stated previously, the filters (4.26) and (4.28) consist of a spatial filter followed by a single-channel postfilter. The obtained narrowband output SNR only depends on the spatial filter part (which is independent of μ). As both filters have the same spatial filter, they can thus be treated together.
By plugging the spatial filter

$$\hat{\mathbf{w}} = P_s\mathbf{R}_v^{-1}\mathbf{a}A_{\text{ref}}^* + \mathbf{R}_v^{-1}\mathbf{\Delta}\mathbf{e}_{\text{ref}} \tag{4.38}$$

[2]The case $\rho\rho^{''} = P_s|\rho^{'}|^2$ occurs if $\mathbf{\Delta}\mathbf{e}_{\text{ref}}$ is a scaled version of the steering vector $\mathbf{a}$. Thus, if the speech correlation matrix is known up to a scaling factor, the optimal performance can be achieved.

into the narrowband output SNR definition (2.25), the following output SNR is obtained:

$$\widehat{\mathrm{SNR}}^{out} = \rho - \frac{\rho\rho^{''} - P_s|\rho^{'}|^2}{P_s|A_{\mathrm{ref}}|^2\rho + P_s A_{\mathrm{ref}}(\rho^{'}) + P_s A_{\mathrm{ref}}^{*}(\rho^{'})^{*} + \rho^{''}} \; . \tag{4.39}$$

Remarkably, this narrowband output SNR is equal to the limit case (4.37) achieved by the SDW-MWF for large values of μ.

Discussion

Under speech correlation matrix estimation errors, and for a scenario with a single target speech source, the R1-MWF (4.8) and SP-MWF (4.17) always achieve the limit narrowband output SNR (4.37) (for any value of μ), whereas the SDW-MWF (4.2) will only achieve this for large μ. In scenarios where only a single target speech source is present and where only limited speech distortions (small μ's) are allowed, the SDW-MWF is therefore outperformed by the other filters. In particular, in order to obtain a similar narrowband output SNR, the SDW-MWF will introduce a higher speech distortion. Experimental results will show that even for moderate values of μ (for example the standard MMSE cost function, $\mu = 1$), a significant performance degradation occurs when using the SDW-MWF. This was also observed in [32], where the SP-MWF (with $\mu = 0$) clearly outperformed the SDW-MWF in both SNR and speech distortion. These practical results are now better explained by the above theoretical analysis.

For a scenario with multiple target speech sources (i.e. speech correlation matrix not a rank-one matrix), the R1-MWF and SP-MWF formulas are no longer theoretically equivalent to the general SDW-MWF formula. A theoretical analysis as for the rank-one case is not straightforward, but it can be expected that there will again be a performance degradation. To investigate this, the next section includes simulations with multiple target speech sources. These simulations show that the R1-MWF and SP-MWF still obtain superior SNRs compared to the SDW-MWF. The R1-MWF will introduce more speech distortion however, especially when the number of target sources is large. Remarkably, the simulations also show that the SP-MWF only introduces distortion in the less dominant target sources. As a consequence, the overall distortion (on the sum of all speech components) will still be low for this filter.

4.5 Experimental results

4.5.1 Setup and stimuli

In the simulations, both a bilateral and a binaural hearing aid configuration are considered. For the binaural configuration, it is assumed that the link imposes no restrictions in terms of bandwidth and power consumption. We therefore assume that all microphone signals are available to the noise reduction procedure, where two microphones are at the left ear and two at the right ear, giving a total of $M = 4$ microphones[3]. The left front and right front microphones are chosen as reference microphones to generate the left and right filters and output signals.

Head-related transfer functions (HRTFs) were measured in two acoustic environments (reverberation times $\mathrm{RT}_{60} = 0.21$s and $\mathrm{RT}_{60} = 0.61$s, [176,177]) on a dummy-head, so that the head-shadow effect is taken into account. The BTE hearing aids have two omnidirectional microphones on each device, with an intermicrophone distance of approximately 1 cm. To generate the microphone signals, the noise and speech signals are convolved with the HRTFs corresponding to their angles of arrival, before being added together.

In all experiments, multi-talker babble (Auditec [6]) was used as noise signal. In the scenarios with multiple noise sources, different time-shifted versions of this signal were generated to obtain uncorrelated noise sources. In the single target speech source scenarios, a signal consisting of 6 instances of speech-shaped noise with periods of silence, was used as target signal (12 seconds of speech, total length 24 seconds), as in [176, 177]. The speech-shaped noise was obtained from the average spectrum of a Dutch male speaker of the VU test material [184]. In the multiple target speech sources scenarios, different lists of the VU test material were used for the different target signals, with each signal consisting of 4 sentences with periods of silence (total length 26 seconds). The different target speech sources are simultaneously active so that the SNR and distortion can be calculated on the global target signal, i.e. the sum of the speech components.

In all experiments, the signals are sampled at $f_s = 20480$ Hz, and the filter length (= DFT size) is $L = 128$. A batch algorithm[4] was implemented where the speech+noise and noise correlation matrices are estimated off-line using the complete microphone signals, i.e. for the speech+noise correlation matrix:

$$\hat{\mathbf{R}}_y = \frac{1}{K} \sum_{k=1}^{K} \mathbf{y}[k]\ \mathbf{y}^H[k]\ , \tag{4.40}$$

[3]We again note that we now consider a single generic microphone array consisting of M microphones, in contrast to the configuration in section 2.1.2.

[4]The R1-MWF and SP-MWF can also be implemented by an adaptive QRD-RLS algorithm as is discussed in [34] and chapter 6.

where k is the frame-index and where the summation only counts the frames where both speech and noise are active. The microphone signals were hereby cut into frames of L samples with 50% overlap, and windowed by a Hann window. The value K in (4.40) is set to the total number of frames where speech is active. The noise correlation matrix is calculated on the noise-only frames in a similar manner. The VAD is hereby assumed to be perfect. The filters (4.2), (4.8) and (4.17) are calculated using the estimated correlation matrices, as explained in section 4.3. Following the approach in [154], the frequency-domain filters are then transformed into corresponding time-domain filters which do not rely on the circularity effect. An extra phase shift is also applied in the IDFT formula as in [154], so that the resulting filters are causal.

The input SNR (measured on the clean signals, i.e. at the loudspeaker) is 0 dB in all experiments. Due to the headshadow effect, the input SNRs at the left and right reference microphones then depend on the spatial scenario, i.e. the positions of the target speech source(s) and noise source(s).

4.5.2 Narrowband SNR improvement

In a first experiment, the narrowband SNR improvement (2.26) is calculated for the different filters. The fifth frequency bin (corresponding to 640 Hz) is selected for the performance comparison. At this frequency, the speech (and noise) signals contain an intermediate amount of energy (see [176] for the average input spectra), while this frequency also has an intermediate importance for speech intelligibility [5]. A single noise source is placed at 120° (with 0° the front direction, 90° to the right of the head), the single target speech source is placed in front of the head at 0°.

The narrowband input and output SNRs are calculated by plugging the resulting filters into expressions (2.24) and (2.25), which make use of the (theoretical) rank one speech correlation matrix (2.23). The steering vector $\mathbf{a}$ is therefore constructed using the HRTF data, and P_s is estimated by calculating the PSD of the clean speech signal. The steering vector is also used to calculate the theoretical optimal filter (4.3). While the steering vector is calculated using a large DFT size, the noise correlation matrix used in the optimal filter is the same as for the other filters, and calculated at $L = 128$. In order to quantify the speech correlation matrix estimation error, the following error measure is proposed:

$$\delta(\hat{\mathbf{R}}_x) = \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \frac{|\mathbf{\Delta}_{ij}|}{|\mathbf{R}_{x,ij}|} \,, \tag{4.41}$$

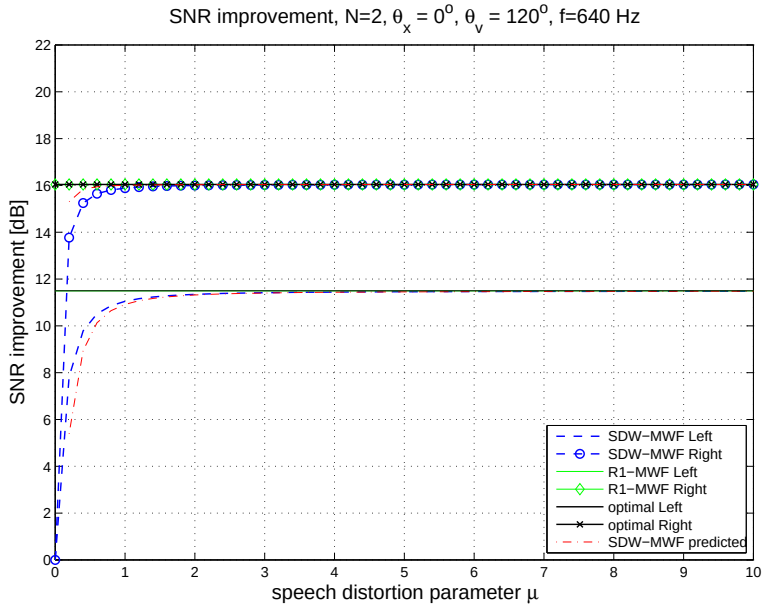
where $\mathbf{\Delta}_{ij}$ is the entry in the ith row and jth column of matrix $\mathbf{\Delta}$, and similarly for $\mathbf{R}_{x,ij}$. For a bilateral setup (i.e. with only 2 used microphones, no binaural link), the estimation errors (average error of the left and right correlation matrix

estimates) are then equal to $\delta(\hat{\mathbf{R}}_x) = 0.01$ for the low-reverberant environment and $\delta(\hat{\mathbf{R}}_x) = 0.05$ for the high-reverberant environment. For a binaural setup (all 4 microphone signals are available), the estimation errors are equal to $\delta(\hat{\mathbf{R}}_x) = 0.07$ for the low-reverberant environment, and $\delta(\hat{\mathbf{R}}_x) = 0.41$ for the high-reverberant environment. In a high-reverberant environment and for a binaural setup, a significant amount of estimation errors are thus introduced, so that a performance degradation can be expected.

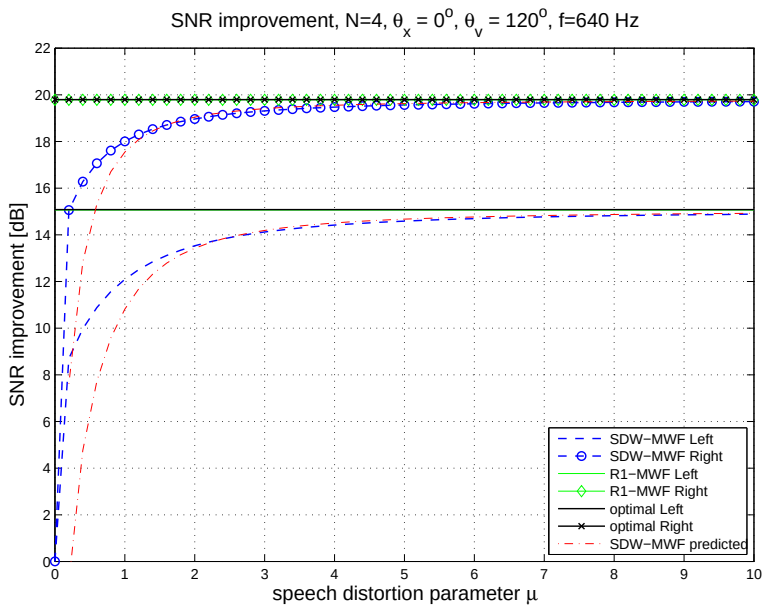
Figures 4.1 (low reverberation) and 4.2 (high reverberation) show the SNR improvements obtained at the left and the right side as a function of the speech distortion parameter μ, for the bilateral and binaural setup. As discussed in section 4.4, there is a fixed performance degradation between the *optimal* performance and the performance of the *R1-MWF* (4.39), especially so for the high-reverberant environment (Figure 4.2). The SP-MWF obtains the same SNR improvement as the R1-MWF, and is therefore omitted from the figures. As was explained, the trade-off parameter μ does not influence the narrowband SNR improvement for the R1-MWF and SP-MWF. For the low-reverberant environment (Figure 4.1), the estimation errors are small so that the R1-MWF performance is very close to the optimal performance. In the high-reverberant environment (Figure 4.2), more errors are introduced so that the R1-MWF performance is visibly degraded compared to the optimal performance.

In contrast to the R1-MWF and SP-MWF, the performance of the *SDW-MWF* (4.2) does depend on μ. For small values of μ, the filter degrades to the trivial filter (no SNR improvement), while for larger values it converges towards the performance of the R1-MWF. The performance degradation for small values of μ is significant for both the low-reverberant and the high-reverberant environment. As a check, the *predicted* output SNR for larger values of μ (4.33) was also calculated and shown on the figures.

When comparing the 4-microphone (binaural) setup to the 2-microphone (bilateral) setup, it can be seen that the SNR improvement is significantly higher when using a binaural setup. By using more microphone signals, a better noise reduction performance is thus effectively achieved. However, it can also be seen that the performance degradation versus the optimal filter (and of the SDW-MWF versus the R1-MWF) is larger for the binaural setup, especially so in the high-reverberant environment. As larger correlation matrices have to be estimated when the number of microphones increases, more estimation errors are introduced, thus leading to a larger degradation.

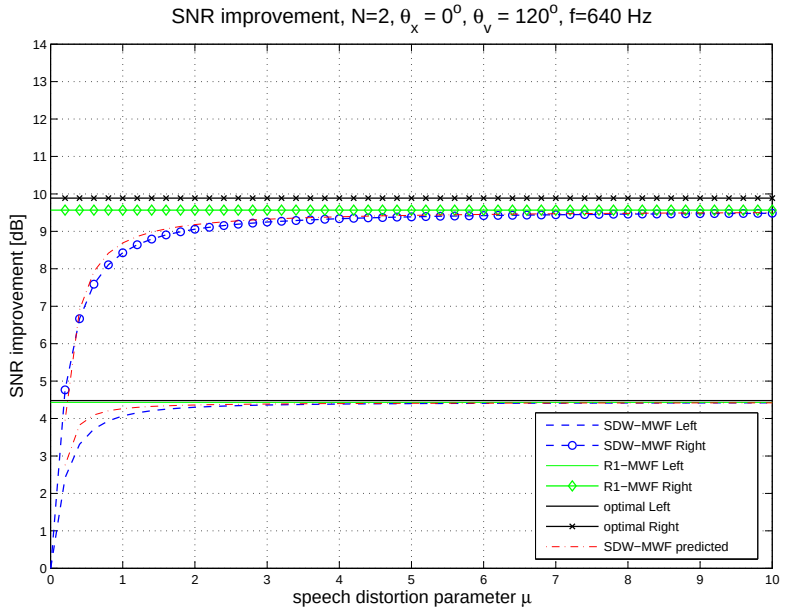


(a) Bilateral setup (2 microphones, no binaural link)

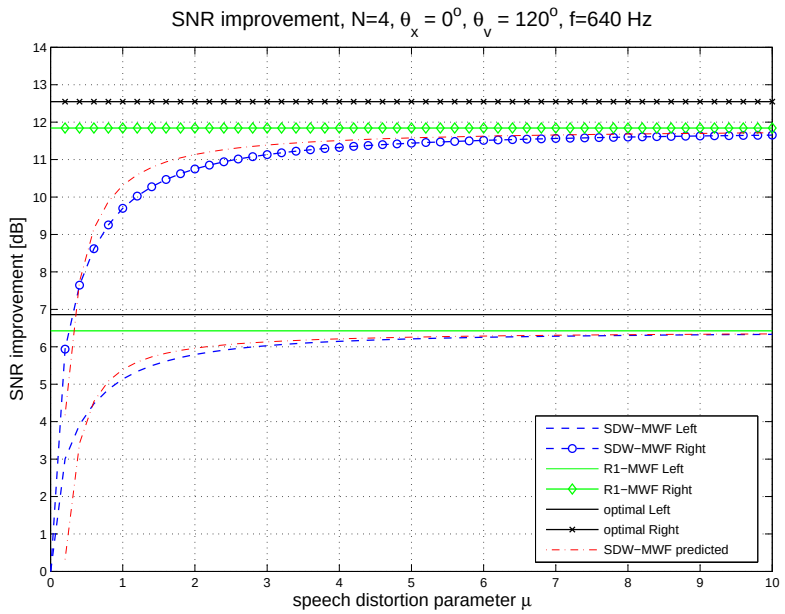


(b) Binaural setup (4 microphones)

Figure 4.1: Narrowband SNR improvement as a function of μ for SDW-MWF (4.2) and R1-MWF (4.8); $RT_{60} = 0.21$s (low reverberation); (a) Bilateral setup (2 microphones, no binaural link), (b) Binaural setup (4 microphones)



(a) Bilateral setup (2 microphones, no binaural link)



(b) Binaural setup (4 microphones)

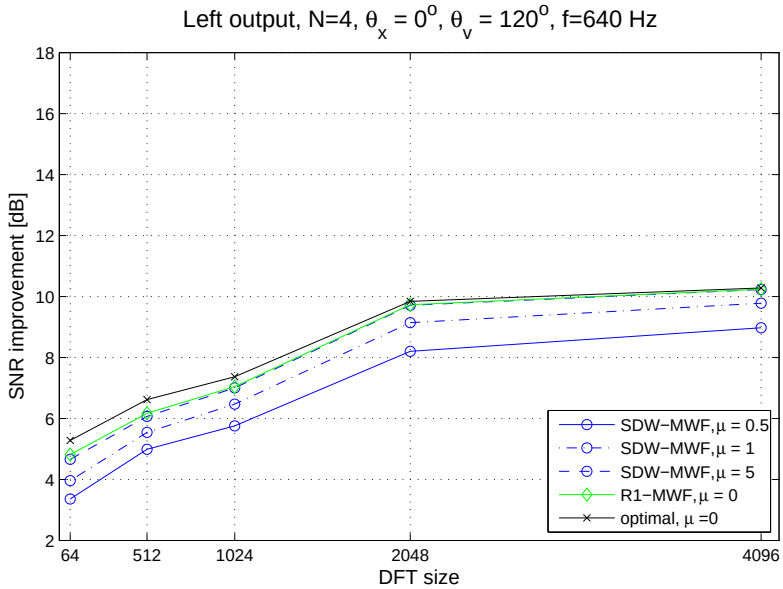
Figure 4.2: Narrowband SNR improvement as a function of μ for SDW-MWF (4.2) and R1-MWF (4.8); $\mathrm{RT}_{60} = 0.61$s (high reverberation); (a) Bilateral setup (2 microphones, no binaural link) , (b) Binaural setup (4 microphones)

Remarkably, for the MSE cost function ($\mu = 1$), the difference between the standard SDW-MWF and R1-MWF is quite large for the binaural setup. In other work using the standard SDW-MWF, a value of $\mu = 5$ was therefore usually selected (e.g. [177] and chapter 3). Figures 4.1 and 4.2 indeed illustrate that this is a sound choice: larger values of μ will not yield much narrowband SNR improvements anymore, but merely introduce more speech distortion.

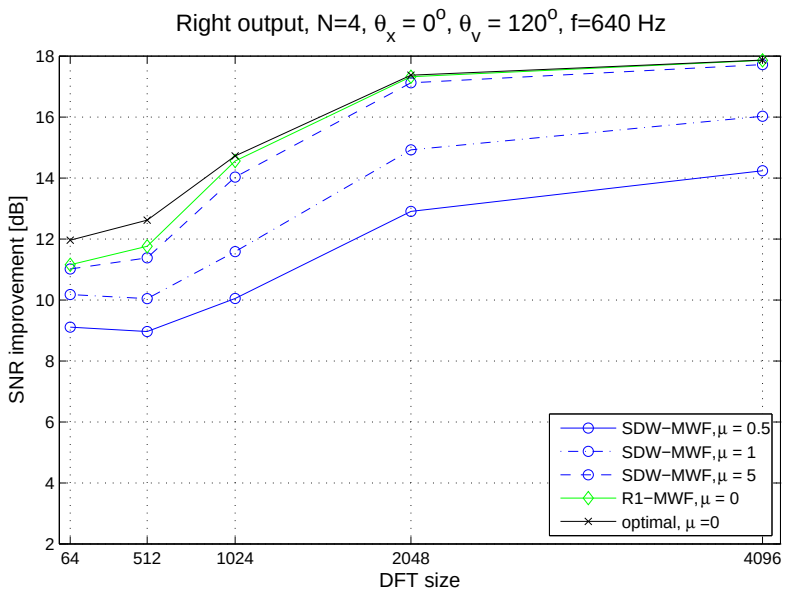
While larger filter lengths (or DFT sizes) than $L = 128$ could be inappropriate for a binaural hearing aid application due to a prohibitive computational complexity and input-output latency, it is interesting from a theoretical point of view to study the effect of increasing L. In Figure 4.3, the left and right narrowband SNR improvements for the 4-microphone setup are shown as a function of L. As before, the appropriate frequency bin (corresponding to 640 Hz) is selected for the performance evaluation, and the results for the SP-MWF are omitted.

As the speech and noise correlation matrices are estimated with a better resolution when L increases, the performances of the different filters (also of the optimal filter) improve. It can also be observed that the performance of the R1-MWF gets closer to the optimal performance at larger values of L. This is due to the fact that the rank-one assumption for the speech correlation matrix is then better satisfied: for $L = 64$, the ratio of the dominant eigenvalue and second-highest eigenvalue is only 2.6, while for $L = 4096$ this ratio is equal to 30.4. The $\hat{\mathbf{R}}_x$ estimation error quantified by (4.41) goes down from $\delta(\hat{\mathbf{R}}_x) = 0.48$ for $L = 64$ to $\delta(\hat{\mathbf{R}}_x) = 0.22$ for $L = 4096$. The remaining errors at $L = 4096$ are mostly scaling errors, which do not degrade the performance according to (4.39), i.e. the numerator $(\rho\rho^{''} - P_s|\rho^{'}|^2)$ is almost equal to zero.

Finally, although the estimated $\hat{\mathbf{R}}_x$ better approximates the ideal $\mathbf{R}_x$ for higher values of L (as was just discussed), Figure 4.3 illustrates that the performance degradation of the SDW-MWF (4.2) is still significant for lower values of μ. In conclusion, even if larger filterlengths could be applied, the R1-MWF still offers a significant benefit over the standard SDW-MWF.



(a) Left output



(b) Right output

Figure 4.3: Narrowband SNR improvement as a function of DFT size for SDW-MWF (4.2) and R1-MWF (4.8); binaural setup (4 microphones); $RT_{60} = 0.61$s (high reverberation); (a) Left output, (b) Right output

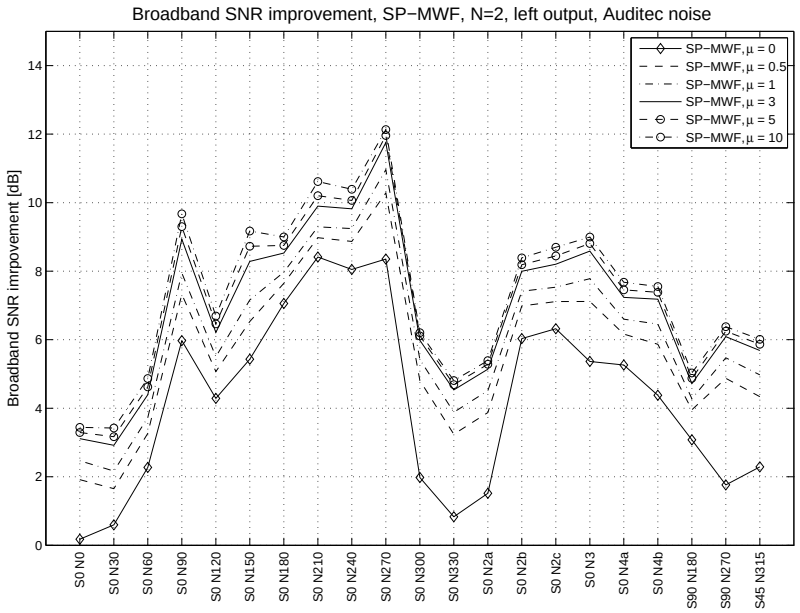
4.5.3 Broadband SNR improvement: effect of the single-channel postfilter

This section features more elaborate experiments for 21 different spatial scenarios, summarized in Table 4.1.

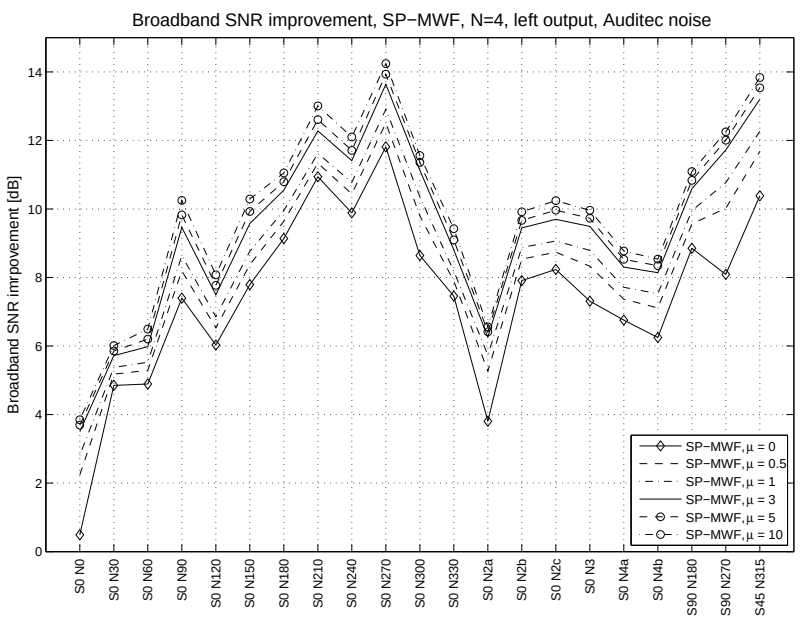
Notation	Description
S0 Nx	target at 0°, single noise source at x° (0° : 30° : 330°)
S0 N2a	target at 0°, noise sources at -60° and 60°
S0 N2b	target at 0°, noise sources at -120° and 120°
S0 N2c	target at 0°, noise sources at 120° and 210°
S0 N3	target at 0°, noise sources at 90°, 180° and 270°
S0 N4a	target at 0°, noise sources at 60°, 120°, 180° and 210°
S0 N4b	target at 0°, noise sources at 60°, 120°, 180° and 270°
S90 N180	target at 90°, single noise source at 180°
S90 N270	target at 90°, single noise source at 270°
S45 N315	target at 45°, single noise source at 315°

Table 4.1: Single target source, spatial scenarios

To illustrate the effect of the single-channel SDW postfilter, the broadband SNR improvement (2.29) is calculated in this experiment. As in [154], the frequency-domain filters are transformed into causal time-domain filters. The (time-domain) speech and noise components of the microphone signals are then filtered separately in order to calculate the broadband SNR. Figure 4.4 shows the broadband SNR improvements obtained by the SP-MWF for a bilateral setup with only 2 used microphones and a binaural setup where all 4 microphones are used to generate the output. Only the left broadband output SNR is shown as the right broadband output SNR leads to similar conclusions. The performance of the SP-MWF is now shown, while the R1-MWF is omitted as it provides similar results. In section 4.5.2 it was already shown that the narrowband SNR performance obtained with the SDW-MWF is dependent on μ. Its broadband SNR improvement is therefore also dependent on μ. As in the narrowband case, it is consistently outperformed by the R1-MWF and SP-MWF, i.e. for a given μ, the broadband SNR improvement is smaller than the SNR improvements of the R1-MWF and SP-MWF. As this performance degradation will also be illustrated in the next section (where intelligibility weighted SNRs are calculated), these results are omitted here.



(a) Bilateral setup (2 microphones, no binaural link)



(b) Binaural setup (4 microphones)

Figure 4.4: Broadband SNR improvement for SP-MWF; left output, RT_{60} = 0.61s, Multi-talker babble noise (Auditec); (a) Bilateral setup (2 microphones, no binaural link) , (b) Binaural setup (4 microphones)

The postfilter attenuates frequency bins with significant residual noise (the denominator of (4.3) is small), so that the broadband output SNR is increased by the postfilter. Larger values of μ lead to larger broadband SNR improvements, as intuitively, the increase of μ globally attenuates the noise power at a higher rate than the target signal [160]. This broadband SNR increase comes at the cost of a higher speech distortion as will be illustrated in the next section. When comparing the 2-microphone to the 4-microphone results, it can be seen that the relative broadband SNR increase by postfiltering is larger in the 2-microphone case for a few particular spatial scenarios (for example S0N30, S0N330). When the angle between speech and noise sources is small, a closely spaced 2-microphone array cannot form a sufficiently narrow beam, so that the performance of the spatial filter is insufficient (see also [177]). In these scenarios, the postfilter can help increase the broadband SNR performance to a larger extend than in the 4-microphone (binaural) case.

As the average spectra of the multi-talker babble noise and speech-shaped VU noise overlap to a great extent [176], the achievable performance increase by postfiltering is limited. For non-overlapping spectra, the improvement by postfiltering can be much larger (see for example [160], where white Gaussian noise was used as interfering noise signal).

4.5.4 SI-weighted SNR improvement and distortion

To assess speech intelligibility improvements, the broadband speech intelligibility (SI) weighted measures (2.30) are calculated. As for the broadband SNR, the (time-domain) speech and noise components of the microphone signals are filtered separately. The signals are then filtered by one-third octave band bandpass filters, and the SNR is calculated per band. Similarly, the intelligibility weighted distortion measure (2.31) is used to quantify the speech distortion. A distortion value of 0 dB corresponds to an undistorted signal, while larger distortion values correspond to more introduced speech distortion.

The SI-weighted SNR improvements for the left output are shown in Figure 4.5, where the spatial scenarios of Table 4.1 are again tested. For the R1-MWF and SP-MWF, the SI-weighted SNR does not depend strongly on μ. This is actually expected as it was shown that the narrowband output SNR, which determines speech intelligibility, is independent of μ for the R1-MWF and SP-MWF. The speech distortion parameter can therefore be set to $\mu = 0$ in this experiment, while keeping in mind that larger values can be used if a larger (not SI-weighted) broadband SNR, or thus listening comfort, is required (cfr. section 4.5.3). The SDW-MWF (4.2) needs larger values of μ in order to achieve the same SNR performance as the other filters, which is in accordance with the theoretical analysis.

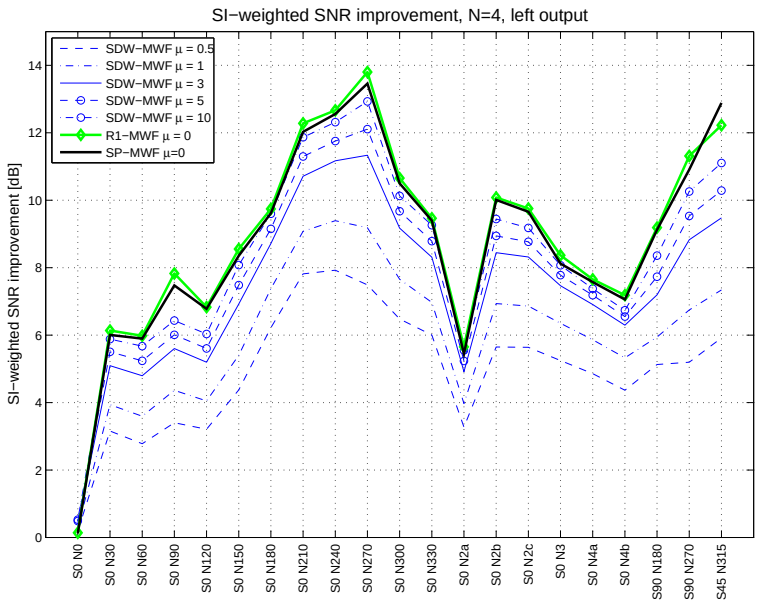


Figure 4.5: SI-weighted SNR improvement for SDW-MWF, R1-MWF, SP-MWF; left output, $\mathrm{RT}_{60} = 0.61$s, 4 microphones

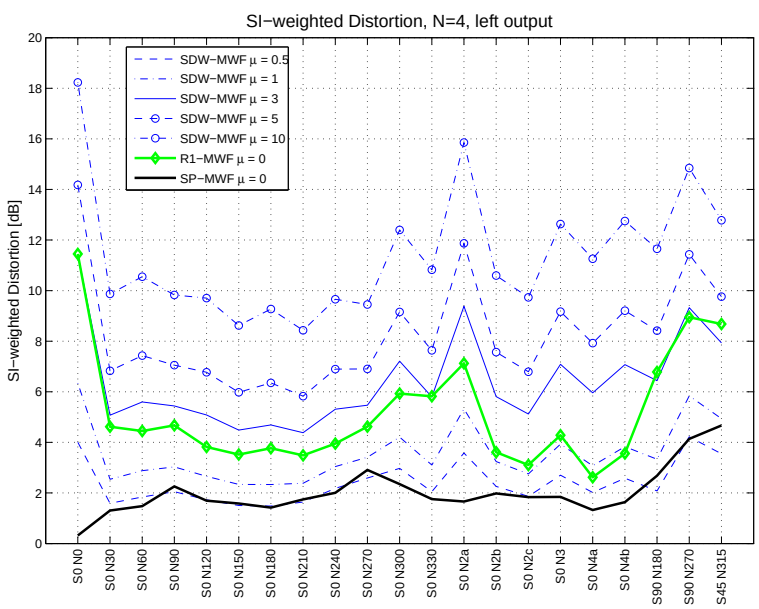


Figure 4.6: SI-weighted Distortion for SDW-MWF, R1-MWF, SP-MWF; left output, $\mathrm{RT}_{60} = 0.61$s, 4 microphones

The introduced SI-weighted distortion is shown in Figure 4.6. If the SDW-MWF is used, a value of $\mu = 5$ (or even larger for some scenarios) should be used in order to achieve the same SNR performance as the other filters. Figure 4.6 shows that this will introduce more speech distortion compared to the other filters (where $\mu = 0$ can be used). Another remarkable observation is that the R1-MWF and SP-MWF, which are theoretically equivalent, introduce different speech distortions. This is due to the fact that their postfilters are different if estimation errors are present (section 4.3). Various simulations have shown that for the same value of μ, the SP-MWF is more conservative than the R1-MWF, thereby introducing less speech distortion.

4.5.5 Multiple target speech sources

The SI-weighted performance of the different filters is now tested in multiple target speech sources scenarios. That is, there are multiple desired speech signals in different directions that should be preserved by the filter. As a consequence, the performance of the R1-MWF and SP-MWF (which implicitly assume a single target source) is expected to degrade.

Sixteen different spatial scenarios are tested, cfr. Table 4.2 for the used notation.

Notation	**Description**
S1	1 target at -30°
S2	2 targets at -30° and 45°
S3	3 targets at -30°, 45° and 150°
S4	4 targets at -90°, -30°, 45° and 150°
N1a	1 noise source at 0°
N1b	1 noise source at -90°
N2	2 noise sources at -60° and 120°
N4	4 noise sources at -90°, -60°, 120°, 180°

Table 4.2: Multiple target sources, Spatial scenarios (all combinations of S and N are made).

The obtained SI-weighted SNR improvements are shown in Figure 4.7. The R1-MWF and SP-MWF still outperform the SDW-MWF, even for scenarios with 4 target speech sources (full-rank speech correlation matrix). There is no noticeable difference between the performance of the R1-MWF and SP-MWF. Again, as larger values of μ do not affect the SI-weighted SNR improvements of the R1-MWF and SP-MWF, only the performance for $\mu = 0$ is shown. Larger values can however be used to increase the broadband SNR as was shown in section 4.5.3.

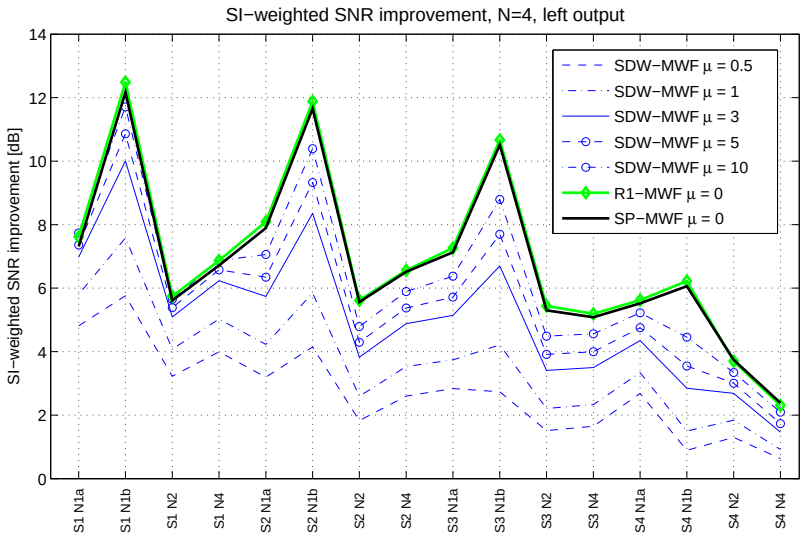


Figure 4.7: SI-weighted SNR improvement for SDW-MWF, R1-MWF, SP-MWF; left output, $\mathrm{RT}_{60} = 0.61$s, 4 microphones

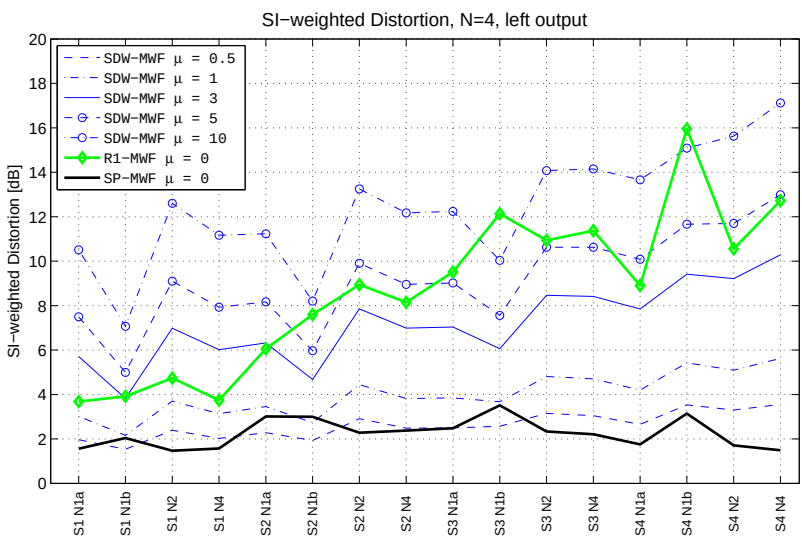
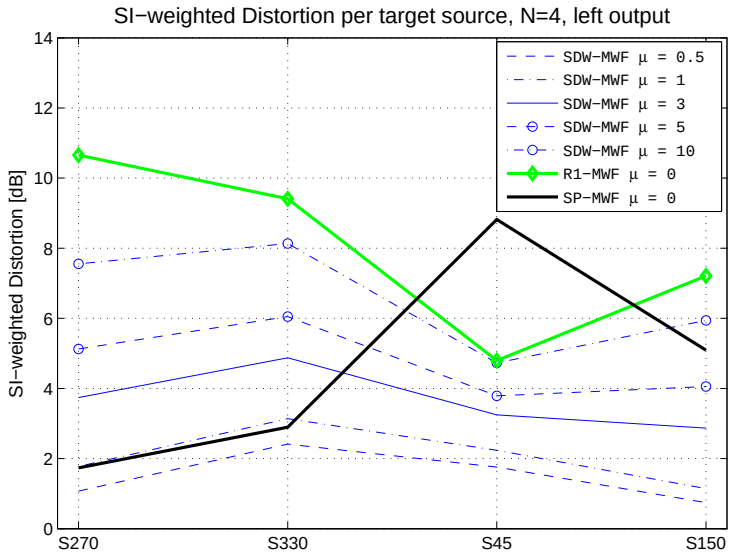


Figure 4.8: SI-weighted Distortion for SDW-MWF, R1-MWF, SP-MWF; left output, $\mathrm{RT}_{60} = 0.61$s, 4 microphones

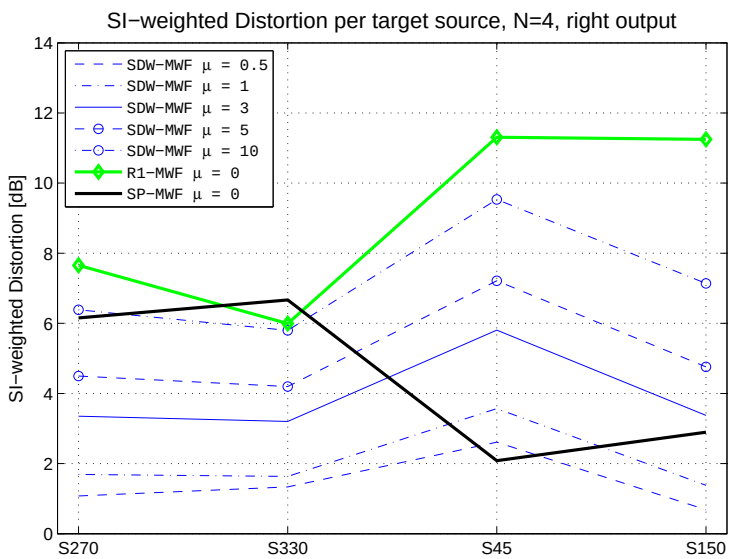
The SI-weighted distortion is shown in Figure 4.8. Again, the distortion increases as μ increases. For the R1-MWF, a performance degradation can be observed: as the number of target sources increases, more speech distortion is introduced. Even for $\mu = 0$, a large increase in distortion can be observed. The violation of the rank-one assumption will thus introduce speech distortion, but not affect the SNR performance. Remarkably, the SP-MWF does not seem to introduce more speech distortion as the number of target sources increases.

To further explain these results, distortion measures were also calculated for each target source separately. The results for the scenario S4N1a (targets at 270, 330, 45 and 150 degrees, noise in front of the head) are shown in Figure 4.9, where (a) shows the distortion in the left hearing aid output, and (b) shows the distortion in the right hearing aid output. The results for the lower reverberation environment ($\mathrm{RT}_{60} = 0.21$s) are used here, because the overall tendencies are more clearly visible in the figures.

The target speech signals were scaled so as to have the same average input power at the loudspeakers, but because of the headshadow effect, the signals have different average input powers at the reference microphones. As a consequence, S270 and S330 will be dominant in the left output (a), whereas S45 and S150 are dominant in the right output (b). Figure 4.9 illustrates that the different filters introduce different distortions in these signals. The R1-MWF introduces a lot of distortion, especially for the dominant sources: S270 and S330 have the largest distortion values in the left output (a), S45 and S150 have the largest values in the right output (b). The SP-MWF on the other hand introduces more distortion in the signals with the lowest input power, whereas the distortion on the dominant sources is very small. This is a beneficial effect, as the distortion introduced on the low power sources will be less audible, while the distortion on the dominant sources should be kept as small as possible. These results also explain why the overall distortion in Figure 4.8 (on the sum of the target signals) increased for the R1-MWF, but remained small for the SP-MWF. Figure 4.9 also shows that the SDW-MWF introduces a more or less even amount of distortion in the different target signals. Larger values of μ can be chosen as the introduced distortions are still reasonable, in contrast to the R1-MWF, where $\mu = 0$ already introduces a large amount of distortion.



(a) Left output



(b) Right output

Figure 4.9: SI-weighted Distortion per target source; S4N1a, RT_{60} = 0.21s, 4 microphones

4.6 Conclusion

The theoretical analysis and the simulations illustrate that for a single target speech source, algorithms based on the R1-MWF and SP-MWF formulations achieve a better noise reduction performance than algorithms based on the standard SDW-MWF formulation. Moreover, they do not lose this performance for smaller values of μ (less speech distortion). For applications where only a single target speech source is present, the R1-MWF and SP-MWF thus have a clear advantage over the SDW-MWF, especially so if only a limited amount of speech distortion is allowed.

