KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Technology
Department of Electrical Engineering

Spatio-Temporal Speech Enhancement in Adverse Acoustic Conditions

Thomas Dietzen

Supervisor:
Prof. dr. ir. T. van Waterschoot
Co-supervisors:
Prof. dr. ir. M. Moonen
Prof. dr. ir. S. Doclo

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Technology (PhD)

September 2019

Spatio-Temporal Speech Enhancement in Adverse Acoustic Conditions

Thomas DIETZEN

Examination committee:
Prof. dr. ir. E. Demeester, chair
Prof. dr. ir. T. van Waterschoot, supervisor
Prof. dr. ir. M. Moonen, co-supervisor
Prof. dr. ir. S. Doclo, co-supervisor
(Carl von Ossietzky University of Oldenburg)
Prof. dr. ir. K. Eneman
Prof. dr. ir. D. Van Compernolle
Dr. ir. Ann Spriet
(NXP Semiconductors Belgium N.V.)
Prof. Dr.-Ing. Nilesh Madhu
(Ghent University)
Prof. Mads Græsbøll Christensen, Ph.D.
(Aalborg University)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Technology (PhD)

September 2019

© 2019 KU Leuven – Faculty of Engineering Technology
Uitgegeven in eigen beheer, Thomas Dietzen, Kasteelpark Arenberg 10, B-3001 Leuven (Belgium)

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

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

Preface

I don't care too much about football – and yet, it oddly relates the start of my PhD. About 20 minutes ahead of time, I arrived for a full-day job interview at NXP Leuven. Ann came down to the reception desk, welcomed me and said I can wait in the meeting room where we would start with the presentation. Clueless as I was, I asked her what the presentation would be about. "Well, *you* have to give the presentation..." she responded. Something went wrong when they sent me the agenda, I thought. So I looked up some old slides on my laptop, skimmed through in the time I had left, and there we go. Later that day, Germany won 7–1 against Brazil in the semi-final of the world cup. I'm not sure if my talk was any good, but Ann and Toon decided to hire me. When Germany won the final, I had a new job. I'm still grateful for that, and will always be.

Throughout my PhD, I have worked in two different countries, at four different institutions, in seven different offices,[1] on nine different desks. A highly dynamic job indeed. During those years, I could rely on my closest ones and came to know many, in and out of office. They all have influenced my research and life on the whole, and thus have their share in this work.

First and foremost, I wish to thank my DREAMS supervisors. Most notably I thank Toon, an imposing character and the in any aspect best promoter I could think of, personally, scientifically, and even financially. I thank Ann and Wouter, who guided me through the start of my PhD when I knew even less than today. Thanks to Simon, who seems to know just about any piece of speech enhancement literature out there, and Marc, who continuously proofs to have an eagle eye on everything. Thanks for your time, fruitful discussions, advice, hints, and reviews. I kindly thank the further members of the examination committee, Koen, Dirk, Nilesh, Mads, as well as the chair, Eric, for thoroughly

[1] the NXP office, the old Building's office, the big office, the Oldenburg office, the Italian office, the Group T office, the big office again, the tower office; with various levels of reverberation, interfering speech, and background noise.

reading the manuscript, their questions, suggestions, and examination. Thanks to the various funding institutions and tax payers who got me through, cf. the Acknowledgements.

Thanks and greetings go to former and current colleagues and friends – Adam, Adel, Ahmed, Alam (for royal nights), Aldona (for hundreds of coffee breaks and moving the cabinet), Alexander, Ana, Andres, Anisa, Ante, Amin, Bahar, Bas, Benjamin, Brian, Bruno, Carlo, Clément, Dan, Daniel, Daryna, David, Denis, Deniz, Dörte, Dominik, Duowei, Ece, Edu, Elisa, Enzo, Fernando, Filip, Filippos, Florian, Francesco, Gert, Giacomo, Gigo, Giuliano, Herr Groß, Hanne, Haxhi, Heleen, Henning, Ina, Jared, Jasper, Jeroen, Johan, John, Jonas, Julian, Julien, Kai, Karen, Kristina K., Leo (for Stroh), Lisa, Maarten J. H. (for the tenor), Maarten T., Magda, Maja (for Dancing Queen), María, Marco (for Verviers progressive rock), Martin, Mathieu, Mina, Naveen, Mohit, Neetha, Nele, Neo, Niccolò, Nico, Oreste, Pablo, Pe, Procrustes (for leg adjustments), Randy (for wild rides in Xi'an, Skateboard & Karaoke in Kyoto), Rich, Robbe, Rolles, Rudi, Santiago, Dr Skorpio, Silvia, Simon V. E., Taewoong, Wouter B., Wouter L., Xueru, and all those that I forgot. Further greetings go to the local bar scenes – 1000Fryd, Auszeit, Blauwe Kater (for Blue Mondays), Booze'n Blues, Boulevard Cafeen, Brasserie, Café Belge, Capri Bar, Fiere Magrit, Fleur de Bière (RIP), Gasthaus Bingert, Giraf, Janet's Bar, Café De Libertad, Metafoor, Nilles, Heartbreak Hotel, the one Mono, the other Mono (for Die Internationale), Bar Del Sol, Tante Anna, Café Wienerhof, and all the folks I have met there.

Besonderer Dank gilt meiner Familie, zuerst meinen Eltern, Marianne und Günter, deren Unterstützung ich mir stets und in all meinen Vorhaben sicher sein konnte, wenn meine Arbeit ihnen auch ein Böhmisches Dorf gewesen sein mag. Ebenso meinen Geschwistern, Corinna, Philipp, Matthias – und insbesondere Stephan und (sister-in-law) Dorota für jedwede Hilfe, wochenlange freie Kost und Logis, sowie schöne und absurde belgische Zeiten. Dank an Ekkehard und Hans, die immer an mich dachten. Danke der Familie Hilpert, Anne, Rüdiger und Alex, unter deren Saarbrücker Dach der größte Teil der Einleitung entstand, als es im Juli unter unserem Leuvener Dach auf die 40° C zuging. Danke allen, die uns rund um die Geburt unseres Sohnes Leander und dessen erste Monate begleitet haben – und nicht zuletzt Leander selbst, who couldn't care less about my PhD, und dessen unvergleichliches Lächeln mich so viele Male zurück auf die Spieldecke der Tatsachen geholt hat.

Und *Kristina*, dir verdanke ich *alles*.

Thomas Dietzen
September 2019

Abstract

Never before has speech been captured as often by electronic devices equipped with one or multiple microphones, serving a variety of applications. It is the key aspect in digital telephony, hearing devices, and voice-driven human-to-machine interaction.

When speech is recorded, the microphones also capture a variety of further, undesired sound components due to *adverse acoustic conditions*. Interfering speech, background noise and reverberation, i.e. the persistence of sound in a room after excitation caused by a multitude of reflections on the room enclosure, are detrimental to the quality and intelligibility of target speech as well as the performance of automatic speech recognition. Hence, *speech enhancement* aiming at estimating the early target-speech component, which contains the direct component and early reflections, is crucial to nearly all speech-related applications presently available.

In this thesis, we compare, propose and evaluate existing and novel approaches in the fields of speech enhancement. At this, we take account of the following technical aspects, which guide the design of the proposed approaches. First, we envisage comprehensive speech enhancement in all varieties of adverse acoustic conditions, which requires dereverberation, interfering speech cancellation and noise reduction. Second, we aim to exploit *spatial* and *temporal* knowledge on the target-speech direction and statistics in form of relative early transfer function (RETF) and time-varying early power-spectral-density (PSD) estimates, and further to acquire such knowledge in a multi-source scenario. Third, we strive for online processing in dynamic acoustic scenarios, and fourth, for moderate computational complexity.

The thesis is introduced by a problem description and a thorough overview on the state of the art. Major parts of the remainder relate to two broad concepts in multi-microphone speech enhancement, namely beamforming and blind deconvolution, specifically by means of the generalized sidelobe canceler (GSC)

and multi-channel linear prediction (MCLP), respectively. While beamforming is a well-established approach to interfering speech cancellation and noise reduction, MCLP may be said to be the presently most popular approach to dereverberation.

As a preparatory step towards comprehensive speech enhancement, we analyze and compare the GSC and MCLP architecture in terms of their potential for dereverberation and noise reduction. They mainly differ in their data-dependent filter path, i.e. the sidelobe cancellation (SC) and the linear prediction (LP) filter paths, which entail spatial and temporal pre-processing by means of a blocking matrix (BM) and a delay, respectively. We show that in case of perfect spatial knowledge, the GSC reaches the same dereverberation performance as MCLP, while obviously performing noise reduction in addition, as opposed to MCLP. In case of deficient spatial knowledge, however, the GSC performs worse than MCLP in terms of dereverberation.

Based on this comparison and the recently common usage of MCLP-and-beamforming cascades, we propose to integrate the GSC and MCLP into a novel architecture referred to as integrated sidelobe cancellation and linear prediction (ISCLP), where the SC filter and the LP filter operate in parallel. We propose to estimate the SC and LP filters jointly and online by means of a single Kalman filter. We further propose a spectral Wiener gain post-processor, relating to the Kalman filter's posterior state estimate. While being computationally less demanding than two state-of-the-art approaches, the ISCLP Kalman filter is shown to perform similar or better in various adverse acoustic conditions.

The ISCLP Kalman filter exploits spatial and temporal target-parameter knowledge to be acquired in a multi-source scenario. To this end, we propose an appropriate online estimation approach, namely square root-based multi-source early PSD estimation and RETF updating. Here, as opposed to the conventional approach, we propose to factorize the early correlation matrix and minimize the approximation error defined with respect to the early-correlation-matrix square root. From the proposed minimization problem, we iteratively obtain estimates of a unitary matrix and the early PSD square roots, which further allow to recursively update the RETF estimate. Evaluation indicates better performance as compared to the conventional approach and convergence in only one iteration.

The ISCLP Kalman filter exhibits a quadratic computational complexity in the number of filter coefficients and the number of channels. We therefore propose low-complexity variants of the ISCLP Kalman filter. The low-complexity variants are obtained by enforcing the state estimation error correlation matrix to assume sparse structures corresponding to the negligence of either temporal, spatial, or all cross-correlations, leading to linear cost in either or both the

number of filter coefficients and the number of channels. The low-complexity ISCLP Kalman filter variants are shown to perform nearly as well as the original variant, thereby permitting far more favourable trade-offs between complexity and performance.

The thesis is concluded by a summary, suggestions for future research and a discussion on industrial valorization.

Korte Inhoud

Nooit eerder werd spraak zo vaak gecapteerd door elektronische toestellen, uitgerust met één of meerdere microfoons, voor uiteenlopende toepassingen. Dit vormt een cruciaal aspect van digitale telefonie, hoortoestellen en stemgestuurde mens-machine-interactie.

Wanneer spraak wordt opgenomen, capteren de microfoons ook verschillende andere, ongewenste geluidscomponenten ten gevolge van *akoestisch ongunstige eigenschappen van de omgeving*. Storende spraak, achtergrondruis en reverberatie, i.e. het nagalmen van geluid in een ruimte veroorzaakt door een groot aantal reflecties tegen de begrenzingen van die ruimte, hebben een nefaste invloed op de spraakkwaliteit en -verstaanbaarheid alsook op de performantie van automatische spraakherkenning. Bijgevolg is *spraakverbetering*, met het oog op het schatten van de zogenaamde vroege spraakcomponent, die naast de directe component ook vroege reflecties bevat, cruciaal in nagenoeg alle spraakgerelateerde toepassingen die vandaag de dag beschikbaar zijn.

In dit proefschrift worden bestaande en nieuwe methodes voor spraakverbetering vergeleken, voorgesteld en geëvalueerd. Hierbij nemen we de volgende technische aspecten in acht, die richting geven aan het ontwerp van de voorgestelde methodes. Ten eerste beogen we een omvattende spraakverbetering in alle mogelijke akoestisch ongunstige omstandigheden, wat dereverberatie alsook onderdrukking van storende spraak en ruis vereist. Ten tweede betrachten we *spatiale* en *temporele* kennis over de richting en statistieken van de gewenste spraakbron aan te wenden in de vorm van schattingen van de relatieve vroege transferfunctie (RETF) en de tijdsvariante vroege spectrale vermogendichtheid (PSD), en die kennis te verwerven in omgevingen met meerdere geluidsbronnen. Ten derde streven we naar onlineverwerking in dynamische akoestische omgevingen en ten vierde ook naar een matige rekencomplexiteit.

Dit proefschrift start met een probleembeschrijving en een diepgaand overzicht

van de state of the art. Grote delen van het verdere proefschrift houden verband met twee concepten uit meermicrofoonspraakverbetering, met name bundelsturing en blinde deconvolutie, specifiek door middel van respectievelijk de veralgemeende zijlob-onderdrukker (GSC) en meerkanaals lineaire predictie (MCLP). Waar bundelsturing een gevestigde aanpak is om storende spraak en ruis te onderdrukken, kan over MCLP gezegd worden dat het momenteel de meest populaire aanpak is voor dereverberatie.

Als voorbereidende stap naar omvattende spraakverbetering, analyseren en vergelijken we de GSC- en MCLP-architecturen in termen van hun potentieel voor dereverberatie en ruisonderdrukking. Ze verschillen hoofdzakelijk in hun data-afhankelijke filterpad, i.e. de filterpaden voor zijlob-onderdrukking (SC) en lineaire predictie (LP), die een spatiale en temporele voorverwerking uitvoeren door middel van respectievelijk een blokkeermatrix (BM) en een vertraging. We tonen aan dat, in het geval van perfecte spatiale kennis, de GSC dezelfde performantie inzake dereverberatie bereikt als MCLP en daarenboven uiteraard ook ruisonderdrukking uitvoert, in tegenstelling tot MCLP. In het geval van ontoereikende spatiale kennis presteert de GSC daarentegen minder goed dan MCLP inzake dereverberatie.

Op basis van deze vergelijking en gemotiveerd door het recente gebruik van serieschakelingen van MCLP en bundelsturing, stellen we voor om de GSC en MCLP te integreren in een nieuwe architectuur die we benoemen als geïntegreerde zijlob-onderdrukking en lineaire predictie (ISCLP), waarin het SC-filter en het LP-filter in parallel werken. We stellen voor om de SC- en LP-filters gezamenlijk en online te schatten door middel van een enkel Kalmanfilter. Verder stellen we een spectrale nabewerking met een Wienerversterking voor, die in verband staat met de a posteriori toestandsschatting van het Kalmanfilter. Terwijl het minder rekenkracht vraagt dan twee state-of-the-art-methodes, vertoont het ISCLP-Kalmanfilter een gelijkaardige of verbeterde performantie in diverse akoestisch ongunstige omgevingen.

Het ISCLP-Kalmanfilter maakt gebruik van spatiale en temporele kennis van parameters die gerelateerd zijn aan de beoogde spraakbron. Teneinde deze kennis te verwerven in een omgeving met meerdere bronnen, stellen we een gepaste methode voor onlineschatting voor die bestaat in een vierkantswortelgebaseerde vroege PSD-schatting voor meerdere bronnen en een RETF-updating. In tegenstelling tot de conventionele aanpak, stellen we hier voor om de vroege-correlatiematrix te factoriseren en de te minimaliseren benaderingsfout te definiëren met betrekking tot de vierkantswortel van de vroege-correlatiematrix. Uit het voorgestelde minimalisatieprobleem bekomen we op een iteratieve manier schattingen van een unitaire matrix en de vierkantswortels van de vroege PSDs, die ons toelaten om vervolgens de RETF-schatting recursief te updaten. Evaluatie toont een betere performantie in vergelijking met de conventionele

aanpak en convergentie in een enkele iteratie.

Het ISCLP-Kalmanfilter vertoont een kwadratische rekencomplexiteit in het aantal filtercoëfficiënten en het aantal kanalen. Om die reden stellen we lage-complexiteitvarianten van het ISCLP-Kalmanfilter voor. De lage-complexiteitvarianten worden bekomen door een spaarse structuur op te leggen aan de correlatiematrix van de toestandsschattingsfout, waardoor een lineaire complexiteit wordt bekomen in ofwel het aantal filtercoëfficiënten, ofwel het aantal kanalen, ofwel beide, naargelang enkel de temporele, enkel de spatiale, of alle kruiscorrelaties worden verwaarloosd. We tonen aan dat de lage-complexiteitvarianten van het ISCLP-Kalmanfilter bijna even goed presteren als de originele variant, waardoor veel gunstigere trade-offs van complexiteit en performantie mogelijk worden.

Het proefschrift wordt afgesloten met een samenvatting, suggesties voor toekomstig onderzoek en een discussie rond industriële valorisatie.

Glossary

Abbreviations and Acronyms

AEC	acoustic echo cancellation
AFC	acoustic feedback control
ASR	automatic speech recognition
BM	blocking matrix
BSI	blind system identification
BSS	blind source separation
CAGR	compound annual growth rate
CD	cepstral distance
cf.	*confere*, compare with, see also
Ch.	Chapter
conv.	convolutive
DoA	direction of arrival
e.g.	*exempli gratia*, for example
FD	frequency domain
Fig.	Figure
FSB	filter-and-sum beamformer
fws	frequency-weighted segmental

GEVD	generalized eigenvalue decomposition
GSC	generalized sidelobe canceler, generalized sidelobe cancellation
i.e.	*id est*, that is
ILD	interaural level difference
IR	impulse response
ISCLP	integrated sidelobe cancellation and linear prediction
ITD	interaural time differences
LP	linear prediction
LS	least squares
MCLP	multi-channel linear prediction
MF	matched filter
MINT	multiple input/output inverse theorem
MIPS	million instructions per second
MPDR	minimum-power distortionless response
MUSHRA	multi-stimulus tests with hidden reference and anchor
MVDR	minimum-variance distortionless response
OEM	original equipment manufacturer
PBFD	partitioned-block frequency domain
PESQ	perceptual evaluation of speech quality
PSD	power spectral density
RETF	relative early transfer function
RIR	room impulse response
RLS	recursive least squares
RTF	room transfer function
SC	sidelobe cancellation

SDR	Ch. 1 – signal-to-diffuse ratio Ch. 4 – signal-to-distortion ratio
Sec.	Section
SAR	signal-to-artifacts ratio
SIR	signal-to-interference ratio
SNR	signal-to-noise ratio
SNRR	signal-to-noise-plus-reverberation ratio
SRR	signal-to-reverberation ratio
SRT	speech reception threshold
STOI	short-time objective intelligibility
STFT	short-time Fourier transform
SVD	singular value decomposition
TD	time domain
TFD	time-frequency domain
VoIP	voice over internet protocol
WPE	weighted prediction error

Mathematical Notation

$\mathbb{R}$	the set of real numbers
$\mathbb{C}$	the set of complex numbers
$\mathbf{A}$ (bold upper case)	a matrix
$\mathbf{a}$ (bold lower case)	a vector
A (upper case)	a positive integer scalar, an ordered set
a (lower case)	a scalar
$[\mathbf{A}]_{i,:}$	the vector defined by row i of $\mathbf{A}$
$[\mathbf{A}]_{i,:}$	the vector defined by column j of $\mathbf{A}$

$[\mathbf{A}]_{i,j}$	the element at row i and column j of $\mathbf{A}$
$[\mathbf{a}]_i$	the i^{th} element of $\mathbf{a}$
$[\mathbf{A}]_{i_1:i_2,:}$	the submatrix spanning rows i_1 to i_2 of $\mathbf{A}$, and similarly for columns
$[\mathbf{A}]_{\in S,:}$	the submatrix composed of the rows of $\mathbf{A}$ with indices in ordered set S, and similarly for columns
$\langle\mathbf{a}\rangle_i$	a vector obtained from $\mathbf{a}$ by nullifying all except the i^{th} element
$\langle\mathbf{a}\rangle_{\in S}$	a vector obtained from $\mathbf{a}$ by nullifying all except the elements in ordered set S
$\text{Diag}[\mathbf{a}]$	a diagonal matrix with the elements of $\mathbf{a}$ on its diagonal
$\text{diag}[\mathbf{A}]$	a column vector composed of the diagonal elements of $\mathbf{A}$
$\text{Diagg}[\mathbf{A}]$	$\text{Diag}\big[\text{diag}[\mathbf{A}]\big]$
$\text{Blkdiag}[\mathbf{A}_1,\ldots,\mathbf{A}_N]$	a block-diagonal matrix with $\mathbf{A}_1,\ldots,\mathbf{A}_N$ on its diagonal
$\text{Tplz}[\mathbf{a},L]$	a Toeplitz matrix of L columns with the first column defined by $\begin{pmatrix}\mathbf{a}^T & \mathbf{0}^{1\times(L-1)}\end{pmatrix}^T$
$\Re[a]$	the real part of a
$\Im[a]$	the imaginary part of a
$\mathbf{A}^*$	the complex conjugate of $\mathbf{A}$
$\mathbf{A}^T$	the transpose conjugate of $\mathbf{A}$
$\mathbf{A}^H$	the complex conjugate transpose or Hermitian of $\mathbf{A}$
$\mathbf{A}^{-1}$	the inverse of $\mathbf{A}$
$\mathbf{A}^P$	the pseudoinverse of $\mathbf{A}$
$\lvert\mathbf{a}\rvert$	the element-wise applied absolute value of $\mathbf{a}$
$\sqrt{\mathbf{a}}$	the element-wise non-negative square root of $\mathbf{a}$
$\mathbf{a}^{1/2}$	a complex vector satisfying $\text{Diag}[\mathbf{a}^{H/2}]\mathbf{a}^{1/2}=\mathbf{a}$

$\max[\mathbf{a}_1, \mathbf{a}_2]$	the element-wise maximum of $\mathbf{a}_1$ and $\mathbf{a}_2$
$\mathrm{E}[\mathbf{A}]$	the expected value of $\mathbf{A}$
$\mathrm{tr}[\mathbf{A}]$	the trace of $\mathbf{A}$
$\|\mathbf{A}\|_F$	the Frobenius norm of $\mathbf{A}$
$\|\mathbf{a}\|_2$	the Euclidian norm of $\mathbf{a}$
$\hat{\mathbf{A}}$	an estimate of $\mathbf{A}$
$\arg\min_{\mathbf{A}} f(\mathbf{A})$	the argument of the minimum of the function $f(\mathbf{A})$ over $\mathbf{A}$
s. t.	subject to
$:=$	simplification, with the expression on the left-hand side replaced by the computationally less expensive expression on the right-hand side
$\mathbf{I}$	an identity matrix, dimensions optionally indicated by superscript
$\mathbf{0}$	a zero matrix or zero vector, dimensions optionally indicated by superscript
$\mathbf{1}$	a vector of ones
$\mathbf{i}$	$[\mathbf{I}]_{:,1}$

Units

°	arc degree
dB	dezibel
Hz	Hertz
m	meter
s	second

Contents

Abstract iii

Korte Inhoud vii

Glossary xi

Contents xvii

List of Figures xxiii

List of Tables xxix

1 Introduction 1

1.1 Adverse Acoustic Conditions 2

1.1.1 Reverberation 4

1.1.2 Interfering Speech and Background Noise 5

1.2 Speech Enhancement 7

1.2.1 Applications and Challenges 7

1.2.2 State of the Art 10

1.2.2.1 Spatio-Temporal Speech Enhancement 10

1.2.2.2 Target-Parameter Estimation 21

1.3 Contributions 21
1.4 Assumptions 24
1.5 Outline of the Thesis 25

I Underlying Architectures 27

2 Comparative Analysis of the GSC and MCLP 28

2.1 Introduction 30
2.2 Signal Model 33
2.2.1 Acoustic Scenario 34
2.2.2 Speech Reference Signal 36
2.2.3 MCLP Filter Input 37
2.2.4 GSC Filter Input 38
2.2.5 Enhanced signal 39
2.3 Filter Estimation 39
2.3.1 Pre-whitened LS 40
2.3.2 Choice of pre-whitening matrix 41
2.3.3 Convergence to Wiener filter solution 42
2.4 MCLP Analysis 42
2.4.1 MCLP Filter Output 42
2.4.2 MCLP Enhancement 43
2.4.2.1 Absence of Incoherent Noise 43
2.4.2.2 Presence of Incoherent Noise 45
2.5 GSC Analysis 45
2.5.1 GSC Filter Output 45
2.5.2 GSC Enhancement 46
2.5.2.1 Absence of Incoherent Noise 46

2.5.2.2 Presence of Incoherent Noise 47
2.6 Comparative Summary 47
2.7 Simulations 49
2.7.1 Simulation Setup 49
2.7.1.1 Acoustic Scenario 49
2.7.1.2 Source Signals 50
2.7.1.3 MCLP and GSC implementation 51
2.7.1.4 Performance Measures 51
2.7.1.5 Varied Parameters 52
2.7.2 Simulation Results 53
2.7.2.1 Time Domain 53
2.7.2.2 STFT Domain 56
2.8 Conclusion 58

II Integrated Architecture and Parameter Estimation 61

3 The ISCLP Kalman Filter 62

3.1 Introduction 64
3.2 Signal Model 66
3.3 Integrated Sidelobe Cancellation and Linear Prediction Kalman Filter 69
3.3.1 ISCLP Signal Path Architecture 69
3.3.2 ISCLP State-Space Model and Kalman Filter Update . 72
3.3.3 Posterior-like Spectral Post-Processing 74
3.4 Implementational Aspects 75
3.4.1 Spatio-Temporal Target Component Leakage 75
3.4.2 Target PSD Estimation and RETF Update 77

3.4.3 Process Equation Parameter Tuning and Initialization . 78
3.5 Simulations 80
3.5.1 Reference Algorithms 80
3.5.1.1 Case A: Alternating Kalman Filters 80
3.5.1.2 Case B: MCLP+GSC Kalman Filter Cascade . 81
3.5.2 Performance Measures 82
3.5.3 Acoustic Scenario 82
3.5.3.1 Case A: Without Interfering Speech 83
3.5.3.2 Case B: With Interfering Speech 83
3.5.4 Algorithmic Settings 83
3.5.5 Results 84
3.5.5.1 Case A 84
3.5.5.2 Case B 89
3.6 Conclusion 91

4 Multi-Source Early PSD Estimation and RETF Update 93

4.1 Introduction 95
4.2 Signal Model 97
4.3 Early PSD Estimation based on the Early Correlation Matrix . 100
4.4 Early PSD Estimation and Recursive RETF Update based on the Early-Correlation-Matrix Square Root 103
4.4.1 Early-Correlation-Matrix Factorization 103
4.4.2 Orthogonal Procrustes-based Early PSD Estimate . . . 104
4.4.3 Recursive RETF Update 107
4.5 Subspace-based Early Correlation Matrix Estimation 109
4.5.1 Correlation Matrix Subspace Decomposition 109
4.5.2 Recursive Correlation Matrix Estimation and Desmoothing 110

4.5.3 Early Correlation Matrix Estimation and Factorization 111
4.6 Simulations . . . 112
4.6.1 Model-based Data . . . 113
4.6.1.1 Performance Measures . . . 113
4.6.1.2 Data Generation . . . 114
4.6.1.3 Algorithmic Settings . . . 115
4.6.1.4 Results . . . 115
4.6.2 Acoustic Data . . . 119
4.6.2.1 Performance Measures . . . 119
4.6.2.2 Acoustic Scenario . . . 120
4.6.2.3 Algorithmic Settings . . . 121
4.6.2.4 Results . . . 124
4.7 Conclusion . . . 126

III Complexity Reduction 127

5 Low-Comlexity ISCLP Kalman Filters 128

5.1 Introduction . . . 129
5.2 Review of the ISCLP Kalman Filter . . . 130
5.2.1 Signal Model . . . 131
5.2.2 The ISCLP Kalman Filter . . . 132
5.3 Complexity Reduction . . . 134
5.3.1 Complexity of the Original ISCLP Kalman Filter . . . 135
5.3.2 Complexity Reduction by Cross-Correlation Negligence 138
5.3.3 Equivalent Multiple Kalman Filters Formulation . . . 140
5.4 Simulations . . . 141
5.4.1 Setup . . . 142

5.4.2 Results 142
5.4.2.1 Case A: Without Interfering Speech 142
5.4.2.2 Case B: With Interfering Speech 147
5.4.2.3 Summary 149
5.5 Conclusion 150

6 Conclusion **151**
6.1 Summary 152
6.2 Suggestions for Future Research 156
6.3 Industrial Valorization 159

A Appendix to Chapter 2 **161**
A.1 GSC Enhancement in Absence of Incoherent Noise – Reformulation 161
A.1.1 GSC Filter Output 161
A.1.2 GSC Bias 163
A.2 The Pesudoinverse of Block-Diagonal Matrices 163

B Appendix to Chapter 4 **164**
B.1 The Orthogonal Procrustes Problem – Reformulation 164
B.2 The Orthogonal Procrustes Problem – Solution 165

Bibliography **169**

Curriculum Vitae **189**

List of Publications **191**

List of Figures

1.1 Adverse acoustic conditions degrading the quality and intelligibility of target speech – the listener has to cope with (a) interfering speech, background noise, and (b) late reverberation, caused by a multitude of late reflections. 3

1.2 Exemplary spectrograms illustrating (a) the early target-speech source image, (b) the reverberant target speech, (c) interfering speech and babble noise, and (d) the superposition of the signals in (b) and (c). 6

2.1 The MCLP framework employing the prediction delay δ in the data-dependent filter path. 31

2.2 The GSC framework employing the blocking matrix $\mathbf{B}$ in the data-dependent filter path. 32

2.3 Schematic of the (sub-band) IR relating $s_n(l)$ and $q_{s_n}(l)$, separated in early part $\mathbf{C}_e\mathbf{H}_n\mathbf{g}$ [——] applied to $\mathbf{s}_n(l)$ and late part $\mathbf{C}_{d|\ell}\mathbf{H}_n\mathbf{g}$ [·····] applied to $\mathbf{s}_n(l-d)$. 38

2.4 Schematic of the correlation matrices $\mathbf{\Psi}_{\tilde{s}_n}$ and $\mathbf{\Psi}_{\tilde{s}_n|d}$ as different submatrices of a larger correlation matrix. 43

2.5 Schematic of the (sub-band) IR relating $s_n(l)$ and $e_{s_n}(l)$, separated in early part $\mathbf{C}_e\mathbf{H}_n\mathbf{g}$ [——] applied to $\mathbf{s}_n(l)$ and bias part $-\boldsymbol{\vartheta}_{n|d}$ [·····] applied to $\mathbf{s}_n(l-d)$. 45

2.6 Dereverberation/noise reduction performance $\Delta SRR/\Delta SNR^{tot}$ versus (a)/(e) SNR_q^{coh}, (b)/(f) SNR_q^{inc}, (c)/(g) L_w^{rel} and (d)/(h) M for colored and white source signals of the MCLP framework, respectively denoted by [——] and [·····], and the GSC framework, respectively denoted by [——] and [·····]. The vertical grid lines indicate the intersection point of the individual subplots. The shaded areas represent the standard deviation. 54

2.7 (a) dereverberation-only/(b) dereverberation-plus-noise-reduction performance in terms of $SRR^{fws}/SNRR^{fws}$ versus SNR_y^{coh} of the MF, the MCLP and the GSC framework, respectively denoted by [-·-·-], [——], and [——]. The shaded areas represent the standard deviation. 57

3.1 The integrated sidelobe cancellation and linear prediction (ISCLP) architecture. 71

3.2 Exemplary spectrograms depicting 2 s of (a) the reference microphone signal $y_1(l)$, and the corresponding outputs of (b) the original alternating Kalman filters, (c) the modified alternating Kalman filters, and (d) the ISCLP Kalman filter for $L = 6$ at $SNR = 10\,\text{dB}$. 85

3.3 (a) $PESQ$, (b) $STOI$, (c) SIR^{fws}, and (d) CD versus SNR for the reference microphone signal [·····], the original alternating Kalman filters [- - -], the modified alternating Kalman filters [–▪–], and the ISCLP Kalman filter [–•–] with $L = 6$ if interfering speech is absent. 87

3.4 (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus L with respect to the reference microphone signal for the original alternating Kalman filters [- - -], the modified alternating Kalman filters [–▪–], and the ISCLP Kalman filter [–•–] at $SNR = 25\,\text{dB}$ if interfering speech is absent. 88

3.5 (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus t with respect to the reference microphone signal for the original alternating Kalman filters [- - -], the modified alternating Kalman filters [–▪–], and the ISCLP Kalman filter [–•–] with $L = 6$ at $SNR = 10\,\text{dB}$ if interfering speech is absent. 89

3.6 (a) *PESQ*, (b) *STOI*, (c) SIR^{fws}, and (d) *CD* versus *SNR* for the reference microphone signal [·····], the MCLP+GSC Kalman filter cascade [–▾–] and the ISCLP Kalman filter [–•–] with $L = 6$ if interfering speech is present. 90

3.7 (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus L with respect to the reference microphone signal for the MCLP+GSC Kalman filter cascade [–▾–] and the ISCLP Kalman filter [–•–] at $SNR = 25\,\text{dB}$ if interfering speech is present. . . . 91

4.1 ε_{φ_s} versus ε_H for conventional MP [–•–] and square-root MP [——] with $\alpha = 10^3$ at $f = 2\,\text{kHz}$. 114

4.2 ε_{φ_s} versus α for conventional MP [–•–] and square-root MP [——] at $\varepsilon_H = -10\,\text{dB}$ and $f = 2\,\text{kHz}$. 114

4.3 ε_{φ_s} versus f for conventional MP [–•–] and square-root MP [——] with $\alpha = 10^3$ at (a) $\varepsilon_H = 0\,\text{dB}$, (b) $\varepsilon_H = -10\,\text{dB}$, and (c) $\varepsilon_H = -20\,\text{dB}$. The graphs denoted by [·····] correspond to $10\log_{10}|\mathbf{h}_n^H\mathbf{h}_{n'}|/M\,\text{dB}$ for $n' \neq n$. 116

4.4 $\varepsilon_{\varphi_s}^{(i)}$ versus ε_H and i for square-root MP with $\alpha = 10^3$ and $\hat{\boldsymbol{\varphi}}_s^{1/2|(0)}$ based upon (a) the sum constraint in (4.13) and (b) the estimator in (4.17)–(4.18) at $f = 2\,\text{kHz}$. 117

4.5 (a) $\varepsilon_H(l+r)$ and (b) $\varepsilon_{\varphi_s}(l+r)$ versus r for conventional MP [–•–] and square-root MP [——] with $\alpha = 10^3$ at $f = 2\,\text{kHz}$ and $\varepsilon_H(l) = 0\,\text{dB}$ if $\mathbf{H}$ changes at $r = 32$ and remains constant otherwise. 118

4.6 Exemplary spectrograms depicting (a.n) estimates $\hat{\varphi}_{s_n}$, $n = 1, 2$, and their decomposition according to Sec. 4.6.2.1, i.e. (b.n) the correctly estimated components $\bar{\varphi}_{s_n}$, (c.n) the interference components $e_{\varphi_{s_n}}^{2|int}$, and (d.$n$) the artifact components $e_{\varphi_{s_n}}^{2|art}$. The reference PSDs $\tilde{\varphi}_{s_1}$ and $\tilde{\varphi}_{s_2}$ originate from a female and a male speaker at $-30°$ and $60°$, respectively, and the estimates $\hat{\varphi}_{s_n}$ are obtained by means of the square-root MP. 123

4.7 (a) *SIR*, (b) *SAR*, and (c) *SDR* in third-octave bands for conventional MP [■], square-root MP without recursive RETF update [▩], and square-root MP with recursive RETF update [▨]. 125

5.1 The (sparse) structure of $\mathbf{\Psi}_{\tilde{w}}(l)$ and the order of non-zero elements [■] for (a) preserved cross-correlations, (b) neglected temporal cross-correlations, (c) neglected spatial cross-correlations, (d) all cross-correlations neglected, exemplarily shown for $N_T = 1$, $M = 3$, and $L = 3$. 139

5.2 (a) *PESQ*, (b) *STOI*, (c) SIR^{fws}, and (d) *CD* versus *SNR* for the reference microphone signal [·····], the $\mathcal{O}(L^2M^2)$-cost [–•–], the $\mathcal{O}(LM^2)$-cost [- ▪ -], the $\mathcal{O}(L^2M)$-cost [- ♦ -], and the $\mathcal{O}(LM)$-cost [- ▾ -] ISCLP Kalman filter with $L = 6$ if interfering speech is absent. 143

5.3 (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus L with respect to the reference microphone signal for the $\mathcal{O}(L^2M^2)$-cost [–•–], the $\mathcal{O}(LM^2)$-cost [- ▪ -], the $\mathcal{O}(L^2M)$-cost [- ♦ -], and the $\mathcal{O}(LM)$-cost [- ▾ -] ISCLP Kalman filter at $SNR = 25\,\text{dB}$ if interfering speech is absent. 144

5.4 (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus L with respect to the reference microphone signal for the $\mathcal{O}(L^2M^2)$-cost [–•–], the $\mathcal{O}(LM^2)$-cost [- ▪ -], the $\mathcal{O}(L^2M)$-cost [- ♦ -], and the $\mathcal{O}(LM)$-cost [- ▾ -] ISCLP Kalman filter with $\bar{\boldsymbol{\psi}}_{w_{SC}} = \mathbf{0}$ if noise and interfering speech are absent. The reference microphone signal scores at $PESQ = 1.97$, $STOI = 0.89$, $SIR^{fws} = 9.36\,\text{dB}$, and $CD = 3.82\,\text{dB}$. 145

5.5 (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus t with respect to the reference microphone signal for the $\mathcal{O}(L^2M^2)$-cost [–•–], the $\mathcal{O}(LM^2)$-cost [- ▪ -], the $\mathcal{O}(L^2M)$-cost [- ♦ -], and the $\mathcal{O}(LM)$-cost [- ▾ -] ISCLP Kalman filter with $L = 6$ at $SNR = 10\,\text{dB}$ if interfering speech is absent. 146

5.6 (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus t with respect to the reference microphone signal for the $\mathcal{O}(L^2M^2)$-cost [–•–], the $\mathcal{O}(LM^2)$-cost [- ▪ -], the $\mathcal{O}(L^2M)$-cost [- ♦ -], and the $\mathcal{O}(LM)$-cost [- ▾ -] ISCLP Kalman filter with $L = 6$ and $\bar{\boldsymbol{\psi}}_{w_{SC}} = \mathbf{0}$ if noise and interfering speech are absent. The reference microphone signal scores at $PESQ = 1.99$, $STOI = 0.91$, $SIR^{fws} = 10.2\,\text{dB}$, and $CD = 3.62\,\text{dB}$. 147

5.7 (a) *PESQ*, (b) *STOI*, (c) SIR^{fws}, and (d) *CD* versus *SNR* for the reference microphone signal [·····], the $\mathcal{O}(L^2M^2)$-cost [–•–], the $\mathcal{O}(LM^2)$-cost [- ▪ -], the $\mathcal{O}(L^2M)$-cost [- ♦ -], and the $\mathcal{O}(LM)$-cost [- ▾ -] ISCLP Kalman filter with $L = 6$ if interfering speech is present. 148

5.8 (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus L with respect to the reference microphone signal for the $\mathcal{O}(L^2M^2)$-cost [–•–], the $\mathcal{O}(LM^2)$-cost [- ▪ -], the $\mathcal{O}(L^2M)$-cost [- ♦ -], and the $\mathcal{O}(LM)$-cost [- ▾ -] ISCLP Kalman filter at $SNR = 25\,\text{dB}$ if interfering speech is present. 149

List of Tables

1.1 Overview of spatio-temporal speech enhancement approaches [56–116] and their properties. Approaches covered by, e.g., Ch. 4 of this thesis are denoted by ⟦4⟧, and similarly for other chapters. A legend to some of the properties is given in the footnotes at the end of this table. 13

2.1 Comparative summary of the MCLP framework versus the GSC framework. 48

5.1 Complexity in terms of required multiplications with input domains $\mathbb{R} \times \mathbb{R}$, $\mathbb{R} \times \mathbb{C}$, and $\mathbb{C} \times \mathbb{C}$ of the Kalman filter update equations for (a) preserved cross-correlations, (b) omitted temporal cross-correlations, (c) omitted spatial cross-correlations, and (d) omitted spatio-temporal cross-correlations. The simplified variant of (5.19) using, e.g., (5.24) is denoted by (5.19|5.24), and similarly for the other simplified variants. . . . 136

Chapter 1

Introduction

Denn das Menschlichste, was wir haben, ist doch die Sprache, und wir haben sie, um zu sprechen. [1]

Unwiederbringlich, Theodor Fontane (1819 – 1898), novelist and poet

[1] For the most human thing we have is language, and we have it to speak.

Speech is the genuine mode of human communication. It allows us to exchange abstract ideas, and as such has been essential to the progress and development of our societies and culture.

For millennia, speech was solely and immediately addressed to the human ear. In this age, however, we are able to record and reproduce speech by means of microphones and loudspeakers, and indeed we may argue that speech recordings have never before been as numerous as today. We make use of this asset while having a phone call, or when we wear hearing aids. And not only that, we may even address machines by means of speech, and so for instance communicate to a virtual assistant.

Naturally, when speech is recorded, the microphones do not record mere speech only, but also a variety of further, undesired sound components – and in particular so in *adverse acoustic conditions*. Background noise and interfering speakers, but also reverberation may render speech unintelligible. Fortunately, thanks to the advances of electronics and digital signal processing, we today possess sophisticated tools and understanding to modify microphone signals towards our needs. If we process microphone signals such as to improve speech quality and intelligibility in adverse acoustic conditions, we refer to this as *speech enhancement*. Clearly, speech enhancement is crucial to nearly all speech-related applications presently available.

Despite substantial progress over the course of the past decades, the research in the fields of speech enhancement is still ongoing, and this thesis is but one contribution. We here compare, propose and evaluate existing and novel approaches to the problem, while taking account of practical challenges.

In this chapter, we introduce the thesis as follows. In Sec. 1.1, we outline the nature of adverse acoustic conditions. In Sec. 1.2, we consider speech enhancement applications and challenges, and review the state of the art. The contributions of this thesis are anticipated in Sec. 1.3, while the underlying assumptions are summarized in Sec. 1.4. In Sec. 1.5, we give an overview on the remainder of the thesis.

1.1 Adverse Acoustic Conditions

In many environments, we face adverse acoustic conditions degrading speech quality and intelligibility [1–17]. Consider, for example, the situation in Fig. 1.1, which may be set in a cafeteria or at a cocktail party. As shown in Fig. 1.1 (a), the listener tries to follow speech A, which is referred to as *target speech*. At the same time, speech B is uttered, which is however not of interest to the

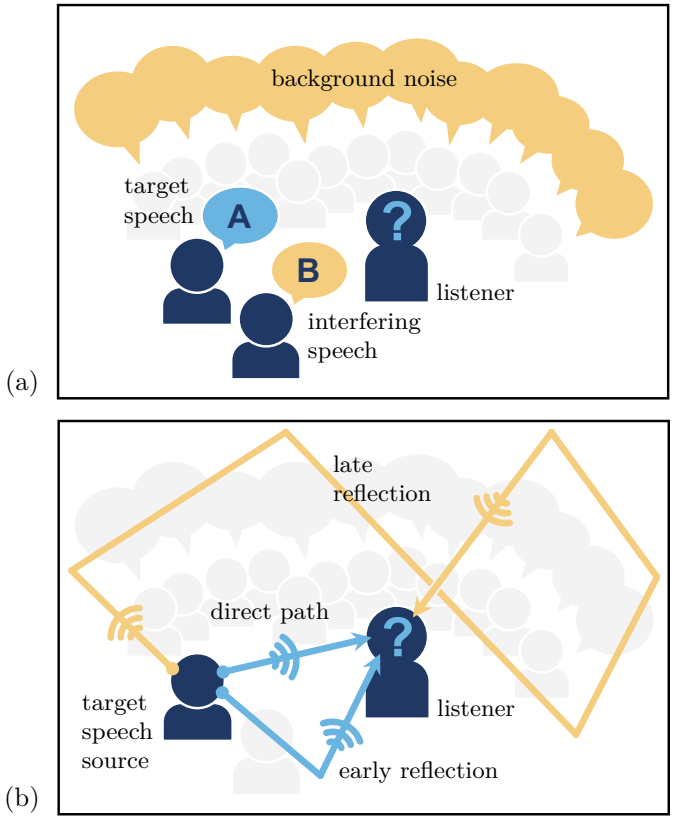


Figure 1.1: Adverse acoustic conditions degrading the quality and intelligibility of target speech – the listener has to cope with (a) interfering speech, background noise, and (b) late reverberation, caused by a multitude of late reflections.

listener and thus referred to as *interfering speech*. The scene is immersed in *background noise*, here produced by a crowd of people chatting, which we refer to as babble noise. As shown in Fig. 1.1 (b), the sound waves of speech A propagate in various directions from the target speech source and reach the listener's ears not only on the direct path, but also via a multitude of early and late reflections on the room enclosure. This acoustic phenomenon of multi-path propagation is referred to as *reverberation*.

We outline the nature of reverberation as well as interfering speech and background noise in Sec. 1.1.1 and Sec. 1.1.2, respectively.

1.1.1 Reverberation

Reverberation is the persistence of sound in a room after excitation. It is commonly quantified by the so-called reverberation time [18, 19], which is defined as the time it takes for the sound pressure level to decay by 60 dB after excitation has ended. In relation to the highly time-varying statistics of speech, which are commonly said to remain stationary for no longer than a few tens of milliseconds [20, 21], the reverberation time may be enormously long. It grows with the volume of the room and the reflectiveness of its surfaces [18, 19], and is typically below 300 ms in living rooms, but may reach several seconds in larger auditoria or churches [18].

Intuitively, reverberation may be explained by the concept of acoustic multi-path propagation from a point source to a point receiver, e.g., a speaker and a listener, in a room with reflective surfaces. From the source, sound waves are emitted in various directions. They reach the receiver first via the direct path, i.e. the line of sight between the source and the receiver, and thereafter via a multitude of reflection paths of different propagation time, cf. Fig. 1.1 (b). A given acoustic channel may be accurately described by the room impulse response (RIR), which embodies the reflection pattern. With increasing propagation time, the density of reflections reaching the receiver grows, but their energy decays exponentially [18], and so the effective length of RIRs is in the range of the reverberation time.

Commonly, reflections are divided into *early* and *late reflections*, cf. Fig. 1.1 (b), causing early and late reverberation, which have different temporal, spatial, and perceptual properties. Early reflections, roughly identified with first-order reflections on walls, the floor and the ceiling, arrive within 50 ms [18] after the direct component and appear sparsely. They cause a directive sound field and are not perceived separately from the direct component, but instead are said to colorize and reinforce it. As such they are considered beneficial to speech intelligibility [18, 22–24]. In the remainder, we refer to the sum of the direct component and early reverberation as the early (speech) source image. Late reflections, associated to higher-order reflections, arrive later but appear densely as a reverberant tail in the RIR. Late reflections cause a spatially fairly homogeneous, often even presumed perfectly diffuse sound field and are perceived as temporal smearing, which in the context of speech is referred to as overlap-masking. As such, late reflections are detrimental to speech intelligibility [1–5, 8, 9, 11, 13–15, 18].

Exemplary spectrograms of an early speech source image and the corresponding reverberant speech [25] at a reverberation time of 610 ms [26] are shown in Fig. 1.2 (a)–(b). In the early speech source image, cf. Fig. 1.2 (a), tonal sounds

such as vowels appear as sets of parallel horizontal lines, while transient sounds such as plosive consonants appear well separated as sharp vertical lines, e.g., at roughly 1 s. If late reverberation is added, cf. Fig. 1.2 (b), the fine temporal structure of the signal is smeared out, and formerly silent periods are filled with reverberation.

1.1.2 Interfering Speech and Background Noise

Apart from late reverberation, we may distinguish between further undesired sound components. In the introductory example of Fig. 1.1, we have introduced interfering speech and background (e.g., babble) noise.

Interfering speech originates from a specific point source, which we may refer to as interfering source. Therewith, it not only causes a directive sound field, but is also subject to late reverberation, cf. Sec. 1.1.1. Like target speech, interfering speech is highly non-stationary. Obviously, interfering sources may also emit other sounds than speech, such as stationary fan noise.[2]

As opposed to interfering speech, background noise is not associated to a specific point source, but is generated in a spatially distributed manner. Similarly to late reflections, it causes a spatially fairly homogeneous, often even presumed perfectly diffuse sound field. Similarly to speech, the statistics of background noise may be highly time-varying, such as in case of babble noise, or may instead be stationary, e.g., in the case of air duct noise. Note that in the case of recordings further noise types of non-acoustic origin take effect, such as sensor and quantization noise generated in microphones and analogue-to-digital converters.

While noise in general is detrimental to the quality and intelligibility of target speech [3, 4, 6–9, 11, 13, 15], speech-like sound components such as interfering speech and babble noise are particularly distracting [3, 7]. An exemplary spectrogram of the superposition of interfering speech [25, 26] and babble noise [27] is shown in Fig. 1.2 (c). The signal energy is densely but irregularly distributed across the spectrogram. Fig. 1.2 (d) shows the superposition of the interfering speech and babble noise in Fig. 1.2 (c) with the reverberant target speech in Fig. 1.2 (b). The fine vertical and horizontal structures of the early target-speech source image in Fig. 1.2 (a) can no longer be recognized.

[2] In Ch. 2, in order to distinguish point-source noise components from other noise components, we refer to them as coherent and incoherent, respectively.

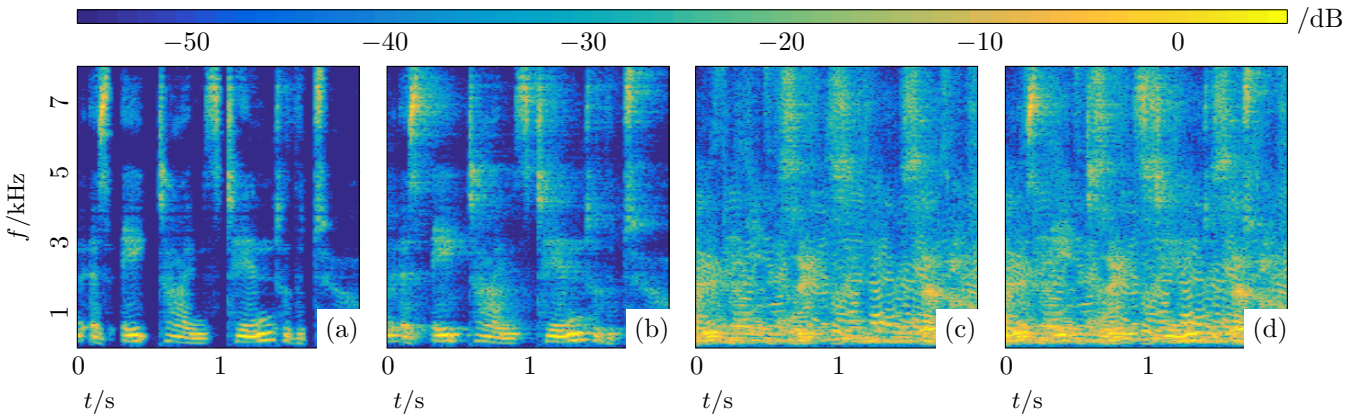


Figure 1.2: Exemplary spectrograms illustrating
(a) the early target-speech source image,
(b) the reverberant target speech,
(c) interfering speech and babble noise, and
(d) the superposition of the signals in (b) and (c).

1.2 Speech Enhancement

The human auditory system often copes remarkably well with adverse acoustic conditions. Normal-hearing listeners are capable of selective hearing, and thus frequently succeed in focusing on a single speaker despite noisy and reverberant surroundings. This phenomenon is often referred to as cocktail party effect [12] and associated with binaural hearing [2–4, 6, 12, 14]. Instead of or in addition to a human listener as in the example in Fig. 1.1, speech is also frequently captured by electronic devices equipped with one or multiple microphones, serving a variety of applications. In nearly all of these applications, however, adverse acoustic conditions act detrimentally, and thus speech enhancement [28–32] is needed to improve speech quality and intelligibility for the end listener, and also the performance of automatic speech recognition. In recent times, due to the pervasive availability of such devices, research and development in speech enhancement has enjoyed broad attention.

We consider typical speech enhancement applications and identify consequential technical challenges in Sec. 1.2.1. The state of the art in speech enhancement is reviewed in Sec. 1.2.2.

1.2.1 Applications and Challenges

The probably longest-standing speech enhancement application is *digital telephony* [33–35], nowadays covering a broad range of technologies such as mobile telephony, the voice over internet protocol (VoIP), and tele-conferencing. In telephony, speech is transmitted from one end to another, with undesired components mixing in due to adverse acoustic conditions at the near end, particularly in distant-speaker scenarios such as in hands-free telephony. The speech signal is usually transmitted monophonically with limited acoustic bandwidth. For data compresion, lossy speech coding [36] is commonly applied, further reducing speech intelligibility and quality – and thereby rendering speech enhancement all the more important.

Another application domain for speech enhancement is *hearing devices*, such as hearing aids or cochlear implants [37–44]. Hearing devices process and reproduce recorded speech signals in order to render them intelligible for their hearing impaired wearer, e.g., by compensating hearing loss through non-linear compression. As hearing impairments typically reduce the ability of selective hearing [6, 7, 9, 15], hearing devices also incorporate speech enhancement in their processing chain.

While telephony and hearing devices concern human-to-human communication, technological advances in *automatic speech recognition* (ASR) [31, 45–49] have made voice-driven human-to-machine communication viable on a mass scale. ASR may be used in, e.g., smart devices such as smart phones or smart speakers, and serve applications as diverse as virtual assistants, home automation, medical transcription, gaming or even military applications. The recognition task may vary across applications between command-and-control and continuous speech recognition. As the performance of ASR also suffers from noise and late reverberation [31, 45–49], especially in distant-speaker scenarios, speech enhancement is often used as a front-end to ASR.

From a technical point of view, the above mentioned applications imply a number of challenges to be tackled in the development of speech enhancement algorithms. These challenges may be identified as follows.

1. **Comprehensive speech enhancement** – While early speech enhancement approaches usually disregarded reverberation, mere dereverberation approaches disregard interfering speech and noise, cf. also Sec. 1.2.2. As the above mentioned applications may be used in diverse acoustic environments, however, simultaneous treatment of all adverse acoustic conditions is generally required in practice. In this thesis, we therefore envisage comprehensive speech enhancement including dereverberation, interfering speech cancellation, and noise reduction.