When the number of target speech sources increases, the simulations show that the R1-MWF loses some of its benefit over the SDW-MWF. It still achieves large (SI-weighted) SNR improvements, but more speech distortion is introduced as the number of target speech sources increases, especially so for the dominant sources. The SP-MWF on the other hand only introduces distortions in the signals with low input powers, which is less audible, in addition to providing a large (SI-weighted) SNR improvement.

Chapter 5

VAD-robust rank-one SDW-MWF techniques

This chapter focuses on second-order statistics estimation errors which are specifically introduced by an erroneous VAD. It was previously shown that the R1-MWF and SP-MWF techniques, which implicitly assume a single target speech source, are more robust against estimation errors in the speech second-order statistics, compared to the standard SDW-MWF. Moreover, it will now be demonstrated that, assuming a single target speech source, they are even inherently insensitive to VAD errors if the noise is stationary. Finally, a novel so-called *VAD-robust* MWF extension is proposed that also achieves a good performance when the noise is (spectrally) nonstationary, even at low input SNRs where the VAD error rate may be high.

Section 5.1 presents the problem statement and some preliminary remarks.

Section 5.2 presents an analysis of the impact of VAD errors, both for a stationary noise scenario as for a (spectrally) nonstationary noise scenario.

Section 5.3 presents the so-called VAD-robust MWF extension which improves the performance in a nonstationary noise scenario.

Section 5.4 validates the new extension by physical evaluations.

Section 5.5 concludes the chapter.

The material presented in this chapter was published in the proceedings of ICASSP [39].

5.1 Introduction

5.1.1 Problem statement

It was demonstrated in chapter 4 that the R1-MWF and SP-MWF, which implicitly assume a single target speech source and decompose the overall filter as a spatial filter followed by a single-channel postfilter, have some benefits over the standard SDW-MWF. In particular, it was proven that the R1-MWF and SP-MWF are more robust to estimation errors in the speech second-order statistics (SOS). Speech estimation errors can occur because of several reasons. However, by using error model (4.18) the impact of speech estimation errors could be studied without considering the actual origin of the errors. Estimation errors in the noise SOS were not included in the analysis.

In this chapter, we now focus on estimation errors in both the speech and the noise SOS, which are specifically introduced by an erroneous VAD. First, it is shown that, assuming a single speech source, the R1-MWF and SP-MWF are inherently insensitive to VAD errors if the noise is stationary. Whereas the spatial information usually changes slowly over time, the spectral information of real-life noise can change quickly. As a result, real-life noise is usually not stationary so that the R1-MWF and SP-MWF performances can degrade, in the presence of VAD errors. By these findings it can thus be argued that (spectral) nonstationary noise is the real issue limiting the performances of these techniques, and not the VAD *per se*. Finally, a so-called *VAD-robust* MWF extension is proposed which improves the SNR performances of the R1-MWF and SP-MWF for (spectrally) nonstationary noise scenarios, even at low input SNRs where the VAD error rate is high.

5.1.2 Preliminaries

As in chapter 4, a generic microphone array consisting of a total of M microphones is considered. In the analysis we will again only consider the narrowband output SNR (2.25). Assuming a single target speech source (2.23) and no SOS estimation errors, the standard SDW-MWF (2.53), R1-MWF (4.8) and SP-MWF (4.17) are all equivalent and equal to expression (4.3), i.e.

$$\mathbf{w}_{\text{opt.}} = \mathbf{R}_v^{-1}\mathbf{a} \; . \; \frac{P_s \; A_{\text{ref}}^*}{\mu + \rho} \; . \tag{5.1}$$

This expression shows that the filter has the conceptually interesting structure of a spatial filter ($\mathbf{R}_v^{-1}\mathbf{a}$) followed by a single-channel Wiener postfilter ($\frac{P_s \; A_{\text{ref}}^*}{\mu+\rho}$). The narrowband output SNR only depends on the spatial filter part and was shown to be equal to $\rho = P_s \mathbf{a}^H \mathbf{R}_v^{-1}\mathbf{a}$. The R1-MWF (4.8) and SP-MWF (4.17) are also

explicitly decomposed into a spatial filter part ($\mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{e}_{\text{ref}}$ for both filters) and a single-channel postfilter part. As we are only interested in the narrowband output SNR, again only the spatial filter part has to be considered so that the R1-MWF and SP-MWF can be treated together in the analysis.

5.2 Impact of VAD errors

5.2.1 Impact of VAD errors under stationary noise

We now focus on the spatial filter part ($\mathbf{R}_v^{-1}\mathbf{a}$) of the theoretical optimal filter (5.1). If the estimated filter is parallel to this filter, the optimal narrowband SNR performance is achieved. We hereby assume that the single-channel postfilter can compensate for any arbitrary scaling factor so that the speech distortion is limited. First, for any values α and β with $\alpha \neq \beta$, and for a single target speech source (2.23), we find that (where $\cong$ means "is parallel[1] to"):

$$\begin{aligned}
\mathbf{R}_v^{-1}\mathbf{a} &\cong \mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{e}_{\text{ref}} \,, \\
&\cong (\mathbf{R}_v + \alpha\mathbf{R}_x)^{-1}(\beta-\alpha)\mathbf{R}_x\mathbf{e}_{\text{ref}} \,, \\
&\cong (\mathbf{R}_v + \alpha\mathbf{R}_x)^{-1}\big[(\mathbf{R}_v + \alpha\mathbf{R}_x) + (\beta-\alpha)\mathbf{R}_x\big]\mathbf{e}_{\text{ref}} - \mathbf{e}_{\text{ref}} \,, \\
&\cong (\mathbf{R}_v + \alpha\mathbf{R}_x)^{-1}(\mathbf{R}_v + \beta\mathbf{R}_x)\mathbf{e}_{\text{ref}} - \mathbf{e}_{\text{ref}} \,. \qquad (5.2)
\end{aligned}$$

This implies that if we have any two different mixes of $\mathbf{R}_v$ and $\mathbf{R}_x$ as in (5.2), a filter which is parallel to the optimal spatial filter ($\mathbf{R}_v^{-1}\mathbf{a}$) can be obtained, so that the optimal narrowband SNR performance is achieved.

When the noise is stationary, VAD errors in the estimated noise and speech+noise correlation matrices (2.16)-(2.17) can be modelled as (cfr. [163]):

$$\hat{\mathbf{R}}_v = \mathbf{R}_v + \delta_1\mathbf{R}_x \quad \text{and} \quad \hat{\mathbf{R}}_y = \mathbf{R}_v + \delta_2\mathbf{R}_x \,, \qquad (5.3)$$

so that the spatial filter part ($\mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{e}_{\text{ref}}$) of the R1-MWF and SP-MWF is estimated as (using $\hat{\mathbf{R}}_x = \hat{\mathbf{R}}_y - \hat{\mathbf{R}}_v$):

$$\begin{aligned}
\hat{\mathbf{w}} &= \hat{\mathbf{R}}_v^{-1}\hat{\mathbf{R}}_y\mathbf{e}_{\text{ref}} - \mathbf{e}_{\text{ref}} \,, \\
&= (\mathbf{R}_v + \delta_1\mathbf{R}_x)^{-1}(\mathbf{R}_v + \delta_2\mathbf{R}_x)\mathbf{e}_{\text{ref}} - \mathbf{e}_{\text{ref}} \,, \qquad (5.4)
\end{aligned}$$

which, by (5.2), is parallel to the optimal spatial filter $\mathbf{R}_v^{-1}\mathbf{a}$. For stationary noise, we thus see that the spatial filter part ($\mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{e}_{\text{ref}}$) of the R1-MWF and SP-MWF, remarkably, is robust against VAD errors.

[1] Vectors $\mathbf{u}$ and $\mathbf{v}$ are parallel if they are equal up to a scalar factor s ($\neq 0$), i.e. $\mathbf{u} = s\mathbf{v}$.

5.2.2 Impact of VAD errors under nonstationary noise

For (spectrally) nonstationary noise[2], (5.3) is inaccurate, as the changing noise power will lead to different weighting factors on $\mathbf{R}_v$, i.e.

$$\hat{\mathbf{R}}_v = \gamma_1 \mathbf{R}_v + \delta_1 \mathbf{R}_x \quad \text{and} \quad \hat{\mathbf{R}}_y = \gamma_2 \mathbf{R}_v + \delta_2 \mathbf{R}_x \; . \tag{5.5}$$

As a result, (5.2) is no longer valid and the spatial filter part ($\mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{e}_{\text{ref}}$) of the R1-MWF and SP-MWF will not be parallel to the optimal filter. By assuming error model (5.5) and a single speech source (2.23), and utilizing the matrix inversion lemma (A.6), we find:

$$\begin{aligned} \hat{\mathbf{w}} &= (\gamma_1 \mathbf{R}_v + \delta_1 \mathbf{R}_x)^{-1}(\gamma_2 \mathbf{R}_v + \delta_2 \mathbf{R}_x)\mathbf{e}_{\text{ref}} - \mathbf{e}_{\text{ref}} \; , \\ &= \left(\frac{\gamma_1\delta_2 - \gamma_2\delta_1}{\gamma_1(\gamma_1 + \delta_1\rho)}\right)\mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{e}_{\text{ref}} \; + \; \left(\frac{\gamma_2}{\gamma_1} - 1\right)\mathbf{e}_{\text{ref}} \; . \end{aligned} \tag{5.6}$$

Expression (5.6) shows that the estimated spatial filter is now a combination of the optimal spatial filter and the trivial filter $\mathbf{e}_{\text{ref}}$, with scaling factors depending on ρ, γ_1, γ_2, δ_1 and δ_2. For the special case $\gamma_1 = \gamma_2$, the scaling factor in front of $\mathbf{e}_{\text{ref}}$ is equal to 0. The estimated filter is then again parallel to the optimal spatial filter, so that the optimal SNR performance is achieved. The condition $\gamma_1 = \gamma_2$ is satisfied for a stationary noise scenario (5.3), for which we have indeed demonstrated robustness to VAD errors in section 5.2.1.

5.3 VAD-robust extension

The analysis of section 5.2.2 demonstrates that if $\hat{\mathbf{R}}_y$ can be scaled so that $\gamma_1 \approx \gamma_2$, the optimal SNR performance can be achieved, even for nonstationary noise. We therefore search for a scalar c so that $(c\hat{\mathbf{R}}_y - \hat{\mathbf{R}}_v) \cong \mathbf{R}_x$. As $\mathbf{R}_x$ is assumed to be a rank-one matrix, the rank of $(c\hat{\mathbf{R}}_y - \hat{\mathbf{R}}_v)$ has to be minimized, which can be done by trace minimization [64]. This results in the following convex optimization problem:

$$\min_c \; \text{Tr}\big\{|c\hat{\mathbf{R}}_y - \hat{\mathbf{R}}_v|\big\} \; , \tag{5.7}$$

where $|.|$ is the absolute value operator. Instead of (5.7), we will however minimize the following convex problem,

$$\min_c \; \text{Tr}\big\{(c\hat{\mathbf{R}}_y - \hat{\mathbf{R}}_v)^2\big\} \; , \tag{5.8}$$

as the optimum can then be found analytically as

$$c_{\text{opt}} = \frac{\text{vec}\{\hat{\mathbf{R}}_v\}^H \text{vec}\{\hat{\mathbf{R}}_y\}}{\text{vec}\{\hat{\mathbf{R}}_y\}^H \text{vec}\{\hat{\mathbf{R}}_y\}} \; , \tag{5.9}$$

[2]By "spectrally" nonstationary noise, we mean that the power and spectral content of the noise changes over time, whereas the spatial characteristics are fixed.

where vec{.} stacks the columns of a matrix into a column vector.

The scaling factor obtained in (5.9) can be wrong if the input SNR is too high, because the trace minimization will then lead to a c for which $(c\hat{\mathbf{R}}_y - \hat{\mathbf{R}}_v) \cong \mathbf{R}_v$. In that case the normal solution (without scaling of $\hat{\mathbf{R}}_y$) is better. Fortunately, at higher input SNRs, there are also less VAD errors so that a good performance is achieved. As a switching criterion, we propose (with $\|.\|$ the matrix 2-norm and TOL a small scalar):

---

if $\quad 1 - \text{TOL} \leq \frac{\|c_{\text{opt}}\hat{\mathbf{R}}_y\|}{\|\hat{\mathbf{R}}_v\|} \leq 1 + \text{TOL} \quad$ **then**

$\quad \hat{\mathbf{w}} \leftarrow \hat{\mathbf{R}}_v^{-1}(c_{\text{opt}}\hat{\mathbf{R}}_y)\mathbf{e}_{\text{ref}} - \mathbf{e}_{\text{ref}}$

else

$\quad \hat{\mathbf{w}} \leftarrow \hat{\mathbf{R}}_v^{-1}\hat{\mathbf{R}}_y\mathbf{e}_{\text{ref}} - \mathbf{e}_{\text{ref}}$

end if

Speech distortion has not been considered in the previous analysis, as it was assumed that arbitrary scaling factors could be compensated by the spectral postfilter. Theoretically the spectral postfilter is given by[3]:

$$\text{postfilter} = \frac{1}{\mu + \rho} = \frac{1}{\mu + \text{SNR}^{out}} \; . \tag{5.10}$$

In this expression, SNR^{out} was previously estimated using only the speech and noise second-order statistics, cfr. the R1-MWF formula (4.8) or the SP-MWF formula (4.17). However, for the current algorithm, where $\hat{\mathbf{w}}$ is scaled depending on the value for c_{opt}, it was observed that estimating SNR^{out} via definition (2.25) leads to the lowest distortions. Namely, $\widehat{\text{SNR}}^{out}$ is found by plugging $\hat{\mathbf{w}}$ into the definition (and with $\hat{\mathbf{R}}_x = \hat{\mathbf{R}}_y - \hat{\mathbf{R}}_v$):

$$\widehat{\text{SNR}}^{out} = \frac{\hat{\mathbf{w}}^H\hat{\mathbf{R}}_y\hat{\mathbf{w}}}{\hat{\mathbf{w}}^H\hat{\mathbf{R}}_v\hat{\mathbf{w}}} - 1 \; , \tag{5.11}$$

where $\hat{\mathbf{R}}_y$ is the original correlation matrix estimate without rescaling with c_{opt}. The estimated postfilter is then found by plugging (5.11) into (5.10), i.e.:

$$\text{postfilter} = \frac{\hat{\mathbf{w}}^H\hat{\mathbf{R}}_v\hat{\mathbf{w}}}{(\mu - 1)(\hat{\mathbf{w}}^H\hat{\mathbf{R}}_v\hat{\mathbf{w}}) + \hat{\mathbf{w}}^H\hat{\mathbf{R}}_y\hat{\mathbf{w}}} \; . \tag{5.12}$$

[3]Compared to the postfilter in (2.59), the factor ρ in the numerator is now already included in the spatial filter part.

5.4 Experimental results

5.4.1 Setup and stimuli

A binaural hearing aid setup with two behind-the-ear hearing aids is considered. There are two omnidirectional microphones per device, and it is assumed that each device exchanges one audiosignal (i.e. the front microphone signal) with the other device. This corresponds to the MWF-Front algorithm, with a total of 3 microphone signals (2 ipsilateral and 1 contralateral) per device.

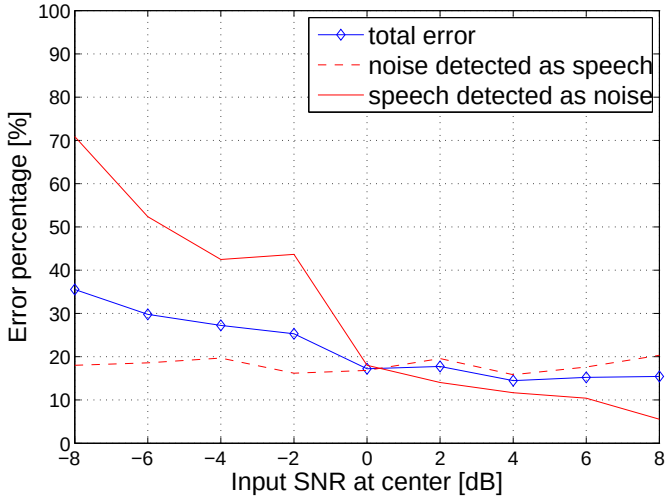
Head-related transfer functions (HRTFs) were measured on the setup described in chapter 3 (SI-weighted reverberation time $\mathrm{RT}_{60} = 0.62s$). Three spatial scenarios are considered: S0N45, S315N90 and S0N90/180/270, where the target speech (S) and interfering noise (N) source(s) are positioned at the specified azimuthal angles (with 0° in front of the head, 90° to the right of the head). As speech stimulus, 4 sentences of the Dutch VU sentence material [184] were used. Multitalker babble noise was used as interfering noise signal(s). To assess the performance, the SI-weighted SNR improvement [74] is calculated on the second part of the signals (last two sentences), to allow the filters to converge.

Adaptive algorithms based on the standard SDW-MWF (2.53), R1-MWF (4.8) and VAD-robust extension (section 5.3), were implemented in a weighted overlap-add (WOLA) filterbank. The signals are sampled at $f_s = 20480$ Hz, segmented in frames of 128 samples with 75% overlap, and windowed by a Hann Window. Other used parameters are: $\mu = 5$, $\lambda_y = \lambda_v = 0.9995$, $\mathrm{TOL} = 0.05$.

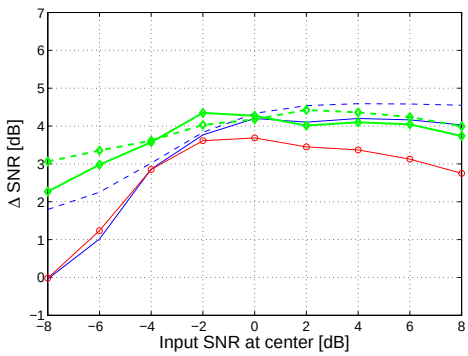
As VAD algorithm, we make a decision fusion (using the wireless link) of log-energy-based VAD algorithms calculated at both sides of the head, together with a cross-correlation based VAD (which assumes a signal from 0° is target speech). This VAD algorithm is denoted as *Fusion-2 VAD* in chapter 9. We note that two of the scenarios pose a difficulty for the cross-correlation VAD: in S315N90 the speech location deviates from 0°, in S0 N90/180/270 one of the noise sources (at 180°) also appears to be from 0°.

5.4.2 Results

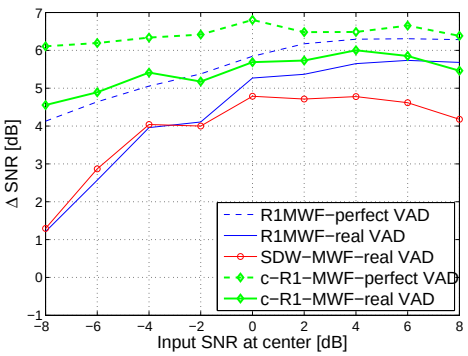
In Figures (5.1)-(5.3), the VAD performance (percentage total error, percentage speech+noise frames erroneously classified as noise-only, and percentage noise-only frames erroneously classified as speech+noise) and left and right SNR improvements (versus the front microphone signals) are shown for input SNRs ranging from -8dB to +8dB (measured in absence of the CORTEX manikin), for the three considered spatial scenarios. The results show that the R1-MWF outperforms the standard SDW-MWF as was already demonstrated in chapter 4,



(a) **S0 N45**: VAD performance



(b) **S0 N45**: Left output

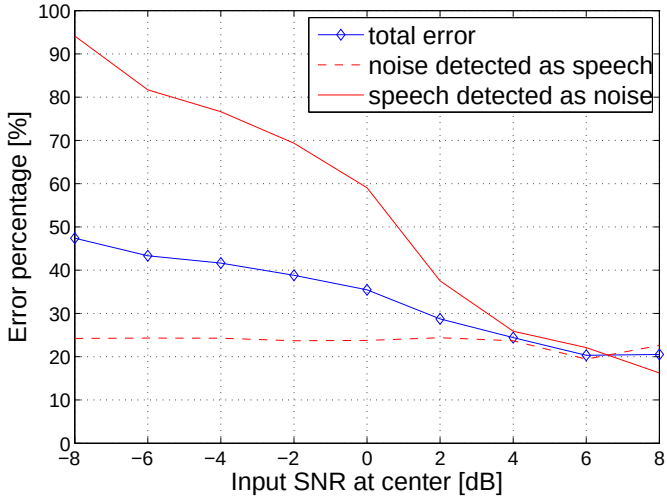


(c) **S0 N45**: Right output

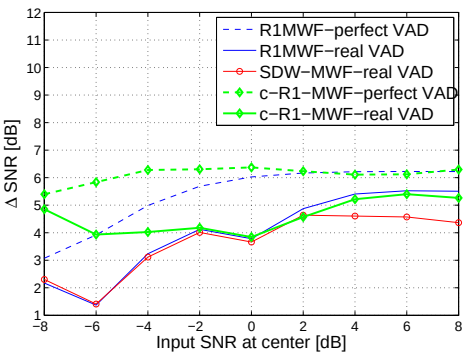
Figure 5.1: **S0 N45**; VAD performance (a), SI-weighted SNR improvement at left device (b) and at right device (c)

especially at higher input SNRs (where $\mathbf{R}_x$ has more weight in the SDW-MWF formula and imperfections have more impact).

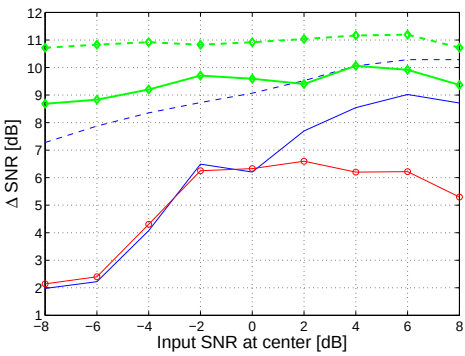
The VAD-robust MWF of section 5.3, denoted as c-R1-MWF, generally leads to a better performance than the R1-MWF. Remarkably, even for a perfect VAD, the c-R1-MWF can achieve performance improvements over the R1-MWF. This can be explained as follows. For the perfect VAD, a binary decision is made for every frame (for the entire frequency range). As a result, even for a perfect VAD, $\hat{\mathbf{R}}_y$ and $\hat{\mathbf{R}}_v$ can be badly scaled at the frequencies with a low average input SNR (where



(a) **S315 N90**: VAD performance



(b) **S315 N90**: Left output

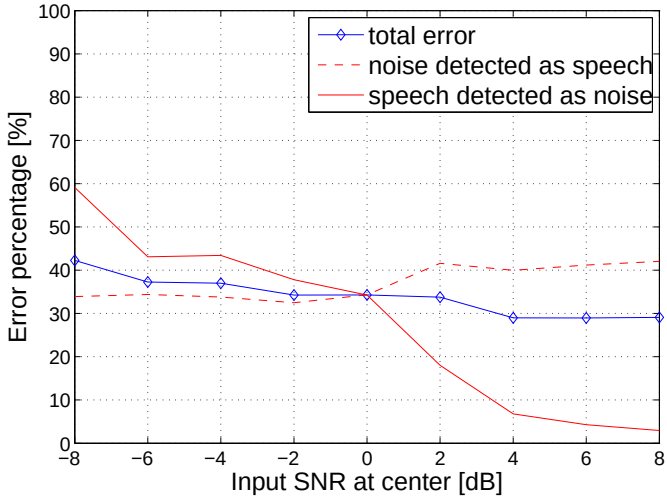


(c) **S315 N90**: Right output

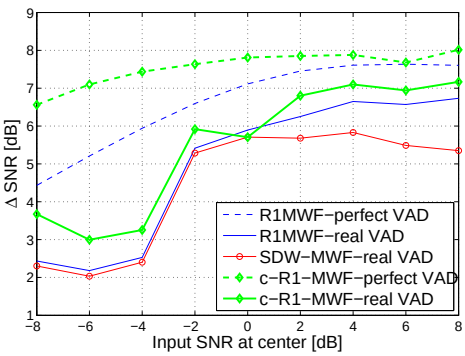
Figure 5.2: **S315 N90**; VAD performance (a), SI-weighted SNR improvement at left device (b) and at right device (c)

speech is mostly absent). This can however be compensated by the c-R1-MWF, thus leading to a better SNR performance.

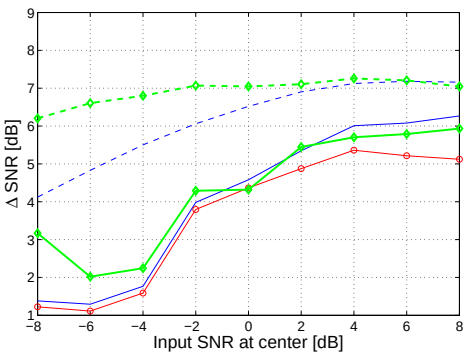
The improvement of the c-R1-MWF is retained when using a real VAD, especially at lower SNRs (high VAD error rate). On the other hand no performance is lost at higher input SNRs as the algorithm then switches back to the normal R1-MWF solution.



(a) **S0 N90/180/270**: VAD performance



(b) **S0 N90/180/270**: Left output



(c) **S0 N90/180/270**: Right output

Figure 5.3: **S0 N90/180/270**; VAD performance (a), SI-weighted SNR improvement at left device (b) and at right device (c)

5.5 Conclusion

In this chapter, it was demonstrated that the rank-one SDW-MWF formulations (R1-MWF and SP-MWF) remarkably are inherently robust to VAD errors, provided the noise is stationary. If performance degradations due to VAD erros do occur in practice, the major underlying issue is that the noise is usually nonstationary in real-life. A novel so-called VAD-robust extension, which can also retain the SNR performance for nonstationary noise, was also proposed.

Chapter 6

QRD-RLS algorithms for rank-one SDW-MWF techniques

In this chapter, a robust and efficient QR Decomposition Recursive Least Squares (QRD-RLS) frequency-domain algorithm for the SDW-MWF is derived. The algorithm is based on the R1-MWF and SP-MWF formulations, which were shown (cfr. chapter 4) to outperform the standard SDW-MWF in the presence of estimation errors in the second-order statistics. As a result, the QRD-RLS algorithm also achieves SI-weighted SNR improvements compared to the reference adaptive SDW-MWF algorithm (evaluated in chapter 3), which was based on the standard SDW-MWF formula. In contrast to previous time-domain QRD-RLS algorithms for the MWF, this algorithm also allows to include a trade-off between speech distortion and noise reduction. Moreover, as all processing is performed in the frequency domain, the computational efficiency is increased. Finally, it is demonstrated that the proposed QRD-RLS algorithms improves the numerical robustness so that the wordlength can be reduced in fixed point processing without loss of numerical accuracy.

Section 6.1 gives an overview of previous adaptive SDW-MWF algorithms and presents the basic idea behind the proposed QRD-RLS algorithm.

Section 6.2 presents the derivation of the QRD-RLS algorithms forthe R1-MWF and SP-MWF.

Section 6.3 presents simulation results obtained with the QRD-RLS algorithms.

Section 6.4 presents the conclusions of the chapter.

The material presented in this chapter was published in the proceedings of EUSIPCO [34].

6.1 Introduction

6.1.1 Adaptive SDW-MWF algorithms

In [54], a class of adaptive noise reduction algorithms is introduced, which are frequency-domain implementations of the SDW-MWF. A Recursive Least Squares (RLS)-type update procedure is adopted, where a weighted sum of a speech and a noise correlation matrix has to be inverted at every filter update. Moreover, an eigenvalue decomposition is calculated to ensure a positive definite speech correlation matrix, so that the algorithm is guaranteed not to diverge. When the number of input microphone signals is large (e.g. in binaural hearing aids), the complexity of these operations increases dramatically. Therefore, some simplifications were also proposed in [54], based on (block) diagonal approximations of the correlation matrices, which however decreases the performance.

In [152], a QR Decomposition Recursive Least Squares (QRD-RLS) based time-domain implementation of the MWF was introduced. Instead of the speech and noise correlation matrices, their Cholesky (square root) factors are stored and updated by a numerically robust procedure based on Givens transformations [73]. As the Cholesky factors have half the dynamic range of the correlation matrices, the wordlength can be reduced in fixed point processing without loss of numerical accuracy. A problem with the QRD-RLS algorithm of [152] is that it does not allow to include a trade-off between speech distortion and noise reduction, as in the SDW-MWF. This explicit trade-off is beneficial, as it allows increasing the global (broadband) output SNR [160]. Additionally, as the algorithm operates in the time domain, which involves large correlation matrices, the computational complexity is prohibitive for a hearing aid application.

In this chapter, it will be shown that the R1-MWF and SP-MWF formulations, which were introduced in chapter 4, enable a frequency-domain implementation based on QRD-RLS. It will be shown that this QRD-RLS algorithm obtains a higher SI-weighted SNR improvement than the algorithm in [54], as was also observed for batch algorithms (cfr. chapter 4). Additionally, in contrast to the implementation in [152], a trade-off can be included between speech distortion and noise reduction. Also, as all processing is performed in the frequency domain (as is usually done in hearing aids), the computational efficiency is increased. Finally, it is demonstrated that the QRD-RLS algorithm indeed improves the numerical robustness so that the wordlength can be reduced in fixed point processing.

6.1.2 Preliminaries

As in chapters 4 and 5, a generic microphone array consisting of M microphones is considered. The standard SDW-MWF formula for a generic microphone array was given by (4.2). By implicitly assuming a single target source, two alternative SDW-MWF formulations, which are now redisplayed for easy referencing, were also derived in chapter 4. These expressions are the R1-MWF (4.8), i.e.

$$\mathbf{w}_{\mathrm{R1-MWF}} = \mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{e}_{\mathrm{ref}} \cdot \frac{1}{\mu + \mathrm{Tr}\{\mathbf{R}_v^{-1}\mathbf{R}_x\}} , \tag{6.1}$$

and the SP-MWF (4.17), i.e.

$$\mathbf{w}_{\mathrm{SP-MWF}} = \mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{e}_{\mathrm{ref}} \cdot \frac{\mathbf{e}_{\mathrm{ref}}^H\mathbf{R}_x\mathbf{e}_{\mathrm{ref}}}{\mu\ \mathbf{e}_{\mathrm{ref}}^H\mathbf{R}_x\mathbf{e}_{\mathrm{ref}} + \mathrm{Tr}\{\mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{e}_{\mathrm{ref}}\mathbf{e}_{\mathrm{ref}}^H\mathbf{R}_x\}} . \tag{6.2}$$

The fact that only $\mathbf{R}_v$ is inverted in these expressions[1] (in contrast to the general formula (4.2)) will be utilized to derive a robust QRD-RLS algorithm. When comparing (6.1) to (6.2), it can be seen that both filters can be decomposed into a spatial filter $\mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{e}_{\mathrm{ref}}$, which is the same for both filters, and a single channel postfilter, which is different for both filters. As the postfilter in (6.2) does not require the full speech correlation matrix (only the reference column), in contrast to the postfilter in (6.1), the SP-MWF allows for a simpler QRD-RLS algorithm.

6.2 QRD-RLS algorithms

6.2.1 QRD-RLS algorithm for R1-MWF

In [152], a QRD-RLS algorithm based on the standard SDW-MWF (4.2) with $\mu = 1$ was proposed. Instead of the speech and noise correlation matrices, their Cholesky factors are stored and updated by a numerically robust procedure[2] based on Givens transformations [73]. As already mentioned, the problem with this approach is that it is derived for the particular case $\mu = 1$, such that effectively $\mathbf{R}_y$ is inverted. For $\mu \neq 1$, large circular noise buffers have to be used, which is not feasible in a hearing aid application. To work around this problem, the R1-MWF (6.1) will now be used as a starting point. As only $\mathbf{R}_v$ is inverted, a QRD updating scheme is then possible even for $\mu \neq 1$.

[1] The VAD-robust extension of chapter 5 is not considered here: the calculation of (5.9) requires estimates of the full correlation matrices, which are not directly available in the current QRD-RLS algorithm.

[2] A review of QRD updating and QRD-RLS can be found in [152].

By plugging $\mathbf{R}_x = \mathbf{R}_y - \mathbf{R}_v$ into (6.1), and by defining $\mathbf{M}_{vy} = \mathbf{R}_v^{-1}\mathbf{R}_y$, the following formula is obtained:

$$\mathbf{w}_{\mathrm{R1-MWF}} = (\mathbf{M}_{vy} - \mathbf{I}_M)\mathbf{e}_{\mathrm{ref}} \; . \; \frac{1}{\mu + \mathrm{Tr}\{\mathbf{M}_{vy}\} - M} \; . \tag{6.3}$$

The R1-MWF formula can thus be split into a spatial beamformer $(\mathbf{M}_{vy} - \mathbf{I}_M)\mathbf{e}_{\mathrm{ref}}$ followed by a (single channel) spectral postfilter, and both parts only depend on the unknown matrix $\mathbf{M}_{vy}$. By defining $\mathbf{R}_v = \mathbf{R}_{v\Delta}^H \mathbf{R}_{v\Delta}$ (i.e. $\mathbf{R}_{v\Delta}$ is the upper triangular Cholesky factor of $\mathbf{R}_v$) and $\mathbf{B} = \mathbf{R}_{v\Delta}^{-H}\mathbf{R}_y$, matrix $\mathbf{M}_{vy}$ is found by solving the following system of equations:

$$\mathbf{R}_{v\Delta} \; \mathbf{M}_{vy} = \mathbf{B} \; . \tag{6.4}$$

As $\mathbf{R}_{v\Delta}$ is triangular, this can be done by backsubstitution. In the next section, it will be shown that $\mathbf{R}_{v\Delta}$ and $\mathbf{B}$ can be efficiently updated together by applying sequences of Givens rotations. As in other Wiener filtering based procedures, there are two modes of operation (noise-only and speech+noise), which will be described separately.

Noise-only frames

In noise-only frames, the noise correlation matrix is updated as in (2.16). First, a standard QRD updating procedure [152] can be used to update the Cholesky factor of the noise correlation matrix estimate (2.16), i.e.

$$\begin{pmatrix} \mathbf{0}_{1\times M} \\ \hdashline \mathbf{R}_{v\Delta}[k] \end{pmatrix} = \mathbf{Q}^H[k] \begin{pmatrix} \sqrt{1-\lambda_v}\; \mathbf{v}^H[k] \\ \hdashline \sqrt{\lambda_v}\; \mathbf{R}_{v\Delta}[k-1] \end{pmatrix} \tag{6.5}$$

where $\mathbf{0}_{1\times M}$ is an all-zero M-dimensional row vector, and where $\mathbf{Q}^H[k]$ can be constructed as a series of M Givens transformations [73, 152]. As the processing is performed in the frequency domain, complex Givens transformations have to be calculated, for example as in [13]. The transformation matrix $\mathbf{Q}$ is then unitary, i.e. $\mathbf{Q}^H\mathbf{Q} = \mathbf{Q}\mathbf{Q}^H = \mathbf{I}_{M+1}$.
The matrix $\mathbf{B}[k-1]$ can be updated to $\mathbf{B}[k]$ using the same matrix $\mathbf{Q}^H[k]$ as in (6.5), which is explained as follows. As $\mathbf{Q}^H[k]$ is unitary, the following expression holds [139]:

$$\begin{pmatrix} \mathbf{0}_{M\times 1} \vdots \frac{1}{\sqrt{\lambda_v}}\mathbf{R}_{v\Delta}^{-1}[k-1] \end{pmatrix} \mathbf{Q}[k] \; \mathbf{Q}^H[k] \begin{pmatrix} \sqrt{1-\lambda_v}\; \mathbf{v}^H[k] \\ \hdashline \sqrt{\lambda_v}\; \mathbf{R}_{v\Delta}[k-1] \end{pmatrix} = \mathbf{I}_M \; . \tag{6.6}$$

By plugging (6.5) into (6.6), the following expression is found:

$$\begin{pmatrix} \mathbf{0}_{M\times 1} \vdots \frac{1}{\sqrt{\lambda_v}}\mathbf{R}_{v\Delta}^{-1}[k-1] \end{pmatrix} \mathbf{Q}[k] = \begin{pmatrix} * \vdots \mathbf{R}_{v\Delta}^{-1}[k] \end{pmatrix} \; , \tag{6.7}$$

where $*$ indicates 'don't care' entries, i.e. values which will not be used. By taking the Hermitian transpose of expression (6.7), and by multiplying with $\mathbf{R}_y[k]$, the following expression is obtained:

$$\left(\begin{array}{c} * \\ \hdashline \mathbf{R}_{v\Delta}^{-H}[k] \end{array}\right) \mathbf{R}_y[k] = \mathbf{Q}^H[k] \left(\begin{array}{c} \mathbf{0}_{1\times M} \\ \hdashline \frac{1}{\sqrt{\lambda_v}} \mathbf{R}_{v\Delta}^{-H}[k-1] \end{array}\right) \mathbf{R}_y[k] \; . \tag{6.8}$$

As in noise-only frames $\mathbf{R}_y[k] = \mathbf{R}_y[k-1]$, an update formula for $\mathbf{B} = \mathbf{R}_{v\Delta}^{-H}\mathbf{R}_y$ has thus been found, i.e.

$$\left(\begin{array}{c} * \\ \hdashline \mathbf{B}[k] \end{array}\right) = \mathbf{Q}^H[k] \left(\begin{array}{c} \mathbf{0}_{1\times M} \\ \hdashline \frac{1}{\sqrt{\lambda_v}} \mathbf{B}[k-1] \end{array}\right) \; . \tag{6.9}$$

In conclusion, $\mathbf{R}_{v\Delta}$ and $\mathbf{B}$ can thus be updated together using a series of M complex Givens rotations, i.e.

$$\left(\begin{array}{c:c} \mathbf{0}_{1\times M} & * \\ \hdashline \mathbf{R}_{v\Delta}[k] & \mathbf{B}[k] \end{array}\right) = \mathbf{Q}^H[k] \left(\begin{array}{c:c} \sqrt{1-\lambda_v}\, \mathbf{v}^H[k] & \mathbf{0}_{1\times M} \\ \hdashline \sqrt{\lambda_v}\, \mathbf{R}_{v\Delta}[k-1] & \frac{1}{\sqrt{\lambda_v}}\, \mathbf{B}[k-1] \end{array}\right) \; . \tag{6.10}$$

With the updated $\mathbf{R}_{v\Delta}$ and $\mathbf{B}$, equation (6.4) can then be solved for $\mathbf{M}_{vy}$, so that the new optimal R1-MWF can be computed.

Speech+noise frames

In speech+noise frames, the speech+noise correlation matrix is updated as in (2.17). However, as we are tracking $\mathbf{B}$ instead of $\mathbf{R}_y$, an update procedure for $\mathbf{B}$ is needed. As the noise correlation matrix is not updated, we can set $\mathbf{R}_{v\Delta}[k] = \mathbf{R}_{v\Delta}[k-1]$, so that

$$\mathbf{B}[k] = \lambda_y\, \mathbf{B}[k-1] + (1-\lambda_y)\, \bar{\mathbf{y}}[k]\, \mathbf{y}^H[k] \; , \tag{6.11}$$

with $\bar{\mathbf{y}}[k] = \mathbf{R}_{v\Delta}^{-H}[k-1]\mathbf{y}[k]$. In this update, $\bar{\mathbf{y}}[k]$ can be efficiently calculated by solving

$$\mathbf{R}_{v\Delta}^{H}[k-1]\, \bar{\mathbf{y}}[k] = \mathbf{y}[k] \tag{6.12}$$

by a single backsubstitution.

Similarly to the reference implementation [54], the filters will be kept fixed in speech+noise frames, however, this need not be the case.

6.2.2 QRD-RLS algorithm for SP-MWF

In a similar manner, the SP-MWF can be realized with a QRD-RLS algorithm. By working out (6.2) as in section 6.2.1, the following expression is found:

$$\mathbf{w}_{\mathrm{SP-MWF}} = (\mathbf{m}_{vy} - \mathbf{e}_{\mathrm{ref}}) \; . \; \frac{\mathbf{r}_x^H \mathbf{e}_{\mathrm{ref}}}{\mathbf{r}_x^H \Big(\mathbf{m}_{vy} + (\mu - 1)\mathbf{e}_{\mathrm{ref}} \Big)} \; , \tag{6.13}$$

where $\mathbf{r}_x$ is a column of the speech correlation matrix ($\mathbf{r}_x = \mathbf{R}_x \mathbf{e}_{\mathrm{ref}}$), and $\mathbf{m}_{vy}$ is a column of $\mathbf{M}_{vy}$ ($\mathbf{m}_{vy} = \mathbf{M}_{vy}\mathbf{e}_{\mathrm{ref}}$). Then, as $\mathbf{r}_x = \mathbf{R}_v(\mathbf{m}_{vy} - \mathbf{e}_{\mathrm{ref}})$, $\mathbf{R}_v = \mathbf{R}_{v\Delta}^H \mathbf{R}_{v\Delta}$ and $\mathbf{R}_{v\Delta}\mathbf{m}_{vy} = \mathbf{B}\mathbf{e}_{\mathrm{ref}} = \mathbf{b}$, this can finally be written as:

$$\mathbf{w}_{\mathrm{SP-MWF}} = (\mathbf{m}_{vy} - \mathbf{e}_{\mathrm{ref}}) \; . \; \frac{< \mathbf{R}_{v\Delta}\mathbf{e}_{\mathrm{ref}} \; , \; \mathbf{b} - \mathbf{R}_{v\Delta}\mathbf{e}_{\mathrm{ref}} >}{< \mathbf{b} + (\mu - 1)\mathbf{R}_{v\Delta}\mathbf{e}_{\mathrm{ref}} \; , \; \mathbf{b} - \mathbf{R}_{v\Delta}\mathbf{e}_{\mathrm{ref}} >} \; , \tag{6.14}$$

where the dotproduct $< \mathbf{a}_1, \mathbf{a}_2 >= \mathbf{a}_2^H \mathbf{a}_1$.

Vector $\mathbf{m}_{vy}$ can be updated during noise-only frames in a similar manner as matrix $\mathbf{M}_{vy}$ is updated for the R1-MWF. However, complexity is reduced here as only one column of $\mathbf{M}_{vy}$ is needed, so that only a single column of $\mathbf{B}$ has to be stored and updated. In contrast, the R1-MWF (6.3) requires the full matrix $\mathbf{M}_{vy}$ in order to calculate $\mathrm{Tr}\{\mathbf{M}_{vy}\}$ in the single-channel postfilter. The single-channel postfilter of the SP-MWF requires the calculation of two dotproducts, using vectors that are easily obtained from the (reduced) QRD-RLS scheme.

6.2.3 Residual extraction

In noise-only mode, it is also possible to obtain the output of the (spatial) filtering $(\mathbf{m}_{vy} - \mathbf{e}_{\mathrm{ref}})^H \mathbf{y} = (\mathbf{m}_{vy} - \mathbf{e}_{\mathrm{ref}})^H \mathbf{v}$ directly from the QRD-RLS scheme, without having to solve (6.4). Namely, by extracting the least squares residuals as in [130], it can be shown that:

$$(\mathbf{m}_{vy} - \mathbf{e}_{\mathrm{ref}})^H \mathbf{v} = -V_{\mathrm{ref}} - \frac{1}{\sqrt{1 - \lambda_v}} \, \epsilon \prod_{m=1}^{M} \cos(\theta_m) \; , \tag{6.15}$$

where the $\cos(\theta_m)$ are found in the Givens rotation matrices, and where ϵ is a by-product of the QRD-RLS scheme, i.e. the value which was indicated with a $*$ in (6.10), above the reference column of $\mathbf{B}$. The final output is then found by multiplying (6.15) with the single-channel postfilter of (6.3) or (6.14). As the postfilter of the R1-MWF (6.3) requires $\mathbf{M}_{vy}$ so that (6.4) still has to be solved, the residual extraction does not yield any benefit. This is however not the case for the SP-MWF, so that the SP-MWF allows for a further reduction of the computational complexity compared to the R1-MWF.

6.3 Experimental results

6.3.1 Setup

A binaural hearing aid configuration is considered. The link is assumed to be ideal in terms of bandwidth and power consumption, so that all microphone signals are assumed to be available as inputs to the noise reduction procedure. Two microphones are used in the left ear device and two in the right ear device, giving a total of $M = 4$. The binaural procedure produces a stereo output, but only the output for the left ear device will be shown. The left-front microphone is then chosen as the reference microphone.

HRTFs were measured in a reverberant room (reverberation time $\mathrm{RT}_{60} = 0.61$s, cfr. [176]) on a dummy-head, so that the head-shadow effect is taken into account. To generate the microphone signals, the noise and speech signals are convolved with the HRTFs corresponding to their angles of arrival, before being added together. 11 different speech-noise configurations were tested, where the azimuthal angles (defined clockwise with 0° as frontal direction) of the noise source(s) are varied. The speech source is always at 0°, except for the last scenario where it is at 270° (to the left of the head). The first six scenarios have a single noise source at an angle between 60° and 300°. Scenarios N2a, N2b and N2c have two noise sources at $[-60°, 60°]$, $[-120°, 120°]$ and $[120°, 210°]$ respectively. Scenario N4 has four noise sources at $60°, 120°, 180°$ and $210°$. Finally, for scenario S270N180 the target speech source is at 270° and the noise source is at 180°.

For the noise (interference) signal(s) multitalker babble noise is used. The target speech signal consists of 6 instances of speech-shaped noise, with periods of silence (12 s of speech, total signal length 24s). Average spectra of the target and interference signals can be found in [176]. The stimuli were scaled to obtain an input SNR of 0 dB. To assess the impact on speech intelligibility, the SI-weighted SNR improvement (2.30) is calculated. The last 12 seconds of the output signals are selected to measure the obtained output SNR, so that the performance after convergence can be assessed.

A comparison is made between the QRD-RLS algorithms proposed in this chapter, and the reference frequency-domain adaptive SDW-MWF algorithm of [54] (unconstrained algorithm with block-structured step size). All algorithms are implemented in a weighted overlap-add (WOLA) filterbank framework [18, 40], as this is a flexible framework suitable for hearing aid applications. The signals are sampled at 20480 Hz, and are processed by 128-point DFTs (with 75% frame-overlap). The MWF-based algorithms are also compared with a (time-domain) implementation of the GSC [76]. The fixed beamformer (delay-and sum beamformer) and blocking matrix of the GSC preprocessing stage are calibrated assuming the target speech source is located at 0°. To avoid speech cancellation,

the GSC filters are only updated in periods where the target speech source is inactive. The filter length was chosen so that the total input-output delay of the GSC algorithm is equal to the input-output delay of the WOLA filterbank. All tested algorithms require a VAD, which will be assumed to be perfect in these simulations.

6.3.2 SI-weighted SNR improvement

In Figure 6.1, the SI-weighted SNR improvement for the 11 different speech-noise scenarios is shown. The curves denoted with *SDW-MWF* are the performances obtained with the algorithm of [54], which is based on the standard SDW-MWF formula (4.2) and was denoted as reference algorithm in section 2.2.4. The curves denoted with *R1-MWF* and *SP-MWF* are the performances obtained with the QRD-RLS algorithms for the filters (6.3) and (6.14) respectively. In order not to overload the figure, only the case $\mu = 5$ is shown for the SP-MWF. Finally, the curve denoted with *GSC* is the performance obtained with the GSC algorithm [76].

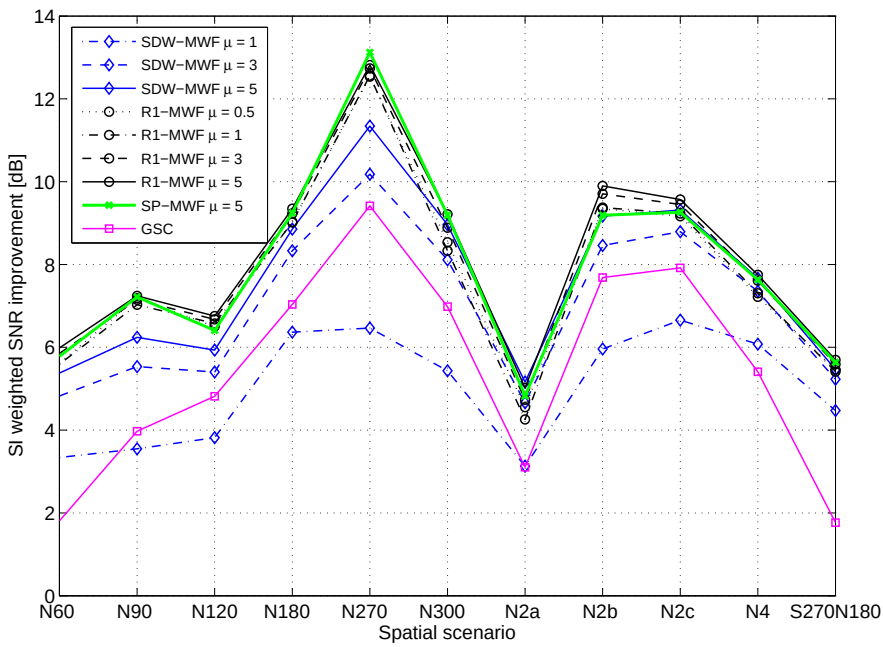


Figure 6.1: SI-weighted SNR improvement at left output

It can be observed that the R1-MWF (and SP-MWF) are insensitive to changes in μ, with respect to speech intelligibility. This was also observed in chapter 4 for the batch implementations, and is a desirable property. In theory the standard SDW-MWF formula is equivalent to the R1-MWF and SP-MWF for a single target

speech source, so that its performance should also be independent of μ. For a batch implementation, it was however demonstrated in chapter 4 that if estimation errors in the SOS are present, the standard SDW-MWF is highly dependent on μ, i.e. if μ is chosen too small, the performance degrades. The same effect is now also observed in the performance of the adaptive algorithm [54].

Figure 6.1 also illustrates that the GSC is outperformed by the MWF-based algorithms. It was demonstrated in [164] that the GSC is particularly sensitive to microphone mismatch, in contrast to the MWF. In practice, microphones are rarely matched in phase and gain, even in a single hearing aid. For a binaural hearing application where the microphone signals of two separate hearing aids are combined, the microphone mismatch may be even more severe, which can explain the lower performance of the GSC in these simulations. Additionally, when the target speech location deviates from the assumed speech location (as for scenario S270 N180), it can be seen from Figure 6.1 that the GSC performance also degrades. Finally, it is noted that algorithms, which make use of a fixed preprocessing stage (such as the GSC implementation in this chapter), may also degrade localization performance [178], whereas a binaural SDW-MWF algorithm enables correct localization (cfr. chapters 7 and 8).

6.3.3 Impact of μ: single-channel postfilter

In Figure 6.2, the broadband SNR improvement (2.29), i.e. the SNR calculated ont the broadband time-domain output signals, without SI-weighting, is shown for the R1-MWF, for different values of μ. The same observations as for the batch implementation (chapter 4) can be made: the speech distortion parameter μ only appears in the single channel spectral postfilter part in (6.3), and therefore acts as in single-microphone spectral subtraction algorithms [118]. If μ is increased, more residual noise is attenuated, hence increasing the broadband SNR by a few dBs. Although speech intelligibility is not improved (cfr. previous section), the listening comfort can then be increased at the cost of more speech distortion.

6.3.4 Robustness: effect of reduced wordlength

Figure 6.3 illustrates the effect of quantizing the values of the noise correlation matrix (or its Cholesky factor), for the spatial scenario S0 N270 and for $\mu = 5$. The QRD-RLS based implementation of the R1-MWF is compared to an algorithm without QRD-RLS, i.e. where the filter is calculated as in (6.1), using the noise correlation matrix estimate (2.16). The SNR performance of the QRD-RLS algorithm stays close to the optimal performance (i.e. the performance obtained without quantization, as shown in Figure 6.2) when the wordlength is reduced, whereas the performance of the algorithm without QRD-RLS degrades.

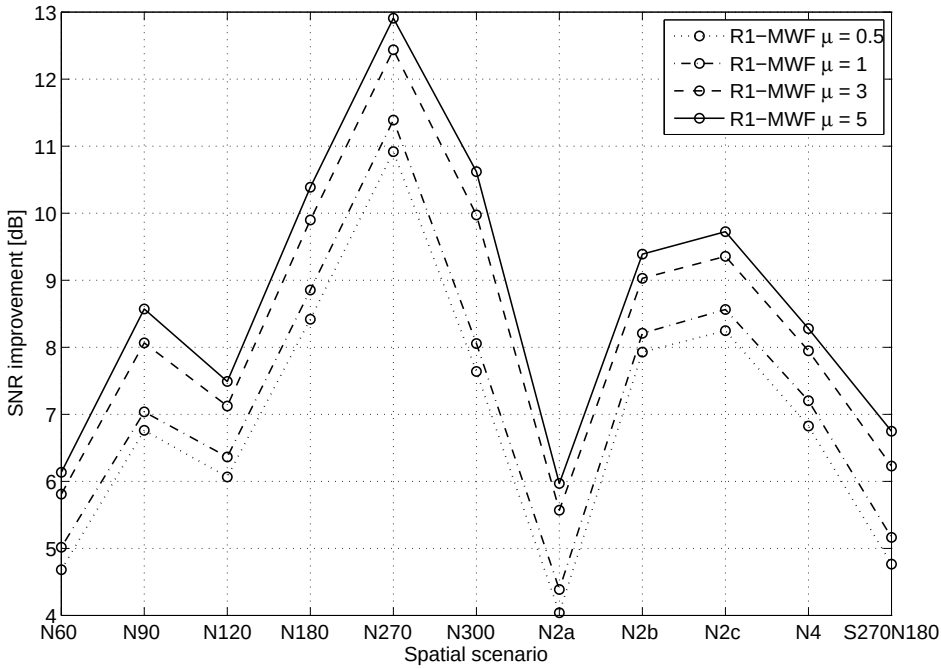


Figure 6.2: Broadband SNR improvement at left output

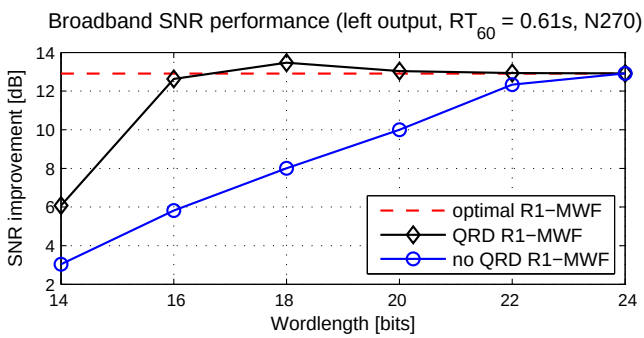


Figure 6.3: Effect of reduced wordlength in the noise correlation matrix

6.4 Conclusion

In this chapter, it was shown that a frequency-domain adaptive SDW-MWF algorithm can be realized with an efficient and robust QRD-RLS updating scheme. Simulations on a binaural 4-microphone hearing aid setup show an improved speech intelligibility weighted SNR compared to the adaptive algorithm of [54], especially so for small values of the trade-off parameter μ. Moreover, in contrast to the algorithm in [152], μ can be different from 1 without needing large circular buffers. The QRD-RLS algorithm can thus be used for smaller values of μ (low distortion beamforming), as it does not suffer from the same performance decrease as [54], but can also be used for larger values of μ, if the broadband SNR (and thus listening comfort) should be increased. Additionally, as the processing is performed in the frequency domain in contrast to the algorithm in [152], computational efficiency is increased. Finally, it was demonstrated that the QRD-RLS algorithm has a higher numerical robustness so that the wordlength can be reduced in fixed point processing without losing accuracy.

Part III

Binaural cue preservation in binaural SDW-MWF techniques

Chapter 7

Binaural cue preservation in full-bandwidth binaural SDW-MWF techniques

Part II of the thesis focused on the noise reduction performance (i.e. speech intelligibility improvements) obtained with binaural SDW-MWF techniques. It is however also desirable to preserve the binaural cues, in particular the interaural time difference (ITD) and interaural level difference (ILD), which are used to localize sounds in the horizontal plane. The binaural cue preservation of the binaural SDW-MWF techniques will therefore be studied in this part of the thesis. It can be shown that the full-bandwidth binaural SDW-MWF (where all microphone signals are exchanged between the devices) only preserves these binaural cues for the target speech source, but distorts the noise binaural cues. Two extensions to also preserve the noise binaural cues have therefore been proposed in previous work, namely the MWF with partial noise estimation (MWF-η) and MWF with interaural transfer function extension (MWF-ITF). In this chapter, the binaural cue preservation of these extensions is analyzed theoretically and tested based on objective performance measures. Both extensions are able to preserve binaural cues for the speech and noise sources, while still achieving significant noise reduction performance.

Section 7.1 presents a literature overview of binaural cue preservation by binaural noise reduction techniques. The MWF-η and MWF-ITF techniques, which were proposed in order to better preserve the noise binaural cues, are also discussed. Finally, some important preliminary remarks are given.

Section 7.2 presents a theoretical analysis of the binaural cue preservation by the binaural SDW-MWF technique.

Section 7.3 reviews the MWF-η technique and presents a theoretical analysis of the binaural cue preservation by this technique.

Section 7.4 reviews the MWF-ITF technique and presents a theoretical analysis of the binaural cue preservation by this technique.

Section 7.5 presents simulations which evaluate the different binaural techniques by objective measures.

Section 7.6 presents the conclusions of the chapter. A direct comparison between the techniques, which indicates the advantages and disadvantages of each technique, is also made.

The material presented in this chapter was published in the proceedings of EUSIPCO [30] and in the IEEE Trans. Audio, Speech, Lang. Process. journal [31].

7.1 Introduction

7.1.1 Binaural cue preservation in binaural noise reduction techniques

As was discussed in chapter 1, current noise reduction techniques in monaural and bilateral hearing aids are not designed to preserve the binaural cues [178,179], in particular the interaural time differences (ITDs) and interaural level difference (ILDs), by which the brain can localize sound sources in the horizontal plane [4, 14, 81]. However, the preservation of these ITDs and ILDs is important, as a mismatch between the perceived sound location and visual cues is disturbing. In traffic situations, incorrect sound localization can even endanger the hearing aid user. In addition to sound localization, binaural cues also improve speech perception in noisy environments (cfr. section 1.2.2). The preservation of binaural cues is thus an important design goal for noise reduction techniques in binaural hearing aids.

As discussed in section 1.4.3, many existing binaural noise reduction techniques apply identical real-valued spectral weights to one microphone signal on the left device and one microphone signal on the right device, so that the ITDs and ILDs are indeed preserved. In essence these techniques can be viewed as spectral filtering approaches, which are known to offer limited or no speech intelligibility improvements [173]. Another class of binaural techniques does perform true beamforming, although there is necessarily a trade-off between noise reduction performance and binaural cue preservation.

The binaural SDW-MWF also performs true beamforming with the microphone signals. If all microphone signals are exchanged between the devices, it can be proven (cfr. section 7.2 and [50]) that the binaural SDW-MWF preserves the binaural cues for the target speech component, but changes noise binaural cues to equal those of the speech component.

In [101], an extension of the binaural SDW-MWF has been proposed, which introduces a new trade-off parameter. This parameter allows a trade-off between noise reduction performance and the preservation of noise binaural cues. Perceptual tests in [176] have shown that this technique, referred to as *MWF with partial noise estimation (MWF-η)*, enables correct localization of both target speech and noise components. Although the noise reduction performance decreases in this approach, this does not necessarily lead to a loss in speech intelligibility due to the compensatory effect of binaural unmasking [177]. In section 7.3, the MWF-η is reviewed and its binaural cue preservation is analyzed theoretically.

In [50] a different extension was proposed to trade-off noise reduction performance for better cue preservation. Here, the SDW-MWF cost function is extended with extra terms related to the *interaural transfer functions (ITF)* of the speech and noise components. Simulations in [50] with this *MWF-ITF* technique have shown that it is then indeed possible to preserve both speech and noise binaural cues. To reduce computational complexity, a simplification was introduced in the MWF-ITF cost function to obtain quadratic cost terms [175]. Remarkably, perceptual tests in [175] showed an improvement in localization performance, even though a simplified cost function was used. In section 7.4 closed form expressions for the optimal MWF-ITF filters are derived. It is proven that this MWF-ITF technique cannot preserve target speech and noise binaural cues simultaneously, but that target speech and noise ITFs are changed into one and the same value, which is a combination of the input target speech and noise ITFs. However, to explain the improvement in the perceptual tests, it is also illustrated that the obtained output ITF is related to the output SNR.

The theoretical analysis and assessment with objective measures in this chapter demonstrate that both SDW-MWF extensions are able to preserve binaural cues for the target speech and noise sources, while still achieving a significant noise reduction performance. A direct comparison between the techniques, which indicates the advantages and disadvantages of each technique, is also made.

7.1.2 Preliminaries

In this chapter, it will be assumed that there are no bandwidth restrictions on the wireless link, so that all microphone signals can be exchanged between the devices. This corresponds to $N = M$ and $\mathbf{F}_{10} = \mathbf{F}_{01} = \mathbf{I}_M$ in the general

binaural processing scheme (section 2.1.2). The binaural cue preservation by reduced-bandwidth SDW-MWF techniques ($N < M$) will be studied in chapter 8.

In the analysis, it will always be assumed that there is a single target speech source, so that (2.23) holds. Under this assumption, the standard SDW-MWF (2.53), R1-MWF (4.8) and SP-MWF (4.17) are completely equivalent and equal to (2.54). The theoretical results presented in this chapter therefore apply to all three filters. However, it is again noted that practical algorithms based on the R1-MWF and SP-MWF perform better than algorithms based on the standard SDW-MWF, under the presence of SOS estimation errors, as was demonstrated in chapter 4.

7.2 Binaural SDW-MWF

The standard SDW-MWF formula was given by (2.53), or alternatively by the R1-MWF (4.8) or SP-MWF (4.17) expressions. In the case of a single target speech source, these expressions are equivalent and equal to (2.54). As it is assumed here that all microphone signals can be exchanged (full-bandwidth), we can set $\mathbf{R}_{vL} = \mathbf{R}_{vR} = \mathbf{R}_v$ and $\mathbf{a}_L = \mathbf{a}_R = \mathbf{a}$, so that

$$\boxed{\begin{aligned} \mathbf{w}_{\mathrm{MWF},L} &= \mathbf{R}_v^{-1}\mathbf{a}\ \frac{P_s}{\mu+\rho}A_{L,\mathrm{ref}}^* \\ \mathbf{w}_{\mathrm{MWF},R} &= \mathbf{R}_v^{-1}\mathbf{a}\ \frac{P_s}{\mu+\rho}A_{R,\mathrm{ref}}^* \end{aligned}} \tag{7.1}$$

with

$$\rho = P_s\ \mathbf{a}^H\mathbf{R}_v^{-1}\mathbf{a}\ . \tag{7.2}$$

This implies that $\mathbf{w}_{\mathrm{MWF},L}$ and $\mathbf{w}_{\mathrm{MWF},R}$ are *parallel* [55], i.e.

$$\boxed{\mathbf{w}_{\mathrm{MWF},L} = \mathrm{ITF}_x^{in,*}\,\mathbf{w}_{\mathrm{MWF},R}}\ , \tag{7.3}$$

with the interaural transfer function of the speech component ITF_x^{in} now equal to

$$\mathrm{ITF}_x^{in} = \frac{X_{L,\mathrm{ref}}}{X_{R,\mathrm{ref}}} = \frac{A_{L,\mathrm{ref}}}{A_{R,\mathrm{ref}}}\ . \tag{7.4}$$

Using definition (2.25) and the fact that $\mathbf{w}_{\mathrm{MWF},L}$ and $\mathbf{w}_{\mathrm{MWF},R}$ are parallel (7.3), the narrowband output SNRs of the left and right hearing aid are the same and equal to

$$\boxed{\mathrm{SNR}_L^{out} = \mathrm{SNR}_R^{out} = \rho = P_s\ \mathbf{a}^H\mathbf{R}_v^{-1}\mathbf{a}}\ . \tag{7.5}$$

Using definition (2.26), the narrowband SNR improvement in the left and the right hearing aid is equal to

$$\boxed{\Delta\text{SNR}_L = \frac{\rho P_{vL}}{P_s|A_{L,\text{ref}}|^2}, \quad \Delta\text{SNR}_R = \frac{\rho P_{vR}}{P_s|A_{R,\text{ref}}|^2}}, \tag{7.6}$$

with

$$P_{vL} = \mathbf{e}_L^H \mathbf{R}_v \mathbf{e}_L, \quad P_{vR} = \mathbf{e}_R^H \mathbf{R}_v \mathbf{e}_R \, . \tag{7.7}$$

As the SDW-MWF filter weight vectors for the left and the right hearing aid are parallel, the ITFs of the output speech and noise components are the same and equal to ITF_x^{in},

$$\boxed{\text{ITF}_x^{out} = \text{ITF}_v^{out} = \text{ITF}_x^{in}}, \tag{7.8}$$

implying that all sounds (including the noise components) are perceived as coming from the speech direction.

7.3 Partial noise estimation extension (MWF-η)

General case

An extension of the binaural SDW-MWF that aims to partly preserve the binaural cues of the noise component has been proposed in [101] and validated through perceptual tests in [176] and [177]. The objective is to produce an MMSE estimate of the sum of the speech component and a scaled version of the noise component in the reference microphones. This partial noise estimation was also presented in [24] and [43] in single-channel noise reduction techniques. The cost function of the SDW-MWF with partial noise estimation (*MWF-η*) in the left device is equal to

$$J_{\text{MWF}\eta}(\mathbf{w}_L) = \mathcal{E}\left\{ \left| X_{L,\text{ref}} - \mathbf{w}_L^H \mathbf{x}_L \right|^2 + \mu \left| \eta V_{L,\text{ref}} - \mathbf{w}_L^H \mathbf{v}_L \right|^2 \right\} \, , \tag{7.9}$$

with the scaling parameter $0 \leq \eta \leq 1$. The cost function for the right device is similarly defined. When $\eta = 0$, this cost function reduces to the standard SDW-MWF cost function in (2.52). When $\eta = 1$, the optimal filters are equal to $\mathbf{w}_{\text{MWF}\eta,L} = \mathbf{e}_L$ and $\mathbf{w}_{\text{MWF}\eta,R} = \mathbf{e}_R$, resulting in perfect preservation of the binaural cues for the speech and the noise component, but no noise reduction. The optimal filters for each hearing aid are equal to

$$\boxed{\begin{aligned} \mathbf{w}_{\text{MWF}\eta,L} &= (\mathbf{R}_{xL} + \mu\mathbf{R}_{vL})^{-1}(\mathbf{R}_{xL} + \eta\mu\mathbf{R}_{vL})\mathbf{e}_L \\ &= (1-\eta)\mathbf{w}_{\text{MWF},L} + \eta\mathbf{e}_L \\ \mathbf{w}_{\text{MWF}\eta,R} &= (\mathbf{R}_{xR} + \mu\mathbf{R}_{vR})^{-1}(\mathbf{R}_{xR} + \eta\mu\mathbf{R}_{vR})\mathbf{e}_R \\ &= (1-\eta)\mathbf{w}_{\text{MWF},R} + \eta\mathbf{e}_R \end{aligned}} \tag{7.10}$$

The MWF-η thus corresponds to a mixing of the output signal of the standard SDW-MWF (weighted with $(1-\eta)$) and the reference microphone signals (weighted with η).

Single target speech source

Using (7.1), in the case of a single speech source and assuming all microphone signals are exchanged (full-bandwidth), the optimal filters are equal to

$$\boxed{\begin{aligned}\mathbf{w}_{\mathrm{MWF}\eta,L} &= (1-\eta)\mathbf{R}_v^{-1}\mathbf{a}\frac{P_s}{\mu+\rho}A_{L,\mathrm{ref}}^* + \eta\ \mathbf{e}_L \\ \mathbf{w}_{\mathrm{MWF}\eta,R} &= (1-\eta)\mathbf{R}_v^{-1}\mathbf{a}\frac{P_s}{\mu+\rho}A_{R,\mathrm{ref}}^* + \eta\ \mathbf{e}_R\end{aligned}} \tag{7.11}$$

Plugging the filters (7.11) in definition (2.25), the narrowband output SNRs for the left and right hearing aid are obtained:

$$\boxed{\begin{aligned}\mathrm{SNR}_L^{out} &= \frac{(\eta\mu+\rho)^2\mathrm{SNR}_L^{in}}{(1-\eta)(2\eta\mu+\rho+\eta\rho)\mathrm{SNR}_L^{in}+\eta^2(\mu+\rho)^2} \\ \mathrm{SNR}_R^{out} &= \frac{(\eta\mu+\rho)^2\mathrm{SNR}_R^{in}}{(1-\eta)(2\eta\mu+\rho+\eta\rho)\mathrm{SNR}_R^{in}+\eta^2(\mu+\rho)^2}\end{aligned}} \tag{7.12}$$

where

$$\mathrm{SNR}_L^{in} = \frac{P_s|A_{L,\mathrm{ref}}|^2}{\mathbf{e}_L^H\mathbf{R}_v\mathbf{e}_L}, \quad \mathrm{SNR}_R^{in} = \frac{P_s|A_{R,\mathrm{ref}}|^2}{\mathbf{e}_R^H\mathbf{R}_v\mathbf{e}_R}\ . \tag{7.13}$$

The narrowband output SNRs for the left and right hearing aid are now only the same if the narrowband input SNRs are the same, or if $\eta = 0$, in which case the narrowband output SNR of the SDW-MWF approach is obtained (i.e. $\mathrm{SNR}_L^{out} = \mathrm{SNR}_R^{out} = \rho$).

Using (7.12) and definition (2.26) and by defining $\Delta\mathrm{SNR}_L^o$ and $\Delta\mathrm{SNR}_R^o$ as the narrowband SNR improvements of the binaural SDW-MWF without extension, cfr. (7.6),

$$\Delta\mathrm{SNR}_L^o = \frac{\rho P_{vL}}{P_s|A_{L,\mathrm{ref}}|^2}, \quad \Delta\mathrm{SNR}_R^o = \frac{\rho P_{vR}}{P_s|A_{R,\mathrm{ref}}|^2}\ , \tag{7.14}$$

the narrowband SNR improvement at the left hearing aid is equal to

$$\Delta\mathrm{SNR}_L = \Delta\mathrm{SNR}_L^o\frac{\left(\frac{\eta\mu+\rho}{\mu+\rho}\right)^2}{\left(\frac{\eta\mu+\rho}{\mu+\rho}\right)^2+(\Delta\mathrm{SNR}_L^o-1)\eta^2}\ . \tag{7.15}$$

If $\eta = 0$ the narrowband SNR improvement is equal to $\Delta\mathrm{SNR}_L^o$, whereas if $\eta = 1$ no SNR improvement is obtained, i.e. $\Delta\mathrm{SNR}_L = 1$. Since the narrowband SNR

improvement ΔSNR_L^o of the binaural SDW-MWF is always larger than or equal to 1 [53], it can easily be shown that

$$\boxed{1 \leq \Delta\text{SNR}_L \leq \Delta\text{SNR}_L^o}\,. \tag{7.16}$$

Similar expressions can be derived for the narrowband SNR improvement at the right hearing aid, also leading to $1 \leq \Delta\text{SNR}_R \leq \Delta\text{SNR}_R^o$.

If ρ is sufficiently large[1], i.e. $\rho >> \mu, \eta, \text{SNR}_L^{in}$, and $\eta \neq 0$, then the narrowband SNR improvement for the left hearing aid in (7.15) becomes approximately equal to:

$$\lim_{\rho\to\infty} \Delta\text{SNR}_L = \frac{1}{\eta^2}\,. \tag{7.17}$$

Similarly, if $\rho >> \mu, \eta, \text{SNR}_R^{in}$, the narrowband SNR improvement in the right hearing aid will be approximately equal to $\frac{1}{\eta^2}$.

From (7.11) it can be seen that the MWF-η filters are generally *not parallel*, such that the ITF of the output speech and noise components are typically different. Using (2.33), the ITF of the output speech component is equal to

$$\text{ITF}_x^{out} = \frac{\mathbf{w}_{\text{MWF}\eta,L}^H\,\mathbf{a}}{\mathbf{w}_{\text{MWF}\eta,R}^H\,\mathbf{a}} = \frac{(1-\eta)\frac{\rho}{\mu+\rho}A_{L,\text{ref}} + \eta A_{L,\text{ref}}}{(1-\eta)\frac{\rho}{\mu+\rho}A_{R,\text{ref}} + \eta A_{R,\text{ref}}} = \frac{A_{L,\text{ref}}}{A_{R,\text{ref}}}\,, \tag{7.18}$$

and so

$$\boxed{\text{ITF}_x^{out} = \text{ITF}_x^{in}}\,, \tag{7.19}$$

such that the ITF of the target speech component is again preserved.

By plugging the filters (7.11) into definition (2.43), the output noise ITF is obtained:

$$\text{ITF}_v^{out} = \frac{\mathbf{w}_L^H\mathbf{R}_v\mathbf{w}_R}{\mathbf{w}_R^H\mathbf{R}_v\mathbf{w}_R} = \frac{\psi(P_s A_{L,\text{ref}}A_{R,\text{ref}}^*) + \eta^2(\mathbf{e}_L^H\mathbf{R}_v\mathbf{e}_R)}{\psi(P_s|A_{R,\text{ref}}|^2) + \eta^2(\mathbf{e}_R^H\mathbf{R}_v\mathbf{e}_R)}\,, \tag{7.20}$$

where

$$\psi = \left[(1-\eta)^2\frac{\rho}{(\mu+\rho)^2} + 2\eta(1-\eta)\frac{1}{(\mu+\rho)}\right]\,. \tag{7.21}$$

By using definitions (2.35), (2.42) and (7.13), it can be shown that (7.20) is equal to:

$$\boxed{\text{ITF}_v^{out} = \frac{\psi\ \text{SNR}_R^{in}}{\psi\ \text{SNR}_R^{in} + \eta^2}\ \text{ITF}_x^{in} + \frac{\eta^2}{\psi\ \text{SNR}_R^{in} + \eta^2}\ \text{ITF}_v^{in}} \tag{7.22}$$

[1]This is the case when the spatial separation between the speech source and the noise sources is sufficiently large, in which case the product $\mathbf{a}^H\mathbf{R}_v^{-1}\mathbf{a}$ is large. A high speech power P_s also leads to a large ρ, but then the assumption $\rho >> \text{SNR}_L^{in}$ may not be valid as the input SNR is also high.

Equation (7.22) shows that ITF_v^{out} is a weighted sum of ITF_x^{in} and ITF_v^{in}. If $\eta = 1$, the factor ψ is equal to 0 so that $\mathrm{ITF}_v^{out} = \mathrm{ITF}_v^{in}$, whereas for $\eta = 0$, $\mathrm{ITF}_v^{out} = \mathrm{ITF}_x^{in}$.

By rearranging the terms in (7.22) and using (7.12), the output ITF can also be related to the narrowband SNR improvement:

$$\mathrm{ITF}_v^{out} = \mathrm{ITF}_x^{in} + \frac{\eta^2(\mu+\rho)^2}{(\eta\mu+\rho)^2}\,\Delta\mathrm{SNR}_R\,(\mathrm{ITF}_v^{in} - \mathrm{ITF}_x^{in})\,. \tag{7.23}$$

Under the assumptions leading to (7.17), the obtained noise output ITF thus becomes:

$$\mathrm{ITF}_v^{out} \approx \mathrm{ITF}_x^{in} + \frac{(\mu+\rho)^2}{(\eta\mu+\rho)^2}\,(\mathrm{ITF}_v^{in} - \mathrm{ITF}_x^{in}) \approx \mathrm{ITF}_v^{in}\,, \tag{7.24}$$

so it is seen that the error on the noise binaural cues will become small for any choice of $\eta \neq 0$ under these assumptions.

7.4 Interaural transfer function extension (MWF-ITF)

General case

To control the binaural cues of the speech and the noise component, it is also possible to extend the binaural SDW-MWF cost function with terms related to the ITF of the speech and the noise components, as has been proposed in [50,100, 175]. When the aim is to preserve the binaural cues of the speech and the noise components, the desired output ITFs are equal to the input ITFs in (2.36) and (2.42).

In [50], the ITF cost function for the noise component is defined as

$$J_{\mathrm{ITF},1}^{v}(\mathbf{w}) = \frac{\mathcal{E}\{|\mathbf{w}_L^H\mathbf{v} - \mathrm{ITF}_v^{in}\,\mathbf{w}_R^H\mathbf{v}|^2\}}{\mathcal{E}\{|\mathbf{w}_R^H\mathbf{v}|^2\}} = \frac{\mathbf{w}^H\mathbf{R}_{vt}\mathbf{w}}{\mathbf{w}^H\mathbf{R}_{v1}\mathbf{w}} \tag{7.25}$$

where the $4M$-dimensional complex stacked weight vector $\mathbf{w}$ is defined as

$$\mathbf{w} = \begin{bmatrix} \mathbf{w}_L \\ \mathbf{w}_R \end{bmatrix}\,, \tag{7.26}$$

and where

$$\mathbf{R}_{vt} = \begin{bmatrix} \mathbf{R}_v & -\mathrm{ITF}_v^{in,*}\,\mathbf{R}_v \\ -\mathrm{ITF}_v^{in}\,\mathbf{R}_v & |\mathrm{ITF}_v^{in}|^2\,\mathbf{R}_v \end{bmatrix} \tag{7.27}$$

$$= \begin{bmatrix} \mathbf{I}_{2M} \\ -\mathrm{ITF}_v^{in}\,\mathbf{I}_{2M} \end{bmatrix} \mathbf{R}_v \begin{bmatrix} \mathbf{I}_{2M} & -\mathrm{ITF}_v^{in,*}\,\mathbf{I}_{2M} \end{bmatrix} \tag{7.28}$$

$$\mathbf{R}_{v1} = \begin{bmatrix} \mathbf{0}_{2M} & \mathbf{0}_{2M} \\ \mathbf{0}_{2M} & \mathbf{R}_v \end{bmatrix} \tag{7.29}$$

For mathematical convenience, we also introduce a *simplified quadratic ITF cost function* which is obtained by removing the denominator in (7.25), corresponding to the output noise power in the right hearing aid, i.e.

$$J_{\mathrm{ITF},2}^{v}(\mathbf{w}) = \mathcal{E}\{|\mathbf{w}_L^H\mathbf{v} - \mathrm{ITF}_v^{in}\,\mathbf{w}_R^H\mathbf{v}|^2\} = \mathbf{w}^H\mathbf{R}_{vt}\mathbf{w} \tag{7.30}$$

The ITF cost function for the speech component is defined similarly as the ITF cost function for the noise component, i.e.

$$J_{\mathrm{ITF},1}^{x}(\mathbf{w}) = \frac{\mathbf{w}^H\mathbf{R}_{xt}\mathbf{w}}{\mathbf{w}^H\mathbf{R}_{x1}\mathbf{w}}\,, \tag{7.31}$$

$$J_{\mathrm{ITF},2}^{x}(\mathbf{w}) = \mathbf{w}^H\mathbf{R}_{xt}\mathbf{w}\,. \tag{7.32}$$

The SDW-MWF with Interaural Transfer Function extension (*MWF-ITF*) thus minimizes the overall cost function which trades off noise reduction, speech distortion and binaural cue preservation, i.e.

$$J_{\mathrm{MWF-ITF}}(\mathbf{w}) = J_{\mathrm{MWF}}(\mathbf{w}) + \gamma J_{\mathrm{ITF}}^{x}(\mathbf{w}) + \delta J_{\mathrm{ITF}}^{v}(\mathbf{w}) \tag{7.33}$$

where the parameters γ and δ allow to put more emphasis on binaural cue preservation for the speech and the noise component. The cost function $J_{\mathrm{MWF}}(\mathbf{w})$ is the standard SDW-MWF cost function (without extension) which was defined in (2.52). For the ITF cost functions we can either use the cost functions defined in (7.25) and (7.31) or the simplified cost functions defined in (7.30) and (7.32).

When using the ITF cost functions $J_{\mathrm{ITF},1}^{x}(\mathbf{w})$ and $J_{\mathrm{ITF},1}^{v}(\mathbf{w})$, no closed-form expression is available for the filter minimizing $J_{\mathrm{MWF-ITF}}(\mathbf{w})$, such that we have to use iterative, e.g. quasi-Newton, optimization techniques [28, 65].

When using the simplified quadratic ITF cost functions $J_{\mathrm{ITF},2}^{x}(\mathbf{w})$ and $J_{\mathrm{ITF},2}^{v}(\mathbf{w})$, the filter minimizing $J_{\mathrm{MWF-ITF}}(\mathbf{w})$ is equal to

$$\boxed{\mathbf{w}_{\mathrm{MWF-ITF},2} = (\mathbf{R} + \gamma\mathbf{R}_{xt} + \delta\mathbf{R}_{vt})^{-1}\mathbf{r}_x}\,, \tag{7.34}$$

with

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_x + \mu\mathbf{R}_v & \mathbf{0}_{2M} \\ \mathbf{0}_{2M} & \mathbf{R}_x + \mu\mathbf{R}_v \end{bmatrix}, \quad \mathbf{r}_x = \begin{bmatrix} \mathbf{R}_x\mathbf{e}_L \\ \mathbf{R}_x\mathbf{e}_R \end{bmatrix}\,. \tag{7.35}$$

Single target speech source

In the case of a single target speech source, the filter $\mathbf{w}_{\mathrm{MWF-ITF},2}$ in (7.34) can be reduced (cfr. Appendix B) to

$$\boxed{\begin{aligned}\mathbf{w}_{\mathrm{MWF-ITF},2,L} &= \mathbf{R}_v^{-1}\mathbf{a}\frac{P_s}{\mu+\rho} \\ &\quad\times\Big\{A^*_{L,\mathrm{ref}} - \xi(A^*_{L,\mathrm{ref}} - \mathrm{ITF}_v^{in,*}A^*_{R,\mathrm{ref}})\big[1-\nu\alpha_t^*(1-\xi\alpha_t)\big]\Big\} \\ \mathbf{w}_{\mathrm{MWF-ITF},2,R} &= \mathbf{R}_v^{-1}\mathbf{a}\frac{P_s}{\mu+\rho} \\ &\quad\times\Big\{A^*_{R,\mathrm{ref}} + \xi\mathrm{ITF}_v^{in}(A^*_{L,\mathrm{ref}} - \mathrm{ITF}_v^{in,*}A^*_{R,\mathrm{ref}})\Big[1-\nu\alpha_t^*\Big(\frac{\mathrm{ITF}_x^{in}}{\mathrm{ITF}_v^{in}}-\xi\alpha_t\Big)\Big]\Big\}\end{aligned}} \tag{7.36}$$

with

$$\xi = \frac{\delta}{\mu+\rho+\delta(1+|\mathrm{ITF}_v^{in}|^2)}\,, \tag{7.37}$$

$$\alpha_t = 1+\mathrm{ITF}_v^{in,*}\mathrm{ITF}_x^{in}\,, \tag{7.38}$$

$$\nu = \frac{\gamma\rho}{\mu+\rho+\gamma\rho(1+|\mathrm{ITF}_x^{in}|^2-\xi|\alpha_t|^2)}\,. \tag{7.39}$$

Hence, the filter $\mathbf{w}_{\mathrm{MWF-ITF},2,L}$ is equal to a scaled version of $\mathbf{w}_{\mathrm{MWF},L}$ in (7.1) and likewise $\mathbf{w}_{\mathrm{MWF-ITF},2,R}$ is equal to a scaled version of $\mathbf{w}_{\mathrm{MWF},R}$. The scaling factors reflect the extensions with the ITF cost functions. As a consequence, the narrowband output SNR will be the same for the left and the right hearing aid, and will be equal to the narrowband output SNR of the SDW-MWF (cfr. 7.5), i.e.

$$\boxed{\mathrm{SNR}_L^{out} = \mathrm{SNR}_R^{out} = \rho = P_s\ \mathbf{a}^H\mathbf{R}_v^{-1}\mathbf{a}}\,, \tag{7.40}$$

and the narrowband SNR improvements are given by

$$\boxed{\Delta\mathrm{SNR}_L = \frac{\rho P_{vL}}{P_s|A_{L,\mathrm{ref}}|^2},\quad \Delta\mathrm{SNR}_R = \frac{\rho P_{vR}}{P_s|A_{R,\mathrm{ref}}|^2}}\,, \tag{7.41}$$

which is the same as (7.6).

As the filters $\mathbf{w}_{\mathrm{MWF-ITF},2,L}$ and $\mathbf{w}_{\mathrm{MWF-ITF},2,R}$ are still *parallel*, the ITFs of the output speech and noise components are the same and result in

$$\boxed{\mathrm{ITF}_x^{out} = \mathrm{ITF}_v^{out} = \frac{\mathrm{ITF}_x^{in} - \xi(\mathrm{ITF}_x^{in}-\mathrm{ITF}_v^{in})\big[1-\nu\alpha_t(1-\xi\alpha_t^*)\big]}{1+\xi\mathrm{ITF}_v^{in,*}(\mathrm{ITF}_x^{in}-\mathrm{ITF}_v^{in})\Big[1-\nu\alpha_t\Big(\frac{\mathrm{ITF}_x^{in,*}}{\mathrm{ITF}_v^{in,*}}-\xi\alpha_t^*\Big)\Big]}} \tag{7.42}$$

It can be shown that for $\gamma \to \infty$ and $\delta \neq 0$, $\text{ITF}_x^{out} = \text{ITF}_v^{out} = \text{ITF}_x^{in}$ and that for $\delta \to \infty$ and $\gamma \neq 0$, $\text{ITF}_x^{out} = \text{ITF}_v^{out} = \text{ITF}_v^{in}$.

Hence, extending the SDW-MWF with the ITF cost functions gives rise to the same narrowband SNR improvement and narrowband output SNR as the binaural SDW-MWF, but changes the ITF of the output components. It is also possible to link the narrowband output SNR with the ITF of the output components. Consider two frequencies, ω_1 and ω_2, where the narrowband output SNR for ω_1 is larger than the narrowband output SNR for ω_2, i.e. $\rho_1 > \rho_2$ (ρ_1 and ρ_2 are given by (7.2) at frequencies ω_1 and ω_2 respectively). Hence, when using the same value for δ, $\xi_1 < \xi_2$ (where ξ_1 and ξ_2 are given by (7.37) at frequencies ω_1 and ω_2 respectively), such that for ω_1 the output ITF is closer to the ITF of the target speech component than for ω_2. This is an advantageous perceptual effect that will be illustrated in simulations in section 7.5.

7.5 Experimental results

In this section, the performance of the MWF-η and MWF-ITF techniques will be tested in a scenario with one target speech source and one noise source. First, the experimental setup will be discussed.

7.5.1 Data model

The sources are located in the far-field of the microphone arrays in a non-reverberant environment. It is assumed that there is one target speech source and one noise source, and that they are located at angles θ_x and θ_v from the head ($\theta = 0°$: front, $\theta = 90°$: right), with an elevation $\phi = 0°$. The speech and noise components of the microphone signals can thus be written as

$$\mathbf{x}(\omega) = \mathbf{d}(\omega, \theta_x)S(\omega), \quad \mathbf{v}(\omega) = \mathbf{d}(\omega, \theta_v)V(\omega), \tag{7.43}$$

with the steering vector $\mathbf{d}(\omega, \theta)$ equal to

$$\mathbf{d}(\omega, \theta) = \left[d_{L,1}(\omega, \theta) \ \ldots \ d_{L,M}(\omega, \theta) \ d_{R,1}(\omega, \theta) \ \ldots \ d_{R,M}(\omega, \theta)\right]^T. \tag{7.44}$$

The (omnidirectional) microphones are located on a head, so the head shadow effect will be taken into account. To achieve this, HRTFs measured on a KEMAR dummy-head [70] are incorporated in the steering vectors. It is assumed that the same HRTF can be used for all the microphones at the left hearing aid, so that for example an entry $d_{L,m}(\omega, \theta)$, representing the mth microphone of the left hearing aid, at an angle θ, can be calculated as

$$d_{L,m}(\omega, \theta) = \text{HRTF}_L(\omega, \theta) \ . \ \exp\Big(-j\omega\tau_{L,m}(\theta)\Big), \tag{7.45}$$

where $\tau_{L,m}(\theta)$ represents the delay between the mth microphone and the reference point at the left hearing aid. The steering vector entries for the right hearing aid are obtained in a similar way (using HRTF_R data and $\tau_{R,m}(\theta)$). The speech and noise correlation matrices are constructed as

$$\mathbf{R}_x(\omega) = P_s\ \mathbf{d}(\omega,\theta_x)\mathbf{d}^H(\omega,\theta_x), \tag{7.46}$$

$$\mathbf{R}_v(\omega) = P_v\ \mathbf{d}(\omega,\theta_v)\mathbf{d}^H(\omega,\theta_v) + P_{vs}\ \mathbf{I}_{2M}\ . \tag{7.47}$$

The parameters P_s, P_v and P_{vs} represent the powers of the speech source, (located) noise source and (internal) sensor noise. Some sensor noise, modelled as spatially uncorrelated noise, is added in (7.47) to add a degree of realism and also to make the noise correlation matrix invertible.