2. **Target-parameter knowledge** – Speech enhancement commonly requires some prior knowledge relating to the target-speech source, which we refer to as target-parameter knowledge. We may distinguish *channel* knowledge, i.e. knowledge on the acoustic channel between the target-speech source and the microphone(s), *spatial* knowledge, i.e. knowledge relating to the target-speech direction,[3] and *temporal* knowledge, i.e. knowledge relating to the time-varying target-speech statistics. In practical speech enhancement applications, the acoustic environment and naturally the target speech signal are unknown, and so target-parameter knowledge needs to be acquired. While accurate channel estimates are difficult to acquire, especially in noisy environments [142, 143], it is known that dereverberation approaches based on explicit acoustic channel equalization are very sensitive to channel estimation errors and fluctuations [73, 75]. In contrast, usage of spatial and temporal knowledge is well established

[3]Note that channel and spatial knowledge relate to each other, however are not equivalent: while spatial knowledge strictly requires multiple microphones and comprises knowledge on directive components only, i.e. the direct component and early reflections, channel knowledge is defined for both a single and multiple microphones and in addition contains knowledge on late reflections.

in practice, cf. also Sec 1.2.2. In this thesis, we therefore do not rely on *channel* knowledge, but exploit *spatial* and *temporal* knowledge, which we strive to obtain in a *multi-source* scenario.

3. **Online processing** – Online processing, also referred to as real-time processing, means the processing of a continuous stream of input data such as to produce a continuous stream of output data. In our case, the input and output data streams are digital audio signals. The relative delay between the input and the output data stream is referred to as algorithmic delay. In most speech enhancement applications, in particular in telephony an hearing devices, but also in continous-speech recognition applications, online processing with small algorithmic delays is strictly required in order to avoid unnatural pauses in human conversations [34, 35, 40, 41]. Online processing further potentially allows [50–52] to cope with *dynamic acoustic scenarios*, i.e. scenarios with time-varying noise fields or even moving sources and microphones. The opposite of online processing is referred to as batch processing, where an entire batch of input data is collected before any output data is produced, disregarding algorithmic delay constraints. Batch processing may be used in, e.g., command-and-control speech recognition applications, but is generally less useful than online processing. In this thesis, we therefore strive for online speech enhancement able to cope with dynamic acoustic scenarios.

4. **Computational Complexity** – Computational complexity is the amount of resources required to execute an algorithm, most notably the number of arithmetic operations, especially multiplications and divisions, and memory usage. While the processing power of integrated hardware has greatly improved over the last decades, low computational complexity is still required in many applications. For instance, mobile devices such as mobile phones and especially hearing devices exhibit small form factors, which limit processing power, memory, and battery life [39–41]. Furthermore, with speech enhancement being only one of them, available resources have to be shared among several processing blocks, e.g., with acoustic echo cancellation (AEC) [53, 54] in telephony, acoustic feedback cancellation (AFC) [38, 44, 54, 55] and non-linear compression in hearing devices, ASR in smart devices, and potentially many other services unrelated to audio. In many signal processing tasks, computational complexity may be reduced at the expense of performance. In this thesis, we therefore strive for reasonable trade-offs between computational complexity and performance of speech enhancement.

1.2.2 State of the Art

From a modeling and algorithmic development perspective, dereverberation is far more challenging as compared to interfering speech cancellation and noise reduction. Dereverberation approaches have to cope with the enormous length of acoustic channels, the time-varying statistics of the late reverberant sound field, and the nearly always simultaneous presence of early speech and late reverberant components. During the past decades, based on different models on the observed reverberant signal and using different kinds of target-parameter knowledge, a large variety of dereverberation approaches has emerged [56–116]. Furthermore, as nearly all approaches to noise reduction have also been adopted in dereverberation in a similar manner, we do not discuss noise reduction separately, but only in conjunction with dereverberation.

Selected speech enhancement approaches are reviewed in Sec. 1.2.2.1. The estimation of required target-parameter knowledge is briefly summarized in Sec. 1.2.2.2.

1.2.2.1 Spatio-Temporal Speech Enhancement

In this section, we discuss *spatial*, *temporal*, and *spatio-temporal* dereverberation and speech enhancement approaches. In this context, the attribute *spatial* may be interpreted in a wide sense, namely as referring to approaches that exploit variations across space, i.e. all approaches employing multiple microphones, or in a strict sense, namely as referring to approaches that exploit explicit spatial knowledge on the target speech source. Similarly, the attribute *temporal* may be interpreted in a wide sense, namely as referring to approaches that exploit variations across time, or in a strict sense, namely as referring to approaches that exploit explicit temporal knowledge on the target speech source.

In the following, we discuss the classes of approaches shown in Table 1.1, which are differently motivated: (a) spectral enhancement [56–70], (b) channel equalization [71–79], (c) beamforming [80–87], (d) multi-channel linear prediction (MCLP) [88–103], (e) signal state-space estimation [104–107], and (f) hybrid approaches [108–116]. Each approach in Table 1.1 is categorized by the following properties (cf. also the table's legend):

- **Multi-channel** – Most approaches rely on multiple channels, and may therefore be classified as *spatial* approaches in the wide sense.
- **Model domain** – Most approaches rely on time domain (TD) or time-frequency domain (TFD) models, while mere frequency domain (FD)

models are rarely used. In the TFD and the FD, individual frequency bins are usually treated independently, which reduces computational complexity.

- **Modeling of (late) reverberation** – Reverberation is modelled as convolutive in the TD or per frequency bin in the TFD [117], and rarely as multiplicative in the FD. Approaches based on convolutive models may be classified as *temporal* approaches in the wide sense. In the FD and the TFD, the RIR respectively corresponds to the room transfer function (RTF) and the convolutive RTF or sub-band RIR [117]. In the TFD, the frame length is usually in the same order as the arrival time span of early reflections, and so it is customary to assume that early reflections arrive within the same frame, while late reflections arrive in subsequent frames. Alternatively, late reverberation may be modelled more coarsely as additive in the TFD, where it is assumed that the early source image and late reverberation are uncorrelated, and so late reverberation may be treated similarly to noise.
- **Treatment of multiple sources and noise** – Commonly, a single source is assumed, such that interfering speech suppression is rarely considered. While some approaches consider noise suppression, others do not, rendering them mere dereverberation approaches.
- **Target-parameter knowledge** – Target-parameter knowledge is either required a priori, or its estimation is an inherent part of the approach. Channel knowledge is usually provided by RIR estimate(s). Spatial knowledge is given by estimates of the direction-of-arrival (DoA) or the relative early transfer functions (RETFs) of the early target-speech source image, in the following referred to as target-speech DoA and target-speech RETFs, respectively. Temporal knowledge is usually given by estimates of the time-varying power spectral density (PSD) of the early target-speech source image, referred to as early target-speech PSD. Approaches using spatial and/or temporal target-parameter knowledge may be classified as *spatial* and/or *temporal* in the strict sense.
- **Online processing** – While some approaches allow online processing of the microphone signals, others rely on batch processing.
- **Computational complexity** – Approaches based on additive late reverberation models may be said to be of relatively low complexity, while approaches based on convolutive models exhibit moderate to high computational complexity.

Note that apart from the approach classes (a)-(f) further approach classes to dereverberation have emerged, such as, e.g., approaches using a speech

production model and residual processing [118–120], approaches based on non-negative matrix factorization [121–123], and deep learning-based approaches [124–128]. As these relate only remotely to the contributions of this thesis, however, we do not further discuss these here, but leave it at mentioning them for completeness.

Table 1.1: Overview of spatio-temporal speech enhancement approaches [56–116] and their properties. Approaches covered by, e.g., Ch. 4 of this thesis are denoted by ⟦4⟧, and similarly for other chapters. A legend to some of the properties is given in the footnotes at the end of this table.

approach class	*multi-channel*	*model domain*	*modeling of (late) reverberation*	*treatment of multiple sources and noise*[1]	*target-parameter knowledge*[2]	*online processing*[3]	*computational complexity*[4]
(a) *spectral enhancement* [56–70]⟦4⟧	✗ [56, 58] ✓ [57–70] ⟦4⟧	TFD	additive	✓✓ [61, 69] ⟦4⟧ ✓ [58–70] ✗ [56, 57]	spatial [57–64, 66, 67] ⟦4⟧ (spatial) [65, 69, 70] ⟦4⟧ (temporal) [56–70] ⟦4⟧	✓	low
(b) *channel equalization* [71–79]	✓ [71–75, 77–79] ✗ [72, 76]	TD [71–75, 77, 78] FD [76] TFD [79]	multiplicative [76] convolutive [71–75, 77–79]	✗	channel [71–79] (temporal) [78]	(✓) [71–77, 79] ✗ [78]	moderate [71–77, 79] high [78]

(c) *beamforming* [80–87] ⟦2⟧	✓	TD [81] ⟦2⟧ FD [80, 82] TFD [83–87] ⟦2⟧	additive [83–87] multiplicative [80, 82] convolutive [81, 87] ⟦2⟧	✓✓ [84] ⟦2⟧ ✓ [82–87] ⟦2⟧ ✗ [81]	channel [80–82, 87] spatial [83–87] ⟦2⟧ temporal [85, 86] ⟦2⟧	✓ [83–86] (✓) [80–82, 87] ✗ ⟦2⟧	low [83–86] moderate [80–82, 87] high ⟦2⟧
(d) *MCLP* [88–103]	✓	TD [88, 89, 91] ⟦2⟧ TFD [90, 92–103] ⟦2⟧	convolutive	✓✓ [94] ✗ [90, 92, 93, 95–103] ⟦2⟧	temporal [92, 98–103] ⟦2⟧ (temporal) [88–91, 93–97]	✓ [92, 98–103] ✗ [88–91, 93–97] ⟦2⟧	moderate [102] high [88–101, 103]
(e) *signal state-space estimation* [104–107]	✓	TD [104, 105] FD [106] TFD [107]	multiplicative [106] convolutive [104, 105, 107]	✓	(channel) [104–107] (temporal) [104–107]	✓	high

(f) *hybrid approaches* *channel equalization & beamforming* [108–110] *MCLP & beamforming* [111–115] ⟦3, 5⟧ *MCLP & signal state-space estimation* [116]	✓	TD [109, 110] FD [108] TFD [111–116] ⟦3, 5⟧	convolutive	✓✓ [111] ⟦3, 5⟧ ✓ [110–116] ⟦3, 5⟧ ✗ [108, 109]	channel [108–110] spatial [111–115] ⟦3, 5⟧ temporal [111, 116] ⟦3, 5⟧ (temporal) [112–115]	✓ [111, 116] ⟦3, 5⟧ (✓) [108–110] ✗ [112–115]	moderate [108–110] ⟦5⟧ high [111–116] ⟦3⟧

[1]The treatment of noise and multiple sources is categorized in the following manner,

- ✓✓ – multiple sources are treated separately,
- ✓ – noise is treated,
- ✗ – neither of the above.

[2]Target-parameter knowledge is categorized in the following manner,

- channel – RIR(s) (TD), RTF(s) (FD), or convolutive RTF(s) (TFD) between the speech source and the microphone(s), estimates presumed given,
- spatial – target-speech DoA (TD or TFD) or RETFs (TFD), estimates presumed given,
- temporal – time-varying early target-speech PSD or magnitude (TFD), associated power ratios (TFD), or linear prediction coefficients of the target-speech source signal (TD), estimates presumed given,
- (—) – as above, but estimation is an inherent part of the approach.

[3]The entry (✓) indicates that online processing is principally possible, but as the required acoustic channel estimate is typically obtained before runtime, dynamic scenarios cannot be tracked.

[4]The computational complexity during run time is categorized in the following manner,

low	– additive model, linear or quadratic cost in the number of channels,
moderate/high	– convolutive or multiplicative model, linear/quadratic cost in the number of channels and filter coefficients or frequency bins, respectively.

(a) Spectral Enhancement Spectral enhancement [30, 56–70] relies on an additive model of late reverberation in the TFD, i.e. late reverberation is treated similarly to noise. A real-valued time- and frequency-varying gain is derived based on some optimalty criterion, e.g., the (multi-channel) Wiener filter ((M)WF) criterion [42, 129, 130], and applied to the (spatially pre-processed, cf. also beamforming) microphone signals. Depending on whether early speech or late reverberation and noise are pre-dominant in a particular time-frequency tile, the derived gain will be closer to one or closer to zero, and thereby suppresses undesired components at the expense of some amount of speech distortion. The gain depends on *temporal* target-parameter knowledge in the form of the early target-speech PSD and further late reverberation and noise PSDs or alternatively the signal-to-(noise-plus-)reverberation ratio (S(N)RR). Generally, spectral enhancement is suitable for online processing and may be said to be of relatively low complexity.

In single-channel spectral enhancement [56, 58], the early target-speech PSD and late reverberant PSD estimation are based on a statistical model for reverberation parametrized by room acoustic measures, e.g., based on Polack's model [131] and the presumed known reverberation time [18]. In multi-channel spectral enhancement [57–70], most approaches [57–64, 66, 67] rely on *spatial* target-parameter knowledge. It is further commonly assumed that late reverberation may be modeled as diffuse [61–70, 132]. In this case, the SRR becomes equivalent to the signal-to-diffuse ratio (SDR) [59, 60], which may be used instead of PSD estimates. Note that multi-channel speech enhancement is commonly applied in conjunction with beamforming [57, 58, 60–68], most notably in the MWF [61–68].

In Ch. 4, we propose a square root-based *multi-source* early PSD estimation and RETF updating approach, cf. also Sec. 1.3, which may similarly be used for spectral enhancement, cf. Ch. 3.

(b) Channel Equalization Channel equalization [31, 71–79] is typically based on a convolutive reverberation model in the TD [71–75, 77, 78] and aims at dereverberation by equalizing the acoustic channel between the target source and the microphone(s), while noise reduction is not considered. While individual RIRs are not invertible due to their non-minimum-phase characteristic [76], the pioneering multiple input/output inverse theorem (MINT) [71] states that using multiple channels, an exact and stable inverse indeed exists if the RTFs are relatively prime, i.e. if they do not share common zeros. Consequently, most channel equalization approaches rely on multiple channels [71–75, 77–79] and may be classified as MINT-based [71, 73–75, 77–79]. Channel knowledge is fundamental and presumed available in form of RIR estimates. The enormous

length of RIRs makes channel equalization very sensitive to estimation errors and acoustic channel fluctuations [73]. As the equalization filter is typically estimated before run time solely based on channel knowledge [71–77,79], dynamic scenarios cannot be tracked, but computational complexity remains moderate.

Instead of complete equalization, partial equalization may be used to maintain early reflections [72, 74, 75]. In [72], channel shortening based on a Rayleigh quotient criterion maximizing the early-to-late energy ratio has been proposed, which can be shown to be closely related to partial MINT equalization [74, 75]. To reduce the detrimental effects of RIR estimation errors and RIR fluctuations during runtime on the equalization performance, Tikhonov regularization [73,75] as well as generalized Rayleigh quotient regularization [73, 75] may be applied. In [78], data-dependent MINT-based equalization promoting sparsity has been proposed.

(c) Beamforming Beamforming [42, 80–87, 133, 134] commonly relies on an additive model of late reverberation in the TFD [83–87], treating reverberation similarly to noise. As a multi-channel filtering technique using explicit *spatial* knowledge on the target source, beamforming aims at combining individual channels such that constructive interference is obtained in the desired direction, while destructive interference is obtained in other directions, thereby suppressing undesired components. This may be achieved by means of data-independent [83] or data-dependent optimality criteria [42, 80–87, 133, 134], e.g., the data-independent super-directive criterion [42,83,134], or the data-dependent minimum-variance distortionless response (MVDR) [42, 80–83, 85–87, 134] criterion. The MWF [42, 61–68, 129, 130] contains the MVDR beamformer as a spatial pre-processor. Online processing is typically feasible, where TFD domain approaches solely based on an additive late reverberation model [83–86] may be said to be of relatively low complexity.

In [83], a cascaded approach is presented, using data-independent, super-directive beamforming for dereverberation, and data-dependent beamforming for noise reduction. The generalized sidelobe canceler (GSC), a popular implementation of the MVDR beamformer using an unconstrained sidelobe cancellation (SC) filter, has been applied in different constellations [85, 87]. In [85], joint dereverberation and noise reduction is performed using a single GSC, while in [87], a nested structure is proposed, employing an inner GSC for dereverberation and an outer GSC for noise reduction. In [84], multiple speech sources are considered, with spatial knowledge based on instantaneous DoA estimates. While typically not aiming at channel inversion, beamforming based on a convolutive reverberation model in the TD has been shown to relate to MINT equalization if RIRs are incorporated in the filter design [81].

In Ch. 2, we analyze and compare the GSC architecture and the MCLP architecture for joint dereverberation and noise reduction both in the TD and the TFD, cf. also Sec. 1.3.

(d) MCLP Multi-channel linear prediction (MCLP) [88–103] is a commonly used deconvolution technique based on a convolutive reverberation model in the TD or TFD, while noise reduction is not targeted. It potentially achieves complete dereverberation under MINT conditions [71]. As opposed to MINT-based equalization, however, prior knowledge on the acoustic channels is not required, but equalization is performed blindly, rendering it particularly attractive in practical applications. MCLP relies on the premise that the late reverberation to be canceled can be modelled as a filtered version of the delayed microphone signals, i.e. as a linear prediction (LP) component. The prediction delay defines the amount of early reflections to be maintained. The sole task in MCLP therefore consists in estimating the multi-channel LP filter from the microphone signals. While batch processing is common [88–91, 93–97], more recent approaches operate online [92, 98–103]. Due to the convolutive reverberation model and the data-dependent LP filter estimation (as opposed to MINT-based equalization), computational complexity may be said to be comparably high [88–101, 103].

For present-day prevailing TFD-based MCLP [90, 92–103], it is known that incorporating *temporal* target-parameter knowledge in the filter estimation, e.g., the early target-speech PSD, greatly improves performance. Batch processing approaches [90, 93–97] typically rely on maximum-likelihood estimation [90, 93, 94, 96, 97] and a time-varying Gaussian [90, 93–95] or sparse prior-based [96, 97] model of the early speech source image, and estimate the LP filter and temporal target-parameter knowledge in an alternating, iterative manner. Online approaches [92, 98–103] in contrast, based on RLS [92, 100, 103] or the Kalman filter [98, 99, 102], require temporal target-parameter knowledge a priori. In [94], MCLP for several point sources is combined with blind source separation (BSS) [135, 136]. A complexity-reduced MCLP Kalman filter neglecting temporal cross-correlations in the state estimation error correlation matrix has been proposed in [102].

In Ch. 2, we analyze and compare the MCLP architecture and the GSC architecture for joint dereverberation and noise reduction both in the TD and the TFD, cf. also Sec. 1.3.

(e) Signal State-Space Estimation Signal state-space estimation [104–107] may be based on a convolutive reverberation model in the TD [104, 105] or the TFD [107], or on a multiplicative model in the FD [106]. Approaches of this kind

interpret the microphone signal model as the measurement equation of a state-space model, with the speech-source signal [104–106] or the early speech source image [107] being the true state to be estimated. All presented approaches [104–107] treat the acoustic channel as unknown, rendering the estimation problem non-linear. The recursive nature of state estimation algorithms allows for online processing, where computational complexity may be said to be comparably high.

In order to tackle the non-linear state estimation problem, the uncented Kalman filter is employed in [104]. In [105], two alternating Kalman filters are used to estimate the speech-source signal and the acoustic channels, with further unknown model parameters sampled in a particle filter. In [106, 107], the Kalman filter is embedded in an expectation-maximization architecture in order to estimate both the state and remaining model parameters.

(f) Hybrid Approaches Hybrid approaches [108–116] combine two of the previously discussed different approach classes in order to compensate for the short-comings of each. We consider approaches combining channel equalization and beamforming [108–110], approaches combining MCLP and beamforming [111–115] and an approach combining MCLP and signal state-space estimation [116]. Another commonly used combination consists of beamforming and spectral enhancement, as, e.g., in the MWF [42, 61–68, 129, 130].

In [108, 109], for merely reverberant scenarios, a trade-off between the dereverberation performance of MINT-based equalization and the robustness of beamforming is obtained by combining the respective cost functions. Similarly, for reverberant and noisy scenarios, a trade-off between dereverberation and noise reduction may be obtained [110]. Instead of MINT-based equalization, also MCLP may be combined with beamforming in order to achieve both dereverberation and noise reduction. Cascades of MCLP and beamforming have been used in [111–115], which was seen to be a commonly adopted approach in the 2018 CHiME-5 challenge [49]. In [115], it has been proposed to unify the cascade, yielding a single cost function. In [116], MCLP is combined with signal state-space estimation, where a pair of alternating Kalman filters respectively estimate the LP filter and the noise-free but reverberant speech component.

In Ch. 3, we propose the integrated sidelobe cancellation and linear prediction (ISCLP) Kalman filter, cf. also Sec. 1.3, which integrates the MCLP and the GSC architecture. In Ch. 5, complexity-reduced variants of the ISCLP Kalman filter are proposed, cf. also Sec. 1.3.

1.2.2.2 Target-Parameter Estimation

The approaches discussed in Sec. 1.2.2.1 commonly require *channel* knowledge, *spatial* knowledge, or *temporal* knowledge on the target speech source. In this section, we briefly summarize corresponding estimation approaches.

Channel Knowledge Channel knowledge, as required in [71–82, 87, 108–110], is usually provided by RIR estimates. For unknown source signals as in speech enhancement, RIR estimation is a blind system identification (BSI) problem. In case of a single source and multiple channels, acoustic BSI approaches typically rely on the cross-relation property [137–141], which denotes the fact that if two microphone signals are filtered with the RIRs to the respective other microphone, the output of both filters must be the same. If the autocorrelation matrix of the source signal has full rank, the RIRs are identifiable under MINT conditions, i.e. if the RTFs do not share common zeros. Note that BSI for RIRs is known to be sensitive to near-common zeros and noise [142, 143].

Spatial and Temporal Knowledge Spatial knowledge, as required in [57–64, 66, 67, 83–87, 111–115], may be provided by target-speech DoA estimates [135, 144–146], e.g., obtained by the well-known MUSIC algorithm [144], or target-speech RETF estimates [65, 69, 70, 147–149], e.g. obtained by subspace-based covariance whitening [65, 147–150]. Temporal knowledge, as required in [85, 86, 92, 98–103, 111, 116], is usually provided by early target-speech PSD estimates. The early target-speech PSD estimate may, e.g., be obtained as in spectral enhancement [56–58, 61–70]. In [86, 100, 111], the early target-speech PSD estimate is obtained based on a statistical model for late reverberation similar to [56, 57], while in [102], the estimator in [68] is used. Alternatively, the early target-speech PSD may be estimated from the enhanced signal obtained in previous frames [99, 116], or by means of a neural network [101, 103]. Approaches to joint estimation of spatial and temporal knowledge have been proposed for both a single source [65, 70] and multiple sources [69].

In Ch. 4, we propose a square root-based *multi-source* early PSD estimation and RETF updating approach, cf. also Sec. 1.3.

1.3 Contributions

The contributions of this thesis may be linked to the challenges identified in Sec. 1.2.1 as outlined below.

1. **Comprehensive speech enhancement** – In this thesis, we envisage comprehensive speech enhancement including dereverberation, interfering speech cancellation, and noise reduction. Presently, the undoubtedly most popular approach class to dereverberation is MCLP. Based on a convolutive reverberation model, MCLP is potentially more performant than spectral enhancement and beamforming, which are usually based on additive late reverberation models. Yet, as compared to other convolutive reverberation model-based approach classes, namely channel equalization and signal state-space estimation, MCLP does not rely on prior or intermediate channel estimation. Unfortunately, however, interfering speech and noise are not suppressed in MCLP. In terms of interfering speech cancellation and noise reduction, instead, beamforming may be identified as the best established approach.

   Therefore, as preparatory step towards comprehensive speech enhancement, we *analyze and compare* the MCLP architecture to a data-dependent beamforming architecture, namely the GSC architecture, cf. Ch. 2. Here, as opposed to most beamforming architectures, also the GSC formulation relies on a convolutive reverberation model. In the signal model, we consider a reverberant target speech component, coherent (i.e. point-source) noise and incoherent noise components. We provide a better understanding of the theoretical performance limitations of both architectures depending on a number of boundary conditions, such as noise levels, filter length and number of microphones. In particular, assuming perfect spatial knowledge, we show that the GSC potentially performs equally well as MCLP in terms of dereverberation, while obviously performing noise reduction in addition. These theoretical findings are confirmed by simulations in the TD. Further, based on TFD simulations, we show that if spatial knowledge is deficient due to modeling and estimation errors, the GSC instead performs worse than MCLP in terms of dereverberation.

   Given the results of Ch. 2 and the success [49] of MCLP-and-beamforming cascades, based on a TFD formulation, we propose to *integrate* the GSC and MCLP into a parallel architecture we refer to as *integrated sidelobe cancellation and linear prediction* (ISCLP), cf. Ch. 3. The signal model comprises several reverberant speech components, whereof some are to be dereverberated and others to be canceled, as well as a diffuse (e.g., babble) noise component to be suppressed. Within the ISCLP architecture, we estimate the SC and LP filters jointly by means of a single Kalman filter. We further augment the ISCLP Kalman filter by a spectral post-processor, which is shown to relate to the Kalman filter's posterior state estimate. In order to demonstrate the effectiveness of the ISCLP Kalman filter, we compare against two state-of-the-art approaches – first the previously

mentioned alternating Kalman filters in [116], and second an MCLP+GSC Kalman filter cascade, conceptually relating to [111–114]. As compared to these two reference algorithms, the ISCLP Kalman filter is computationally less expensive, yet it is shown to perform similarly or better as compared to the alternating Kalman filters, and to outperform the MCLP+GSC Kalman filter cascade.

2. **Target-parameter knowledge** – The proposed TFD-based ISCLP Kalman filter, cf. Ch. 3, does not rely on channel knowledge, but requires *spatial* and *temporal* target-parameter knowledge, namely target-speech RETFs and the early target-speech PSD.

   In this thesis, we strive to obtain this knowledge in a *multi-source* scenario. In Ch. 4, we propose a square root-based multi-source early PSD estimation and RETF update approach in reverberant environments. Here, as opposed to the conventional approach in the manner of [61,64,66,84,151], we propose to factorize the early correlation matrix and minimize the approximation error defined with respect to the early-correlation-matrix *square root*. The proposed minimization problem seeks a unitary matrix and the square roots of the early PSDs up to an arbitrary complex argument, making conventionally used non-negative thresholding or non-negative inequality constraints redundant. The estimated unitary matrix and early PSD square roots further allow to recursively update the RETF estimate, which is not inherently possible in the conventional approach. Simulation results indicate better performance as compared to the conventional approach in terms of the relative squared PSD estimation error and the signal-to-interference ratio measuring the source-component separation.

3. **Online processing** – In this thesis, we strive for online speech enhancement able to cope with dynamic acoustic scenarios. While the comparative analysis of the GSC and MCLP in Ch. 2 relies on batch processing, the ISCLP Kalman filter proposed in Ch. 3, operates recursively by nature. In the square root-based multi-source early PSD estimation and RETF updating approach proposed in Ch. 4, the early PSDs are estimated independently per frame, while the RETF update is performed recursively. Both proposed approaches may therefore be implemented in an online processing chain. Both the state of the ISCLP state-space model in Ch. 3 and the RETFs in Ch. 4 are assumed time-varying, which allows to track dynamic scenarios.

4. **Computational Complexity** – In this thesis, we strive for reasonable trade-offs between computational complexity and performance of speech enhancement.

Although being cheaper than comparable state-of-the-art approaches [112, 113, 116], the ISCLP Kalman filter proposed in Ch. 3 exhibits a quadratic cost in the number of filter coefficients and channels per frequency bin. In Ch. 5, we consider complexity reduction of the ISCLP Kalman filter by neglecting cross-correlations in the state estimation error correlation matrix, thereby enforcing a sparse structure. Specifically, we consider neglecting temporal, spatial, and all cross-correlations, obtaining linear cost in the number of filter coefficients, linear cost in the number of channels, and linear cost in both the number of filter coefficients and channels, respectively. In our experimental validation adopting the same simulation setup as for the original ISCLP Kalman filter in Ch. 3, we show that the simplified variants of the ISCLP Kalman filter perform nearly as well as the original variant, thereby permitting far more favourable trade-offs between complexity and performance.

Note that MATLAB implementations of the ISCLP Kalman filter in Ch. 3, the multi-source early PSD estimation and RETF updating approach in Ch. 4, and the complexity-reduced Kaman filters in Ch. 5 are available online [152–154].

1.4 Assumptions

The contributions outlined in Sec. 1.3 are based on a number of commonly and successfully employed assumptions regarding the observed microphone signals in the TFD. These are summarized as follows.

- The signal is assumed to be uncorrelated across sub-bands. For convolutive sub-band models, this corresponds to an approximation of the TD convolution in the TFD [117], ignoring cross-band dependencies. It is used throughout the thesis and in the vast majority of TFD-based speech enhancement approaches [56–70, 79, 83–87, 90, 92–97, 99–103, 107, 111–116], as it allows to treat sub-bands independently and thereby reduce computational complexity.
- Both late reverberation and background noise are assumed to be diffuse. Lacking detailed knowledge on the acoustic environment, this is a commonly made [42, 61–70, 83, 132, 134, 167], albeit not necessarily accurate assumption. Along with the subsequent assumption, it is employed in Ch. 4 in order to separate the microphone signal correlation matrix into directive and presumed diffuse components.
- Within the limits defined by the reverberation time, we assume that late reverberation is correlated to early source image in previous frames,

but uncorrelated to the early source image *within* the current frame. Again, albeit not necessarily accurate, this assumption is commonly made [61–70, 83, 132]. Along with the previous assumption, it is, e.g., employed in Ch. 4 in order to separate the microphone signal correlation matrix into directive and presumed diffuse components.

- The early source image is assumed to be temporally uncorrelated across frames. Note that with late reverberation being correlated to the early source image in previous frames, this assumption in fact is a pre-requisite to the previous assumption, namely that late reverberation is uncorrelated to the early source image *within* the current frame. It is, e.g., further used in the ISCLP state-space model in Ch. 3, where the early source image takes the role of presumed temporally uncorrelated measurement noise.

Naturally, in realistic acoustic scenarios, the above assumptions are approximately valid only. Nonetheless, as they result in a coarse but fairly simple signal model, they serve as a practical and solid foundation for algorithmic development and may be said to be an accepted standard in the state of the art. Within the individual chapters of this thesis, the above assumptions are treated in more detail and introduced in a formal manner.

1.5 Outline of the Thesis

The remainder of this thesis is organized in three parts and a conclusion.

Part I covers **Ch. 2**. In Ch. 2, we discuss the the comparative analysis of the GSC and MCLP, which underlie the ISCLP architecture. Ch. 2 is based on the publication

- **T. Dietzen**, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, "Comparative analysis of generalized sidelobe cancellation and multi-channel linear prediction for speech dereverberation and noise reduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 3, pp. 544–558, Mar. 2019.

Part II covers **Ch. 3** and **Ch. 4**. In Ch. 3, we discuss the proposed ISCLP Kalman filter, which integrates the GSC and MCLP architectures. In Ch. 4, we discuss the proposed approach to mutli-source early PSD estimation and RETF updating, which serves the target-parameter estimation in the ISCLP Kalman filter. **Ch. 3** and **Ch. 4** are based on the respective submissions

- **T. Dietzen**, S. Doclo, M. Moonen, and T. van Waterschoot, "Integrated sidelobe cancellation and linear prediction Kalman filter for joint multi-microphone dereverberation, interfering speech cancellation, and noise reduction," ESAT-STADIUS Tech. Rep. TR 19-70, KU Leuven, Belgium, submitted for publication, June 2019.
- **T. Dietzen**, S. Doclo, M. Moonen, and T. van Waterschoot, "Square root-based multi-source early PSD estimation and recursive RETF update in reverberant environments by means of the orthogonal Procrustes problem," ESAT-STADIUS Tech. Rep. TR 19-69, KU Leuven, Belgium, submitted for publication, June 2019.

Part III covers **Ch. 5**. In Ch. 5, complexity reduction of the ISCLP Kalman filter is discussed.

In **Ch. 6**, the thesis is concluded by a summary, suggestions for future research and a discussion on industrial valorization.

Part I

Underlying Architectures

Chapter 2

Comparative Analysis of the GSC and MCLP

Comparative Analysis of Generalized Sidelobe Cancellation and Multi-Channel Linear Prediction for Speech Dereverberation and Noise Reduction

Thomas Dietzen, Ann Spriet, Wouter Tirry, Simon Doclo, Marc Moonen, and Toon van Waterschoot

Published in *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 3, pp. 544–558, Mar. 2019.

© 2019 IEEE. Reprinted, with permission, from:
T. Dietzen, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, Comparative analysis of generalized sidelobe cancellation and multi-channel linear prediction for speech dereverberation and noise reduction, *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 3, pp. 544–558, Mar. 2019.
Changes include layout, representation, minor additions, and minor editing aspects.

The candidate's contributions as first author include: literature study, co-derivation of the presented theory, co-development of the presented algorithms, co-design of the evaluation experiments, software implementation and computer simulations, co-formulation of the conclusions, text redaction and editing.

Abstract

For blind speech dereverberation, two frameworks are commonly used: on the one hand, the multi-channel linear prediction (MCLP) framework, and on the other hand, data-dependent beamforming, e.g., the generalized sidelobe canceler (GSC) framework. The MCLP framework is designed to perform deconvolution and hence has gained increased prominence in blind speech dereverberation. The GSC framework is commonly used for noise reduction, but may be applied for dereverberation as well. In previous work we have shown that for the noiseless case, MCLP and the GSC yield in theory mathematically equivalent results in terms of dereverberation. In this paper, we assume additional coherent- as well as incoherent-noise components and formally analyze and compare both frameworks in terms of dereverberation and noise reduction performance. Both the theoretical analysis and time domain simulation results demonstrate that unlike the GSC, MCLP expectably shows limited performance in terms of noise reduction, while both perform equally well in terms of dereverberation, provided that the GSC blocking matrix achieves complete blocking of the early reverberant-speech component and sufficiently many microphones are available. In case of incomplete blocking, however, the GSC performs inferior to MCLP in terms of dereverberation, as shown in short-time Fourier transform (STFT) domain simulations.

Index terms — Multi-channel linear prediction, data-dependent beamforming, dereverberation, noise reduction.

2.1 Introduction

It is well known that reverberation, caused by reflections against room boundaries and objects, and background noise may have a deteriorating effect on the quality and intelligibility of a speech signal recorded by a microphone [15]. Speech dereverberation accompanied by noise reduction is therefore needed in many applications ranging from hands-free mobile telephony to distant automatic speech recognition.

Dereverberation approaches based on multiple microphones take advantage of spatial diversity and, according to the multiple input/output inverse theorem (MINT) [71], theoretically allow complete inversion of the (presumed time-invariant) room impulse responses (RIRs) between the speech source and the microphone array, provided that the corresponding transfer functions do not share common zeros. In practical applications however, the RIRs are unknown – and since MINT is very sensitive to RIR estimation errors [73], which are

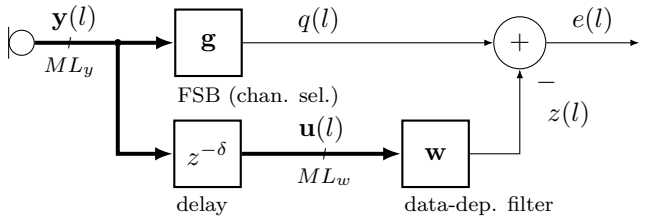


Figure 2.1: The MCLP framework employing the prediction delay δ in the data-dependent filter path.

unavoidable in practice, especially in noisy environments [142, 143], explicit inversion is not favorable. In recent years instead, assuming *no or limited prior knowledge* on the RIRs, multi-channel linear prediction (MCLP) [88, 89, 91, 93–96, 99, 100, 102, 111, 116], beamforming [81, 82, 85, 87, 108, 110, 155, 156] and combinations thereof [112, 113, 157] have been most commonly and successfully used for (blind) speech dereverberation, while partly including noise reduction [85,87,110,112,113,116,157]. In the following, we briefly review these approaches.

The MCLP framework is designed to perform deconvolution, and is hence suited for dereverberation, while noise reduction is *not* targeted. It operates blindly on the microphone signals, i.e. does not require any prior knowledge on the RIRs. A block diagram of MCLP is shown in Fig. 2.1. The framework relies on the premise that the reverberant component to be canceled can be modeled as a filtered version of the delayed microphone signals, i.e. as a linear prediction component. The prediction delay is a design parameter defining the number of early reflections to be maintained. The sole task in MCLP therefore consists in estimating the multi-channel prediction filter from the microphone signals. When the prediction filter is of sufficient order, MCLP is theoretically able to completely equalize the RIRs [89]. Nowadays, MCLP is commonly implemented in frequency sub-bands using the short-time Fourier transform (STFT) [91, 93–96, 99, 100, 102, 111–113, 116, 157]. Incorporating the power spectral density (PSD) of the speech-source signal in the cost function has been shown to be beneficial, as, e.g., in the weighted prediction error (WPE) method [94, 95], where the speech-source signal is modeled as time-varying Gaussian [91, 93–95] or using sparse priors [96]. Adaptive approaches based on recursive least squares [100, 111] and the Kalman filter [99, 102, 116, 157] have been proposed. In [116], given *noisy* microphone signals, the reverberant-speech component and the prediction-filter coefficients are estimated in an alternating fashion. To reduce noise after dereverberation, it has been proposed to cascade MCLP with minimum-variance distortionless response (MVDR) beamforming [112], [113], which became a popular approach in the recent CHiME-5 challenge [49]. Beamforming is designed to perform spatial filtering, and is hence commonly used for noise reduction, but may *also* be applied

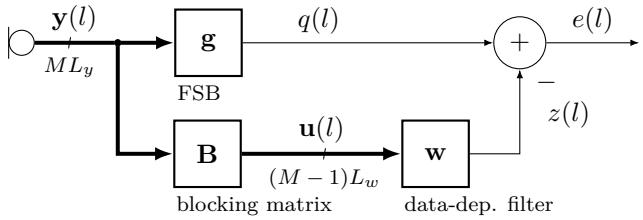


Figure 2.2: The GSC framework employing the blocking matrix $\mathbf{B}$ in the data-dependent filter path.

for dereverberation [31]. One can distinguish between data-independent (e.g., superdirective) beamforming and data-dependent (e.g., MVDR) beamforming. Although beamforming traditionally does *not* target channel inversion, it may be considered equivalent to MINT if the (presumably known) RIRs are incorporated in the filter design [81]. The so-called MINTFormer [108] provides a trade-off between the performance of MINT and the robustness of beamforming. In [110], a MINT-based multi-channel Wiener filter for joint dereverberation and noise reduction has been proposed. The analysis in [82] shows that for MVDR beamforming incorporating known RIRs, an inherent performance trade-off exists between dereverberation and noise reduction in case of incoherent as well as mixed coherent- and incoherent-noise fields. In this work, we are mainly concerned with the generalized sidelobe canceller (GSC) [32, 158], an implementation of the minimum-power distortionless response (MPDR) beamformer and widely employed in noise reduction. Fig. 2.2 depicts a block diagram of the GSC, which consists of three components: a filter-and-sum beamformer (FSB) steering a beam into the target direction, a blocking matrix blocking the speech component, and a data-dependent filter minimizing the output power and thereby suppressing residual noise components. The ability to block the speech component is essential to the GSC, as speech leakage through the blocking matrix may lead to partial speech cancellation. Employing the GSC for dereverberation, the blocking matrix should block the early (but not the late) reverberant-speech component. In [85] therefore a blocking matrix is used incorporating the relative early transfer functions (RETFs) of the speech source in order to jointly perform dereverberation and noise reduction. In the nested GSC [87], an inner GSC is employed for dereverberation and an outer GSC for noise reduction. In [157], we have proposed to integrate the GSC and MCLP in a parallel manner, and compared to the corresponding MCLP-GSC cascade, cf. also [112, 113].

A comparison of the block diagrams in Fig. 2.1 and 2.2 readily reveals the major difference between the two frameworks, which is due to their different objective. Where MCLP – designed for deconvolution – applies a simple delay to the microphone signals in the data-dependent filter path, the GSC instead – designed

for spatial filtering – applies a blocking matrix. On the one hand, regarding dereverberation, the need for a blocking matrix is certainly a drawback of the GSC as compared to MCLP, as its design requires prior knowledge. On the other hand, regarding noise reduction, the blocking matrix distinguishes the speech source from potential localized noise sources, which is not possible in MCLP. For the noiseless dereverberation task, we have shown in [156] that the MCLP and GSC framework theoretically lead to the mathematically equivalent results for stationary source signals. In practice, additional noise may always be present. In this paper therefore, using pre-whitened least squares (LS) filter estimates, we formally analyze and compare the behavior of both frameworks in case of noise, both in terms of dereverberation and noise reduction. The main intention is to provide a better understanding of the theoretical performance limitations of both frameworks depending on a number of boundary conditions, such as noise levels, filter length and number of microphones, which cannot be done by naive comparison. In our theoretical analysis, we assume *complete blocking* of the early reverberant-speech component in the GSC blocking matrix, which requires prior knowledge of the early part of the speech-source RIRs or the RETFs. We derive that if the number of microphones is sufficiently large, the GSC theoretically achieves complete coherent-noise cancellation if incoherent noise is absent, while MCLP cancels the late coherent-noise components only, as expected by design. Further, in case of *complete blocking*, the GSC performs equally well as MCLP in terms of dereverberation; theoretically achieving complete reverberation cancellation if incoherent noise is absent. These theoretical findings are confirmed by time domain simulations. In addition, in case of *incomplete blocking*, based on STFT domain simulations using estimated RETFs, we show that the GSC instead performs inferior to MCLP in terms of dereverberation.

In Sec. 2.2, the signal model for both frameworks is presented. In Sec. 2.3, the filter estimation is discussed. Sec. 2.4 and 2.5 proceed with the performance analysis of the MCLP and the GSC framework, respectively. A comparative summary of the two frameworks is presented in Sec. 2.6, followed by simulation results in Sec. 2.7.

2.2 Signal Model

In this section, we define the signal model for both MCLP and the GSC. For simplicity, we employ the same notation for those signals and filters that correspond in both frameworks, cf. Fig. 2.1 and 2.2. As outlined before, the major difference between both consists in the use of a prediction delay δ in MCLP (cf. Fig. 2.1) and a blocking matrix $\mathbf{B}$ in the GSC (cf. Fig. 2.2) in the data-dependent filter path. In addition, the GSC speech reference is typically

created by applying an FSB, whereas in MCLP a particular microphone signal is traditionally selected [88, 89, 91, 93–96, 99, 100, 102, 111–113]. Both cases are covered generically by the filter $\mathbf{g}$ (cf. Fig. 2.1 and 2.2). The signal model equivalently applies in the time domain and the STFT domain, where l respectively denotes the time or frame index. In case of the STFT domain, throughout the paper, the frequency sub-band index is omitted as we treat all frequency sub-bands independently. Subsequently, vectors are denoted by lower-case boldface letters, matrices by upper-case boldface letters, $\mathbf{I}^{L\times L}$ and $\mathbf{0}^{L_1\times L_2}$ denote identity and zero matrices with the (optional) superscript indicating their dimensions, $\mathbf{A}^*$, $\mathbf{A}^T$, $\mathbf{A}^H$, $\mathbf{A}^P$, and $\mathrm{E}\big[\mathbf{A}\big]$ denote the complex conjugate, the transpose, the complex conjugate transpose, the pseudoinverse and the expected value of a matrix $\mathbf{A}$, $\mathrm{Blkdiag}\big[\mathbf{A}_1, \ldots, \mathbf{A}_N\big]$ constructs a block-diagonal matrix from its arguments, and $\mathrm{Tplz}\big[\mathbf{a}, L\big]$ creates a Toeplitz matrix of L columns with the first column defined by the vector $\big(\mathbf{a}^T\ \mathbf{0}^{1\times(L-1)}\big)^T$.

The acoustic scenario is presented in Sec. 2.2.1, while in Sec. 2.2.2 the speech reference signal and its individual components are defined. In Sec. 2.2.3 and Sec. 2.2.4, the data-dependent filter input signal is discussed for MCLP and the GSC, respectively. In Sec. 2.2.5, the filter output and the enhanced signal are generically defined.

2.2.1 Acoustic Scenario

We assume an acoustic scenario comprising one speech source emitting the signal $s_1(l)$, and $N-1$ localized noise sources emitting the signals $s_n(l)$, $n = 2\ldots N$, in a reverberant environment with M microphones. The m^{th} microphone signal $y_m(l)$, $m = 1\ldots M$, consists of the reverberant-speech component, reverberant-noise components, referred to as coherent-noise components hereafter, as well as an incoherent-noise component (originating from spatially uncorrelated noise, e.g., sensor noise), i.e.

$$y_m(l) = \sum_{n=1}^{N} \underbrace{\sum_{l'=0}^{L_h-1} h_{n,m}^*(l')s_n(l-l')}_{x_{n,m}(l)} + v_m(l), \tag{2.1}$$

with $h_{n,m}(l')$ denoting the time-invariant (sub-band) RIR between the n^{th} source and the m^{th} microphone of length L_h (neglecting the dead time common to all RIRs), l' the tap index, $x_{n,m}(l)$ the reverberant components (reverberant-speech and coherent-noise components), and $v_m(l)$ the incoherent-noise component. Note that in the STFT case, the sub-band convolution model in (2.1) poses an approximation of the time-domain convolution [117], where the sub-band RIR

length L_h is roughly R_{STFT} times smaller than the corresponding time domain RIR length, with R_{STFT} denoting the hop size in the STFT analysis [117]. We define the stacked multi-microphone vector $\mathbf{y}(l) \in \mathbb{C}^{ML_y}$,

$$\mathbf{y}(l) = \Big(\mathbf{y}_1^T(l) \ \cdots \ \mathbf{y}_M^T(l)\Big)^T, \tag{2.2}$$

$$\mathbf{y}_m(l) = \Big(y_m(l) \ \cdots \ y_m(l-L_y+1)\Big)^T, \tag{2.3}$$

with L_y the number of samples/frames per microphone. With $\mathbf{x}_n(l)$ and $\mathbf{v}(l)$ defined in a similar manner as in (2.2), we obtain

$$\mathbf{y}(l) = \sum_{n=1}^{N} \mathbf{x}_n(l) + \mathbf{v}(l) = \mathbf{x}(l) + \mathbf{v}(l). \tag{2.4}$$

With the blockwise Toeplitz matrix $\mathbf{H} \in \mathbb{C}^{NL_s \times ML_y}$ and the stacked source-signal vector $\mathbf{s}(l) \in \mathbb{C}^{NL_s}$ defined by

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{1,1} & \cdots & \mathbf{H}_{1,M} \\ \vdots & & \vdots \\ \mathbf{H}_{N,1} & \cdots & \mathbf{H}_{N,M} \end{pmatrix} = \begin{pmatrix} \mathbf{H}_1 \\ \vdots \\ \mathbf{H}_N \end{pmatrix}, \tag{2.5}$$

$$\mathbf{H}_{n,m} = \mathrm{Tplz}\Big[\Big(h_{n,m}(0) \ \cdots \ h_{n,m}(L_h-1)\Big)^T, \ L_y\Big], \tag{2.6}$$

$$\mathbf{s}(l) = \Big(\mathbf{s}_1^T(l) \ \cdots \ \mathbf{s}_N^T(l)\Big)^T, \tag{2.7}$$

$$\mathbf{s}_n(l) = \Big(s_n(l) \ \cdots \ s_n(l-L_s+1)\Big)^T, \tag{2.8}$$

$$L_s = L_h + L_y - 1, \tag{2.9}$$

the vector $\mathbf{x}(l)$ can then be written as

$$\mathbf{x}(l) = \sum_{n=1}^{N} \mathbf{H}_n^H \mathbf{s}_n(l) = \mathbf{H}^H \mathbf{s}(l). \tag{2.10}$$

We assume $\mathbf{s}_n(l)$ and $\mathbf{v}(l)$ to be mutually uncorrelated, i.e. with the correlation matrices $\mathbf{\Psi}_{s_n}(l) = \mathrm{E}\big[\mathbf{s}_n(l)\mathbf{s}_n^H(l)\big]$ and $\mathbf{\Psi}_v(l)$ equivalently, using (2.4), (2.7), (2.10), we find

$$\mathbf{\Psi}_s(l) = \mathrm{Blkdiag}\Big[\mathbf{\Psi}_{s_1}(l), \ \ldots, \ \mathbf{\Psi}_{s_N}(l)\Big], \tag{2.11}$$

$$\mathbf{\Psi}_x(l) = \mathbf{H}^H \mathbf{\Psi}_s(l)\mathbf{H}, \tag{2.12}$$

$$\mathbf{\Psi}_y(l) = \mathbf{\Psi}_x(l) + \mathbf{\Psi}_v(l). \tag{2.13}$$

The matrices $\mathbf{\Psi}_{s_n}(l)$ are assumed to be invertible, such that $\mathbf{\Psi}_s^{-1}(l) = \text{Blkdiag}\left[\mathbf{\Psi}_{s_1}^{-1}(l),\ \ldots,\ \mathbf{\Psi}_{s_N}^{-1}(l)\right]$. Note that in the STFT domain, it is commonly assumed that $\mathrm{E}\left[s_1(l)s_1^*(l-l')\right] = 0$ for $l' \neq 0$ if the STFT hop size is sufficiently large, i.e. $\mathbf{\Psi}_{s_1}(l)$ becomes a diagonal matrix. Similar assumptions could be made for other source signals, but are not required in our analysis.

2.2.2 Speech Reference Signal

With the filter $\mathbf{g} \in \mathbb{C}^{ML_y}$, we define the speech reference signal $q(l)$ for both frameworks as

$$\begin{aligned} q(l) &= \mathbf{g}^H \mathbf{y}(l) \\ &= \sum_{n=1}^{N} \underbrace{(\mathbf{H}_n\mathbf{g})^H \mathbf{s}_n(l)}_{q_{s_n}(l)} + \underbrace{\mathbf{g}^H \mathbf{v}(l)}_{q_v(l)}, \end{aligned} \tag{2.14}$$

where $q_{s_n}(l)$ and $q_v(l)$ denote the individual source components of $q(l)$. Defining the parameter $d \ll L_h$ as the boundary between early and late reverberation, the reverberant-speech and coherent-noise components $q_{s_n}(l)$ may further be decomposed into early and late components $q_{s_n|e}(l)$ and $q_{s_n|\ell}(l)$, i.e.

$$q_{s_n}(l) = \underbrace{(\mathbf{C}_e\mathbf{H}_n\mathbf{g})^H \mathbf{s}_n(l)}_{q_{s_n|e}(l)} + \underbrace{(\mathbf{C}_\ell\mathbf{H}_n\mathbf{g})^H \mathbf{s}_n(l)}_{q_{s_n|\ell}(l)}, \tag{2.15}$$

with $\mathbf{C}_e \in \mathbb{C}^{L_s \times L_s}$ and its complement $\mathbf{C}_\ell$ defined as

$$\mathbf{C}_e = \begin{pmatrix} \mathbf{I}^{d\times d} & \mathbf{0}^{d\times(L_s-d)} \\ \mathbf{0}^{(L_s-d)\times d} & \mathbf{0}^{(L_s-d)\times(L_s-d)} \end{pmatrix}, \tag{2.16}$$

$$\mathbf{C}_\ell = \mathbf{I}^{L_s\times L_s} - \mathbf{C}_e. \tag{2.17}$$

For later derivations throughout Sec. 2.4 and Sec. 2.5, we note that $q_{s_n|\ell}(l)$ in (2.15) may alternatively be expressed as

$$q_{s_n|\ell}(l) = (\mathbf{C}_{d|\ell}\mathbf{H}_n\mathbf{g})^H \mathbf{s}_n(l-d), \tag{2.18}$$

where $\mathbf{C}_{d|\ell}$ is derived from $\mathbf{C}_\ell$ by shifting d rows upwards , i.e.

$$\mathbf{C}_{d|\ell} = \begin{pmatrix} \mathbf{0}^{(L_s-d)\times d} & \mathbf{I}^{(L_s-d)\times(L_s-d)} \\ \mathbf{0}^{d\times d} & \mathbf{0}^{d\times(L_s-d)} \end{pmatrix}. \tag{2.19}$$

Based on the above definitions, the (sub-band) impulse response (IR) relating $s_n(l)$ and $q_{s_n}(l)$ is graphically represented in Fig. 2.3. The parameter d can be controlled by design choices in the MCLP and the GSC framework, as shown in Sec. 2.2.3 and Sec. 2.2.4, respectively.

We now define the early reverberant-speech component $q_{s_1|e}(l)$ as the target component to be maintained, and the remaining late reverberant-speech component plus all noise components $q_{s_1|\ell}(l) + \sum_{n=2}^{N} q_{s_n}(l) + q_v(l)$ as the component to be canceled. Note that in a *different* acoustic scenario, e.g., with N speech sources instead of one speech source plus $N-1$ noise sources, the target component could be defined differently, e.g., by $\sum_{n=1}^{N} q_{s_n|e}(l)$.

2.2.3 MCLP Filter Input

In the MCLP framework, the filter input signal $\mathbf{u}(l) \in \mathbb{C}^{ML_w}$ is a delayed version of the microphone signals $\mathbf{y}(l)$. The prediction delay δ is chosen as $\delta = d$, i.e.

$$\boxed{\begin{aligned} \mathbf{u}(l) &= \mathbf{y}(l-d) \\ &= \mathbf{H}^H \mathbf{s}(l-d) + \mathbf{v}(l-d). \end{aligned}} \tag{2.20}$$

Hence, the length L_y in (2.9) equals the length L_w of a single filter channel of the data-dependent filter $\mathbf{w}$, i.e.

$$L_y = L_w. \tag{2.21}$$

With (2.9), (2.21), we determine that $\mathbf{H} \in \mathbb{C}^{NL_s \times ML_y}$ is a fat matrix if the MCLP filter length L_w satisfies the condition

$$\boxed{L_w \geq \frac{N(L_h-1)}{M-N},} \tag{2.22}$$

which obviously requires $M > N$ microphones. If L_w is chosen according to (2.22) and the (sub-band) RIRs meet the MINT requirements (i.e. no common zeros), which is commonly assumed [71, 89], then the system is invertible and $\mathbf{H}$ has full row rank [71]. As it is crucial for our derivations in Sec. 2.4.2, full row rank of $\mathbf{H}$ is assumed in the remainder. Since our simulation results in Sec. 2.7 support our theoretical conclusions in Sec. 2.4, we consider this assumption to be reasonable.