The experiments are performed using a speech source at $-5°$ and a noise source at $40°$. A 2-microphone array is used on both the left and the right hearing aid. The microphone distance on the left hearing aid is 2 cm, whereas the right hearing aid has a microphone distance of 1.5cm. The algorithms are tested at a frequency of $\omega = 2\pi \times 2000$ rad/s, the HRTF data is sampled at 44100 Hz.

7.5.2 Performance of MWF-η

The performance of the MWF-η is now evaluated with the data model of the previous section. As the ITF of the speech component is always preserved for any choice of η, as was shown in (7.19), only the noise ITD and ILD errors will be shown. The performance measures (2.48) and (2.50) are used for the ILD and ITD errors. The ILD error can thus be expressed in dB, while the ITD error is a relative error, which lies between 0 and 1.

Dependence on parameter η

The speech and noise powers in (7.46) and (7.47) will be fixed to $P_s = P_v = 1$, the sensor noise power is fixed to $P_{vs} = 0.01$. In Figure 7.1, the output narrowband SNRs, narrowband SNR improvements (for left and right hearing aid) and the noise cue errors are shown for different values of the parameter η.

The case $\eta = 0$ corresponds to the standard binaural SDW-MWF. It can be seen that the ITD/ILD errors of the noise component are large. It was indeed shown in section 7.2 and in [50] that the standard binaural SDW-MWF distorts the noise cues.

For larger values of η, the ILD/ITD errors of the noise component decrease. However, as more noise is mixed into the output signals, the obtained narrowband SNR improvement will also decrease, so that there is a trade-off between

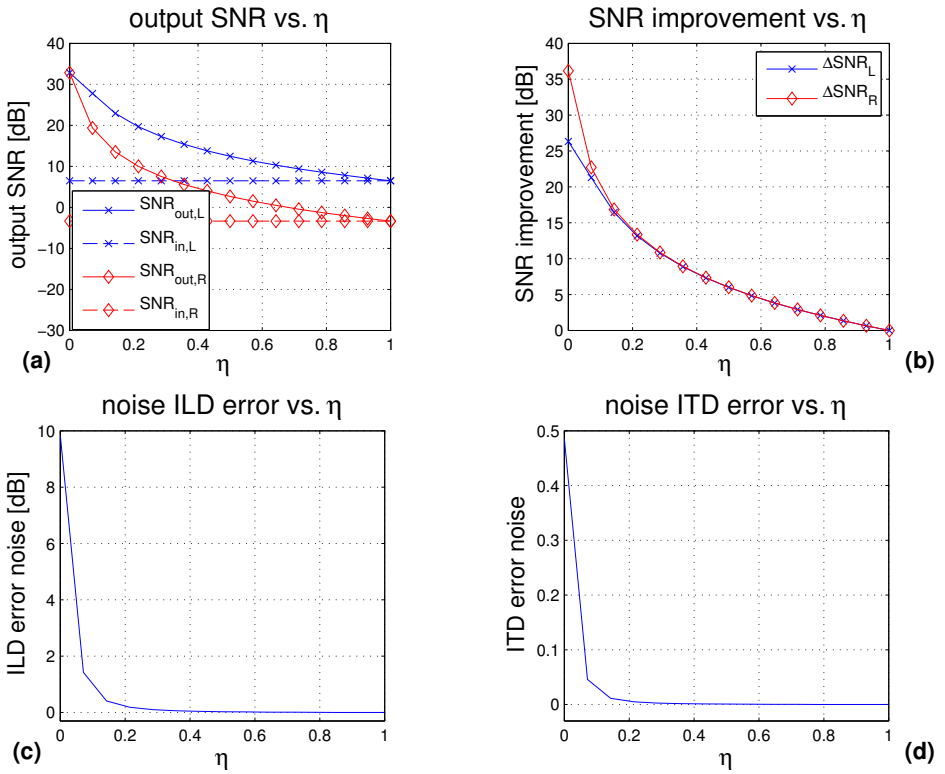


Figure 7.1: Dependence of output SNR (a), SNR improvement (b) and noise binaural cues (ILD (c) and ITD (d)) on partial noise parameter η. Larger values of η lead to lower ILD/ITD errors, but also decrease the noise reduction performance.

narrowband SNR improvement and binaural cue preservation. In these simulations the factor ρ (7.2) is sufficiently large so that the assumptions in section 7.3 are valid and the narrowband SNR improvement (for both hearing aids) is approximately equal to $\Delta\text{SNR}_{L,R} = \frac{1}{\eta^2}$, as in (7.17).

A large loss in narrowband SNR improvement can be seen for a choice of $\eta = 0.2$ compared to the SDW-MWF case ($\eta = 0$). It does not seem possible for the binaural unmasking effect [194] to compensate for this loss. The narrowband SNR improvement obtained in a more realistic setup is however smaller than the theoretical limits shown here, so that it is expected that the real-life absolute SNR loss will be smaller than the loss shown here, and so that the compensatory effect

of the binaural unmasking can be sufficient. Perceptual tests using this parameter setting [177] have indeed shown an improvement in speech intelligibility for some of the speech-noise configurations.

Dependence on signal power P_s

In Figure 7.2 the signal power P_s is varied (dB values are relative to noise power P_v), while P_v and P_{vs} are fixed to 1 and 0.01 respectively. The partial noise parameter η is fixed to 0.2.

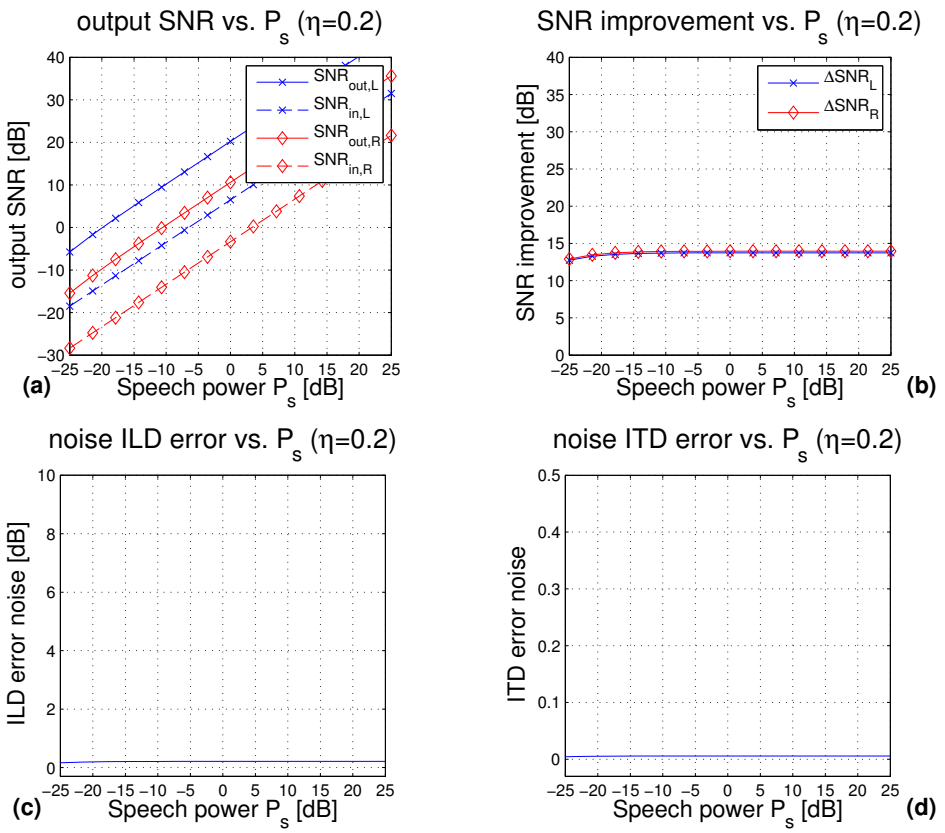


Figure 7.2: Dependence of output SNR (a), SNR improvement (b) and noise binaural cues (ILD (c) and ITD (d)) on signal power P_s. The signal power has only a small influence on the noise reduction performance and on the binaural noise cue preservation.

The results show that for the ILD/ITD errors and for the narrowband SNR improvement, there is no significant dependence on the signal power P_s. The obtained narrowband SNR improvement can be predicted by (7.17), which gives a narrowband SNR improvement of $10\log_{10}\left(\frac{1}{0.2^2}\right) = 13.98$ dB. The ITD/ILD errors are relatively small for all values of P_s, so that $\eta = 0.2$ seems to be sufficient for this scenario. Perceptual tests in [176] indeed show that for $\eta = 0.2$, both speech and noise sources are localized correctly.

Dependence on noise power P_v

In Figure 7.3 the noise power P_v is varied (dB values are relative to signal power P_s), while P_s and P_{vs} are fixed to 1 and 0.01 respectively. The partial noise parameter η is fixed to 0.2.

Higher noise powers P_v lead to lower input and narrowband output SNRs whereas the narrowband SNR improvement is again given by (7.17). The ITD/ILD errors of the noise component are small when the obtained narrowband output SNR is low, which is in fact an advantageous perceptual effect: in frequency bins with a lot of residual noise, this noise will be perceived in the correct direction.

At lower noise powers, the assumptions leading to (7.17) are no longer valid (the input SNRs are high) so that the narrowband SNR improvement is lower than predicted in (7.17). As seen in (7.23), ITF_v^{out} will be shifted towards ITF_x^{in} if the narrowband SNR improvement is low, which leads to higher ILD/ITD errors. However, these higher errors are still acceptable because the narrowband output SNR is high. So, the residual noise will then be masked by the speech signal.

7.5.3 Performance of MWF-ITF

In this section, the performance of the MWF-ITF is evaluated. In [50], it was shown that the algorithm with non-quadratic ITF extension (7.25) allows preservation of both speech as noise binaural cues. However, as for the MWF-η, the obtained narrowband SNR improvement will decrease if more emphasis is put on cue preservation.

Because the MWF-ITF with non-quadratic extension is not computationally feasible for a hearing aid application, we will use the quadratic ITF extension in (7.30) from now on.

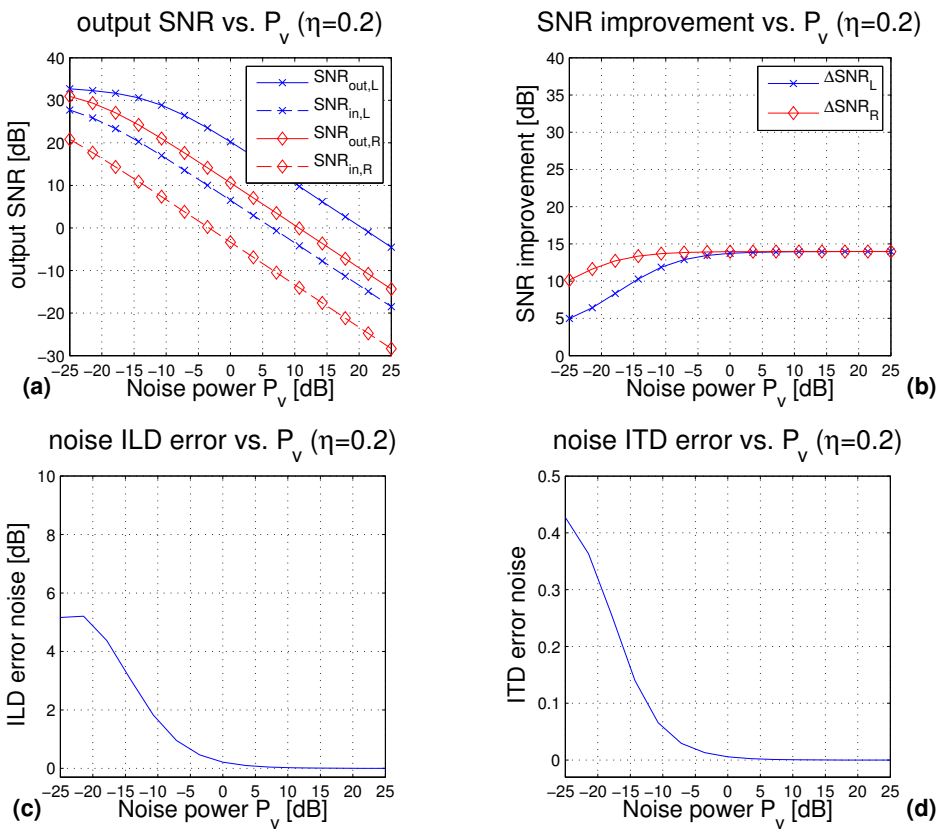


Figure 7.3: Dependence of output SNR (a), SNR improvement (b) and noise binaural cues (ILD (c) and ITD (d)) on noise power P_v. When the output SNR is low, the binaural noise cues are better preserved, which is an advantageous perceptual effect.

ITD and ILD error for quadratic ITF extension

In Figure 7.4 the ILD errors of speech and noise components for the SDW-MWF with quadratic ITF extension (7.30) are shown. The ILD error for the noise component is shown on the left, the ILD error for the speech component is shown on the right. The ITD errors are omitted here, but lead to similar conclusions [30]. For some choices of the ITF parameters γ and δ, the ILD errors on the noise component can be made arbitrarily small. However, the ILD errors on the target speech component will then become large. On the other hand, when the ILD

errors on the target speech component are made small, the ILD errors on the noise component are large. An optimal choice of parameters, where the ILD errors of both noise and speech are small appears to be impossible. These results are in accordance with the theoretical discussion in section 7.4, where it was shown that the output ITFs of speech and noise component are equal, so that speech and noise cues cannot be preserved simultaneously.

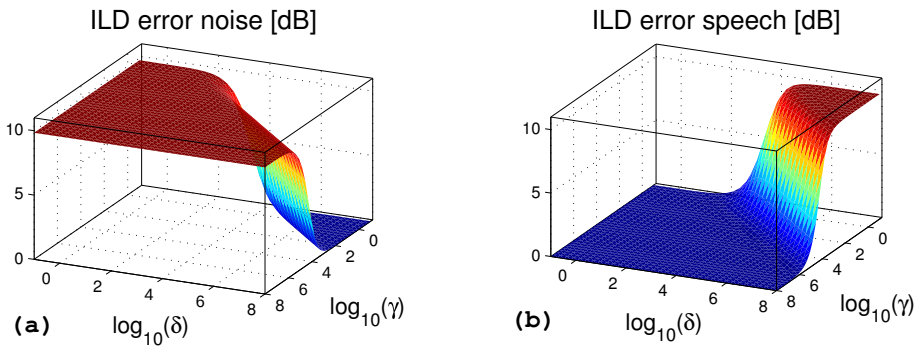


Figure 7.4: MWF-ITF with quadratic ITF extension; The ILD errors for the noise (a) and speech components (b) are shown for different values of γ and δ.

Output ITF versus narrowband output SNR

The theoretical discussion and the previous simulations showed that it is impossible to preserve speech and noise binaural cues simultaneously. As such, this approach may seem inappropriate as a binaural noise reduction algorithm, as it would be impossible to correctly localize both the speech and the noise source. However, in [175] the MWF-ITF algorithm (with quadratic ITF cost terms), was validated perceptually. An improvement in the total localization performance (speech+noise) was observed, which seems to contradict the previous discussion. To explain the improvement, the relationship between the narrowband output SNR and the output ITF will be analyzed.

In Figure 7.5, the narrowband output SNR and the noise ITF parameter δ are varied. γ is fixed to 0 in this simulation. To vary the narrowband output SNR, the signal power P_s will be varied which changes the narrowband output SNR as seen in (7.2). As in the previous section, the noise and speech ILD errors are shown. It can be seen that for certain values of the ITF parameter δ, the ILD error of the noise component is small at low narrowband output SNRs, while the speech ILD error is small at high narrowband output SNRs. The ITD errors, which

are omitted here, behave similarly [30]. These results show that the output ITF is shifted towards the input ITF of the noise component when the narrowband output SNR is low, while the output ITF is shifted towards the input ITF of the speech component in high SNR regions.

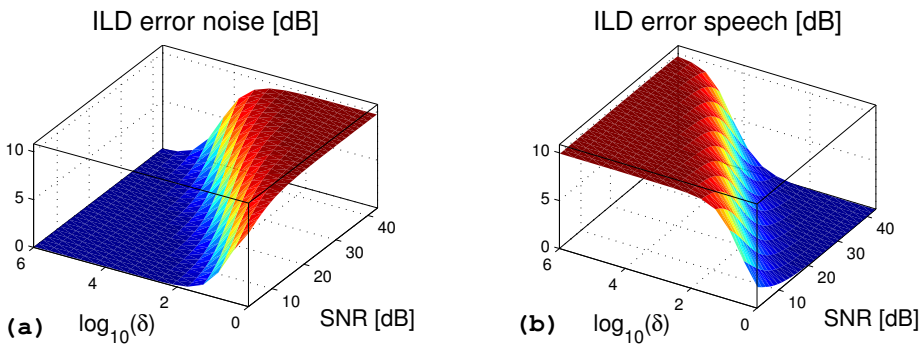


Figure 7.5: MWF-ITF with quadratic ITF extension; The ILD errors for the noise (a) and speech (b) components are shown as a function of the ITF parameter δ and the output SNR.

When the technique is applied on broadband signals, as in [175], the obtained narrowband output SNRs will vary in the different frequency bins, and similarly, the output ITFs will vary. This in fact represents an advantageous perceptual effect: in frequency bins with a low narrowband output SNR, the ITF is shifted towards the noise ITF, so that the residual noise in the output signals can still be heard in the noise direction, and vice versa for the speech component.

Although MWF-ITF thus allows for correct localization, a benefit in speech intelligibility due to binaural unmasking was not observed in the perceptual tests in [175]. This can be expected as the binaural unmasking effect relies on differences in the binaural cues of target and interferer in the same frequency bin, which is not achieved by MWF-ITF. Only when the interfering source is speech, perceiving target and interferer in different locations (for example by the advantageous perceptual effect of MWF-ITF) leads to speech intelligibility improvements [41], but not for speech-shaped noise as interfering source, as was the case in [175].

7.5.4 Comparison of SDW-MWF, MWF-ITF and MWF-η in a reverberant room

The broadband performances of the SDW-MWF, MWF-η and MWF-ITF techniques are now compared in a reverberant environment. HRTFs were measured in a reverberant room ($RT_{60} = 0.61s$ [176, 177]) on a binaural microphone array mounted on a dummy-head, so that the head-shadow effect is taken into account. To generate the microphone signals, the speech and noise signals are convolved with the measured HRTFs corresponding to their angles of arrival ($\theta_x = -15°$, $\theta_v = 45°$), and then added together. To further increase the degree of realism, the correlation matrices are estimated in a batch procedure (assuming a perfect VAD) on the speech and noise signals instead of being constructed with the steering vectors. The correlation matrix estimates are then plugged into (2.53), (7.10) and (7.34) to obtain the optimal filter coefficients of SDW-MWF, MWF-η and MWF-ITF respectively. The speech signal consists of 4 sentences of the English Hearing In Noise Test (HINT) [135], the noise signal is multitalker babble noise [6]. The microphone signals have a total duration of 26 seconds, and are sampled at 8000 Hz. The signals are processed by 64-point DFTs, and the performance measures defined in section 2.1.6 are calculated for every frequency bin independently. The speech distortion trade-off parameter μ is set to 5, the MWF-ITF parameters are $\gamma = 100$ and $\delta = 300$, and the MWF-η parameter is $\eta = 0.2$.

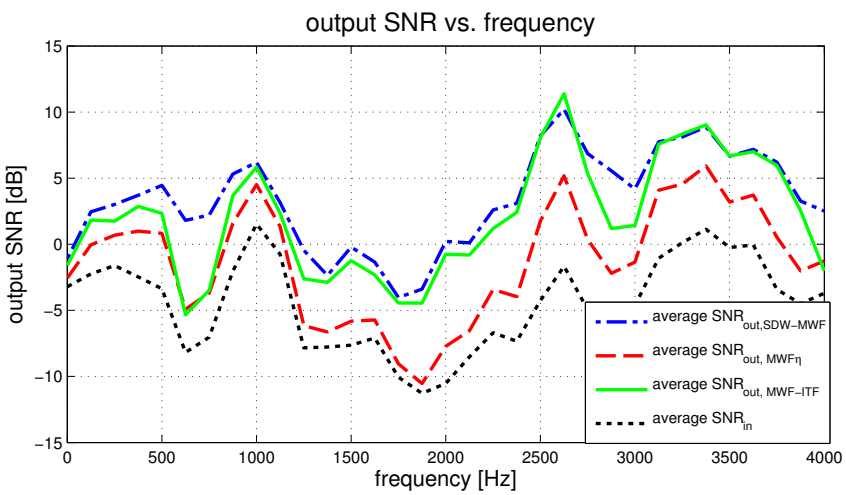


Figure 7.6: Performance comparison of MWF-ITF ($\gamma = 100$ and $\delta = 300$), MWF-η ($\eta = 0.2$) and SDW-MWF in a reverberant room. The average narrowband output SNRs and average narrowband input SNR are shown as a function of frequency.

Figure 7.6 shows the average narrowband output SNR as a function of frequency for the SDW-MWF, MWF-ITF and MWF-η. The SNRs of the left and right output signals were averaged to obtain the average SNRs shown in the figure. As a reference, the (unprocessed) input SNR of this setup is also shown. It can be observed that every procedure achieves a narrowband SNR improvement over the unprocessed case. As expected, the MWF-η obtains the lowest narrowband output SNR. The SDW-MWF and MWF-ITF procedures obtain similar narrowband output SNRs as predicted by the theory, except for a few bins where some dips in the narrowband output SNR curves occur (around 500, 1250, 1750 and 3000 Hz).

Figure 7.7 shows the relative speech and noise ITD errors (2.50) for the three techniques, as a function of frequency. According to (7.8) and (7.19), both SDW-MWF as MWF-η do not distort the speech binaural cues. The speech ITD errors in Figure 7.7 (a) are indeed small for these techniques. In addition, MWF-η is expected to obtain the smallest noise ITD error whereas SDW-MWF is expected to obtain a large noise ITD error, which is illustrated in Figure 7.7 (b). In the theoretical analysis, it was furthermore shown that the output ITF of the MWF-ITF depends on the obtained narrowband output SNR. This advantageous perceptual effect is also illustrated by Figure 7.7: at the frequency bins where dips in the obtained narrowband output SNR in Figure 7.6 occur, a large speech ITD error is seen, whereas the noise ITD error is small. Thus, the output ITF is effectively shifted towards the noise ITF at frequencies with a low narrowband output SNR, which improves localization of the noise source.

7.6 Conclusion

In this chapter, a class of binaural noise reduction techniques has been discussed. To fully exploit the advantage of binaural hearing, the preservation of binaural cues is an important objective, in addition to noise reduction. Perceptual tests in previous work have shown that two extensions of the SDW-MWF, namely MWF-ITF and MWF-η, can lead to localization and speech intelligibility improvements. In this chapter, the binaural cue preservation performance of these techniques has been analyzed theoretically and validated with objective performance measures.

The binaural SDW-MWF changes the binaural cues of the residual noise to equal the cues of the target speech. Both the target speech and residual noise signals will thus be perceived in the direction of the target speech source, so that the hearing aid user cannot rely on binaural unmasking to improve speech intelligibility.

The binaural MWF-η technique is an extension that mixes a fraction of the original noise signal into the output signals. As a drawback, the narrowband SNR improvement decreases. However, this approach makes it possible to preserve both target speech and residual noise binaural cues. These findings support

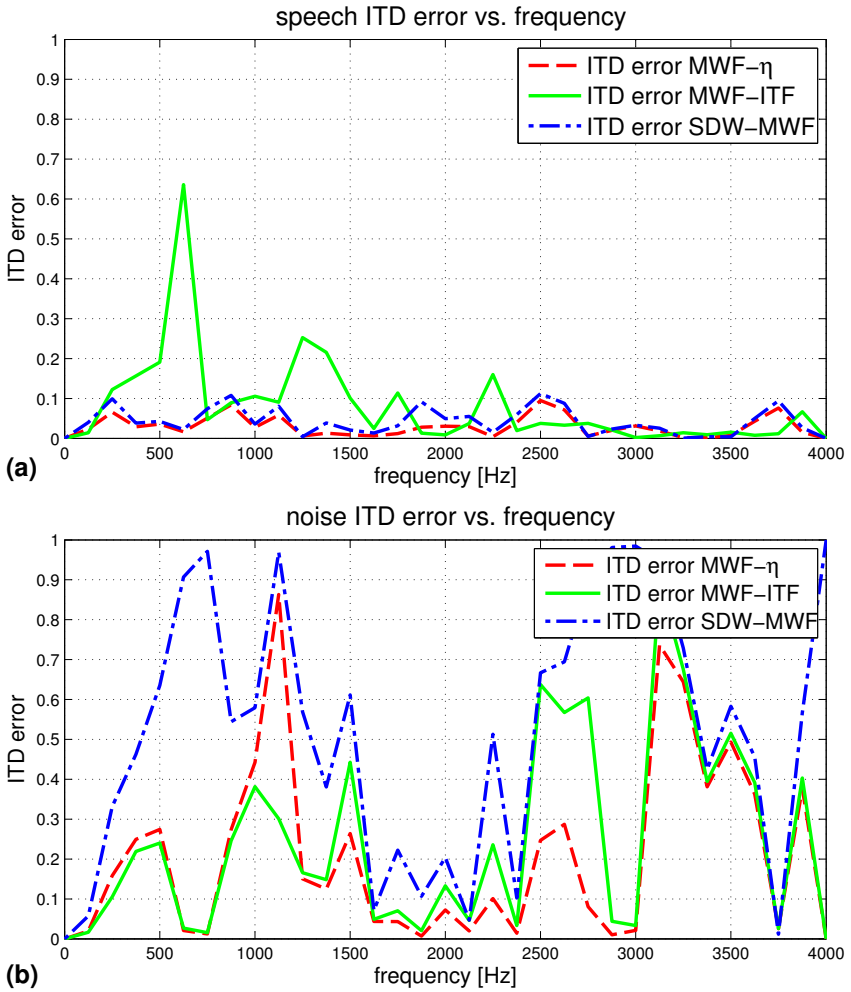


Figure 7.7: Performance comparison of MWF-ITF ($\gamma = 100$ and $\delta = 300$), MWF-η ($\eta = 0.2$) and SDW-MWF in a reverberant room. The speech ITD errors (a) and noise ITD errors (b) are shown as a function of frequency.

the perceptual tests in [176], where an improvement in localization performance was observed. Furthermore, correct localization can lead to a benefit in speech intelligibility if the compensatory effect of binaural unmasking is larger than the SNR loss [177].

The MWF-ITF technique extends the binaural SDW-MWF cost function with terms related to the ITFs of the target speech and noise component. These terms are simplified to quadratic terms to make the algorithm computationally feasible. It has been shown that the obtained narrowband output SNR remains equal to the narrowband output SNR of the binaural SDW-MWF. While it is impossible to preserve the speech and noise cues at the same time, there is an advantageous perceptual effect: if the narrowband output SNR is high, the obtained ITF will be shifted towards the ITF of the target speech component, whereas for a low narrowband output SNR, the ITF will be shifted towards the ITF of the noise component. This effect makes a correct localization possible, and explains the localization performance improvement in prior perceptual tests [175]. As the binaural unmasking effect requires different speech and noise ITFs in a single frequency bin, a speech intelligibility improvement due to binaural unmasking is not possible however. Another drawback of MWF-ITF is the fact that it is not straightforward to find an optimal γ and δ parameter setting for different scenarios.

The advantages and disadvantages of the different techniques are summarized in Table 7.1. The MWF-η appears to be the best choice from a practical point-of-view (versus the MWF-ITF with difficult parameter tuning). Furthermore, the MWF-ITF only allows for correct sound localization but not for binaural unmasking, as noted above. Finally, the MWF-η is also easily applicable if not all microphone signals can be exchanged (reduced-bandwidth), which is discussed in chapter 8.

	SDW-MWF	**MWF-η**	**MWF-ITF**
Advantages	• High noise reduction performance • No parameter tuning • Preservation of speech cues	• Preservation of both speech and noise cues • Easy parameter tuning	• High noise reduction performance • Advantageous perceptual effect
Disadvantages	• Speech and noise components perceived in same direction • Noise cues cannot be preserved	• Lower noise reduction performance	• Speech and noise components with the same ITF's • Difficult parameter tuning

Table 7.1: Comparison of SDW-MWF, MWF-η, MWF-ITF

Chapter 8

Binaural cue preservation in reduced-bandwidth bilateral and binaural SDW-MWF techniques

In the previous chapter it was shown that the SDW-MWF is a suitable choice for binaural noise reduction: the target speech binaural cues are perfectly preserved if all microphone signals can be exchanged, while extensions to the cost function also enable preserving the residual noise cues. However, due to bandwidth limitations it may not be feasible to exchange all microphone signals between the devices so that in practice, reduced-bandwidth techniques have to be utilized. In this chapter, it is shown that a straightforward reduced-bandwidth SDW-MWF technique still preserves the target speech ITD cues, but distorts the target speech ILD cues. These conflicting cues are detrimental to the localization ability of the hearing aid user. Novel reduced-bandwidth SDW-MWF techniques, which significantly reduce the ILD distortion, are therefore introduced. The novel techniques are based on the rank-one SDW-MWF fomulations introduced in chapter 4, which are explicitly structured as an MVDR beamformer followed by a spectral postfilter. By using a common spectral postfilter at the left and at the right device, the ILD distortion can be eliminated without severely degrading the noise reduction performance. Experiments in a reverberant environment confirm the superior performances of the reduced-bandwidth SDW-MWF techniques with common postfilter.

Section 8.1 presents the problem statement and some important preliminary remarks.

Section 8.2 presents a theoretical analysis of the speech ILD distortion introduced by the reduced-bandwidth MWF techniques.

Section 8.3 presents the two novel reduced-bandwidth MWF techniques with common postfilter.

Section 8.4 evaluates the new techniques by experiments in a reverberant environment.

Section 8.5 finally presents the overall conclusions of the chapter.

Part of the material presented in this chapter was published in the proceedings of IWAENC [33]. A journal paper version has been submitted to IEEE Trans. Audio, Speech, Lang. Process. [37].

8.1 Introduction

8.1.1 Problem statement

In chapter 7, the binaural cue preservation by the binaural SDW-MWF and its extensions was studied. If all microphone signals are exchanged between the devices (full-bandwidth), the SDW-MWF preserves the target speech binaural cues, while the MWF-η and MWF-ITF extensions enable preservation of the residual noise binaural cues. Perceptual evaluations have illustrated that both a significant intelligibility improvement [177] and a good localization performance [176] can be achieved with the binaural SDW-MWF and its extensions.

In chapter 7, it was assumed that there are no bandwidth restrictions on the wireless link, so that all microphone signals can indeed be exchanged between the left and the right device. Power consumption will however always be a limiting factor, so that most probably, only a subset of microphone signals can be exchanged in practice. Several reduced-bandwidth techniques have therefore been presented in [55], where only filtered combinations of microphone signals are exchanged (cfr. section 2.2.3 of chapter 2). The problem has also been studied from an information-theoretic point-of-view in [153, 169]. Although these previous works show the potential SNR improvements of the reduced-bandwidth binaural SDW-MWF techniques, the binaural cue preservation of these approaches has not been studied.

In this chapter, the reduced-bandwidth framework of chapter 2 is utilized to study the speech binaural cue preservation of several reduced-bandwidth binaural SDW-MWF techniques. It is shown that the reduced-bandwidth techniques do not necessarily preserve the target speech binaural cues, as was the case for the full-bandwidth binaural SDW-MWF where all microphone signals are exchanged

(chapter 7). In particular, it is shown theoretically and through experiments that the speech ITD cues are preserved by all reduced-bandwidth MWF techniques (even the bilateral SDW-MWF), but that the speech ILD cues can be significantly distorted in certain scenarios. Although a strong cue can dominate over a weaker cue in rivalry experiments, all cues should be kept consistent for optimal sound reproduction [4]: if the conflicts are too large, the resulting source image can be perceived to be diffuse or in the wrong location [4,81]. The speech ILD distortion introduced by the reduced-bandwidth SDW-MWF techniques should therefore be reduced.

It is illustrated that by adapting the speech distortion parameter or by including a partial noise estimation, the ILD distortion can sometimes be reduced, but never completely eliminated. Two novel reduced-bandwidth techniques, which in principle completely eliminate the speech ILD distortion, are therefore introduced, namely a technique for a low-bandwidth binaural link which does not allow for full audio-streaming, and a technique where one microphone signal can be streamed in full-duplex. The techniques make use of the rank-one SDW-MWF fomulations introduced in chapter 4, which are explicitly structured as an MVDR beamformer followed by a spectral postfilter. It is shown theoretically and through experiments that by using a common spectral postfilter at the left and right device, the speech ILD distortion is effectively eliminated.

8.1.2 Preliminaries

Chapter 2 introduced the binaural SDW-MWF (2.53). Chapter 4 introduced alternative formulations, i.e. the R1-MWF (4.8) and SP-MWF (4.17), which implicitly assume a single target speech source. Under this assumption, (4.8) and (4.17) are theoretically equivalent to (2.53), but it was demonstrated in chapters 4 and 6 that implementations based on (4.8) and especially (4.17) are numerically favorable. As the SP-MWF will be utilized in this chapter, it is reformulated here for easy referencing, i.e.

$$\boxed{\begin{aligned}\mathbf{w}_{\mathrm{SP-MWF},L} &= \mathbf{R}_{vL}^{-1}\mathbf{R}_{xL}\mathbf{e}_L \cdot \frac{\mathbf{e}_L^H\mathbf{R}_{xL}\mathbf{e}_L}{\mu\ \mathbf{e}_L^H\mathbf{R}_{xL}\mathbf{e}_L + \mathrm{Tr}\{\mathbf{R}_{vL}^{-1}\mathbf{R}_{xL}\mathbf{e}_L\mathbf{e}_L^H\mathbf{R}_{xL}\}}\,, \\ \mathbf{w}_{\mathrm{SP-MWF},R} &= \mathbf{R}_{vR}^{-1}\mathbf{R}_{xR}\mathbf{e}_R \cdot \frac{\mathbf{e}_R^H\mathbf{R}_{xR}\mathbf{e}_R}{\mu\ \mathbf{e}_R^H\mathbf{R}_{xR}\mathbf{e}_R + \mathrm{Tr}\{\mathbf{R}_{vR}^{-1}\mathbf{R}_{xR}\mathbf{e}_R\mathbf{e}_R^H\mathbf{R}_{xR}\}}\,.\end{aligned}} \tag{8.1}$$

In chapter 7, the binaural SDW-MWF cost function was extended by means of a so-called partial noise estimation parameter η in order to preserve the residual

noise binaural cues. The optimal so-called MWF-η filters are equal to:

$$\boxed{\begin{array}{rcl} \mathbf{w}_{\mathrm{MWF}\eta,L} & = & (1-\eta)\ \mathbf{w}_{\mathrm{MWF},L} + \eta\ \mathbf{e}_L\ , \\ \mathbf{w}_{\mathrm{MWF}\eta,R} & = & (1-\eta)\ \mathbf{w}_{\mathrm{MWF},R} + \eta\ \mathbf{e}_R\ . \end{array}} \tag{8.2}$$

The MWF-η thus corresponds to a mixing of the SDW-MWF output signal (of which the noise binaural cues are distorted) with the unprocessed reference microphone signals (with correct binaural cues). Expression (8.2) can be applied to any of the presented filter formulations, i.e. the standard SDW-MWF (2.53), R1-MWF (4.8) or SP-MWF (8.1).

In the case of a single target speech source, (2.53), (4.17) and (8.1) are all equivalent and equal to

$$\mathbf{w}_{\mathrm{MWF},L} = \mathbf{R}_{vL}^{-1}\mathbf{a}_L\ \frac{P_s}{\mu+\rho_L} A_{L,\mathrm{ref}}^*\ , \tag{8.3}$$

$$\mathbf{w}_{\mathrm{MWF},R} = \mathbf{R}_{vR}^{-1}\mathbf{a}_R\ \frac{P_s}{\mu+\rho_R} A_{R,\mathrm{ref}}^*\ , \tag{8.4}$$

with

$$\rho_L = P_s \mathbf{a}_L^H \mathbf{R}_{vL}^{-1}\mathbf{a}_L\ , \tag{8.5}$$

$$\rho_R = P_s \mathbf{a}_R^H \mathbf{R}_{vR}^{-1}\mathbf{a}_R\ . \tag{8.6}$$

Using definition (2.25), the narrowband[1] output SNRs obtained at the left and right device are also equal to ρ_L and ρ_R, i.e.

$$\mathrm{SNR}_L^{out} = P_s \mathbf{a}_L^H \mathbf{R}_{vL}^{-1}\mathbf{a}_L = \rho_L\ , \tag{8.7}$$

$$\mathrm{SNR}_R^{out} = P_s \mathbf{a}_R^H \mathbf{R}_{vR}^{-1}\mathbf{a}_R = \rho_R\ . \tag{8.8}$$

As this chapter focuses on the speech binaural cue preservation for a single target speech source, a single analysis using (8.3) and (8.4) can be made for all filter formulations.

As shown in [157,161] and (2.59), formulas (8.3) and (8.4) can be decomposed into a spatial filter, which is equivalent to the MVDR beamformer [8,21] followed by a (single-channel) spectral postfilter, i.e.

$$\mathbf{w}_{\mathrm{MWF},L} = \underbrace{\mathbf{R}_{vL}^{-1}\mathbf{a}_L\ \frac{A_{L,\mathrm{ref}}^*}{\mathbf{a}_L^H\mathbf{R}_{vL}^{-1}\mathbf{a}_L}}_{\mathrm{MVDR}} \cdot \underbrace{\frac{\rho_L}{\mu+\rho_L}}_{\mathrm{postfilter}}\ , \tag{8.9}$$

[1] It is again noted that the term *narrowband SNR* indicates that the SNR is calculated in a single frequency bin, cfr. Section 2.1.5. It is not related to the bandwidth of the wireless link, and should be distinguished from the term *reduced-bandwidth SDW-MWF*.

and similarly for $\mathbf{w}_{\mathrm{MWF},R}$. As (8.9) requires explicit estimation (or prior knowledge) of the speech power P_s and of the steering vectors $\mathbf{a}_L$ and $\mathbf{a}_R$, it does not allow for an easy practical implementation. However, the R1-MWF (4.8) and SP-MWF (8.1) are also explicitly decomposed into a spatial filter and spectral postfilter (see also section 8.3.1), but still only require a relatively easy estimation of the speech and noise correlation matrices as for the general SDW-MWF (2.53). Hence, they provide (equivalent) practical formulas to compute the spatial filter and spectral postfilter. It will be shown in the following that the decomposed structure of the R1-MWF and in particular the SP-MWF can be exploited in order to better preserve the speech ILD cues.

8.2 Speech binaural cue preservation by reduced bandwidth SDW-MWF and MWF-η

8.2.1 Output speech binaural cues

By using the general framework of section 2.1.2, and by assuming a single target speech source as in section 2.1.4, the output speech ITF, ILD and ITD can be calculated for any $\mathbf{C}_L$ and $\mathbf{C}_R$. By plugging the optimal filters (8.3) and (8.4) into the ITF, ILD and ITD definitions of section 2.1.6 and using definitions (8.5), (8.6), we obtain:

$$\mathrm{ITF}_x^{out} = \frac{\frac{P_s}{\mu+\rho_L}\ \mathbf{a}_L^H\mathbf{R}_{vL}^{-1}\mathbf{a}_L\ A_{L,\mathrm{ref}}}{\frac{P_s}{\mu+\rho_R}\ \mathbf{a}_R^H\mathbf{R}_{vR}^{-1}\mathbf{a}_R\ A_{R,\mathrm{ref}}} = \frac{1+\frac{\mu}{\rho_R}}{1+\frac{\mu}{\rho_L}}\ \mathrm{ITF}_x^{in}\ , \tag{8.10}$$

$$\mathrm{ILD}_x^{out} = \frac{(1+\frac{\mu}{\rho_R})^2}{(1+\frac{\mu}{\rho_L})^2}\ \mathrm{ILD}_x^{in}\ , \tag{8.11}$$

$$\mathrm{ITD}_x^{out} = \angle\left(\frac{P_s\ \rho_L\ \rho_R}{(\mu+\rho_L)(\mu+\rho_R)}A_{L,\mathrm{ref}}A_{R,\mathrm{ref}}^*\right) = \mathrm{ITD}_x^{in}\ . \tag{8.12}$$

Formula (8.12) shows that the speech ITD cues are never distorted by the reduced-bandwidth SDW-MWF techniques, regardless of which signals are exchanged over the wireless link, or even for the bilateral case where no signals are exchanged (2.9), (2.10). Formula (8.11) however shows that the output speech ILD is scaled compared to the input speech ILD, and so that the speech ILD cues are indeed distorted by the reduced-bandwidth SDW-MWF.

For the special case $\rho_L = \rho_R$, which means that the obtained left and right output SNRs are equal, formula (8.11) shows that the speech ILD cues are also undistorted. This is e.g. the case if all microphone signals can be exchanged over the wireless

link, i.e. $\mathbf{C}_L = \mathbf{C}_R = \mathbf{I}_{2M}$. This result matches the result obtained in chapter 7, i.e. formula (7.8).

However, if not all microphone signals are exchanged ($N < M$), then in general $\rho_L \neq \rho_R$, leading to ILD distortions. By increasing the number of exchanged signals, the obtained output SNRs ρ_L and ρ_R generally increase while the mismatch $|\rho_L - \rho_R|$ decreases [55]. As a result, the ILD distortion will also decrease (cfr. Appendix C.1). It can therefore be concluded that, in addition to a better noise reduction performance, a configuration using a binaural link will generally introduce a lower speech ILD distortion compared to a bilateral configuration.

8.2.2 Influence of μ

As can be seen from (8.11), $\text{ILD}_x^{out} = \text{ILD}_x^{in}$ for $\mu = 0$. From (8.9) it is seen that this parameter choice corresponds to an MVDR where the speech distortion is enforced to 0, so that the speech binaural cues will indeed not be distorted. For $\mu > 0$, the reduced-bandwidth SDW-MWF performs additional spectral postfiltering, so that the residual noise is further attenuated. Although the narrowband output SNRs (8.7) and (8.8) are independent of μ, the broadband output SNRs can then be increased as in single-channel spectral subtraction algorithms [118], thereby increasing listening comfort. It can be shown (cfr. Appendix C.1) that the speech ILD error strictly increases as μ increases. Improving the noise reduction performance therefore comes at a cost: as more emphasis is put on noise reduction, the speech ILD cues will be more distorted.

In previous perceptual evaluations of the (full-bandwidth) binaural MWF [176, 177], μ was always fixed to a constant value (i.e. $\mu = 5$), in each time-frequency point. As explained, this parameter choice leads to speech ILD distortion in a reduced-bandwidth scheme. The ILD distortion can however be reduced by decreasing μ in time-frequency points with a high speech power, while larger values of μ are kept in time-frequency points with a low speech power (so that the noise is still sufficiently attenuated). By making μ dependent on the estimated segmental SNR (as in [11]), so that the spectral postfilter effectively becomes similar to classical single-channel spectral subtraction, the speech ILD distortion could thus be reduced.