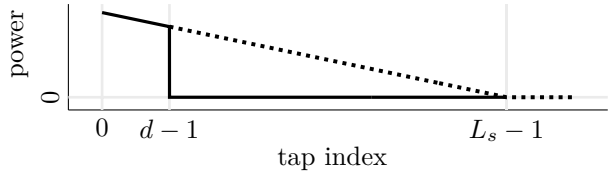


Figure 2.3: Schematic of the (sub-band) IR relating $s_n(l)$ and $q_{s_n}(l)$, separated in early part $\mathbf{C}_e\mathbf{H}_n\mathbf{g}$ [——] applied to $\mathbf{s}_n(l)$ and late part $\mathbf{C}_{d|\ell}\mathbf{H}_n\mathbf{g}$ [·····] applied to $\mathbf{s}_n(l-d)$.

2.2.4 GSC Filter Input

In the GSC framework, the filter input signal $\mathbf{u}(l) \in \mathbb{C}^{(M-1)L_w}$ is constructed by applying a blocking matrix $\mathbf{B} \in \mathbb{C}^{ML_y \times (M-1)L_w}$ to the microphone signal, i.e.

$$\boxed{\begin{aligned}\mathbf{u}(l) &= \mathbf{B}^H\mathbf{y}(l) \\ &= (\mathbf{H}\mathbf{B})^H\mathbf{s}(l) + \mathbf{B}^H\mathbf{v}(l),\end{aligned}} \tag{2.23}$$

where L_w again describes the length of a single filter channel of the filter $\mathbf{w}$. Eq. (2.23) is the GSC counterpart to (2.20) for MCLP. We intend to completely block the component in $\mathbf{x}_1(l) = \mathbf{H}_1^H\mathbf{s}_1(l)$ depending on the early part of $\mathbf{H}_1$. A matrix $\mathbf{B}$ satisfying[1] this condition may be defined in the following manner,

$$\mathbf{B} = \begin{pmatrix} -\mathbf{H}_{1,2|e} & \cdots & -\mathbf{H}_{1,M|e} \\ \text{Blkdiag}\left[\mathbf{H}_{1,1|e}, & \ldots, & \mathbf{H}_{1,1|e}\right] \end{pmatrix}, \tag{2.24}$$

$$\mathbf{H}_{1,m|e} = \text{Tplz}\left[\begin{pmatrix} h_{1,m}(0) & \cdots & h_{1,m}(L_b-1) \end{pmatrix}^T,\ L_w\right], \tag{2.25}$$

where L_b denotes the length of the blocking filters, such that in the GSC, we find for L_y in (2.9),

$$L_y = L_b + L_w - 1. \tag{2.26}$$

The definition in (2.24)–(2.25) ensures that all components corresponding to the first $d \geq L_b$ taps of the speech-source (sub-band) RIRs $h_{1,m}(l')$ are nullified, where the case $d > L_b$ occurs if the first L_b taps of $h_{1,m}(l')$ are succeeded by

[1] Many definitions of $\mathbf{B}$ achieving complete blocking exist. In the STFT domain, for $L_b = 1$, the blocking matrix may also be defined using RETFs [85]. If the target component is instead defined as $\sum_{n=1}^{N} q_{s_n|e}(l)$ as in the *different* acoustic scenario mentioned in Sec. 2.2.2, then also the definition of $\mathbf{B}$ needs to change accordingly.

one or more zeros. The product $\mathbf{HB}$ takes the form

$$\mathbf{HB} = \begin{pmatrix} \mathbf{0}^{d\times(M-1)L_w} \\ \mathbf{H}_B \end{pmatrix}, \tag{2.27}$$

where, using (2.9), (2.26), $\mathbf{H}_B \in \mathbb{C}^{NL_s-d\times(M-1)L_w}$ is a fat matrix if the GSC filter length L_w satisfies the condition

$$\boxed{L_w \geq \frac{N(L_h-2)+(N-1)d}{M-N-1},} \tag{2.28}$$

which obviously requires $M > N+1$ microphones. If L_w is chosen according to (2.28) and the $M-1$ (sub-band) impulse responses in $\mathbf{HB}$ meet the MINT requirements, i.e. the nullity of $(\mathbf{HB})^H$ does not exceed d, then $\mathbf{H}_B$ has full row rank according to the rank-nullity theorem [159]. As it is crucial for our derivations in Sec. 2.5.2.1, full row rank of $\mathbf{H}_B$ is assumed in the remainder. Since our simulation results in Sec. 2.7 support our theoretical conclusions in Sec. 2.5, we consider this assumption to be reasonable. Comparing L_w for the GSC and MCLP in (2.28) and (2.22), respectively, we find that the GSC requires longer filters. Note however that the GSC employs one filter channel less.

2.2.5 Enhanced signal

For both frameworks, the filter output signal $z(l)$ and the enhanced signal $e(l)$ are given by

$$z(l) = \mathbf{w}^H\mathbf{u}(l), \tag{2.29}$$

$$e(l) = q(l) - z(l), \tag{2.30}$$

with $\mathbf{u}(l)$ given by (2.20) in MCLP or (2.23) in the GSC. For MCLP, $z(l)$ is the linear prediction of $q(l)$, and $e(l)$ accordingly the linear prediction residual. The estimation of the filter $\mathbf{w}$ is discussed in Sec. 2.3.

2.3 Filter Estimation

We now present the pre-whitened LS estimate of $\mathbf{w}$ in Sec. 2.3.1 and discuss the choice of the pre-whitening matrix in Sec. 2.3.2. In Sec. 2.3.3, we present the corresponding Wiener solution, which is then used in the theoretical analysis in the subsequent Sec. 2.4 and Sec. 2.5.

2.3.1 Pre-whitened LS

With $l = 0 \dots L_{obs} - 1$ and L_{obs} denoting the number of observations used in the filter estimation, let $\underline{\mathbf{q}} \in \mathbb{C}^{1 \times L_{obs}}$ and $\underline{\mathbf{U}} \in \mathbb{C}^{ML_w \times L_{obs}}$ denote correspondingly stacked versions of $q(l)$ and $\mathbf{u}(l)$, i.e.

$$\underline{\mathbf{q}} = \Big(q(0) \ \cdots \ q(L_{obs}-1) \Big), \tag{2.31}$$

$$\underline{\mathbf{U}} = \Big(\mathbf{u}(0) \ \cdots \ \mathbf{u}(L_{obs}-1) \Big), \tag{2.32}$$

and let $\underline{\mathbf{Y}}$, $\underline{\mathbf{S}}$, $\underline{\mathbf{S}}_n$, and $\underline{\mathbf{V}}$ be defined equivalently to $\underline{\mathbf{U}}$ in (2.32). Further, let $\underline{\mathbf{\Omega}}^{-1/2} \in \mathbb{C}^{L_{obs} \times L_{obs}}$ denote some pre-whitening matrix with $\underline{\mathbf{\Omega}} = \underline{\mathbf{\Omega}}^{H/2}\underline{\mathbf{\Omega}}^{1/2}$ to be defined explicitly in Sec. 2.3.2. Let $\underline{\tilde{\mathbf{U}}}$ and $\underline{\tilde{\mathbf{q}}}$ denote correspondingly pre-whitened versions of $\underline{\mathbf{U}}$ and $\underline{\mathbf{q}}$, i.e.

$$\underline{\tilde{\mathbf{q}}} = \underline{\mathbf{q}\mathbf{\Omega}}^{-1/2} = \Big(\tilde{q}(0) \ \cdots \ \tilde{q}(L_{obs}-1) \Big)^T, \tag{2.33}$$

$$\underline{\tilde{\mathbf{U}}} = \underline{\mathbf{U}\mathbf{\Omega}}^{-1/2} = \Big(\tilde{\mathbf{u}}(0) \ \cdots \ \tilde{\mathbf{u}}(L_{obs}-1) \Big), \tag{2.34}$$

and let $\underline{\tilde{\mathbf{Y}}}$, $\underline{\tilde{\mathbf{S}}}$, $\underline{\tilde{\mathbf{S}}}_n$, and $\underline{\tilde{\mathbf{V}}}$ as well as their respective column vectors $\tilde{\mathbf{y}}(l)$, $\tilde{\mathbf{s}}(l)$, $\tilde{\mathbf{s}}_n(l)$, and $\tilde{\mathbf{v}}(l)$ be defined equivalently to (2.34). Based on these definitions, the pre-whitened data $\tilde{q}(l)$ and $\tilde{\mathbf{u}}(l)$ may be expressed equivalently to $q(l)$ in (2.14) and $\mathbf{u}(l)$ in (2.20), (2.23), where $\tilde{\mathbf{y}}(l)$, $\tilde{\mathbf{s}}(l)$, and $\tilde{\mathbf{v}}(l)$ replace $\mathbf{y}(l)$, $\mathbf{s}(l)$, and $\mathbf{v}(l)$, respectively. Based on (2.29)–(2.30), (2.33)–(2.34), we generically define the LS cost function,

$$f_{LS}(\mathbf{w}) = \big\| \underbrace{\underline{\tilde{\mathbf{q}}} - \mathbf{w}^H \underline{\tilde{\mathbf{U}}}}_{\underline{\tilde{\mathbf{e}}} = \Big(\tilde{e}(0) \ \cdots \ \tilde{e}(L_{obs}-1) \Big)} \big\|_2^2, \tag{2.35}$$

leading to the the LS filter estimate $\hat{\mathbf{w}}_{LS}$,

$$\begin{aligned} \hat{\mathbf{w}}_{LS} &= \arg\min_{\mathbf{w}} f_{LS}(\mathbf{w}) \\ &= \big(\underline{\tilde{\mathbf{U}}\tilde{\mathbf{U}}}^H \big)^{-1} \underline{\tilde{\mathbf{U}}\tilde{\mathbf{q}}}^H \\ &= \big(\underline{\mathbf{U}\mathbf{\Omega}}^{-1} \underline{\mathbf{U}}^H \big)^{-1} \underline{\mathbf{U}\mathbf{\Omega}}^{-1} \underline{\mathbf{q}}^H. \end{aligned} \tag{2.36}$$

Note that with (2.33)–(2.34), $\hat{\mathbf{w}}_{LS}$ in (2.36) may alternatively be written as

$$\hat{\mathbf{w}}_{LS} = \left(\sum_{l=0}^{L_{obs}-1} \tilde{\mathbf{u}}(l) \tilde{\mathbf{u}}^H(l) \right)^{-1} \sum_{l=0}^{L_{obs}-1} \tilde{\mathbf{u}}(l) \tilde{q}^*(l). \tag{2.37}$$

2.3.2 Choice of pre-whitening matrix

Despite the prediction delay in MCLP and the blocking matrix in the GSC, due to the coloration of $s_1(l)$, we generally have $\mathrm{E}[\mathbf{u}(l)q^*_{s_1|e}(l)] \neq \mathbf{0}$, cf. $q^*_{s_1|e}(l)$ in (2.15) and $\mathbf{u}(l)$ in (2.20) and (2.23)–(2.25), respectively. Hence, the estimator in (2.36)–(2.37) is biased for the case $\underline{\mathbf{\Omega}} = \mathbf{I}$. In order to mitigate an estimation bias, we may use the pre-whitening matrix $\underline{\mathbf{\Omega}}^{-1/2}$. In particular, for the two cases $\underline{\boldsymbol{\varepsilon}} \in \{\underline{\mathbf{q}}_{s_1|e}, \underline{\mathbf{s}}_1\}$ with $\underline{\mathbf{q}}_{s_1|e}$ and $\underline{\mathbf{s}}_1$ defined equivalently to (2.31), an unbiased estimate is achieved if

$$\underline{\mathbf{\Omega}} = \underline{\mathbf{\Psi}}_{\varepsilon}, \tag{2.38}$$

where $\underline{\mathbf{\Psi}}_{\varepsilon} = \mathrm{E}\big[\underline{\boldsymbol{\varepsilon}}^H\underline{\boldsymbol{\varepsilon}}\big]$. Eq. (2.36) then corresponds to the generalized LS estimator of $\mathbf{w}$ for the data model $\underline{\mathbf{q}} = \mathbf{w}^H\underline{\mathbf{U}} + \underline{\boldsymbol{\varepsilon}}$, where $\underline{\boldsymbol{\varepsilon}}$ resembles the observation noise with $\mathrm{E}\big[\underline{\mathbf{U}}\underline{\boldsymbol{\varepsilon}}^H\big] \neq \mathbf{0}$ but $\mathrm{E}\big[\underline{\mathbf{U}}\underline{\mathbf{\Psi}}_{\varepsilon}^{-1}\underline{\boldsymbol{\varepsilon}}^H\big] = \mathbf{0}$.

In the time domain, $\underline{\mathbf{\Psi}}_{q_{s_1|e}}$ and $\underline{\mathbf{\Psi}}_{s_1}$ are generally non-diagonal. The choice $\underline{\mathbf{\Omega}} = \underline{\mathbf{\Psi}}_{s_1}$ here corresponds to the pre-whitening paradigms proposed in [88] for MCLP and [155] for the GSC. In Sec. 2.7.2.1, we present time domain simulations for both the unbiased and the biased case with $\underline{\mathbf{\Omega}} = \mathbf{I}$.

In the STFT domain, $\underline{\mathbf{\Psi}}_{q_{s_1|e}}$ may be modeled as a matrix with $2d-1$ non-zero diagonals and $\underline{\mathbf{\Psi}}_{s_1}$ may be modeled as a fully diagonal matrix, cf. Sec. 2.2.1, where the l^{th} diagonal element of $\underline{\mathbf{\Psi}}_{s_1}$ corresponds to the PSD $\psi_{s_1}(l) = \mathrm{E}\big[|s_1(l)|^2\big]$. In this case, with (2.31)–(2.32), $\hat{\mathbf{w}}_{LS}$ in (2.36) may therefore be written as

$$\hat{\mathbf{w}}_{LS} = \left(\sum_{l=0}^{L_{obs}-1} \frac{\mathbf{u}(l)\mathbf{u}^H(l)}{\psi_{s_1}(l)}\right)^{-1} \sum_{l=0}^{L_{obs}-1} \frac{\mathbf{u}(l)q^*(l)}{\psi_{s_1}(l)}, \tag{2.39}$$

i.e. each frame $q(l)$ and $\mathbf{u}(l)$ is weighted by the inverse of $\psi_{s_1}(l)$, which, in case of MCLP, corresponds to the WPE criterion [94, 95]. Note that $\psi_{s_1}(l)$ varies over time for non-stationary source signals. In Sec. 2.7.2.2, we present STFT-domain simulations for $d = 1$ and $\underline{\mathbf{\Omega}} = \underline{\mathbf{\Psi}}_{q_{s_1|e}} \propto \underline{\mathbf{\Psi}}_{s_1}$. Herein, prior to estimating $\mathbf{w}$ according to (2.36), the PSDs $\psi_{q_{s_1|e}}(l)$ on the diagonal of $\underline{\mathbf{\Psi}}_{q_{s_1|e}}$ are estimated per frame l as proposed in [65, 157]. To this end, it is assumed that the spatial coherence matrix of the late reverberant component may be modeled as diffuse, which allows to obtain the PSD estimates by means of the generalized eigenvalue decomposition (GEVD) of the spatial correlation matrix of the microphone signals and the diffuse coherence matrix, cf. Sec. 2.7.1.3.

Note that in the *different* acoustic scenario mentioned in Sec. 2.2.2 with the target component defined as $\sum_{n=1}^{N} q_{s_n|e}(l)$ instead of $q_{s_1|e}(l)$, in order to achieve an unbiased filter estimate, one has to change $\underline{\mathbf{\Omega}}$ accordingly, e.g., using $\underline{\boldsymbol{\varepsilon}} = \sum_{n=1}^{N} \underline{\mathbf{q}}_{s_n|e}$ in (2.38).

2.3.3 Convergence to Wiener filter solution

For the purpose of the analysis in Sec. 2.4 and Sec. 2.5, we assume wide-sense stationarity for the pre-whitened signals $\tilde{\mathbf{u}}(l)$ and $\tilde{q}(l)$, i.e. their statistics are independent of l. Then, for $L_{obs} \to \infty$, the estimate $\hat{\mathbf{w}}_{LS}$ in (2.37) converges to the Wiener filter solution $\hat{\mathbf{w}}_{WF}$,

$$\hat{\mathbf{w}}_{WF} = \boldsymbol{\Psi}_{\tilde{u}}^{P} \boldsymbol{\psi}_{\tilde{u}\tilde{q}}, \tag{2.40}$$

with $\boldsymbol{\Psi}_{\tilde{u}} = \mathrm{E}\big[\tilde{\mathbf{u}}(l)\tilde{\mathbf{u}}^{H}(l)\big]$ and $\boldsymbol{\psi}_{\tilde{u}\tilde{q}} = \mathrm{E}\big[\tilde{\mathbf{u}}(l)\tilde{q}^{*}(l)\big]$. Here, the inverse in (2.37) is replaced by the pseudoinverse, as in the GSC, $\boldsymbol{\Psi}_{\tilde{u}}$ becomes rank-deficient in absence of incoherent noise and in case of complete blocking, i.e. if (2.27) holds, cf. Sec. 2.5.2.1 and Appendix A.1.1.

2.4 MCLP Analysis

For $\mathbf{w} = \hat{\mathbf{w}}_{WF}$, we now derive the MCLP filter output signal $z(l)$ in Sec. 2.4.1 and then derive and discuss the enhanced signal $e(l)$ under different noise conditions in Sec. 2.4.2.

2.4.1 MCLP Filter Output

Using (2.14), (2.20), and noting that $\mathrm{E}\big[\tilde{\mathbf{y}}(l-d)\tilde{\mathbf{y}}^{H}(l-d)\big] = \mathrm{E}\big[\tilde{\mathbf{y}}(l)\tilde{\mathbf{y}}^{H}(l)\big]$, the terms $\boldsymbol{\Psi}_{\tilde{u}}$ and $\boldsymbol{\psi}_{\tilde{u}\tilde{q}}$ in (2.40) become

$$\boldsymbol{\Psi}_{\tilde{u}} = \boldsymbol{\Psi}_{\tilde{y}}, \tag{2.41}$$

$$\boldsymbol{\psi}_{\tilde{u}\tilde{q}} = \boldsymbol{\Psi}_{\tilde{y}|d}\mathbf{g}, \tag{2.42}$$

$$\boldsymbol{\Psi}_{\tilde{y}|d} = \mathrm{E}\big[\tilde{\mathbf{y}}(l-d)\tilde{\mathbf{y}}^{H}(l)\big]. \tag{2.43}$$

Inserting (2.20) in (2.29) and substituting $\mathbf{w}$ by $\hat{\mathbf{w}}_{WF}$ in (2.40), we obtain for the filter output signal,

$$z(l) = (\boldsymbol{\Psi}_{\tilde{y}}^{P}\boldsymbol{\Psi}_{\tilde{y}|d}\mathbf{g})^{H}\mathbf{y}(l-d). \tag{2.44}$$

Let the shifted correlation matrices $\boldsymbol{\Psi}_{\tilde{s}_n|d}$, $\boldsymbol{\Psi}_{\tilde{s}|d}$, and $\boldsymbol{\Psi}_{\tilde{v}|d}$ be defined equivalently to $\boldsymbol{\Psi}_{\tilde{y}|d}$ in (2.43), with relations equivalent to (2.11)–(2.13). We now introduce the following relation between $\boldsymbol{\Psi}_{\tilde{s}_n|d}$ and $\boldsymbol{\Psi}_{\tilde{s}_n}$, which is used in the subsequent derivations in Sec. 2.4.2. For this, note that we can interpret $\boldsymbol{\Psi}_{\tilde{s}_n|d}$ and $\boldsymbol{\Psi}_{\tilde{s}_n}$ as different submatrices of a larger correlation matrix, as shown in Fig. 2.4.



(a) correlation matrix $\mathbf{\Psi}_{\tilde{s}_n}$ (b) correlation matrix $\mathbf{\Psi}_{\tilde{s}_n|d}$

Figure 2.4: Schematic of the correlation matrices $\mathbf{\Psi}_{\tilde{s}_n}$ and $\mathbf{\Psi}_{\tilde{s}_n|d}$ as different submatrices of a larger correlation matrix.

The submatrix defining $\mathbf{\Psi}_{\tilde{s}_n|d}$ is shifted left by d columns as compared to the submatrix defining $\mathbf{\Psi}_{\tilde{s}_n}$. Noting that the autocorrelation width of $\tilde{s}_n(l)$ is typically much smaller than L_h in both the time and STFT domain, we assume that the autocorrelation of $\tilde{s}_n(l)$ is zero for lags greater than $L_s - d$, where $L_s - d \geq {}^{M}/_{M-N}(L_h - 1) - d$ and $d \ll L_h$, cf. (2.9), (2.21)–(2.22). Using (2.16), (2.19), we then express $\mathbf{\Psi}_{\tilde{s}_n|d}$ in terms of $\mathbf{\Psi}_{\tilde{s}_n}$ by

$$\mathbf{\Psi}_{\tilde{s}_n|d} = \mathbf{\Psi}_{\tilde{s}_n}\mathbf{C}_{d|\ell} + \mathbf{C}_{d|\ell}\mathbf{\Psi}_{\tilde{s}_n}\mathbf{C}_e. \tag{2.45}$$

The product $\mathbf{\Psi}_{\tilde{s}_n}\mathbf{C}_{d|\ell}$ shifts the elements in $\mathbf{\Psi}_{\tilde{s}_n}$ right by d columns. The product $\mathbf{C}_{d|\ell}\mathbf{\Psi}_{\tilde{s}_n}\mathbf{C}_e$ replaces the resulting zero columns by the first d columns of $\mathbf{\Psi}_{\tilde{s}_n}$ shifted up by d rows.

2.4.2 MCLP Enhancement

We now analyze the behavior of MCLP considering two scenarios: absence and presence of incoherent noise.

2.4.2.1 Absence of Incoherent Noise

The absence of incoherent noise corresponds to $\mathbf{v}(l) = \mathbf{0}$, i.e. $\mathbf{y}(l) = \mathbf{x}(l)$. In this case, using (2.10) and relations equivalent to (2.11)–(2.13), the individual terms in (2.44) are equal to $\mathbf{y}(l-d) = \mathbf{H}^H\mathbf{s}(l-d)$, $\mathbf{\Psi}_{\tilde{y}} = \mathbf{H}^H\mathbf{\Psi}_{\tilde{s}}\mathbf{H}$, and $\mathbf{\Psi}_{\tilde{y}|d} = \mathbf{H}^H\mathbf{\Psi}_{\tilde{s}|d}\mathbf{H}$. Inserting these in (2.44) and noting that $\mathbf{H}^P = \mathbf{H}^H(\mathbf{H}\mathbf{H}^H)^{-1}$ and hence $\mathbf{H}\mathbf{H}^P = \mathbf{I}$ since $\mathbf{H}$ is assumed to have full row rank yields

$$z(l) = (\mathbf{\Psi}_{\tilde{s}}^P\mathbf{\Psi}_{\tilde{s}|d}\mathbf{H}\mathbf{g})^H\mathbf{s}(l-d), \tag{2.46}$$

which, using (2.11), (2.45), (2.14)–(2.15), may be written as

$$z(l) = \sum_{n=1}^{N}\left(q_{s_n|\ell}(l) + \boldsymbol{\vartheta}_{n|d}^H\mathbf{s}_n(l-d)\right), \tag{2.47}$$

with $\boldsymbol{\vartheta}_{n|d} = \mathbf{\Psi}_{\tilde{s}_n}^{-1}\mathbf{C}_{d|\ell}\mathbf{\Psi}_{\tilde{s}_n}\mathbf{C}_e\mathbf{H}_n\mathbf{g}.$ (2.48)

As apparent from (2.47)–(2.48), all reverberant source components are treated *mutually independently* and *equally*. This holds as long as (2.22) is satisfied and $\mathbf{H}$ has full row rank. Inserting (2.47) into (2.30) yields the MCLP output signal,

$$\boxed{e(l) = \sum_{n=1}^{N} \Big(\underbrace{q_{s_n|\mathrm{e}}(l) - \boldsymbol{\vartheta}_{n|d}^{H} \mathbf{s}_n(l-d)}_{e_{s_n}(l)} \Big).} \tag{2.49}$$

From (2.49), we observe that $e(l)$ equals the sum of the early components $q_{s_n|\mathrm{e}}(l)$ and a (potential) bias term $-\boldsymbol{\vartheta}_{n|d}^{H}\mathbf{s}_n(l-d)$ per source, with $\boldsymbol{\vartheta}_{n|d} \in \mathbb{C}^{L_s}$ and $L_s = L_h + L_w - 1$ according to (2.9), (2.26). Therefore, as only the late components $q_{s_n|\ell}(l)$ are canceled, the MCLP framework suits best in the *different* acoustic scenario mentioned in Sec. 2.2.2 with the target component defined as $\sum_{n=1}^{N} q_{s_n|\mathrm{e}}(l)$ instead of $q_{s_1|\mathrm{e}}(l)$. Combining (2.15) and (2.18), we can compare the individual components $e_{s_n}(l)$ in (2.49) to

$$q_{s_n}(l) = q_{s_n|\mathrm{e}}(l) + (\mathbf{C}_{d|\ell}\mathbf{H}_n\mathbf{g})^{H}\mathbf{s}_n(l-d), \tag{2.50}$$

i.e. the bias term replaces the late component $q_{s_n|\ell}(l) = (\mathbf{C}_{d|\ell}\mathbf{H}_n\mathbf{g})^{H}\mathbf{s}_n(l-d)$. Similarly as for the (sub-band) IR $\mathbf{H}_n\mathbf{g}$ relating $s_n(l)$ and $q_{s_n}(l)$ in Fig. 2.3, we visualize the (sub-band) IR relating $s_n(l)$ and $e_{s_n}(l)$, composed of the early part $\mathbf{C}_e\mathbf{H}_n\mathbf{g}$ and the bias part $-\boldsymbol{\vartheta}_{n|d}$, in Fig. 2.5. In the following we interpret the bias term in more detail, which has also partly been done in our previous work [156]. Firstly, from $\mathbf{C}_e\mathbf{H}_n\mathbf{g}$ in (2.48), we observe that the bias term $-\boldsymbol{\vartheta}_{n|d}^{H}\mathbf{s}_n(l-d)$ depends on the first d taps of $\mathbf{H}_n\mathbf{g}$ only, i.e. on its early part, but *not* its late part. Secondly, we note that $\boldsymbol{\vartheta}_{n|d}$ depends on the correlation matrix $\boldsymbol{\Psi}_{\tilde{s}_n}$ of the pre-whitened version $\tilde{\mathbf{s}}_n(l)$ of $\mathbf{s}_n(l)$, cf. Sec. 2.3.1. We can hence argue that for $\underline{\boldsymbol{\Omega}} = \underline{\boldsymbol{\Psi}}_{q_{s_1|e}}$ as defined in (2.38), with $q_{s_1|e}(l) = (\mathbf{C}_e\mathbf{H}_1\mathbf{g})^{H}\mathbf{s}_1(l)$, cf. (2.15), the coloration of the pre-whitened speech-source signal $\tilde{\mathbf{s}}_1(l)$ is inverse to the filter $\mathbf{C}_e\mathbf{H}_1\mathbf{g}$, such that only the first element of the vector $\boldsymbol{\Psi}_{\tilde{s}_1}\mathbf{C}_e\mathbf{H}_1\mathbf{g}$ is non-zero. Similarly, we can argue that for $\underline{\boldsymbol{\Omega}} = \underline{\boldsymbol{\Psi}}_{s_1}$, the matrix $\boldsymbol{\Psi}_{\tilde{s}_1}$ becomes diagonal, such that only the first d elements of $\boldsymbol{\Psi}_{\tilde{s}_1}\mathbf{C}_e\mathbf{H}_1\mathbf{g}$ are non-zero. In both cases, with $\mathbf{C}_{d|\ell}$ as in (2.19), we find $\mathbf{C}_{d|\ell}\boldsymbol{\Psi}_{\tilde{s}_1}\mathbf{C}_e\mathbf{H}_1\mathbf{g} = \mathbf{0}$, and therefore $\boldsymbol{\vartheta}_{1|d} = \mathbf{0}$ in (2.48) and finally $e_{s_1}(l) = q_{s_1|e}(l)$ in (2.49). Hence, the estimator is indeed unbiased for $\underline{\boldsymbol{\Omega}} \in \{\underline{\boldsymbol{\Psi}}_{q_{s_1|e}}, \underline{\boldsymbol{\Psi}}_{s_1}\}$, as anticipated in Sec. 2.3.

Note that the remaining early components may still be biased, i.e. $\boldsymbol{\vartheta}_{n|d} \neq \mathbf{0}$ for $n \neq 1$. In general, for $\underline{\boldsymbol{\Omega}} = \mathbf{I}$, the term $\boldsymbol{\vartheta}_{n|d}^{H}\mathbf{s}_n(l-d)$ in (2.49) represents a (delayed) linear prediction component of $q_{s_n|e}(l)$, i.e. the output signal component $e_{s_n}(l)$ may be understood as a (partially) whitened version of $q_{s_n|e}(l)$. This effect is also known as excessive whitening [89].

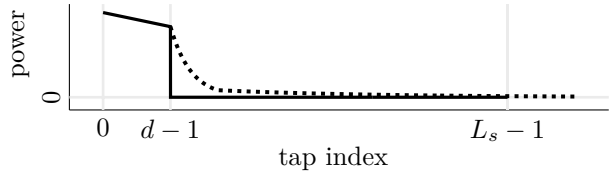


Figure 2.5: Schematic of the (sub-band) IR relating $s_n(l)$ and $e_{s_n}(l)$, separated in early part $\mathbf{C}_e\mathbf{H}_n\mathbf{g}$ [——] applied to $\mathbf{s}_n(l)$ and bias part $-\boldsymbol{\vartheta}_{n|d}$ [·····] applied to $\mathbf{s}_n(l-d)$.

2.4.2.2 Presence of Incoherent Noise

If additional incoherent noise $\mathbf{v}(l) \neq \mathbf{0}$ is present, the pseudoinverse of the sum $\boldsymbol{\Psi}_{\tilde{y}} = \boldsymbol{\Psi}_{\tilde{x}} + \boldsymbol{\Psi}_{\tilde{v}}$ in the filter $\boldsymbol{\Psi}_{\tilde{y}}^{P}\boldsymbol{\Psi}_{\tilde{y}|d}\mathbf{g}$ in (2.44) cannot be decomposed into its individual components, such that further simplification of (2.44) is not possible. In this more general case, MCLP cancels the linear prediction of the sum of $\mathbf{g}^H\mathbf{x}(l)$ and $\mathbf{g}^H\mathbf{v}(l)$. Noting that the incoherent noise acts as M additional independent sources, we find that the condition (2.22) for complete linear prediction in the MCLP framework, where M is required to exceed the number of independent sources N, cannot be fulfilled, resulting in decreased performance.

2.5 GSC Analysis

Similarly to Sec. 2.4, for $\mathbf{w} = \hat{\mathbf{w}}_{WF}$, we now derive the GSC filter output signal $z(l)$ in Sec. 2.5.1 and then derive and discuss the enhanced signal $e(l)$ under different noise conditions in Sec. 2.5.2.

2.5.1 GSC Filter Output

Following a derivation similar to Sec. 2.4.1, using (2.14) and (2.23), $\boldsymbol{\Psi}_{\tilde{u}}$ and $\boldsymbol{\psi}_{\tilde{u}\tilde{q}}$ in (2.40) can be written as

$$\boldsymbol{\Psi}_{\tilde{u}} = \mathbf{B}^H\boldsymbol{\Psi}_{\tilde{y}}\mathbf{B}, \tag{2.51}$$

$$\boldsymbol{\psi}_{\tilde{u}\tilde{q}} = \mathbf{B}^H\boldsymbol{\Psi}_{\tilde{y}}\mathbf{g}. \tag{2.52}$$

Inserting (2.23) in (2.29) and substituting $\mathbf{w}$ by $\hat{\mathbf{w}}_{WF}$ in (2.40), we obtain for the filter output signal,

$$z(l) = \left(\mathbf{B}(\mathbf{B}^H\boldsymbol{\Psi}_{\tilde{y}}\mathbf{B})^{P}\mathbf{B}^H\boldsymbol{\Psi}_{\tilde{y}}\mathbf{g}\right)^H\mathbf{y}(l). \tag{2.53}$$

2.5.2 GSC Enhancement

Similarly to Sec. 2.4.2, we now analyze the behavior of the GSC, again considering two scenarios: absence and presence of incoherent noise.

2.5.2.1 Absence of Incoherent Noise

For the case $\mathbf{v}(l) = \mathbf{0}$, i.e. $\mathbf{y}(l) = \mathbf{x}(l)$, using (2.10) and relations equivalent to (2.11)–(2.13), the individual terms in (2.53) are equal to $\mathbf{y}(l) = \mathbf{H}^H\mathbf{s}(l)$ and $\mathbf{\Psi}_{\tilde{y}} = \mathbf{H}^H\mathbf{\Psi}_{\tilde{s}}\mathbf{H}$. Inserting in (2.53) yields

$$z(l) = \Big(\mathbf{HB}\big((\mathbf{HB})^H\mathbf{\Psi}_{\tilde{s}}\mathbf{HB}\big)^P(\mathbf{HB})^H\mathbf{\Psi}_{\tilde{s}}\mathbf{Hg}\Big)^H\mathbf{s}(l). \tag{2.54}$$

In Appendix A.1.1, assuming complete blocking such that (2.27) holds, it is shown that (2.54) can be reformulated as

$$\boxed{\begin{aligned} z(l) &= q_{s_1|\ell}(l) + \boldsymbol{\vartheta}_{1|d}^H\mathbf{s}_1(l-d) + \sum_{n=2}^{N} q_{s_n}(l), && (2.55)\\ \text{with}\quad \boldsymbol{\vartheta}_{1|d} &= (\mathbf{C}_{d|\ell}\mathbf{\Psi}_{\tilde{s}_1}\mathbf{C}_{d|\ell}{}^H)^P\mathbf{C}_{d|\ell}\mathbf{\Psi}_{\tilde{s}_1}\mathbf{C}_e\mathbf{H}_1\mathbf{g}. && (2.56)\end{aligned}}$$

Inserting (2.55) into (2.30) yields the GSC output signal,

$$\boxed{e(l) = q_{s_1|e}(l) - \boldsymbol{\vartheta}_{1|d}^H\mathbf{s}_1(l-d).} \tag{2.57}$$

Eqs. (2.55)–(2.57) form the GSC counterpart to (2.47)–(2.49) for MCLP. From (2.57), we observe that $e(l)$ consists of two terms: the target component $q_{s_1|e}(l)$ and a bias term $-\boldsymbol{\vartheta}_{1|d}^H\mathbf{s}_1(l-d)$. This implies that not only the late reverberant-speech component, but also the coherent-noise components are completely canceled in the GSC, which is in contrast to MCLP, where only the late, but not the early coherent-noise components could be canceled. In Appendix A.1.2, it is shown that $\boldsymbol{\vartheta}_{1|d}$ in (2.56) for the GSC is indeed equal to $\boldsymbol{\vartheta}_{1|d}$ in (2.48) for MCLP. Hence, the discussion on the bias term in MCLP in Sec. 2.4.2.1 similarly applies to the GSC, implying that for $\underline{\mathbf{\Omega}} \in \{\underline{\mathbf{\Psi}}_{q_{s_1|e}}, \underline{\mathbf{\Psi}}_{s_1}\}$ in (2.36), we find $\boldsymbol{\vartheta}_{1|d} = \mathbf{0}$ and $e(l) = q_{s_1|e}(l)$, i.e. we achieve complete and unbiased cancellation. Note that if the target component was defined as $\sum_{n=1}^{N} q_{s_n|e}(l)$ instead of $q_{s_1|e}(l)$ as in the *different* acoustic scenario mentioned in Sec. 2.2.2 and $\mathbf{B}$ was changed accordingly, for the same $\underline{\mathbf{\Omega}}$, the GSC would yield the same result as MCLP. Also note that the above conclusions hold for complete blocking, i.e. if (2.27) is satisfied. For *incomplete blocking*, partial speech cancellation may appear. In Sec. 2.7, we simulate both cases.

2.5.2.2 Presence of Incoherent Noise

Similarly to the MCLP framework, if additional incoherent noise $\mathbf{v}(l) \neq \mathbf{0}$ is present, a simplification of (2.53) is not possible. We may therefore apply the same reasoning as in Sec. 2.4.2.2. Noting that the incoherent noise acts as M additional independent noise sources, we find that the condition for complete cancellation in the GSC framework (2.28), where $M-1$ is required to exceed the number of independent sources N, cannot be fulfilled, resulting in decreased performance. These conclusions are compliant with the analysis in [82], which demonstrates that in MVDR beamforming, there is an inherent trade-off between dereverberation and noise reduction for incoherent and mixed-coherent-plus-incoherent noise fields.

2.6 Comparative Summary

Table 2.1 summarizes the theoretical findings from Sec. 2.4 and Sec. 2.5. For a given pre-whitening matrix $\underline{\mathbf{\Omega}}^{-1/2}$, MCLP does not further distinguish between the speech source and the localized noise sources, while the GSC does so by means of the spatial pre-processing in the blocking matrix. MCLP hence treats all source components $q_{s_n}(l)$ the same, suppressing the late components $q_{s_n|\ell}(l)$, but none of the early components $q_{s_n|e}(l)$. By contrast, the GSC suppresses all coherent-noise components $q_{s_n}(l)$ on the one hand, and the late reverberant-speech component $q_{s_1|\ell}(l)$ on the other hand, provided that the blocking matrix $\mathbf{B}$ achieves *complete blocking* of the early reverberant-speech component. In both frameworks, an unbiased speech component with $\boldsymbol{\vartheta}_{1|d} = \mathbf{0}$ may be obtained by pre-whitening $q(l)$ and $\mathbf{u}(l)$ accordingly. The presence of incoherent noise decreases the performance.

The spatial pre-processing of the GSC naturally comes at a cost. The blocking matrix requires spatial information, which needs to be acquired in practice. Further, as to be demonstrated in simulations, cf. Sec. 2.7.2, in case of *incomplete blocking*, the GSC performs inferior in terms of dereverberation as compared to MCLP. Compared to MCLP, the minimum number of required microphones is increased by one, as the blocking matrix creates $M-1$ independent output signals only from M input signals. In the GSC, the number of filter channels is accordingly decreased by one, where a higher filter length L_w is required per channel.

Table 2.1: Comparative summary of the MCLP framework versus the GSC framework.

Property	*MCLP framework*	*GSC framework*
spatial knowledge	not required	required in $\mathbf{B}$, cf. (2.24)–(2.25)
filter input signal	$\mathbf{u}(l) = \mathbf{y}(l-d)$	$\mathbf{u}(l) = \mathbf{B}^H\mathbf{y}(l)$
filter length required for complete[1] cancellation	$L_w \geq \dfrac{N(L_h-1)}{M-N}$, requires $M > N$	$L_w \geq \dfrac{N(L_h-2)+(N-1)d}{M-N-1}$, requires $M > N+1$
output signal[1]	$e(l) = \sum_{n=1}^{N} q_{s_n\|e}(l) - \boldsymbol{\vartheta}_{n\|d}^H \mathbf{s}_n(l-d)$	$e(l) = q_{s_1\|e}(l) - \boldsymbol{\vartheta}_{1\|d}^H \mathbf{s}_1(l-d)$

[1]If incoherent noise absent (reduced performance otherwise).

2.7 Simulations

In this section, we present simulation results comparing MCLP and the GSC in terms of dereverberation and noise reduction performance. The simulation setup is described in 2.7.1, and the results are discussed in 2.7.2.

2.7.1 Simulation Setup

In order to confirm the theory in Sec. 2.4 to Sec. 2.5 and to further assess the practical relevance we respectively perform simulations for time domain and STFT domain implementations. The reasons for this are as follows: in the time domain, using oracle knowledge on the early RIRs, complete blocking can be simulated for the GSC, cf. Sec. 2.7.1.3. Further, unweighted global power-ratio measures can be well defined and evaluated, cf. Sec. 2.7.1.4. Therefore, in order to confirm the theory, we perform simulations on the time domain implementation, employing an ideal blocking matrix in the GSC yielding *complete blocking*, and evaluate the performance using unweighted global power-ratio measures. In the STFT domain, complete blocking cannot be simulated, since the sub-band convolution model in (2.1) poses an approximation of the time-domain convolution only [117]. Instead of using oracle knowledge in the blocking matrix, we estimate the RETFs from the microphone signals, such that the GSC performance also depends on the estimation quality of the RETFs, cf. Sec. 2.7.1.3. As, due to incomplete blocking, power-ratio measures equivalent to those used in the time domain cannot be well defined, and as unweighted global power-ratio measures are further known to relate comparably poorly to the perceived speech quality, we instead use perceptually motivated freqeuncy-weighted segmental power-ratio measures [160]. Therefore, in order to address the practical relevance, we perform simulations on the STFT domain implementation, employing an estimated blocking matrix in the GSC yielding *incomplete blocking*, and evaluate the performance using weighted segmental power-ratio measures.

2.7.1.1 Acoustic Scenario

Multi-channel RIRs are generated using the randomized image method [161] at a sampling frequency of 16 kHz, with the image sources randomly displaced within a sphere of 8 cm. A fractional delay low-pass filter with a relative cut-off frequency of 0.9 and a length of 11 taps is applied, such that the energy of each acoustic wave, i.e. of the direct component and each reflection, is spread over 11 samples. The room dimensions are $5\times4\times3$ m, the reverberation time

is 0.5 s. The room impulse responses are truncated after 8000 taps. A linear array of 8 microphones with inter-microphone distances of (4, 4, 4, 8, 4, 4, 4) cm is used. The simulations comprise one speech source and one localized noise source. In total, 128 scenarios are generated. In each scenario, the position and orientation of the microphone array is randomized. The speech source is located at a random position in broadside direction at 2 m distance to the center of the microphone array (i.e. on a circle around its axis). The position of the localized noise source is randomized, with the constraint that the distance to the center of the microphone array is at least 1 m and the angle between the localized noise source and the speech source, seen from the center of the microphone array, is at least 15 °.

2.7.1.2 Source Signals

We define two different source signal settings. In the first one, both the speech-source signal and the localized noise source signal are chosen to be temporally correlated, i.e. *colored* signals. The (non-stationary) speech-source signal $s_1(l)$ is composed of male and female speech of in total 51 s duration [25], while for the localized noise source signal $s_2(l)$, stationary pink noise is used. This setting is evaluated for both the time domain and the STFT domain implementations. In the time domain, for the chosen setup, cf. Sec. 2.7.1.3, the coloration of the source signals causes a biased filter estimate. In the second setting, both $s_1(l)$ and $s_2(l)$ are chosen to be temporally uncorrelated, i.e. *white*, stationary signals, and have been generated independently from the source signals in the first setting. This setting is evaluated in the time domain implementations only, leading to an unbiased filter estimate and serving as a reference in order to illustrate the effect of the bias in the first setting. Since sensor noise is always present in practice, spatially and temporally uncorrelated noise $\mathbf{v}(l)$ is added in all simulations. Note that due to the incoherent noise, the time domain simulation results may at most approximately reach the theoretical limits discussed in Sec. 2.4.2.1 and Sec. 2.5.2.1.

The power of the noise components is defined via the signal-to-coherent-noise ratio SNR_y^{coh} and the signal-to-incoherent-noise ratio SNR_y^{inc} in the first microphone, i.e.

$$SNR_y^{coh} = 10\log_{10}\frac{\sum_l |x_{1,1}(l)|^2}{\sum_l |x_{2,1}(l)|^2}\,\text{dB}, \tag{2.58}$$

$$SNR_y^{inc} = 10\log_{10}\frac{\sum_l |x_{1,1}(l)|^2}{\sum_l |v_1(l)|^2}\,\text{dB}, \tag{2.59}$$

where the reverberant-speech component $x_{1,1}(l)$ in the first microphone is considered to be the useful signal.

2.7.1.3 MCLP and GSC implementation

Time Domain In the time domain, we define the direct speech component to be the target component, i.e. we choose $\delta = L_b = 11$ samples, corresponding to the energy spread of a single acoustic wave, cf. Sec. 2.7.1.1, yielding $d = 11$ for MCLP and $d \geq 11$ for the GSC, cf. Sec. 2.2.3 and Sec. 2.2.4. An ideal GSC blocking matrix $\mathbf{B}$ was designed, cf. (2.24)–(2.25). The filter $\mathbf{g}$ is chosen to be a matched filter (MF) such that $\mathbf{B}^H\mathbf{g} = \mathbf{0}$, both for the GSC and MCLP. We choose $\underline{\mathbf{\Omega}} = \mathbf{I}$ in (2.36), leading to a biased filter estimate for colored source signals. The effect of the bias is shown by comparing the performance for both colored and white source signals.

STFT Domain In the STFT domain, using square-root-Hann windows of 512 samples with 50% overlap, we choose $\delta = L_b = 1$ frame. The GSC blocking matrix $\mathbf{B}$ uses an estimate of the RETFs, which we obtain as presented in [65, 147, 157]: we estimate the *average* spatial correlation matrix of the microphone signals using the whole batch, and the spatial correlation matrix of the stationary noise components using 5 s noise-only frames, such that the spatial speech-component correlation matrix can be estimated by subtraction. Then, assuming that the spatial coherence matrix of the late reverberant-speech component in frame l may be modeled as diffuse, the RETF relative to the first microphone is estimated by from the GEVD of the spatial speech-component correlation matrix estimate and the diffuse coherence matrix [65, 157]. Again, the filter $\mathbf{g}$ is chosen to be an MF with $\mathbf{B}^H\mathbf{g} = \mathbf{0}$, i.e. $\mathbf{g}$ is a normalized version of the RETF estimate. For $\underline{\mathbf{\Omega}}$ in (2.36), we use an estimate of $\underline{\mathbf{\Psi}}_{q_{s_1|e}}$, which in the STFT domain is diagonal for $d = 1$, cf. Sec. 2.3.2. Since $\mathbf{g}$ is a normalized version of the RETF estimate, estimating the PSDs $\psi_{q_{s_1|e}}(l)$ in $\underline{\mathbf{\Psi}}_{q_{s_1|e}}$ corresponds to estimating the early-reverberant speech component in the first microphone. Again, this can be done using the GEVD [65,157], now applied to a *recursive* estimate of the spatial speech-component correlation matrix in frame l and the diffuse coherence matrix. See, e.g., [65, 157] for a detailed and more formal discussion on GEVD-based RETF and PSD estimation.

2.7.1.4 Performance Measures

Time Domain In the time domain, equivalently to $q_{s_1}(l)$, $q_{s_2}(l)$, $q_v(l)$ and $q_{s_1|e}(l)$, we define the individual source components of $e(n)$ as $e_{s_1}(l)$, $e_{s_2}(l)$,

$e_v(l)$ and $e_{s_1|e}(l)$, where $e_{s_1|e}(l) = q_{s_1|e}(l)$, cf. (2.49), (2.57). With $\sigma \in \{q, e\}$, the signal-to-coherent-noise ratio SNR_σ^{coh}, the signal-to-incoherent-noise ratio SNR_σ^{inc}, the signal-to-total-noise ratio SNR_σ^{tot}, and the signal-to-reverberation ratio SRR_σ at the MF output and the MCLP and GSC output are defined as

$$SNR_\sigma^{tot} = 10\log_{10}\frac{\sum_l |\sigma_{s_1e}(l)|^2}{\sum_l |\sigma_{s_2}(l) + \sigma_v(l)|^2}\,\text{dB}, \tag{2.60}$$

$$SRR_\sigma = 10\log_{10}\frac{\sum_l |\sigma_{s_1e}(l)|^2}{\sum_l |\sigma_{s_1}(l) - \sigma_{s_1|e}(l)|^2}\,\text{dB}, \tag{2.61}$$

where the component $\sigma_{s_1e}(l)$ is considered to be the useful signal. Please note that $q(l)$, and hence for $\sigma = q$ also the measures in (2.58)–(2.61), are independent of the particular framework. Further, note that in the denominator of (2.61), for $\sigma = q$, the difference $q_{s_1}(l) - q_{s_1|e}(l)$ equals the late reverberant-speech component $q_{s_1|\ell}(l)$, while for $\sigma = e$, the difference $e_{s_1}(l) - e_{s_1|e}(l)$ comprises not only residual reverberation, but also a bias term in the general case. For evaluation, we use the improvement in SNR^{tot} and SRR, i.e.

$$\Delta SNR^{tot} = SNR_e^{tot} - SNR_q^{tot}, \tag{2.62}$$

$$\Delta SRR = SRR_e - SRR_q. \tag{2.63}$$

STFT Domain In the STFT domain, the target component $q_{s_1|e}(l)$ cannot be observed separately, since the sub-band convolution model in (2.1) poses an approximation of the time-domain convolution only [117]. Further, due to overlapping frames in the STFT processing and incomplete blocking in the GSC, the target component $q_{s_1|e}(l)$ may not be completely maintained in $e(l)$, such that the measures in (2.60)–(2.63) are not suitable in the STFT domain. Instead, we define the direct-component in $q(l)$ as a reference signal, which *cannot* be assumed to be equivalent to $q_{s_1|e}(l)$. Then, with $\sigma \in \{q, e\}$, for $\sigma(l)$ and $\sigma_{s_1}(l)$, respectively, we compute the frequency-weighted segmental signal-to-noise-plus-reverberation ratio and the frequency-weighted segmental signal-to-reverberation ratio [160], denoted as $SNRR^{fws}$ and SRR^{fws} and indicating the dereverberation-plus-noise-reduction performance and the dereverberation-only performance.

2.7.1.5 Varied Parameters

In the time domain, simulations are carried out for different values of the following parameters: SNR_y^{coh}, SNR_y^{inc}, L_w, and M. The filter length L_w is presented relatively to the theoretical minimum given in (2.22), (2.28), denoted

by L_w^{rel}. While one parameter is varied, the others are fixed at $SNR_q^{coh} = 0\,\text{dB}$, $SNR_q^{inc} = 90\,\text{dB}$, $L_w^{\text{rel}} = 1$, and $M = 8$, i.e. all simulations intersect at this point. For $N = 2$, the minimum number of microphones required by MCLP and the GSC is given by $M = 3$ and $M = 4$, respectively, cf. (2.22), (2.28). If the number of microphones M falls below this required minimum, the filter length is computed setting the denominators in (2.22), (2.28) to one. Simulations posing nearly ideal conditions, i.e. sufficiently high SNR_q^{inc}, $L_w^{\text{rel}} \geq 1$ and sufficiently many microphones M, validate the theoretical results in Sec. 2.4 and Sec. 2.5, with minor deviations occurring due to the LS approximation in (2.36) of the Wiener solution in (2.40) and remaining low-level incoherent noise.

In the STFT domain, simulations are carried out for different values of SNR_y^{coh} only, with $SNR_q^{inc} = 90\,\text{dB}$, $L_w^{\text{rel}} = 1$, and $M = 8$. Since complete blocking is not achieved in the STFT domain, decreased performance is expected for the GSC.

2.7.2 Simulation Results

We now discuss the time and STFT domain simulation results.

2.7.2.1 Time Domain

The performance of both frameworks in terms of ΔSRR and ΔSNR^{tot} are shown in Fig. 2.6. We first discuss the dereverberation performance, followed by the noise reduction performance.

Dereverberation From Fig. 2.6 (a)–(d) we observe that under favorable conditions with predominantly late-reverberant-speech interference, i.e. for high SNR_y^{coh}, high SNR_y^{inc}, $L_w^{\text{rel}} \geq 1$ and sufficiently high M, the SRR improvement of both MCLP and GSC converge to the same value of around 31 dB for white source signals, respectively denoted by [·····] and [·····]. This upper limit is determined by the LS approximation (2.36) of the Wiener Solution of the (2.40). In all conditions, for colored source signals, the target component $q_{s_1|e}(l)$ is partially whitened due to the biased filter estimate, leading to a performance drop for both MCLP [——] and the GSC [——], cf. Sec. 2.4.2.1 and Sec. 2.5.2.1 and Table 2.1. The GSC reaches up to 18 dB ΔSRR, outperforming MCLP by 3 dB. This is due to the potential delay between the direct component and the first reflection, increasing d for the GSC and thereby decreasing the bias, cf. Sec. 2.7.1.3. The higher standard deviation of ΔSRR for the GSC is a result of the variation of this delay over different source and microphone array positions.