8.2.3 Influence of η

By plugging (8.2)-(8.4) into the performance measures of section 2.1.6, we again find that the speech ITD is undistorted by the MWF-η, while the output speech

ILD is equal to

$$\mathrm{ILD}_x^{out} = \frac{(1+\frac{\mu}{\rho_R})^2\ (1+\frac{\eta\mu}{\rho_L})^2}{(1+\frac{\mu}{\rho_L})^2\ (1+\frac{\eta\mu}{\rho_R})^2}\ \mathrm{ILD}_x^{in}\ . \tag{8.13}$$

For $\eta = 1$, the MWF-η output signals are equal to the reference microphone signals so that $\mathrm{ILD}_x^{out} = \mathrm{ILD}_x^{in}$, whereas for $\eta = 0$, formula (8.11) is obviously obtained. In Appendix C.1, it is shown that for intermediate values of η, the ILD error strictly decreases if η increases.

In addition to better preserving the residual noise binaural cues, the MWF-η thus also reduces the speech ILD error, at the cost of a lower noise reduction performance. In [176, 177], a (small) value ($\eta = 0.2$) was proposed as a good parameter setting: the residual noise ILD and ITD errors are small, while the noise reduction performance is still adequate. For this setting, the speech ILD error can however not be eliminated, as this would only be achieved for the limit case $\eta = 1$, where no noise reduction is performed. Furthermore, if there are no clear noise ITD/ILD cues (for example in a diffuse noise field), it would also not be appropriate to use the MWF-η solution. The MWF-η by itself is therefore not a sufficient solution for eliminating the speech ILD distortion.

8.3 Reduced-bandwidth SDW-MWF with common postfilter

8.3.1 Introduction

The previous analysis reveals that the speech ILD distortion is introduced by the spectral postfilter of the SDW-MWF, which depends on the speech distortion parameter μ (8.9). The ILD distortion can be reduced by adapting μ based on the segmental SNR and/or by using a partial noise estimation extension (MWF-η), but with these approaches it is not possible to completely eliminate the distortion. Moreover, the noise reduction performance may be affected by these approaches, as μ may be erroneously adapted when the segmental SNR is poorly estimated, or because some of the unprocessed (low SNR) signal is mixed with the MWF output signal by the MWF-η.

An alternative solution, where the speech ILD distortion can be completely eliminated without significantly compromising the noise reduction performance, is to use the same spectral postfilter at the left and right device, which is then applied to the outputs of the MVDR-filtered left and right output signals. This approach is somewhat similar to [97, 108, 121, 148, 192], where a common spectral postfilter is applied to the raw microphone signals of the left and right device. The difference is that in the proposed approach, an MVDR beamforming, which utilizes

all the ipsilateral microphone signals and possibly also contralateral signal(s), is performed at both devices separately in a first stage. This MVDR stage is then followed by the common spectral postfiltering stage. In [97,108,121,148,192], only a spectral filtering of one microphone signal in the left device and one microphone signal in the right device is performed, whereas no beamforming is performed in the direct signal path.

Two common postfilter (CP) techniques, where the same spectral postfilter part (8.9) is used in the left and right device, will be proposed in the following subsections. As previously mentioned, formula (8.9) does not allow for an easy implementation, while the standard SDW-MWF expressions (2.53) do not allow for manipulating the postfilter part. The R1-MWF (4.8) or SP-MWF (8.1) formulas provide good alternatives as these only rely on the speech and noise SOS as in (2.53), but are also explicitly structured as a spatial filter followed by a spectral postfilter as in (8.9). Of the two alternatives, the SP-MWF was found to be superior to the R1-MWF in terms of speech distortion (cfr. chapter 4), especially when the rank-one model (2.23) is not satisfied. The SP-MWF is therefore used in the remainder of this chapter.

The SP-MWF (8.1) can be structured as an MVDR (by setting $\mu = 0$) followed by a postfilter depending on μ as in (8.9), which may then be replaced by a CP in the left and right device, leading to

$$\mathbf{w}_{\text{SP-MWF-CP},L} = \mathbf{R}_{vL}^{-1}\mathbf{R}_{xL}\mathbf{e}_L \cdot \frac{\mathbf{e}_L^H\mathbf{R}_{xL}\mathbf{e}_L}{\text{Tr}\{\mathbf{R}_{vL}^{-1}\mathbf{R}_{xL}\mathbf{e}_L\mathbf{e}_L^H\mathbf{R}_{xL}\}} \cdot \text{CP}\ (\mu)\ , \quad (8.14)$$

$$\mathbf{w}_{\text{SP-MWF-CP},R} = \underbrace{\mathbf{R}_{vR}^{-1}\mathbf{R}_{xR}\mathbf{e}_R \cdot \frac{\mathbf{e}_R^H\mathbf{R}_{xR}\mathbf{e}_R}{\text{Tr}\{\mathbf{R}_{vR}^{-1}\mathbf{R}_{xR}\mathbf{e}_R\mathbf{e}_R^H\mathbf{R}_{xR}\}}}_{\text{MVDR}} \cdot \text{CP}\ (\mu)\ , \quad (8.15)$$

We propose two schemes for constructing the CP, namely a scheme for a low-bandwidth binaural link which does not allow to exchange microphone signals between the devices (bilateral MWF with CP), and a scheme where one microphone signal can be exchanged between the devices (binaural MWF-Front with CP).

8.3.2 Bilateral MWF with common postfilter

If the binaural link does not allow to exchange microphone signals between the devices (cfr. formulas (2.9) and (2.10)), it can still be utilized to derive a CP for the two devices. We consider the scheme of Figure 8.1, where only the ipsilateral microphone signals are available for bilateral MVDR filtering (MWF with $\mu = 0$), but where an estimate of the narrowband output SNRs is exchanged between the

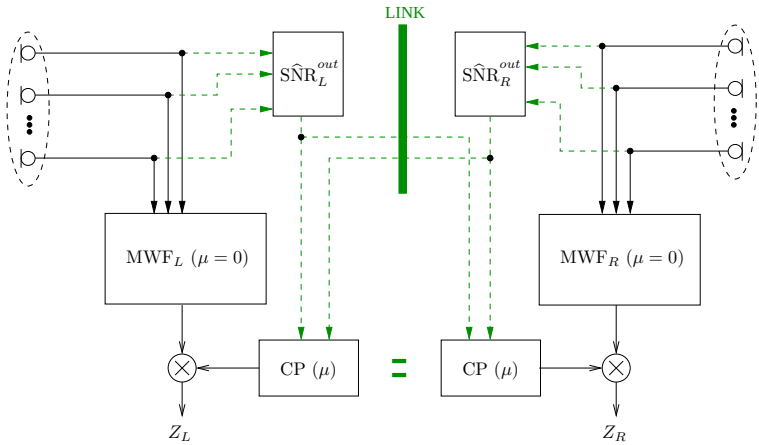


Figure 8.1: Bilateral MWF with common postfilter (CP)

devices. The CP is then derived in each frequency bin as:

$$\text{CP }(\mu) = \frac{\max\{\widehat{\text{SNR}}_L^{out},\ \widehat{\text{SNR}}_R^{out}\}}{\mu + \max\{\widehat{\text{SNR}}_L^{out},\ \widehat{\text{SNR}}_R^{out}\}}\ , \tag{8.16}$$

where the estimation of $\widehat{\text{SNR}}_L^{out}$ and $\widehat{\text{SNR}}_R^{out}$ is a by-product of the MVDR in (8.14) and (8.15), i.e.

$$\widehat{\text{SNR}}_L^{out} = \frac{\text{Tr}\{\mathbf{R}_{vL}^{-1}\mathbf{R}_{xL}\mathbf{e}_L\mathbf{e}_L^H\mathbf{R}_{xL}\}}{\mathbf{e}_L^H\mathbf{R}_{xL}\mathbf{e}_L}\ , \tag{8.17}$$

$$\widehat{\text{SNR}}_R^{out} = \frac{\text{Tr}\{\mathbf{R}_{vR}^{-1}\mathbf{R}_{xR}\mathbf{e}_R\mathbf{e}_R^H\mathbf{R}_{xR}\}}{\mathbf{e}_R^H\mathbf{R}_{xR}\mathbf{e}_R}\ . \tag{8.18}$$

By selecting the maximum output SNR estimate in (8.16), the performance at the best-ear side (which is crucial for intelligibility) is not compromised, whereas the noise will be less attenuated at the worst-ear side (compared to the case where each side has its own postfilter).

The required bandwidth for the bilateral MWF with CP can be kept low in a number of ways:

- The CP is only enforced at a subset of frequencies, for example only at frequencies above 1000 Hz (as ILD cues are then the dominant cues [14]).
- The output SNRs are exchanged and updated at a rate lower than the frame rate, i.e. the same SNR value is used in a number of subsequent frames.

- The frequency bins are grouped in bands, where one average output SNR estimate is used for an entire band, thus reducing the number of exchanged values.
- If only simplex instead of duplex transmission is feasible, a CP can still be obtained if the receiving device only uses the contralateral SNR estimates instead of its own SNR estimates to compute its postfilter. The approach can be improved by monitoring which is the best-ear side (for example based on simple broadband SNR estimates), so that the device at the best-ear side can be chosen as transmitter.

8.3.3 Binaural MWF-Front with common postfilter

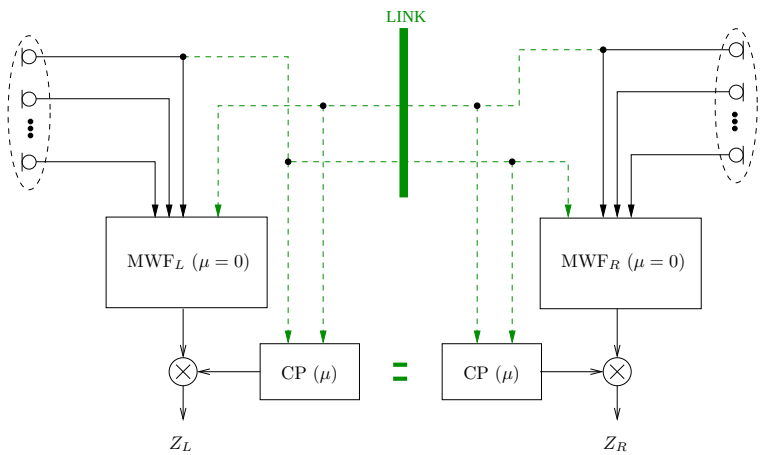


Figure 8.2: Binaural MWF-Front with common postfilter (CP)

In Figure 8.2, a binaural reduced-bandwidth SDW-MWF technique with CP is depicted. In this scheme, the front microphone signal of each device is transmitted to the other device (in full-duplex). This was referred to as MWF-Front in [55] and section 2.2.3, and corresponds to

$$\mathbf{F}_{10}^H = \mathbf{F}_{01}^H = \Big[1 \underbrace{0 \dots 0}_{M-1}\Big] \tag{8.19}$$

in the general framework of section 2.1.2. Each device can thus construct an $(M+1)$-microphone MVDR using its own M microphone signals and the front microphone signal of the other device.

As both devices have access to both front microphone signals, each device can additionally compute a two-microphone MVDR with these two front microphone signals (whereby each device selects the same best-ear side microphone signal as reference). The narrowband output SNR of this two-microphone MVDR can then be computed in both devices as:

$$\widehat{\mathrm{SNR}}_{\mathrm{front}}^{out} = \frac{\mathrm{Tr}\{\mathbf{R}_{v,\mathrm{front}}^{-1}\mathbf{R}_{x,\mathrm{front}}\mathbf{e}_{\mathrm{front}}\mathbf{e}_{\mathrm{front}}^{H}\mathbf{R}_{x,\mathrm{front}}\}}{\mathbf{e}_{\mathrm{front}}^{H}\mathbf{R}_{x,\mathrm{front}}\mathbf{e}_{\mathrm{front}}} , \tag{8.20}$$

where $\mathbf{e}_{\mathrm{front}}$ selects the best-ear side front microphone signal as reference, and where $\mathbf{R}_{v,\mathrm{front}}$ and $\mathbf{R}_{x,\mathrm{front}}$ are defined as:

$$\mathbf{R}_{v,\mathrm{front}} = \mathcal{E}\{\begin{bmatrix} V_{0,0} \\ V_{1,0} \end{bmatrix} . \begin{bmatrix} V_{0,0}^{H} & V_{1,0}^{H} \end{bmatrix}\}, \quad \mathbf{R}_{x,\mathrm{front}} = \mathcal{E}\{\begin{bmatrix} X_{0,0} \\ X_{1,0} \end{bmatrix} . \begin{bmatrix} X_{0,0}^{H} & X_{1,0}^{H} \end{bmatrix}\} . \tag{8.21}$$

The CP is then calculated in each device as

$$\mathrm{CP}\ (\mu) = \frac{\widehat{\mathrm{SNR}}_{\mathrm{front}}^{out}}{\mu + \widehat{\mathrm{SNR}}_{\mathrm{front}}^{out}} . \tag{8.22}$$

By making use of the QRD-RLS algorithm of chapter 6, it is not even required to calculate a separate two-microphone SNR estimate as in (8.20) to obtain the CP. In Appendix C.2, it is shown that the CP can be extracted from the $(M+1)$-microphone MVDR solutions with only a few simple operations.

It is noted that the required bandwidth can be further reduced if only the low frequency part (e.g. up to 4000 Hz) of the front microphone signals is exchanged as in [167]. Although the bilateral array achieves sufficient noise reduction at the frequencies above 4000 Hz [167], it can still introduce a significant speech ILD distortion, which is moreover the dominant localization cue at these frequencies [14]. At the frequencies above 4000 Hz, the low-bandwidth bilateral MWF with CP of section 8.3.2 should therefore be used.

8.4 Experimental results

8.4.1 ILD estimation

In chapter 7, the speech input and output ILDs were measured using the (theoretical) performance measures of section 2.1.6. These measures quantify the narrowband, long-term average ILD. As the effect of different spectral postfilters and their introduced frame-by-frame ILD fluctuations are now also under investigation, a short-term per-frame ILD estimate is needed. Unfortunately,

such a per-frame ILD estimate is unreliable, especially in the presence of reverberation [62]. As suggested in [62, 94], a solution for reducing the estimation variance is to only select the frames with an interaural coherence above a certain threshold, which is justified by the precedence effect[2] [14]. The determination of the optimal threshold is non-trivial as it depends on both the frequency and the direction-of-arrival of the source signal [94]. It was however illustrated in [94] that a good estimate for the threshold can be obtained by fixing it to the 90th percentile of the interaural coherence values, and this procedure will therefore be utilized in this chapter.

In addition to the per-frame ILD calculation and selection, a model of the middle and inner ear is also incorporated so that the physical ILD levels are converted to physiological levels. As in [94], the input and output signals are resampled to 16 kHz, divided into frames of 20 ms and decomposed into 24 critical bands using a Gammatone cochlear filterbank [142] with center frequencies according to the Glasberg and Moore model [71]. As in [62], each critical band signal is subsequently processed using the envelope compression model of [10], in order to perform a neural transduction. The different models were implemented using the downloadable binaural toolboxes from Akeroyd [3] and Slaney [158]. The ILD (in dB) per frame is then equal to (cfr. [62] and [94]):

$$\text{ILD}\ [i,k] = 10\ \log_{10}\left(\frac{E_L[i,k]}{E_R[i,k]}\right)\ , \tag{8.23}$$

where i is the critical band index, k is the frame index, and where $E_L[i,k]$ and $E_R[i,k]$ are the smoothed energies of the left and right compressed critical band signals [62]. By this definition, a source arriving from the left side of the head obtains a positive ILD (in dB).

8.4.2 Setup and stimuli

A binaural hearing aid configuration with two behind-the-ear devices is considered. Each device has two omnidirectional microphones ($M = 2$), with an intermicrophone distance of approximately 1 cm.

Head-related transfer functions (HRTFs) were measured in a reverberant room ($\text{RT}_{60} = 0.61s$ [176, 177]) on a CORTEX MK2 manikin, so that the head-shadow effect is taken into account. Four spatial scenarios are considered, i.e. S0N60, S0N120, S0 N[60/120/180/210] and S90N270, where the target speech (S) and interfering noise (N) source(s) are positioned at the specified azimuthal angles (with 0° in front of the head, 90° to the right of the head). To generate the

[2]Precedence effect or law of the first wavefront: if a sound arrives at a listener by different paths (e.g. because of wall reflections), only the direction of the first arriving signal is perceived.

microphone signals, the noise and speech signals were convolved with the HRTFs corresponding to their angles of arrival, before being added together.

In all experiments, five consecutive sentences (with five-second periods of silence between sentences) of the English Hearing-In-Noise Test (HINT) [135] were used as speech stimulus. Multitalker babble (Auditec [6]) is used as noise stimulus. In the scenario with multiple noise sources, different time-shifted versions of this signal were generated to obtain uncorrelated noise sources. The signals were scaled so that the input SNR is 0 dB in the reference microphone signal of the best ear (left ear in S0N60, S0N120 and S0 N[60/120/180/210], right ear in S90 N270).

To assess the performance (both SNR improvement as speech ILD distortion), only the data frames corresponding to the last three sentences are selected, to allow the filters to converge.

8.4.3 Algorithms

Notation	Description
BIL−μ=5	Bilateral MWF ($M = 2$, $N = 0$), fixed μ
BIL−μ=5−CP	Bil. MWF with common postfilter, fixed μ
BIL−μ=adapt	Bil. MWF, adaptive μ
BIL−μ=5 − η	Bil. MWF, fixed μ, partial noise estimation
FRONT−μ=5	Binaural MWF-Front ($M = 2$, $N = 1$), fixed μ
FRONT−μ=5−CP	Bin. MWF-Front with common postfilter, fixed μ
FRONT−μ=adapt	Bin. MWF-Front, adaptive μ
FRONT−μ=5 − η	Bin. MWF-Front, fixed μ, partial noise estimation
ALL−μ=5	Bin. MWF-All, ($M = N = 2$)

Table 8.1: Algorithms

Nine different algorithms are considered, a description is found in Table 8.1. The algorithms denoted with *BIL* (=bilateral) assume that exchanging microphone signals is not possible, so that each device can only use its own two microphone signals for noise reduction. The algorithms denoted with *FRONT* assume that the front microphone signals can be exchanged, so that three microphone signals are available for noise reduction in each device. The algorithm denoted with *ALL* is the reference condition where all microphone signals are available in each device, as in chapter 7. The speech distortion parameter is always fixed to $\mu = 5$, except for two algorithmic conditions where μ is adapted according to the segmental SNR as in [11]. To improve the performance for colored noise, the segmental SNR is estimated in different frequency bands as in [96], so that a different μ can be used in different bands. The implementation of this multi-band approach was based on the code provided with [118]. The frequency spectrum was hereby divided in eight equal bands (bandwidth of 1250 Hz) as in [96]. Furthermore, in two other

algorithmic conditions, a partial noise estimation is performed as in (8.2), where the partial noise parameter is set to $\eta = 0.2$ as in [176, 177]. Finally, in two algorithmic conditions the proposed common postfilter (CP) techniques of section 8.3.2 (for bilateral MWF) and section 8.3.3 (for MWF-Front) are tested. For the bilateral CP, we keep in mind that the required bandwidth can be reduced as explained in section 8.3.2, although these alternatives are not tested in these experiments. To derive the MWF-Front CP, the efficient scheme described in Appendix C.2 is used.

In all experiments, the signals are sampled at $f_s = 20480$ Hz and the filter length (=DFT size) is set to $L = 128$. The algorithms are implemented in a Weighted Overlap-Add (WOLA) filterbank [40], whereby the microphone signals are segmented into frames of L samples with 50% overlap, and windowed by a Hann window.

8.4.4 Batch versus adaptive implementation

In chapter 7, a batch implementation of the SDW-MWF was evaluated. In the batch algorithm, the correlation matrices are estimated in an off-line procedure using the complete microphone signals, which results in an optimal average filter. It is however possible to recursively update the correlation matrices and filters in an online procedure, which allows tracking a moderately changing spatial scenario [54]. In this adaptive SDW-MWF algorithm, the correlation matrix estimates $\hat{\mathbf{R}}_{vL}$ and $\hat{\mathbf{R}}_{yL}$ at the left device (similarly for the right device) are updated as indicated by expressions (2.16)-(2.17) in section 2.1.3, where the exponential forgetting factors are set to $\lambda_v = \lambda_y = 0.999$. The speech correlation matrix estimates are found as $\hat{\mathbf{R}}_{xL} = \hat{\mathbf{R}}_{yL} - \hat{\mathbf{R}}_{vL}$ and $\hat{\mathbf{R}}_{xR} = \hat{\mathbf{R}}_{yR} - \hat{\mathbf{R}}_{vR}$.

A perfect VAD will be assumed in these simulations. Although estimation errors because of VAD errors are therefore eliminated, it is still possible that the speech correlation matrix is badly estimated at certain frequencies, e.g. because of speech absence [160]. As a result, the values in the denominators of the R1-MWF or SP-MWF might decay to zero, so that they have to be kept above certain thresholds [160]. When using the MDVR spatial filter parts of (8.14), (8.15), it is also possible to use precalibrated steering vectors $\mathbf{a}_L$ and $\mathbf{a}_R$ instead of $\mathbf{R}_{xL}\mathbf{e}_L$ and $\mathbf{R}_{xR}\mathbf{e}_R$, at the time-frequency points where the left or right speech correlation matrices are badly estimated (a bad estimation is detected if any of the diagonal entries of $\mathbf{R}_{xL}$ or $\mathbf{R}_{xR}$ are negative). The MVDR does not depend on the speech power P_s as can be seen from (8.9), which simplifies the precalibration. To construct the steering vector, the HRTF corresponding to 0° is used. As in practice the actual reverberation time of the environment will be unknown, an HRTF measured in an anechoic room was used for the precalibration.[3]

[3]Simulations indicate that there is no significant performance difference if a steering vector, calibrated in the actual reverberant environment, is used instead.

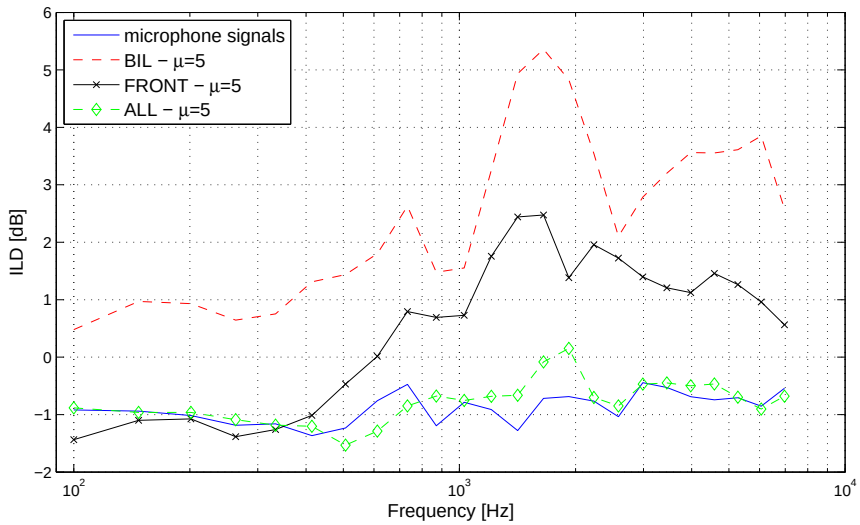
8.4.5 Speech ILD per critical band and normalized cross-correlation function

In Figure 8.3 (a), the obtained mean ILDs per critical band are shown for the (batch) algorithms BIL$-\mu = 5$ (MWF-bil), Front$-\mu = 5$ (MWF-Front) and ALL$-\mu$=5 (MWF-All), for the S0 N60 spatial scenario. The ILD was measured on a frame-by-frame basis as outlined in section 8.4.1, whereby the average standard deviations in the ILD estimation are indicated in the figure caption. The ILD measured on the unprocessed front microphone signals is also shown as a reference. It can be seen that the bilateral MWF obtains a high positive ILD for most frequencies, which means that the apparent sound location is shifted towards the left side of the head. For the MWF-Front, an ILD distortion is still observed, although the effect is less than for the bilateral case. It is still significant however, as by [81], a difference as small as 0.5 dB can be noticeable. The MWF-All does not introduce a noticeable ILD distortion, which is in agreement with chapter 7.

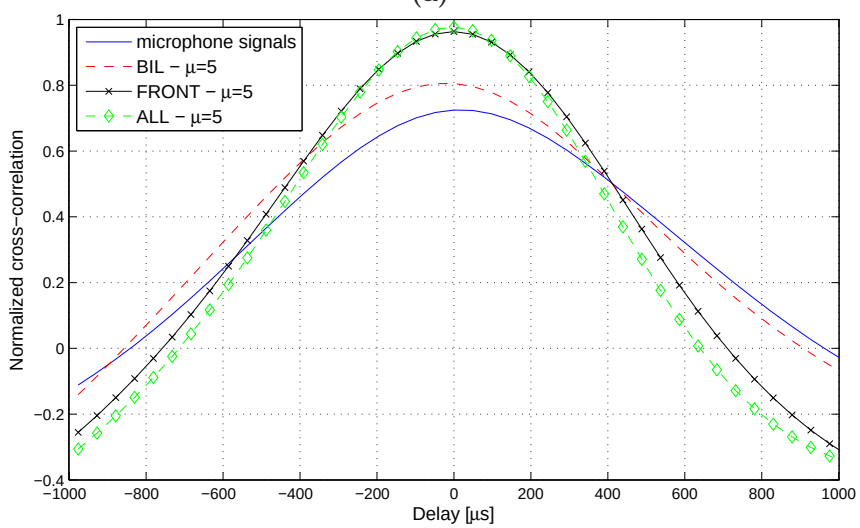
In Figure 8.3 (b) the normalized cross-correlation function, calculated between the left and right output signals, is shown. As in [176], the signals were filtered by a low-pass filter with a cut-off frequency of 1000 Hz, as only the lower frequencies are important for the ITD [14]. The delay for which the maximum of the cross-correlation function occurs gives an estimate of the ITD. It is thus seen that the different reduced-bandwidth algorithms obtain a correct ITD, which is in agreement with the theoretical analysis.

In Figure 8.4, the ILD per band and normalized cross-correlation functions for the S90 N270 scenario are shown. Again, the reduced-bandwidth techniques introduce ILD distortions, while the ITD is preserved. It is however also observed that even the MWF-All does not obtain the same ILD values as the unprocessed microphone signals, at the higher frequencies. This effect can be explained by the fact that, in addition to noise reduction, the MWF also achieves a certain amount of dereverberation. This is illustrated by the fact that the normalized cross-correlation function of the MWF-All obtains a higher maximum value (more coherence), and by the fact that the standard deviations in the ILD estimation are smaller, compared to the unprocessed microphone signals. As reverberation decreases the magnitude of the ILDs [156], the MWF-All obtains ILDs with a higher magnitude due to its inherent dereverberation effect.

To further corroborate, the experiment was also repeated for a low reverberation time ($\mathrm{RT}_{60} = 0.21s$, [176, 177]), cfr. Figure 8.5. As now the unprocessed microphone signals contain less reverberation, the dereverberation effect has less impact. As a result, it can be observed that the ILD values indeed coincide better. The bilateral MWF however still obtains ILD values with unnaturally large magnitudes.

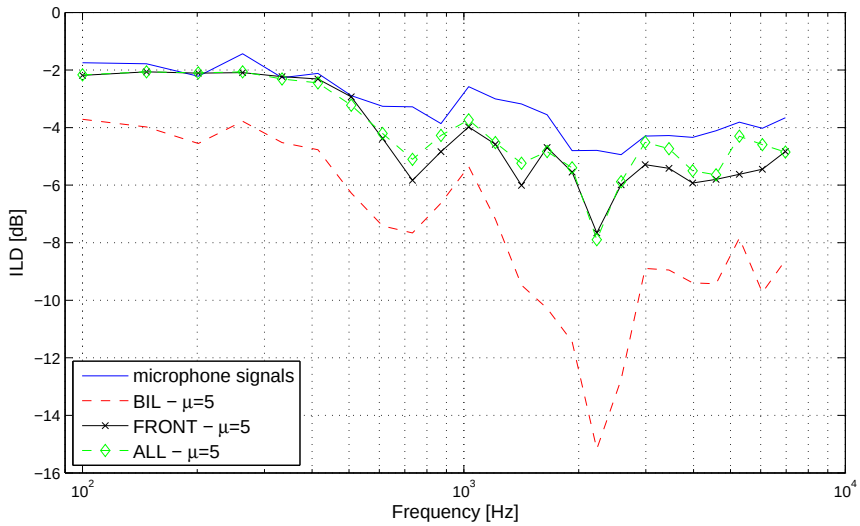


(a)

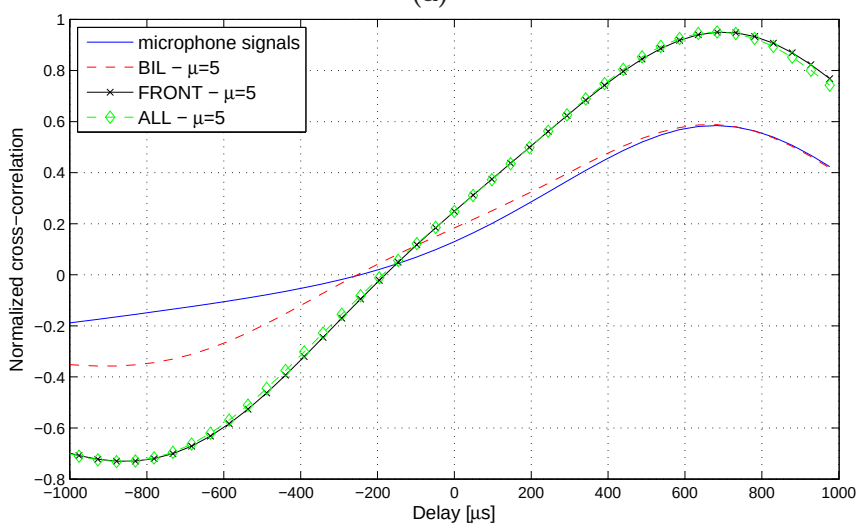


(b)

Figure 8.3: **S0 N60**; ILD per critical band (a) and normalized cross-correlation (b) of the speech component. The average ILD standard deviations are 0.71 dB (microphone signals), 0.73 dB (BIL), 0.41 dB (FRONT) and 0.28 dB (ALL).

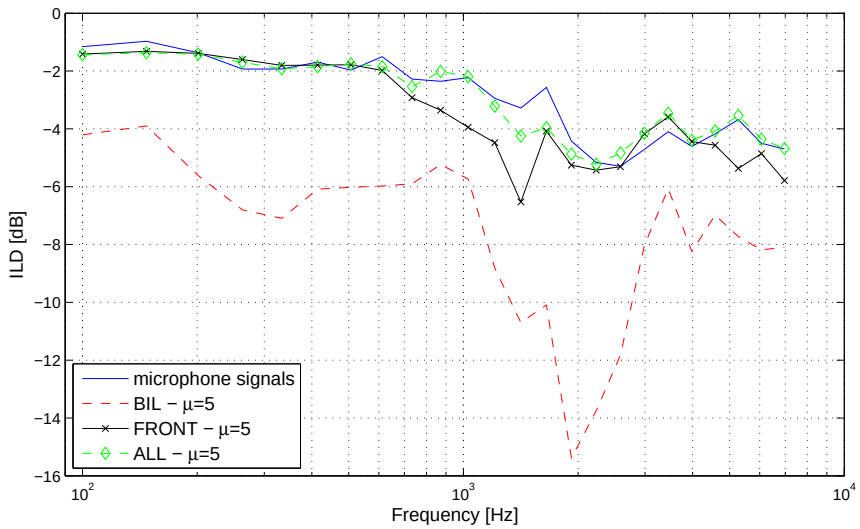


(a)

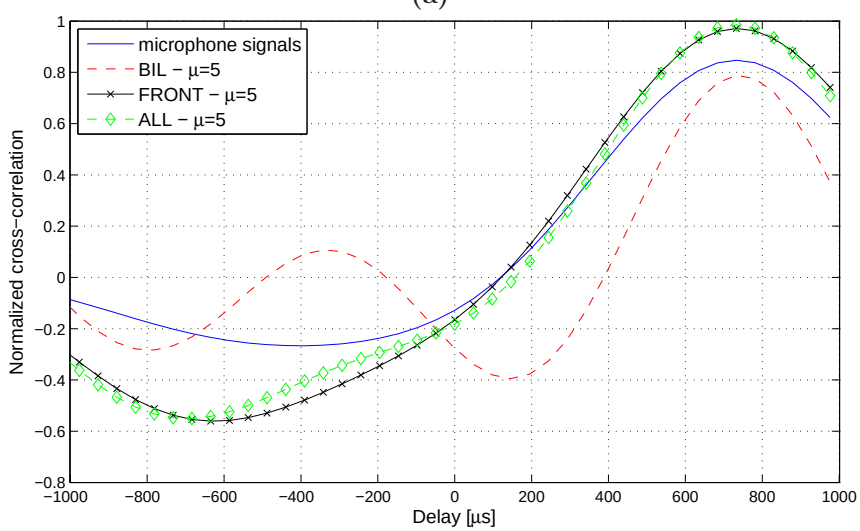


(b)

Figure 8.4: **S90 N270**; ILD per critical band (a) and normalized cross-correlation (b) of the speech component. The average ILD standard deviations are 1.22 dB (microphone signals), 1.27 dB (BIL), 0.43 dB (FRONT) and 0.42 dB (ALL).



(a)



(b)

Figure 8.5: **S90 N270**; Low reverberation; ILD per critical band (a) and normalized cross-correlation (b) of the speech component. The average ILD standard deviations are 0.61 dB (microphone signals), 0.84 dB (BIL), 0.28 dB (FRONT) and 0.21 dB (ALL).

8.4.6 Average speech ILD error and broadband SNR improvement

In this section, the performances of all the considered algorithms of Table 8.1 are compared, both for a batch as for an adaptive implementation. In Figures 8.6-8.9, the average mean ILD errors (in absolute value) and standard deviations (indicated by errorbars) are shown for the different speech-in-noise scenarios, together with the left and right broadband SNR improvements (compared to the unprocessed front microphone signals). As the ILD cues are dominant at higher frequencies [14], only the critical bands with center frequencies over 1000 Hz are included in the average.

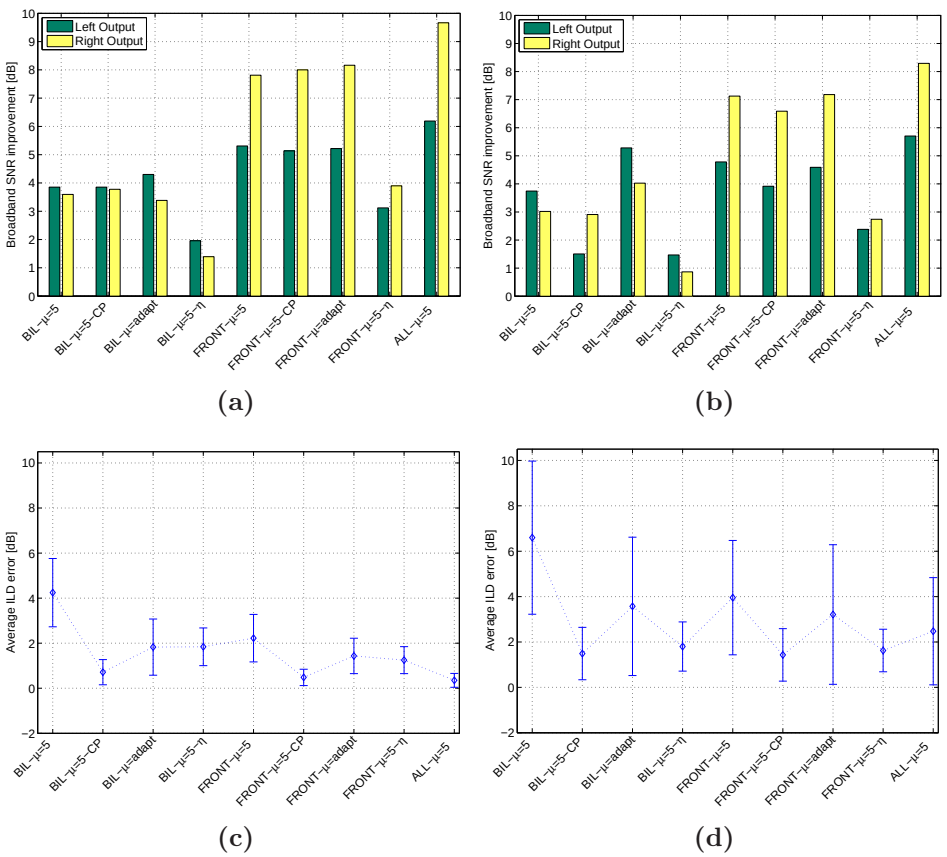


Figure 8.6: **S0 N60**; Broadband SNR improvement for batch (a) and adaptive (b) MWF; average speech ILD error for batch (c) and adaptive (d) MWF

In Figure 8.6, the performances of the batch and adaptive implementations of the nine considered algorithms are shown for the S0N60 scenario. As in [177], adding contralateral microphone signal(s) to the noise reduction procedure (FRONT and ALL) leads to significant SNR improvements over the bilateral case, both for the batch as for the adaptive implementation. For the batch implementation (Figure 8.6 (a),(c)), it can be seen that by adapting μ or using a partial noise estimate (MWF-η), the average ILD error is decreased compared to the BIL-μ=5 or FRONT-μ=5 algorithms, but not completely eliminated. The CP algorithms are best at preserving the ILD cues, while they are not outperformed by BIL-$\mu = 5$ or FRONT-$\mu = 5$ in terms of SNR improvement. If the batch results are compared with the adaptive results (Figure 8.6 (b),(d)), it can be observed that

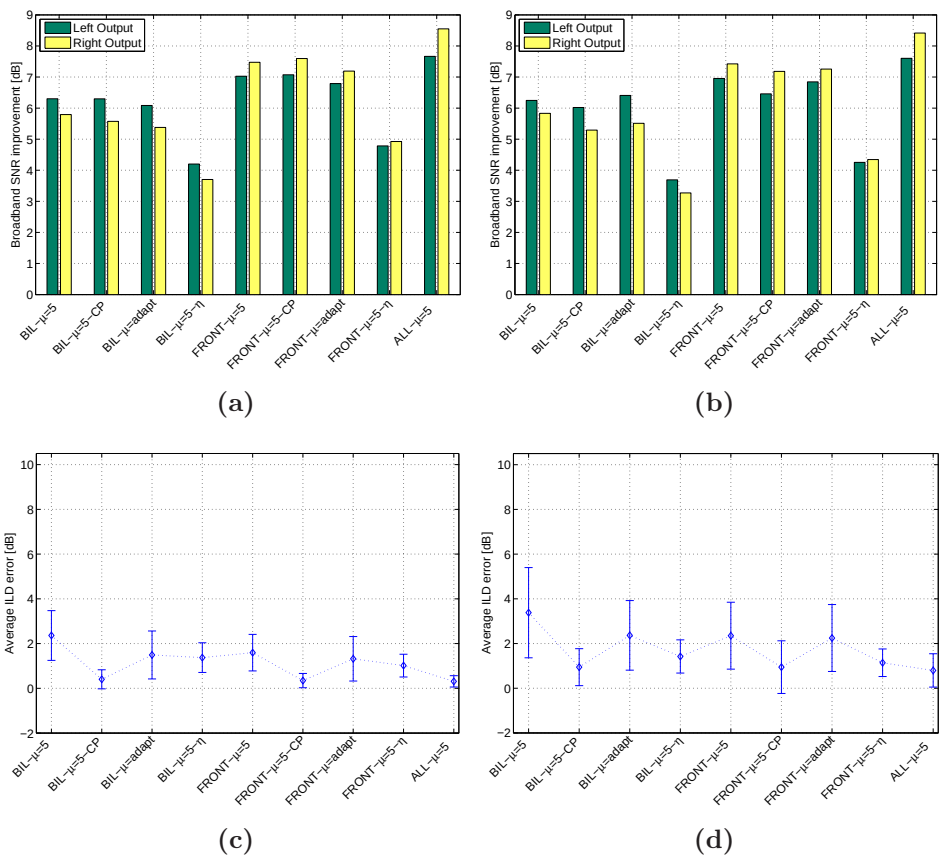


Figure 8.7: **S0 N[60/120/180/210]**; Broadband SNR improvement for batch (a) and adaptive (b) MWF; average speech ILD error for batch (c) and adaptive (d) MWF

the average ILD error is higher for the adaptive implementations, and that there are more fluctuations in the results as the standard deviations are larger. The CP algorithms still preserve the ILD cues adequately, but show a degraded SNR improvement compared to the BIL-μ=5 or FRONT-μ=5 algorithms (although in the case of the MWF-Front, the difference is within 1 dB).

In Figure 8.7, the performances for a scenario with four babble noise interferers (at 60, 120, 180 and 210 degrees) is shown. Overall, the same observations as for the S0N60 scenario can be made. As four babble noises are active at the same time, the overall noise level is more stationary than for a single noise interferer. As a result, there is not much performance difference between the batch and adaptive implementations, in contrast to the S0N60 scenario.

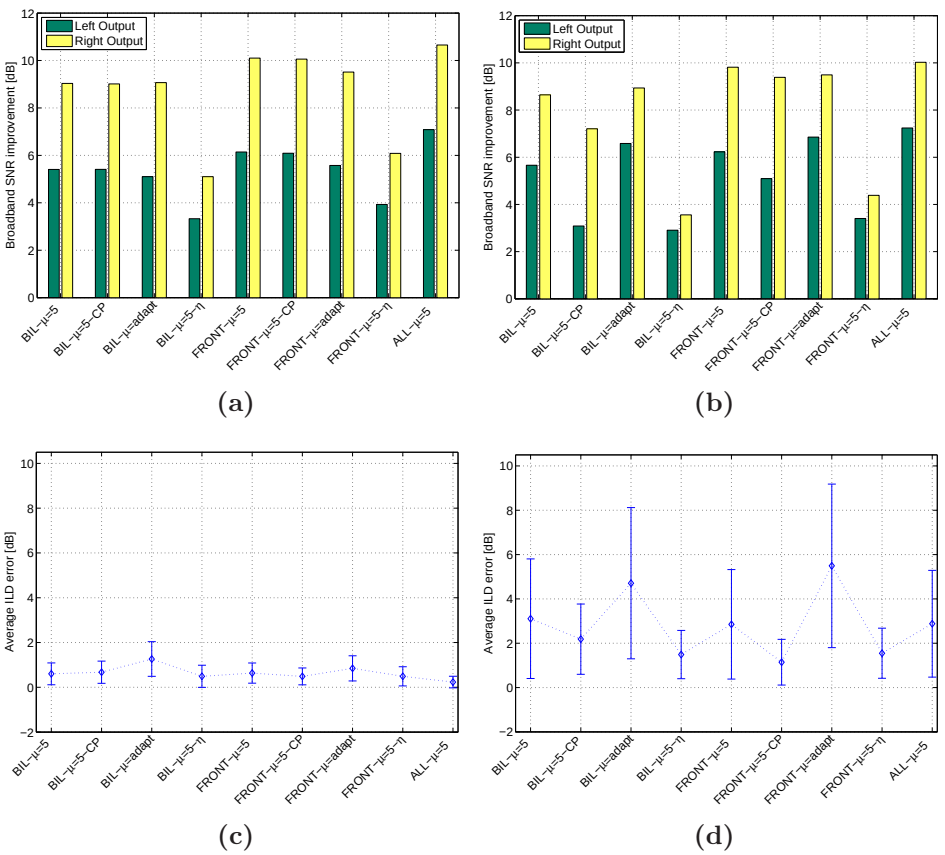


Figure 8.8: **S0 N120**; Broadband SNR improvement for batch (a) and adaptive (b) MWF; average speech ILD error for batch (c) and adaptive (d) MWF

In Figure 8.8, the performances for the S0N120 scenario are shown. As the speech and noise source are now farther apart, a large SNR improvement is found, even for the bilateral MWF. As in [33], none of the batch algorithms introduce a significant speech ILD error. It was therefore concluded in [33] that speech ILD errors only occur for scenarios, where the obtained output SNRs at the left and right devices are very different. These are in fact the scenarios where a bilateral (close-spaced) microphone array is insufficient to reduce the noise, namely scenarios where (some of) the speech and noise sources are too close together, or alternatively, scenarios where speech and noise are positioned in two different hemispheres (such as S90N270) [177]. For the adaptive implementations (Figure 8.8 (d)), it is seen that ILD errors can still occur, especially for the BIL$-\mu$=adapt and FRONT$-\mu$=adapt algorithms. Again the speech ILD error is lowest for the CP algorithms.