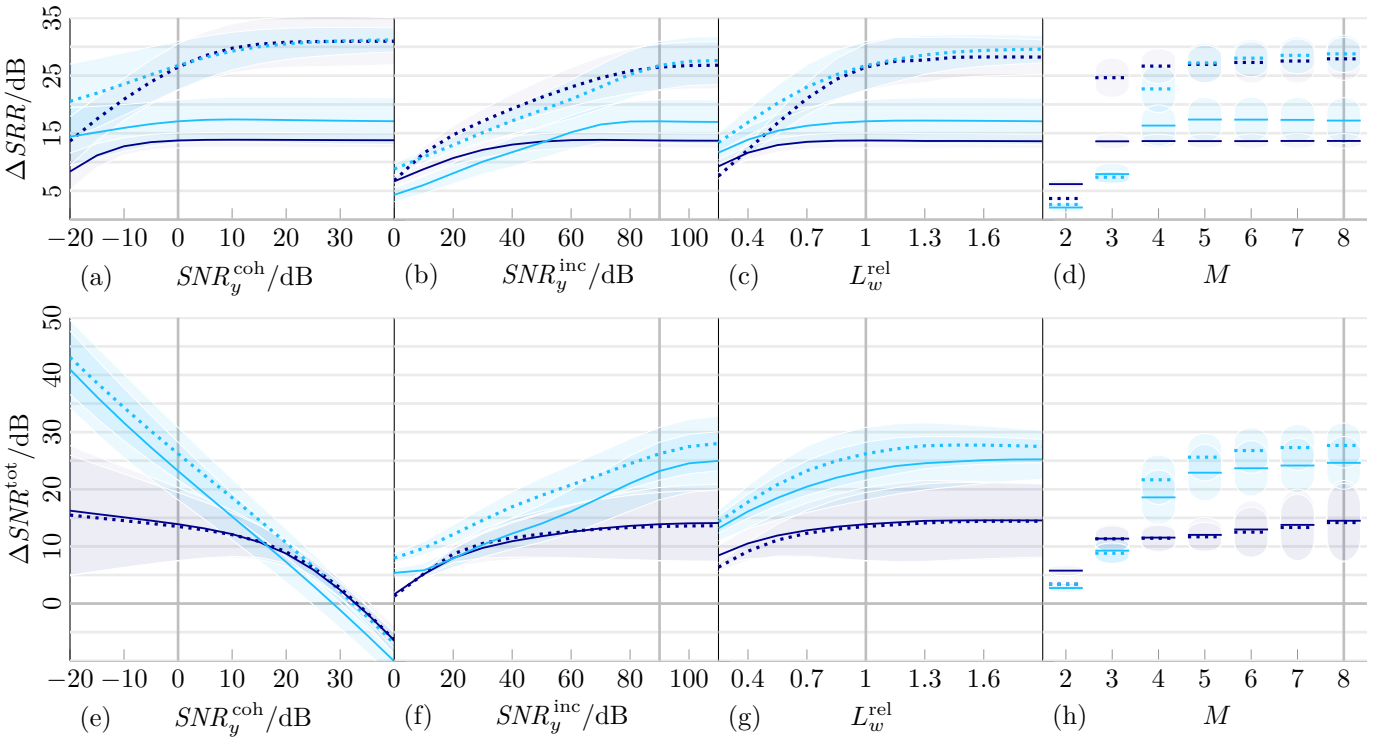


Figure 2.6: Dereverberation/noise reduction performance $\Delta SRR/\Delta SNR^{tot}$ versus (a)/(e) SNR_q^{coh}, (b)/(f) SNR_q^{inc}, (c)/(g) L_w^{rel} and (d)/(h) M for colored and white source signals of the MCLP framework, respectively denoted by [——] and [·····], and the GSC framework, respectively denoted by [——] and [·····]. The vertical grid lines indicate the intersection point of the individual subplots. The shaded areas represent the standard deviation.

As can be seen in Fig. 2.6 (a), for white source signals, MCLP shows a rather high sensitivity towards coherent noise for $SNR_y^{coh} < 10\,\text{dB}$, whereas the GSC is somewhat less sensitive. This can be explained by the limited number of observations L_{obs} used in the LS approximation (2.36) of the Wiener solution (2.40), causing LS to focus on the suppression of components with higher power, i.e. here on (late) coherent-noise suppression. For colored source signals, the effect is less pronounced in both frameworks.

While it can be observed from Fig. 2.6 (b) that both MCLP and GSC are highly sensitive to incoherent noise, the results also indicate an up to 2.5 dB lower performance of the GSC as compared to MCLP for $SNR_y^{inc} < 90\,\text{dB}$ and $< 55\,\text{dB}$ for white and colored source signals, respectively. The reason for this may lie in the GSC blocking matrix, which by construction causes a cross-correlation of the incoherent-noise components in the data-dependent filter input $\mathbf{u}(l)$, as opposed to the mere delay in MCLP. Hence, for the GSC, not only the autocorrelation submatrices of $\mathbf{\Psi}_{\tilde{u}}$ are affected by incoherent noise, but also the cross-correlation submatrices.

As shown in Fig. 2.6 (c), ΔSRR saturates for both MCLP and GSC above $L_w^{\text{rel}} = 1$, both for white and colored source signals, as expected from theory. The GSC performs slightly better than MCLP if $L_w^{\text{rel}} < 1$. We may however state that for both frameworks, undermodeling is not extremely critical, as even at $L_w^{\text{rel}} = 0.7$, we obtain ΔSRR values above 22 dB for white source signals, while the performance is hardly affected for colored source signals.

From Fig. 2.6 (d), we note that ΔSRR drops sharply for both MCLP and GSC if the number of microphones is smaller than required, i.e. $M < 3$ and $M < 4$, respectively. This holds for both white and colored source signals. MCLP reaches saturation at $M = 3$, while the GSC saturates at $M = 5$ only instead of $M = 4$. This may be caused by remaining low-level incoherent noise and possibly nearly common zeros in the transfer functions corresponding to $\mathbf{H}_B$ in (2.27) for $M = 4$.

Noise Reduction From Fig. 2.6 (e)–(h) we observe that under favorable conditions with predominantly coherent-noise interference, i.e. for low SNR_y^{coh}, high SNR_y^{inc}, $L_w^{\text{rel}} \geq 1$ and sufficiently high M, the GSC [·····, ——] shows increasing improvement in terms of ΔSNR^{tot} for decreasing values of SNR_y^{coh}, while for MCLP [·····, ——], ΔSNR^{tot} is limited to at most 15 dB. the GSC [·····, ——] clearly outperforms MCLP [·····, ——] in terms of ΔSNR^{tot}. This is due to the GSC suppressing the entire coherent-noise component $q_{s_2}(l)$, while MCLP suppresses the late coherent-noise component $q_{s_2|\ell}(l)$ only, cf. (2.47)–(2.49), (2.55)–(2.57) and Table 2.1. Note that MCLP exhibits a stronger standard

deviation in ΔSNR^{tot} than the GSC. This is caused by the varying power of the early coherent-noise component $q_{s_2|\mathrm{e}}(l)$, as the power of the individual direct components at the output of the MF may be distributed over a range potentially larger than d, depending on the angle between the speech source and the coherent-noise source. In all conditions, the GSC performs somewhat worse for colored signals than white signals, while no significant difference is found for MCLP.

Fig. 2.6 (e) indicates that the GSC exceeds MCLP for $SNR_y^{coh} < 20\,\mathrm{dB}$, while both frameworks perform similarly for high SNR_y^{coh} values. For the GSC, ΔSNR^{tot} decreases at a rate of slightly less than 10 dB ΔSNR^{tot} per 10 dB SNR_y^{coh}, such that the noise power at the output is almost constant throughout the simulated range. This implies that for $SNR_y^{coh} \geq 35\,\mathrm{dB}$, the total noise power is in fact even boosted as compared to the output of the MF, both for MCLP and the GSC. Again, this effect can be explained by the limited number of observations L_{obs} in the LS estimate (2.36), here causing LS to focus on reverberant-speech suppression.

As can be seen from Fig. 2.6 (f), for both white and colored source signals, the GSC exceeds MCLP for higher SNR_y^{inc} values, while the difference reduces for lower values.

As shown in Fig. 2.6 (g), for both white and colored source signals, both MCLP and the GSC again saturate for $L_w^{\mathrm{rel}} \geq 1$. Again, undermodeling does not appear to be extremely critical.

From Fig. 2.6 (h), for both white and colored source signals, we once more find a sharp performance drop for both MCLP and the GSC if $M < 3$ and $M < 4$, respectively. Again, saturation is reached at $M = 3$ and $M = 5$, respectively.

2.7.2.2 STFT Domain

The performance of both frameworks in terms of SRR^{fws} and $SNRR^{\mathrm{fws}}$ are shown in Fig. 2.7, where the performance of the MF serves as a reference.

From Fig. 2.7 (a), we note that the dereverberation-only performance of the MF [- · - · -] in terms of SRR^{fws} remains almost constant around 4.1 dB. At high SNR_y^{coh} values with predominantly late-reverberant-speech interference, MCLP [——] and the GSC [——] outperform the MF by up to 3.1 dB and 4.1 dB, respectively. Note that in theory, for *complete blocking*, i.e. if (2.27) is satisfied, the GSC is expected to perform as effectively as MCLP in terms of dereverberation, cf. (2.47)–(2.49), (2.55)–(2.57), Table 2.1, and the time domain simulation results in Sec. 2.7.2.1. However, since the sub-band convolution

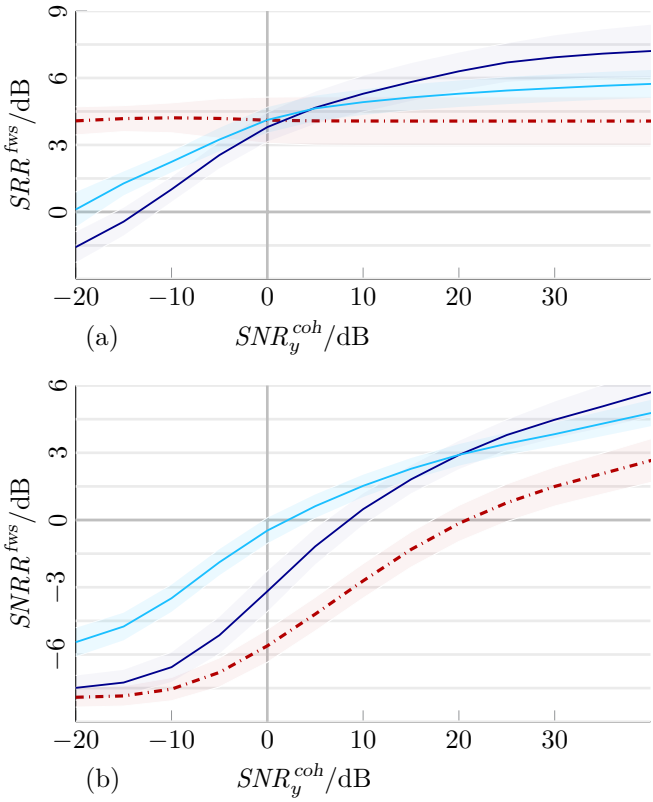


Figure 2.7: (a) dereverberation-only/(b) dereverberation-plus-noise-reduction performance in terms of $SRR^{fws}/SNRR^{fws}$ versus SNR_y^{coh} of the MF, the MCLP and the GSC framework, respectively denoted by [-·-·-], [——], and [——]. The shaded areas represent the standard deviation.

model in (2.1) poses an approximation of the time-domain convolution only, and since the RETFs used in the GSC blocking matrix are subject to estimation errors, cf. Sec. 2.7.1.3, complete blocking is not achieved in the STFT domain. Due to *incomplete blocking*, the GSC hence suffers from some amount of early reverberant-speech cancellation and in addition from incomplete prediction of the late reverberant-speech component $q_{s_1|\ell}(l)$, leading to reduced dereverberation performance in comparison to MCLP. At low SNR_y^{coh} values with predominantly coherent-noise interference, we find that both MCLP and the GSC perform worse than the MF, indicating speech distortion. Again, this effect can be explained by the limited number of observations L_{obs} used in the LS estimate (2.36), causing LS to focus on (late) coherent-noise suppression, cf. Sec. 2.7.2.1. Since the GSC is able to suppress the early coherent-noise component $q_{s_2|e}(l)$

also, the GSC performance is less affected.

From Fig. 2.7 (b), we note that the dereverberation-plus-noise-reduction performance of the MF in terms of $SNRR^{fws}$ ranges between $-8\,\text{dB}$ for $SNR_y^{coh} = -20\,\text{dB}$ and $2.6\,\text{dB}$ for $SNR_y^{coh} = 40\,\text{dB}$, where the upper limit is still affected by the noise component, as can be seen by comparison to the dereverberation-only performance in Fig. 2.7 (a). At high SNR_y^{coh} values MCLP and the GSC outperform the MF by up to $32.8\,\text{dB}$ and $2.1\,\text{dB}$, respectively. At low SNR_y^{coh} values, MCLP performs only somewhat better than the MF, while the GSC in contrast outperforms the MF by up to $5.1\,\text{dB}$. This difference at low SNR_y^{coh} values is expected as MCLP suppresses the late coherent-noise component $q_{s_2|e}(l)$ only, while the GSC suppresses the entire coherent-noise component $q_{s_2}(l)$, cf. (2.47)–(2.49), (2.55)–(2.57), Table 2.1, and the time domain simulation results in Sec. 2.7.2.1. Audio examples of the STFT domain simulations are available online [162].

2.8 Conclusion

In this paper, we formally analyzed and compared the MCLP and GSC frameworks in terms of blind dereverberation and noise reduction performance. Both frameworks are theoretically able to perform complete dereverberation if incoherent noise is absent. Due to the use of a blocking matrix, the GSC is theoretically able to completely cancel coherent noise in the absence of incoherent noise, while MCLP cancels the late coherent-noise component only. For complete cancellation, the GSC requires one additional microphone as compared to MCLP. Furthermore, the blocking matrix design requires spatial information in form of the early speech-source RIR or the RETF, which needs to be acquired in practice. In order to confirm the theory and to assess the practical relevance of the theoretical findings, we carried out time domain simulations using oracle knowledge on the early RIRs, resulting in *complete blocking* of the early reverberant-speech component, and STFT domain simulations using estimated REFTs, resulting in *incomplete blocking*.

The simulation results confirm that in terms of noise reduction, as opposed to the GSC performance, the performance of MCLP is limited. In terms of dereverberation, the GSC performs equally well if *complete blocking* is achieved, as expected from the theoretical analysis, but performs inferior for *incomplete blocking*. Both MCLP and the GSC exhibit strong sensitivity to incoherent noise. For both frameworks, dereverberation and noise reduction performance reach their maximum at a relative filter length of about one, while moderate undermodeling of the filter length does not appear to be extremely critical. The

simulations further confirm that for one coherent-noise component, the GSC requires four microphones, while MCLP requires three microphones only.

In summary, we can state that if sufficiently many microphones are available and complete blocking is achieved, the GSC performs superior to MCLP in terms of noise reduction and equally well in terms of dereverberation, but inferior in terms of dereverberation for incomplete blocking. In practice therefore, in acoustic conditions with only mild noise but predominantly late-reverberant-speech interference, MCLP is to be preferred, while in case of predominantly noise but mild to moderate late-reverberant-speech interference, the GSC is to be preferred. In acoustic conditions with both strong reverberation and strong noise, combined schemes may be most appropriate.

Part II

Integrated Architecture and Parameter Estimation

Chapter 3

The ISCLP Kalman Filter

Integrated Sidelobe Cancellation and Linear Prediction Kalman Filter for Joint Multi-Microphone Speech Dereverberation, Interfering Speech Cancellation, and Noise Reduction

Thomas Dietzen, Simon Doclo, Marc Moonen, and Toon van Waterschoot

ESAT-STADIUS Tech. Rep. TR 19-70, KU Leuven, Belgium, submitted for publication, June 2019.

The candidate's contributions as first author include: literature study, co-derivation of the presented theory, co-development of the presented algorithms, co-design of the evaluation experiments, software implementation and computer simulations, co-formulation of the conclusions, text redaction and editing.

Abstract

In multi-microphone speech enhancement, reverberation as well as additive noise and/or interfering speech are commonly suppressed by deconvolution and spatial filtering, e.g., using multi-channel linear prediction (MCLP) on the one hand and beamforming, e.g., a generalized sidelobe canceler (GSC), on the other hand. In this paper, we consider several reverberant speech components, whereof some are to be dereverberated and others to be canceled, as well as a diffuse (e.g., babble) noise component to be suppressed. In order to perform both deconvolution and spatial filtering, we integrate MCLP and the GSC into a novel architecture referred to as integrated sidelobe cancellation and linear prediction (ISCLP), where the sidelobe-cancellation (SC) filter and the linear prediction (LP) filter operate in parallel, but on different microphone signal frames. Within ISCLP, we estimate both filters jointly by means of a single Kalman filter. We further propose a spectral Wiener gain post-processor, which is shown to relate to the Kalman filter's posterior state estimate. The presented ISCLP Kalman filter is benchmarked against two state-of-the-art approaches, namely first a pair of alternating Kalman filters respectively performing dereverberation and noise reduction, and second an MCLP+GSC Kalman filter cascade. While the ISCLP Kalman filter is roughly M^2 times less expensive than both reference algorithms, where M denotes the number of microphones, it is shown to perform similarly as compared to the former, and to outperform the latter.

Index terms — Dereverberation, interfering speech cancellation, noise reduction, beamforming, multi-channel linear prediction, Kalman filter.

3.1 Introduction

In many wide-spread speech processing applications such as hands-free telephony and distant automatic speech recognition, reverberation as well as additive noise and/or interfering speech impinging on a microphone may deteriorate the quality and intelligibility of the speech recordings [15]. The demanding tasks of dereverberation, noise reduction and/or interfering speech cancellation, and in particular the conjunction of these therefore remain a subject of ongoing research, with multi-microphone-based approaches exploiting spatial diversity receiving particular interest [83, 85, 87, 89, 91, 92, 94–96, 98–100, 102, 103, 111–113, 115, 116, 157, 163]. In this context, we below briefly discuss two broad concepts in multi-microphone speech enhancement, namely spatial filtering and deconvolution.

As a spatial filtering technique, beamforming is commonly used in noise reduction and interfering speech cancellation, but may as well be applied for dereverberation [83, 85, 87]. In order to perform both dereverberation and noise reduction, several beamforming schemes have been proposed. In [83], a cascaded approach is presented, using data-independent, super-directive beamforming for dereverberation, and data-dependent, e.g., minimum-variance distortionless response (MVDR) beamforming, for noise reduction. The generalized sidelobe canceler (GSC), a popular implementation of the MVDR beamformer, has been applied in different constellations [85, 87]. In [85], joint dereverberation and noise reduction is performed using a single GSC, while in [87], a nested structure is proposed, employing an inner GSC for dereverberation and an outer GSC for noise reduction. The GSC is composed of two parallel signal paths: a reference path and a sidelobe-cancellation (SC) path. The reference path traditionally employs a matched filter (MF), while the SC path cascades a blocking matrix (BM), blocking either the entire or the early-reverberant speech component, and an SC filter, minimizing the output power and thereby suppressing residual nuisance components in the reference path, i.e. either residual noise or both residual noise and reverberation components.

As a deconvolution technique, multi-channel linear prediction (MCLP) [89, 91, 92, 94–96, 98–100, 102, 103, 111–113, 115, 116, 157, 163] recently prevailed in blind speech dereverberation, while noise reduction is not targeted. As opposed to beamforming, MCLP does not require spatial information on the speech source. Instead, for each microphone, the reverberation component to be canceled is modeled as a linear prediction (LP) component, i.e. as a filtered version of the delayed microphone signals, with the LP filter to be estimated. Besides iterative LP filter estimation approaches such as [91, 94–96, 112, 113], also adaptive approaches based on recursive least squares (RLS) [92, 100, 103, 111] as well as the Kalman filter [98, 99, 102] have been proposed in the past years. In order to reduce noise after dereverberation, multiple-output MCLP has been cascaded with MVDR beamforming in [112, 113], which was seen to be a commonly adopted approach in the 2018 CHiME-5 challenge [49]. In [115], the cascade in [112, 113] is unified. In [116], joint MCLP-based dereverberation and noise reduction is performed using a pair of alternating Kalman filters respectively estimating the LP filter and the noise-free reverberant speech component.

In [163, cf. Ch. 2], we have presented a comparative analysis of the GSC and MCLP. In another previous paper [157], instead of cascading MCLP and beamforming or relying on beamforming only, we have proposed to integrate the GSC and MCLP by employing an SC path and LP path in parallel, resulting in an architecture we refer to as integrated sidelobe cancellation and linear prediction (ISCLP). Within this novel architecture, we have estimated the SC and LP filters jointly by means of a single Kalman filter. Here, the spatial pre-processing blocks

MF and BM require an estimate of the relative early transfer functions (RETFs), cf. also [85], while the Kalman filter requires an estimate of the power spectral density (PSD) of the desired early speech component, cf. also [98, 99, 102]. In this paper, the work in [157] is extended in the following manner. We generalize the short-time Fourier transform (STFT) domain-based signal model, which now comprises several (spatio-temporally correlated) reverberant speech components, whereof some are to be dereverberated and others to be canceled, as well as a (spatially, but not temporally correlated) diffuse (e.g., babble) noise component to be suppressed. This generalized acoustic scenario necessitates (non-stationary) multi-source early PSD estimation and RETF updates, which is achieved by means of the algorithm recently proposed in [164, cf. Ch. 4]. We further augment the proposed approach by a spectral Wiener gain post-processor, which is shown to relate to the Kalman filter's posterior state estimate. In order to demonstrate the effectiveness of the ISCLP Kalman filter, we compare against two state-of-the-art approaches – first the previously mentioned alternating Kalman filters in [116], and second an MCLP+GSC Kalman filter cascade, conceptually relating to [112,113]. As compared to these two reference algorithms, the ISCLP Kalman filter is computationally roughly M^2 times less expensive, where M denotes the number of microphones. Yet, the ISCLP Kalman filter is shown to perform similarly as compared to the alternating Kalman filters, and to outperform the MCLP+GSC Kalman filter cascade. A MATLAB implementation is available [152].

The paper is organized as follows. In Sec. 3.2, we present the signal model in the STFT domain. In Sec. 3.3, the ISCLP Kalman filter is described. Implementational aspects are discussed in Sec. 3.4, followed by simulations in Sec. 3.5.

3.2 Signal Model

Throughout the paper, we use the following notation: vectors are denoted by lower-case boldface letters, matrices by upper-case boldface letters, $\mathbf{I}$ and $\mathbf{0}$ denote an identity and zero matrix, $\mathbf{1}$ denotes a vector of ones, $\mathbf{A}^*$, $\mathbf{A}^T$, $\mathbf{A}^H$, and $\mathrm{E}[\mathbf{A}]$ denote the complex conjugate, the transpose, the complex conjugate transpose or Hermitian, and the expected value of a matrix $\mathbf{A}$. The operation $\mathrm{Diag}[\mathbf{a}]$ creates a diagonal matrix with the elements of $\mathbf{a}$ on its diagonal, and $\mathrm{tr}[\mathbf{A}]$ denotes the trace of $\mathbf{A}$. Submatrices are referenced either by index ranges or alternatively by sets of indices, e.g., the submatrix of $\mathbf{A}$ spanning all rows and the columns j_1 to j_2 is denoted as $[\mathbf{A}]_{:,j_1:j_2}$, and the submatrix composed of all rows and the columns of $\mathbf{A}$ with indices in the ordered set T is denoted as $[\mathbf{A}]_{:,\in T}$.

In the short-time Fourier transform (STFT) domain, with l and k indexing the frame and the frequency bin, respectively, let $y_m(l,k)$ with $m = 1, \ldots, M$ denote the m^{th} microphone signal, with M the number of microphones. In the following, we treat all frequency bins independently and hence omit the frequency index. We define the stacked microphone signal vector $\mathbf{y}(l) \in \mathbb{C}^M$,

$$\mathbf{y}(l) = \Big(y_1(l) \ \cdots \ y_M(l)\Big)^T \tag{3.1}$$

composed of the mutually uncorrelated reverberant speech components $\mathbf{x}_n(l)$ with $n = 1, \ldots, N$ originating from $N < M$ point speech sources and the noise component $\mathbf{v}(l)$, defined similarly to (3.1), i.e.

$$\mathbf{y}(l) = \sum_{n=1}^{N} \mathbf{x}_n(l) + \mathbf{v}(l). \tag{3.2}$$

Here, the reverberant speech components $\mathbf{x}_n(l)$ may be decomposed into the early and late-reverberant speech components $\mathbf{x}_{n|e}(l)$ and $\mathbf{x}_{n|\ell}(l)$, i.e.

$$\mathbf{x}_n(l) = \mathbf{x}_{n|e}(l) + \mathbf{x}_{n|\ell}(l), \tag{3.3}$$

which are commonly parted by the arrival time of the therein contained reflections and assumed to have distinct spatio-temporal properties as outlined below. Let $\mathbf{x}_e(l) = \sum_{n=1}^{N} \mathbf{x}_{n|e}(l)$ and $\mathbf{x}_\ell(l) = \sum_{n=1}^{N} \mathbf{x}_{n|\ell}(l)$ denote the sum of the early and late-reverberant speech components, respectively, such that $\mathbf{y}(l)$ in (3.2)–(3.3) may alternatively be written as

$$\mathbf{y}(l) = \mathbf{x}_e(l) + \mathbf{x}_\ell(l) + \mathbf{v}(l). \tag{3.4}$$

Early reflections are assumed to arrive within the same frame, where the early components in $\mathbf{x}_{n|e}(l)$ are related by the RETFs in $\mathbf{h}_n(l) \in \mathbb{C}^M$ as $\mathbf{x}_{n|e}(l) = \mathbf{h}_n(l) s_n(l)$. Here, without loss of generality, the RETFs are assumed to be relative to the first microphone, i.e. $[\mathbf{h}_n(l)]_1 = 1$, and $s_n(l) = [\mathbf{x}_{n|e}(l)]_1$ denotes the early component in the first microphone originating from the n^{th} source, in the following referred to as early source image. We stack $\mathbf{h}_n(l)$ and $s_n(l)$ into $\mathbf{H}(l) \in \mathbb{C}^{M \times N}$ and $\mathbf{s}(l) \in \mathbb{C}^N$, respectively, i.e.

$$\mathbf{H}(l) = \Big(\mathbf{h}_1(l) \ \cdots \ \mathbf{h}_N(l)\Big), \tag{3.5}$$

$$\mathbf{s}(l) = \Big(s_1(l) \ \cdots \ s_N(l)\Big)^T, \tag{3.6}$$

such that $\mathbf{x}_e(l)$ is expressed by

$$\mathbf{x}_e(l) = \mathbf{H}(l)\mathbf{s}(l). \tag{3.7}$$

In the following, let $N_T \leq N$ early speech source images $s_n(l)$ be defined as the target source images, and let T denote the set of the corresponding $|T| = N_T$ target-source indices. Let T' denote the complement set to T, with $|T'| = N - N_T$. In order to distinguish the target components in $\mathbf{y}(l)$ as well as their complements, we introduce the short-hand notations similar to (3.5)–(3.7),

$$\mathbf{H}_T(l) = [\mathbf{H}(l)]_{:,\in T}, \tag{3.8}$$

$$\mathbf{s}_T(l) = [\mathbf{s}(l)]_{\in T}, \tag{3.9}$$

$$\mathbf{x}_{e|T}(l) = \mathbf{H}_T(l)\mathbf{s}_T(l), \tag{3.10}$$

and $\mathbf{H}_{T'}(l)$, $\mathbf{s}_{T'}(l)$, and $\mathbf{x}_{e|T'}(l)$ similarly, such that $\mathbf{x}_e(l)$ in (3.4) becomes

$$\mathbf{x}_e(l) = \mathbf{x}_{e|T}(l) + \mathbf{x}_{e|T'}(l). \tag{3.11}$$

Our objective is to estimate

$$s_T(l) = \sum_{n \in T} s_n(l) = \mathbf{1}^T \mathbf{s}_T(l) \tag{3.12}$$

from $\mathbf{y}(l)$ by means of the ISCLP Kalman filter. To this end, we rely on assumptions on the spatio-temporal behavior of the individual microphone signal components. We assume that $s_n(l)$ is temporally uncorrelated across frames, i.e. we have $\mathrm{E}[s_n(l-l')s_n^*(l)] = 0$ for $l' > 0$, and with $\mathbf{x}_{n|e}(l) = \mathbf{h}_n(l)s_n(l)$ consequently

$$\mathrm{E}[\mathbf{x}_{n|e}(l-l')\mathbf{x}_{n|e}^H(l)] = \mathbf{0} \quad \text{for} \quad l' > 0. \tag{3.13}$$

For speech signals, this assumption can be considered approximately justified if the STFT window length and window shift are sufficiently large. Within the limits defined by the reverberation time, we assume that the late-reverberant speech component $\mathbf{x}_{n|\ell}(l)$ is correlated to previous early source images $s_n(l-l')$ with $l' > 0$, but not to the current early source image $s_n(l)$, i.e. we have

$$\boxed{\mathrm{E}[\mathbf{x}_{n|e}(l-l')\mathbf{x}_{n|\ell}^H(l)] \neq \mathbf{0} \quad \text{for} \quad l' > 0,} \tag{3.14}$$

$$\mathrm{E}[\mathbf{x}_{n|e}(l)\mathbf{x}_{n|\ell}^H(l)] = \mathbf{0}. \tag{3.15}$$

Note that (3.14) also implies $\mathrm{E}[\mathbf{x}_{n|\ell}(l-l')\mathbf{x}_{n|\ell}^H(l)] \neq \mathbf{0}$ for all l'. With (3.14), we may predict $\mathbf{x}_{n|\ell}(l)$ from $\mathbf{x}_n(l-l')$, which indeed is the fundamental assumption of MCLP-based dereverberation [89, 91, 92, 94–96, 98–100, 102, 103, 111–113, 116]. Assumptions (3.13) and (3.15) allow for unbiased filter estimation [163, cf. Ch. 2] in MCLP-based dereverberation [89,91,92,94–96,98–100,102,103,111–113,116]

and GSC-based dereverberation and noise reduction [85,87], respectively. Hence, all three assumptions (3.13)–(3.15) are equally essential in the derivation of the ISCLP Kalman filter, cf. Sec. 3.3. Similarly to $s_n(l)$, the noise component $\mathbf{v}(l)$ is assumed to be temporally uncorrelated, i.e.

$$\mathrm{E}[\mathbf{v}(l-l')\mathbf{v}^H(l)] = \mathbf{0} \quad \text{for} \quad l' > 0, \tag{3.16}$$

and is therefore not predictable.

Within frame l, i.e. for $l' = 0$, we further make assumptions on the spatial behavior of $\mathbf{x}_{n|\ell}(l)$ and $\mathbf{v}(l)$, namely that both may be modeled as spatially diffuse. However, as these assumptions are irrelevant in the derivation of the ISCLP Kalman filter itself, cf. Sec. 3.3, but required only for parameter estimation based on [164, cf. Ch. 4], i.e. the estimation of the RETFs $\mathbf{H}_T(l)$ and the PSD $\varphi_{s_T}(l) = \mathrm{E}[s_T(l)s_T^*(l)]$, we treat them in the corresponding section only, cf. Sec. 3.4.2.

3.3 Integrated Sidelobe Cancellation and Linear Prediction Kalman Filter

We strive to estimate the target component $s_T(l)$ from the microphone signals $\mathbf{y}(l)$ defined in Sec. 3.2. For this purpose, we introduce the ISCLP architecture. In Sec. 3.3.1, we describe the SC and LP signal paths and filter constellations, which require spatio-temporal pre-processing of $\mathbf{y}(l)$. In Sec. 3.3.2, striving for recursive filter estimation, we define an ISCLP state-space model for the SC and the LP filter, wherefrom a Kalman filter is derived. The Kalman filter yields a (prior) estimate $e(l) = \hat{s}_T(l)$ of $s_T(l)$, which may further be spectrally post-processed, as shown in Sec. 3.3.3.

3.3.1 ISCLP Signal Path Architecture

A block-diagram of the ISCLP architecture is depicted in Fig. 3.1. It integrates the GSC and MCLP and hence consists of three signal paths: a reference path employing an MF, an SC path, composed of a BM and an SC filter, and a LP path, composed of a delay and an LP filter. While the MF, the BM and the SC filter are multiplicative (mult.), i.e. they operate on a single frame, the LP filter is convolutive (conv.), i.e. it operates across frames. The MF and the BM perform spatial pre-processing, serving unconstrained estimation of the SC filter, while the delay may analogously be considered as temporal pre-processing, serving unconstrained estimation of the LP filter. Structurally, one

may interpret ISCLP either as MCLP with the conventional reference channel selection replaced by a GSC, or alternatively as a GSC employing a generalized BM (composed of a traditional BM and a delay line), and a convolutive filter (composed of the SC and the LP filter). In the following, we formally discuss the individual signal paths.

In oder to maintain the target component $s_T(l)$ in (3.12), the MF $\mathbf{g} \in \mathbb{C}^M$ must satisfy [32, 42]

$$\mathbf{g}^H(l)\mathbf{H}_T(l) = \mathbf{1}^T, \tag{3.17}$$

where a commonly used [32, 42] choice for $\mathbf{g}(l)$ adhering to (3.17) is

$$\mathbf{g}(l) = \mathbf{H}_T(l)\big(\mathbf{H}_T^H(l)\mathbf{H}_T(l)\big)^{-1}\mathbf{1}, \tag{3.18}$$

with $\mathbf{H}_T(l)\big(\mathbf{H}_T^H(l)\mathbf{H}_T(l)\big)^{-1}$ the pseudoinverse of $\mathbf{H}_T^H(l)$. In practice, we hence require an estimate $\hat{\mathbf{H}}_T(l)$ of $\mathbf{H}_T(l)$, cf. also Sec. 3.4.2. With $\mathbf{y}(l)$ as in (3.4), combining (3.10)–(3.12), the MF output $q(l)$ becomes

$$\begin{aligned} \boxed{q(l) = \mathbf{g}^H(l)\mathbf{y}(l)} \\ = s_T(l) + \mathbf{g}^H(l)\big(\mathbf{x}_{e|T'}(l) + \mathbf{x}_\ell(l) + \mathbf{v}(l)\big). \end{aligned} \tag{3.19}$$

The BM $\mathbf{B}(l) \in \mathbb{C}^{M\times M-N_T}$ must be orthogonal to $\mathbf{H}_T(l)$, i.e.

$$\mathbf{B}^H(l)\mathbf{H}_T(l) = \mathbf{0}, \tag{3.20}$$

and with (3.18) hence $\mathbf{B}^H(l)\mathbf{g}(l) = \mathbf{0}$. One may, e.g., choose $\mathbf{B}(l)$ based on the first $M - N_T$ columns of the rank-$(M - N_T)$ projection matrix to the null space of $\mathbf{H}_T(l)$ [32], i.e.

$$\mathbf{B}(l) = \big[\mathbf{I} - \mathbf{H}_T(l)\big(\mathbf{H}_T^H(l)\mathbf{H}_T(l)\big)^{-1}\mathbf{H}_T^H(l)\big]_{:,1:M-N_T}, \tag{3.21}$$

with $\mathbf{H}_T(l)\big(\mathbf{H}_T^H(l)\mathbf{H}_T(l)\big)^{-1}\mathbf{H}_T^H(l)$ the projection matrix to the column space of $\mathbf{H}_T(l)$. With $\mathbf{y}(l)$ as in (3.4), combining (3.10)–(3.11), the SC-filter input $\mathbf{u}_{SC}(l) \in \mathbb{C}^{M-N_T}$ is then given by

$$\begin{aligned} \boxed{\mathbf{u}_{SC}(l) = \mathbf{B}^H(l)\mathbf{y}(l)} \\ = \mathbf{B}^H(l)\big(\mathbf{x}_{e|T'}(l) + \mathbf{x}_\ell(l) + \mathbf{v}(l)\big), \end{aligned} \tag{3.22}$$

whereby the target component $\mathbf{x}_{e|T}(l) = \mathbf{H}_T(l)\mathbf{s}_T(l)$ is canceled. Using a delay of one[1] frame, the LP-filter input $\mathbf{u}_{LP}(l) \in \mathbb{C}^{(L-1)M}$ is defined by stacking $\mathbf{y}(l)$

[1] In MCLP literature, delays of more than one frame are commonly used [94–96, 99, 100, 103, 111–113, 116] in order to avoid temporal target component leakage due to overlapping windows

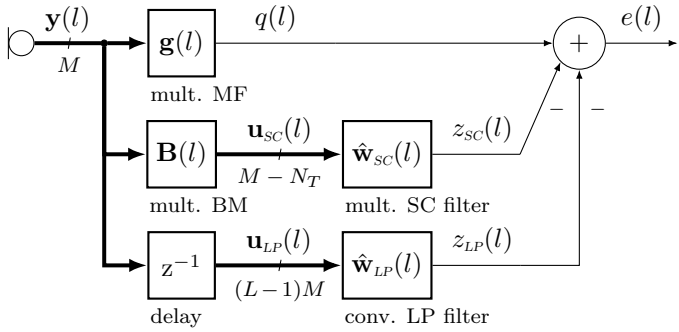


Figure 3.1: The integrated sidelobe cancellation and linear prediction (ISCLP) architecture.

over the past $L-1$ frames, i.e.

$$\boxed{\mathbf{u}_{LP}(l) = \Big(\mathbf{y}^T(l-1) \ \cdots \ \mathbf{y}^T(l-L+1)\Big)^T.} \tag{3.23}$$

With the SC filter $\hat{\mathbf{w}}_{SC}(l) \in \mathbb{C}^{M-N_T}$ and the LP filter $\hat{\mathbf{w}}_{LP}(l) \in \mathbb{C}^{(L-1)M}$, the enhanced signal $e(l) = \hat{s}_T(l)$ at the output of ISCLP, also referred to as error signal in the remainder, is given by

$$e(l) = \hat{s}_T(l) = q(l) - z_{SC}(l) - z_{LP}(l), \tag{3.24}$$

$$\text{with} \quad z_{SC}(l) = \hat{\mathbf{w}}_{SC}^H(l)\mathbf{u}_{SC}(l), \tag{3.25}$$

$$z_{LP}(l) = \hat{\mathbf{w}}_{LP}^H(l)\mathbf{u}_{LP}(l). \tag{3.26}$$

At this point, given $q(l)$, $\mathbf{u}_{SC}(l)$, and $\mathbf{u}_{LP}(l)$, our task consists in obtaining the filters $\hat{\mathbf{w}}_{SC}(l)$ and $\hat{\mathbf{w}}_{LP}(l)$ as estimates of some yet to be defined associated true states $\mathbf{w}_{SC}(l)$ and $\mathbf{w}_{LP}(l)$, cf. Sec 3.3.2. In this respect, let us first discuss the mutual relations between the target component $s_T(l)$ in $q(l)$ and the signals $\mathbf{u}_{SC}(l)$ and $\mathbf{u}_{LP}(l)$, as well as the consequences thereof for the filter estimation. Note that due to the delay in the LP path, the filter estimates $\hat{\mathbf{w}}_{SC}(l)$ and $\hat{\mathbf{w}}_{LP}(l)$ do not operate on the same input-data frame at the same time. The SC-filter input $\mathbf{u}_{SC}(l)$ in (3.22) depends on the current frame $\mathbf{y}(l)$ only, such that $\hat{\mathbf{w}}_{SC}(l)$ will exploit spatial correlations within the current frame. Due to the cancellation of $\mathbf{x}_{e|T}(l)$ at the BM output and (3.15), we have $\mathrm{E}[\mathbf{u}_{SC}(l)s_T^*(l)] = \mathbf{0}$. This allows for unconstrained, recursive estimation of $\mathbf{w}_{SC}(l)$, which is indeed the general

in the STFT processing, cf. Sec. 3.4.1. As we here also consider interfering reverberant speech components to be canceled, larger delays in the LP filter path however call for a convolutive SC filter [163, cf. Ch. 2] instead. The design proposed in this contribution did not show sensitivity to leakage effects, cf. Sec. 3.4.1 and Sec. 3.5.

incentive behind the usage of GSC-like structures [32, 42]. In contrast, the LP-filter input $\mathbf{u}_{LP}(l)$ in (3.23) depends on the $L-1$ previous frames $\mathbf{y}(l-l')$ with $l'=1,\dots,L-1$, such that $\hat{\mathbf{w}}_{LP}(l)$ will exploit spatio-temporal correlations between the current and the previous frames (but not within the current frame). Due to this delay and (3.13), we have $\mathrm{E}[\mathbf{u}_{LP}(l)s_T^*(l)]=\mathbf{0}$, likewise allowing for unconstrained, recursive estimation of $\mathbf{w}_{LP}(l)$. However, with both $\mathbf{u}_{SC}(l)$ and $\mathbf{u}_{LP}(l)$ containing (late-)reverberant components, the two inputs are *not* independent, i.e.

$$\boxed{\mathrm{E}[\mathbf{u}_{LP}(l)\mathbf{u}_{SC}^H(l)]\neq\mathbf{0},} \tag{3.27}$$

cf. (3.14), and as a consequence also $\mathrm{E}[z_{LP}(l)z_{SC}^*(l)]\neq 0$. In other words, a change in $\hat{\mathbf{w}}_{SC}(l)$ requires a change in $\hat{\mathbf{w}}_{LP}(l)$, and vice versa. We therefore strive to *jointly* estimate both filters.

3.3.2 ISCLP State-Space Model and Kalman Filter Update

In order to recursively estimate the SC and LP filter, we employ a Kalman filter, which has also been applied successfully to MCLP in previous works [98, 99, 102, 116]. Hereby, we interpret $\hat{\mathbf{w}}_{SC}(l)$ and $\hat{\mathbf{w}}_{LP}(l)$ as estimates of the true states $\mathbf{w}_{SC}(l)$ and $\mathbf{w}_{LP}(l)$, which are defined by a state-space model comprising the so-called measurement equation and the process equation. In the following, we first define the state-space model, and then present the corresponding Kalman filter update equations, which recursively estimate the true state.

As we intend to estimate $\mathbf{w}_{SC}(l)$ and $\mathbf{w}_{LP}(l)$ jointly, cf. Sec. 3.3.1, we stack the SC and LP filter path into $\mathbf{u}(l)\in\mathbb{C}^{LM-N_T}$ and $\mathbf{w}(l)\in\mathbb{C}^{LM-N_T}$, i.e.

$$\mathbf{u}(l)=\begin{pmatrix}\mathbf{u}_{SC}^T(l) & \mathbf{u}_{LP}^T(l)\end{pmatrix}^T, \tag{3.28}$$

$$\mathbf{w}(l)=\begin{pmatrix}\mathbf{w}_{SC}^T(l) & \mathbf{w}_{LP}^T(l)\end{pmatrix}^T. \tag{3.29}$$

and $\hat{\mathbf{w}}(l)$ defined similarly to (3.29). The true state $\mathbf{w}(l)$ is considered a random variable with zero mean and correlation matrix $\mathbf{\Psi}_w(l)=\mathrm{E}[\mathbf{w}(l)\mathbf{w}^H(l)]$. We assume that $\mathbf{w}(l)$ leads to complete cancellation[2] of $\mathbf{g}^H(l)\big(\mathbf{x}_{e|T'}(l)+\mathbf{x}_\ell(l)+\mathbf{v}(l)\big)$, and therefore yielding $e(l)=s_T(l)$, cf. (3.19) and (3.24)–(3.26). Reformulating (3.24)–(3.26) using (3.28)–(3.29), inserting $e(l)=s_T(l)$ and rearranging yields the so-called measurement equation,

$$\boxed{q^*(l)=\mathbf{u}^H(l)\mathbf{w}(l)+s_T^*(l).} \tag{3.30}$$

[2]Note that complete cancellation may not necessarily be possible, e.g., if $\mathbf{v}(l)\neq\mathbf{0}$ [163, cf. Ch. 2], and so the true state does not necessarily exist. Nonetheless, lacking detailed knowledge on the true system, we assume that it lies in the model set.

In Kalman filter terminology, we refer to $q^*(l)$ as the measurement and to $s_T^*(l)$ as the (presumed zero-mean and temporally uncorrelated, cf. also Sec. 3.2) measurement noise with PSD $\varphi_{s_T}(l) = \mathrm{E}[s_T(l)s_T^*(l)]$. In practice, in order to implement the Kalman filter update equations, an estimate $\hat{\varphi}_{s_T}(l)$ of $\varphi_{s_T}(l)$ is required, cf. Sec. 3.4.2.

The true state $\mathbf{w}(l)$ is assumed time-varying, which accounts for potential time variations in the room impulse responses (RIRs), e.g., caused by time-varying source and microphone-array positions, as well as time-varying activity of individual sources and noise powers. The so-called process equation models the evolution of the true state $\mathbf{w}(l)$ in the form of a first-order difference equation, i.e.

$$\boxed{\mathbf{w}(l) = \mathbf{A}^H(l)\mathbf{w}(l-1) + \mathbf{w}_\Delta(l).} \tag{3.31}$$

where $\mathbf{A}(l)$ models the state transition from one frame to the next, and the process noise $\mathbf{w}_\Delta(l)$ models a random (presumed zero-mean and temporally uncorrelated) variation component with correlation matrix $\mathbf{\Psi}_{w_\Delta}(l) = \mathrm{E}[\mathbf{w}_\Delta(l)\mathbf{w}_\Delta^H(l)]$. Lacking detailed knowledge on the exact evolution of the true state, both $\mathbf{A}(l)$ and $\mathbf{\Psi}_{w_\Delta}(l)$ are commonly considered design parameters to be tuned [99, 102, 116, 165], cf. Sec. 3.4.3.

The true state $\mathbf{w}(l)$ modeled by (3.30)–(3.31) may be estimated recursively by means of the Kalman filter update equations [51, 52], which are commonly presented as two distinct sets of updates per recursion, namely an a-priori time update reflecting the state evolution, cf. (3.31), and an a-posteriori measurement update reflecting the current measurement, cf. (3.30). Specifically, let $\hat{\mathbf{w}}(l)$ and $\hat{\mathbf{w}}^+(l)$ denote the yet to be defined prior and posterior state estimates of $\mathbf{w}(l)$, respectively, and let $\tilde{\mathbf{w}}(l)$ and $\tilde{\mathbf{w}}^+(l)$ denote the associated state estimation errors, i.e.

$$\tilde{\mathbf{w}}(l) = \hat{\mathbf{w}}(l) - \mathbf{w}(l), \tag{3.32}$$

$$\tilde{\mathbf{w}}^+(l) = \hat{\mathbf{w}}^+(l) - \mathbf{w}(l), \tag{3.33}$$

with the associated state estimation error correlation matrices $\mathbf{\Psi}_{\tilde{w}}(l)$ and $\mathbf{\Psi}_{\tilde{w}}^+(l)$. Then, based upon (3.31) and (3.30), respectively, the prior and posterior state estimates $\hat{\mathbf{w}}(l)$ and $\hat{\mathbf{w}}^+(l)$ shall recursively minimize the expected squared Euclidian norm of the associated state estimation error, i.e. $\mathrm{E}\big[\|\tilde{\mathbf{w}}(l)\|^2\big] = \mathrm{tr}[\mathbf{\Psi}_{\tilde{w}}(l)]$ and $\mathrm{E}\big[\|\tilde{\mathbf{w}}^+(l)\|^2\big] = \mathrm{tr}[\mathbf{\Psi}_{\tilde{w}}^+(l)]$. This leads to the celebrated Kalman filter update equations [51, 52],

$$\hat{\mathbf{w}}(l) = \mathbf{A}^H(l)\hat{\mathbf{w}}^+(l-1), \tag{3.34}$$

$$\mathbf{\Psi}_{\tilde{w}}(l) = \mathbf{A}^H(l)\mathbf{\Psi}_{\tilde{w}}^+(l-1)\mathbf{A}(l) + \mathbf{\Psi}_{w_\Delta}(l), \tag{3.35}$$

$$e^*(l) = q^*(l) - \mathbf{u}^H(l)\hat{\mathbf{w}}(l), \tag{3.36}$$

$$\varphi_e(l) = \mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l)\mathbf{u}(l) + \varphi_{s_T}(l), \tag{3.37}$$

$$\mathbf{k}(l) = \mathbf{\Psi}_{\tilde{w}}(l)\mathbf{u}(l)\varphi_e^{-1}(l), \tag{3.38}$$

$$\hat{\mathbf{w}}^+(l) = \hat{\mathbf{w}}(l) + \mathbf{k}(l)e^*(l), \tag{3.39}$$

$$\mathbf{\Psi}_{\tilde{w}}^+(l) = \mathbf{\Psi}_{\tilde{w}}(l) - \mathbf{k}(l)\mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l), \tag{3.40}$$

where the time and the measurement update are given by (3.34)–(3.35) and (3.39)–(3.40), respectively. In the time update, cf. (3.34)–(3.35), the previously acquired posterior quantities $\hat{\mathbf{w}}^+(l-1)$ and $\mathbf{\Psi}_{\tilde{w}}^+(l-1)$ are propagated according to the evolution of the state $\mathbf{w}(l)$, cf. (3.31), yielding the prior quantities $\hat{\mathbf{w}}(l)$ and $\mathbf{\Psi}_{\tilde{w}}(l)$. Then, given $\hat{\mathbf{w}}(l)$ and $\mathbf{\Psi}_{\tilde{w}}(l)$, the complex conjugate error signal $e^*(l)$, its PSD $\varphi_e(l)$, and the Kalman gain $\mathbf{k}(l)$ are computed, cf. (3.36)–(3.38), thereby leveraging new information in terms of the measurement $q^*(l)$ and the measurement noise PSD $\varphi_{s_T}(l)$, cf. (3.30). Finally, in the measurement update, cf. (3.39)–(3.40), $e^*(l)$ and $\mathbf{k}(l)$ are utilized to update $\hat{\mathbf{w}}(l)$ and $\mathbf{\Psi}_{\tilde{w}}(l)$, yielding the posterior quantities $\hat{\mathbf{w}}^+(l)$ and $\mathbf{\Psi}_{\tilde{w}}^+(l)$. The error signal $e(l)$ in (3.36) thereby represents the Kalman filter estimate of $s_T(l)$, cf. also (3.24)–(3.26). As the Kalman filter minimizes $\mathrm{tr}[\mathbf{\Psi}_{\tilde{w}}(l)]$ during convergence, it is easily seen that also $\varphi_e(l) = \mathrm{E}\big[|e(l)|^2\big]$ in (3.37) is minimized. The Kalman filter requires initialization, which we consider in Sec. 3.4.3.

3.3.3 Posterior-like Spectral Post-Processing

With $\hat{\mathbf{w}}(l)$ a prior estimate of $\mathbf{w}(l)$, we may consider $e(l) = \hat{s}_T(l)$ in (3.36) a prior estimate of $s_T(l)$. After the measurement update in (3.39), yielding the posterior estimate $\hat{\mathbf{w}}^+(l)$ of $\mathbf{w}(l)$, we may accordingly define a posterior estimate $e^+(l) = \hat{s}_T^+(l)$ similar to (3.36) by

$$e^{*|+}(l) = \hat{s}_T^{*|+}(l) = q^*(l) - \mathbf{u}^H(l)\hat{\mathbf{w}}^+(l). \tag{3.41}$$

Interestingly, $e^+(l)$ in (3.41) can be shown to be a spectrally post-processed version of $e(l)$. Specifically, inserting (3.39) while using (3.36), inserting (3.38) and finally (3.37), we find

$$\begin{aligned} e^+(l) = \hat{s}_T^+(l) &= \big(1 - \mathbf{u}^H(l)\mathbf{k}(l)\big)^* e(l) \\ &= \big(1 - \mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l)\mathbf{u}(l)\varphi_e^{-1}(l)\big)^* e(l) \\ &= \frac{\varphi_{s_T}(l)}{\varphi_e(l)} e(l), \end{aligned} \tag{3.42}$$

where $\gamma(l) = \varphi_{s_T}(l)/\varphi_e(l)$ can be recognized as the spectral Wiener gain minimizing $\mathrm{E}\big[|s_T(l) - \gamma(l)e(l)|^2\big]$. In practice, where we rely on potentially highly non-stationary estimates $\hat{\varphi}_{s_T}(l)$, cf. Sec. 3.4.1 and Sec. 3.4.2, one may prefer slowly decaying gains for perceptual reasons [30]. Therefore, instead of using (3.42), we propose to alternatively define $\gamma(l)$ and $e^+(l)$ by

$$\gamma(l) = \max\left[\frac{\varphi_{s_T}(l)}{\varphi_e(l)}, \beta\gamma(l-1)\right], \tag{3.43}$$

$$e^+(l) = \hat{s}_T^+(l) = \gamma(l)e(l), \tag{3.44}$$

with the tuning parameter $\beta \in [0, 1]$ limiting the gain decay. Note that (3.43)–(3.44) reduce to (3.42) for $\beta = 0$, and to (3.36) for $\beta = 1$ and $\gamma(0) = 1$ as initial gain, since $\varphi_{s_T}(l)/\varphi_e(l) \leq 1$ due to (3.37).

3.4 Implementational Aspects

Kalman filters perform optimally if the assumed state-space model matches the true system [51, 52]. In a practical implementation, the here presented ISCLP Kalman filter derived from the ISCLP state-space model in (3.30)–(3.31) is subject to modeling errors, requires parameter estimation, and, where detailed knowledge on the underlying system dynamics is not available, parameter tuning. These implementational aspects are discussed in the following. In Sec. 3.4.1, we qualitatively discuss the potential target component leakage due to imperfect spatio-temporal pre-processing in ISCLP and its impact on the proposed ISCLP Kalman filter. In Sec. 3.4.2, we summarize a recently proposed approach to early PSD estimation and recursive RETF updating, which we employ in conjunction with the Kalman filter. In Sec. 3.4.3, we discuss the process equation parameter tuning and Kalman filter initialization.

3.4.1 Spatio-Temporal Target Component Leakage

The previously made assumptions that $\mathrm{E}[\mathbf{u}_{SC}(l)s_T^*(l)] = \mathbf{0}$ and $\mathrm{E}[\mathbf{u}_{LP}(l)s_T^*(l)] = \mathbf{0}$, cf. Sec. 3.3.1, may not be strictly satisfied in a practical implementation, which we refer to as *target component leakage*. Leakage may occur due to the following reasons. The spatial pre-processing components MF and BM rely on spatial information in terms of the RETFs $\mathbf{H}_T(l)$, cf. (3.18) and (3.21), which needs to be estimated in practice. The estimate $\hat{\mathbf{H}}_T(l)$ commonly contains estimation errors, i.e. we have $\hat{\mathbf{H}}_T(l) \neq \mathbf{H}_T(l)$. Further, the RETF-based data model in (3.7) itself may be erroneous, e.g., due to dependencies across frequency bins [117].

Finally, the assumption in (3.15) that $\mathbf{x}_{n|e}(l)$ and $\mathbf{x}^{H}_{n|\ell}(l)$ are uncorrelated may be violated, e.g., due to overlapping windows in the STFT processing. In general, these estimation and modeling errors cause incomplete blocking and therefore target component leakage through the BM, such that $\mathrm{E}[\mathbf{u}_{SC}(l)s^{*}_{T}(l)] \neq \mathbf{0}$, cf. (3.19), (3.22). This may be referred to as *spatial target component leakage*. Similarly, if $s_T(l)$ is temporally correlated such that (3.13) is violated, e.g., due to overlapping windows in the STFT processing or to too small window lengths and shifts, we find $\mathrm{E}[\mathbf{u}_{LP}(l)s^{*}_{T}(l)] \neq \mathbf{0}$, cf. (3.19), (3.23), which may be referred to as *temporal target component leakage*.