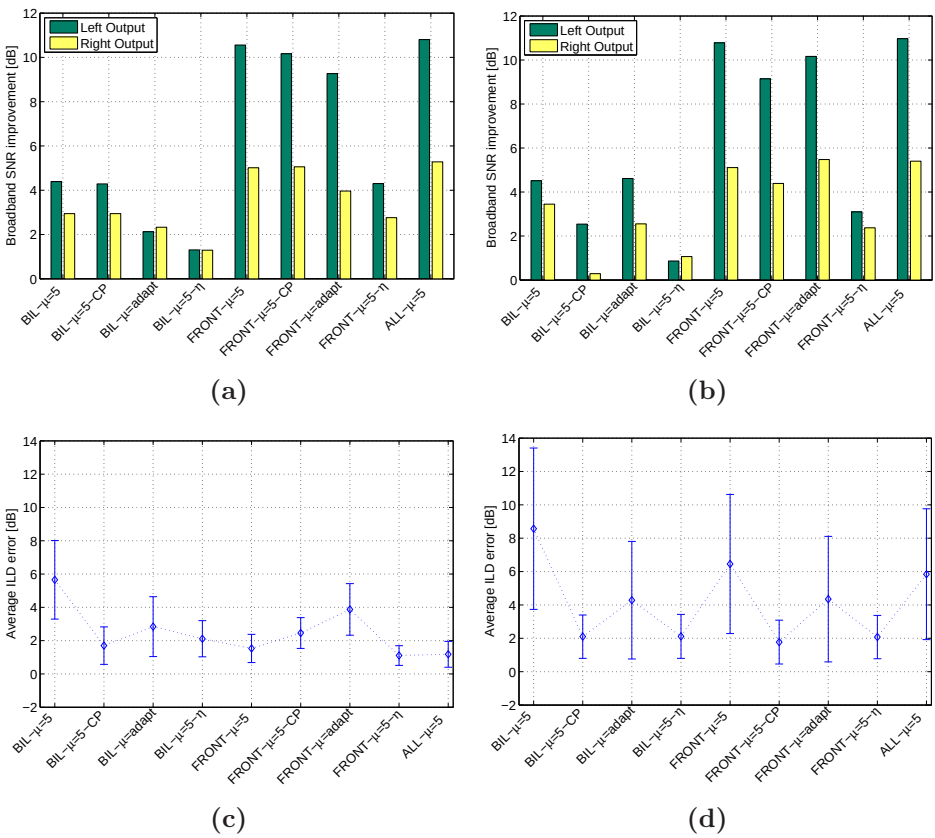


Figure 8.9: **S90 N270**; Broadband SNR improvement for batch (a) and adaptive (b) MWF; average speech ILD error for batch (c) and adaptive (d) MWF

Finally, the performances of the S90N270 scenario are shown in Figure 8.9. Although the precalibrated steering vector (which is used if estimation errors occur) assumes a frontal speech source, the performance does not completely degrade for the adaptive implementations with CP. Again, especially the FRONT$-\mu$=5$-$CP algorithm achieves a good SNR improvement, while at the same time only a low speech ILD error is introduced.

8.5 Conclusion

In this chapter, it has been proven theoretically and shown through experiments that reduced-bandwidth SDW-MWF techniques preserve the speech ITD cues, but distort the speech ILD cues. By adapting μ, or including a partial noise estimation parameter η, the speech ILD distortion can be reduced for some scenarios, but it can never be completely eliminated. Two novel reduced-bandwidth techniques with common postfilter, which have different bandwidth requirements, have also been proposed. The experiments indicate that significant speech intelligibility improvements are obtained with the common postfilter techniques, while the speech ILD cues are preserved adequately, even in a realistic adaptive implementation.

Part IV

Binaural VAD algorithms

Chapter 9

Binaural VAD based on decision fusion

As is the case for other single- and multi-microphone noise reduction techniques, the SDW-MWF relies on a voice activity detection (VAD) algorithm, which classifies frames as noise-only or speech+noise frames. In this chapter, two novel binaural fusion VADs are proposed, which are extensions of a fusion VAD originally proposed for a wireless sensor network application. By making use of the exchange of information over the wireless link of the binaural hearing aid, the binaural VADs perform a decision fusion of energy VADs calculated in the left and right device and, if the available bandwidth also allows for transmitting audio signals, a cross-correlation based VAD. The superior performances of the proposed binaural VADs are assessed (in terms of receiver operating characteristics and also by evaluating the impact on the SDW-MWF performance) by experiments with a binaural hearing aid setup.

Section 9.1 presents the introduction to the chapter.

Section 9.2 reviews the single- and multi-microphone VAD algorithms which are utilized in this chapter.

Section 9.3 reviews the fusion VAD algorithm which was previously proposed for a wireless sensor network application. Two novel binaural fusion VADs are proposed.

Section 9.4 evaluates the different VAD algorithms by simulations. Both the VAD performance (in terms of receiver operating characteristics) and the impact on the noise reduction performance of the SDW-MWF is evaluated.

Section 9.5 concludes the chapter.

The material presented in this chapter was published in the proceedings of EUSIPCO [35].

9.1 Introduction

The SDW-MWF relies on a voice activity detection (VAD) algorithm, which classifies frames as either speech+noise or noise-only frames. In fact, many other single- and multi-microphone noise reduction techniques require a VAD to detect speech pauses, as a spectral or spatial estimate of the noise component is needed. An algorithm such as the GSC also requires a VAD to control the filter adaptation, as adaptation during speech activity significantly degrades the performance. A robust VAD algorithm is therefore crucial and highly sought for.

In this chapter, we focus on a VAD algorithm for binaural hearing aids. The wireless link is utilized to derive a decision fusion of different VADs, in order to obtain a superior fusion VAD. First, the decisions of single-microphone energy VADs, calculated on a reference microphone signal in the left and right device, are fused using only a small-bandwidth exchange of SNR estimates as in [9]. Second, if the available bandwidth also allows for transmitting an audio signal, a two-microphone cross-correlation VAD can also be applied. Two novel fusion VADs, which fuse the energy and cross-correlation VADs, are proposed and are shown to further increase the VAD performance. In contrast to statistical model-based VAD algorithms such as [141], the proposed fusion VADs still only require a (relatively simple) SNR estimation in each device.

9.2 Overview VAD algorithms

In this section, a single-microphone energy based VAD and a two-microphone cross-correlation based VADs are reviewed. These VAD algorithms serve as *local* VAD decisions, which are subsequently fused by the proposed decision fusion algorithms to obtain a superior *fusion* VAD decision. It should however be noted that the fusion algorithms are general in the sense that they could also be applied to other single- or multi-microphone VADs instead of the VADs of this section.

9.2.1 Single-microphone energy VAD

There are many single-microphone VAD algorithms available, which are for example based on short-term energy, zero-crossing rate, speech and noise

probability distributions, or combinations of properties (cfr. [48] for a summary of some algorithms). In this chapter, the log-energy based VAD of [180] will be used (and denoted by *energy VAD*), as it was shown in [48] that it achieved the best performance (over different scenarios, input SNRs and performance measures) compared to the other single-microphone VADs. The energy VAD tracks the short-term log-energy of a reference microphone signal, on a frame-by-frame basis. A histogram of the log-energy shows two clusters (corresponding to noise-only frames and speech+noise frames), which can be fitted by a bimodal Gaussian distribution (cfr. Figure 9.1). The optimal threshold is then found at the intersection of the two Gausssians. An online approximation of this histogram-based approach is also proposed in [180] and used in the simulations in this chapter. In the online method, the noise variance is tracked adaptively. When the short-term frame energy is significantly higher than the noise variance estimate, the frame is classified as speech+noise. As a result, the energy VAD does not perform well at low input SNRs or for nonstationary noise. On the other hand, no prior information about the target speech location is required.

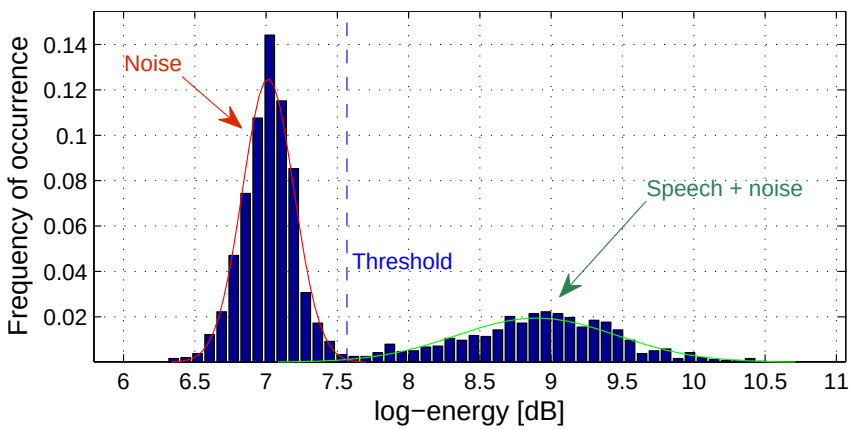


Figure 9.1: Log-energy histogram with bimodal Gaussian fit, for four VU sentences in multitalker babble noise.

9.2.2 Two-microphone cross-correlation based VAD

The considered two-microphone VAD (denoted as *cross-corr VAD*) is based on a per-frame time difference of arrival (TDOA) estimation. The instantaneous TDOA at frame-index k can be estimated by finding the delay corresponding to the maximum of the cross-correlation function [88]:

$$\widehat{\tau}_{12}[k] = \underset{\tau}{\operatorname{argmax}} \; r_{12}[\tau, k] \;, \tag{9.1}$$

where $r_{12}[\tau, k]$ is the instantaneous cross-correlation function between microphones 1 and 2, at frame-index k. $r_{12}[\tau, k]$ is calculated as:

$$r_{12}[\tau, k] = \int_{-\infty}^{\infty} Y_1[\omega, k]\ Y_2^H[\omega, k]\ \exp(j\omega\tau)\ \mathrm{d}\omega\ , \tag{9.2}$$

where $Y_1[\omega, k]$ and $Y_2[\omega, k]$ are the instantaneous (smoothed) spectra of microphone 1 and 2 at frame-index k. Speech is then detected if the cross-correlation value corresponding to the assumed speech TDOA is sufficiently close to the cross-correlation value corresponding to the estimated TDOA:

if $r_{12}[\tau_{\text{assumed}}, k] > T_{\text{corr}} * r_{12}[\widehat{\tau}_{12}, k]$ **then**
  $\text{VAD}_{\text{corr}}\ [k] \leftarrow 1$
else
  $\text{VAD}_{\text{corr}}\ [k] \leftarrow 0$
end if (9.3)

Simulations (cfr. section 9.4.2) indicate that a threshold value $T_{\text{corr}} = 0.7$ is a good setting and so this value will be used from now on (unless otherwise specified).

Correlation-based techniques are not suitable for use in a single hearing aid device (with close-spaced microphones), as they assume the intermicrophone distance is sufficiently large (> 7 cm) [109], or only work if the interferers are located in the rear half-plane [170]. However, if the hearing aids are connected by a wireless link, each device can stream a microphone signal to the contralateral device, so that each device can calculate (9.1) and (9.2) using one of its own microphone signals and the received microphone signal of the contralateral device. A drawback of the cross-correlation VAD is the fact that the speech source location (τ_{assumed}) has to be known *a priori*. We will always choose $\tau_{\text{assumed}} = 0$ in this chapter, i.e. the target speaker is located in the facing direction of the hearing aid user (a frequently occurring case).

9.3 Binaural fusion VADs

9.3.1 Decision fusion of energy VADs: Fusion-1

The energy VAD can be applied to the reference microphone signal of each device. If a small bandwidth wireless link is available so that it is feasible to transmit parameters, the energy VAD decisions and estimated (left and right) local SNRs

can be exchanged between the devices. As the energy VAD is more robust for higher SNRs, the estimated local SNR is an indication for the reliability of the energy VAD decision, and this information can be used to obtain a superior fusion VAD. The SNR estimates are calculated as in [9]:

if $\mathrm{VAD}_{\mathrm{energy}}[k] == 1$ (speech + noise) **then**

$S[k] \leftarrow \alpha * S[k-1] + (1-\alpha) * E[k]$
$N[k] \leftarrow N[k-1]$

else

$S[k] \leftarrow S[k-1]$
$N[k] \leftarrow \beta * N[k-1] + (1-\beta) * E[k]$

end if

$\hat{\mathrm{SNR}}[k] \leftarrow \frac{S[k]}{N[k]}$ (9.4)

$S[k]$ is the estimated local signal level (speech+noise), $N[k]$ is the estimated local noise level and $E[k]$ is the frame-based energy, i.e.

$$E[k] = \frac{1}{L}\sum_{l=0}^{L-1}(y_{\mathrm{ref}}[kL+l])^2 \,, \tag{9.5}$$

where L is the frame size and y_{ref} is the (time-domain) reference microphone signal. The forgetting factors α and β will both be set to 0.9.

In the fusion algorithm of [9] (denoted here as *fusion-1 VAD*), the energy VAD decisions and local SNR estimates of different sensor nodes in a wireless network are transmitted to a fusion center, where a weighted sum of the local VAD decisions is calculated and compared to a threshold:

$$w_p[k] = \frac{\hat{\mathrm{SNR}}_p[k]}{\sum_{p=1}^{P}\hat{\mathrm{SNR}}_p[k]} \,, \tag{9.6}$$

if $\sum_{p=1}^{P}\left(w_p[k] * \mathrm{VAD}_{\mathrm{energy},p}\,[k]\right) > T_{\mathrm{fusion}-1}$ **then**

$\mathrm{VAD}_{\mathrm{fusion}-1}\,[k] \leftarrow 1$

else

$\mathrm{VAD}_{\mathrm{fusion}-1}\,[k] \leftarrow 0$

end if (9.7)

where P is the number of nodes. The binaural hearing aid application corresponds to a simple two-node ($P = 2$) network. If the local SNR estimates and energy VAD

decisions are exchanged between the devices, (9.6) and (9.7) can be calculated in each device. Because of the headshadow effect, there is often a best-ear side (i.e. high SNR) and a worst-ear side (low SNR). If the received SNR estimate is larger than the local SNR estimate, the other device is probably at the best-ear side. An alternative fusion strategy is therefore to use the VAD decision of the other device instead of the own VAD decision, if the received SNR estimate is larger than the own SNR estimate. This fusion strategy actually corresponds to a threshold $T_{\text{fusion}-1} = 0.5$ in (9.7), and will be used from now on (unless otherwise specified).

9.3.2 Decision fusion of energy and cross-correlation VADs: Fusion-2

If microphone signals can be streamed over the wireless link, both devices can calculate the energy VAD and SNR estimates as in the previous section (as now, both devices have access to signals from both sides of the head). In addition, the two-microphone cross-correlation VAD can also be calulated in both devices. We now propose to fuse the two energy VAD decisions with the cross-correlation based VAD, in a similar manner as in the previous approach. As the cross-correlation VAD is not linked to a local SNR estimate, $\hat{\text{SNR}}_{\text{corr}}[k]$ is set equal to the average of the other SNRs:

$$\hat{\text{SNR}}_{\text{corr}}[k] = \frac{\hat{\text{SNR}}_L[k] + \hat{\text{SNR}}_R[k]}{2} \,, \tag{9.8}$$

where $\hat{\text{SNR}}_L[k]$ and $\hat{\text{SNR}}_R[k]$ are the estimated local SNRs at the left and right device. As a consequence, the cross-correlation VAD gets an intermediate weight in the decision fusion. The decision fusion rule (denoted as *fusion-2*) is then equal to:

$$w_1[k] = \frac{\hat{\text{SNR}}_L[k]}{\hat{\text{SNR}}_L[k] + \hat{\text{SNR}}_R[k] + \hat{\text{SNR}}_{\text{corr}}[k]} \,, \tag{9.9}$$

$$w_2[k] = \frac{\hat{\text{SNR}}_R[k]}{\hat{\text{SNR}}_L[k] + \hat{\text{SNR}}_R[k] + \hat{\text{SNR}}_{\text{corr}}[k]} \,, \tag{9.10}$$

$$w_3[k] = \frac{\hat{\text{SNR}}_{\text{corr}}[k]}{\hat{\text{SNR}}_L[k] + \hat{\text{SNR}}_R[k] + \hat{\text{SNR}}_{\text{corr}}[k]} \,, \tag{9.11}$$

if $w_1[k] * \text{VAD}_{\text{energy},L}\ [k] + w_2[k] * \text{VAD}_{\text{energy},R}\ [k]$
$+w_3[k] * \text{VAD}_{\text{corr}}\ [k] > T_{\text{fusion}-2}$ **then**

$\text{VAD}_{\text{fusion}-2}\ [k] \leftarrow 1$

else

$\text{VAD}_{\text{fusion}-2}\ [k] \leftarrow 0$

end if (9.12)

A threshold value $T_{\text{fusion}-2} = 0.45$ was found to be a good setting, as will be illustrated in section 9.4.2.

9.3.3 Decision fusion of energy and cross-correlation VADs based on optimal AND rule: Fusion-opt

An alternative for fusing the energy and cross-correlation VAD is to combine the VAD decisions by an AND rule. In [172], it is shown that the AND rule can be applied in an optimal way, by making use of the Receiver Operating Characteristics (ROC) graphs [63] of the separate VAD algorithms. We propose here to apply the AND rule to the output of the fusion-1 VAD (fusion of energy VADs) together with the output of the cross-correlation VAD. To generate the corresponding ROC graphs, the thresholds $T_{\text{fusion}-1}$ and T_{corr} are varied from zero to one in steps of 0.05.

In Figure 9.2, the average ROC of the fusion-1 and cross-correlation VAD are shown. In the ROC graph, the false positive rate (number of noise-only frames erroneously classified as speech+noise, divided by the total number of noise-only frames) is plotted on the X axis and the true positive rate (number of speech+noise frames correctly classified as speech+noise, divided by the total number of speech+noise frames) is plotted on the Y axis. The performance of the VADs was averaged over a range of input SNRs and over several spatial scenarios. Both scenarios with speech in the frontal facing direction as in other directions were included. More information about the setup and test stimuli is given in section 9.4.

By measuring the performance obtained with the AND rule for every possible combination of $T_{\text{fusion}-1}$ and T_{corr}, where both parameters are varied from zero to one in steps of 0.05, the points marked as *candidates for AND-fusion* are found. By definition, the optimal AND rule ROC is then the convex hull of these candidate points. It is indeed observed that the area under the convex hull ROC is larger than the areas of the fusion-1 and cross-correlation VADs, so that in principle, a better performance is obtained. From now on, the point closest to the top-left corner

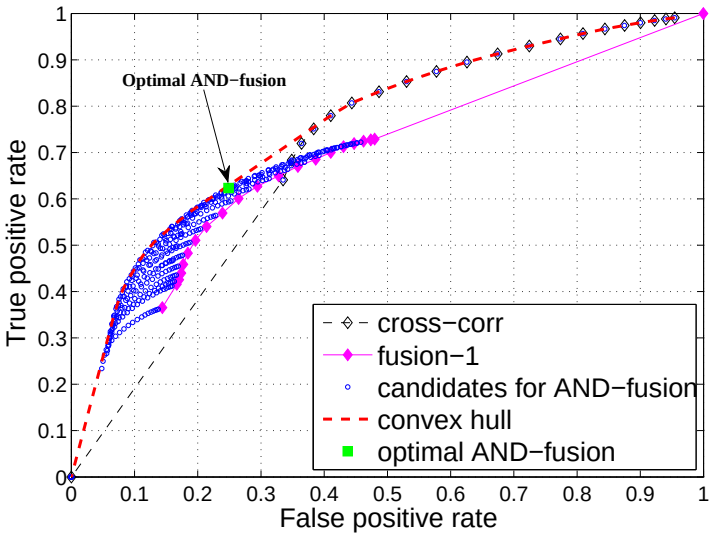


Figure 9.2: Average ROC graphs for cross-correlation VAD and Fusion-1 VAD, and optimal AND-fusion. The average performance over various spatial scenarios and input SNRs is shown.

is chosen as the optimal fusion VAD (denoted *fusion-opt*). The corresponding thresholds are $T_{\text{corr}} = 0.55$ and $T_{\text{fusion}-1} = 0.25$.

9.4 Experimental results

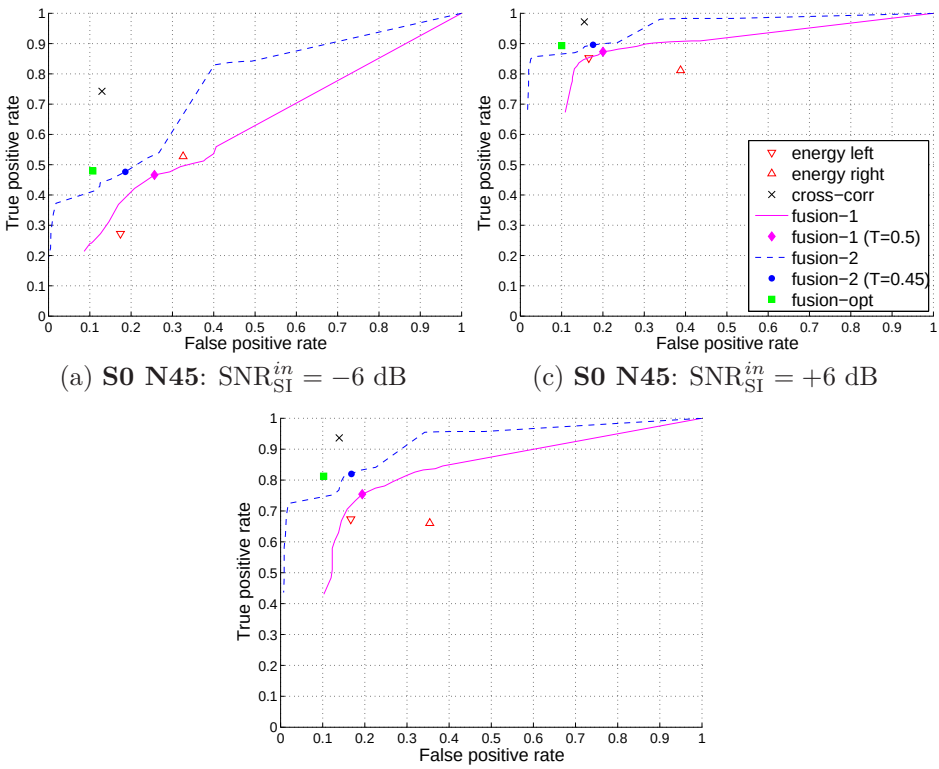
9.4.1 Setup and stimuli

A binaural hearing aid setup with two behind-the-ear hearing aids is considered. There are two omnidirectional microphones per device, and it is assumed that the link allows for transmitting one audio signal (i.e. the front microphone signal) to the other device (full-duplex).

Head-related transfer functions (HRTFs) were measured on the setup described in chapter 3 (SI-weighted reverberation time $\text{RT}_{60} = 0.62s$). The considered acoustic scenarios have a target speech (S) source and interfering noise (N) source(s) at specified azimuthal angles (with 0° in front of the head, 90° to the right of the head).

As speech stimulus, 4 consecutive sentences of the Dutch VU sentence material [184] were used. Multitalker babble noise was used as interfering noise signal(s). The signals are sampled at $f_s = 20480$ Hz and processed by a weighted overlap-add (WOLA) filterbank. The signals are segmented in frames of $L = 128$ samples with 75% overlap, and windowed by a Hann Window.

9.4.2 VAD performance: ROC graphs



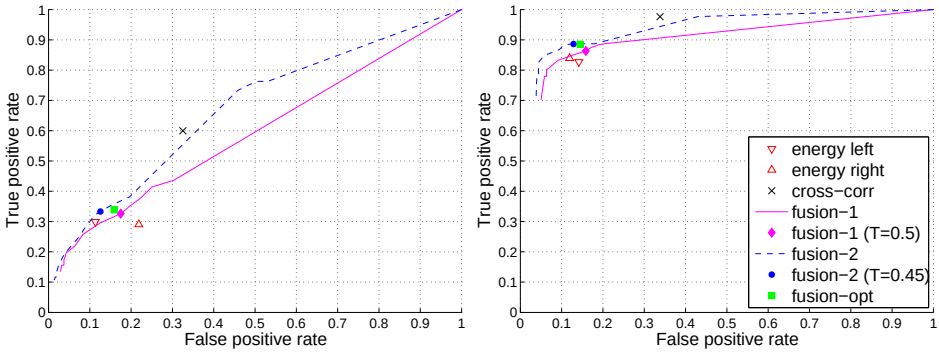
(a) **S0 N45**: $\text{SNR}_{\text{SI}}^{in} = -6$ dB

(c) **S0 N45**: $\text{SNR}_{\text{SI}}^{in} = +6$ dB

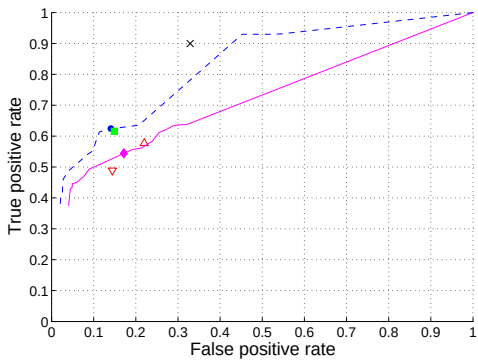
(b) **S0 N45**: $\text{SNR}_{\text{SI}}^{in} = 0$ dB

Figure 9.3: ROC graphs for fusion-1 and fusion-2 VADs, S0 N45.

In Figures 9.3-9.5, the ROC graphs of the VAD algorithms are plotted for three different spatial scenarios, and three different input SNRs per scenario. The input SNR is measured in absence of the CORTEX manikin, at the center point of the setup. Because of the head-shadow effect, the actual input SNRs at the left and right device are therefore dependent on the spatial scenario, cfr. Figure 9.6 of the next section. The ROC graphs were generated by varying the decision thresholds



(a) **S0 N[135/225]**: $\mathrm{SNR}_{\mathrm{SI}}^{in} = -6$ dB (c) **S0 N[135/225]**: $\mathrm{SNR}_{\mathrm{SI}}^{in} = +6$ dB

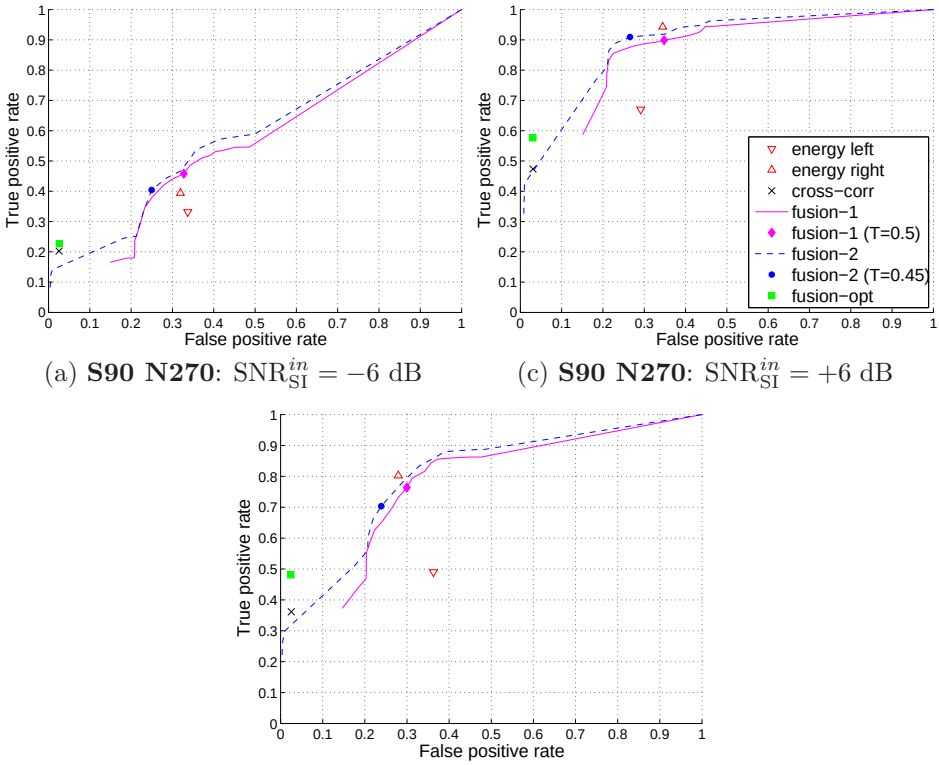


(b) **S0 N[135/225]**: $\mathrm{SNR}_{\mathrm{SI}}^{in} = 0$ dB

Figure 9.4: ROC graphs for fusion-1 and fusion-2 VADs, S0 N[135/225].

$T_{\mathrm{fusion}-1}$ and $T_{\mathrm{fusion}-2}$ from 0 to 1. The cross-correlation VAD threshold is set to $T_{\mathrm{corr}} = 0.7$ and the optimal AND-fusion VAD (fusion-opt) is fixed as in section 9.3.3. As the VADs perform better at higher input SNRs, the ROC graphs indeed shift towards the top-left corner (i. e. have a larger area below the ROC graph) as the input SNR increases.

Because of the head-shadow effect, it can be observed that the energy VAD performs better (i.e. more towards the top-left corner in the ROC space) in the left device in the S0 N45 scenario (left ear is the best ear), whereas it performs better in the right device in the S90 N270 scenario (right ear is the best ear). It can be seen that the fusion-1 VAD generally gives an improvement over the individual energy VADs. The ROC point corresponding to $T_{\mathrm{fusion}-1} = 0.5$ which is used in subsequent experiments, is also indicated on the curve. The cross-correlation VAD generally performs well if the speech source is located in the assumed direction (i. e. 0°), for S90 N270 a low true positive rate is however observed. The fusion-2



(a) **S90 N270**: $\mathrm{SNR}_{\mathrm{SI}}^{in} = -6$ dB

(c) **S90 N270**: $\mathrm{SNR}_{\mathrm{SI}}^{in} = +6$ dB

(b) **S90 N270**: $\mathrm{SNR}_{\mathrm{SI}}^{in} = 0$ dB

Figure 9.5: ROC graphs for fusion-1 and fusion-2 VADs, S90 N270.

VAD generally gives a better performance than the fusion-1 VAD, and even for S90 N270 the performance is not degraded as is the case for the cross-correlation VAD. The ROC point corresponding to $T_{\mathrm{fusion}-2} = 0.45$ which is used in subsequent experiments, is also indicated on the curve. The fusion-opt VAD of section 9.3.3 generally gives a performance similar to the fusion-2 VAD, but is however degraded in a similar way as the cross-correlation VAD for the S90 N270 scenario. It should also be noted that the three spatial scenarios of this section were not included as training scenarios to derive the optimal AND rule fusion.

9.4.3 SDW-MWF performance

In this section, the performance of the binaural SDW-MWF, using the different discussed VAD algorithms, is assessed by objective (physical) measures[1]. The results for six spatial scenarios, for an average input SNR of 0 dB (measured in the center point of the setup, in absence of the CORTEX manikin) are presented. Because of the headshadow effect, the actual input SNRs at the left and right devices depend on the spatial scenario, as is illustrated in Figure 9.6

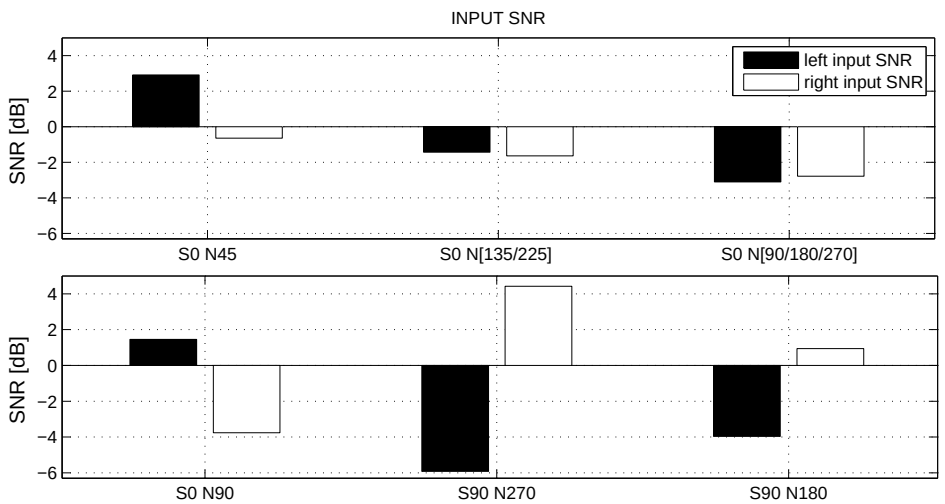
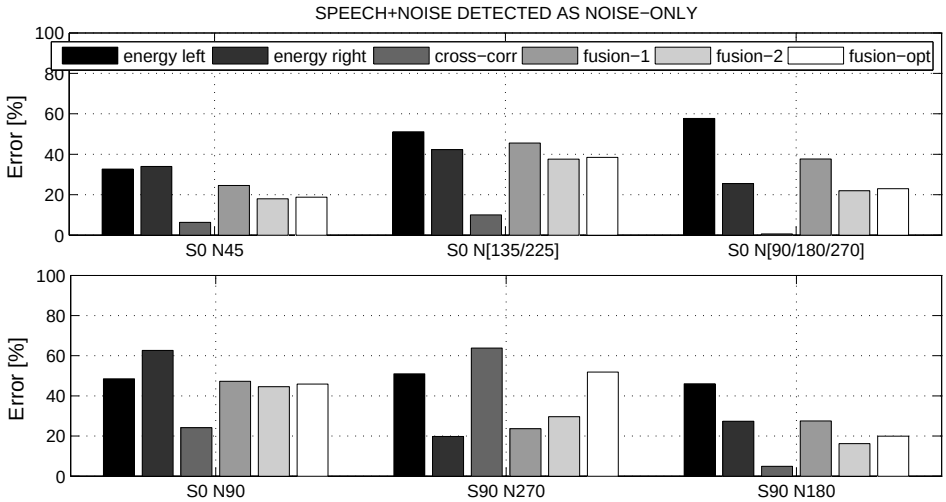


Figure 9.6: Average input SNRs at the left and right device (mounted on CORTEX MK2), for an average input SNR of 0 dB (measured in absence of the CORTEX manikin).

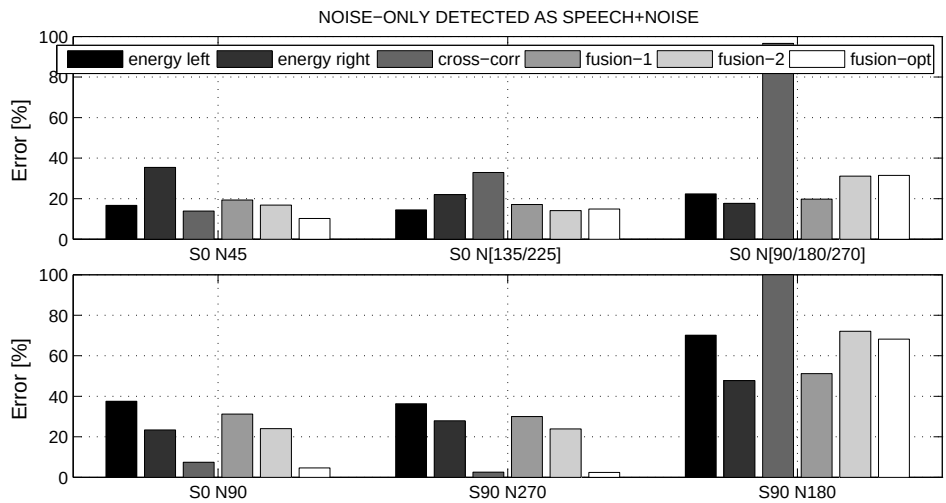
In Figure 9.7, the VAD errors introduced by the different VAD algorithms are shown for the six considered spatial scenarios. A distinction is made between the ratio of speech+noise frames erroneously detected as noise-only (equal to one minus the true positive rate), and the ratio of noise-only frames erroneously detected as speech+noise (equal to false positive rate). The considered VAD algorithms are the log-energy based VADs (calculated at the left or right device), the cross-correlation based VAD, and the three discussed fusion VADs (fusion-1, fusion-2 and fusion-opt).

The SDW-MWF performance (SI-weighted SNR improvement and distortion in dB) for five different VADs is shown in Figures 9.8 (left device) and 9.9 (right device). The performance with a perfect VAD is also shown as a reference point.

[1]For the fusion-2 VAD, the impact on subjective performances was previously presented in chapter 3.

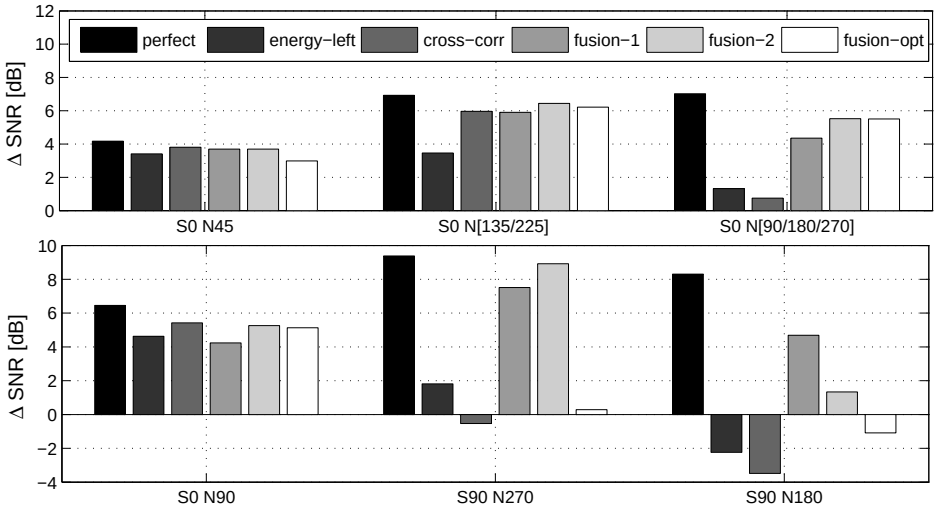


(a)

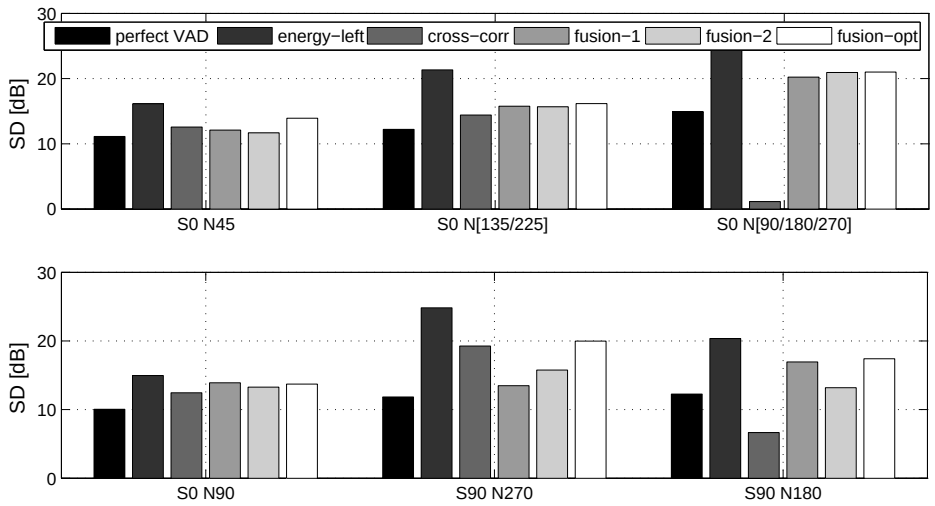


(b)

Figure 9.7: Percentage speech+noise frames erroneously detected as noise-only (=1− true positive rate) (a) and percentage noise-only frames erroneously detected as speech+noise (=false positive rate) (b), for an average input SNR of 0 dB (measured in absence of the CORTEX manikin).

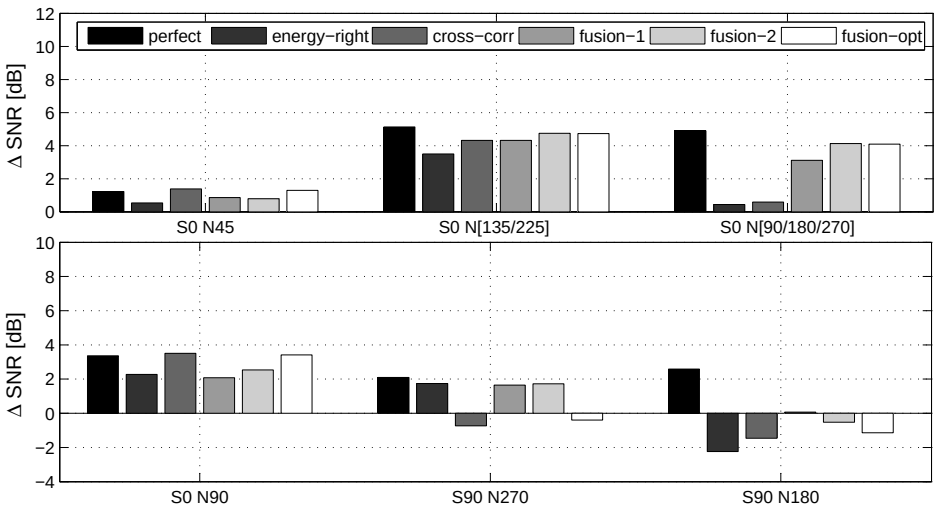


(a)

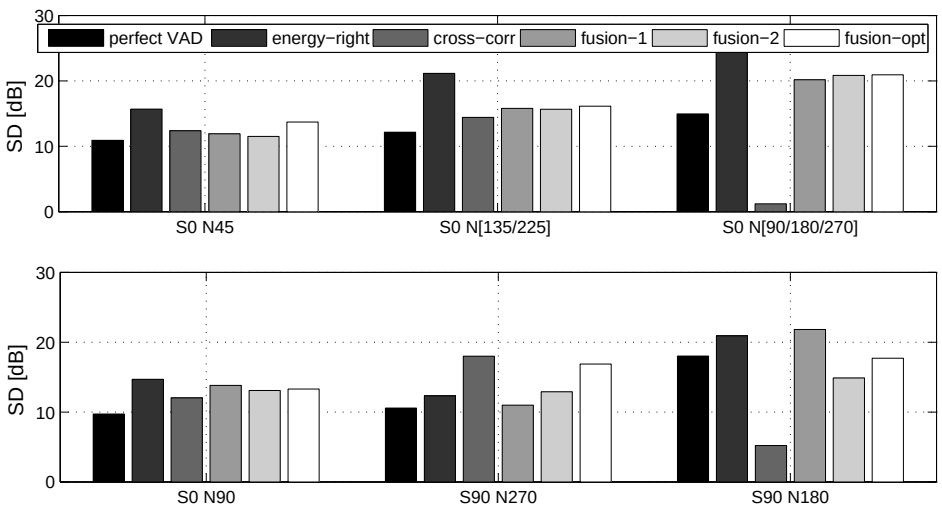


(b)

Figure 9.8: SI-weighted SNR improvement (a) and speech distortion (b) for left device, for an average input SNR of 0 dB (measured in absence of the CORTEX manikin).



(a)



(b)

Figure 9.9: SI-weighted SNR improvement (a) and speech distortion (b) for right device, for an average input SNR of 0 dB (measured in absence of the CORTEX manikin).

The reference adaptive SDW-MWF algorithm which was discussed in section 2.2.4, and implemented on a real-time platform in chapter 3, is considered. It is assumed that the front microphone signal can be exchanged over the wireless link so that each device has access to a total of 3 microphone signals ($M = 2$ ipsilateral signals and $N = 1$ contralateral signal), which corresponds to the MWF-Front technique. The speech distortion parameter was set to $\mu = 5$, the exponential forgetting factors were set to $\lambda_y = \lambda_v = 0.9995$.