Potentially, spatial and temporal leakage cause a biased [163, cf. Ch. 2] filter estimate $\hat{\mathbf{w}}(l)$, which leads to partial suppression of $s_T(l)$, also referred to as speech cancellation in GSC terminology [32, 42], or excessive whitening in MCLP terminology [89]. However, note that the Kalman filter offers inherent robustness towards target-component leakage. To see this, consider the measurement update terms in (3.39)–(3.40), respectively given by $\mathbf{k}(l)e^{*}(l)$ and $\mathbf{k}(l)\mathbf{u}^{H}(l)\mathbf{\Psi}_{\tilde{w}}(l)$. Using (3.30) and (3.32), we may express $e^{*}(l)$ in (3.36) in terms of $s^{*}_{T}(l)$, while using (3.37), we may similarly express $\mathbf{k}(l)$ in (3.38) in terms of $\varphi_{s_T}(l)$, i.e.

$$e^{*}(l) = s^{*}_{T}(l) - \mathbf{u}^{H}(l)\tilde{\mathbf{w}}(l), \tag{3.45}$$

$$\mathbf{k}(l) = \frac{\mathbf{\Psi}_{\tilde{w}}(l)\mathbf{u}(l)}{\mathbf{u}^{H}(l)\mathbf{\Psi}_{\tilde{w}}(l)\mathbf{u}(l) + \varphi_{s_T}(l)}. \tag{3.46}$$

From (3.45)–(3.46), we note that $\varphi_{s_T}(l) = \mathrm{E}[s_T(l)s^{*}_{T}(l)]$ acts as a regularization parameter in both update terms $\mathbf{k}(l)e^{*}(l)$ and $\mathbf{k}(l)\mathbf{u}^{H}(l)\mathbf{\Psi}_{\tilde{w}}(l)$. Consequently, strong target powers inhibit the measurement update, while weak target powers promote it. Put differently, in terms of robustness towards target-component leakage and convergence, the Kalman filter benefits from non-stationarities and sparsity in $\varphi_{s_T}(l)$ across time. Note that in recursive MCLP implementations based on the weighted prediction error (WPE) criterion and RLS [92, 100, 103, 111], the target-component PSD similarly appears as a regularization term in the update equations.

In practice, we rely on estimates $\hat{\varphi}_{s_T}(l)$, which should hence maintain non-stationarities. In WPE RLS literature, the target-component PSD estimate is obtained, e.g., directly from the plain microphone signals [92], based on a late-reverberant PSD estimate obtained by means of an exponential decay model [100, 111], or using a neural network [103]. Here, as we consider a more generic signal model comprising several reverberant speech components and diffuse noise, cf. Sec. 3.2, we instead estimate $\hat{\varphi}_{s_T}(l)$ by means of [164, cf. Ch. 4], cf. Sec. 3.4.2.

3.4.2 Target PSD Estimation and RETF Update

We require an RETF estimate $\hat{\mathbf{H}}_T(l)$ of $\mathbf{H}_T(l)$, cf. (3.18) and (3.21), and a PSD estimate $\hat{\varphi}_{s_T}(l)$ of $\varphi_{s_T}(l)$, cf. (3.37) and (3.43). To this end, we use an algorithm recently proposed in [164, cf. Ch. 4] by the authors of this paper, which computes early PSD estimates and recursively updates the RETF estimates for all N point sources. The algorithm [164, cf. Ch. 4] is summarized as follows.

Let $\mathbf{\Psi}_{x_e}(l) = \mathrm{E}[\mathbf{x}_e(l)\mathbf{x}_e^H(l)]$ denote the correlation matrix of $\mathbf{x}_e(l)$ within frame l, which generally has rank N and is given by

$$\mathbf{\Psi}_{x_e}(l) = \mathbf{H}(l)\,\mathrm{Diag}[\boldsymbol{\varphi}_s(l)]\mathbf{H}^H(l), \tag{3.47}$$

$$\boldsymbol{\varphi}_s(l) = \Big(\varphi_{s_1}(l) \;\cdots\; \varphi_{s_N}(l)\Big)^T, \tag{3.48}$$

with $\varphi_{s_n}(l)$ denoting the PSD of the early speech source image $s_n(l)$. Instead of directly using the conventional early correlation matrix model in (3.47), the algorithm in [164, cf. Ch. 4] is based on its factorization, i.e. it relies on the square-root model

$$\mathbf{\Psi}_{x_e}^{1/2}(l)\mathbf{\Omega}(l) = \mathbf{H}(l)\,\mathrm{Diag}[\boldsymbol{\varphi}^{1/2}(l)], \tag{3.49}$$

where $\mathbf{\Psi}_{x_e}^{1/2}(l) \in \mathbb{C}^{M\times N}$ and $\boldsymbol{\varphi}^{1/2} \in \mathbb{C}^N$ are some square roots of $\mathbf{\Psi}_{x_e}(l)$ and $\boldsymbol{\varphi}_s(l)$ such that $\mathbf{\Psi}_{x_e}^{1/2}(l)\mathbf{\Psi}_{x_e}^{H/2}(l) = \mathbf{\Psi}_{x_e}(l)$ and $\mathrm{Diag}[\boldsymbol{\varphi}^{H/2}(l)]\boldsymbol{\varphi}^{1/2}(l) = \boldsymbol{\varphi}_s(l)$, respectively, and $\mathbf{\Omega}(l)$ is a unitary matrix, i.e. $\mathbf{\Omega}(l)\mathbf{\Omega}^H(l) = \mathbf{I}$, which accounts for the non-uniqueness of both square-roots. Note that right-multiplying each side of (3.49) with its Hermitian yields (3.47). In the estimation, we distinguish the prior and posterior RETF estimates $\hat{\mathbf{H}}(l)$ and $\hat{\mathbf{H}}^+(l)$, respectively, and assume that initial RETF estimates $\hat{\mathbf{H}}(0)$ are available, which may be based on, e.g., initial single-source RETF estimates acquired from segments with distinctly active sources [147], or some initial knowledge or estimates of the associated dicrections of arrival (DoAs) [145, 146]. Given a (to be obtained) square-root estimate $\hat{\mathbf{\Psi}}_{x_e}^{1/2}(l)$ and a prior RETF estimate $\hat{\mathbf{H}}(l)$, which is propagated from the previous posterior, i.e. $\hat{\mathbf{H}}(l) = \hat{\mathbf{H}}^+(l-1)$, we first obtain the unitary and diagonal estimates $\hat{\mathbf{\Omega}}(l)$ and $\mathrm{Diag}[\hat{\boldsymbol{\varphi}}^{1/2}]$, yielding $\hat{\boldsymbol{\varphi}}_s(l) = \mathrm{Diag}[\hat{\boldsymbol{\varphi}}^{H/2}]\hat{\boldsymbol{\varphi}}^{1/2}$, and based on these estimates second update the RETF estimate, yielding the posterior $\hat{\mathbf{H}}^+(l)$, whereat the recursion is closed. Here, both steps are based on approximation error minimization with respect to the square-root model in (3.49). Given $\hat{\boldsymbol{\varphi}}_s(l)$ and $\hat{\mathbf{H}}^+(l)$, we extract $\hat{\varphi}_{s_T}(l)$ and $\hat{\mathbf{H}}_T(l)$ as $\hat{\varphi}_{s_T}(l) = \mathbf{1}^T[\hat{\boldsymbol{\varphi}}_s(l)]_{\in T}$ and $\hat{\mathbf{H}}_T(l) = [\hat{\mathbf{H}}^+(l)]_{\in T}$, cf. Sec. 3.2.

The said required square root $\mathbf{\Psi}_{x_e}^{1/2}(l)$ is estimated in the following manner. While $\mathbf{x}_{n|\ell}(l)$ and $\mathbf{v}(l)$ exhibit a fundamentally different temporal behavior

across frames, cf. Sec. 3.2, we assume that their spatial behavior within frame l is the same. Specifically, we model both $\mathbf{x}_{n|\ell}(l)$ and $\mathbf{v}(l)$ as spatially diffuse with coherence matrix $\mathbf{\Gamma} \in \mathbb{C}^{M\times M}$, which may be computed from the microphone array geometry [132,166] and is therefore assumed to be known. For the late reverberant component $\mathbf{x}_{n|\ell}(l)$, this is a commonly made assumption [67,68,132]. For the noise component $\mathbf{v}(l)$, the assumption is commonly made for noise types such as, e.g., babble noise [167], which we use in our simulations, cf. Sec. 3.5. Based on these assumptions, the microphone signal correlation matrix $\mathbf{\Psi}_y(l) = \mathrm{E}[\mathbf{y}(l)\mathbf{y}^H(l)]$ may be written as

$$\mathbf{\Psi}_y(l) = \mathbf{\Psi}_{x_e}(l) + \varphi_d(l)\mathbf{\Gamma}, \tag{3.50}$$

with $\varphi_d(l) = \sum_{n=1}^{N} \varphi_{x_{n|\ell}}(l) + \varphi_v(l)$ and $\varphi_{x_{n|\ell}}(l)$ and $\varphi_v(l)$ denoting the PSDs of the late-reverberant speech components and the diffuse noise component, respectively. We obtain a subspace representation of (3.50) by means of the generalized eigenvalue decomposition (GEVD) of $\mathbf{\Psi}_y(l)$ and $\mathbf{\Gamma}$. Based on the generalized eigenvectors and generalized eigenvalues, $\mathbf{\Psi}_y(l)$ may be decomposed into a diffuse component, cf. also the diffuse PSD estimator in [68], and a factorized early rank-N component $\mathbf{\Psi}_{x_e}(l) = \mathbf{\Psi}_{x_e}^{1/2}(l)\mathbf{\Psi}_{x_e}^{H/2}(l)$. A temporally smooth estimate $\hat{\mathbf{\Psi}}_{y|sm}(l)$ of $\mathbf{\Psi}_y(l)$ itself is obtained from the microphone signals by recursively averaging $\mathbf{y}^H(l)\mathbf{y}(l)$. In order to restore non-stationarities, we desmooth[3] the generalized eigenvalues of $\hat{\mathbf{\Psi}}_{y|sm}(l)$ and $\mathbf{\Gamma}$ and thereby yield non-stationary PSD estimates in the subsequent processing steps, as favored in the Kalman filter, cf. Sec 3.4.1. For further details, we refer the interested reader to [164, cf. Ch. 4].

3.4.3 Process Equation Parameter Tuning and Initialization

The tracking and convergence behavior of the Kalman filter depends on its process equation parameter tuning and initialization. The process equation models the evolution of the state by means of the parameters $\mathbf{A}(l)$ and $\mathbf{\Psi}_{w_\Delta}(l)$, cf. (3.31) and (3.34)–(3.35). In practice, only limited knowledge of the state evolution is available, such that $\mathbf{A}(l)$ and $\mathbf{\Psi}_{w_\Delta}(l)$ are commonly left to tuning [99, 102, 116, 165]. Typically, both $\mathbf{A}(l)$ and $\mathbf{\Psi}_{w_\Delta}(l)$ are chosen to be scaled identities, with $\mathbf{A}(l)$ commonly time-invariant [99, 102, 116, 165] and acting as a forgetting factor [102, 165], and $\mathbf{\Psi}_{w_\Delta}(l)$ either time-variant [99, 116, 165] or time-invariant [102]. Here, we set $\mathbf{A}(l)$ and $\mathbf{\Psi}_{w_\Delta}(l)$ based on the assumption that the state correlation matrix $\mathbf{\Psi}_w(l)$ is time-invariant, i.e. $\mathbf{\Psi}_w(l) = \mathbf{\Psi}_w$. Unfortunately, $\mathbf{\Psi}_w$ is unknown and not available in practice, however, we may

[3] Considering recursive averaging as an invertible recursive filtering operation, the generalized eigenvalues may be desmoothed by means of the corresponding inverse filter.

define a rough guess $\bar{\mathbf{\Psi}}_w$. Given such a guess $\bar{\mathbf{\Psi}}_w$, by means of a forgetting factor $\alpha \in (0,1)$, we may account for a steadily time-varying acoustic scenario and true state $\mathbf{w}(l)$ by setting

$$\mathbf{A}(l) = \sqrt{\alpha}\,\mathbf{I}, \tag{3.51}$$

$$\mathbf{\Psi}_{w_\Delta}(l) = (1-\alpha)\bar{\mathbf{\Psi}}_w, \tag{3.52}$$

such that if $\bar{\mathbf{\Psi}}_w = \mathbf{\Psi}_w$, we rightly have $\mathbf{\Psi}_w = \alpha\mathbf{\Psi}_w + (1-\alpha)\mathbf{\Psi}_w$ from (3.31). While $\bar{\mathbf{\Psi}}_w$ may rather be defined by design than by truly estimating $\mathbf{\Psi}_w$, the notion of $\bar{\mathbf{\Psi}}_w$ being a rough guess of $\mathbf{\Psi}_w$ may nonetheless guide its definition to some extent. Here, we choose a diagonal matrix with distinct diagonal elements. With $\bar{\mathbf{\Psi}}_w = \mathrm{Diag}[\bar{\boldsymbol{\psi}}_w]$, let $\bar{\boldsymbol{\psi}}_{w_{SC}} \in \mathbb{R}^{M-N_T}$ and $\bar{\boldsymbol{\psi}}_{w_{LP}} \in \mathbb{R}^{(L-1)M}$ denote the subvectors of $\bar{\boldsymbol{\psi}}_w$ associated to the SC and the LP filter, respectively, which we treat separately. Expecting lower values for later prediction coefficients in the LP filter, we choose the power of the diagonal elements in $\bar{\boldsymbol{\psi}}_{w_{LP}}$ to drop exponentially each M elements, i.e. we set

$$\bar{\boldsymbol{\psi}}_{w_{SC}} = \bar{\psi}_{w_{SC}}\mathbf{1}, \tag{3.53}$$

$$\bar{\boldsymbol{\psi}}_{w_{LP}} = \left(\bar{\psi}_{w_{LP}}^{1}\mathbf{1}^T \;\ldots\; \bar{\psi}_{w_{LP}}^{L-1}\mathbf{1}^T\right)^T, \tag{3.54}$$

with $\bar{\psi}_{w_{SC}} > 0$ and $\bar{\psi}_{w_{LP}} \in (0,1)$ further adjustable.

The matrix $\bar{\mathbf{\Psi}}_w$ may also be used to initialize the Kalman filter. With the commonly chosen initial state estimate $\hat{\mathbf{w}}(0) = \mathbf{0}$, we have $\tilde{\mathbf{w}}(0) = \mathbf{w}(0)$ in (3.32), such that the true initial state estimation error correlation matrix $\mathbf{\Psi}_{\tilde{w}}(0)$ becomes $\mathbf{\Psi}_{\tilde{w}}(0) = \mathbf{\Psi}_w(0) = \mathbf{\Psi}_w$. Therefore, we initialize the Kalman filter by

$$\hat{\mathbf{w}}(0) = \mathbf{0}, \tag{3.55}$$

$$\hat{\mathbf{\Psi}}_{\tilde{w}}(0) = \bar{\mathbf{\Psi}}_w, \tag{3.56}$$

in (3.34)–(3.35), where $\hat{\mathbf{\Psi}}_{\tilde{w}}(0)$ in (3.56) is an estimate if $\bar{\mathbf{\Psi}}_w \neq \mathbf{\Psi}_w$. Finally, note that the process equation parameter tuning in (3.51)–(3.52) may also be considered from a (re-)initialization perspective. In case of meaningful measurement updates, the Kalman filter tracks $\mathbf{w}(l)$, but otherwise tends to return to its initial condition due to (3.51)–(3.52), such that explicit re-initialization as, e.g., in case of a sudden change in the acoustic environment, is not necessary. To see this, consider the case where, e.g., $\mathbf{u}(l) = \mathbf{0}$ for a period of time, such that no measurement update is performed. In this case, regardless of their current values, we have $\hat{\mathbf{w}}(l)$ slowly converging to $\mathbf{0}$ and $\hat{\mathbf{\Psi}}_{\tilde{w}}(l)$ slowly converging to $\bar{\mathbf{\Psi}}_w$, cf. (3.34)–(3.35). Note that if desired, explicit re-initialization may still easily be incorporated in the proposed concept, namely by defining α time-variant and setting it to zero at the determined re-initialization point.

3.5 Simulations

In order to demonstrate the effectiveness of the presented ISCLP Kalman filter, we define two case studies, case A and case B. In case A, we compare to the (computationally more demanding) alternating Kalman filters proposed in [116]. Here, we consider one reverberant speech and a babble noise component, $\mathbf{x}_1(l)$ and $\mathbf{v}(l)$, with $\mathbf{x}_1(l)$ containing the target component $\mathbf{x}_{e|T}(l) = \mathbf{x}_{1|e}(l)$. In case B, we compare to a (computationally more demanding) MCLP+GSC Kalman filter cascade, which conceptually relates to [112,113] in that it cascades linear prediction and beamforming. Here, we consider two reverberant speech components and a babble noise component, $\mathbf{x}_1(l)$, $\mathbf{x}_2(l)$, and $\mathbf{v}(l)$, with $\mathbf{x}_1(l)$ again containing the target component $\mathbf{x}_{e|T}(l) = \mathbf{x}_{1|e}(l)$, and $\mathbf{x}_2(l)$ an interfering speech component to be canceled. In both cases, we investigate the algorithms' behavior depending on the signal-to-noise ratio, *SNR*, which is defined as the power ratio of $\mathbf{x}_1(l)$ to $\mathbf{v}(l)$, and depending on the filter length L. In case A, we additionally investigate the convergence behavior.

In what follows, we describe the two reference algorithms in more detail in Sec. 3.5.1, the performance measures in Sec. 3.5.2, the acoustic scenario in Sec. 3.5.3, the algorithmic settings in 3.5.4, and finally the simulation results in Sec. 3.5.5.

3.5.1 Reference Algorithms

We discuss the alternating Kalman filters in Sec. 3.5.1.1 and the MCLP+GSC Kalman filter cascade in Sec. 3.5.1.2.

3.5.1.1 Case A: Alternating Kalman Filters

In [116], MCLP-based dereverberation and noise reduction is performed in each microphone channel using two alternating Kalman filters. The Kalman filter dedicated to dereverberation estimates a multiple-output LP filter, and the Kalman filter dedicated to noise reduction estimates the noise-free reverberant speech component. The enhanced signal is computed from the posterior[4] state estimates of both Kalman filters. The two state vectors have dimensions $M^2(L-1)$ and $M(L-1)$, respectively, while the ISCLP Kalman filter requires a single state vector with dimension $ML - N_T$ only with $N_T = 1$ in case A, cf. (3.29). Since the Kalman filter in general exhibits a quadratic computational cost in the state vector dimension and the alternating Kalman filters requires

[4] which corresponds to spectral post-processing, cf. Sec. 3.3.3.

a roughly M times larger state vector, the alternating Kalman filters are computationally roughly M^2 times as demanding as the ISCLP Kalman filter.[5] The two state space models do not provide a spatial distinction between point sources (and therefore do not require RETF estimates, as opposed to the ISCLP Kalman filter) and further do not consider temporally correlated interference components such as interfering reverberant speech. We hence set $\mathbf{x}_2(l) = \mathbf{0}$ when comparing to [116], i.e. interfering speech is absent, cf. Sec. 3.5.3.2.

The alternating Kalman filters require correlation matrix estimates of the measurement and process noises, more precisely of the random variation of the multiple-output LP filter state, comparable to $\mathbf{\Psi}_{w_\Delta}(l)$ in the ISCLP Kalman filter, cf. (3.31), the early component $\mathbf{\Psi}_{x_e|T}(l) = \mathbf{\Psi}_{x_e}(l)$, the early-plus-noise component $\mathbf{\Psi}_{x_e}(l) + \mathbf{\Psi}_v(l)$, and the noise component $\mathbf{\Psi}_v(l)$ [116]. In the original implementation in [116], a time-invariant estimate $\hat{\mathbf{\Psi}}_v$ is assumed to be available, which we here compute in an oracle fashion from $\mathbf{v}(l)$ directly, while the other correlation matrices are estimated based on the previous state estimates and error signals of the alternating Kalman filters. For the sake of a fair and more meaningful comparison, we implement two versions of [116]. The first version is implemented as proposed in [116] and discussed above, subsequently referred to as the original alternating Kalman filters. In the second version, we align the parameter estimation and tuning towards the proposed approach, i.e. $\mathbf{\Psi}_{x_e}(l)$ is instead estimated based on [164, cf. Ch. 4], cf. Sec. 3.4.2, and the process equation parameters modeling the evolution of the multiple-output LP filter state are defined similarly to Sec. 3.4.3, subsequently referred to as the modified alternating Kalman filters.

3.5.1.2 Case B: MCLP+GSC Kalman Filter Cascade

In [112, 113], multiple-output MCLP based on the (iterative) WPE criterion [94, 95] is cascaded with MVDR beamforming in order to reduce noise after dereverberation, which became a popular approach in the CHiME-5 challenge [49]. For the sake of a close comparison, however, we here instead compare to a (recursive) multiple-output MCLP-based Kalman filter cascaded with a (recursive) GSC-based Kalman filter, subsequently referred to as MCLP+GSC. Herein, we estimate the LP and SC filters independently. The enhanced signal at the GSC output is computed using spectral post-processing of the same kind as in (3.43)–(3.44). The two state vectors have dimensions $M^2(L-1)$ and $M - N_T$, respectively, while the ISCLP Kalman filter requires a single state vector with dimension $ML - N_T$ only with $N_T = 1$ in case B, cf. (3.29). Since the Kalman filter in general exhibits a quadratic computational cost in the

[5] leaving aside the computational complexity of parameter estimation, which depends on the employed parameter estimation algorithms.

state vector dimension and the MCLP+GSC Kalman filter cascade requires a roughly M times larger state vector, the MCLP+GSC Kalman filter cascade is computationally roughly M^2 times as demanding as the ISCLP Kalman filter. The GSC state space model does provide a spatial distinction between point sources (based on an RETF estimate, as the ISCLP Kalman filter). We hence set $\mathbf{x}_2(l) \neq \mathbf{0}$ when comparing to the MCLP+GSC Kalman filter cascade, i.e. interfering speech is present, cf. Sec. 3.5.3.2.

The MCLP and GSC Kalman filters require correlation matrix estimates of their respective measurement and process noises, more precisely of the random variation of the multiple-output LP filter and SC filter state, respectively, defined similarly to the corresponding SC and LP submatrices of $\mathbf{\Psi}_{w_\Delta}(l)$ in the ISCLP Kalman filter, cf. (3.31), and the early components $\mathbf{\Psi}_{x_e|T}(l)$ and $\varphi_{s_T}(l)$, respectively, computed based on [164, cf. Ch. 4] as in the proposed ISCLP Kalman filter, cf. Sec. 3.4.2.

3.5.2 Performance Measures

As performance measures, we choose the perceptual evaluation of speech quality [168], *PESQ*, with mean opinion scores of objective listening quality $\in [1, 4.5]$, the short-time objective intelligibility [169], *STOI*, with scores $\in [0, 1]$, the frequency-weighted segmental signal-to-interference ratio [16,30], SIR^{fws}, in dB, and the cepstral distance [16,30], *CD*, in dB. While high values are preferable for *PESQ*, *STOI*, and SIR^{fws}, low values are preferred for *CD*. These intrusive measures require a clean reference signal $\tilde{s}_T(l)$, which approximates the target signal $s_T(l)$ in (3.12). In order to generate $\tilde{s}_T(l)$, we convolve the target speech source signal with the early part of the RIR to the first microphone, cf. Sec 3.5.3, whereat we define the first N_{STFT} samples of the RIR as its early part, with N_{STFT} the analysis and synthesis window length of the STFT processing corresponding to 32 ms, cf. Sec. 3.5.4. Note that due to modeling errors in the RETF-model in (3.7), we generally have $\tilde{s}_T(l) \neq s_T(l)$. When investigating the dependency on *SNR* or L, we compute the measures from 4 s to 10 s, i.e. roughly after convergence. When investigating the convergence behavior, we compute the measures within sliding windows of 2 s each. The computed measures are averaged over several individual simulations, cf. Sec. 3.5.3.

3.5.3 Acoustic Scenario

We describe the acoustic scenarios without and with interfering speaker in Sec. 3.5.3.1 and Sec. 3.5.3.2, respectively.

3.5.3.1 Case A: Without Interfering Speech

In case A, the microphone signals are composed of one reverberant speech and a babble noise component, $\mathbf{x}_1(l)$ and $\mathbf{v}(l)$, with $\mathbf{x}_1(l)$ containing the target component $\mathbf{x}_{e|T}(l) = \mathbf{x}_{1|e}(l)$. To generate $\mathbf{x}_1(l)$, we use measured RIRs of 0.61 s reverberation time to a linear microphone array with $M = 5$ microphones and 8 cm inter-microphone distance [26]. When investigating the dependency on *SNR* or L, the speech source remains positioned in 2 m distance of the microphone array at 0° relative to the broad-side direction during 10 s of simulation. When investigating the convergence behavior, the speech source remains positioned in 2 m distance at 0° for the first 8 s, then jumps to 15°, where it remains for another 10 s. Both female and male speech [25] are used as speech source signals. The babble noise component is generated using [27, 167]. From the speech source signal files and the babble noise file [27], we randomly select individual segments, yielding individual simulations to be averaged in the performance evaluation, cf. Sec. 3.5.2. In total, when investigating the dependency on *SNR* or L, we generate 64 individual simulations per condition. When investigating the convergence behavior, we generate 128 individual simulations.

3.5.3.2 Case B: With Interfering Speech

In case B, the microphone signals are composed of two reverberant speech components and a noise component, $\mathbf{x}_1(l)$, $\mathbf{x}_2(l)$, and $\mathbf{v}(l)$, with $\mathbf{x}_1(l)$ again containing the target component $\mathbf{x}_{e|T}(l) = \mathbf{x}_{1|e}(l)$, and $\mathbf{x}_2(l)$ an interfering speech component. We investigate the dependency on *SNR* and L, and generate $\mathbf{x}_1(l)$ and $\mathbf{v}(l)$ in the same manner as in case A, cf. Sec. 3.5.3.1. To generate $\mathbf{x}_2(l)$, we use the same set of RIR measurements, where the associated source is positioned in 2 m distance at either $\{30, 60, 90\}°$. If $\mathbf{x}_1(l)$ contains female speech, then $\mathbf{x}_2(l)$ contains male speech [25] and vice versa. On average, $\mathbf{x}_1(l)$ and $\mathbf{x}_2(l)$ have roughly the same power. From the speech source signal files and the babble noise file, we randomly select individual segments, generating $3 \cdot 64 = 192$ individual simulations per condition to be averaged in the performance evaluation, cf. Sec. 3.5.2.

3.5.4 Algorithmic Settings

In our simulations, the sampling frequency is $f_s = 16$ kHz, and the STFT analysis and synthesis uses square-root Hann windows of $N_{STFT} = 512$ samples with 50% overlap. When investigating the dependency on *SNR* and the convergence behavior, we set $L = 6$ in (3.23). The estimates $\hat{\boldsymbol{\varphi}}_s(l)$ and $\hat{\mathbf{H}}^{+}(l)$, required in

(3.37) and (3.18), (3.21) are obtained by means of [164, cf. Ch. 4], cf. Sec. 3.4.2. In (3.51)–(3.52), we set α such that $10\log_{10}(1-\alpha) = -25\,\text{dB}$. Expecting lower values for SC filter coefficients at higher frequencies due to generally reduced spatial correlations between individual microphones, we choose $\bar{\psi}_{w_{SC}}$ in (3.53) to be frequency-dependent with $10\log_{10}\bar{\psi}_{w_{SC}}$ decreasing linearly from 0 dB at 0 kHz to -15 dB at 8 kHz. In (3.54), we set $10\log_{10}\bar{\psi}_{w_{LP}} = -4\,\text{dB}$. In (3.43), we set β such that $20\log_{10}\beta = -2\,\text{dB}$, and $\gamma(0) = 1$.

3.5.5 Results

We discuss the results in case A and B in Sec. 3.5.5.1 and Sec. 3.5.5.2, respectively. Audio examples are available at [152].

3.5.5.1 Case A

Consider the spectrograms in Fig. 3.2 depicting 2 s of (a) the reference microphone signal $y_1(l)$, and the corresponding outputs of (b) the original alternating Kalman filters, (c) the modified alternating Kalman filters, and (d) the ISCLP Kalman filter for $L = 6$ in an exemplary simulation at $SNR = 10\,\text{dB}$. As can be seen by comparison with (a), all three algorithms in (b)–(d) considerably reduce reverberation and noise. Yet, their spectrograms exhibit slightly different features. As opposed to the modified alternating Kalman filters and the ISCLP Kalman filter (c)–(d), the original alternating Kalman filters (b) show some amount of temporal smearing resembling musical noise [116]. This is due to errors in the correlation matrix estimates used to update the alternating Kalman filters, which in turn are computed recursively based on the alternating Kalman filters' previous state estimates and error signals [116]. In contrast, in the modified alternating Kalman filters and the ISCLP Kalman filter, the required correlation matrix and PSD estimates are computed directly from the microphone signals while maintaining non-stationarities, cf. Sec. 3.4.2 and Sec. 3.5.1.1. As compared to the modified alternating Kalman filters (c), the signal power in the ISCLP Kalman filter (d) decays somewhat less quickly after transient speech components, which is due to $\beta > 0$ in (3.43), cf. Sec. 3.5.4, resulting in a perceptually somewhat more pleasant sound image [152].

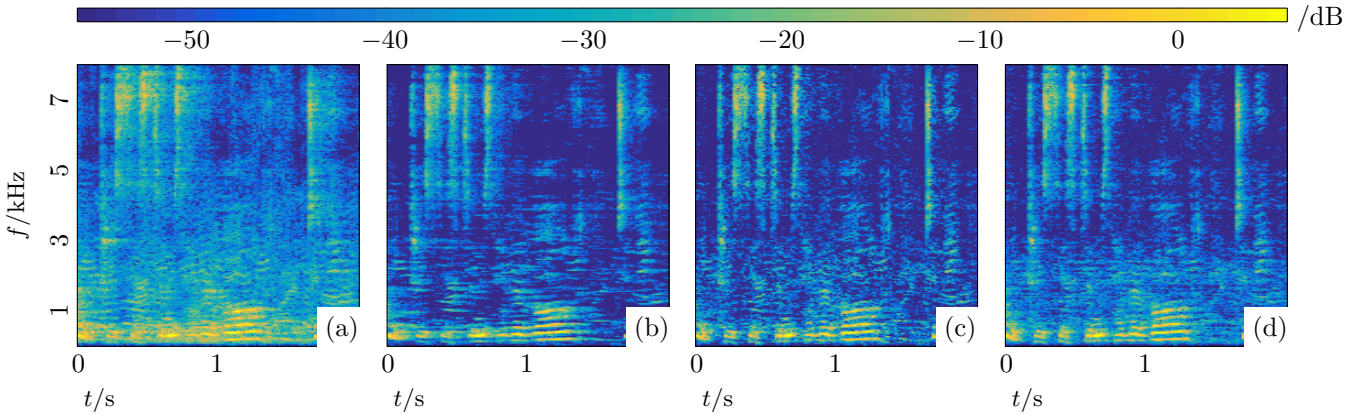


Figure 3.2: Exemplary spectrograms depicting 2 s of
(a) the reference microphone signal $y_1(l)$,
and the corresponding outputs of
(b) the original alternating Kalman filters,
(c) the modified alternating Kalman filters, and
(d) the ISCLP Kalman filter for $L = 6$
at $SNR = 10\,\mathrm{dB}$.

Fig. 3.3 shows the performance in terms of (a) *PESQ*, (b) *STOI*, (c) SIR^{fws}, and (d) *CD* versus *SNR* for the reference microphone signal [·····], the original alternating Kalman filters [- - -], the modified alternating Kalman filters [–■–], and the ISCLP Kalman filter [–•–] with $L = 6$. In this and the following figures, the graphs denote medians over all individual simulations, cf. Sec. 3.5.3, and the shaded areas indicate the range from the first to the third quartile. Overall, the measures show a high degree of agreement. As expected, the reference microphone signal reaches better scores at higher *SNR* values in all measures. Above roughly $SNR = -5\,\text{dB}$, all three algorithms show a significant improvement over the reference microphone signal in all measures, least pronounced in *STOI*. The modified alternating Kalman filters generally outperform the original alternating Kalman filters, validating the modified parameter estimation and tuning aligned to the proposed ISCLP Kalman filter, cf. Sec. 3.5.1.1. In terms of *PESQ*, *STOI*, and *CD*, the ISCLP Kalman filter reaches very similar scores as compared to the modified alternating Kalman filters. In terms of SIR^{fws}, the ISCLP Kalman filter performs somewhat worse than the modified alternating Kalman filters above $SNR = 20\,\text{dB}$, which is due to a small amount of speech cancellation caused by the SC filter, cf. Sec. 3.4.1. Note that in this *SNR* range, the babble noise component $\mathbf{v}(l)$ becomes negligible, i.e. reverberant interference is pre-dominant, which can be handled by the LP filter only. The SC filter therefore becomes superfluous in this case. Further simulations showed that the ISCLP Kalman filter may reach similar SIR^{fws} scores as compared to the modified alternating Kalman filters if the SC filter variance $\bar{\psi}_{w_{SC}}$ in (3.53) is set depdending on the *SNR*, which allows to essentially switch off the SC filter at high *SNR* values, and thereby avoid unnecessary speech cancellation.

Fig. 3.4 depicts the performance improvement in terms of (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus L with respect to the reference microphone signal for the original alternating Kalman filters [- - -], the modified alternating Kalman filters [–■–], and the ISCLP Kalman filter [–•–] at $SNR = 25\,\text{dB}$. Note that in Fig. 3.4 and in the following figures presenting performance improvements, the resolution of the vertical axes is twice as large as in Fig. 3.3. Again, the measures show a high degree of agreement. We find that in all measures, the original alternating Kalman filters generally yield less improvement and in addition show a stronger dependency on L as compared to the modified alternating Kalman filters and the ISCLP Kalman filter. The improvement for both the modified alternating Kalman filters and the ISCLP Kalman filter saturates at roughly $L = 6$. The original alternating Kalman filters reach the largest improvement between $L = 8$ and $L = 10$. In terms of (c) ΔSIR^{fws} and (d) ΔCD, however, as opposed to the other two algorithms, its performance decays again for larger values of L [116]. Further simulations showed that for all three algorithms, the dependency on L decreases with

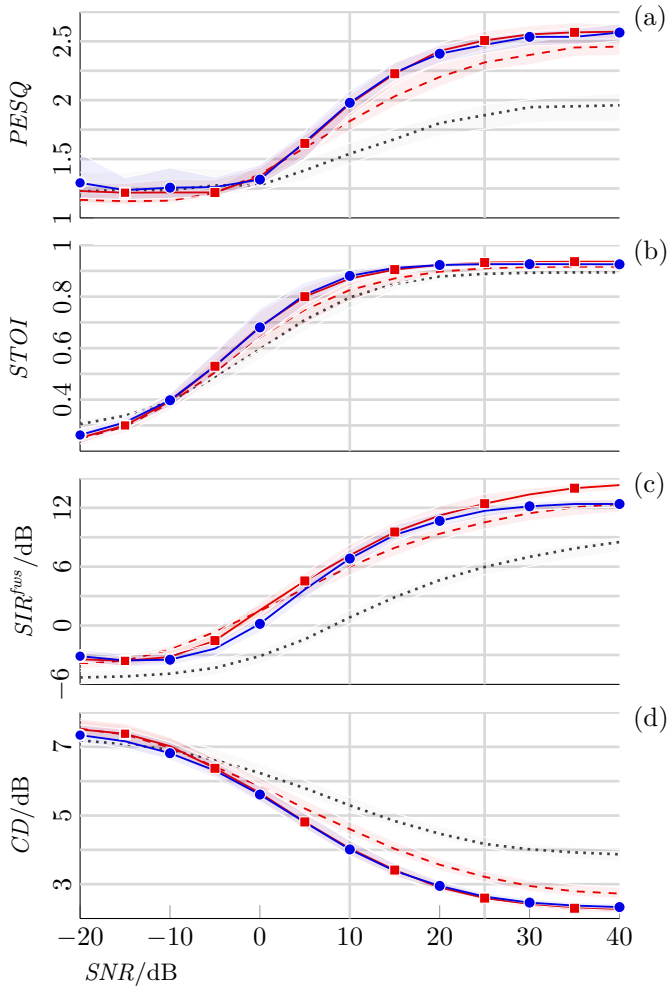


Figure 3.3: (a) *PESQ*, (b) *STOI*, (c) SIR^{fws}, and (d) *CD* versus *SNR* for the reference microphone signal [·····], the original alternating Kalman filters [- - -], the modified alternating Kalman filters [–■–], and the ISCLP Kalman filter [–•–] with $L = 6$ if interfering speech is absent.

decreasing *SNR* values. This is expected since at low *SNR* values, the babble noise component $\mathbf{v}(l)$ becomes pre-dominant, which is temporally uncorrelated, cf. Sec. 3.2, and may therefore not be suppressed by the LP filter.

Fig. 3.5 shows the performance improvement in terms of (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus time t with respect to the reference

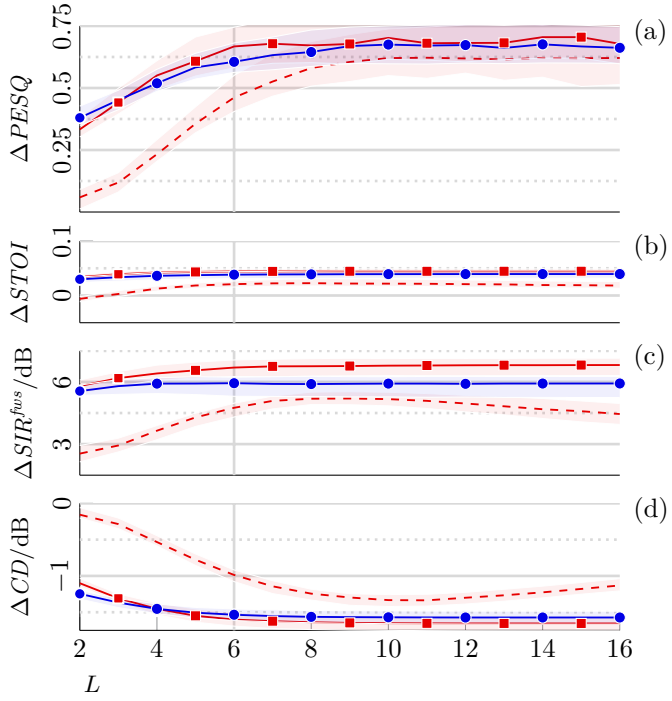


Figure 3.4: (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus L with respect to the reference microphone signal for the original alternating Kalman filters [- - -], the modified alternating Kalman filters [–■–], and the ISCLP Kalman filter [–●–] at $SNR = 25\,\text{dB}$ if interfering speech is absent.

microphone signal for the original alternating Kalman filters [- - -], the modified alternating Kalman filters [–■–], and the ISCLP Kalman filter [–●–] with $L = 6$ at $SNR = 10\,\text{dB}$. Again, the measures largely agree. We find that after initialization, all algorithms converge after roughly 4 s. The speech source position changes at 8 s, cf. Sec. 3.5.3.1, such that the three algorithms have to re-adapt. In case of the ISCLP Kalman filter, this does not only require adaptation of $\hat{\mathbf{w}}(l)$, but also of the estimate $\hat{\mathbf{H}}_T(l)$, cf. (3.18), (3.21), and Sec. 3.4.2. Note that none of the three algorithms is re-initialized after $t = 8\,\text{s}$, but re-adapt themselves, cf. also Sec. 3.4.2 for the ISCLP Kalman filter. However, we find that for all three algorithms, convergence speed after the speech source position change is somewhat reduced as compared to the initial convergence stage.

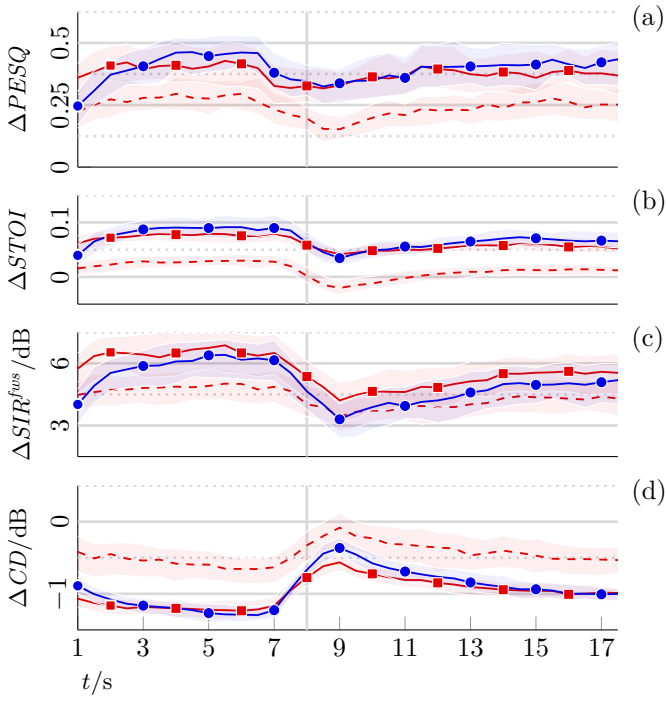


Figure 3.5: (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus t with respect to the reference microphone signal for the original alternating Kalman filters [- - -], the modified alternating Kalman filters [—■—], and the ISCLP Kalman filter [—•—] with $L = 6$ at $SNR = 10\,\text{dB}$ if interfering speech is absent.

3.5.5.2 Case B

Fig. 3.6 shows the performance in terms of (a) *PESQ*, (b) *STOI*, (c) SIR^{fws}, and (d) *CD* versus *SNR* for the reference microphone signal [·····], the MCLP+GSC Kalman filter cascade [—▾—] and the ISCLP Kalman filter [—•—] with $L = 6$. Also here, the measures show a high degree of agreement. As in case A, cf. Fig. 3.3, the reference microphone signal reaches better scores at higher *SNR* values in all measures. The curves are, however, generally flatter as compared to those in Fig. 3.3, which is due to the now additional interfering speech component $\mathbf{x}_2(l)$, cf. Sec. 3.5.3. Above roughly $SNR = -5\,\text{dB}$, both algorithms show a significant improvement over the reference microphone signal in all measures, with the ISCLP Kalman filter clearly outperforming the MCLP+GSC cascade. For the ISCLP Kalman filter, as compared to case A where $\mathbf{x}_2(l) = \mathbf{0}$, cf. Fig. 3.3, *PESQ* now predicts less improvement, while *STOI* predicts more improvement, indicating different sensitivity of both measures to the additional

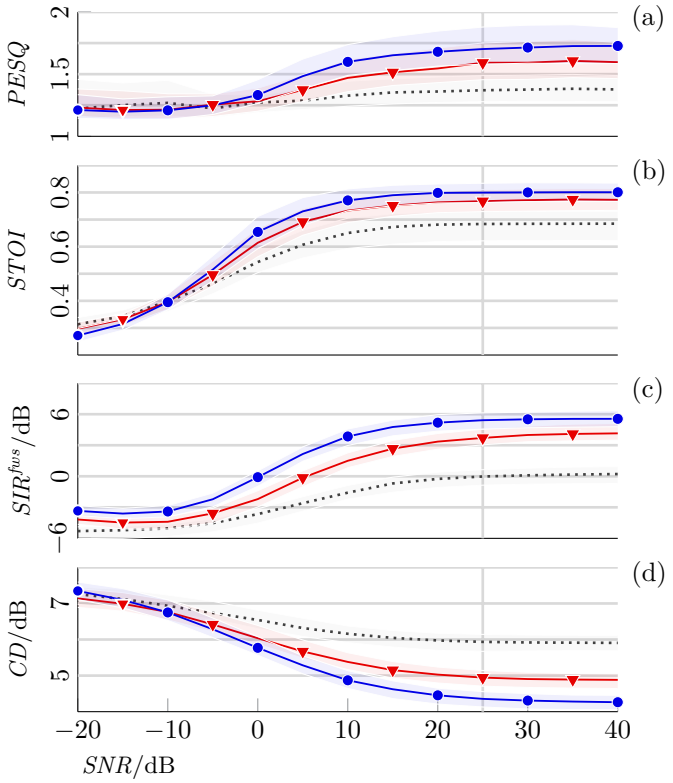


Figure 3.6: (a) *PESQ*, (b) *STOI*, (c) SIR^{fws}, and (d) *CD* versus *SNR* for the reference microphone signal [·····], the MCLP+GSC Kalman filter cascade [–▾–] and the ISCLP Kalman filter [–●–] with $L = 6$ if interfering speech is present.

interfering speech component $\mathbf{x}_2(l)$.

Fig. 3.7 depicts the performance improvement in terms of (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus L with respect to the reference microphone signal for the MCLP+GSC Kalman filter cascade [–▾–] and the ISCLP Kalman filter [–●–] at $SNR = 25\,\text{dB}$. Again, the ISCLP Kalman filter clearly outperforms the MCLP+GSC Kalman filter cascade in the simulated range. For the ISCLP Kalman filter, as compared to case A where $\mathbf{x}_2(l) = \mathbf{0}$, cf. Fig. 3.4, the improvement shows a stronger dependency on L and saturates somewhat later, indicating that longer filters are required in case of additional temporally correlated components such as $\mathbf{x}_2(l)$, which is in line with the findings in [163, cf. Ch. 2]. As in case A, further simulations showed that for both algorithms, the dependency on L decreases with decreasing *SNR* values.

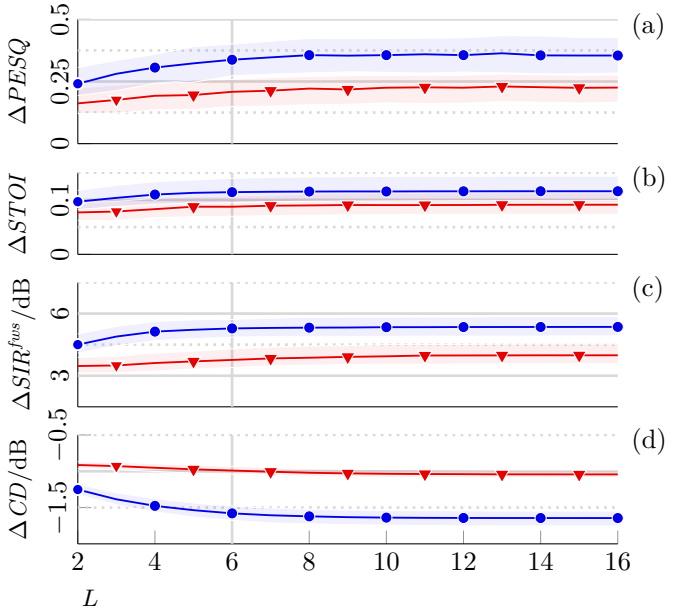


Figure 3.7: (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus L with respect to the reference microphone signal for the MCLP+GSC Kalman filter cascade [–▾–] and the ISCLP Kalman filter [–●–] at $SNR = 25\,\text{dB}$ if interfering speech is present.

3.6 Conclusion

In this paper, in order to jointly perform deconvolution and spatial filtering, allowing for dereverberation, interfering speech cancellation and noise reduction, we have presented the ISCLP Kalman filter, which integrates MCLP and the GSC. Hereat, the SC filter and the LP filter operate in parallel but on different input-data frames, and are estimated jointly. We further have proposed a spectral Wiener gain post-processor, relating to the Kalman filter's posterior state estimate. Implementational aspects such as spatio-temporal target component leakage, target PSD estimation and RETF updates, as well as process equation parameter tuning and initialization have been discussed. The presented ISCLP Kalman filter has been benchmarked in terms of its dependency on the SNR and the filter length L, as well as in terms of its convergence behavior. With M the number of microphones, the ISCLP Kalman filter is roughly M^2 times less expensive than both reference algorithms, namely first a pair of alternating Kalman filters in an original and a modified version, and second an MCLP+GSC Kalman filter cascade. Nonetheless, simulation

results indicate better or similar performance as compared to the original or modified version of the former, and better performance as compared to the latter.

Chapter 4

Multi-Source Early PSD Estimation and RETF Update

Square Root-based Multi-Source Early PSD Estimation and Recursive RETF Update in Reverberant Environments by Means of the Orthogonal Procrustes Problem

Thomas Dietzen, Simon Doclo, Marc Moonen, and Toon van Waterschoot

ESAT-STADIUS Tech. Rep. TR 19-69, KU Leuven, Belgium, submitted for publication, June 2019.

The candidate's contributions as first author include: literature study, co-derivation of the presented theory, co-development of the presented algorithms, co-design of the evaluation experiments, software implementation and computer simulations, co-formulation of the conclusions, text redaction and editing.

Abstract

Multi-channel short-time Fourier transform (STFT) domain-based processing of reverberant microphone signals commonly relies on power-spectral-density (PSD) estimates of early source images, where early refers to reflections contained within the same STFT frame. State-of-the-art approaches to multi-source early PSD estimation, given an estimate of the associated relative early transfer functions (RETFs), conventionally minimize the approximation error defined with respect to the early correlation matrix, requiring non-negative inequality constraints on the PSDs. Instead, we here propose to factorize the early correlation matrix and minimize the approximation error defined with respect to the early-correlation-matrix square root. The proposed minimization problem – constituting a generalization of the so-called orthogonal Procrustes problem – seeks a unitary matrix and the square roots of the early PSDs up to an arbitrary complex argument, making non-negative inequality constraints redundant. A solution is obtained iteratively, requiring one singular value decomposition (SVD) per iteration. The estimated unitary matrix and early PSD square roots further allow to recursively update the RETF estimate, which is not inherently possible in the conventional approach. An estimate of the said early-correlation-matrix square root itself is obtained by means of the generalized eigenvalue decomposition (GEVD), where we further propose to restore non-stationarities by desmoothing the generalized eigenvalues in order to compensate for inevitable recursive averaging. Simulation results indicate fast convergence of the proposed multi-source early PSD estimation approach in only one iteration if initialized appropriately, and better performance as compared to the conventional approach.

Index terms — Early PSD estimation, RETF estimation, orthogonal Procrustes problem, unitary constraint, singular value decomposition, generalized eigenvalue decomposition.

4.1 Introduction

In many multi-microphone signal processing applications, the recorded microphone signals constitute a mixture of several spatially diverse components, originating from different sources, bearing reverberation and noise. As far as speech is concerned, one typically admits early reflections, while late reverberant components deteriorate the perceived quality and intelligibility [15]. In order to process the various mixture components, many techniques heavily rely on estimates of their power spectral densities (PSDs) [30, 134, 170].

In recent years, a number of multi-microphone approaches to the estimation of early speech PSDs, late reverberant PSDs, and/or noise PSDs have been proposed, which rely on a spatial correlation matrix model in the short-time Fourier transform (STFT) domain [61–64, 66–69, 84, 151, 171]. In order to estimate these PSDs, some parameters of the correlation matrix model are assumed to be known or estimated beforehand, such as the direction(s) of arrival (DoA(s)) or the relative early transfer function(s) (RETF(s)) associated to the source(s) [61–64, 66, 67, 84, 151, 171], or the spatial coherence matrix of the noise or the late reverberant component, where in particular the latter is commonly modeled as a spatially diffuse sound field [61–64, 66–69, 84, 132]. It should be noted that the majority of these approaches consider a single source [62–64, 66–68, 171], while only some consider multiple sources [61, 69, 84, 151], which is the focus of this paper.

In [62, 63], the early speech and late reverberant PSD estimates are obtained by maximum-likelihood estimation, where in [62], both are estimated jointly, and in [63], the late reverberant PSD estimation relies on blocking the early speech component. Given particular coherence matrix estimates, e.g., defined from DoA or RETF estimates (for point sources) or assumptions on the spatial nature of the sound field (for noise and late reverberation), other estimators rely on Frobenius-norm minimization of the approximation error defined with respect to an estimate of the associated correlation matrix component [61, 64, 66, 69, 84, 151, 171]. Specifically, in [171], the speech PSD is estimated by minimizing the approximation error defined with respect to an estimate of the speech-only correlation matrix component (while reverberation is not considered). In a similar manner, in [84], considering multiple sources, the early PSDs are estimated from an estimate of the early correlation matrix component. In [61], the late reverberant PSD is estimated from an estimate of a blocking-based correlation matrix, generated by blocking the direct components, while the multiple early PSDs are estimated as in [84]. Likewise, one may also jointly estimate several PSDs associated to different kinds of coherence matrices, e.g., one may jointly estimate early speech PSD(s), the late reverberant PSD, and noise PSD(s) [64, 66, 151]. In [69], joint estimation of the RETFs, the early speech PSDs, the late reverberant PSD, and the noise PSDs is proposed using simultaneous confirmatory factor analysis in multiple frames, i.e. by jointly minimizing a number of approximation errors defined over several frames, during which the RETFs are assumed to be stationary. Note that the PSD estimates based on this type of minimization problem are not inherently guaranteed to be non-negative, requiring either non-negative thresholding, or, alternatively, non-negative inequality constraints. In [68], the estimation of the late reverberant PSD is based on a subspace decomposition, outperforming the late reverberant PSD estimators in [61–63], while the early speech PSD estimate is obtained from the decision-directed approach [172].