In general, it can be observed that the three binaural fusion VADs achieve a superior perfomance over the individual energy and cross-correlation VADs. The cross-correlation VAD fails in scenarios where speech is not coming from the assumed frontal direction (S90 N270 and S90 N180) and/or when there is a noise source at 180° (S90 N180 and S0 N[90/180/270]) as for this location the same TDOA is obtained as for a source coming from 0°. The fusion-2 VAD stays robust in these scenarios, whereas the performance with the fusion-opt VAD also degrades. Overall, it can be seen that the fusion-2 VAD usually achieves the best performance (in terms of SNR improvement and speech distortion).

9.5 Conclusion

In this chapter, two binaural fusion VADs (fusion-2 and fusion-opt) have been proposed, which are extensions of a fusion VAD for wireless sensor networks (fusion-1). The binaural fusion VADs make use of the wireless link in a binaural hearing aid to combine energy VAD decisions calculated in the left and the right device with a cross-correlation based VAD decision, in order to obtain a VAD performance which is superior to the individual VAD algorithms. The performance was assessed by ROC graphs and by evaluating the impact on the noise reduction performance of the binaural SDW-MWF.

Part V

Conclusion

Chapter 10

Conclusions and further research

10.1 Conclusions and summary

In this thesis, we have designed and evaluated noise reduction techniques for binaural hearing aids. **Chapter 1** gave a general introduction and identified three research objetcives, i.e. improving the noise reduction performance, preserving the binaural cues and deriving a binaural VAD. **Chapter 2** presented a theoretical binaural noise reduction framework and objective performance measures. The binaural Speech-Distortion-Weighted Multi-channel Wiener Filter (SDW-MWF), which is adopted throughout the thesis, was also reviewed.

We now revisit the three main research objectives and formulate conclusions for each objective.

Noise reduction performance

A first major goal of this work was to improve the noise reduction performance in binaural hearing aids, as poor speech intelligibility in noise is one of the top complaints of hearing aid users. More in particular, by exchanging microphone signals between the devices, so that each device also has access to one or more microphone signals from the other device, the noise reduction performance (and speech intelligibility) should significantly improve.

In **Chapter 3** the practical feasibility of binaural noise reduction, based on the binaural SDW-MWF, was assessed by a perceptual evaluation. Adaptive SRT

tests were conducted in a total of four scenarios, including a realistic cafeteria scenario, on 20 normal hearing subjects and 8 hearing impaired subjects. In contrast to previous work, online adaptive algorithms for the bilateral and binaural SDW-MWF, together with a bilateral fixed beamformer, were worked out and implemented on a real-time platform. As the SDW-MWF relies on a VAD, a realistic VAD algorithm was also included. The main conclusion of the study was that the binaural SDW-MWF retains its benefit over the bilateral algorithms, even in these challenging and realistic conditions. For the hearing aid users, who obtained their SRT at higher input SNR levels, using a real VAD did not significantly degrade the performance compared to algorithms using a perfect VAD. In conclusion, the perceptual study demonstrates the feasibility of the SDW-MWF for binaural noise reduction, and forms a good motivation for further developments and optimizations described in this thesis.

Although the perceptual study showed that the SDW-MWF performed adequately at SNRs around the SRT level, improving the robustness against estimation errors in the second-order statistics is still very valuable. **Chapter 4** therefore presented a theoretical performance analysis of the standard SDW-MWF in the presence of estimation errors. Two novel expressions, namely the R1-MWF and SP-MWF, were also introduced and analyzed. The R1-MWF and SP-MWF implicitly assume a single target speech source (rank-one speech correlation matrix), but can also be calculated in a more general case. The theoretical analysis and simulations illustrated that for a single speech source, the R1-MWF and SP-MWF achieve a better noise reduction performance than the standard SDW-MWF implementation. Moreover, simulations also demonstrated that the R1-MWF and especially the SP-MWF retains a better performance if the number of target speech sources is increased. A second benefit of the novel expressions is that they have a conceptually appealing structure. Namely, they are explicitly decomposed into a spatial filter and single-channel spectral postfilter, where the SDW trade-off parameter μ only appears in the postfilter part. This structure can be exploited in a number of ways, as demonstrated in the following chapters. In **Chapter 5**, it was theoretically proven that the R1-MWF and SP-MWF remarkably are inherently robust to VAD errors, provided the noise is stationary. The main underlying reason for practical performance degradations is then that the noise is not stationary in real-life. A novel VAD-robust extension, which exploits the decomposed filter structure of the R1-MWF and SP-MWF, was subsequently proposed. It was demonstrated that the performance can be improved in certain nonstationary noise scenarios. In **Chapter 6**, the decomposed structure is exploited in order to derive a QR Decomposition Recursive Least Squares (QRD-RLS) algorithm for the R1-MWF and SP-MWF. It was demonstrated that this algorithm is numerically favorable, so that (for example) the wordlength can be reduced in fixed processing without affecting the performance.

Binaural cue preservation

A second major goal of this work was to design the noise reduction so that the binaural cues (ITDs and ILDs) of the target speech and residual noise are preserved, while simultaneously achieving sufficient noise reduction performance. The binaural cues allow to localize sound sources, which is also very important to hearing aid users (e.g. in traffic). The binaural cues, however, also improve speech intelligibility in noise by the so-called binaural unmasking effect.

In **Chapter 7**, full-bandwidth binaural SDW-MWF techniques were considered, i.e. where all microphone signals can be exchanged between the devices. The binaural cue preservation of the SDW-MWF and two previously proposed extensions, namely the MWF-η and MWF-ITF, was analyzed theoretically and validated with objective performance measures. The binaural SDW-MWF changes the binaural cues of the residual noise to equal the cues of the target speech. Both the speech and residual noise signals will thus be perceived in the direction of the speech source, so that the hearing aid user cannot rely on binaural unmasking to improve speech intelligibility. Both the MWF-η and MWF-ITF can offer improvements in preserving the binaural cues of the residual noise. The MWF-ITF includes extra constraints related to the interaural transfer functions of the speech and the noise. It was shown that, for a good parameter setting, correct localization of both speech and noise is indeed possible, but the binaural unmasking effect can not be exploited by the MWF-ITF. The MWF-η mixes a fraction of the original noise signal into the output signal, which thus necessarily decreases the obtained SNR improvement. In contrast to the MWF-ITF, the MWF-η can exploit the binaural unmasking effect, which even compensates for the SNR loss in some scenarios. Finally, the MWF-η does not involve a difficult parameter setting, in contrast to the MWF-ITF. In conclusion, the MWF-η thus seems to be the best alternative as binaural cue preservation technique.

In **Chapter 8**, reduced-bandwidth binaural SDW-MWF techniques were considered, i.e. where not every microphone signal can be exchanged between the devices. The impact on the binaural cue preservation was assessed, particularly the binaural cues of the target speech source. It was first proven theoretically that the reduced-bandwidth MWF techniques no longer preserve the speech binaural cues, in contrast to the full-bandwidth case. More in particular, the algorithms preserve the ITD cues but not the ILD cues. By adapting μ, or using the MWF-η, the speech ILD distortion can be reduced for some scenarios, but it can never be completely eliminated. Two novel reduced-bandwidth techniques with common postfilter (CP), which have different bandwidth requirements, were also proposed. These techniques again make use of the decomposed filter structure of the R1-MWF and SP-MWF expressions, as discussed above. Experiments in a reverberant environment indicated that significant speech intelligibility improvements can be obtained with the CP techniques, while the speech ILD cues are preserved

adequately, even in a realistic adaptive implementation.

Binaural VAD

A third goal of this work was to derive an improved VAD algorithm. As many noise reduction algorithms (including the SDW-MWF) rely on a VAD, a robust VAD algorithm is highly sought for. Our aim here was to derive a binaural VAD, which utilizes the wireless link in binaural hearing aids.

In **Chapter 9**, two binaural fusion VADs (fusion-2 and fusion-opt) were proposed, which are extensions of a fusion VAD for wireless sensor networks (fusion-1). The binaural fusion VADs make use of the wireless link to combine energy VAD decisions calculated in the left and the right device and a cross-correlation based VAD decision, in order to obtain a VAD performance superior to the individual VAD algorithms. The performance was assessed by Receiver Operating Characteristics and by evaluating the impact on the noise reduction performance of the binaural SDW-MWF. The fusion-2 VAD obtained the best overall performance and was therefore selected as "real VAD" reference in other parts of this thesis (chapters 3 and 5).

Summary

Table 10.1 summarizes the best performing noise reduction (NR) and VAD algorithms proposed in this thesis, for two different power consumption modes. The two defined modes are a *Low-power mode* where the devices only exchange parameters at a low data rate, and a *High-power mode* where the devices exchange audio signals at a high data rate. In the Low-power mode, the bilateral MWF with common postfilter (cfr. Chapter 8) is utilized, which only requires an exchange of the narrowband output SNRs. In the High-power mode, it is assumed that one microphone signal can be exchanged (in full-duplex), which enables the MWF-Front with common postfilter (cfr. Chapter 8). Ideally, the High-power mode should only be activated when a significant speech intelligibility improvement would be achieved compared to the Low-power mode (so that battery life is saved). This would be the case for speech-in-noise scenarios similar to those of Chapter 3, i.e. scenarios where the speech and noise sources are closely spaced or complex scenarios with multiple noise sources. This could thus be taken into account in the design of automatic classification algorithms.

The noise reduction techniques should be implemented using the QRD-RLS algorithm based on the SP-MWF (cfr. Chapter 6), both for the Low-power and for the High-power mode. Similarly, the partial noise estimation extension (MWF-η), which improves the noise cue preservation (cfr. Chapter 7), can be applied in both modes.

Finally, it was demonstrated (cfr. Chapter 9) that the exchange of microphone signals (High-power mode) enables a superior fusion (denoted as fusion-2 VAD) of energy-based VAD decisions and a cross-correlation based VAD decision. A fusion of energy-based VAD decisions only (denoted as fusion-1 VAD) is however also possible in the Low-power mode, as only broadband output SNR estimates and local VAD decisions have to be exchanged between the devices. It was demonstrated (cfr. Chapter 9) that the fusion-1 VAD then still leads to a significant performance improvement compared to using only the local VAD decisions.

	Low-power mode	**High-power mode**
NR technique	Bilateral MWF with CP	Binaural MWF-Front with CP
NR implementation	QRD-RLS based on SP-MWF	QRD-RLS based on SP-MWF
Noise cues	MWF-η extension	MWF-η extension
Binaural VAD	Fusion-1	Fusion-2

Table 10.1: Summary noise reduction (NR) and VAD algorithms

10.2 Suggestions for further research

In this section, suggestions for further research are presented. Some of these suggestions directly build on the binaural SDW-MWF work presented in this thesis. Other suggestions are more general in the sense that, although they are also concerned with SDW-MWF based noise reduction, they are not necessarily targeted towards a binaural hearing aid application.

Perceptual evaluation of rank-one binaural SDW-MWF algorithms

The presented R1-MWF and SP-MWF expressions (chapter 4) were shown to be theoretically more robust in the presence of estimation errors. In a reduced-bandwidth bilateral or binaural implementation, they can moreover be extended with a CP (chapter 8) so that the ILD distortions are reduced. These findings

however still have to be validated by a perceptual evaluation. In work by another research group [126], an implementation of the SP-MWF was evaluated on ten normal hearing subjects. Although physical SNR results were favorable for the SP-MWF, this could not be reproduced in the perceptual results. As a possible reason, it is reported that annoying spatial artefacts were perceived by some of the subjects. It should be noted that the CP extension was not implemented in [126], so that these observations may thus be in agreement with the results in chapter 8. A perceptual evaluation of the bilateral and binaural SP-MWF, including CP extension, is therefore planned in our own future work. In a first stage, an adaptive SRT test similar to the evaluation in chapter 3 will be conducted. A localization experiment, similar to the evaluation in [176], could however also be conducted in a second stage.

Effect of audio coding and transmission delay

In this thesis, we have taken into account that the available transmission bandwidth may not allow for the exchange of all microphone signals, so that reduced-bandwidth SDW-MWF algorithms should be considered. Some other non-ideal aspects of the wireless link were however not considered. For example, in practice the transmitted audio signals will be compressed to further reduce the required transmission bandwidth. As lossy compression may degrade the signal, this could have an impact on the achievable noise reduction performance. In some of the literature [51, 153, 169], performance bounds based on information theory were therefore derived for several binaural noise reduction schemes. An upper bound for the SNR performance is given as function of the available bitrate. These results can give interesting insights. However, it should be noted that an ideal codec is assumed so that these are only theoretical bounds. It is therefore still an open question which performance could be achieved using practical audio codecs. A second non-ideality, which was not considered, is that the wireless transmission will introduce a certain transmission delay in the signals. This delay may have a negative impact on the noise reduction performance. It should be investigated if these delays also introduce ILD distortions in the reduced-bandwidth SDW-MWF techniques with CP, proposed in chapter 8.

SDW-MWF based noise reduction for cochlear implants

Cochlear implants are provided for people with profound hearing losses. A cochlear implant is a surgically implanted device which restores hearing through direct electrical stimulation of the auditory nerve. Many hearing aid functionalities are also included in the speech processor of cochlear implants. For example, multi-microphone noise reduction is also applied in state of the art devices (e.g. [166]). It is not unthinkable that cochlear implants will follow the same trends

as hearing aids, so that binaurally linked cochlear implants may emerge in the future. The presented binaural SDW-MWF algorithms in this thesis are then also applicable in these binaural cochlear implants. As cochlear implants apply an additional postprocessing, i.e. to transform the audio signal into electrical stimuli, this could however be taken into account in the design of the SDW-MWF. In particular, in each stimulation cycle the cochlear implant stimulates only a few frequency bands or channels out of the total frequency range. For example, the Advanced Combination Encoder (ACE) strategy [182] selects the 8 channels (out of 22) with the highest energy in each stimulation cycle. One of the consequences is that for cochlear implant users (in contrast to hearing aid users), perceptual speech intelligibility improvements can be achieved with single-microphone noise reduction [42, 85]. This suggests that the typical distortions and artefacts introduced by single-microphone noise reduction are less of an issue, probably because they are no longer perceivable after the ACE processing. In summary, the noise reduction can thus be much more aggressive than what is allowable for hearing aids. If we consider the R1-MWF and SP-MWF techniques of this thesis (chapter 4), we see that the spectral postfilter is calculated based on long-term output SNR estimates (which are obtained from the speech and noise second-order statistics). For cochlear implant users, it could however be investigated whether a more aggressive spectral postfilter, based on short-term estimates of the output SNR, can lead to a better performance.

Perception-based multi-microphone noise reduction

A common trade-off in single-microphone and multi-microphone noise reduction techniques is the trade-off between speech distortion and noise reduction. In the SDW-MWF expressions considered in this thesis, this trade-off is directly controllable by the μ parameter, whereby higher values of μ allow for more noise reduction, at the cost of a higher introduced speech distortion. In fact, the presented R1-MWF and SP-MWF expressions (chapter 4) allow for an easier manipulation of this trade-off due to their explicit decomposed structure, as the μ parameter only occurs in a single-channel spectral postfilter. It should however be noted that in this context, the definitions of speech distortion and noise reduction are based on the MMSE criterion, which may not necessarily be relevant from a psycho-acoustic point-of-view. In single-microphone noise reduction, some approaches therefore incorporate the estimated psycho-acoustic masking threshold of the clean speech [86, 92, 185]. It could be investigated whether such approaches can also be extended to multi-microphone noise reduction, particularly to SDW-MWF based noise reduction. A possible approach to incorporate the masking threshold into the SDW-MWF, is to reformulate the SDW-MWF criterion in an optimization framework, similar to the perception-based clipping algorithm of [44].

Improved estimation and tracking of second-order statistics

As demonstrated in chapter 3, a realistic implementation of the SDW-MWF with real VAD does not show significant performance degradations around the SRT level, for a group of hearing aid users. Even so, a higher robustness against estimation errors, particularly those introduced by the VAD, is still very valuable. This thesis therefore presented filter expressions which are theoretically more robust to estimation errors in the speech second-order statistics (chapter 4). It was also demonstrated that the robustness to VAD errors could be improved in some scenarios (chapter 5). It should however be noted that there is still a remaining issue: either only the speech+noise correlation matrix is updated (in speech+noise frames), or else the noise correlation matrix is updated (in noise-only frames), while the other matrix is kept fixed. As a result, the tracking performance of the algorithm may be limited in highly nonstationary noise scenarios. A solution would be to somehow continuously update the speech+noise and noise correlation matrices, similar to single-microphone noise estimation techniques [27,127]. Some attempts to solve this problem have indeed been reported recently. In the approach of [146], it is shown that the noise correlation matrix can be updated during speech+noise periods under the assumption that the noise field is diffuse (unfortunately this is often not the case). In the approach of [83], the noise correlation matrix is continuously updated assuming the speech steering vector is known and one noise reference per microphone pair is available (or obtained using a classic noise estimation algorithm [27,127]). Finally, in [134] it is proposed to exploit prior knowlegde of both correlation matrices combined with a continuous updating approach. The weighting between prior knowledge and new data could then (for example) be based on an estimate of the conditional speech presence probability (which basically provides a continuous output value between 0 and 1, instead of a binary VAD decision). Although initial results are promising, further research is still required.

Appendix A

Appendix to Chapter 2

A.1 Optimal SDW-MWF

The SDW-MWF cost funtion for the left device (2.52) can be rewritten as [55]:

$$J_{\text{MWF}}(\mathbf{w}_L) = P_L - \mathbf{r}_{xL}^H \mathbf{w}_L - \mathbf{w}_L^H \mathbf{r}_{xL} + \mathbf{w}_L^H (\mathbf{R}_{xL} + \mu \mathbf{R}_{vL}) \mathbf{w}_L \ , \tag{A.1}$$

with

$$P_L = \mathcal{E}\{|X_{L,\text{ref}}|^2\} = \mathbf{e}_L^H \mathbf{R}_{xL} \mathbf{e}_L, \quad \mathbf{r}_{xL} = \mathcal{E}\{\mathbf{x}_L X_{L,\text{ref}}^*\} = \mathbf{R}_{xL} \mathbf{e}_L \ . \tag{A.2}$$

The optimal $\mathbf{w}_L$ which minimizes the cost function (A.1) is then found by setting the gradient with respect to $\mathbf{w}_L$ equal to zero, i.e.

$$\nabla J_{\text{MWF}}(\mathbf{w}_L) = -2\mathbf{R}_{xL}\mathbf{e}_L + 2(\mathbf{R}_{xL} + \mu \mathbf{R}_{vL})\mathbf{w}_L = \mathbf{0} \ , \tag{A.3}$$

so that the optimal $\mathbf{w}_L$ is equal to

$$\mathbf{w}_L = (\mathbf{R}_{xL} + \mu \mathbf{R}_{vL})^{-1} \mathbf{R}_{xL} \mathbf{e}_L \ . \tag{A.4}$$

A similar derivation can be made for the right device, resulting in:

$$\mathbf{w}_R = (\mathbf{R}_{xR} + \mu \mathbf{R}_{vR})^{-1} \mathbf{R}_{xR} \mathbf{e}_R \ . \tag{A.5}$$

A.2 Single target speech source SDW-MWF

For the special case of a single target speech source (section 2.1.4), the SDW-MWF (2.53) can be rewritten into a conceptually interesting form. The matrix inversion

lemma (also known as Woodbury identity) will be utilized in the derivations. It is given by

$$(\mathbf{U}\mathbf{C}\mathbf{V}^H + \mathbf{A})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left[\mathbf{C}^{-1} + \mathbf{V}^H\mathbf{A}^{-1}\mathbf{U}\right]^{-1}\mathbf{V}^H\mathbf{A}^{-1}\,, \tag{A.6}$$

where $\mathbf{U}$, $\mathbf{C}$, $\mathbf{V}$ and $\mathbf{A}$ are matrices of the correct size, e.g. where $\mathbf{U}$ and $\mathbf{V}$ are $P \times Q$, $\mathbf{C}$ is $Q \times Q$ and $\mathbf{A}$ is $P \times P$. Obviously, it is assumed that $\mathbf{C}$ and $\mathbf{A}$ are full-rank matrices. For the special case of a rank-one update of $\mathbf{A}$, (A.6) can be applied to find

$$(\mathbf{u}\mathbf{v}^H + \mathbf{A})^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^H\mathbf{A}^{-1}}{1 + \mathbf{v}^H\mathbf{A}^{-1}\mathbf{u}}\,, \tag{A.7}$$

where $\mathbf{u}$ and $\mathbf{v}$ are P-dimensional column vectors.

By using (2.23) and applying the matrix inversion lemma (A.6)-(A.7), it can be seen that (for the left device):

$$\begin{aligned}(\mathbf{R}_{xL} + \mu\mathbf{R}_{vL})^{-1} &= (P_s\mathbf{a}_L\mathbf{a}_L^H + \mu\mathbf{R}_{vL})^{-1}\,,\\ &= \frac{1}{\mu}\left(\mathbf{R}_{vL}^{-1} - \frac{P_s\mathbf{R}_{vL}^{-1}\mathbf{a}_L\mathbf{a}_L^H\mathbf{R}_{vL}^{-1}}{\mu + \rho_L}\right)\,,\end{aligned} \tag{A.8}$$

where

$$\rho_L = P_s\mathbf{a}_L^H\mathbf{R}_{vL}^{-1}\mathbf{a}_L\,. \tag{A.9}$$

The optimal left SDW-MWF (2.53) is then equal to

$$\begin{aligned}\mathbf{w}_L &= (\mathbf{R}_{xL} + \mu\mathbf{R}_{vL})^{-1}\mathbf{R}_{xL}\mathbf{e}_L\,,\\ &= \frac{1}{\mu}\left(\mathbf{R}_{vL}^{-1} - \frac{P_s\mathbf{R}_{vL}^{-1}\mathbf{a}_L\mathbf{a}_L^H\mathbf{R}_{vL}^{-1}}{\mu + \rho_L}\right)P_s\mathbf{a}_L\mathbf{a}_L^H\mathbf{e}_L\,,\\ &= \mathbf{R}_{vL}^{-1}\mathbf{a}_L\,\frac{P_s}{\mu + \rho_L}A_{L,\mathrm{ref}}^*\,.\end{aligned} \tag{A.10}$$

A similar derivation can be made for the right device, resulting in:

$$\mathbf{w}_R = \mathbf{R}_{vR}^{-1}\mathbf{a}_R\,\frac{P_s}{\mu + \rho_R}A_{R,\mathrm{ref}}^*\,. \tag{A.11}$$

Appendix B

Appendix to Chapter 7

In this section, the MWF-ITF optimal filters (7.34) are worked out for the special case of a single speech source. First, the derivation will be made for the special case $\gamma = 0$. This result will then be re-used in the derivation of the general case $\gamma \neq 0$.

B.1 MWF-ITF optimal filters for extension with noise ITF ($\gamma = 0$)

The filter in (7.34) reduces to

$$\mathbf{w}_{\mathrm{MWF-ITF},2} = (\mathbf{R} + \delta\mathbf{R}_{vt})^{-1}\mathbf{r}_x \ . \tag{B.1}$$

For easier notation, we define $\alpha_v = -\mathrm{ITF}_v^{in}$, such that (7.28) can be written as

$$\mathbf{R}_{vt} = \begin{bmatrix} \mathbf{I}_{2M} \\ \alpha_v \mathbf{I}_{2M} \end{bmatrix} \mathbf{R}_v \begin{bmatrix} \mathbf{I}_{2M} & \alpha_v^* \mathbf{I}_{2M} \end{bmatrix} \ . \tag{B.2}$$

Using the matrix inversion lemma (A.6) it can be shown that $(\mathbf{R} + \delta\mathbf{R}_{vt})^{-1}$ is equal to

$$\begin{aligned} (\mathbf{R} + \delta\mathbf{R}_{vt})^{-1} \quad = \quad & \mathbf{R}^{-1} - \mathbf{R}^{-1} \begin{bmatrix} \mathbf{I}_{2M} \\ \alpha_v \mathbf{I}_{2M} \end{bmatrix} \\ & \times \left(\begin{bmatrix} \mathbf{I}_{2M} & \alpha_v^* \mathbf{I}_{2M} \end{bmatrix} \mathbf{R}^{-1} \begin{bmatrix} \mathbf{I}_{2M} \\ \alpha_v \mathbf{I}_{2M} \end{bmatrix} + \frac{\mathbf{R}_v^{-1}}{\delta} \right)^{-1} \\ & \times \begin{bmatrix} \mathbf{I}_{2M} & \alpha_v^* \mathbf{I}_{2M} \end{bmatrix} \mathbf{R}^{-1} \ . \end{aligned} \tag{B.3}$$

By defining

$$\mathbf{P} = \mathbf{I}_{2M} - \frac{P_s \mathbf{R}_v^{-1} \mathbf{a}\mathbf{a}^H}{\mu + \rho} \tag{B.4}$$

the matrix $\mathbf{R}^{-1}$ can be written as

$$\mathbf{R}^{-1} = \frac{1}{\mu} \begin{bmatrix} \mathbf{P}\mathbf{R}_v^{-1} & \mathbf{0}_{2M} \\ \mathbf{0}_{2M} & \mathbf{P}\mathbf{R}_v^{-1} \end{bmatrix} , \tag{B.5}$$

and the following useful expression can be formulated:

$$\mathbf{P}\mathbf{R}_v^{-1}\mathbf{a} = \frac{\mu}{\mu + \rho}\mathbf{R}_v^{-1}\mathbf{a} . \tag{B.6}$$

Using (B.5), it can be shown that

$$\mathbf{R}^{-1} \begin{bmatrix} \mathbf{I}_{2M} \\ \alpha_v \mathbf{I}_{2M} \end{bmatrix} = \frac{1}{\mu} \begin{bmatrix} \mathbf{P}\mathbf{R}_v^{-1} \\ \alpha_v \mathbf{P}\mathbf{R}_v^{-1} \end{bmatrix} \tag{B.7}$$

$$\begin{bmatrix} \mathbf{I}_{2M} & \alpha_v^* \mathbf{I}_{2M} \end{bmatrix} \mathbf{R}^{-1} \begin{bmatrix} \mathbf{I}_{2M} \\ \alpha_v \mathbf{I}_{2M} \end{bmatrix} = \frac{1 + |\alpha_v|^2}{\mu} \mathbf{P}\mathbf{R}_v^{-1} \tag{B.8}$$

such that (B.3) reduces to

$$\begin{aligned} (\mathbf{R} + \delta\mathbf{R}_{vt})^{-1} &= \mathbf{R}^{-1} - \frac{\delta}{\mu^2} \begin{bmatrix} \mathbf{P} \\ \alpha_v \mathbf{P} \end{bmatrix} \\ &\times \left(\mathbf{I}_{2M} + \frac{\delta(1 + |\alpha_v|^2)}{\mu} \mathbf{P} \right)^{-1} \begin{bmatrix} \mathbf{P}\mathbf{R}_v^{-1} & \alpha_v^* \mathbf{P}\mathbf{R}_v^{-1} \end{bmatrix} . \end{aligned} \tag{B.9}$$

The filter in (B.1) is then equal to

$$\mathbf{w}_{\mathrm{MWF-ITF},2} = (\mathbf{R} + \delta\mathbf{R}_{vt})^{-1} P_s \begin{bmatrix} \mathbf{a} A_{L,\mathrm{ref}}^* \\ \mathbf{a} A_{R,\mathrm{ref}}^* \end{bmatrix} , \tag{B.10}$$

$$\begin{aligned} &= \mathbf{R}^{-1} P_s \begin{bmatrix} \mathbf{a} A_{L,\mathrm{ref}}^* \\ \mathbf{a} A_{R,\mathrm{ref}}^* \end{bmatrix} - \frac{\delta}{\mu^2} \begin{bmatrix} \mathbf{P} \\ \alpha_v \mathbf{P} \end{bmatrix} \\ &\quad \times \left(\mathbf{I}_{2M} + \frac{\delta(1 + |\alpha_v|^2)}{\mu} \mathbf{P} \right)^{-1} \\ &\quad \times \mathbf{P}\ \mathbf{R}_v^{-1}\mathbf{a}\ P_s (A_{L,\mathrm{ref}}^* + \alpha_v^* A_{R,\mathrm{ref}}^*) . \end{aligned} \tag{B.11}$$

Using the matrix inversion lemma (A.6), it can be shown that

$$(\mathbf{I}_{2M} + c\mathbf{P})^{-1} = \frac{1}{c+1} \left[\mathbf{I}_{2M} + \frac{c P_s \mathbf{R}_v^{-1} \mathbf{a}\mathbf{a}^H}{(c+1)\mu + \rho} \right] , \tag{B.12}$$

with c an arbitrary constant, such that

$$(\mathbf{I}_{2M} + c\mathbf{P})^{-1}\mathbf{R}_v^{-1}\mathbf{a} = \frac{\mu+\rho}{(c+1)\mu+\rho}\mathbf{R}_v^{-1}\mathbf{a} \tag{B.13}$$

Using (7.1), (B.6) and (B.13), the filter in (B.11) is equal to

$$\begin{aligned}\mathbf{w}_{\mathrm{MWF-ITF},2} &= \frac{P_s}{\mu+\rho}\begin{bmatrix} A_{L,\mathrm{ref}}^*\mathbf{R}_v^{-1}\mathbf{a} \\ A_{R,\mathrm{ref}}^*\mathbf{R}_v^{-1}\mathbf{a} \end{bmatrix} \\ &\quad -\frac{\delta}{\mu}\frac{P_s(A_{L,\mathrm{ref}}^* + \alpha_v^* A_{R,\mathrm{ref}}^*)}{\mu+\rho+\delta(1+|\alpha_v|^2)}\begin{bmatrix} \mathbf{P} \\ \alpha_v\mathbf{P} \end{bmatrix}\mathbf{R}_v^{-1}\mathbf{a}\,, \end{aligned} \tag{B.14}$$

such that, using (B.6), and by defining

$$\xi = \frac{\delta}{\mu+\rho+\delta(1+|\alpha_v|^2)}\,, \tag{B.15}$$

the optimal MWF-ITF filter reduces to expression

$$\boxed{\begin{aligned}\mathbf{w}_{\mathrm{MWF-ITF},2} &= (\mathbf{R}+\delta\mathbf{R}_{vt})^{-1}P_s\begin{bmatrix} \mathbf{a}A_{L,\mathrm{ref}}^* \\ \mathbf{a}A_{R,\mathrm{ref}}^* \end{bmatrix} \\ &= \frac{P_s}{\mu+\rho}\begin{bmatrix} \left[A_{L,\mathrm{ref}}^* - \xi(A_{L,\mathrm{ref}}^* + \alpha_v^* A_{R,\mathrm{ref}}^*)\right]\mathbf{R}_v^{-1}\mathbf{a} \\ \left[A_{R,\mathrm{ref}}^* - \xi\alpha_v(A_{L,\mathrm{ref}}^* + \alpha_v^* A_{R,\mathrm{ref}}^*)\right]\mathbf{R}_v^{-1}\mathbf{a} \end{bmatrix} \end{aligned}} \tag{B.16}$$

B.2 MWF-ITF optimal filters for extension with speech and noise ITFs ($\gamma \neq 0$)

For easier notation, we define $\alpha_x = -\mathrm{ITF}_x^{in}$, such that $\mathbf{R}_{xt}$ is equal to

$$\begin{aligned}\mathbf{R}_{xt} &= \begin{bmatrix} \mathbf{I}_{2M} \\ \alpha_x\mathbf{I}_{2M} \end{bmatrix}\mathbf{R}_x\begin{bmatrix} \mathbf{I}_{2M} & \alpha_x^*\mathbf{I}_{2M} \end{bmatrix} \\ &= P_s\begin{bmatrix} \mathbf{a} \\ \alpha_x\mathbf{a} \end{bmatrix}\begin{bmatrix} \mathbf{a}^H & \alpha_x^*\mathbf{a}^H \end{bmatrix}\,. \end{aligned} \tag{B.17}$$

Hence, using the matrix inversion lemma (A.6), the filter in (7.34) is equal to

$$\begin{aligned}\left[(\mathbf{R}+\delta\mathbf{R}_{vt})+\gamma\mathbf{R}_{xt}\right]^{-1}\mathbf{r}_x &= (\mathbf{R}+\delta\mathbf{R}_{vt})^{-1}\mathbf{r}_x \\ &\quad -\frac{\gamma P_s(\mathbf{R}+\delta\mathbf{R}_{vt})^{-1}\begin{bmatrix} \mathbf{a} \\ \alpha_x\mathbf{a} \end{bmatrix}\begin{bmatrix} \mathbf{a}^H & \alpha_x^*\mathbf{a}^H \end{bmatrix}(\mathbf{R}+\delta\mathbf{R}_{vt})^{-1}\mathbf{r}_x}{1+\gamma P_s\begin{bmatrix} \mathbf{a}^H & \alpha_x^*\mathbf{a}^H \end{bmatrix}(\mathbf{R}+\delta\mathbf{R}_{vt})^{-1}\begin{bmatrix} \mathbf{a} \\ \alpha_x\mathbf{a} \end{bmatrix}} \end{aligned} \tag{B.18}$$

Similarly to (B.16), by setting $A^*_{L,\text{ref}} = 1$ and $A^*_{R,\text{ref}} = \alpha_x$, it can be shown that

$$P_s(\mathbf{R} + \delta\mathbf{R}_{vt})^{-1} \begin{bmatrix} \mathbf{a} \\ \alpha_x\,\mathbf{a} \end{bmatrix} = \frac{P_s}{\mu+\rho} \begin{bmatrix} (1-\xi\alpha_t)\mathbf{R}_v^{-1}\mathbf{a} \\ (\alpha_x - \alpha_v\xi\alpha_t)\mathbf{R}_v^{-1}\mathbf{a} \end{bmatrix} \tag{B.19}$$

with

$$\alpha_t = 1 + \alpha_v^*\alpha_x \,. \tag{B.20}$$

Using (B.19), it can be shown that

$$P_s \begin{bmatrix} \mathbf{a}^H & \alpha_x^*\,\mathbf{a}^H \end{bmatrix} (\mathbf{R} + \delta\mathbf{R}_{vt})^{-1} \begin{bmatrix} \mathbf{a} \\ \alpha_x\,\mathbf{a} \end{bmatrix} = \frac{\rho}{\mu+\rho}(1 + |\alpha_x|^2 - \xi|\alpha_t|^2) \,. \tag{B.21}$$

Using (B.16) and $\alpha_x = -\frac{A_{L,\text{ref}}}{A_{R,\text{ref}}}$, it can be shown that

$$\begin{bmatrix} \mathbf{a}^H & \alpha_x^*\,\mathbf{a}^H \end{bmatrix} (\mathbf{R} + \delta\mathbf{R}_{vt})^{-1}\mathbf{r}_x = -\frac{\rho\xi\alpha_t^*}{\mu+\rho}(A^*_{L,\text{ref}} + \alpha_v^* A^*_{R,\text{ref}}) \,. \tag{B.22}$$

By plugging (B.16), (B.19), (B.21) and (B.22) into (B.18), and by defining

$$\nu = \frac{\gamma\rho}{\mu + \rho + \gamma\rho(1 + |\alpha_x|^2 - \xi|\alpha_t|^2)} \,, \tag{B.23}$$

the filter $\mathbf{w}_{\text{MWF}-\text{ITF},2}$ reduces to expression (7.36), i.e.

$$\begin{aligned}
\mathbf{w}_{\text{MWF}-\text{ITF},2,L} \;&=\; \mathbf{R}_v^{-1}\mathbf{a}\frac{P_s}{\mu+\rho} \\
&\quad\times\Big\{A^*_{L,\text{ref}} - \xi(A^*_{L,\text{ref}} - \text{ITF}_v^{in,*}A^*_{R,\text{ref}})\big[1 - \nu\alpha_t^*(1-\xi\alpha_t)\big]\Big\} \\
\mathbf{w}_{\text{MWF}-\text{ITF},2,R} \;&=\; \mathbf{R}_v^{-1}\mathbf{a}\frac{P_s}{\mu+\rho} \\
&\quad\times\Big\{A^*_{R,\text{ref}} + \xi\text{ITF}_v^{in}(A^*_{L,\text{ref}} - \text{ITF}_v^{in,*}A^*_{R,\text{ref}})\Big[1 - \nu\alpha_t^*\Big(\frac{\text{ITF}_x^{in}}{\text{ITF}_v^{in}} - \xi\alpha_t\Big)\Big]\Big\}
\end{aligned} \tag{B.24}$$

Appendix C

Appendix to Chapter 8

C.1 Relative speech ILD error of SDW-MWF and MWF-η

In this section, expressions for the relative speech ILD error introduced by the binaural SDW-MWF and MWF-η are given. The influence of parameters η and μ on the relative ILD error is then studied.

The output speech ILD of the MWF-η (8.13) can be rewritten as:

$$\mathrm{ILD}_x^{out} = \left(1 + \frac{(\rho_L - \rho_R)\ \mu\ (1-\eta)\Big(\mu(1+\eta)(\rho_L+\rho_R) + 2\rho_L\rho_R + 2\mu^2\eta\Big)}{(\mu+\rho_L)^2\ (\mu\eta+\rho_R)^2}\right) \mathrm{ILD}_x^{in} \quad \text{(C.1)}$$

The relative ILD error is then equal to:

$$\Delta\mathrm{ILD}_x = \frac{|\mathrm{ILD}_x^{out} - \mathrm{ILD}_x^{in}|}{\mathrm{ILD}_x^{in}}\,, \quad \text{(C.2)}$$

$$= \frac{|\rho_L - \rho_R|\ \mu\ (1-\eta)\Big(\mu(1+\eta)(\rho_L+\rho_R) + 2\rho_L\rho_R + 2\mu^2\eta\Big)}{(\mu+\rho_L)^2\ (\mu\eta+\rho_R)^2}\,, \quad \text{(C.3)}$$

where $(1-\eta) \geq 0$ for $0 \leq \eta \leq 1$. As the partial derivative of (C.3) with respect to η is negative for $\rho_L \neq \rho_R$, $\rho_L > 0$, $\rho_R > 0$ and $\mu > 0$, i.e.

$$\frac{\delta}{\delta\eta}\left(\Delta\mathrm{ILD}_x\right) = \frac{-2\ |\rho_L - \rho_R|\ \mu\ (\mu+\rho_R)^2(\mu\eta+\rho_L)}{(\mu+\rho_L)^2(\mu\eta+\rho_R)^3} < 0\,, \quad \text{(C.4)}$$

the ILD error strictly decreases as η increases. Thus, by sacrificing noise reduction performance, the speech ILD error can be reduced by the MWF-η (in addition to reducing the noise ITD and ILD errors, cfr. chapter 7 and [31]).

The relative ILD error of the SDW-MWF is found by setting $\eta = 0$ in (C.3):

$$\Delta\text{ILD}_{x,\ \eta=0} = \frac{|\rho_L - \rho_R|\ \mu\ (\mu(\rho_L + \rho_R) + 2\rho_L\rho_R)}{(\mu + \rho_L)^2\ \rho_R^2}\ . \tag{C.5}$$

By calculating the derivative from (C.5) with respect to μ,

$$\frac{\delta}{\delta\mu}\left(\Delta\text{ILD}_{x,\ \eta=0}\right) = \frac{2\ |\rho_L - \rho_R|\ (\mu + \rho_R)\ \rho_L^2}{(\mu + \rho_L)^3\ \rho_R^2} > 0\ , \tag{C.6}$$

it is seen that the speech ILD error strictly increases as μ increases. Increasing the noise reduction performance (increasing μ) therefore comes at the cost of introducing more speech ILD distortion.

C.2 Common postfilter for MWF-Front

In this section it is shown that the CP can be efficiently calculated for the binaural MWF-Front scheme, by making use of the QRD-RLS algorithm proposed in chapter 6 and [34]. In the QRD-RLS algorithm, the upper triangular Cholesky factors $\mathbf{R}_{v\Delta L}$ and $\mathbf{R}_{v\Delta R}$ are stored and updated instead of the noise correlation matrices (with $\mathbf{R}_{vL} = \mathbf{R}_{v\Delta L}^H\mathbf{R}_{v\Delta L}$ and similarly for $\mathbf{R}_{vR}$). The column vectors $\mathbf{b}_L$ and $\mathbf{b}_R$ are also defined as:

$$\mathbf{b}_L = \mathbf{R}_{v\Delta L}^{-H}\mathbf{R}_{yL}\mathbf{e}_L, \quad \mathbf{b}_R = \mathbf{R}_{v\Delta R}^{-H}\mathbf{R}_{yR}\mathbf{e}_R\ . \tag{C.7}$$

It was shown in chapter 6 that $\mathbf{b}_L$ and $\mathbf{b}_R$ can be updated together with $\mathbf{R}_{v\Delta L}$ and $\mathbf{R}_{v\Delta R}$ by applying sequences of unitary Givens rotations. Using these definitions, the SP-MWF (8.1) is found to be equivalent to

$$\begin{aligned}\mathbf{w}_{\text{SP}-\text{MWF},L} &= (\mathbf{m}_{vyL} - \mathbf{e}_L) \\ &\quad \times \frac{< \mathbf{R}_{v\Delta L}\mathbf{e}_L\ ,\ \mathbf{b}_L - \mathbf{R}_{v\Delta L}\mathbf{e}_L >}{< \mathbf{b}_L + (\mu - 1)\mathbf{R}_{v\Delta L}\mathbf{e}_L\ ,\ \mathbf{b}_L - \mathbf{R}_{v\Delta L}\mathbf{e}_L >}\ , \end{aligned} \tag{C.8}$$

$$\begin{aligned}\mathbf{w}_{\text{SP}-\text{MWF},R} &= (\mathbf{m}_{vyR} - \mathbf{e}_R) \\ &\quad \times \frac{< \mathbf{R}_{v\Delta R}\mathbf{e}_R\ ,\ \mathbf{b}_R - \mathbf{R}_{v\Delta R}\mathbf{e}_R >}{< \mathbf{b}_R + (\mu - 1)\mathbf{R}_{v\Delta R}\mathbf{e}_R\ ,\ \mathbf{b}_R - \mathbf{R}_{v\Delta R}\mathbf{e}_R >}\ , \end{aligned} \tag{C.9}$$

where the dotproduct is defined as $< \mathbf{a}_1, \mathbf{a}_2 >= \mathbf{a}_2^H \mathbf{a}_1$. These expressions can again be structured as an MVDR followed by a spectral postfilter as in (8.9), i.e.

$$\mathbf{w}_{\mathrm{SP-MWF},L} = \underbrace{(\mathbf{m}_{vyL} - \mathbf{e}_L) \, . \, \frac{< \mathbf{R}_{v\Delta L}\mathbf{e}_L \, , \, \mathbf{b}_L - \mathbf{R}_{v\Delta L}\mathbf{e}_L >}{||\mathbf{b}_L - \mathbf{R}_{v\Delta L}\mathbf{e}_L||^2}}_{\text{MVDR}} \times \underbrace{\frac{||\mathbf{b}_L - \mathbf{R}_{v\Delta L}\mathbf{e}_L||^2}{< \mathbf{b}_L + (\mu - 1)\mathbf{R}_{v\Delta L}\mathbf{e}_L \, , \, \mathbf{b}_L - \mathbf{R}_{v\Delta L}\mathbf{e}_L >}}_{\text{postfilter}} \quad \text{(C.10)}$$

and similarly for $\mathbf{w}_{\mathrm{SP-MWF},R}$. In expressions (C.8)-(C.10), $\mathbf{m}_{vyL}$ and $\mathbf{m}_{vyR}$ are found by solving the following systems of equations, which can be done by a simple backsubstitution:

$$\mathbf{R}_{v\Delta L} \, \mathbf{m}_{vyL} = \mathbf{b}_L, \quad \mathbf{R}_{v\Delta R} \, \mathbf{m}_{vyR} = \mathbf{b}_R \, . \quad \text{(C.11)}$$

The signal vectors $\mathbf{y}_L$ and $\mathbf{y}_R$ are now redefined as

$$\mathbf{y}_L = \begin{bmatrix} Y_{0,0} \\ Y_{1,0} \\ \hdashline Y_{0,1} \\ \vdots \\ Y_{0,M-1} \end{bmatrix}, \quad \mathbf{y}_R = \begin{bmatrix} Y_{0,0} \\ Y_{1,0} \\ \hdashline Y_{1,1} \\ \vdots \\ Y_{1,M-1} \end{bmatrix}, \quad \text{(C.12)}$$

so that the front microphone signals are now the first two entries of each vector (in the same order). $\mathbf{e}_L$ and $\mathbf{e}_R$ are also modified accordingly. As a result of this reordering, the upper-left 2-by-2 submatrices in $\mathbf{R}_{v\Delta L}$ and $\mathbf{R}_{v\Delta R}$ are equal, and by definition (C.7), it can be seen that the first two entries of vectors $\mathbf{b}_L$ and $\mathbf{b}_R$ only depend on the front microphone signals (as the upper-left 2-by-2 submatrices of $\mathbf{R}_{v\Delta L}^{-H}$ and $\mathbf{R}_{v\Delta R}^{-H}$ are also equal and only depend on the front microphone signals). As a result, these quantities can be calculated in both devices and moreover, they are identical to the values which would be obtained by calculating a separate (QRD-RLS) SP-MWF on only the front microphone signals. To derive the CP, it is thus only necessary to store and update (the first two entries of) $\mathbf{b}_L$ and $\mathbf{b}_R$ in both devices, and then select the same vector at both sides (for example the vector of the best ear). For example, assuming the left side is the best ear, the CP is found as:

$$\mathrm{CP}\ (\mu) = \frac{||\mathbf{b}_L^{''} - \mathbf{R}_{v\Delta L}^{''}\mathbf{e}_L||^2}{< \mathbf{b}_L^{''} + (\mu - 1)\mathbf{R}_{v\Delta L}^{''}\mathbf{e}_L \, , \, \mathbf{b}_L^{''} - \mathbf{R}_{v\Delta L}^{''}\mathbf{e}_L >} \quad \text{(left device)} \, , \quad \text{(C.13)}$$

$$= \frac{||\mathbf{b}_L^{''} - \mathbf{R}_{v\Delta R}^{''}\mathbf{e}_L||^2}{< \mathbf{b}_L^{''} + (\mu - 1)\mathbf{R}_{v\Delta R}^{''}\mathbf{e}_L \, , \, \mathbf{b}_L^{''} - \mathbf{R}_{v\Delta R}^{''}\mathbf{e}_L >} \quad \text{(right device)} \, , \quad \text{(C.14)}$$

where

$$\mathbf{R}''_{v\Delta L} = \left[\begin{array}{c|c} \mathbf{I}_2 & \mathbf{0}_{2\times(M-1)} \end{array}\right] \mathbf{R}_{v\Delta L}, \quad \mathbf{b}''_L = \left[\begin{array}{c|c} \mathbf{I}_2 & \mathbf{0}_{2\times(M-1)} \end{array}\right] \mathbf{b}_L \tag{C.15}$$

are the first two rows of $\mathbf{R}_{v\Delta L}$ and $\mathbf{b}_L$. The CP can thus be extracted from the $(M+1)$-microphone MVDR solution by calculating two simple dotproducts.