In this contribution, we are mainly concerned with early PSD estimation and recursive RETF updates for multiple sources in reverberant environments, given initial estimates of the associated RETFs. Instead of minimizing the approximation error defined with respect to an estimate of the early correlation matrix as in the manner of [64, 66, 84, 151], however, we propose to factorize the early correlation matrix and minimize the approximation error defined with respect to the early-correlation-matrix square root. Instead of directly estimating the early PSDs, the proposed minimization problem seeks a unitary matrix and the square roots of the early PSDs up to an arbitrary complex argument, making non-negative thresholding or non-negative inequality constraints redundant. The proposed minimization problem constitutes a generalization [173] of the so-called orthogonal Procrustes problem [174,175] and may be solved iteratively, requiring one singular value decomposition (SVD) per iteration. The estimated unitary matrix and early PSD square roots further allow us to recursively update the RETF estimate, which is not inherently possible in the conventional approach. An estimate of the said early-correlation-matrix square root itself is obtained from an estimate of the microphone signal correlation matrix and the diffuse coherence matrix by means of the generalized eigenvalue decomposition (GEVD). Hereat, in order to compensate for the inevitable recursive averaging in the microphone-signal-correlation-matrix estimation, we further propose to restore non-stationarities by desmoothing the generalized eigenvalues. Simulation results indicate fast convergence of the proposed multi-source early PSD estimation approach in only one iteration if initialized appropriately, and better performance as compared to the conventional approach in terms of the relative squared PSD estimation error and the signal-to-interference ratio [176] measuring the source-component separation. A MATLAB implementation is available at [153].

The remainder of this paper is organized as follows. In Sec. 4.2, we introduce the signal model. Given an estimate of the early correlation matrix component, some state-of-the-art approaches to early PSD estimation are reviewed in Sec. 4.3, while the proposed approach, given an estimate of the early-correlation-matrix square root, is presented in Sec. 4.4. In Sec. 4.5, we discuss the estimation of the required early correlation matrix component and its factorization. The proposed approach is evaluated in Sec. 4.6, followed by a conclusion in Sec. 4.7.

4.2 Signal Model

Throughout the paper, we use the following notation: vectors are denoted by lower-case boldface letters, matrices by upper-case boldface letters, $\mathbf{I}$ and $\mathbf{0}$ denote identity and zero matrices, $\mathbf{i}$ and $\mathbf{1}$ denote the first column of $\mathbf{I}$

and a vector of ones, respectively, $\mathbf{A}^T$, $\mathbf{A}^H$, and $\mathrm{E}[\mathbf{A}]$ denote the transpose, the complex conjugate transpose or Hermitian, and the expected value of a matrix $\mathbf{A}$. The operation $\mathrm{diag}[\mathbf{A}]$ creates a column vector from the diagonal elements of a square matrix $\mathbf{A}$, $\mathrm{Diag}[\mathbf{a}]$ and $\mathrm{Diag}[\mathbf{a}^T]$ create a diagonal matrix with the elements of $\mathbf{a}$ on its diagonal, $\mathrm{Diagg}[\mathbf{A}] = \mathrm{Diag}\big[\mathrm{diag}[\mathbf{A}]\big]$ zeros the off-diagonal elements of $\mathbf{A}$, and $\mathrm{tr}[\mathbf{A}]$ denotes the trace of $\mathbf{A}$. For non-negative $\mathbf{a} \in \mathbb{R}^N$, $\mathbf{a}^{1/2} \in \mathbb{C}^N$ denotes a complex vector with arbitrary complex argument that satisfies $\mathrm{Diag}[\mathbf{a}^{H/2}]\mathbf{a}^{1/2} = \mathbf{a}$, and hence $\big|\mathbf{a}^{1/2}\big| = \sqrt{\mathbf{a}}$, with absolute value and non-negative square-root applied element-wise. The operation $\max[\mathbf{a}_1, \mathbf{a}_2]$ returns a vector of the element-wise maxima of $\mathbf{a}_1$ and $\mathbf{a}_2$. $\|\mathbf{A}\|_F$ denotes the Frobenius norm of $\mathbf{A}$, whereas $\|\mathbf{a}\|_2$ denotes the Euclidian norm of $\mathbf{a}$. Row i and column j of $\mathbf{A}$ are denoted as $[\mathbf{A}]_{i,:}$ and $[\mathbf{A}]_{:,j}$, respectively, the element at their intersection as $[\mathbf{A}]_{i,j}$, and submatrices spanning rows i_1 to i_2 or columns j_1 to j_2 as $[\mathbf{A}]_{i_1:i_2,:}$ and $[\mathbf{A}]_{:,j_1:j_2}$, respectively. $\Re[a]$ and $\Im[a]$ denote the real and imaginary part of $a \in \mathbb{C}$.

In the STFT domain, with l and k indexing the frame and the frequency bin, respectively, let $x_m(l,k)$ with $m = 1, \ldots, M$ denote the m^{th} microphone signal, with M the number of microphones. In the following, we treat all frequency bins independently and hence omit the frequency index. We define the stacked microphone signal vector $\mathbf{x}(l) \in \mathbb{C}^M$,

$$\mathbf{x}(l) = \Big(x_1(l) \;\cdots\; x_M(l)\Big)^T \tag{4.1}$$

composed of the reverberant signal components $\mathbf{x}_n(l)$ with $n = 1, \ldots, N$ originating from N point sources, defined equivalently to (4.1), i.e.

$$\mathbf{x}(l) = \sum_{n=1}^{N} \mathbf{x}_n(l). \tag{4.2}$$

Each reverberant signal component $\mathbf{x}_n(l)$ may be decomposed into the early and late reverberant component $\mathbf{x}_{n|e}(l)$ and $\mathbf{x}_{n|\ell}(l)$, i.e.

$$\mathbf{x}_n(l) = \mathbf{x}_{n|e}(l) + \mathbf{x}_{n|\ell}(l), \tag{4.3}$$

which are commonly parted by the arrival time of the therein contained reflections and assumed to have distinct spatial properties as outlined below. Early reflections are assumed to arrive within the same frame, with the early components in $\mathbf{x}_{n|e}(l)$ related by the RETF in $\mathbf{h}_n(l) \in \mathbb{C}^M$, i.e.

$$\mathbf{x}_{n|e}(l) = \mathbf{h}_n(l) s_n(l). \tag{4.4}$$

Here, without loss of generality, the RETF $\mathbf{h}_n(l)$ is assumed to be relative to the first microphone, i.e. $\mathbf{i}^T\mathbf{h}_n(l) = [\mathbf{h}_n(l)]_1 = 1$, and $s_n(l) = [\mathbf{x}_{n|e}(l)]_1$ denotes

the early component in the first microphone originating from the n^{th} source, in the following referred to as early source image. We define the stacked RETF matrix $\mathbf{H}(l) \in \mathbb{C}^{M\times N}$, yielding

$$\mathbf{H}(l) = \Big(\mathbf{h}_1(l) \ \cdots \ \mathbf{h}_N(l)\Big), \tag{4.5}$$

$$\mathbf{i}^T\mathbf{H}(l) = [\mathbf{H}(l)]_{1,:} = \mathbf{1}^T. \tag{4.6}$$

Similarly, we stack $s_n(l)$ into $\mathbf{s}(l) \in \mathbb{C}^N$, i.e.

$$\mathbf{s}(l) = \Big(s_1(l) \ \cdots \ s_N(l)\Big)^T, \tag{4.7}$$

such that the sum of the early components $\mathbf{x}_{n|\mathrm{e}}(l)$ may be expressed more compactly as

$$\sum_{n=1}^{N} \mathbf{x}_{n|\mathrm{e}}(l) = \mathbf{H}(l)\mathbf{s}(l). \tag{4.8}$$

Further, we assume that $\mathbf{x}_{n|\mathrm{e}}(l)$ and $\mathbf{x}_{n|\ell}(l)$ are mutually uncorrelated within frame l. Let $\mathbf{\Psi}_x(l) = \mathrm{E}[\mathbf{x}(l)\mathbf{x}^H(l)] \in \mathbb{C}^{M\times M}$ denote the microphone signal correlation matrix, and let the early and late reverberant correlation matrix $\mathbf{\Psi}_{x_e}(l)$ and $\mathbf{\Psi}_{x_\ell}(l)$ be similarly defined. With (4.3)–(4.8), we then find

$$\mathbf{\Psi}_x(l) = \mathbf{\Psi}_{x_e}(l) + \mathbf{\Psi}_{x_\ell}(l), \tag{4.9}$$

wherein $\mathbf{\Psi}_{x_e}(l)$ generally has rank N and is expressed by

$$\boxed{\mathbf{\Psi}_{x_e}(l) = \mathbf{H}(l)\mathbf{\Phi}_s(l)\mathbf{H}^H(l),} \tag{4.10}$$

$$\mathbf{\Phi}_s(l) = \mathrm{Diag}[\boldsymbol{\varphi}_s(l)], \tag{4.11}$$

$$\boldsymbol{\varphi}_s(l) = \Big(\varphi_{s_1}(l) \ \cdots \ \varphi_{s_N}(l)\Big)^T, \tag{4.12}$$

with $\varphi_{s_n}(l)$ denoting the PSD of the early source image $s_n(l)$. Note that applying (4.6) to (4.10)–(4.11) while using $\mathbf{1}^T\mathbf{\Phi}_s(l)\mathbf{1} = \mathbf{1}^T\boldsymbol{\varphi}_s(l)$, we find that

$$\boxed{\mathbf{i}^T\mathbf{\Psi}_{x_e}(l)\mathbf{i} = [\mathbf{\Psi}_{x_e}(l)]_{1,1} = \mathbf{1}^T\boldsymbol{\varphi}_s(l),} \tag{4.13}$$

i.e. the sum of the early PSDs $\varphi_{s_n}(l)$ equals $[\mathbf{\Psi}_{x_e}(l)]_{1,1}$. Assuming that $\mathbf{x}_{n|\ell}(l)$ may be modeled as diffuse [61–64, 66–69, 84, 132] with coherence matrix $\mathbf{\Gamma} \in \mathbb{C}^{M\times M}$, which may be computed from the microphone array geometry [132]

and is therefore considered to be known in the remainder, we may write $\mathbf{\Psi}_{x_\ell}(l)$ as

$$\mathbf{\Psi}_{x_\ell}(l) = \varphi_{x_\ell}(l)\mathbf{\Gamma}, \tag{4.14}$$

$$\text{with} \quad \varphi_{x_\ell}(l) = \sum_{n=1}^{N} \varphi_{x_{n|\ell}}(l), \tag{4.15}$$

and $\varphi_{x_{n|\ell}}(l)$ denoting the PSD of the late reverberant component $\mathbf{x}_{n|\ell}(l)$. The PSDs $\boldsymbol{\varphi}_s(l)$ and $\varphi_{x_\ell}(l)$ may be highly non-stationary, especially if the point sources are speech sources, while the associated coherence matrices $\mathbf{h}_n(l)\mathbf{h}_n^H(l)$ and $\mathbf{\Gamma}$ are commonly assumed to be comparably slowly time-varying or even time-invariant.

Note that with (4.14)–(4.15), one may easily include further diffuse components, e.g., babble noise, without formally changing the signal model. However, since in this paper, we are mainly concerned with the estimation of the early PSDs $\boldsymbol{\varphi}_s(l)$ and the recursive updating of the estimate of the RETFs $\mathbf{H}(l)$, we restrict the discussion and simulations, cf. Sec. 4.6, to the example of late reverberation for the sake of conciseness.

Further, note that while the above signal model is commonly and effectively used [61–64, 66–69, 84] due to its simplicity, it may be said to be deficient in a number of aspects. The assumption that $\mathbf{x}_{n|e}(l)$ and $\mathbf{x}_{n|\ell}(l)$ are mutually uncorrelated within frame l may be violated due to overlapping windows in the STFT-processing or source signals remaining correlated over several frames. The assumption that $\mathbf{\Psi}_{x_e}(l)$ in (4.10) has rank N implicitly relies on the assumption that the frequency bins may be treated independently, ignoring cross-bin dependencies [117]. Finally, related to that, there may be components that may be modeled neither by the rank-N component $\mathbf{\Psi}_{x_e}(l)$ in (4.10) nor by the diffuse component $\mathbf{\Psi}_{x_\ell}(l)$ in (4.14), depending on the geometry and physical properties of the acoustic environment.

In the remainder, as we mostly consider the single frame l only, we also drop the frame index for conciseness and refer back to it only where necessary, namely when we differentiate the frames l and $l-1$ in recursive equations.

4.3 Early PSD Estimation based on the Early Correlation Matrix

In this section, we discuss some state-of-the-art approaches [64, 66, 69, 84, 151] to the estimation of the early PSDs $\boldsymbol{\varphi}_s$ based on the signal model in (4.10)–(4.13).

In the following, we refer to (4.10)–(4.13) as the conventional signal model. We develop our discussion from the premise that estimates $\hat{\mathbf{\Psi}}_{x_e}$ and $\hat{\mathbf{H}}$ of the early correlation matrix $\mathbf{\Psi}_{x_e}$ and the RETFs $\mathbf{H}$ in (4.10) are readily available. Throughout the paper, despite being irrelevant to the approaches discussed in this section, we consider $\hat{\mathbf{\Psi}}_{x_e}$ to generally have rank N, similar to $\mathbf{\Psi}_{x_e}$. A rank-N estimator of $\mathbf{\Psi}_{x_e}$ is described in Sec. 4.5. Further, we assume that $\hat{\mathbf{H}}$ satisfies $\mathbf{i}^T\hat{\mathbf{H}} = \mathbf{1}^T$, cf. $\mathbf{H}$ in (4.6).

Given the estimates of a early correlation matrix $\hat{\mathbf{\Psi}}_{x_e}$ and the therein superimposed coherence matrices $\hat{\mathbf{h}}_n\hat{\mathbf{h}}_n^H$, one may estimate the associated PSDs φ_{s_n}, cf. (4.5), (4.10)–(4.11), as described in [64, 66, 84, 151].[1] Adopting this approach, we define the approximation error as a function of $\boldsymbol{\varphi}_s$ as

$$\mathbf{E}_c(\boldsymbol{\varphi}_s) = \hat{\mathbf{\Psi}}_{x_e} - \hat{\mathbf{H}}\,\mathrm{Diag}[\boldsymbol{\varphi}_s]\hat{\mathbf{H}}^H, \tag{4.16}$$

where the subscript c stands for conventional. The early PSDs $\boldsymbol{\varphi}_s$ can then be estimated by Frobenius-norm minimization of the approximation error followed by non-negative thresholding, i.e.

$$\hat{\boldsymbol{\varphi}}_s' = \arg\min_{\boldsymbol{\varphi}_s} \left\|\mathbf{E}_c(\boldsymbol{\varphi}_s)\right\|_F^2, \tag{4.17}$$

$$\hat{\boldsymbol{\varphi}}_s = \max[\hat{\boldsymbol{\varphi}}_s', \mathbf{0}]. \tag{4.18}$$

The non-negative thresholding in (4.18) is necessary as the elements of $\hat{\boldsymbol{\varphi}}_s'$ in (4.17) may in fact be negative, conflicting with the notion of $\boldsymbol{\varphi}_s$ being a vector of PSDs. If $\hat{\mathbf{H}}^H\hat{\mathbf{H}}$ has full rank, which (without sufficiency) requires $N \leq M$, the problem in (4.17) has a unique solution given by

$$\hat{\boldsymbol{\varphi}}_s' = \mathbf{A}_{c_0}^{-1}\mathbf{b}_{c_0}, \tag{4.19}$$

where $\mathbf{A}_{c_0} \in \mathbb{R}^{N\times N}$ and $\mathbf{b}_{c_0} \in \mathbb{R}^N$ are defined by

$$[\mathbf{A}_{c_0}]_{n,n'} = |\hat{\mathbf{h}}_n^H\hat{\mathbf{h}}_{n'}|^2, \tag{4.20}$$

$$\mathbf{b}_{c_0} = \mathrm{diag}[\hat{\mathbf{H}}^H\hat{\mathbf{\Psi}}_{x_e}\hat{\mathbf{H}}]. \tag{4.21}$$

Alternatively, instead of simple thresholding after solving (4.17), one may solve the minimization problem subject to the non-negative inequality constraint $\boldsymbol{\varphi}_s \geq \mathbf{0}$, as proposed in [69]. In addition to this, one may further impose a soft constraint on $\mathbf{1}^T\boldsymbol{\varphi}_s$ corresponding to (4.13), i.e. one may define a soft-constraint error as a function of $\boldsymbol{\varphi}_s$ to be penalized as

$$e_c(\boldsymbol{\varphi}_s) = [\hat{\mathbf{\Psi}}_{x_e}]_{1,1} - \mathbf{1}^T\boldsymbol{\varphi}_s$$

[1] In [84], as in our case, point-source coherence matrices of rank one are considered, while in [64, 66, 151], without rendering a difference in the principle approach, general coherence matrices are considered.

$$= [\mathbf{E}_c(\boldsymbol{\varphi}_s)]_{1,1}. \tag{4.22}$$

The resulting minimization problem can then be written as

$$\boxed{\begin{aligned} \hat{\boldsymbol{\varphi}}_s = \arg\min_{\boldsymbol{\varphi}_s} \left\|\mathbf{E}_c(\boldsymbol{\varphi}_s)\right\|_F^2 + \alpha\left|e_c(\boldsymbol{\varphi}_s)\right|^2 \\ \text{s.t.} \quad \boldsymbol{\varphi}_s \geq \mathbf{0}, \end{aligned}} \tag{4.23}$$

where α is the penalty factor. For $\alpha \to \infty$, a hard constraint $\mathbf{1}^T\boldsymbol{\varphi}_s = [\hat{\boldsymbol{\Psi}}_{x_e}]_{1,1}$ is introduced, which may however not be desirable due to potential estimation errors in $[\hat{\boldsymbol{\Psi}}_{x_e}]_{1,1}$. Note that in [69], instead of a soft constraint, a box constraint on $\mathbf{1}^T\boldsymbol{\varphi}_s$ has been used. For the sake of comparison to the algorithm proposed in Sec. 4.4, however, we restrict our discussion to the soft constraint. Due to the inequality constraint $\boldsymbol{\varphi}_s \geq \mathbf{0}$, the problem in (4.23) does not have a closed-form solution and may require several iterations in order to be solved. Using the proximal gradient method [177,178], we may solve (4.23) by iterating the below set of equations until convergence is reached,

$$\hat{\boldsymbol{\varphi}}_s'^{(i)} = \hat{\boldsymbol{\varphi}}_s^{(i-1)} + \mu\big(\mathbf{b}_c - \mathbf{A}_c\hat{\boldsymbol{\varphi}}_s^{(i-1)}\big), \tag{4.24}$$

$$\hat{\boldsymbol{\varphi}}_s^{(i)} = \max[\hat{\boldsymbol{\varphi}}_s'^{(i)}, \mathbf{0}], \tag{4.25}$$

where i is the iteration index, μ the step-size, and $\mathbf{b}_c - \mathbf{A}_c\hat{\boldsymbol{\varphi}}_s^{(i-1)}$ the gradient with $\mathbf{A}_c \in \mathbb{R}^{N\times N}$ and $\mathbf{b}_c \in \mathbb{R}^N$ defined by

$$\mathbf{A}_c = \mathbf{A}_{c_0} + \alpha\mathbf{1}\mathbf{1}^T, \tag{4.26}$$

$$\mathbf{b}_c = \mathbf{b}_{c_0} + \alpha[\hat{\boldsymbol{\Psi}}_{x_e}]_{1,1}\mathbf{1}, \tag{4.27}$$

and $\mathbf{A}_{c_0}$ and $\mathbf{b}_{c_0}$ defined in (4.20)–(4.21). As initial value, it is straight-forward to choose $\hat{\boldsymbol{\varphi}}_s^{(0)} = \mathbf{A}_c^{-1}\mathbf{b}_c$, which yields the global minimum if $\hat{\boldsymbol{\varphi}}_s^{(0)} \geq \mathbf{0}$. In this case therefore, convergence is reached after one iteration of (4.24)–(4.25). In any case, for $\hat{\boldsymbol{\varphi}}_s^{(0)} = \mathbf{A}_c^{-1}\mathbf{b}_c$ and $\alpha = 0$, the estimate obtained after one iteration of (4.24)–(4.25) corresponds to the estimate defined by (4.17)–(4.19). We therefore use (4.23) as a reference for comparison in the remainder. In the following, we refer to (4.23) as the conventional minimization problem (conventional MP).

4.4 Early PSD Estimation and Recursive RETF Update based on the Early-Correlation-Matrix Square Root

In this section, in order to estimate the early PSDs $\boldsymbol{\varphi}_s$, instead of defining the approximation error to be minimized with respect to $\hat{\boldsymbol{\Psi}}_{x_e}$ as in (4.16), we propose to define the approximation error with respect to the square root $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2} \in \mathbb{C}^{M\times N}$ of $\hat{\boldsymbol{\Psi}}_{x_e}$, satisfying $\hat{\boldsymbol{\Psi}}_{x_e} = \hat{\boldsymbol{\Psi}}_{x_e}^{1/2}\hat{\boldsymbol{\Psi}}_{x_e}^{H/2}$. As to be shown, instead of directly estimating the diagonal of $\boldsymbol{\Phi}_s = \mathrm{Diag}[\boldsymbol{\varphi}_s]$, the resulting minimization problem now consists in estimating a unitary matrix $\boldsymbol{\Omega} \in \mathbb{C}^{N\times N}$ and the diagonal of $\boldsymbol{\Phi}_s^{1/2} = \mathrm{Diag}[\boldsymbol{\varphi}_s^{1/2}]$, which constitutes a generalization [173] of the so-called orthogonal Procrustes problem [174, 175]. Since the early PSDs herein are represented by $\boldsymbol{\varphi}_s = \mathrm{Diag}[\boldsymbol{\varphi}_s^{H/2}]\boldsymbol{\varphi}_s^{1/2}$, the corresponding estimate $\hat{\boldsymbol{\varphi}}_s$ is guaranteed to be non-negative, such that a non-negative inequality constraint as in (4.23) is not required. Further, we show that the obtained estimates $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\varphi}}_s^{1/2}$ may be used to recursively update the RETF estimate $\hat{\mathbf{H}}$, which is not inherently possible from the estimate $\hat{\boldsymbol{\varphi}}_s$ given by (4.23).

In Sec. 4.4.1, as a pre-requisite to our derivation, we discuss the factorization of the conventional signal model in (4.10)–(4.13), yielding the square-root signal model. In Sec. 4.4.2, based upon the square-root signal model and given the estimates $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$ and $\hat{\mathbf{H}}$, we then define and solve the square-root minimization problem (square-root MP). In Sec. 4.4.3, we discuss the recursive updating of the RETF estimate $\hat{\mathbf{H}}$.

4.4.1 Early-Correlation-Matrix Factorization

We consider the factorization of the rank-N matrices on both sides of (4.10). On the left-hand side of (4.10), we define the square root $\boldsymbol{\Psi}_{x_e}^{1/2} \in \mathbb{C}^{M\times N}$ such that $\boldsymbol{\Psi}_{x_e}^{1/2}\boldsymbol{\Psi}_{x_e}^{H/2} = \boldsymbol{\Psi}_{x_e}$. Note that the product is invariant to right-multiplication of a particular square root with any unitary matrix, and so $\boldsymbol{\Psi}_{x_e}^{1/2}$ is not unique. On the right-hand side of (4.10), with $\boldsymbol{\Phi}_s^{1/2} = \mathrm{Diag}[\boldsymbol{\varphi}_s^{1/2}]$ and $\mathrm{Diag}[\boldsymbol{\varphi}_s^{H/2}]\boldsymbol{\varphi}_s^{1/2} = \boldsymbol{\varphi}_s$, we define the square root $\mathbf{H}\boldsymbol{\Phi}_s^{1/2}$ such that $\mathbf{H}\boldsymbol{\Phi}_s^{1/2}\boldsymbol{\Phi}_s^{H/2}\mathbf{H}^H = \mathbf{H}\boldsymbol{\Phi}_s\mathbf{H}^H$. Note that while the magnitude of the elements in $\boldsymbol{\varphi}_s^{1/2} \in \mathbb{C}^N$ is well-defined, namely by $\left|\boldsymbol{\varphi}_s^{1/2}\right| = \sqrt{\boldsymbol{\varphi}_s}$, their complex argument may be chosen arbitrarily, and so $\boldsymbol{\Phi}_s^{1/2}$ is not unique. The non-uniqueness of both square roots implies that while their respective products on both sides of (4.10) coincide, the said square roots themselves generally do not, i.e. we have $\boldsymbol{\Psi}_{x_e}^{1/2} \neq \mathbf{H}\boldsymbol{\Phi}_s^{1/2}$. Hence, for a particular

$\boldsymbol{\Psi}_{x_e}^{1/2}$ and $\boldsymbol{\Phi}_s^{H/2}$, we introduce the unitary matrix $\boldsymbol{\Omega} \in \mathbb{C}^{N\times N}$, which is such that $\boldsymbol{\Psi}_{x_e}^{1/2}\boldsymbol{\Omega}$ and $\mathbf{H}\boldsymbol{\Phi}_s^{1/2}$ do coincide, i.e. we may summarize

$$\boxed{\boldsymbol{\Psi}_{x_e}^{1/2}\boldsymbol{\Omega} = \mathbf{H}\boldsymbol{\Phi}_s^{1/2},} \tag{4.28}$$

$$\boldsymbol{\Phi}_s^{1/2} = \mathrm{Diag}[\boldsymbol{\varphi}_s^{1/2}], \tag{4.29}$$

$$\boldsymbol{\Omega}\boldsymbol{\Omega}^H = \mathbf{I}, \tag{4.30}$$

where right-multiplying each side of (4.28) with its Hermitian yields (4.10). At this point, in order to stress the meaning of (4.28)–(4.30), we add that the column vectors $[\boldsymbol{\Psi}_{x_e}^{1/2}]_{:,n}$ and $[\mathbf{H}\boldsymbol{\Phi}_s^{1/2}]_{:,n} = \mathbf{h}_n\varphi_s^{1/2}$ form generally different bases[2] of the same vector space, and hence $\boldsymbol{\Omega}$ implements a change of basis.

Applying (4.6) to (4.28)–(4.29) and noting that $\mathbf{1}^T\,\mathrm{Diag}[\boldsymbol{\varphi}_s^{1/2}] = \boldsymbol{\varphi}_s^{T/2}$, we find that $\boldsymbol{\varphi}_s^{1/2}$ and $\boldsymbol{\Omega}$ satisfy

$$\boxed{\mathbf{i}^T\boldsymbol{\Psi}_{x_e}^{1/2}\boldsymbol{\Omega} = [\boldsymbol{\Psi}_{x_e}^{1/2}\boldsymbol{\Omega}]_{1,:} = \boldsymbol{\varphi}_s^{T/2},} \tag{4.31}$$

where right-multiplying each side of (4.31) with its Hermitian yields (4.13). We further note that if $\boldsymbol{\Omega}$ was known for a given square root $\boldsymbol{\Psi}_{x_e}^{1/2}$, then $\boldsymbol{\varphi}_s^{1/2}$ could be obtained from (4.31) immediately. In the following, we refer to (4.28)–(4.31) as the square-root signal model.

4.4.2 Orthogonal Procrustes-based Early PSD Estimate

In this section, based on the square-root signal model in (4.28)–(4.31), we seek unitary and diagonal estimates $\hat{\boldsymbol{\Omega}}$ and $\mathrm{Diag}[\hat{\boldsymbol{\varphi}}_s^{1/2}]$ of $\boldsymbol{\Omega}$ and $\mathrm{Diag}[\boldsymbol{\varphi}_s^{1/2}]$. Similarly to Sec. 4.3, we develop our discussion from the premise that estimates $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$ and $\hat{\mathbf{H}}$ of the early-correlation-matrix square root $\boldsymbol{\Psi}_{x_e}^{1/2}$ and the RETF $\mathbf{H}$ in (4.28) are readily available, with $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$ generally of rank N and $\mathbf{i}^T\hat{\mathbf{H}} = \mathbf{1}^T$. An estimator of $\boldsymbol{\Psi}_{x_e}^{1/2}$ is described in Sec. 4.5.2, while Sec. 4.4.3 describes a recursive update scheme for $\hat{\mathbf{H}}$.

Similarly to Sec. 4.3, now based on the square-root signal model in (4.28)–(4.30) instead of the conventional signal model in (4.10), we define the approximation

[2] A particular case is obtained for $N = 1$, where $\boldsymbol{\Omega}$ and $\boldsymbol{\Phi}_s^{1/2}$ are scalar, while $\mathbf{H} = \mathbf{h}$ and $\boldsymbol{\Psi}_{x_e}^{1/2} = \boldsymbol{\psi}_{x_e}^{1/2}$ are proportional column vectors. In this case, given an estimate $\hat{\boldsymbol{\psi}}_{x_e}^{1/2}$, we may even estimate $\mathbf{h}$ by $\hat{\mathbf{h}} = \hat{\boldsymbol{\psi}}_{x_e}^{1/2}/[\hat{\boldsymbol{\psi}}_{x_e}^{1/2}]_1$, satisfying $[\hat{\mathbf{h}}]_1 = 1$, cf. (4.6). In essence, despite somewhat different derivation and terminology, this is equivalent to the approach taken in subspace-based single-source RETF estimation [147].

error as a function of $\boldsymbol{\Omega}$ and $\boldsymbol{\varphi}_s^{1/2}$, i.e.

$$\mathbf{E}_{sq}(\boldsymbol{\Omega}, \boldsymbol{\varphi}_s^{1/2}) = \hat{\boldsymbol{\Psi}}_{x_e}^{1/2}\boldsymbol{\Omega} - \hat{\mathbf{H}}\,\mathrm{Diag}[\boldsymbol{\varphi}_s^{1/2}], \tag{4.32}$$

which is akin to $\mathbf{E}_c(\boldsymbol{\varphi}_s)$ in (4.16), and where the subscript sq stands for square root. Further, now based on the square-root signal model in (4.31) instead of the conventional signal model in (4.13), we define a soft-constraint error as a function of $\boldsymbol{\Omega}$ and $\boldsymbol{\varphi}_s^{1/2}$ to be penalized as

$$\begin{aligned}\mathbf{e}_{sq}(\boldsymbol{\Omega}, \boldsymbol{\varphi}_s^{1/2}) &= [\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}\boldsymbol{\Omega}]_{1,:}^T - \boldsymbol{\varphi}_s^{1/2} \\ &= [\mathbf{E}_{sq}(\boldsymbol{\Omega}, \boldsymbol{\varphi}_s^{1/2})]_{1,:}^T,\end{aligned} \tag{4.33}$$

which is akin to $e_c(\boldsymbol{\varphi}_s)$ in (4.22). Note that while $e_c(\boldsymbol{\varphi}_s)$ defines a error on the sum of the early PSDs, $\mathbf{e}_{sq}\big(\boldsymbol{\Omega}, \boldsymbol{\varphi}_s^{1/2}\big)$ instead defines an error on each of the early PSD square roots and is therefore more informative. Based on (4.32), (4.33), and the unitary constraint in (4.30), we define the minimization problem,

$$\boxed{\begin{aligned}\{\hat{\boldsymbol{\Omega}}, \hat{\boldsymbol{\varphi}}_s^{1/2}\} = \underset{\boldsymbol{\Omega},\, \boldsymbol{\varphi}_s^{1/2}}{\arg\min}\, &\big\|\mathbf{E}_{sq}(\boldsymbol{\Omega}, \boldsymbol{\varphi}_s^{1/2})\big\|_F^2 + \alpha\big\|\mathbf{e}_{sq}(\boldsymbol{\Omega}, \boldsymbol{\varphi}_s^{1/2})\big\|_2^2 \\ \text{s.t.} \quad &\boldsymbol{\Omega}\boldsymbol{\Omega}^H = \mathbf{I},\end{aligned}} \tag{4.34}$$

which is akin to the conventional MP in (4.23) and referred to as the square-root minimization problem (square-root MP) in the following. While the unitary constraint in (4.34) does not have an equivalent in (4.23), the inequality constraint $\boldsymbol{\varphi}_s \geq \mathbf{0}$ used in (4.23) is not required in (4.34), as in the square-root signal model, we find that $\boldsymbol{\varphi}_s = \mathrm{Diag}[\boldsymbol{\varphi}_s^{H/2}]\boldsymbol{\varphi}_s^{1/2}$, and therefore the corresponding estimate $\hat{\boldsymbol{\varphi}}_s$ is guaranteed to be non-negative. Problems of the kind as in (4.34), i.e. Frobenius-norm minimization problems seeking a unitary and a diagonal matrix, here $\boldsymbol{\Omega}$ and $\mathrm{Diag}[\boldsymbol{\varphi}_s^{1/2}]$, constitute a generalization [173] of the so-called orthogonal Procrustes problem [173–175], which seeks a unitary matrix only. As outlined in the following, under a specific rank condition, the orthogonal Procrustes problem has a unique closed-form solution, which is found by means of the SVD [174, 175]. The generalized orthogonal Procrustes problem, on the contrary, does not have a unique closed-form solution, but may be solved iteratively [173]. In particular, along the lines of [173], we propose to solve (4.34) by alternatingly (re-)estimating $\boldsymbol{\Omega}$ and $\boldsymbol{\varphi}_s^{1/2}$ until convergence is reached, namely by solving the orthogonal Procrustes problem and the soft-constrained convex problem, respectively,

$$\hat{\boldsymbol{\Omega}}^{(i)} = \underset{\boldsymbol{\Omega}}{\arg\min}\, \big\|\mathbf{E}_{sq}(\boldsymbol{\Omega}, \hat{\boldsymbol{\varphi}}_s^{1/2|(i-1)})\big\|_F^2$$

$$\text{s.t.} \quad \boldsymbol{\Omega}\boldsymbol{\Omega}^H = \mathbf{I}, \tag{4.35}$$

$$\hat{\boldsymbol{\varphi}}_s^{1/2|(i)} = \arg\min_{\boldsymbol{\varphi}_s^{1/2}} \left\|\mathbf{E}_{sq}(\hat{\boldsymbol{\Omega}}^{(i)}, \boldsymbol{\varphi}_s^{1/2})\right\|_F^2 + \alpha\left\|\mathbf{e}_{sq}(\hat{\boldsymbol{\Omega}}^{(i)}, \boldsymbol{\varphi}_s^{1/2})\right\|_2^2, \tag{4.36}$$

where the soft constraint is applied in (4.36) only, i.e. once per iteration. Using (4.32), by expansion of the Frobenius norm in (4.35) as in Appendix B.1, it is easily shown [174, 175] that (4.35) is equivalent to

$$\hat{\boldsymbol{\Omega}}^{(i)} = \arg\max_{\boldsymbol{\Omega}} \Re\big\{\mathrm{tr}[\boldsymbol{\Omega}\mathbf{C}_{sq}^{(i-1)}]\big\} \quad \text{s.t.} \quad \boldsymbol{\Omega}\boldsymbol{\Omega}^H = \mathbf{I}, \tag{4.37}$$

$$\text{with} \quad \mathbf{C}^{(i-1)} = \mathrm{Diag}[\hat{\boldsymbol{\varphi}}_s^{H/2|(i-1)}]\hat{\mathbf{H}}^H\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}. \tag{4.38}$$

If $\mathbf{C}^{(i-1)}$ has full rank, which (without sufficiency) requires $N \leq M$, the problem in (4.35) has a unique closed-form solution, which is based on the SVD of $\mathbf{C}^{H|(i-1)}$ [174, 175]. Precisely, if we decompose $\mathbf{C}^{H|(i-1)}$ as

$$\mathbf{C}^{H|(i-1)} = \mathbf{U}_L\boldsymbol{\Sigma}\mathbf{U}_R^H, \tag{4.39}$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times N}$ is a diagonal matrix of singular values and both $\mathbf{U}_L \in \mathbb{C}^{N\times N}$ and $\mathbf{U}_R \in \mathbb{C}^{N\times N}$ are unitary, then $\hat{\boldsymbol{\Omega}}^{(i)}$ is given by

$$\hat{\boldsymbol{\Omega}}^{(i)} = \mathbf{U}_L\mathbf{U}_R^H, \tag{4.40}$$

as shown in Appendix B.2. With (4.32) and (4.33), the solution to (4.36) is easily found as

$$\hat{\boldsymbol{\varphi}}_s^{1/2|(i)} = \mathbf{A}_{sq}^{-1}\mathbf{b}_{sq}^{(i)}, \tag{4.41}$$

with $\mathbf{A}_{sq} \in \mathbb{R}^{N\times N}$ and $\mathbf{b}_{sq}^{(i)} \in \mathbb{C}^N$ defined by

$$\mathbf{A}_{sq} = \mathrm{Diagg}[\hat{\mathbf{H}}^H\hat{\mathbf{H}}] + \alpha\mathbf{I}, \tag{4.42}$$

$$\begin{aligned}\mathbf{b}_{sq}^{(i)} &= \mathrm{diag}[\hat{\mathbf{H}}^H(\mathbf{I} + \alpha\mathbf{i}\mathbf{i}^T)\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}\hat{\boldsymbol{\Omega}}^{(i)}] \\ &= \mathrm{diag}[\hat{\mathbf{H}}^H\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}\hat{\boldsymbol{\Omega}}^{(i)}] + \alpha[\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}\hat{\boldsymbol{\Omega}}^{(i)}]_{1,:}^T.\end{aligned} \tag{4.43}$$

The set of equations (4.42)–(4.43) is akin to (4.26)–(4.27) for the conventional MP. Note that for $\alpha \to \infty$, the soft constraint in the square-root MP in (4.36) becomes a hard constraint and, moreover, solely determines $\hat{\boldsymbol{\varphi}}_s^{1/2|(i)}$, namely as

$\hat{\boldsymbol{\varphi}}_s^{1/2|(i)} = [\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}\hat{\boldsymbol{\Omega}}^{(i)}]_{1,:}^T$ according to (4.41)–(4.43). This is not the case for the soft constraint in the conventional MP in (4.23).

Note that since the problem in (4.34) is non-convex, the iteration in (4.35)–(4.36) is not guaranteed to converge to a global minimum [173]. The initial value $\hat{\boldsymbol{\varphi}}_s^{1/2|(0)}$ of the iteration may, e.g., be chosen based on the sum constraint in (4.13) as $\hat{\boldsymbol{\varphi}}_s^{1/2|(0)} = \sqrt{[\hat{\boldsymbol{\Psi}}_{x_e}]_{1,1}/N}\,\mathbf{1}$, or based on the comparably lowly complex estimator in (4.17)–(4.18), here denoted by $\hat{\boldsymbol{\varphi}}_{s|c_0}$, as $\hat{\boldsymbol{\varphi}}_s^{1/2|(0)} = \sqrt{\hat{\boldsymbol{\varphi}}_{s|c_0}}$. Here, the latter provides faster convergence, cf. Sec. 4.6.1.4.

4.4.3 Recursive RETF Update

Based upon the square-root model in (4.28), the estimates $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\varphi}}_s^{1/2}$ obtained as discussed in Sec. 4.4.2 may also be used to recursively update the RETF estimate $\hat{\mathbf{H}}$. In the following, we differentiate the prior and posterior estimates $\hat{\mathbf{H}}$ and $\hat{\mathbf{H}}^+$, and propose to simply propagate the posterior in the previous frame to the prior in the current frame, i.e.

$$\hat{\mathbf{H}}(l) = \hat{\mathbf{H}}^+(l-1). \tag{4.44}$$

In each frame, we use $\hat{\mathbf{H}}$ to obtain $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\varphi}}_s^{1/2}$ with (4.35)–(4.36), and then use $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\varphi}}_s^{1/2}$ to obtain $\hat{\mathbf{H}}^+$, where we again resort to the square-root signal model in (4.28). To this end, we define the approximation error as a function of $\mathbf{H}$,

$$\mathbf{E}_{sq}(\mathbf{H}) = \hat{\boldsymbol{\Psi}}_{x_e}^{1/2}\hat{\boldsymbol{\Omega}} - \mathbf{H}\,\mathrm{Diag}[\hat{\boldsymbol{\varphi}}_s^{1/2}], \tag{4.45}$$

which is similar to (4.32). Based upon (4.45) and the constraint in (4.6), we define the minimization problem,

$$\boxed{\begin{aligned}\hat{\mathbf{H}}^+ = \arg\min_{\mathbf{H}} \big\|\mathbf{E}_{sq}(\mathbf{H})\big\|_F^2 + \big\|(\hat{\mathbf{H}} - \mathbf{H})\,\mathrm{Diag}\big[\sqrt{\boldsymbol{\beta}}\,\big]\big\|_F^2 \\ \text{s. t.} \quad \mathbf{i}^T\mathbf{H} = \mathbf{1}^T,\end{aligned}} \tag{4.46}$$

where the penalty term $\big\|(\hat{\mathbf{H}}-\mathbf{H})\,\mathrm{Diag}\big[\sqrt{\boldsymbol{\beta}}\,\big]\big\|_F^2$ relates to Levenberg-Marquardt regularization [179, 180] in that it penalizes deviation from the previous (i.e., the prior) estimate $\hat{\mathbf{H}}$. Here, we leave $\boldsymbol{\beta}$ subject to tuning, as outlined below. In this respect, recall that according to (4.28), both $\boldsymbol{\Psi}_{x_e}^{1/2}$ and $\mathbf{H}$ span the same column space. However, due to modeling and estimation errors, this is not necessarily true for the corresponding estimates $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$ and $\hat{\mathbf{H}}$. In particular, if

the n^{th} source image has a comparably low early PSD φ_{s_n} or is inactive, then the associated subspace dimension will not be well or not at all be represented in $\hat{\mathbf{\Psi}}_{x_e}^{1/2}$, and both $[\hat{\mathbf{\Omega}}_s]_{:,n}$ and $\hat{\varphi}_{s_n}^{1/2}$ may exhibit comparably large estimation errors. Further, the estimate $\hat{\varphi}_{s_n}^{1/2}$ may contain residual late reverberation due to erroneous separation of $\hat{\mathbf{\Psi}}_x$ into $\hat{\mathbf{\Psi}}_{x_e}$ and $\mathbf{\Psi}_{x_\ell}$, cf. (4.9), Sec. 4.5 and Sec. 4.6. In such a case, one would preferably rely on the prior estimate $\hat{\mathbf{h}}_n$ instead of updating based on $[\hat{\mathbf{\Omega}}_s]_{:,n}$ and $\hat{\varphi}_{s_n}^{1/2}$. Considering the solution to (4.46), which is given by

$$[\hat{\mathbf{H}}^{+}]_{1,:} = \mathbf{1}^T, \tag{4.47}$$

$$[\hat{\mathbf{H}}^{+}]_{2:M,:} = \big[(\hat{\mathbf{\Psi}}_{x_e}^{1/2}\hat{\mathbf{\Omega}}\,\mathrm{Diag}[\hat{\boldsymbol{\varphi}}_s^{H/2}] + \hat{\mathbf{H}}\,\mathrm{Diag}[\boldsymbol{\beta}]\big) \cdot \mathrm{Diag}^{-1}[\hat{\boldsymbol{\varphi}}_s + \boldsymbol{\beta}]\big]_{2:M,:}, \tag{4.48}$$

we indeed find that the smaller $\hat{\varphi}_{s_n}$ as compared to $\beta_n = [\boldsymbol{\beta}]_n$, the more $\hat{\mathbf{h}}_n^+$ relies on $\hat{\mathbf{h}}_n$, as desired. In order to further increase robustness against modeling and estimation errors, source inactivity and residual late reverberation in $\hat{\varphi}_{s_n}$, we propose to make β_n time-varying with binary values. More precisely, we base β_n on the power ratio

$$\boldsymbol{\xi} = \hat{\boldsymbol{\varphi}}_s/(\mathbf{1}^T\hat{\boldsymbol{\varphi}}_s + \varphi_{reg}), \tag{4.49}$$

where $\xi_n = [\boldsymbol{\xi}]_n \in [0,\,1]$. Here, φ_{reg} may be used for regularization, e.g., we may choose $\varphi_{reg} = \varphi_{x_\ell}$ in order to limit ξ_n in frames where pre-dominantly late reverberation is estimated. Given ξ_n, we set β_n as

$$\beta_n \begin{cases} = \beta & \text{if} \quad \xi_n \geq \xi_{th}, \\ \to \infty & \text{else}, \end{cases} \tag{4.50}$$

and thereby resort to $\hat{\mathbf{h}}_n^+ = \hat{\mathbf{h}}_n$ if ξ_n is smaller than the pre-defined threshold ξ_{th}. The value β, used if $\xi_n \geq \xi_{th}$, should scale in relation to the dynamic range of φ_{s_n} and may be chosen depending on the (estimated) probability density function of the complex STFT coefficients s_n, cf. Sec 4.6.

Note that in order to start the recursion defined by (4.44), (4.35)–(4.36), and (4.46), an initial estimate $\hat{\mathbf{H}}(0)$ is required, which may be based on, e.g., initial single-source RETF estimates acquired from segments with mutual-exclusively active sources [147], or some initial knowledge or estimates of the associated DoAs [84, 145, 146].

4.5 Subspace-based Early Correlation Matrix Estimation

In Sec. 4.3 and Sec. 4.4, we respectively assumed that the early-correlation-matrix estimate $\hat{\mathbf{\Psi}}_{x_e}$ and its square root $\hat{\mathbf{\Psi}}_{x_e}^{1/2}$ of rank N are available. In this section, we discuss how to obtain these estimates from the microphone signals $\mathbf{x}$. We estimate $\mathbf{\Psi}_x = \mathrm{E}[\mathbf{x}\mathbf{x}^H]$ by recursively averaging $\mathbf{x}\mathbf{x}^H$, yielding the *smooth* estimate $\hat{\mathbf{\Psi}}_{x|sm}$ and its equally *smooth* subspace representation based on the GEVD. From the latter, we first define a *desmoothed* estimate $\hat{\mathbf{\Psi}}_x$, and second extract the early component $\hat{\mathbf{\Psi}}_{x_e}$ and its square root $\hat{\mathbf{\Psi}}_{x_e}^{1/2}$.

In Sec. 4.5.1, we introduce the subspace model of $\mathbf{\Psi}_x$. In Sec. 4.5.2, we obtain the smooth and desmoothed estimates $\hat{\mathbf{\Psi}}_{x|sm}$ and $\hat{\mathbf{\Psi}}_x$, respectively. In Sec. 4.5.3, given $\hat{\mathbf{\Psi}}_x$, we then retrieve subspace-based rank-N estimates $\hat{\mathbf{\Psi}}_{x_e}$ and $\hat{\mathbf{\Psi}}_{x_e}^{1/2}$.

4.5.1 Correlation Matrix Subspace Decomposition

In each frame l, we define the GEVD [150] of $\mathbf{\Psi}_x$ and the diffuse coherence matrix $\mathbf{\Gamma}$, cf. (4.14), i.e.

$$\mathbf{\Psi}_x\mathbf{P} = \mathbf{\Gamma}\mathbf{P}\mathbf{\Lambda}_x, \tag{4.51}$$

$$\text{with} \quad \mathbf{\Lambda}_x = \mathrm{Diag}[\boldsymbol{\lambda}_x], \tag{4.52}$$

where $\boldsymbol{\lambda}_x \in \mathbb{R}^M$ comprises the generalized eigenvalues, and the columns of $\mathbf{P} \in \mathbb{C}^{M\times M}$ comprise the associated generalized eigenvectors. In the GEVD, the generalized eigenvectors in $\mathbf{P}$ are uniquely defined up to a scaling factor, and for any factorization $\mathbf{\Gamma} = \mathbf{\Gamma}^{1/2}\mathbf{\Gamma}^{H/2}$, we find that $\mathbf{\Gamma}^{H/2}\mathbf{P}$ is column-wise orthogonal due to $\mathbf{\Psi}_x$ being Hermitian. In the following, without loss of generality, we assume the eigenvectors to be scaled such that $\mathbf{\Gamma}^{H/2}\mathbf{P}$ is unitary, i.e.

$$\mathbf{P}^H\mathbf{\Gamma}\mathbf{P} = \mathbf{I}, \tag{4.53}$$

and therefore, combining (4.51) and (4.53),

$$\mathbf{P}^H\mathbf{\Psi}_x\mathbf{P} = \mathbf{\Lambda}_x. \tag{4.54}$$

An alternative, but mathematically equivalent formulation to the GEVD in (4.51) is given by the EVD of the pre-whitened matrix $\mathbf{\Psi}'_x = \mathbf{\Gamma}^{-1/2}\mathbf{\Psi}_x\mathbf{\Gamma}^{-H/2}$ [68, 147, 149], which is defined by $\mathbf{\Psi}'_x\mathbf{P}' = \mathbf{P}'\mathbf{\Lambda}'_x$. By comparison with (4.51), we find $\mathbf{\Lambda}'_x = \mathbf{\Lambda}_x$ and $\mathbf{P}' = \mathbf{\Gamma}^{H/2}\mathbf{P}$, provided that the respective (generalized)

eigenvalues are sorted in the same order, and the (generalized) eigenvectors are scaled accordingly.

For convenience of presentation, assume that the generalized eigenvalues in $\boldsymbol{\lambda}_x$ are sorted in a descending order, and the generalized eigenvectors in $\mathbf{P}$ are sorted accordingly. Then, inserting $\boldsymbol{\Psi}_x = \boldsymbol{\Psi}_{x_e} + \boldsymbol{\Psi}_{x_\ell}$ with $\boldsymbol{\Psi}_{x_\ell} = \varphi_{x_\ell}\boldsymbol{\Gamma}$, cf. (4.9) and (4.14), into (4.54) while making use of (4.53) yields

$$\boldsymbol{\Lambda}_x = \mathbf{P}^H \boldsymbol{\Psi}_{x_e} \mathbf{P} + \varphi_{x_\ell} \mathbf{I}, \tag{4.55}$$

wherein $\boldsymbol{\Psi}_{x_e}$ and in consequence $\mathbf{P}^H \boldsymbol{\Psi}_{x_e} \mathbf{P}$ generally have rank N, and the latter in addition is diagonal, i.e. if $N < M$ we have

$$\mathbf{P}^H \boldsymbol{\Psi}_{x_e} \mathbf{P} = \begin{pmatrix} \boldsymbol{\Lambda}_{x_e} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{4.56}$$

$$\text{with} \quad \boldsymbol{\Lambda}_{x_e} = \text{Diag}[\boldsymbol{\lambda}_{x_e}], \tag{4.57}$$

and $\boldsymbol{\lambda}_{x_e} \in \mathbb{R}^N$.

4.5.2 Recursive Correlation Matrix Estimation and Desmoothing

We compute a smooth estimate $\hat{\boldsymbol{\Psi}}_{x|sm}$ of $\boldsymbol{\Psi}_x$ by recursively averaging $\mathbf{x}\mathbf{x}^H$ using some pre-defined forgetting factor $\zeta \in (0,\, 1)$, i.e.

$$\hat{\boldsymbol{\Psi}}_{x|sm}(l) = \zeta \hat{\boldsymbol{\Psi}}_{x|sm}(l-1) + (1-\zeta)\mathbf{x}(l)\mathbf{x}^H(l), \tag{4.58}$$

and perform the GEVD $\hat{\boldsymbol{\Psi}}_{x|sm}\hat{\mathbf{P}} = \boldsymbol{\Gamma}\hat{\mathbf{P}}\hat{\boldsymbol{\Lambda}}_{x|sm}$ similar to (4.51)–(4.54), with $\hat{\mathbf{P}}$ an estimate of $\mathbf{P}$ and $\hat{\boldsymbol{\Lambda}}_{x|sm} = \text{Diag}[\hat{\boldsymbol{\lambda}}_{x|sm}]$ a smooth estimate of $\boldsymbol{\Lambda}_x$. Note that in order to excite all subspace dimensions and the associated generalized eigenvalues and hence to achieve a meaningful decomposition, $\hat{\boldsymbol{\Psi}}_{x|sm}$ needs to be well-conditioned, and so ζ must be sufficiently close to one. As discussed in Sec. 4.2, the PSDs φ_{s_n} and φ_{x_ℓ} may be highly non-stationary, while the associated coherence matrices $\mathbf{h}_n\mathbf{h}_n^H$ and $\boldsymbol{\Gamma}$ are commonly assumed to be comparably slowly time-varying or even time-invariant. In theory, a linear combination of the PSDs φ_{s_n} and φ_{x_ℓ} is rendered by the *unknown* generalized eigenvalues $\boldsymbol{\lambda}_x$ of $\boldsymbol{\Psi}_x$ and $\boldsymbol{\Gamma}$, i.e. also $\boldsymbol{\lambda}_x$ may be highly non-stationary. In contrast, due to the (inevitable) recursive averaging in (4.58), the *computed* generalized eigenvalues $\hat{\boldsymbol{\lambda}}_{x|sm}$ of $\hat{\boldsymbol{\Psi}}_{x|sm}$ and $\boldsymbol{\Gamma}$ are slowly time-varying if ζ is sufficiently large, i.e. non-stationarities are to some extent smoothed, and so would be PSD estimates based on $\hat{\boldsymbol{\lambda}}_{x|sm}$ or $\hat{\boldsymbol{\Psi}}_{x|sm}$. While smooth PSD estimates are

commonly preferred in some applications (e.g., for perceptual reasons, in the computation of spectral gains in speech enhancement [30]), others may exploit non-stationarities (such as, e.g., the Kalman filter [51], where PSD estimates of the observation noise act as a regularization term in the recursive update of the state estimate [181, cf. Ch. 3]). Depending on the application, we therefore propose to restore non-stationarities by desmoothing $\hat{\boldsymbol{\lambda}}_{x|sm}$, yielding an estimate $\hat{\boldsymbol{\lambda}}_x$ of $\boldsymbol{\lambda}_x$.