Bibliography

[1] R. Aichner, H. Buchner, F. Yan, and W. Kellermann. A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments. *Signal Process.*, 86(6):1260–1277, Jun. 2006.

[2] R. Aichner, H. Buchner, M. Zourub, and W. Kellermann. Multi-channel source separation preserving spatial information. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages I5–I8, Honolulu, HI, USA, Apr. 2007.

[3] M. Akeroyd. A binaural cross-correlogram toolbox for MATLAB. *"www.ihr.mrc.ac.uk/downloads/products/matlab/binauraltoolbox"*, 2001.

[4] V.R. Algazi and R.O. Duda. Headphone-based spatial sound. *IEEE Signal Process. Mag.*, 28(1):33–42, Jan. 2011.

[5] ANSI-SII. ANSI S3.5-1997 American National Standard Methods for Calculation of the Speech Intelligibility Index. *Acoust. Soc. Amer., Melville, NY*, Jun. 1997.

[6] Auditec. Auditory tests (revised). *Compact Disc, Auditec, St. Louis*, 1997.

[7] D. R. Begault. Auditory and non-auditory factors that potentially influence virtual acoustic imagery. In *Proc. AES Int. Conf. Spatial Sound Reproduction*, Rovaniemi, Finland, Apr. 1999.

[8] J. Benesty, J. Chen, and Y. Huang. *Noncausal (frequency-domain) optimal filters*, chapter 6 in "Microphone Array Signal Processing", pages 115–137. Springer-Verlag, 2008.

[9] V. Berisha, H. Kwon, and S. Spanias. Real-time implementation of a distributed voice activity detector. In *IEEE Workshop on Sensor Array and Multichannel Process.*, pages 659–662, Waltham, MA, USA, Jul. 2006.

[10] L. R. Bernstein, S. van de Par, and C. Trahiotis. The normalized interaural correlation: Accounting for NoSπ thresholds obtained with Gaussian and "low-noise" masking noise. *J. Acoust. Soc. Amer.*, 106(2):870–876, Aug. 1999.

[11] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, volume 4, pages 208–211, Apr. 1979.

[12] A. Bertrand and M. Moonen. Robust distributed noise reduction in hearing aids with external acoustic sensor nodes. *EURASIP J. Adv. Signal Process.*, 14 pages, 2009.

[13] D. Bindel, J. Demmel, W. Kahan, and O. Marques. On computing Givens rotations reliably and efficiently. *ACM Transactions on Mathematical Software (TOMS)*, 28(2):206–238, Jun. 2002.

[14] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localisation* (Revised Edition). MIT Press, 1996.

[15] S. F. Boll. Suppression of Acoustic Noise in Speech Using Spectral Subtraction. *IEEE Trans. Acoust., Speech, Signal Process.*, 27(2):113–120, Apr. 1979.

[16] A. Boothroyd, K. Fitz, J. Kindred, S. Kochkin, H. Levitt, B. Moore, and J. Yantz. Hearing aids and wireless technology. *Hearing Rev.*, 14(6):44–48, Jun. 2007.

[17] A. S. Bregman. *Auditory scene analysis*. Cambridge: MIT Press., 1990.

[18] R. Brennan and T. Schneider. A flexible filterbank structure for extensive signal manipulations in digital hearing aids. In *Proc. IEEE Int. Symposium on Circuits and Systems (ISCAS)*, volume 6, pages 569–572, May-Jun. 1998.

[19] 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.*, 83(4):1508–1516, Apr. 1988.

[20] H. Buchner, R. Aichner, and W. Kellermann. A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics. *IEEE Trans. Speech Audio Process.*, 13(1):120–134, Jan. 2005.

[21] J. Capon. High-resolution frequency-wavenumber spectrum analysis. *Proc. IEEE*, 57(8):1408–1418, Aug. 1969.

[22] J. Catic, T. Dau, J. M. Buchholz, and F. Gran. The effect of a voice activity detector on the speech enhancement performance of the binaural multichannel Wiener filter. *EURASIP J. Audio, Speech, Music Process.*, 12 pages, 2010.

[23] J. Chen, J. Benesty, and Y. Huang. A minimum distortion noise reduction algorithm with multiple microphones. *IEEE Trans. Audio, Speech, Lang. Process.*, 16(3):481–493, Mar. 2008.

[24] J. Chen, J. Benesty, Y. Huang, and S. Doclo. New insights into the noise reduction Wiener filter. *IEEE Trans. Audio, Speech, Lang. Process.*, 14(4):1218–1234, Jul. 2006.

[25] E. C. Cherry. Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Amer.*, 25(5):975–979, Sep. 1953.

[26] P. Clarkson and S. Bahgat. Envelope expansion methods for speech enhancement. *J. Acoust. Soc. Amer.*, 89(3):1378–1382, Mar. 1991.

[27] I. Cohen. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. *IEEE Trans. Audio, Speech, Lang. Process.*, 11(5):466–475, Sep. 2003.

[28] T. Coleman, M. A. Branch, and A. Grace. *MATLAB Optimization Toolbox User's Guide*. The Mathworks Inc., Natick MA, USA, Jan. 1999.

[29] P. Comon. Independent component analysis, A new concept? *Signal Process.*, 36(3):287–314, Apr. 1994.

[30] B. Cornelis, S. Doclo, T. Van Den Bogaert, M. Moonen, and J. Wouters. Analysis of localization cue preservation by Multichannel Wiener Filtering based binaural noise reduction in hearing aids. In *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008.

[31] 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.*, 18(2):342–355, Feb. 2010.

[32] B. Cornelis, M. Moonen, and J. Wouters. Comparison of frequency domain noise reduction strategies based on Multichannel Wiener Filtering and Spatial Prediction. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 129–132, Tai Pei, Taiwan, Apr. 2009.

[33] B. Cornelis, M. Moonen, and J. Wouters. Binaural cue preservation in binaural hearing aids with reduced-bandwidth multichannel Wiener filter based noise reduction. In *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Tel-Aviv, Israel, Sep. 2010.

[34] B. Cornelis, M. Moonen, and J. Wouters. A QRD-RLS based frequency-domain Multichannel Wiener Filter algorithm for noise reduction in hearing aids. In *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, pages 1953–1957, Aalborg, Denmark, Aug. 2010.

[35] B. Cornelis, M. Moonen, and J. Wouters. Binaural voice activity detection for MWF-based noise reduction in binaural hearing aids. In *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, pages 486–490, Barcelona, Spain, Aug. 2011.

[36] B. Cornelis, M. Moonen, and J. Wouters. Performance analysis of multichannel Wiener filter based noise reduction in hearing aids under second order statistics estimation errors. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(5):1368–1381, Jul. 2011.

[37] B. Cornelis, M. Moonen, and J. Wouters. Reduced-bandwidth Multichannel Wiener Filter based binaural noise reduction and localization cue preservation in binaural hearing aids. *Submitted*, 2011.

[38] B. Cornelis, M. Moonen, and J. Wouters. Speech intelligibility improvements with hearing aids using bilateral and binaural adaptive multichannel wiener filtering based noise reduction. *Submitted*, 2011.

[39] B. Cornelis, M. Moonen, and J. Wouters. A VAD-robust Multichannel Wiener Filter algorithm for noise reduction in hearing aids. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 281–284, Prague, Czech Republic, May 2011.

[40] R. Crochiere. A weighted overlap-add method of short-time Fourier analysis/synthesis. *IEEE Trans. Acoust., Speech, Signal Process.*, 28(1):99–102, Feb. 1980.

[41] C. J. Darwin. Contributions of binaural information to the separation of different sound sources. *Int. J. Audiol.*, 45(Suppl. 1):S20–S24, 2006.

[42] P. Dawson, S. Mauger, and A. Hersbach. Clinical evaluation of signal-to-noise ratio based noise reduction in Nucleus cochlear implant recipients. *Ear Hear.*, 32(3):382–390, May-Jun. 2011.

[43] B. de Vries and R. A. J. de Vries. An integrated approach to hearing aid algorithm design for enhancement of audibility, intelligibility and comfort. In *Proc. of the IEEE Benelux Signal Process. Symp. (SPS2004)*, pages 65–68, Hilvarenbeek, The Netherlands, Apr. 2004.

[44] B. Defraene, T. van Waterschoot, H. J. Ferreau, M. Diehl, and M. Moonen. Perception-based clipping of audio signals. In *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, pages 517–521, Aalborg, Denmark, Aug. 2010.

[45] M. Dendrinos, S. Bakamidis, and G. Carayannis. Speech enhancement from noise : A regenerative approach. *Speech Commun.*, 10(2):45–57, Feb. 1991.

[46] J.G. Desloge, W.M. Rabinowitz, and P.M. Zurek. Microphone-array hearing aids with binaural output–Part I: Fixed-processing systems. *IEEE Trans. Speech Audio Process.*, 5(6):529–542, Nov. 1997.

[47] H. Dillon. *Hearing Aids*. Boomerang Press, Australia, 2001.

[48] S. Doclo. *Multi-microphone noise reduction and dereverberation techniques for speech applications*. PhD thesis, ESAT, Katholieke Universiteit Leuven, Belgium, May 2003.

[49] S. Doclo, S. Gannot, M. Moonen, and A. Spriet. *Acoustic Beamforming for Hearing Aid Applications*, chapter 9 in "Handbook on Array Processing and Sensor Networks" (Haykin, S. and Ray Liu, K. J., Eds.), pages 269–302. John Wiley and Sons, Inc., 2009.

[50] S. Doclo, T. J. Klasen, T. Van den Bogaert, J. Wouters, and M. Moonen. Theoretical analysis of binaural cue preservation using multi-channel Wiener filtering and interaural transfer functions. In *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Paris, France, Sep. 2006.

[51] S. Doclo, T. C. Lawin-Ore, and T. Rohdenburg. Rate-constrained binaural MWF-based noise reduction algorithms. In *ITG Fachtagung Speech Communication*, Bochum, Germany, Oct. 2010.

[52] S. Doclo and M. Moonen. GSVD-based optimal filtering for single and multimicrophone speech enhancement. *IEEE Trans. Signal Process.*, 50(9):2230–2244, Sep. 2002.

[53] S. Doclo and M. Moonen. On the Output SNR of the Speech-Distortion Weighted Multichannel Wiener Filter. *IEEE Signal Process. Lett.*, 12(12):809–811, Dec. 2005.

[54] S. Doclo, A. Spriet, J. Wouters, and M. Moonen. Frequency-Domain Criterion for Speech Distortion Weighted Multichannel Wiener Filter for Robust Noise Reduction. *Speech Commun.*, 49(7–8):636–656, Jul.-Aug. 2007.

[55] S. Doclo, T. Van den Bogaert, J. Wouters, and M. Moonen. Reduced-bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids. *IEEE Trans. Audio, Speech, Lang. Process.*, 17(1):38–51, Jan. 2009.

[56] B. Edwards. The future of hearing aid technology. *Trends Amplif.*, 11(1):31–46, Mar. 2007.

[57] G. W. Elko. Microphone array systems for hands-free telecommunication. *Speech Commun.*, 20(3-4):229–240, Dec. 1996.

[58] K. Eneman, A. Leijon, S. Doclo, A. Spriet, M. Moonen, and J. Wouters. *Auditory-profile-based physical evaluation of multi-microphone noise reduction techniques in hearing instruments*, chapter 15 in "Advances in Digital Speech Transmission ", (Martin, R. and Heute, U. and Antweiller, C., Eds.), pages 431–458. John Wiley and Sons, Inc., 2008.

[59] Y. Ephraim and D. Malah. Speech Enhancement Using a Minimun Mean-Square Error Short-Time Spectral Amplitude Estimator. *IEEE Trans. Acoust., Speech, Signal Process.*, 32(6):1109–1121, Dec. 1984.

[60] Y. Ephraim and H. L. Van Trees. A Signal Subspace Approach for Speech Enhancement. *IEEE Trans. Speech Audio Process.*, 3(4):251–266, Jul. 1995.

[61] D. Fabry, H. Mulder, and E. Dijkstra. Acceptance of the wireless microphone as a hearing aid accessory for adults. *Hearing J.*, 60(11):32–36, Nov. 2007.

[62] C. Faller and J. Merimaa. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *J. Acoust. Soc. Amer.*, 116(5):3075–3089, Nov. 2004.

[63] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, Jun. 2006.

[64] M. Fazel, H. Hindi, and S. Boyd. Rank minimization and applications in system theory. In *Proc. Amer. Control Conf.*, volume 4, pages 3273–3278, Jun. 2004.

[65] R. Fletcher. *Practical Methods of Optimization*. Wiley, New York, 1987.

[66] T. Francart, T. Van den Bogaert, M. Moonen, and J. Wouters. Amplification of interaural level differences improves sound localization in acoustic simulations of bimodal hearing. *J. Acoust. Soc. Amer.*, 126(6):3209–3213, Dec. 2009.

[67] T. Francart, A. van Wieringen, and J. Wouters. APEX 3: a multi-purpose test platform for auditory psychophysical experiments. *J. Neurosci. Methods*, 172(2):283–293, Jul. 2008.

[68] O. L. Frost III. An Algorithm for Linearly Constrained Adaptive Array Processing. *Proc. IEEE*, 60(8):926–935, Aug. 1972.

[69] S. Gannot, D. Burshtein, and E. Weinstein. Signal Enhancement Using Beamforming and Non-Stationarity with Applications to Speech. *IEEE Trans. Signal Process.*, 49(8):1614–1626, Aug. 2001.

[70] B. Gardner and K. Martin. HRTF measurements of a KEMAR dummy-head microphone. *MIT Media Lab Perceptual Computing, Tech. Rep. #280*, May 1994.

[71] B. Glasberg and B. Moore. Derivation of auditory filter shapes from notched-noise data. *Hear. Res.*, 47(1-2):103–138, Aug. 1990.

[72] S. Goetze, T. Rohdenburg, V. Hohmann, B. Kollmeier, and K. D. Kammeyer. Direction of arrival estimation based on the dual delay line approach for binaural hearing aid microphone arrays. In *Proc. Int. Symp. Intell. Signal Process. Communic. Systems (ISPACS)*, pages 84–87, Dec. 2007.

[73] G. H. Golub and C. F. Van Loan. *Matrix Computations*. MD : John Hopkins University Press, Baltimore, 3rd edition, 1996.

[74] 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.*, 94(5):3009–3010, Nov. 1993.

[75] J. E. Greenberg and P. M. Zurek. Evaluation of an adaptive beamforming method for hearing aids. *J. Acoust. Soc. Amer.*, 91(3):1662–1676, Mar. 1992.

[76] L. J. Griffiths and C. W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. Antennas Propagat.*, 30(1):27–34, Jan. 1982.

[77] E. 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.*, 18(1):158–170, Jan. 2010.

[78] V. Hamacher. Comparison of advanced monaural and binaural noise reduction algorithms for hearing aids. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 4008–4011, Orlando, FL, USA, May 2002.

[79] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass. Signal processing in high-end hearing aids: state of the art, challenges, and future trends. *EURASIP J. Appl. Signal Process.*, pages 2915–2929, 2005.

[80] V. Hamacher, U. Kornagel, T. Lotter, and H. Puder. *Binaural signal processing in hearing aids: Technologies and algorithms*, chapter 14 in "Advances in Digital Speech Transmission ", (Martin, R. and Heute, U. and Antweiller, C., Eds.), pages 401–429. John Wiley and Sons, Inc., 2008.

[81] W. Hartmann. How we localize sound. *Physics Today*, 52(11):24–29, Nov. 1999.

[82] S. Haykin and Z. Chen. The cocktail party problem. *Neural Computation*, 17(9):1875–1902, Sep. 2005.

[83] R. C. Hendriks and T. Gerkmann. Estimation of the noise correlation matrix. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 4740–4743, Prague, Czech Republic, May 2011.

[84] W. Herbordt, T. Trini, and W. Kellermann. Robust spatial estimation of the signal-to-interference ratio for non-stationary mixtures. In *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, pages 247–250, Kyoto, Japan, Sep. 2003.

[85] I. Hochberg, A. Boothroyd, M. Weiss, and S. Hellman. Effects of noise and noise suppression on speech perception by cochlear implant users. *Ear Hear.*, 13(4):263–271, Aug. 1992.

[86] Y. Hu and P. C. Loizou. Incorporating a psycho-acoustical model in frequency-domain speech enhancement. *IEEE Signal Process. Lett.*, 11(2):270–273, Feb. 2004.

[87] Y. Hu and P. C. Loizou. A comparative intelligibility study of single-microphone noise reduction algorithms. *J. Acoust. Soc. Amer.*, 122(3):1777–1786, Sep. 2007.

[88] Y. Huang, J. Benesty, and J. Chen. *Time delay estimation and source localization*, chapter 51 part I in "Springer handbook of speech processing" (Benesty, J. and Sondhi, M. and Huang, Y., Eds.), pages 1043–1064. Springer-Verlag, 2007.

[89] Y. Huang, J. Benesty, and J. Chen. Analysis and comparison of multichannel noise reduction methods in a common framework. *IEEE Trans. Audio, Speech, Lang. Process.*, 16(5):957–968, Jul. 2008.

[90] A. Hyvarinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, Jun. 2000.

[91] ISO3382. Acoustics - measurements of the reverberation time of rooms with reference to other acoustical parameters. *Internat. Org. for Standardization*, 1997.

[92] F. Jabloun and B. Champagne. Incorporating the human hearing properties in the signal subspace approach for speech enhancement. *IEEE Trans. Speech Audio Process.*, 11(6):700–708, Nov. 2003.

[93] J. Jensen and R. C. Hendriks. Spectral magnitude minimum mean-square error binary masks for DFT based speech enhancement. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 4736–4739, Prague, Czech Republic, May 2011.

[94] M. Jeub, M. Schafer, T. Esch, and P. Vary. Model-based dereverberation preserving binaural cues. *IEEE Trans. Audio, Speech, Lang. Process.*, 18(7):1732–1745, Sep. 2010.

[95] M. Jeub and P. Vary. Binaural dereverberation based on a dual-channel Wiener filter with optimized noise field coherence. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 4710–4713, Dallas, TX, USA, Mar. 2010.

[96] S. Kamath and P. C. Loizou. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Orlando, Florida, May 2002.

[97] A. H. Kamkar-Parsi and M. Bouchard. Instantaneous binaural target PSD estimation for hearing aid noise reduction in complex acoustic environments. *IEEE Trans. Instr. Meas.*, 60(4):1141–1154, Apr. 2011.

[98] G. Keidser, K Rohrseitz, H. Dillon, V. Hamacher, L. Carte, 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. *Int. J. Audiol.*, 45(10):563–579, Oct. 2006.

[99] H. Keppler, I. Dhooge, L. Maes, W. D'haenens, A. Bockstael, B. Philips, F. Swinnen, and B. Vinck. Short-term auditory effects of listening to an MP3 player. *Arch. Otolaryngol. Head Neck Surg.*, 136(6):538–548, Jun. 2010.

[100] T. J. Klasen, S. Doclo, T. Van den Bogaert, M. Moonen, and J. Wouters. Binaural multi-channel Wiener filtering for hearing aids: preserving interaural time and level differences. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 145–148, Toulouse, France, May 2006.

[101] T. J. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters. Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues. *IEEE Trans. Signal Process.*, 55(4):1579–1585, Apr. 2007.

[102] S. Kochkin. MarkeTrak VI: Consumers rate improvements sought in hearing instruments. *Hearing Rev.*, 9(11):18–20,22, Nov. 2002.

[103] S. Kochkin. MarkeTrak VII: Customer satisfaction with hearing instruments in the digital age. *Hearing J.*, 58(9):30–43, Sep. 2005.

[104] S. Kochkin. MarkeTrak VII: Hearing loss population tops 31 million people. *Hearing Rev.*, 12(7):16–29, Jul. 2005.

[105] S. Kochkin. Increasing hearing aid adoption through multiple environmental listening utility. *Hearing J.*, 60(11):28–29, Nov. 2007.

[106] S. Kochkin. MarkeTrak VIII: 25-year trends in the hearing health market. *Hearing Rev.*, 16(11):12–31, Oct. 2009.

[107] S. Kochkin and F. Kuk. The binaural advantage: evidence from subjective benefit and customer satisfaction data. *Hearing Rev.*, 4(4):29–34, Apr. 1997.

[108] B. Kollmeier, J. Peissig, and V. Hohmann. Real-time multiband dynamic compression and noise reduction for binaural hearing aids. *J. Rehabil. Res. Develop.*, 30(1):82–94, 1993.

[109] A. Koul and J. Greenberg. Using intermicrophone correlation to detect speech in spatially separated noise. *EURASIP J. Appl. Signal Process.*, 14 pages, 2006.

[110] F. Kuk, B. Crose, T. Kyhn, M. Morkelbjerg, M. L. Rank, M. Norgaard, and H. Pontoppidan. Digital wireless hearing aids, Part 3: Audiological benefits. *Hearing Rev.*, 18(8):48–56, Aug. 2011.

[111] N. Li and P. C. Loizou. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. *J. Acoust. Soc. Amer.*, 123(3):1673–1682, Mar. 2008.

[112] Y. Li and D. L. Wang. On the optimality of ideal binary time-frequency masks. *Speech Commun.*, 51(3):230–239, Mar. 2009.

[113] E. R. Libby. The search for the binaural advantage, revisited. *Hearing Rev.*, 14(12):22–26, Nov. 2007.

[114] J. S. Lim and A. V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proc. IEEE*, 67(12):1586–1604, Dec. 1979.

[115] G. Lindley. Accessing the far world: a new age of connectivity in hearing aids. *Hearing Rev.*, 14(5):54–59, May 2007.

[116] N. Linh-Trung, A. Belouchrani, K. Abed-Meraim, and B. Boashash. Separating more sources than sensors using time-frequency distributions. *EURASIP J. Appl. Signal Process.*, 2005:2828–2847, Jan. 2005.

[117] M. Lockwood, D. Jones, R. Bilger, C. Lansing, W. O'Brien, B. Wheeler, and A. Feng. Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms. *J. Acoust. Soc. Amer.*, 115(1):379–391, Jan. 2004.

[118] P. C. Loizou. *Speech enhancement: Theory and Practice*. CRC press, New York, USA, 2007.

[119] P. C. Loizou and G. Kim. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(1):47–56, Jan. 2011.

[120] A. Lombard, Y. Zheng, H. Buchner, and W. Kellermann. TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband Independent Component Analysis. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(6):1490–1503, Aug. 2011.

[121] T. Lotter and P. Vary. Dual-channel speech enhancement by superdirective beamforming. *EURASIP J. Appl. Signal Process.*, 14 pages, 2006.

[122] F. Luo, J. Yang, C. Pavlovic, and A. Nehorai. Adaptive null-forming scheme in digital hearing aids. *IEEE Trans. Signal Process.*, 50(7):1583–1590, Jul. 2002.

[123] H. Luts, K. Eneman, J. Wouters, M. Schulte, M. Vormann, M. Buechler, N. Dillier, R. Houben, W. A. Dreschler, M. Froehlich, H. Puder, G. Grimm, V. Hohmann, A. Leijon, A. Lombard, D. Mauler, and A. Spriet. Multicenter evaluation of signal enhancement algorithms for hearing aids. *J. Acoust. Soc. Amer.*, 127(3):1491–1505, Mar. 2010.

[124] H. Luts, J. B. Maj, W. Soede, and J. Wouters. Better speech perception in noise with an assistive multimicrophone array for hearing aids. *Ear Hear.*, 25(5):411–420, Oct. 2004.

[125] S. Makino, S. Araki, S. Winter, and H. Sawada. *Underdetermined Blind Source Separation using acoustic arrays*, chapter 10 in "Handbook on Array Processing and Sensor Networks" (Haykin, S. and Ray Liu, K. J., Eds.), pages 303–341. John Wiley and Sons, Inc., 2009.

[126] D. Marquardt, V. Hohmann, and S. Doclo. Performance comparison of binaural beamforming and MWF-based noise reduction algorithms for hearing aids. In *Fortschritte der Akustik - DAGA*, pages 639–640, Düsseldorf, Germany, Mar. 2011.

[127] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Audio, Speech, Lang. Process.*, 9(5):504–512, Jul. 2001.

[128] MathWorks. *xPC Target*. The Mathworks Inc., Natick MA, USA, 2000.

[129] R. J. McAulay and M. L. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust., Speech, Signal Process.*, 28(2):137–145, Apr. 1980.

[130] J. G. McWirther. Recursive least squares mimimization using a systolic array. *Proc. SPIE Real Time Signal Processing IV*, 431:105–112, 1983.

[131] J. Mejia, S. Carlille, and H. Dillon. Method and system for enhancing the intelligibility of sounds. *U.S. Patent 2009/0304188 A1*, May 2007.

[132] L. H. Mens. Speech understanding in noise with an eyeglass hearing aid: asymmetric fitting and the head shadow benefit of anterior microphones. *Int. J. Audiol.*, 50(1):27–33, Jan. 2011.

[133] I. Merks, M. Boone, and A. Berkhout. Design of a broadside array for a binaural hearing aid. In *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 1997.

[134] K. Ngo. *Digital signal processing algorithms for noise reduction, dynamic range compression and feedback cancellation in hearing aids*. PhD thesis, ESAT, Katholieke Universiteit Leuven, Belgium, Jul. 2011.

[135] M. Nilsson, S. D. Soli, and A. Sullivan. Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *J. Acoust. Soc. Amer.*, 95(2):1085–1099, Feb. 1994.

[136] R. Nishimura, Y. Suzuki, and F. Asano. A new adaptive binaural microphone array system using a weighted least squares algorithm. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 1925–1928, Orlando, FL, USA, May 2002.

[137] W. Noble, S. Sinclari, and D. Byrne. Improvement in aided sound localization with open earmolds: observations in people with high-frequency loss. *J. Amer. Acad. Aud.*, 9(1):25–34, Feb. 1998.

[138] M. Nyffeler. Field study on user control of directional focus: Benefits of hearing the facets of a full life. *Hearing Rev.*, 16(1):24–28, Jan. 2009.

[139] C.T. Pan and R.J. Plemmons. Least squares modifications with inverse factorization: parallel implications. *J. Computational and Applied Mathematics*, 27(1-2):109–127, Sep. 1989.

[140] A. Pandey and V. J. Mathews. Low-delay signal processing for digital hearing aids. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(4):699–710, May 2011.

[141] J. H. Park, M. H. Shin, and H. K. Kim. *Statistical Model-Based Voice Activity Detection Using Spatial Cues and Log Energy for Dual-Channel Noisy Speech Recognition*, volume 120 in "Communication and Networking", ser. Communications in Computer and Information Science (Kim, T. and Vasilakos, T. and Sakurai, K. and Xiao, Y. and Zhao, G. and Izak, D., Eds.), pages 172–179. Springer Berlin Heidelberg, 2010.

[142] R. Patterson, M. Allerhand, and C. Gigure. Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *J. Acoust. Soc. Amer.*, 98(4):1890–1894, Oct. 1995.

[143] R. Plomp. Auditory handicap of hearing impairment and the limited benefit of hearing aids. *J. Acoust. Soc. Amer.*, 63(2):533–549, Feb. 1978.

[144] R. Plomp and A. M. Mimpen. Improving the reliability of testing the speech reception threshold for sentences. *Audiology*, 18(1):43–52, 1979.

[145] T. A. Powers and P. Burton. Wireless binaural synchronization in hearing aids: Questions and answers. *Hearing Rev.*, 12(1):28–30,89, Jan. 2005.

[146] M. Rahmani, A. Akbari, B. Ayad, and B. Lithgow. Noise cross PSD estimation using phase information in diffuse noise field. *Signal Process.*, 89(5):703–709, May 2009.

[147] L. Rayleigh. On our perception of sound direction. *Philosoph. Mag.*, 13:214–232, 1907.

[148] K. Reindl, Y. Zheng, and W. Kellermann. Analysis of two generic Wiener filtering concepts for binaural speech enhancement in hearing aids. In *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, pages 989–993, Aalborg, Denmark, Aug. 2010.

[149] T. Ricketts. Directional hearing aids. *Trends Amplif.*, 5(4):139–176, Dec. 2001.

[150] T. Ricketts and P. Henry. Evaluation of an adaptive directional microphone hearing aid. *Int. J. Audiol.*, 41(2):100–112, Mar. 2002.

[151] T. Rohdenburg. *Development and Objective Perceptual Quality Assessment of Monaural and Binaural Noise Reduction Schemes for Hearing Aids*. PhD thesis, University of Oldenburg, Germany, Dec. 2008.

[152] G. Rombouts and M. Moonen. QRD-based unconstrained optimal filtering for acoustic noise reduction. *Signal Process.*, 83(9):1889–1904, Sep. 2003.

[153] O. Roy and M. Vetterli. Rate-constrained collaborative noise reduction for wireless hearing aids. *IEEE Trans. Signal Process.*, 57(2):645–657, Feb. 2009.

[154] H. Sawada, R. Mukai, S. Araki, and S. Makino. *Frequency-domain blind source separation*, chapter 13 in "Speech Enhancement" (Benesty, J. and Makino, S. and Chen, J., Eds.), pages 299–327. Springer-Verlag, 2005.

[155] B. Shield. Evaluation of the social and economic costs of hearing impairment. *Report for Hear-it AISBL*, 2006.

[156] B. G. Shinn-Cunningham, N. Kopco, and T. J. Martin. Localizing nearby sound sources in a classroom: Binaural room impulse responses. *J. Acoust. Soc. Amer.*, 117(5):3100–3115, May 2005.

[157] K. U. Simmer, J. Bitzer, and C. Marro. *Post-Filtering Techniques*, chapter 3 in "Microphone Arrays: Signal Processing Techniques and Applications" (Brandstein, M. S. and Ward, D. B., Eds.), pages 39–60. Springer-Verlag, May 2001.

[158] M. Slaney. Auditory toolbox version 2. *Technical Report no.1998-010, Interval Research Corp.*, Palo Alto, CA, 1998, *"http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010/"*.

[159] W. Soede, A. J. Berkhout, and F. A. Bilsen. Development of a directional hearing instrument based on array technology. *J. Acoust. Soc. Amer.*, 94(2):785–798, Aug. 1993.

[160] M. Souden, J. Benesty, and S. Affes. On optimal frequency-domain multichannel linear filtering for noise reduction. *IEEE Trans. Audio, Speech, Lang. Process.*, 18(2):260–276, Feb. 2010.

[161] A. Spriet, S. Doclo, M. Moonen, and J. Wouters. A unification of adaptive multi-microphone noise reduction systems. In *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Paris, France, Sep. 2006.

[162] A. Spriet, M. Moonen, and J. Wouters. Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction. *Signal Process.*, 84(12):2367–2387, Dec. 2004.

[163] 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 Process.*, 85(6):1073–1088, Jun. 2005.

[164] A. Spriet, M. Moonen, and J. Wouters. Robustness analysis of multi-channel Wiener filtering and Generalized Sidelobe Cancellation for multi-microphone noise reduction in hearing aid applications. *IEEE Trans. Speech Audio Process.*, 13(4):487–503, Jul. 2005.

[165] A. Spriet, M. Moonen, and J. Wouters. Stochastic gradient-based implementation of spatially preprocessed speech distortion weighted multichannel Wiener filtering for noise reduction in hearing aids. *IEEE Trans. Signal Process.*, 53(3):911–925, Mar. 2005.

[166] 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 Hear.*, 28(1):62–72, Feb. 2007.

[167] S. Srinivasan. Low-bandwidth binaural beamforming. *Electronics Letters*, 44(22):1292–1293, Oct. 2008.

[168] S. Srinivasan. Noise reduction in binaural hearing aids: Analyzing the benefit over monaural systems. *J. Acoust. Soc. Amer.*, 124(6):EL353–EL359, 2008.

[169] S. Srinivasan and A. den Brinker. Rate-constrained beamforming in binaural hearing aids. *EURASIP J. Adv. Signal Process.*, 9 pages, 2009.

[170] S. Srinivasan and K. Janse. Spatial audio activity detection for hearing aids. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 4021–4024, Las Vegas, USA, Apr. 2008.

[171] M. Stone and B. Moore. Tolerable hearing aid delays. I. Estimation of limits imposed by the auditory path alone using simulated hearing losses. *Ear Hear.*, 20(3):182–192, Jun. 1999.

[172] Q. Tao, R. van Rootseler, R. Veldhuis, S. Gehlen, and F. Weber. Optimal decision fusion and its application on 3D face recognition. In *Proc. of the Spec. Int. Group on Biometr. and Electr. Signatures*, pages 15–24, Darmstadt, Germany, Jul. 2007.

[173] T. D. Trine and D. Van Tasell. Digital hearing aid design: Fact vs. fantasy. *Hear. J.*, 55(2):36–38, 40–42, Feb. 2002.

[174] D. Van Compernolle. Switching Adaptive Filters for Enhancing Noisy and Reverberant Speech from Microphone Array Recordings. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 833–836, Albuquerque NM, USA, Apr. 1990.

[175] T. Van den Bogaert, S. Doclo, M. Moonen, and J. Wouters. Binaural cue preservation for hearing aids using an interaural transfer function multichannel Wiener filter. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 565–568, Honolulu HI, USA, Apr. 2007.

[176] T. Van den Bogaert, S. Doclo, M. Moonen, and J. Wouters. The effect of multi-microphone noise reduction systems on sound source localization in binaural hearing aids. *J. Acoust. Soc. Amer.*, 124(1):484–497, Jul. 2008.

[177] T. Van den Bogaert, S. Doclo, M. Moonen, and J. Wouters. Speech enhancement with multichannel Wiener filter techniques in multi-microphone binaural hearing aids. *J. Acoust. Soc. Amer.*, 125(1):360–371, Jan. 2009.

[178] T. Van den Bogaert, T. J. Klasen, M. Moonen, L. Van Deun, and J. Wouters. Horizontal localisation with bilateral hearing aids: without is better than with. *J. Acoust. Soc. Amer.*, 119(1):515–526, Jan. 2006.

[179] T. Van den Bogaert, J Wouters, T. J. Klasen, and M Moonen. Distortion of interaural time cues by directional noise reduction systems in modern digital hearing aids. In *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, pages 57–60, New Paltz, NY, Oct. 2005.

[180] S. Van Gerven and F. Xie. A Comparative Study of Speech Detection Methods. In *Proc. EUROSPEECH*, volume 3, pages 1095–1098, Rhodos, Greece, Sep. 1997.

[181] B. D. Van Veen and K. M. Buckley. Beamforming: A Versatile Approach to Spatial Filtering. *IEEE ASSP Mag.*, 5(2):4–24, Apr. 1988.

[182] A. E. Vandali, L. A. Whitford, K. L. Plant, and G. M. Clark. Speech perception as a function of electrical stimulation rate: Using the Nucleus 24 Cochlear implant system. *Ear Hear.*, 21(6):608–624, Dec. 2000.

[183] J. Vanden Berghe and J. Wouters. An adaptive noise canceller for hearing aids using two nearby microphones. *J. Acoust. Soc. Amer.*, 103(6):3621–3626, Jun. 1998.

[184] N. J. Versfeld, L. Daalder, J. M. Festen, and T. Houtgast. Method for the selection of sentence materials for efficient measurement of the speech reception threshold. *J. Acoust. Soc. Amer.*, 107(3):1671–1684, Mar. 2000.

[185] N. Virag. Single-channel speech enhancement based on masking properties of the human auditory system. *IEEE Trans. Speech Audio Process.*, 7(2):126–137, Mar. 1999.

[186] D. L. Wang. *On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis*, chapter in "Speech Separation by Humans and Machines" (Divenyi, P., Ed.), pages 181–197. Springer US, 2005.

[187] D. L. Wang. Time-Frequency masking for speech separation and its potential for hearing aid design. *Trends Amplif.*, 12(4):332–353, Dec. 2008.

[188] S. Wehr, R. Zourub, R. Aichner, and W. Kellermann. Post-processing for BSS algorithms to recover spatial cues. In *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Paris, France, Sep. 2006.

[189] M. R. Weiss, E. Aschkenasy, and T. W. Parsons. Processing speech signals to attenuate interference. In *Proc. IEEE Symp. Speech Recogn.*, pages 292–293, Pittsburgh, PA, USA, Apr. 1974.

[190] D. P. Welker, J. E. Greenberg, J. G. Desloge, and P. M. Zurek. Microphone-array hearing aids with binaural output–Part II: A two-microphone adaptive system. *IEEE Trans. Speech Audio Process.*, 5(6):543–551, Nov. 1997.

[191] B. Widrow, J. R. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Ziedler, E. J. Dong, and R. C. Goodlin. Adaptive noise cancellation: principles and applications. *Proc. IEEE*, 63(12):1692–1716, Dec. 1975.

[192] T. Wittkop and V. Hohmann. Strategy-selective noise reduction for binaural digital hearing aids. *Speech Commun.*, 39(1-2):111–138, Jan. 2003.

[193] W. S. Woods, I. Merks, T. Zhang, K. Fitz, and B. Edwards. Assessing the benefit of adaptive null-steering using real-world signals. *Int. J. Audiol.*, 49(6):434–443, Jun. 2010.

[194] P.M. Zurek. *Binaural advantages and directional effects in speech intelligibility*, chapter 15 in "Acoustical factors affecting hearing aid performance", pages 255–276. 2nd Ed., Boston, MA, Allyn and Bacon, 1992.

List of publications

International Journal Papers

1. B. Cornelis, S. Doclo, T. Van den Bogaert, M. Moonen, and J. Wouters "Theoretical analysis of binaural multi-microphone noise reduction techniques", *IEEE Trans. Audio Speech Lang. Process.*, 18(2):342-355, Feb. 2010

2. B. Cornelis, M. Moonen, and J. Wouters "Performance analysis of Multi-channel Wiener Filter based noise reduction in hearing aids under second-order statistics estimation errors", *IEEE Trans. Audio Speech Lang. Process.*, 19(5):1368-1381, Jul. 2011

3. B. Cornelis, M. Moonen, and J. Wouters "Reduced-bandwidth Multi-channel Wiener Filter based binaural noise reduction and localization cue preservation in binaural hearing aids", *submitted to IEEE Trans. Audio Speech Lang. Process.*, 2011

4. B. Cornelis, M. Moonen, and J. Wouters "Speech intelligibility improvements with hearing aids using bilateral and binaural adaptive Multi-channel Wiener Filtering based noise reduction", *submitted to J. Acoust. Soc. Amer.*, 2011

International Conference Papers

1. B. Cornelis, S. Doclo, T. Van den Bogaert, M. Moonen, and J. Wouters "Analysis of localization cue preservation by Multi-channel Wiener Filtering based binaural noise reduction in hearing aids", In *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008

2. B. Cornelis, M. Moonen, and J. Wouters "Comparison of frequency-domain noise reduction strategies based on Multi-channel Wiener Filtering and Spatial Prediction", In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 129-132, Tai Pei, Taiwan, Apr. 2009

3. B. Cornelis, M. Moonen, and J. Wouters "A QRD-RLS based frequency-domain Multi-channel Wiener Filter algorithm for noise reduction in hearing aids", In *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, pages 1953-1957, Aalborg, Denmark, Aug. 2010

4. B. Cornelis, M. Moonen, and J. Wouters "Binaural cue preservation in binaural hearing aids with reduced-bandwidth Multi-channel Wiener Filter based noise reduction", In *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Tel-Aviv, Israel, Sep. 2010

5. B. Cornelis, M. Moonen, and J. Wouters "A VAD-robust Multi-channel Wiener Filter algorithm for noise reduction in hearing aids", In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 281-284, Prague, Czech Republic, May 2011

6. B. Cornelis, M. Moonen, and J. Wouters "Binaural voice activity detection for MWF-based noise reduction in binaural hearing aids", *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, pages 486-490, Barcelona, Spain, Aug. 2011

International Conference Abstracts

1. B. Cornelis, M. Moonen, and J. Wouters "Speech intelligibility improvements in hearing aids using adaptive bilateral and binaural Multi-channel Wiener Filtering based noise reduction", *Int. Hearing Aid Research Conf. (IHCON)*, Lake Tahoe CA, USA, Aug. 2010

2. J. Wouters, B. Cornelis, T. Francart, A. Lenssen, and M. Moonen "Evaluation of hearing aid signal processing algorithms", *European Federation of Audiology Societies (EFAS) Congress*, Warsaw, Poland, Jun. 2011

National Journal Papers

1. A. Bertrand, B. Cornelis, J. Wouters, and M. Moonen "Hoe zegt u? - Beter spraakverstaan met binaurale hoorapparaten en akoestische sensornetwerken", *Het Ingenieursblad, Thema geluid en akoestiek*, 80:28-31, Sep. 2011

Curriculum Vitae



Bram Cornelis was born in Bornem, Belgium, in 1984. He received the M.Sc. degree in electrical engineering (magna cum laude) from the K.U. Leuven, Belgium, in 2007. Since then he has been a Ph.D. candidate at the SISTA Lab, Department of Electrical Engineering (ESAT), K.U. Leuven, and the ExpORL Lab, Department of Neurosciences, K.U. Leuven, under supervision of Prof. Marc Moonen and Prof. Jan Wouters.

He received a Ph.D. scholarship of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen) during 2008-2011, and was recipient of a Student Travel Grant Award to attend the 2010 International Hearing Aid Research Conference (IHCON), Lake Tahoe CA, USA. His research interests are in array signal processing, beamforming, sound localization, speech enhancement, voice activity detection and signal processing algorithms for hearing aids and cochlear implants.