To this end, we note that the recursive averaging in (4.58) may be considered an element-wise filtering operation with $\mathbf{x}(l)\mathbf{x}^H(l)$ as the input, $\hat{\boldsymbol{\Psi}}_{x|sm}(l)$ as the output, and the (all-pole) z-domain transfer function given by $(1-\zeta)/(1-\zeta z^{-1})$. Therefore, in order to desmooth $\hat{\boldsymbol{\lambda}}_{x|sm}(l)$, we propose to apply the corresponding (all-zero) inverse transfer function given by $(1-\zeta z^{-1})/(1-\zeta)$ followed by non-negative thresholding, i.e.

$$\hat{\boldsymbol{\lambda}}'_x(l) = \frac{\hat{\boldsymbol{\lambda}}_{x|sm}(l) - \zeta\hat{\boldsymbol{\lambda}}_{x|sm}(l-1)}{1-\zeta}, \tag{4.59}$$

$$\hat{\boldsymbol{\lambda}}_x(l) = \max[\hat{\boldsymbol{\lambda}}'_x(l),\, \mathbf{0}], \tag{4.60}$$

where the thresholding in (4.60) avoids negative eigenvalue estimates, which otherwise may appear in a limited number of frames due to modeling and estimation errors. Note that the desmoothing operation requires the associated generalized eigenvalues in $\hat{\boldsymbol{\lambda}}_{x|sm}(l)$ and $\hat{\boldsymbol{\lambda}}_{x|sm}(l-1)$ to be sorted correspondingly. This may be ensured by sorting $\hat{\mathbf{P}}(l)$ such that $\hat{\mathbf{P}}^H(l-1)\boldsymbol{\Gamma}\hat{\mathbf{P}}(l) \approx \mathbf{I}$, cf. (4.53), and $\hat{\boldsymbol{\lambda}}_{x|sm}(l)$ accordingly, which can be done easily for large ζ and the therewith slowly time-varying GEVD [153]. Alternatively, recursive sorting may be avoided if the GEVD is estimated recursively, e.g., by means of the power method [182,183]. One may then define the corresponding desmoothed estimate $\hat{\boldsymbol{\Psi}}_x$ via its decomposition

$$\hat{\boldsymbol{\Psi}}_x\hat{\mathbf{P}} = \boldsymbol{\Gamma}\hat{\mathbf{P}}\hat{\boldsymbol{\Lambda}}_x, \tag{4.61}$$

$$\text{with} \quad \hat{\boldsymbol{\Lambda}}_x = \text{Diag}[\hat{\boldsymbol{\lambda}}_x], \tag{4.62}$$

where $\hat{\mathbf{P}}$ remains unchanged.

4.5.3 Early Correlation Matrix Estimation and Factorization

Given $\hat{\mathbf{P}}$ and $\hat{\boldsymbol{\Lambda}}_x$ in (4.61)–(4.62), we now retrieve the subspace-based rank-N estimates $\hat{\boldsymbol{\Psi}}_{x_e}$ and $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$. To this end, based on (4.55)–(4.57), we note that $\boldsymbol{\lambda}_{x_e}$ may be estimated as

$$\hat{\boldsymbol{\lambda}}_{x_e} = [\hat{\boldsymbol{\lambda}}_x]_{1:N} - \hat{\varphi}_{x_\ell}\mathbf{1}, \tag{4.63}$$

where $\hat{\varphi}_{x_\ell}$ in turn may be obtained by averaging the last $M-N$ generalized eigenvalues in $[\hat{\boldsymbol{\lambda}}_x]_{N+1:M}$ [68]. Considering (4.56)–(4.57), given $\hat{\boldsymbol{\Lambda}}_{x_e} = \mathrm{Diag}[\hat{\boldsymbol{\lambda}}_{x_e}]$ from (4.63) and $\hat{\mathbf{P}}^{-1} = \hat{\mathbf{P}}^H \boldsymbol{\Gamma}$ from (4.53), we can define a rank-N estimate of $\boldsymbol{\Psi}_{x_e}$ as

$$\begin{aligned}\hat{\boldsymbol{\Psi}}_{x_e} &= \boldsymbol{\Gamma}\hat{\mathbf{P}} \begin{pmatrix} \hat{\boldsymbol{\Lambda}}_{x_e} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \hat{\mathbf{P}}^H \boldsymbol{\Gamma} \\ &= \boldsymbol{\Gamma}[\hat{\mathbf{P}}]_{:,1:N} \hat{\boldsymbol{\Lambda}}_{x_e} [\hat{\mathbf{P}}]_{:,1:N}^H \boldsymbol{\Gamma}. \end{aligned} \tag{4.64}$$

From (4.64), we can further easily derive a square root $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$ as

$$\hat{\boldsymbol{\Psi}}_{x_e}^{1/2} = \boldsymbol{\Gamma}[\hat{\mathbf{P}}]_{:,1:N} \hat{\boldsymbol{\Lambda}}_{x_e}^{1/2} \tag{4.65}$$

$$\text{with} \quad \hat{\boldsymbol{\Lambda}}_{x_e}^{1/2} = \mathrm{Diag}[\hat{\boldsymbol{\lambda}}_{x_e}^{1/2}], \tag{4.66}$$

with arbitrary complex arguments of the elements in $\hat{\boldsymbol{\lambda}}_{x_e}^{1/2}$.

Note that as opposed to the order presented in Sec. 4.5.2 and this section, we may also apply desmoothing only after obtaining a smooth estimate of the early correlation matrix and its square root, which showed to yield comparable results in our simulations.

4.6 Simulations

In this section, we compare the algorithms based on the conventional and the square-root MP as presented in Sec. 4.3 and Sec. 4.4, respectively. We assume that an (initial) RETF estimate $\hat{\mathbf{H}}$ is available, and that $\hat{\boldsymbol{\Psi}}_{x_e}$ and $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$ are obtained as described in Sec. 4.5.

Apart from estimation errors in $\hat{\boldsymbol{\Psi}}_{x_e}$, $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$, and $\hat{\mathbf{H}}$, the performance of both algorithms is subject to modeling errors, cf. Sec. 4.2. Unfortunately, due to the model deficiencies in (4.9)–(4.15), exact and observable ground truth early PSDs $\boldsymbol{\varphi}_s$ and ground truth RETFs $\mathbf{H}$ do not exist in a practical setup based on realistic acoustic data. Therefore, in order to yield a broader understanding of the algorithms' behavior, we perform two kinds of simulations. In the first kind, instead of generating time-domain data and estimating $\boldsymbol{\Psi}_x$ in the STFT domain, we generate $\hat{\boldsymbol{\Psi}}_x = \boldsymbol{\Psi}_x$ directly based on (4.9)–(4.14) and assumed geometric and physical properties, i.e. $\hat{\boldsymbol{\Psi}}_x$ is free of modeling and estimation errors. This way, we are able to define exact ground truth early PSDs $\boldsymbol{\varphi}_s$ and ground truth RETFs $\mathbf{H}$ that may be used to define exact performance measures.

Further, the estimates $\hat{\mathbf{\Psi}}_{x_e}$ and $\hat{\mathbf{\Psi}}_{x_e}^{1/2}$ obtained as described in Sec. 4.5 will be free of estimation errors, such that the performance of both algorithms depends on the RETF estimation error in $\hat{\mathbf{H}}$ and the algorithmic settings in Sec. 4.3 and Sec. 4.4 only. We refer to these simulations as the *model-based-data* case. In the second kind of simulations, we generate acoustic data in the time domain from recorded speech signals and measured room impulse responses (RIRs), and estimate $\mathbf{\Psi}_x$ in the STFT domain. This way, the setup becomes more practical, however, evaluation becomes less trivial in terms of the definition of performance measures, such that we need to define and rely on an approximate ground truth early PSD $\tilde{\boldsymbol{\varphi}}_s$ as a reference. We refer to these simulations as the *acoustic-data* case. The model-based-data case and the acoustic-data case are discussed in Sec. 4.6.1 and Sec. 4.6.2, respectively.

4.6.1 Model-based Data

We define our performance measures in Sec. 4.6.1.1, discuss the data-generation in Sec. 4.6.1.2, the algorithmic settings in Sec. 4.6.1.3, and the evaluation results in Sec. 4.6.1.4.

4.6.1.1 Performance Measures

We define the RETF estimation error,

$$\mathbf{E}_H = \hat{\mathbf{H}} - \mathbf{H}, \tag{4.67}$$

where $\mathbf{i}^T\mathbf{E}_H = [\mathbf{E}_H]_{1,:} = \mathbf{0}^T$ since both $\mathbf{H}$ and $\hat{\mathbf{H}}$ satisfy (4.6), and based on that the relative squared RETF estimation error,

$$\varepsilon_H = 10\log_{10}\frac{\mathrm{tr}[\mathbf{E}_H^H\mathbf{E}_H]}{\mathrm{tr}[\mathbf{H}^H\mathbf{H}] - N}\mathrm{dB}, \tag{4.68}$$

where we subtract N in the denominator in order to compensate for the fact that the first row of $\mathbf{H}$ is known. Since the early PSDs $\boldsymbol{\varphi}_s$ are already a second-order property of the underlying signal $\mathbf{s}$, we define the PSD estimation error with respect to the non-negative square root of $\hat{\boldsymbol{\varphi}}_s$ and $\boldsymbol{\varphi}_s$, i.e.

$$\mathbf{e}_{\varphi_s} = \sqrt{\hat{\boldsymbol{\varphi}}_s} - \sqrt{\boldsymbol{\varphi}_s}, \tag{4.69}$$

and based on that the relative squared PSD estimation error,

$$\varepsilon_{\varphi_s} = 10\log_{10}\frac{\mathbf{e}_{\varphi_s}^T\mathbf{e}_{\varphi_s}}{\mathbf{1}^T\boldsymbol{\varphi}_s}\,\mathrm{dB}. \tag{4.70}$$

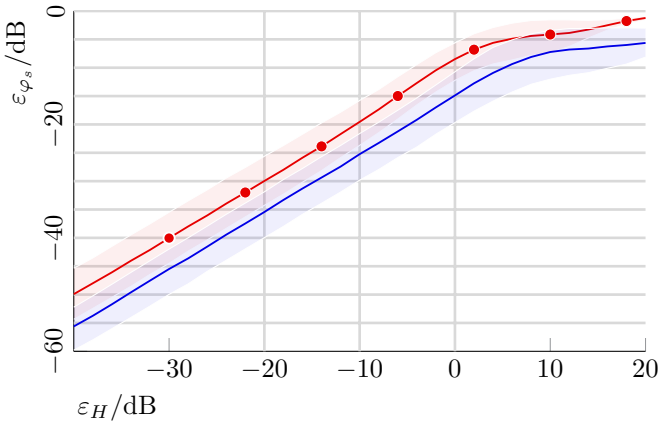


Figure 4.1: ε_{φ_s} versus ε_H for conventional MP [—•—] and square-root MP [——] with $\alpha = 10^3$ at $f = 2\,\text{kHz}$.

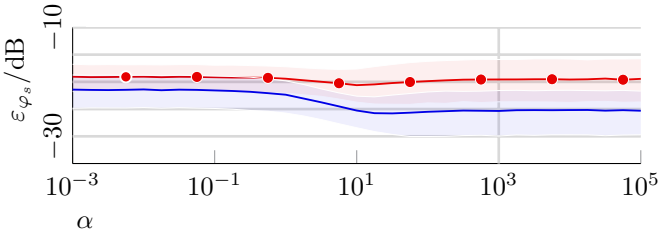


Figure 4.2: ε_{φ_s} versus α for conventional MP [—•—] and square-root MP [——] at $\varepsilon_H = -10\,\text{dB}$ and $f = 2\,\text{kHz}$.

4.6.1.2 Data Generation

Let $\hat{\mathbf{\Psi}}_x$ be available and free of modeling and estimation errors, i.e. we have $\hat{\mathbf{\Psi}}_x = \mathbf{\Psi}_x$ with $\mathbf{\Psi}_x$ adhering to (4.9)–(4.14). We generate $\mathbf{\Psi}_x$ based on assumed geometric and physical properties. We assume a linear microphone array of $M = 5$ microphones with inter-microphone distance of 8 cm and the speed of sound to be 340 m/s. Further, we assume $N = 3$ sources, positioned at $(-30, 0, 60)^\circ$ relative to the broadside direction of the microphone array. The RETFs $\mathbf{H}$ are generated assuming omnidirectional microphones of equal gain as well as free- and far-field propagation for the early components, i.e. $\mathbf{H}$ depends on the DoAs only and is fully defined by the corresponding phase shifts between microphones. The estimate $\hat{\mathbf{H}}$ is generated by adding an error component $\mathbf{E}_H$ according to (4.67), where the elements $[\mathbf{E}_H]_{:,2:M}$ are drawn from independent complex Gaussian distributions, yielding a particular ε_H according to (4.68). The diffuse coherence matrix $\mathbf{\Gamma}$ is computed assuming a spherical-isotropic

sound field. The early PSDs $\boldsymbol{\varphi}_s$ are generated in the following manner. We draw the real and imaginary parts of the elements of $\mathbf{s}$ from independent Laplace distributions, which is a commonly assumed distribution for STFT coefficients of speech [184, 185], i.e. we have $\Re[s_n] \sim (1/b)e^{-2|\Re[s_n]|/b}$ and $\Im[s_n] \sim (1/b)e^{-2|\Im[s_n]|/b}$, where the scaling parameter b is referred to as diversity. Then, we define $\boldsymbol{\varphi}_s = \mathrm{Diag}[\mathbf{s}^H]\mathbf{s}$, i.e. $\boldsymbol{\varphi}_s$ is the squared magnitude of $\mathbf{s}$. Given the above, we set $\boldsymbol{\Psi}_{x_e} = \mathbf{H}\boldsymbol{\varphi}_s\mathbf{H}^H$ according to (4.10). Note that since $\hat{\boldsymbol{\Psi}}_x = \boldsymbol{\Psi}_x$ is free of modeling errors, where $\boldsymbol{\Psi}_x = \boldsymbol{\Psi}_{x_e} + \boldsymbol{\Psi}_{x_\ell}$ with $\boldsymbol{\Psi}_{x_\ell} = \varphi_{x_\ell}\boldsymbol{\Gamma}$, cf. (4.9) and (4.14), the component $\boldsymbol{\Psi}_{x_e}$ may be perfectly estimated from $\hat{\boldsymbol{\Psi}}_x$ by means of the GEVD as described in Sec. 4.5.3, yielding $\hat{\boldsymbol{\Psi}}_{x_e} = \boldsymbol{\Psi}_{x_e}$ independently of φ_{x_ℓ}. Further, note that next to $\mathbf{H}$ and $\boldsymbol{\varphi}_s$, via the GEVD, also $\boldsymbol{\Gamma}$ influences the shape of the square root $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2} = \boldsymbol{\Psi}_{x_e}^{1/2}$ in the sense of defining the basis for a given vector space, cf. Sec. 4.5.3. For each data-point in the evaluation, cf. Sec. 4.6.1.4, we simulate 2^{14} realizations of $\hat{\boldsymbol{\Psi}}_{x_e}$, $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$ and $\hat{\mathbf{H}}$.

4.6.1.3 Algorithmic Settings

In the model-based-data case, as opposed to the acoustic-data case, cf. Sec. 4.6.2.3, the sampling frequency and STFT-processing parameters are irrelevant since we generate $\hat{\boldsymbol{\Psi}}_x$ directly in the STFT-domain, cf. Sec. 4.6.1.2. Regardless, we simulate frequencies up to $f = 8\,\mathrm{kHz}$, corresponding to a virtual sampling frequency of $f_s = 16\,\mathrm{kHz}$. The soft-constraint penalty factor α in the conventional MP in (4.23) and the square-root MP in (4.34) is simulated in the range $\alpha \in [10^{-3}, 10^5]$. We perform at most $i_{max} = 20$ iterations of the associated iterative algorithms in (4.24)–(4.25) and (4.35)–(4.36). All but one of our simulations consider a single frame l only. In the one simulation considering recursive behavior, we do not update $\hat{\mathbf{H}}$ for the conventional MP, but we do update $\hat{\mathbf{H}}$ recursively for the square-root MP as described in Sec. 4.4.3. In the latter case, in (4.49), since $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2} = \boldsymbol{\Psi}_{x_e}^{1/2}$ is free of modeling and estimation errors and therefore free of residual late reverberation, cf. Sec. 4.6.1.2, we set $\varphi_{reg} = 0$. In (4.50), the threshold ξ_{th} is set as $10\log_{10}\xi_{th} = -2\,\mathrm{dB}$ and β is set as $\beta = 20b^2$, with b the diversity of the Laplace distributions used in the generation of $\boldsymbol{\varphi}_s$, cf. Sec. 4.6.1.2.

4.6.1.4 Results

Fig. 4.1 shows the PSD estimation performance in terms of the relative squared PSD estimation error ε_{φ_s} for different values of the relative squared RETF estimation error ε_H for the algorithms based on the conventional MP [–•–] and the square-root MP [——] with $\alpha = 10^3$ at $f = 2\,\mathrm{kHz}$ within a single frame

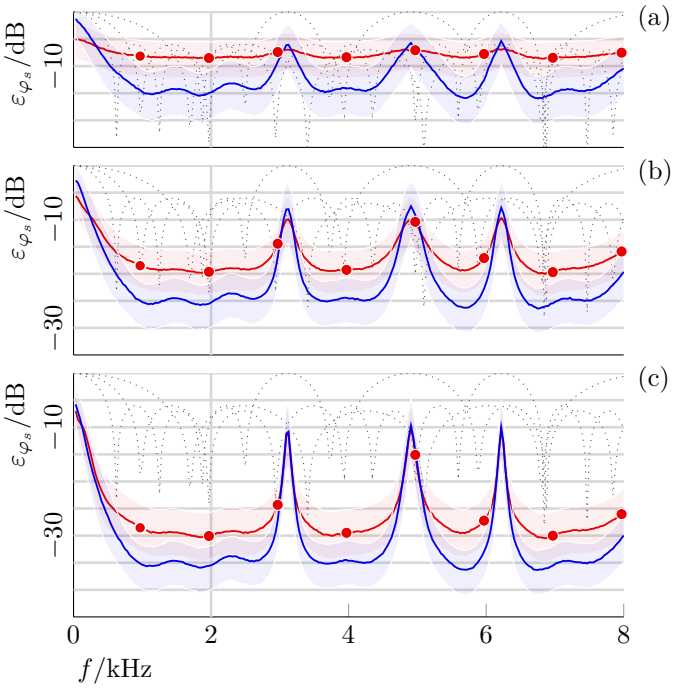


Figure 4.3: ε_{φ_s} versus f for conventional MP [–•–] and square-root MP [——] with $\alpha = 10^3$ at (a) $\varepsilon_H = 0\,\text{dB}$, (b) $\varepsilon_H = -10\,\text{dB}$, and (c) $\varepsilon_H = -20\,\text{dB}$. The graphs denoted by [·····] correspond to $10\log_{10}|\mathbf{h}_n^H\mathbf{h}_{n'}|/M\,\text{dB}$ for $n' \neq n$.

l. In this figure and similar ones in the following, the graphs denote medians over all 2^{14} realizations, cf. Sec. 4.6.1.2, and the shaded areas denote the range from the first to the third quartile. As can be seen, for both the conventional MP and the square-root MP, ε_{φ_s} increases at a rate of about 10 dB per 10 dB increase in ε_H until roughly $\varepsilon_H = 0\,\text{dB}$ and $\varepsilon_H = 5\,\text{dB}$ is reached, respectively, after which ε_{φ_s} begins to saturate. This saturation is due to the fact that both algorithms yield non-negative estimates $\hat{\boldsymbol{\varphi}}_s \geq \mathbf{0}$, which limits the estimation error at high values of ε_H. The square-root MP outperforms the conventional MP by at least 5.7 dB for $\varepsilon_H \leq 0\,\text{dB}$, and by somewhat less for $\varepsilon_H \geq 5\,\text{dB}$.

Fig. 4.2 illustrates ε_{φ_s} for different values of the soft constraint penalty factor α for the conventional MP [–•–] and the square-root MP [——] at $\varepsilon_H = -10\,\text{dB}$ and $f = 2\,\text{kHz}$ within a single frame l. We note that while α hardly impacts the performance of the conventional MP, we generally reach larger improvements for higher values of α in the square-root MP. Recall that the soft constraint in the conventional MP is scalar-based, cf. (4.22), while the soft constraint in the square-root MP is vector-based, cf. (4.33), and is therefore more informative.

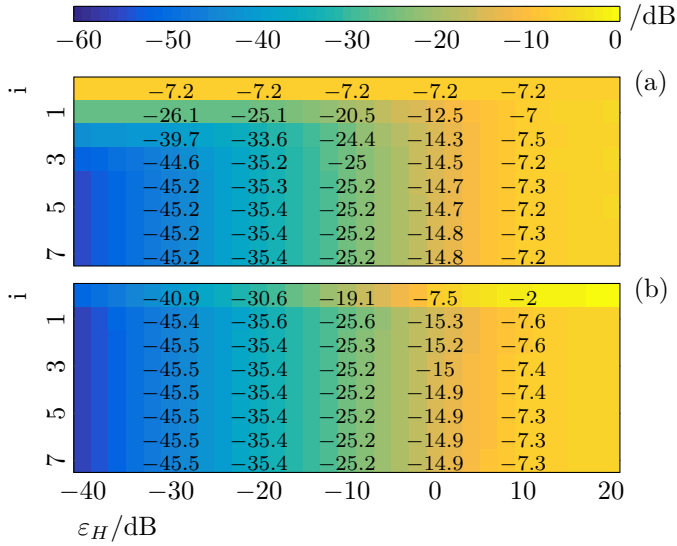


Figure 4.4: $\varepsilon_{\varphi_s}^{(i)}$ versus ε_H and i for square-root MP with $\alpha = 10^3$ and $\hat{\boldsymbol{\varphi}}_s^{1/2|(0)}$ based upon (a) the sum constraint in (4.13) and (b) the estimator in (4.17)–(4.18) at $f = 2\,\mathrm{kHz}$.

The square-root MP outperforms the conventional MP by 2.5 dB at low values of α, and by 5.7 dB at high values of α. Interestingly, for both algorithms, despite $\hat{\boldsymbol{\Psi}}_{x_e}$ and $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$ being free of estimation errors, the minimum of ε_{φ_s} does not occur at the highest values of α, but at around $\alpha = 10^1$. As compared to higher values, the improvement is however mild.

Fig. 4.3 illustrates ε_{φ_s} for different frequencies f for the conventional MP [−•−] and the square-root MP [——] with $\alpha = 10^3$ at (a) $\varepsilon_H = 0\,\mathrm{dB}$, (b) $\varepsilon_H = -10\,\mathrm{dB}$, and (c) $\varepsilon_H = -20\,\mathrm{dB}$ within a single frame l. Note that at some frequencies, due to spatial aliasing, which occurs for two different DoAs if their phase difference in each microphone is a multiple of 2π, the two corresponding DoA-based RETFs in $\mathbf{H}$, cf. Sec. 4.6.1.2, will be identical, and therefore $\mathbf{H}$ itself and consequently also $\boldsymbol{\Psi}_{x_e}$ and $\boldsymbol{\Psi}_{x_e}^{1/2}$ will be rank-deficient. In our setup, this situation occurs for $f \in \{3.11, 4.91, 6.22\}\,\mathrm{kHz}$, cf. also the dotted lines [·····] corresponding to $10\log_{10}|\mathbf{h}_n^H\mathbf{h}_{n'}|/M\,\mathrm{dB}$ for $n' \neq n$, which reach 0 dB if $\mathbf{h}_{n'} = \mathbf{h}_n$. As expected, by comparing Fig. 4.3 (a) to Fig. 4.3 (c), neither of the two algorithms performs well in the proximity of these frequencies, independent of ε_H. At other frequencies, however, the square-root MP outperforms the conventional MP by roughly 5 to 7 dB.

Fig. 4.4 demonstrates the effect in the median of the initial estimate $\hat{\boldsymbol{\varphi}}_s^{1/2|(0)}$

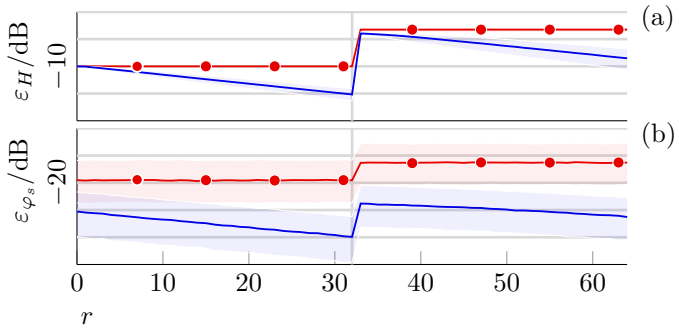


Figure 4.5: (a) $\varepsilon_H(l+r)$ and (b) $\varepsilon_{\varphi_s}(l+r)$ versus r for conventional MP [–•–] and square-root MP [——] with $\alpha = 10^3$ at $f = 2\,\text{kHz}$ and $\varepsilon_H(l) = 0\,\text{dB}$ if $\mathbf{H}$ changes at $r = 32$ and remains constant otherwise.

on the convergence behavior in terms of the relative squared PSD estimation error $\varepsilon_{\varphi_s}^{(i)}$ at iteration i for different values of ε_H of the iterative algorithm in (4.35)–(4.36) solving the square-root MP with $\alpha = 10^3$ at $f = 2\,\text{kHz}$. The initial value is based on (a) the sum constraint in (4.13) as $\hat{\boldsymbol{\varphi}}_s^{1/2|(0)} = \sqrt{[\hat{\boldsymbol{\Psi}}_{x_e}]_{1,1}/N}\,\mathbf{1}$, and (b) the estimator in (4.17)–(4.18), here denoted by $\hat{\boldsymbol{\varphi}}_{s|c_0}$, as $\hat{\boldsymbol{\varphi}}_s^{1/2|(0)} = \sqrt{\hat{\boldsymbol{\varphi}}_{s|c_0}}$. In both cases, the algorithm converges to almost the same final value of ε_{φ_s}. However, we find that in (a), convergence is reached at around $i = 3$ to $i = 4$, while in (b), due to the improved initial estimate, convergence is reached at $i = 1$ already. Hence, while the computation of the initial estimate in (b) is somewhat more expensive, we save 2 to 3 iterations as compared to (a).

Fig. 4.5 demonstrates the recursive behavior in terms of (a) $\varepsilon_H(l+r)$ and (b) $\varepsilon_{\varphi_s}(l+r)$ with r the recursion index for the conventional MP [–•–] and the square-root MP [——] with $\alpha = 10^3$ at $f = 2\,\text{kHz}$ and $\varepsilon_H(l) = 0\,\text{dB}$. Here, the source positioned at $-30°$ transitions to $-40°$ at $r = 32$, resulting in a transient change in the otherwise constant RETF $\mathbf{H}$. While no update of the estimate $\hat{\mathbf{H}}$ is performed for the conventional MP, we do update $\hat{\mathbf{H}}$ recursively for square-root MP as described in Sec. 4.4.3. For the conventional MP, we expectably find that $\varepsilon_H(l+r)$ and $\varepsilon_{\varphi_s}(l+r)$ remain constant except for a transient increase of 6.8 dB and 3.2 dB at $r = 33$, respectively. For the square-root MP, due to the recursive update of $\hat{\mathbf{H}}$, we find that $\varepsilon_H(l+r)$ and $\varepsilon_{\varphi_s}(l+r)$ decrease by 5.2 dB and 4.7dB over the course of the first 32 recursions, followed by an increase of 11.2 dB and 6.1 dB at $r = 33$, respectively, and a subsequent decrease at roughly the same rate.

4.6.2 Acoustic Data

We define the performance measures in Sec. 4.6.2.1, discuss the acoustic scenario in Sec. 4.6.2.2, the algorithmic settings in Sec. 4.6.2.3, and the evaluation results in Sec. 4.6.2.4.

4.6.2.1 Performance Measures

In the acoustic-data case, due to the model deficiencies in (4.9)–(4.15), cf. Sec. 4.2, exact and observable ground truth early PSDs $\boldsymbol{\varphi}_s$ and ground truth RETFs $\mathbf{H}$ do unfortunately not exist, and so the performance measures in (4.67)–(4.70) cannot be used. However, one may define approximate ground truth early PSDs $\tilde{\boldsymbol{\varphi}}_s$ as a reference for evaluation. To this end, given the source signals and RIRs of a particular acoustic scenario, cf. Sec. 4.6.2.2, we convolve the source signals with only the early part of the RIR to the first microphone and transform to the STFT-domain, yielding $\tilde{\mathbf{s}}$, and set $\tilde{\boldsymbol{\varphi}}_s = \text{Diag}[\tilde{\mathbf{s}}^H]\tilde{\mathbf{s}}$, i.e. $\tilde{\boldsymbol{\varphi}}_s$ is the squared magnitude[3] of $\tilde{\mathbf{s}}$. Note that the definition of the early part of the RIR is somewhat arbitrary due to the weighted and overlapping windows in the STFT-processing. For STFT windows of N_{STFT} samples with 50% overlap, one may, e.g., choose the first N_{STFT} or the first $N_{STFT}/2$ taps of the RIR. Here, we have chosen the first N_{STFT} samples corresponding to 32 ms, cf. Sec. 4.6.2.3. In our setup, we have found that different choices result in quantitatively different performance, but not qualitatively different conclusions.

Given a segment of L frames of $\tilde{\boldsymbol{\varphi}}_s$ and $\hat{\boldsymbol{\varphi}}_s$, we decompose $\sqrt{\hat{\boldsymbol{\varphi}}_s}$ according to [176] as

$$\sqrt{\hat{\boldsymbol{\varphi}}_s} = \sqrt{\bar{\boldsymbol{\varphi}}_s} + \mathbf{e}_{\varphi_s}^{int} + \mathbf{e}_{\varphi_s}^{art}, \tag{4.71}$$

where $\sqrt{\bar{\varphi}_{s_n}}$ is the component of $\sqrt{\hat{\varphi}_{s_n}}$ associated to $\sqrt{\tilde{\varphi}_{s_n}}$, i.e. the correctly estimated component, $e_{\varphi_{s_n}}^{int} = [\mathbf{e}_{\varphi_s}^{int}]_n$ contains components associated to $\sqrt{\tilde{\varphi}_{s_{n'}}}$ with $n' \neq n$, i.e. erroneously estimated leakage or interference components across sources, and $e_{\varphi_{s_n}}^{art} = [\mathbf{e}_{\varphi_s}^{art}]_n$ contains components not associated to any $\sqrt{\tilde{\varphi}_{s_n}}$, i.e. erroneously estimated artifact components. Exemplary spectrograms illustrating the decomposition in (4.71) are shown in Fig. 4.6, cf. also the discussion in Sec. 4.6.2.4.

Given L frames of $\sqrt{\bar{\boldsymbol{\varphi}}_s}$, $\mathbf{e}_{\varphi_s}^{int}$ and $\mathbf{e}_{\varphi_s}^{art}$, we define the signal-to-interference ratio $SIR(\kappa)$, the signal-to-artifacts ratio $SAR(\kappa)$, and the signal-to-distortion ratio

[3] If subspace-based desmoothing, cf. Sec. 4.5.2, is not applied in the computation of $\hat{\boldsymbol{\varphi}}_s$, one may instead choose a recursively averaged version of the squared magnitude as a reference.

$SDR(\kappa)$ per third-octave band κ along the lines of [176] as

$$SIR(\kappa) = 10\log_{10}\frac{\sum_{k,l}\left\|\sqrt{\bar{\boldsymbol{\varphi}}_s}(k,l)\right\|_2^2}{\sum_{k,l}\left\|\mathbf{e}_{\varphi_s}^{int}(k,l)\right\|_2^2}\,\text{dB}, \tag{4.72}$$

$$SAR(\kappa) = 10\log_{10}\frac{\sum_{k,l}\left\|\sqrt{\bar{\boldsymbol{\varphi}}_s}(k,l)+\mathbf{e}_{\varphi_s}^{int}(k,l)\right\|_2^2}{\sum_{k,l}\left\|\mathbf{e}_{\varphi_s}^{art}(k,l)\right\|_2^2}\,\text{dB}, \tag{4.73}$$

$$SDR(\kappa) = 10\log_{10}\frac{\sum_{k,l}\left\|\sqrt{\bar{\boldsymbol{\varphi}}_s}(k,l)\right\|_2^2}{\sum_{k,l}\left\|\mathbf{e}_{\varphi_s}^{int}(k,l)+\mathbf{e}_{\varphi_s}^{art}(k,l)\right\|_2^2}\,\text{dB}, \tag{4.74}$$

with $k = k_\kappa^-, \ldots, k_\kappa^+$ and k_κ^- and k_κ^+ the frequency-bin indices of the lower and upper band limits of third-octave-band κ, and $l = 0, \ldots, L-1$.

The decomposition in (4.71) relies on a segment of L frames of $\tilde{\boldsymbol{\varphi}}_s$ and $\hat{\boldsymbol{\varphi}}_s$ and is done in the following manner. Let $\hat{\boldsymbol{\varphi}}_{s_n}$ be a vector stacking the early PSD estimates φ_{s_n} of source n over L observed frames, i.e. $\hat{\boldsymbol{\varphi}}_{s_n} = \left(\hat{\varphi}_{s_n}(0)\ \cdots\ \hat{\varphi}_{s_n}(L-1)\right)^T$, and let $\tilde{\boldsymbol{\varphi}}_{s_n}$, $\bar{\boldsymbol{\varphi}}_{s_n}$, $\underline{\mathbf{e}}_{\varphi_{s_n}}^{int}$ and $\underline{\mathbf{e}}_{\varphi_{s_n}}^{art}$ be defined equivalently, such that $\sqrt{\tilde{\boldsymbol{\varphi}}_{s_n}} = \sqrt{\bar{\boldsymbol{\varphi}}_{s_n}} + \underline{\mathbf{e}}_{\varphi_{s_n}}^{int} + \underline{\mathbf{e}}_{\varphi_{s_n}}^{art}$, similarly to (4.71). Then, we perform the orthonormal projection of each individual vector $\sqrt{\hat{\boldsymbol{\varphi}}_{s_n}}$ onto the one-dimensional subspace spanned by the corresponding vector $\sqrt{\tilde{\boldsymbol{\varphi}}_{s_n}}$, yielding $\sqrt{\bar{\boldsymbol{\varphi}}_{s_n}}$ with $\sqrt{\bar{\boldsymbol{\varphi}}_{s_n}} \propto \sqrt{\tilde{\boldsymbol{\varphi}}_{s_n}}$, as well as onto the N-dimensional subspace spanned by all N vectors $\sqrt{\tilde{\boldsymbol{\varphi}}_{s_n}}$, yielding $\sqrt{\bar{\boldsymbol{\varphi}}_{s_n}} + \underline{\mathbf{e}}_{\varphi_{s_n}}^{int}$, which then allows us to explicitly compute $\underline{\mathbf{e}}_{\varphi_{s_n}}^{int}$ and $\underline{\mathbf{e}}_{\varphi_{s_n}}^{art}$. For further details, we refer the interested reader to [176].

4.6.2.2 Acoustic Scenario

We use RIRs of 0.61 s reverberation time to a physical linear microphone array of $M = 5$ microphones with an inter-microphone distance of 8 cm [26], similar to the assumed microphone array in Sec. 4.6.1.2. We simulate $N = 2$ sources, using female and male speech [25] as source signals. The sources are assigned to two out of three possible source positions in 2 m distance of the microphone array at $\{-30, 0, 60\}^\circ$ relative to the broad-side direction, yielding six different speaker-source-position combinations. From the two source signal files, we randomly select 32 segments of 5 s each. Per segment-pair, we generate microphone signals for each speaker-source-position combination.

4.6.2.3 Algorithmic Settings

In the acoustic-data case, the sampling frequency is $f_s = 16\,\text{kHz}$, and the STFT-analysis and synthesis is based on square-root Hann windows of $N_{STFT} = 512$ samples (corresponding to 32 ms) with 50% overlap, resulting in $L = 312$ frames per segment. The desmoothed correlation matrix estimate $\hat{\mathbf{\Psi}}_x$ (cf. Sec. 4.5.1 and Sec. 4.5.2) is computed using $\zeta = e^{-N_{STFT}/2f_s\tau}$ with $\tau = 160\,\text{ms}$. As in Sec. 4.6.1.2, $\mathbf{\Gamma}$ is computed assuming a spherical-isotropic sound field. Given $\hat{\mathbf{\Psi}}_x$ and $\mathbf{\Gamma}$, we compute the estimates $\hat{\varphi}_{x_\ell}$, $\hat{\mathbf{\Psi}}_{x_e}$ and $\hat{\mathbf{\Psi}}_{x_e}^{1/2}$ as described in Sec. 4.5.3. We assume that the DoAs are known [84, 145, 146], and compute the (initial) estimate $\hat{\mathbf{H}}$ based on that. Note that in a reverberant environment, where the free-field assumption does not hold, the RETFs are generally not only defined by the DoA, but also by early reflections, and therefore we generally have $\hat{\mathbf{H}} \neq \mathbf{H}$ in our setup. Similarly to the model-based data case, cf. Sec. 4.6.1.3, the penalty factor α in the conventional MP in (4.23) and the square-root MP in (4.34) is simulated in the range $\alpha \in [10^{-3}, 10^{5}]$. We perform at most $i_{max} = 20$ iterations of the associated iterative algorithms in (4.24)–(4.25) and (4.35)–(4.36). While we do not update $\hat{\mathbf{H}}$ for the conventional MP in Sec. 4.3, we consider two cases for the square-root MP in Sec. 4.4, namely first where we do not update $\hat{\mathbf{H}}$, and second where we update $\hat{\mathbf{H}}$ recursively as described in Sec. 4.4.3. In the latter case, in (4.49), since $\hat{\mathbf{\Psi}}_{x_e}^{1/2}$ is subject to modeling and estimation errors and contains residual late reverberation, we set $\varphi_{reg} = \hat{\varphi}_{x_\ell}$. In (4.50), the threshold ξ_{th} is again set as $10\log_{10}\xi_{th} = -2\,\text{dB}$ and β is set per third-octave band κ as $\beta(\kappa) = 20\hat{b}^2(\kappa)$, with $\hat{b}(\kappa)$ pre-defined as the diversity of the Laplace distributions fitted to the real and imaginary parts of the STFT coefficients of a training signal within third-octave band κ. Here, the training signal is generated from the entire female and male speech source signals, cf. Sec. 4.6.2.2, by convolving the early part of the RIR of the first microphone corresponding to a source at 2 m distance at 0° relative to the broadside direction, cf. also the similar segment-wise definition of the reference signal $\tilde{s}_n$ in Sec. 4.6.2.1. Note that while $\hat{b}(\kappa)$ is pre-computed using all STFT coefficients of both male and female speech within third-octave band κ, the actual distributions may vary across speakers, across source positions, across individual frequency bins, and across individual segments, cf. also Sec. 4.6.2.2.

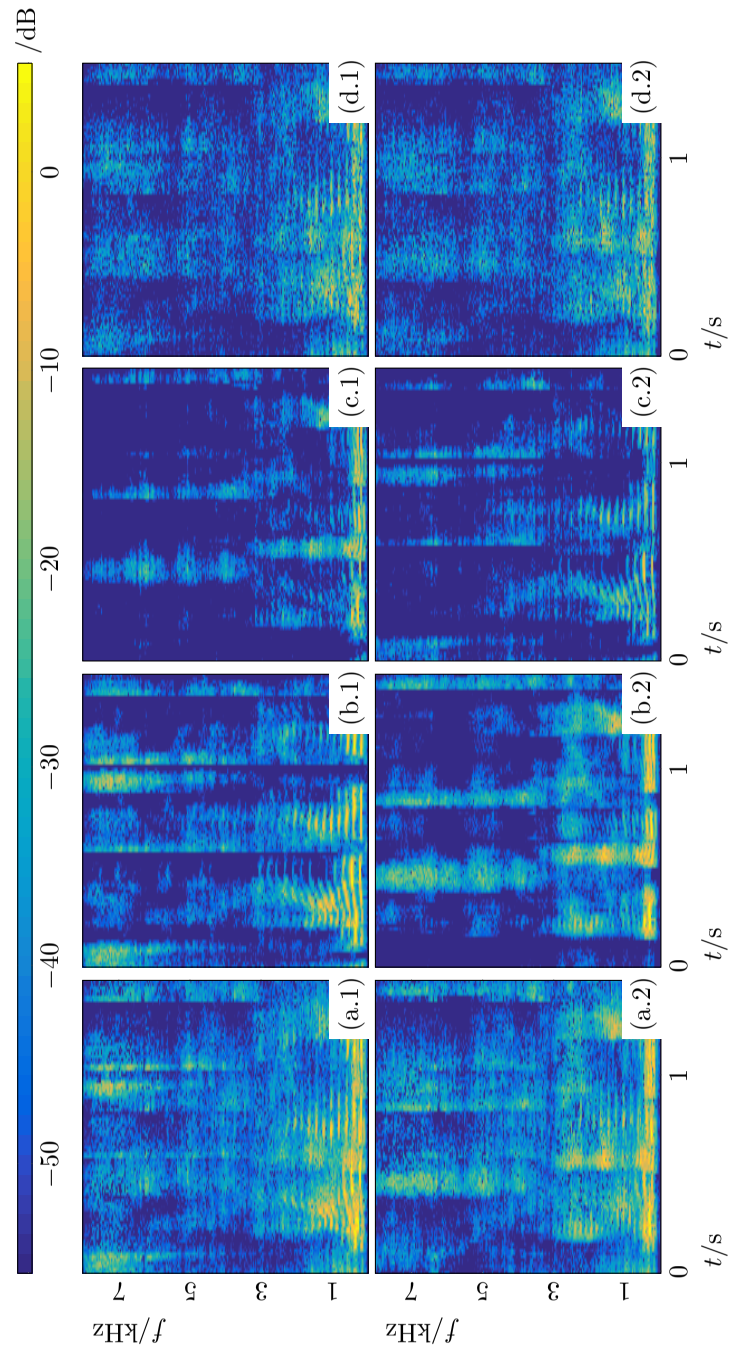

/dB
0
−10
−20
−30
−40
−50
(a.1)
(b.1)
(c.1)
(d.1)
(a.2)
(b.2)
(c.2)
(d.2)
f/kHz
1
3
5
7
t/s
0
1

Figure 4.6: Exemplary spectrograms depicting

(a.n) estimates $\hat{\varphi}_{s_n}$, $n = 1, 2$,

and their decomposition according to Sec. 4.6.2.1, i.e.

(b.n) the target components $\bar{\varphi}_{s_n}$,

(c.n) the interference components $e^{2|int}_{\varphi_{s_n}}$, and

(d.n) the artifacts components $e^{2|art}_{\varphi_{s_n}}$.

The reference PSDs $\tilde{\varphi}_{s_1}$ and $\tilde{\varphi}_{s_2}$ originate from a female and a male speaker at $-30°$ and $60°$, respectively, and the estimates $\hat{\varphi}_{s_n}$ are obtained by means of the square-root MP.

4.6.2.4 Results

Before discussing the performance of the conventional MP and the square-root MP in terms of the measures *SIR*, *SAR*, and *SDR*, we first consider the examplary spectrograms in Fig. 4.6 visualizing the decomposition of $\sqrt{\hat{\boldsymbol{\varphi}}_s}$ upon which these measures are based. In this example, the microphone signals $\mathbf{x}$ and the reference PSDs $\tilde{\varphi}_{s_1}$ and $\tilde{\varphi}_{s_2}$ originate from a female and a male speaker at $-30°$ and $60°$, respectively, and the estimates $\hat{\varphi}_{s_1}$ and $\hat{\varphi}_{s_2}$ in Fig. 4.6 (a.1) and Fig. 4.6 (a.2) are obtained by means of the square-root MP. The correctly estimated components $\bar{\varphi}_{s_1}$ and $\bar{\varphi}_{s_2}$ in Fig. 4.6 (b.1) and Fig. 4.6 (b.2) are frequency-bin-wise scaled versions of the reference PSDs $\tilde{\varphi}_{s_1}$ and $\tilde{\varphi}_{s_2}$, respectively, cf. Sec. 4.6.2.1. As can be seen, the leakage or interference components in $e^{2|int}_{\varphi_{s_1}}$ and $e^{2|int}_{\varphi_{s_2}}$ in Fig. 4.6 (c.1) and Fig. 4.6 (c.2) relate to the opposing reference PSDs, cf. Fig. 4.6 (b.2) and Fig. 4.6 (b.1), respectively. Finally, the artifact components $e^{2|art}_{\varphi_{s_1}}$ and $e^{2|art}_{\varphi_{s_2}}$ in Fig. 4.6 (d.1) and Fig. 4.6 (d.2) do not relate to any of the reference PSDs, but rather to residual late reverberation in the estimate $\hat{\boldsymbol{\Psi}}^{1/2}_{x_e}$, cf. also Sec. 4.5.3, which is due to modeling errors in (4.9)–(4.14) and a potential deviation of the late reverberant sound field from the spatial coherence matrix $\boldsymbol{\Gamma}$. Note that in $e^{2|art}_{\varphi_{s_1}}$ and $e^{2|art}_{\varphi_{s_2}}$, the energy is concentrated in the same spectro-temporal regions, indicating a similar spatial sound field of these components.

Fig. 4.7 shows the median over all segments and speaker-source-combinations, cf. Sec. 4.6.2.2, of (a) *SIR*, (b) *SAR*, and (c) *SDR* in third-octave bands for the conventional MP [■], the square-root MP without recursive RETF update [■], and the square-root MP with recursive RETF update [■]. Here, in each third-octave band κ, we have selected $\alpha(\kappa)$ such that $SIR(\kappa)$ is maximized for each algorithm, i.e. the figure indicates their upper performance limit in terms of $SIR(\kappa)$ with respect to the tuning of $\alpha(\kappa)$. Note that in our setup, selecting $\alpha(\kappa)$ to maximize $SAR(\kappa)$ or $SDR(\kappa)$ does not lead to qualitatively substantial differences. For the conventional MP, we have found values of $\alpha(\kappa) \ll 1$ to be preferable in all third-octave bands κ, indicating that the soft-constraint penalty in (4.23) is not very useful in practice. For the square-root MP, with and without recursive RETF update, we have found $\alpha(\kappa) \gg 1$ to be preferable in third-octave bands below 0.5kHz, and $\alpha(\kappa) \leq 1$ to be preferable above 0.5kHz. From Fig. 4.7 (a), we find that the square-root MP clearly outperforms the conventional MP in terms of *SIR* in third-octave bands above 0.25 kHz, with improvements of 1 dB to 6 dB, indicating better source-component separation performance. Further, for the square-root MP, we find that the recursive RETF update mildly improves the performance by up to 1 dB. Recall that the initial RETF estimate $\hat{\mathbf{H}}$ is based on the correct DoAs, but does not consider early reflections, cf. Sec. 4.6.2.3. From Fig. 4.7 (b), we note that for all algorithms,

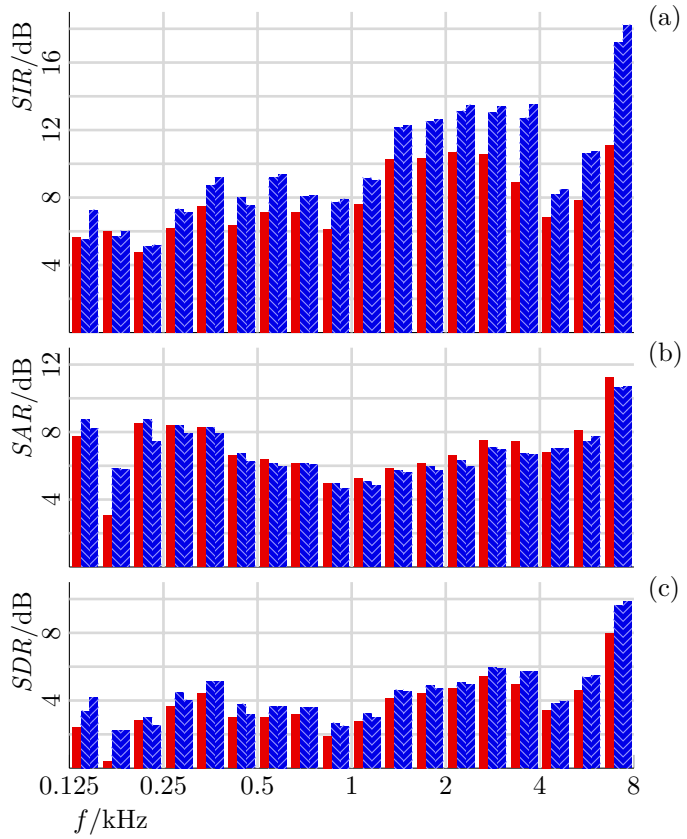


Figure 4.7: (a) *SIR*, (b) *SAR*, and (c) *SDR* in third-octave bands for conventional MP [■], square-root MP without recursive RETF update [▧], and square-root MP with recursive RETF update [▨].

we have $SAR(\kappa) < SIR(\kappa)$ in third-octave bands above 0.5 kHz, indicating comparably strong residual late reverberation. The square-root MP performs slightly worse than the conventional MP in terms of *SAR* in third-octave bands above 0.25 kHz, with degradations of less than 1 dB. In the square-root MP, recursive RETF updating results in minor differences only. As can be seen from Fig. 4.7 (c), we find that the square-root MP outperforms the conventional MP in terms of *SDR*, however, due to the comparably strong residual late reverberation, by much less than in terms of *SIR*. Again, in the square-root MP, recursive RETF updating results in minor differences only.

4.7 Conclusion

We have discussed early PSD estimation and recursive RETF updating in the STFT domain for multiple sources in reverberant environments, based on a commonly used multi-microphone correlation matrix model, given (initial) RETF estimates. State-of-the-art approaches to early PSD estimation formulate a minimization problem on the approximation error with respect to an estimate of the early correlation matrix, referred to as conventional MP. Instead, we here have factorized the early correlation matrix model and formulated a corresponding minimization problem on the approximation error with respect to an estimate of the early-correlation-matrix square root, which we referred to as the square-root MP. The square-root MP seeks a unitary matrix and the square roots of the early PSDs up to an arbitrary complex argument, and therewith constitutes a generalization of the orthogonal Procrustes problem. As opposed to the conventional MP, non-negative inequality constraints are not required in the square-root MP. The square-root MP may be solved iteratively, requiring one SVD per iteration. Based on the estimated unitary matrix and early PSD square roots, we are further able to recursively update the RETF estimate, which is not inherently possible in the conventional approach. The respectively required estimates of the early correlation matrix and the early-correlation-matrix square root may be obtained from an estimate of the microphone signal correlation matrix and the diffuse coherence matrix by means of the GEVD. Hereat, in order to compensate for inevitable recursive averaging, we have restored non-stationarities by desmoothing the generalized eigenvalues.

In order to evaluate the proposed approach, we have performed two kinds of simulations. In the first kind, the data is generated based on the microphone signal correlation matrix model and assumed geometric and physical properties, excluding modeling errors from the evaluation. This is referred to as model-based-data case. In the second kind, the data is generated from recorded speech and measured RIRs, creating a more practical setup. This is referred to as acoustic-data case. In the model-based-data case, the simulation results indicate better performance of the square-root MP as compared to the conventional MP in terms of the relative squared PSD estimation error. If initialized accordingly, the square-root MP can be solved in only one iteration. In the acoustic-data case, the simulation results indicate better performance of the square-root MP as compared to the conventional MP in terms of the source-component separation measured by the signal-to-interference ratio. Both the square-root MP and the conventional MP suffer somewhat from residual late reverberation in the early-correlation-matrix estimate.

Part III

Complexity Reduction

Chapter 5

Low-Comlexity ISCLP Kalman Filters

Abstract

The previously introduced short-time Fourier transform (STFT) domain-based integrated sidelobe cancellation and linear prediction (ISCLP) Kalman filter exhibits a quadratic computational cost in the number of filter coefficients and channels per frequency bin. In this chapter, as low computational complexity is required in many applications, we strive to simplify the ISCLP Kalman filter update equations. Specifically, we propose to enforce the state estimation error correlation matrix to assume sparse structures by negligence of either temporal, spatial, or all cross-correlations, leading to linear cost in the number of filter coefficients, the number of channels, or both the number of filter coefficients and channels, respectively. In simulations, we show that the simplified variants of the ISCLP Kalman filter perform nearly as well as the original variant, thereby permitting far more favourable trade-offs between complexity and performance.

Index terms — Complexity reduction, Kalman filter, integrated sidelobe cancellation and linear prediction.

5.1 Introduction

In Ch. 3, we have proposed the short-time Fourier transform (STFT) domain-based integrated sidelobe cancellation and linear prediction (ISCLP) Kalman filter for joint dereverberation, interfering speech cancellation and noise reduction. In terms of computational complexity, although being cheaper than comparable state-of-the-art approaches [112, 113, 116], cf. Sec. 3.5.1, the ISCLP Kalman filter exhibits a quadratic cost in the number of filter coefficients and channels per frequency bin. In many applications however, such as mobile telephony [34, 35] and hearing devices [39–41], low computational complexity is required due to device-specific hardware constraints such as limited processing power, memory and battery life.

In adaptive filtering, several low-complexity variants of recursive least squares (RLS) exist, e.g., fast transveral RLS algorithms [50, 186] or numerically better-behaved fast QR decomosition-based RLS algorithms [187, 188]. These algorithms reach linear instead of quadratic cost in the number of filter coefficients, where complexity reduction is achieved by exploiting the structure of the filter input vector, which is commonly composed by shifting consecutively delayed samples of the same signal at each recursion.

Here instead, we consider complexity reduction of the Kalman filter by negligence of cross-correlations in the state estimation error correlation matrix,

thereby enforcing it to assume a sparse structure. In audio signal processing, simplifications of this kind have been proposed in [98, 102, 189–192], likewise reaching linear instead of quadratic cost in the number of filter coefficients [102, 189] or, in case of partitioned-block frequency domain (PBFD)-based processing, in the number of filter partitions [98, 190–192]. In the time domain Kalman filter in [189], the state estimation error correlation matrix has been simplified by replacing the outer product of the Kalman gain and filter input vector in the associated update term by a corresponding inner product. In PBFD-based Kalman filtering [98, 190–192], state estimation error correlations across partitions are commonly neglected. Akin to that, we have proposed the STFT domain-based low-complexity multi-channel linear prediction (MCLP) Kalman filter [102], where temporal cross-correlations have been neglected.

This chapter may be viewed as an extension of our previous work on the low-complexity MCLP Kalman filter [102]. We propose to apply the same concept to the ISCLP Kalman filter, and in addition to neglecting temporal cross-correlations obtaining linear cost in the number of filter coefficients, we also consider neglecting spatial cross-correlations as well as neglecting all cross-correlations, obtaining linear cost in the number of channels, and linear cost in both the number of filter coefficients and channels, respectively. In our experimental validation adopting the same simulation setup as for the original ISCLP Kalman filter in Ch. 3, we show that the simplified variants of the ISCLP Kalman filter perform nearly as well as the original variant, thereby permitting far more favourable trade-offs between complexity and performance. A MATLAB implementation is available at [154].

The chapter is organized as follows. In Sec. 5.2, we review the ISCLP Kalman filter. The complexity reduction schemes for the ISCLP Kalman filter are outlined in Sec. 5.3. In the simulations in Sec. 5.4, we examine the achievable trade-off between complexity and performance for the simplified variants of the ISCLP Kalman filter. The chapter is concluded in Sec. 5.5.

5.2 Review of the ISCLP Kalman Filter

Throughout the chapter, we use the following notation: vectors are denoted by lower-case boldface letters, matrices by upper-case boldface letters, $\mathbf{1}$ denotes a vector of ones, $\mathbf{A}^*$, $\mathbf{A}^T$, and $\mathbf{A}^H$, denote the complex conjugate, the transpose, and the complex conjugate transpose of a matrix $\mathbf{A}$. The operation diag[$\mathbf{A}$] creates a column vector from the diagonal elements of a square matrix $\mathbf{A}$, Diag[$\mathbf{a}$] creates a diagonal matrix with the elements of $\mathbf{a}$ on its diagonal, and and tr[$\mathbf{A}$] denotes the trace of $\mathbf{A}$. Subvectors are referenced by sets of indices,

i.e. the subvector of $\mathbf{a}$ composed of all its elements with indices in the ordered set S is denoted as $[\mathbf{a}]_{\in S}$. Similarly, masked vectors are denoted by $\langle \mathbf{a} \rangle_{\in S}$, with all except the elements in the ordered set S nullified, or by $\langle \mathbf{a} \rangle_i$, with all except the i^{th} element nullified. The operator $:=$ denotes a simplification, where the expression on the left-hand side is replaced by the computationally less expensive expression on the right-hand side.

In this section, we define the signal model in Sec. 5.2.1, and review the ISCLP Kalman filter in Sec. 5.2.2.

5.2.1 Signal Model

The ISCLP Kalman filter relies on the signal model in Sec. 3.2, which is not repeated here in all detail for the sake of conciseness. Instead, it shall suffice to note that the multi-microphone signal is represented in the STFT domain by the vector $\mathbf{y}(l) \in \mathbb{C}^M$ per frequency bin, where the frequency bin is not explicitly indexed, M denotes the number of microphones, and l the frame index. Here, $\mathbf{y}(l)$ contains N mutually uncorrelated reverberant speech components $\mathbf{x}_n(l)$ as well as the diffuse (e.g., babble) noise component $\mathbf{v}(l)$, i.e.

$$\mathbf{y}(l) = \sum_{n=1}^{N} \mathbf{x}_n(l) + \mathbf{v}(l). \tag{5.1}$$

Let $\mathbf{x}_n(l)$ be composed of the early and late reverberant components $\mathbf{x}_{n|e}(l) = \mathbf{h}_n(l)s_n(l)$ and $\mathbf{x}_{n|\ell}(l)$, respectively, i.e.

$$\mathbf{x}_n(l) = \mathbf{x}_{n|e}(l) + \mathbf{x}_{n|\ell}(l), \tag{5.2}$$

where

$$\mathbf{x}_{n|e}(l) = \mathbf{h}_n(l)s_n(l) \tag{5.3}$$

with $s_n(l)$ the early source images, whereof $N_T \leq N$ early source images are defined as target source images, and with $\mathbf{h}_n(l) \in \mathbb{C}^M$ the associated RETFs. Let $\mathbf{s}_T(l) \in \mathbb{C}^{N_T}$ and $\mathbf{H}_T(l) \in \mathbb{C}^{M \times N_T}$ respectively stack the early target source images and their RETFs, and let

$$\mathbf{x}_{e|T}(l) = \mathbf{H}_T(l)\mathbf{s}_T(l). \tag{5.4}$$

We seek to estimate the target component defined by

$$s_T(l) = \mathbf{1}^T\mathbf{s}_T(l). \tag{5.5}$$

We assume the reverberant speech components $\mathbf{x}_n(l)$ to be spatio-temporally correlated due to reverberation, but the early source images $s_n(l)$ to be on the

one hand temporally uncorrelated and on the other hand uncorrelated to the late reverberant component $\mathbf{x}_{n|\ell}(l)$ in frame l. The diffuse noise component $\mathbf{v}(l)$ is assumed spatially, but not temporally correlated. For a more detailed discussion of the signal model, we refer the interested reader to Sec. 3.2.

5.2.2 The ISCLP Kalman Filter

In order to estimate $s_T(l)$, the ISCLP Kalman filter embeds adaptive filtering in the ISCLP architecture, which integrates the generalized sidelobe canceller (GSC) and multi-channel linear prediction (MCLP). As a pre-requisite to adaptive filtering, the ISCLP architecture performs spatio-temporal pre-processing of $\mathbf{y}(l)$ by means of a matched filter (MF), a blocking matrix (BM), and a delay. Precisely, cf. Sec. 3.3.1, based on the L most recent frames of $\mathbf{y}(l)$, we define the spatio-temporally pre-processed signals $q(l)$ and $\mathbf{u}(l) \in \mathbb{C}^{LM-N_T}$ as

$$q(l) = \mathbf{g}^H(l)\mathbf{y}(l), \tag{5.6}$$

$$\mathbf{u}(l) = \begin{pmatrix} \mathbf{B}^H(l)\mathbf{y}(l) \\ \mathbf{y}(l-1) \\ \vdots \\ \mathbf{y}(l-L+1) \end{pmatrix}, \tag{5.7}$$

where the MF $\mathbf{g}(l) \in \mathbb{C}^M$ and the BM $\mathbf{B}(l) \in \mathbb{C}^{M \times M-N_T}$ are such that [32, 42]

$$\mathbf{g}^H(l)\mathbf{H}_T(l) = \mathbf{1}^T, \tag{5.8}$$

$$\mathbf{B}^H(l)\mathbf{H}_T(l) = \mathbf{0}. \tag{5.9}$$

In practice, the design of $\mathbf{g}(l)$ and $\mathbf{B}(l)$, cf. Sec. 3.3.1, requires an estimate of $\mathbf{H}_T(l)$, which we assume to be given, cf. Sec. 3.4.2 and Ch. 4. With $\mathbf{x}_{e|T}(l) = \mathbf{H}_T(l)\mathbf{s}_T(l)$ and (5.6)–(5.9), we note that $q(l)$ contains the target component $s_T(l)$ in the current frame l, while $\mathbf{u}(l)$ contains $\mathbf{H}_T(l-l')\mathbf{s}_T(l-l')$ for $l' > 0$, but not for $l' = 0$. Both $q(l)$ and $\mathbf{u}(l)$ further contain correlated (residual) nuisance components. Given $q(l)$ and $\mathbf{u}(l)$, we then define the enhanced signal $e(l) = \hat{s}_T(l)$ in the ISCLP architecture as

$$e(l) = \hat{s}_T(l) = q(l) - \hat{\mathbf{w}}^H(l)\mathbf{u}(l), \tag{5.10}$$

with $\hat{\mathbf{w}}(l) \in \mathbb{C}^{LM-N_T}$ the adaptive filter of L (multi-channel) coefficients. Based on the spatio-temporal structure of $\mathbf{u}(l)$ in (5.7), the adaptive filter $\hat{\mathbf{w}}(l)$ may be

split into the multiplicative sidelobe cancellation (SC) filter $\hat{\mathbf{w}}_{SC}(l) \in \mathbb{C}^{M-N_T}$ applied to $\mathbf{B}^H(l)\mathbf{y}(l)$, and the convolutive linear prediction (LP) filter $\hat{\mathbf{w}}_{LP}(l) \in \mathbb{C}^{(L-1)M}$ applied to $\mathbf{y}(l-l')$ for $l' > 0$. Note that due to the spatio-temporal preprocessing in (5.7) and the assumption that $s_n(l)$ is on the one hand temporally uncorrelated and on the other hand uncorrelated to $\mathbf{x}_{n|\ell}(l)$, cf. Sec. 5.2.1, we have $\mathrm{E}[\mathbf{u}(l)s_T^*(l)] = \mathbf{0}$, which allows for unconstrained, recursive adaptation of $\hat{\mathbf{w}}(l)$.

The adaptation of $\hat{\mathbf{w}}(l)$ is performed by means of the Kalman filter, cf. Sec. 3.3.2. To this end, we interpret $\hat{\mathbf{w}}(l)$ as the estimate of the true state $\mathbf{w}(l)$. The true state is defined by a state-space model comprising the process equation and measurement equation, respectively given by

$$\mathbf{w}(l) = \sqrt{\alpha}\mathbf{w}(l-1) + \mathbf{w}_\Delta(l), \tag{5.11}$$

$$q^*(l) = \mathbf{u}^H(l)\mathbf{w}(l) + s_T^*(l). \tag{5.12}$$

Here, the parameters of the process equation in (5.11) are considered subject to tuning, with $\alpha \in (0,1)$ a forgetting factor, and $\mathbf{w}_\Delta(l)$ the process noise with correlation matrix $\mathbf{\Psi}_{w_\Delta}(l)$. We define $\mathbf{\Psi}_{w_\Delta}(l) = (1-\alpha)\bar{\mathbf{\Psi}}_w$, where we recall that the diagonal matrix $\bar{\mathbf{\Psi}}_w$ is a rough guess of the presumed time-invariant correlation matrix $\mathbf{\Psi}_w$ of $\mathbf{w}(l)$. The measurement equation in (5.12) is defined by the premise that $e(l) = \hat{s}_T(l) = s_T(l)$ if $\hat{\mathbf{w}}(l) = \mathbf{w}(l)$, cf. also (5.10). In (5.12), $s_T^*(l)$ resembles the measurement noise with PSD $\varphi_{s_T}(l)$, whereof we assume an estimate to be given, cf. Sec. 3.4.2 and Ch. 4. In the Kalman filter, we distinguish the prior and posterior estimates $\hat{\mathbf{w}}(l)$ and $\hat{\mathbf{w}}^+(l)$ based on (5.11) and (5.12), respectively. With $\mathbf{\Psi}_{\tilde{w}}(l)$ and $\mathbf{\Psi}_{\tilde{w}}^+(l)$ denoting the correlation matrices of the corresponding state estimation errors $\tilde{\mathbf{w}}(l) = \hat{\mathbf{w}}(l) - \mathbf{w}(l)$ and $\tilde{\mathbf{w}}^+(l) = \hat{\mathbf{w}}^+(l) - \mathbf{w}(l)$, minimization of $\mathrm{tr}[\mathbf{\Psi}_{\tilde{w}}(l)]$ and $\mathrm{tr}[\mathbf{\Psi}_{\tilde{w}}^+(l)]$ leads to the Kalman filter update equations [51, 52]

$$\hat{\mathbf{w}}(l) = \sqrt{\alpha}\hat{\mathbf{w}}^+(l-1), \tag{5.13}$$

$$\mathbf{\Psi}_{\tilde{w}}(l) = \alpha\mathbf{\Psi}_{\tilde{w}}^+(l-1) + (1-\alpha)\bar{\mathbf{\Psi}}_w, \tag{5.14}$$

$$e^*(l) = q^*(l) - \mathbf{u}^H(l)\hat{\mathbf{w}}(l), \tag{5.15}$$

$$\varphi_e(l) = \mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l)\mathbf{u}(l) + \varphi_{s_T}(l), \tag{5.16}$$

$$\mathbf{k}(l) = \mathbf{\Psi}_{\tilde{w}}(l)\mathbf{u}(l)\varphi_e^{-1}(l), \tag{5.17}$$

$$\hat{\mathbf{w}}^+(l) = \hat{\mathbf{w}}(l) + \mathbf{k}(l)e^*(l), \tag{5.18}$$

$$\boxed{\mathbf{\Psi}_{\tilde{w}}^+(l) = \mathbf{\Psi}_{\tilde{w}}(l) - \mathbf{k}(l)\mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l),} \tag{5.19}$$

where (5.13)–(5.14) and (5.18)–(5.19) are referred to as the time and the measurement update[1], respectively, and (5.15) is equivalent to (5.10). With $\hat{\mathbf{w}}(l)$ being a prior estimate of $\mathbf{w}(l)$, we may consider $e(l) = \hat{s}_T(l)$ in (5.15) a prior estimate of $\hat{s}_T(l)$. A posterior-like estimate $e^+(l) = \hat{s}_T^+(l)$ can be defined, cf. Sec. 3.3.3, as a spectrally post-processed version of $e(l)$ by

$$\gamma(l) = \max\left[\frac{\varphi_{s_T}(l)}{\varphi_e(l)}, \beta\gamma(l-1)\right], \tag{5.20}$$

$$e^+(l) = \hat{s}_T^+(l) = \gamma(l)e(l), \tag{5.21}$$

with $\beta \in [0, 1]$ limiting the gain decay.

The update equations (5.13)–(5.19) are initialized by $\hat{\mathbf{w}}(0) = 0$ and $\mathbf{\Psi}_{\tilde{w}}(0) = \bar{\mathbf{\Psi}}_w$, where we recall that the diagonal matrix $\bar{\mathbf{\Psi}}_w$ is a rough guess of the presumed time-invariant correlation matrix $\mathbf{\Psi}_w$ of $\mathbf{w}(l)$. In this respect, it should be noted that the ISCLP Kalman filter essentially contains the MCLP Kalman filter in [102] as a special case, which is obtained if the SC filter component $\hat{\mathbf{w}}_{SC}^{(+)}(l)$ of $\hat{\mathbf{w}}^{(+)}(l)$ remains zero at all times. With $\bar{\boldsymbol{\psi}}_w = \mathrm{diag}[\bar{\mathbf{\Psi}}_w]$ and $\bar{\boldsymbol{\psi}}_{w_{SC}} \in \mathbb{R}^{M-N_T}$ denoting the subvector of $\bar{\boldsymbol{\psi}}_w$ associated to the SC filter, this may be ensured by setting $\bar{\boldsymbol{\psi}}_{w_{SC}} = \mathbf{0}$, which corresponds to assuming a purely reverberant scenario.

We complete this review by stressing that in a practical implementation, the performance of the ISCLP Kalman filter is subject to modeling errors due to deficiencies in the microphone signal model, cf. Sec. 3.2 and Sec. 3.4.1, and the ISCLP state-space model, cf. Sec. 3.3.2 and Sec. 3.4.3, as well as parameter estimation errors, i.e. errors in the estimates of $\mathbf{H}_T(l)$ and $\varphi_{s_T}(l)$, cf. Sec. 3.4.2 and Ch. 4. For further details on the motivation, the derivation, and the implementation of the ISCLP Kalman filter, we refer the interested reader to Sec. 3.3 and Sec. 3.4.

5.3 Complexity Reduction

Given the ISCLP Kalman filter update equations in (5.13)–(5.19), we strive to reduce their computational complexity by introducing reasonable simplifications. We consider computational complexity in terms of multiplications only, disregarding the inversion of the scalar $\varphi_e(l)$ in (5.17). In what follows, we

[1]Note that with (5.17), the update term $\mathbf{k}(l)\mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l)$ in (5.19) may also be written as $\mathbf{k}(l)\mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l) = \mathbf{k}(l)\mathbf{k}^H(l)\varphi_e(l)$ and hence is Hermitian, as expected for a correlation matrix. If the Kalman filter is implemented in finite-precision arithmetic, however, the numerically better-behaved but computationally more complex so-called Joseph form [51,52] of the update term may be preferred over $\mathbf{k}(l)\mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l)$. Here, we assume infinite precision and therefore base the following discussion on (5.19), but note that it similarly applies to the Joseph form.

make use of the operator $\cdot$ in order to explicitly denote products of interest. We discuss the complexity of the original ISCLP Kalman update equations (5.13)–(5.19) in Sec. 5.3.1, define three kinds of simplifications in Sec. 5.3.2, and present an equivalent multiple Kalman filters formulation of the simplified ISCLP Kalman filters in Sec. 5.3.3.

5.3.1 Complexity of the Original ISCLP Kalman Filter

We may distinguish products with computational costs of $\mathcal{O}(LM)$ and $\mathcal{O}(L^2M^2)$ multiplications, where the latter obviously dominate the total cost. Products of $\mathcal{O}(LM)$ cost are

- $\sqrt{\alpha} \cdot \hat{\mathbf{w}}^+(l-1)$ in (5.13),
- $\mathbf{u}^H(l) \cdot \hat{\mathbf{w}}(l)$ in (5.15),
- $\mathbf{u}^H(l) \cdot \mathbf{\Psi}_{\tilde{w}}(l)\mathbf{u}(l)$ in (5.16), where $\mathbf{\Psi}_{\tilde{w}}(l)\mathbf{u}(l)$ is a vector,
- $\mathbf{\Psi}_{\tilde{w}}(l)\mathbf{u}(l) \cdot \varphi_e^{-1}(l)$ in (5.17),
- and $\mathbf{k}(l) \cdot e^*(l)$ in (5.18).

Products of $\mathcal{O}(L^2M^2)$ cost are

- $\alpha \cdot \mathbf{\Psi}_{\tilde{w}}^+(l-1)$ in (5.14),
- $\mathbf{\Psi}_{\tilde{w}}(l) \cdot \mathbf{u}(l)$ in (5.16), which also appears in (5.17) and (5.19) but is assumed to be computed only once,
- and $\mathbf{k}(l) \cdot \mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l)$ in (5.19).

The exact number of required multiplications can be found in Table 5.1 for case (a). In this table, products are distinguished by the input domains of their factors, i.e. $\mathbb{R} \times \mathbb{R}$, $\mathbb{R} \times \mathbb{C}$, and $\mathbb{C} \times \mathbb{C}$, which may result in different usage of million instructions per second (MIPS) depending on the employed soft- and hardware architecture.

Table 5.1: Complexity in terms of required multiplications with input domains $\mathbb{R} \times \mathbb{R}$, $\mathbb{R} \times \mathbb{C}$, and $\mathbb{C} \times \mathbb{C}$ of the Kalman filter update equations for (a) preserved cross-correlations, (b) omitted temporal cross-correlations, (c) omitted spatial cross-correlations, and (d) omitted spatio-temporal cross-correlations. The simplified variant of (5.19) using, e.g., (5.24) is denoted by (5.19|5.24), and similarly for the other simplified variants.

Eq. \ *Domain*		$\mathbb{R} \times \mathbb{R}$	$\mathbb{R} \times \mathbb{C}$	$\mathbb{C} \times \mathbb{C}$
(5.13)	(a)–(d)	0	$LM - N_T$	0
(5.14)	(a)	$LM - N_T$	$\frac{1}{2}L^2M^2 - (\frac{1}{2} + N_T)\, LM + \frac{1}{2}N_T^2 + \frac{1}{2}N_T$	0
	(b)		$\frac{1}{2}LM^2 - (\frac{1}{2}L + N_T)\, M + \frac{1}{2}N_T^2 + \frac{1}{2}N_T$	
	(c)		$\frac{1}{2}L^2M - L\,(\frac{1}{2}M + N_T) + N_T$	
	(d)		0	
(5.15)	(a)–(d)	0	0	$LM - N_T$
(5.16)	(a)	0	$LM - N_T$	$L^2M^2 - 2N_TLM + N_T^2$
	(b)			$LM^2 - 2N_TM + N_T^2$
	(c)			$L^2M - 2N_TL + N_T$
	(d)			$LM - N_T$
(5.17)	(a)–(d)	0	$LM - N_T$	0
(5.18)	(a)–(d)	0	0	$LM - N_T$
(5.19)	(a)	0	0	$\frac{1}{2}L^2M^2 + (\frac{1}{2} - N_T)LM + \frac{1}{2}N_T^2 - \frac{1}{2}N_T$
(5.19|5.24)	(b)			$\frac{1}{2}LM^2 + (\frac{1}{2}L - N_T)\, M + \frac{1}{2}N_T^2 - \frac{1}{2}N_T$
(5.19|5.25)	(c)			$\frac{1}{2}L^2M - L\,(\frac{1}{2}M - N_T)$
(5.19|5.26)	(d)			$LM - N_T$

$\sum$	(a)	$LM - N_T$	$1/2L^2M^2 + (3/2 - N_T)\,LM + 1/2N_T^2 - 3/2N_T$	$3/2L^2M^2 + (5/2 - 3N_T)LM + 3/2N_T^2 - 5/2N_T$
	(b)		$1/2LM^2 + (3/2L - N_T)\,M + 1/2N_T^2 - 3/2N_T$	$3/2LM^2 + (5/2L - 3N_T)M + 3/2N_T^2 - 5/2N_T$
	(c)		$1/2L^2M + L\,(3/2M - N_T) - N_T$	$3/2L^2M + L(5/2M - 3N_T) - N_T$
	(d)		$2LM - 2N_T$	$4LM - 4N_T$

5.3.2 Complexity Reduction by Cross-Correlation Negligence

In order to reduce the computational complexity of the Kalman filter, we may simplify the products of $\mathcal{O}(L^2M^2)$ cost, namely $\alpha \cdot \mathbf{\Psi}^{+}_{\tilde{w}}(l-1)$ in (5.14), $\mathbf{\Psi}_{\tilde{w}}(l) \cdot \mathbf{u}(l)$ in (5.16), and $\mathbf{k}(l) \cdot \mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l)$ in (5.19). At this point, we note that the cost of the products $\alpha \cdot \mathbf{\Psi}^{+}_{\tilde{w}}(l-1)$ and $\mathbf{\Psi}_{\tilde{w}}(l) \cdot \mathbf{u}(l)$ directly depends on the number of non-zero elements in $\mathbf{\Psi}^{+}_{\tilde{w}}(l-1)$ and $\mathbf{\Psi}_{\tilde{w}}(l)$, which in general are full matrices and therefore have $\mathcal{O}(L^2M^2)$ non-zero elements. This may easily be verified by noting that the update term $\mathbf{k}(l) \cdot \mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l)$ in (5.19) in fact is an outer vector product, and therefore necessarily is a full matrix of rank one. Hence, in order to simplify the Kalman filter update equations, we propose to simplify the update term $\mathbf{k}(l)\mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l)$, i.e. to perform sparse updates instead and thereby keep $\mathbf{\Psi}^{(+)}_{\tilde{w}}(l)$ sparse.

Particularly, we define the sparse update based on the assumption that some of the off-diagonal elements in $\mathbf{\Psi}^{(+)}_{\tilde{w}}(l)$, which represent cross-correlations of the state estimation error $\tilde{\mathbf{w}}^{+}(l)$, are negligible[2] and may hence be omitted. More specifically, as in the MCLP Kalman filter in [102], we may assume that, e.g., the temporal cross-correlations are negligible, yielding $\mathcal{O}(LM^2)$ non-zero elements in $\mathbf{\Psi}^{(+)}_{\tilde{w}}(l)$. Note that while the MCLP [102] and the ISCLP Kalman filter operate in the (weighted overlap-add-based) STFT domain, this assumption directly corresponds to an assumption previously used in (overlap-save-based) PBFD Kalman filtering [98, 190–192], namely that the state estimation errors in different filter partitions are mutually uncorrelated. Alternatively, we may assume that the spatial cross-correlations are negligible, or even that all cross-correlations are negligible, yielding $\mathcal{O}(L^2M)$ and $\mathcal{O}(LM)$ non-zero elements in $\mathbf{\Psi}^{(+)}_{\tilde{w}}(l)$, respectively. The corresponding sparse structures of $\mathbf{\Psi}^{(+)}_{\tilde{w}}(l)$ are shown in Fig. 5.1 (b)–(d), next to the full structure in Fig. 5.1 (a).

Given the desired sparse structure of $\mathbf{\Psi}^{(+)}_{\tilde{w}}(l)$, we impose the same sparse structure on the update term, i.e. the simplification of $\mathbf{k}(l)\mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l)$, and thereby simplify the ISCLP Kalman filter update equations. With $i = 1, \ldots, LM - N_T$, we define the ordered sets

$$\Theta_{l'} = \left\{ i : \left\lceil \frac{i + N_T}{M} \right\rceil = l' \right\}, \tag{5.22}$$

$$\Xi_m = \left\{ i : M \left\lfloor \frac{i + N_T - m}{M} \right\rfloor = i + N_T - m \right\}, \tag{5.23}$$

[2]Here, negligible may be interpreted in a strict sense, i.e. we may assume that the cross-correlations are of negligibly low level, which is the case if the recursive average of the corresponding elements in $\mathbf{k}(l)\mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l)$ is negligible, cf. (5.14), or in a wide sense, i.e. we may assume that their impact on the convergence behavior of the ISCLP Kalman filter is negligible.

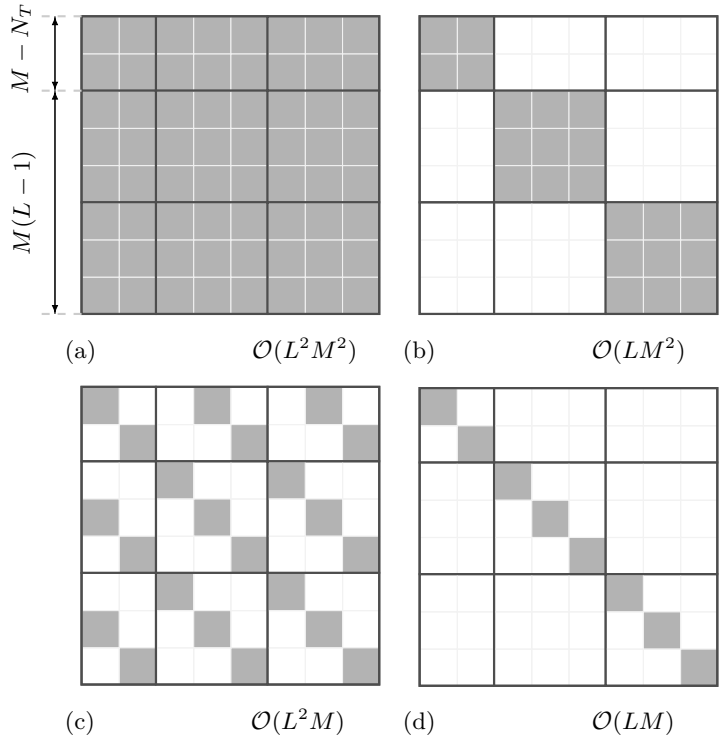


Figure 5.1: The (sparse) structure of $\mathbf{\Psi}_{\tilde{w}}(l)$ and the order of non-zero elements [■] for (a) preserved cross-correlations, (b) neglected temporal cross-correlations, (c) neglected spatial cross-correlations, (d) all cross-correlations neglected, exemplarily shown for $N_T = 1$, $M = 3$, and $L = 3$.

where $\Theta_{l'}$ and Ξ_m collect the indices of vector $\mathbf{w}(l)$ associated to the (multi-channel) filter coeffient l' and to channel m, respectively. Given these sets, the

simplified update terms may be written as

$$\mathbf{k}(l)\mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l) := \sum_{l'=1}^{L} \langle \mathbf{k}(l) \rangle_{\in \Theta_{l'}} \langle \mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l) \rangle_{\in \Theta_{l'}}, \tag{5.24}$$

$$\mathbf{k}(l)\mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l) := \sum_{m=1}^{M} \langle \mathbf{k}(l) \rangle_{\in \Xi_{m}} \langle \mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l) \rangle_{\in \Xi_{m}}, \tag{5.25}$$

$$\begin{aligned} \mathbf{k}(l)\mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l) &:= \sum_{i=1}^{LM-N_T} \langle \mathbf{k}(l) \rangle_i \langle \mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l) \rangle_i \\ &= \mathrm{Diag}[\mathbf{k}(l)]\,\mathrm{Diag}[\mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l)], \end{aligned} \tag{5.26}$$

where (5.24), (5.25), and (5.26) correspond to neglecting temporal cross-correlations, neglecting spatial cross-correlations, and neglecting all cross-correlations, respectively, and the update terms are of rank L, rank M, and full rank, yielding total computational costs of $\mathcal{O}(LM^2)$, $\mathcal{O}(L^2M)$, and $\mathcal{O}(LM)$. In the latter case, i.e. the least complex variant, the computational cost is of the same order, but not exactly the same,[3] as for (propotionate) normalized least mean squares ((P)NLMS) algorithms [50, 51, 193, 194]. The exact number of required multiplications for the simplified Kalman filter update equations can be found in Table 5.1 for case b)–d). In the following, we refer to the different complexity variants by their cost, e.g., we refer to (5.13)–(5.19) with simplification (5.24) as the $\mathcal{O}(LM^2)$-cost ISCLP Kalman filter, and similarly for the other variants.

In practice, the assumptions underlying the presented simplifications of the ISCLP Kalman filter, namely that temporal, spatial, or all cross-correlations are negligible, may be well or less well justified, leading to some degree of suboptimality. The simplifications hence imply a trade-off between complexity and performance, which we investigate in the simulations in Sec. 5.4.

5.3.3 Equivalent Multiple Kalman Filters Formulation

The three simplified ISCLP Kalman filter variants may equivalently be represented by multiple Kalman filters sharing the same error signal $e(l)$ and

[3] NLMS differs from the Kalman filter in that it does not consider a state evolution, which corresponds to $\alpha = 1$ in (5.13)–(5.14), and that it applies a scalar gain to $\mathbf{u}(l)e^*(l)$ in the filter update [50, 51, 193] instead of the gain matrix $\mathbf{\Psi}_{\tilde{w}}(l)\varphi_e^{-1}(l)$, cf. (5.17)–(5.18). A thorough disussion on the relations between NLMS and the Kalman filter can be found in [193]. In PNLMS instead, similar to the $\mathcal{O}(LM)$-variant of the ISCLP Kalman filter, a diagonal gain matrix is applied [194], albeit computed in a somewhat different manner.

error signal PSD $\varphi_e(l)$, as outlined in this section. Specifically, consider, e.g., the $\mathcal{O}(LM^2)$-cost ISCLP Kalman filter. For ease of presentation, let $\mathbf{u}_{l'}(l) = [\mathbf{u}(l)]_{\in\Theta_{l'}}$ denote the non-zero subvector of $\langle \mathbf{u}(l)\rangle_{\in\Theta_{l'}}$, and let $\hat{\mathbf{w}}_{l'}^{(+)}(l)$, $\mathbf{k}_{l'}(l)$, $\mathbf{\Psi}_{\tilde{w}|l'}^{(+)}(l)$, and $\bar{\mathbf{\Psi}}_{w|l'}$ be similarly defined. Then, due to the block-diagonal structure of $\mathbf{\Psi}_{\tilde{w}}^{(+)}(l)$ as shown in Fig. 5.1 (b) with the submatrices $\mathbf{\Psi}_{\tilde{w}|l'}^{(+)}(l)$ on the diagonal, the update equations in (5.13)–(5.18) are equivalent to

$$\hat{\mathbf{w}}_{l'}(l) = \sqrt{\alpha}\hat{\mathbf{w}}_{l'}^{+}(l-1), \tag{5.27}$$

$$\mathbf{\Psi}_{\tilde{w}|l'}(l) = \alpha\mathbf{\Psi}_{\tilde{w}|l'}^{+}(l-1) + (1-\alpha)\bar{\mathbf{\Psi}}_{w|l'}, \tag{5.28}$$

$$e^*(l) = q^*(l) - \sum_{l'=1}^{L} \mathbf{u}_{l'}^H(l)\hat{\mathbf{w}}_{l'}(l), \tag{5.29}$$

$$\varphi_e(l) = \sum_{l'=1}^{L} \mathbf{u}_{l'}^H(l)\mathbf{\Psi}_{\tilde{w}|l'}(l)\mathbf{u}_{l'}(l) + \varphi_{s_T}(l), \tag{5.30}$$

$$\mathbf{k}_{l'}(l) = \mathbf{\Psi}_{\tilde{w}|l'}(l)\mathbf{u}_{l'}(l)\varphi_e^{-1}(l), \tag{5.31}$$

$$\hat{\mathbf{w}}_{l'}^{+}(l) = \hat{\mathbf{w}}_{l'}(l) + \mathbf{k}_{l'}(l)e^*(l), \tag{5.32}$$

$$\mathbf{\Psi}_{\tilde{w}|l'}^{+}(l) = \mathbf{\Psi}_{\tilde{w}|l'}(l) - \mathbf{k}_{l'}(l)\mathbf{u}_{l'}^H(l)\mathbf{\Psi}_{\tilde{w}|l'}(l), \tag{5.33}$$

where obviously (5.27)–(5.28) need to be computed for all l' before proceeding with (5.29)–(5.32). Similar equivalences hold for the $\mathcal{O}(L^2M)$-cost and the $\mathcal{O}(LM)$-cost ISCLP Kalman filter.

5.4 Simulations

In this section, we investigate the impact of the proposed simplifications, i.e. we examine the trade-off between complexity and performance for the $\mathcal{O}(LM^2)$-cost, the $\mathcal{O}(L^2M)$-cost and the $\mathcal{O}(LM)$-cost ISCLP Kalman filter, where the original $\mathcal{O}(LM^2)$-cost ISCLP Kalman filter serves as a reference. Note again that next to these simplifications, also modeling deficiencies and parameter estimation errors are sources of suboptimality equally applying to all complexity variants of the ISCLP Kalman filter, cf. Sec. 5.2.2. We discuss the simulation setup in Sec. 5.4.1, and the results in Sec. 5.4.2. Audio examples are available at [154].

5.4.1 Setup

In order to compare the different complexity variants, we adopt the simulation setup in Sec. 3.5 designed to benchmark the original $\mathcal{O}(L^2M^2)$-cost ISCLP Kalman filter, which is not repeated here in detail for the sake of conciseness. Instead, we note that we use the same performance measures, the same acoustic scenarios, and the same algorithmic settings, cf. Sec. 3.5.2, Sec. 3.5.3, and Sec. 3.5.4, respectively. We recall that two acoustic scenarios were defined, referred to as case A and case B. In case A, we consider one reverberant speech and a babble noise component, $\mathbf{x}_1(l)$ and $\mathbf{v}(l)$, with $\mathbf{x}_1(l)$ containing the target component $\mathbf{x}_{e|T}(l) = \mathbf{x}_{1|e}(l)$. Here, as opposed to Sec. 3.5.3, we now also explicitly consider the noiseless scenario with $\mathbf{v}(l) = \mathbf{0}$ as a special case, posing a mere dereverberation task. In that, we set $\bar{\psi}_{w_{SC}} = \mathbf{0}$ such that the ISCLP filter corresponds to the MCLP Kalman filter in [102]. In case B, we consider two reverberant speech components and a babble noise component, $\mathbf{x}_1(l)$, $\mathbf{x}_2(l)$, and $\mathbf{v}(l)$, with $\mathbf{x}_1(l)$ again containing the target component $\mathbf{x}_{e|T}(l) = \mathbf{x}_{1|e}(l)$, and $\mathbf{x}_2(l)$ an interfering speech component to be canceled. For further details on the setup, the interested reader is referred to Sec. 3.5.2 to Sec. 3.5.4.

5.4.2 Results

We present the results in case A and case B in Sec. 5.4.2.1 and Sec. 5.4.2.2, respectively, followed by a summary of the simulation results in Sec. 5.4.2.3.

5.4.2.1 Case A: Without Interfering Speech

In case A, interfering speech is absent.

Fig. 5.2 shows the performance in terms of (a) *PESQ*, (b) *STOI*, (c) SIR^{fws}, and (d) *CD* versus *SNR* for the reference microphone signal [·····], the $\mathcal{O}(L^2M^2)$-cost [—●—], the $\mathcal{O}(LM^2)$-cost [- ■ -], the $\mathcal{O}(L^2M)$-cost [- ♦ -], and the $\mathcal{O}(LM)$-cost [- ▼ -] ISCLP Kalman filter with $L = 6$. In this and the following figures, the graphs denote medians over all individual simulations, cf. Sec. 3.5.3, and the shaded areas indicate the range from the first to the third quartile. As previously shown for the same acoustic scenario in Fig. 3.3, the reference microphone signal expectably reaches better scores at higher *SNR* values, compared to which the $\mathcal{O}(L^2M^2)$-cost ISCLP Kalman filter yields significant improvements above roughly $SNR = -5\,\text{dB}$. In all measures, only small differences can be found across the different complexity variants of the ISCLP Kalman filter, with largest differences observable in *PESQ*, and smallest differences in *CD*. This indicates a generally good trade-off between complexity and performance for

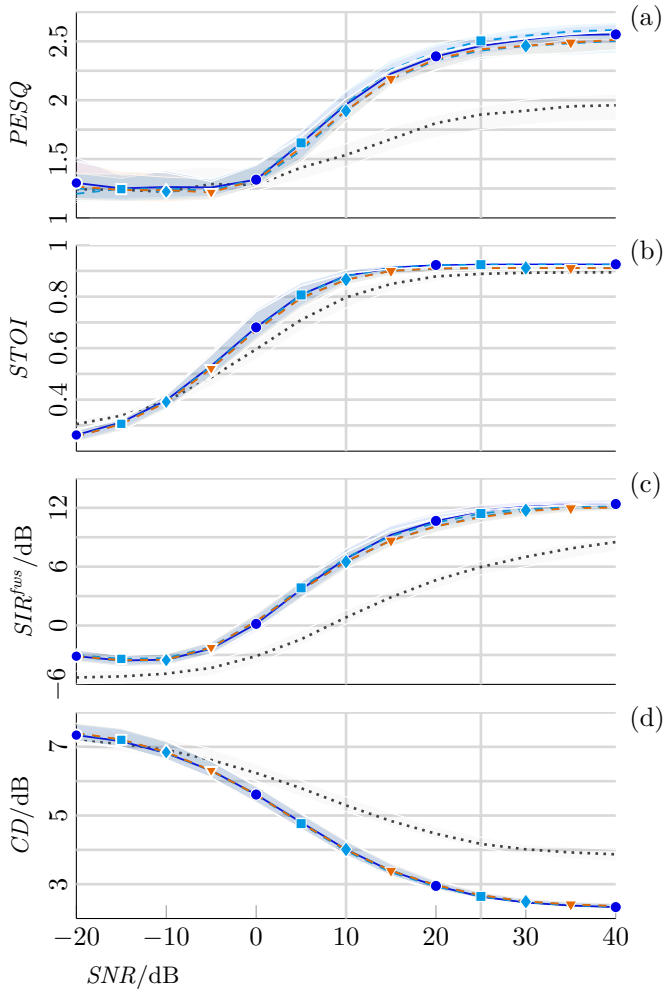


Figure 5.2: (a) *PESQ*, (b) *STOI*, (c) SIR^{fws}, and (d) *CD* versus *SNR* for the reference microphone signal [·····], the $\mathcal{O}(L^2M^2)$-cost [–●–], the $\mathcal{O}(LM^2)$-cost [- ■ -], the $\mathcal{O}(L^2M)$-cost [- ♦ -], and the $\mathcal{O}(LM)$-cost [- ▼ -] ISCLP Kalman filter with $L = 6$ if interfering speech is absent.

all three simplified variants, i.e. the $\mathcal{O}(LM^2)$-cost, the $\mathcal{O}(L^2M)$-cost, and the $\mathcal{O}(LM)$-cost variant.

Fig. 5.3 depicts the performance improvement in terms of (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus L with respect to the reference microphone signal for the $\mathcal{O}(L^2M^2)$-cost [–●–], the $\mathcal{O}(LM^2)$-cost [- ■ -], the

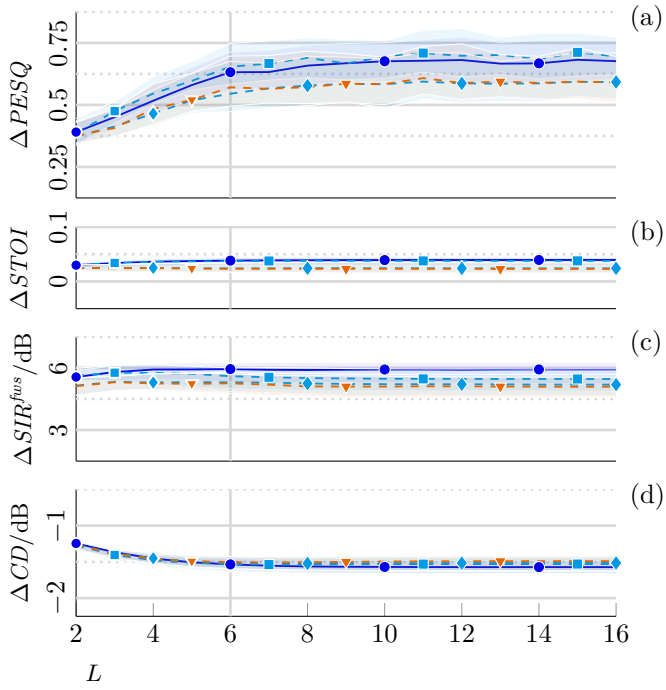


Figure 5.3: (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus L with respect to the reference microphone signal for the $\mathcal{O}(L^2M^2)$-cost [–●–], the $\mathcal{O}(LM^2)$-cost [- ■ -], the $\mathcal{O}(L^2M)$-cost [- ♦ -], and the $\mathcal{O}(LM)$-cost [- ▼ -] ISCLP Kalman filter at $SNR = 25\,\text{dB}$ if interfering speech is absent.

$\mathcal{O}(L^2M)$-cost [- ♦ -], and the $\mathcal{O}(LM)$-cost [- ▼ -] ISCLP Kalman filter at $SNR = 25\,\text{dB}$. Note that in Fig. 5.3 and in the following figures presenting performance improvements, the resolution of the vertical axes is twice as large as in Fig. 5.2. As previously shown for the same acoustic scenario in Fig. 3.4, the improvement for the $\mathcal{O}(L^2M^2)$-cost ISCLP Kalman filter saturates at roughly $L = 6$. Due to the higher resolution of the vertical axes, differences across the simplified variants, i.e. the $\mathcal{O}(LM^2)$-cost, the $\mathcal{O}(L^2M)$-cost, and the $\mathcal{O}(LM)$-cost ISCLP Kalman filter, are now somewhat better visible. We note that among these three, the $\mathcal{O}(LM^2)$-cost variant generally performs best and, except for ΔSIR^{fws}, equally well as compared to the $\mathcal{O}(L^2M^2)$-cost variant. Differences between the the $\mathcal{O}(L^2M)$-cost and the $\mathcal{O}(LM)$-cost variants are hardly visible.

Fig. 5.4 shows the performance improvement if the experiment of Fig. 5.3 is repeated for the noiseless case (i.e. at $SNR = \infty\,\text{dB}$) with $\bar{\boldsymbol{\psi}}_{w_{SC}} = \mathbf{0}$, such that the ISCLP Kalman filter corresponds to the MCLP Kalman filter in [102]. As compared to Fig. 5.3, for $L \geq 6$, we find the same improvement in $PESQ$ and

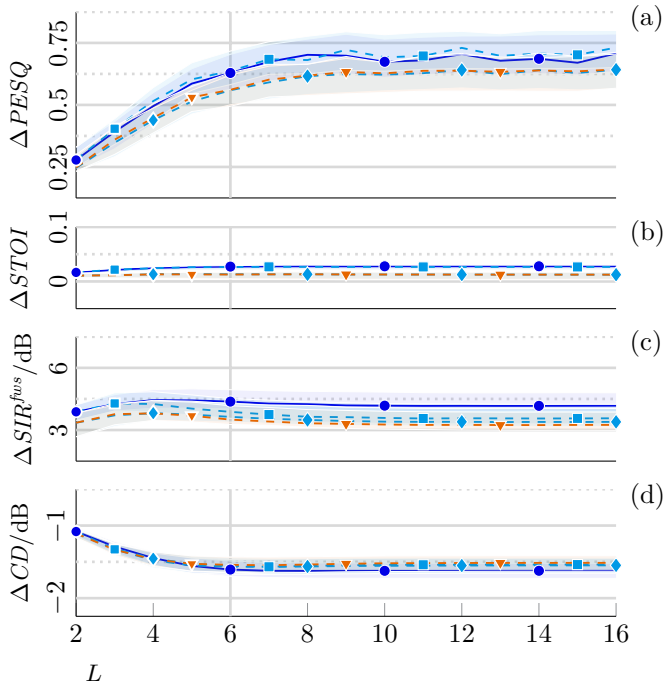


Figure 5.4: (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus L with respect to the reference microphone signal for the $\mathcal{O}(L^2M^2)$-cost [–●–], the $\mathcal{O}(LM^2)$-cost [- ■ -], the $\mathcal{O}(L^2M)$-cost [- ♦ -], and the $\mathcal{O}(LM)$-cost [- ▾ -] ISCLP Kalman filter with $\bar{\boldsymbol{\psi}}_{w_{SC}} = \mathbf{0}$ if noise and interfering speech are absent. The reference microphone signal scores at $PESQ = 1.97$, $STOI = 0.89$, $SIR^{fws} = 9.36\,\text{dB}$, and $CD = 3.82\,\text{dB}$.

CD, but less improvement in $STOI$ and SIR^{fws}, indicating less sensitivity of the latter two towards reverberation than towards noise. Again compared to Fig. 5.3, below $L = 6$, we further find a larger dependency on L in terms of $\Delta PESQ$ and ΔCD. This is expected since the babble noise component $\mathbf{v}(l)$, which is temporally uncorrelated and may therefore not be suppressed by the LP filter, has stronger impact on the improvement scores at lower SNR values. No substantial difference between the two figures can be found in terms of the relative performance of the different complexity variants of the ISCLP Kalman filter.

Fig. 5.5 depicts the performance improvement in terms of (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus time t with respect to the reference microphone signal for the $\mathcal{O}(L^2M^2)$-cost [–●–], the $\mathcal{O}(LM^2)$-cost [- ■ -], the $\mathcal{O}(L^2M)$-cost [- ♦ -], and the $\mathcal{O}(LM)$-cost [- ▾ -] ISCLP Kalman filter at $SNR =$

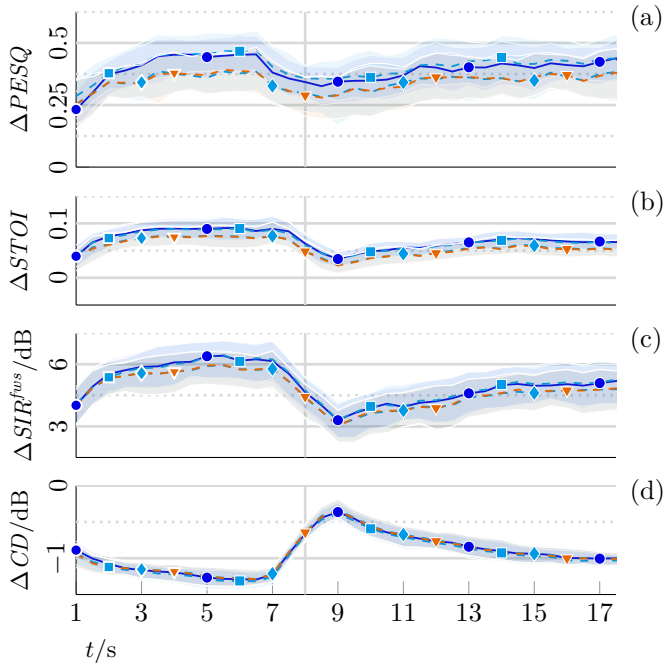


Figure 5.5: (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus t with respect to the reference microphone signal for the $\mathcal{O}(L^2M^2)$-cost [–●–], the $\mathcal{O}(LM^2)$-cost [- ■ -], the $\mathcal{O}(L^2M)$-cost [- ♦ -], and the $\mathcal{O}(LM)$-cost [- ▼ -] ISCLP Kalman filter with $L = 6$ at $SNR = 10\,\text{dB}$ if interfering speech is absent.

25 dB. We find that after initialization, all complexity variants of the ISCLP Kalman filter converge at the same rate, reaching saturation after roughly 4 s. The speech source position changes at 8 s, cf. Sec. 3.5.3, such that the Kalman filters have to re-adapt. After the speech source position change, convergence speed is somewhat reduced as compared to the initial convergence stage.

Fig. 5.6 shows the the performance improvement if the experiment of Fig. 5.5 is repeated for the noiseless case (i.e. at $SNR = \infty\,\text{dB}$) with $\bar{\boldsymbol{\psi}}_{w_{SC}} = \mathbf{0}$, such that the ISCLP Kalman filter corresponds to the MCLP Kalman filter in [102]. We note that as compared to Fig. 5.5, as expected in the absence of noise, all complexity variants of the ISCLP Kalman filter converge faster, now reaching saturation after roughly 2 s. Again, the convergence speed after the speech source position change is somewhat reduced as compared to the initial convergence stage. In terms of the relative performance of the different complexity variants of the ISCLP Kalman filter, no substantial difference can be found between the two figures.

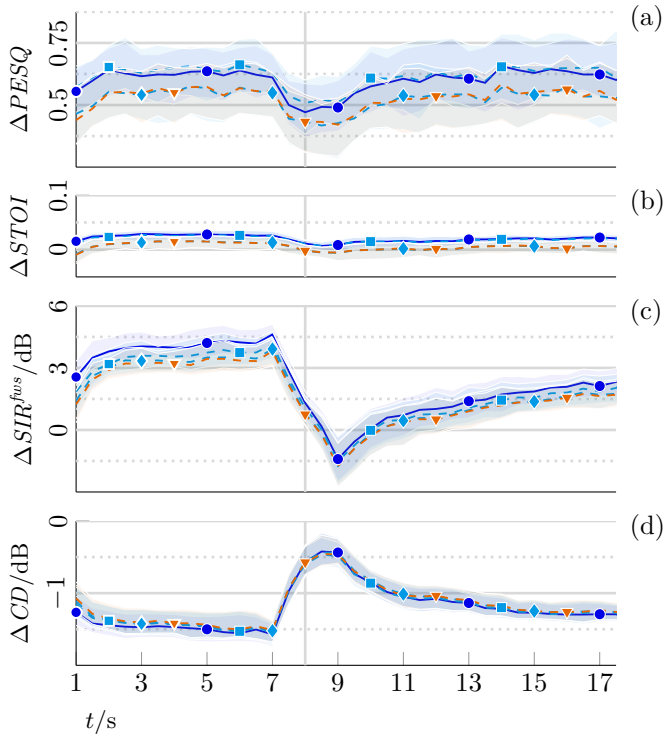


Figure 5.6: (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus t with respect to the reference microphone signal for the $\mathcal{O}(L^2M^2)$-cost [—•—], the $\mathcal{O}(LM^2)$-cost [- ▪ -], the $\mathcal{O}(L^2M)$-cost [- ♦ -], and the $\mathcal{O}(LM)$-cost [- ▾ -] ISCLP Kalman filter with $L = 6$ and $\bar{\psi}_{w_{SC}} = \mathbf{0}$ if noise and interfering speech are absent. The reference microphone signal scores at $PESQ = 1.99$, $STOI = 0.91$, $SIR^{fws} = 10.2\,\text{dB}$, and $CD = 3.62\,\text{dB}$.

5.4.2.2 Case B: With Interfering Speech

In case B, interfering speech is present.

Fig. 5.7 shows the performance in terms of (a) $PESQ$, (b) $STOI$, (c) SIR^{fws}, and (d) CD versus SNR for the reference microphone signal [·····], the $\mathcal{O}(L^2M^2)$-cost [—•—], the $\mathcal{O}(LM^2)$-cost [- ▪ -], the $\mathcal{O}(L^2M)$-cost [- ♦ -], and the $\mathcal{O}(LM)$-cost [- ▾ -] ISCLP Kalman filter with $L = 6$. As shown previously for the same acoustic scenarios, cf. Sec. 3.5.5, the curves are generally flatter as compared to those in Fig. 5.2 due to the now additional interfering speech component $\mathbf{x}_2(l)$. Differences across the different complexity variants of the ISCLP Kalman filter are now even less pronounced, which may relate to the generally worse

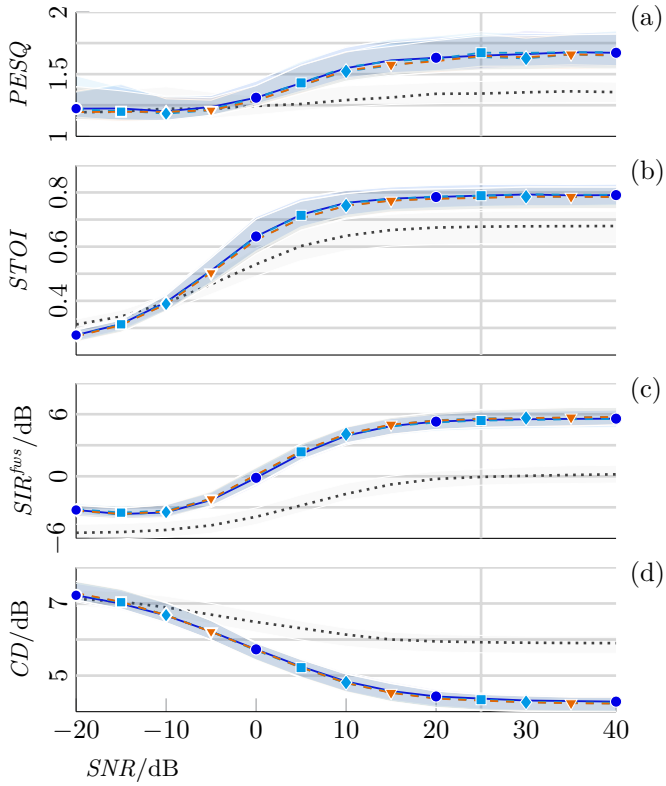


Figure 5.7: (a) *PESQ*, (b) *STOI*, (c) SIR^{fws}, and (d) *CD* versus *SNR* for the reference microphone signal [·····], the $\mathcal{O}(L^2M^2)$-cost [—•—], the $\mathcal{O}(LM^2)$-cost [- ▪ -], the $\mathcal{O}(L^2M)$-cost [- ♦ -], and the $\mathcal{O}(LM)$-cost [- ▾ -] ISCLP Kalman filter with $L = 6$ if interfering speech is present.

scores due to the more challenging acoustic scenario.

Fig. 5.8 depicts the performance improvement in terms of (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus L with respect to the reference microphone signal for the $\mathcal{O}(L^2M^2)$-cost [—•—], the $\mathcal{O}(LM^2)$-cost [- ▪ -], the $\mathcal{O}(L^2M)$-cost [- ♦ -], and the $\mathcal{O}(LM)$-cost [- ▾ -] ISCLP Kalman filter at $SNR = 25\,\text{dB}$. For the ISCLP Kalman filter, As previously shown for the same acoustic scenarios, cf. Sec. 3.5.5, as compared to the case $\mathbf{x}_2(l) = \mathbf{0}$, cf. Fig. 5.8, the improvements saturate somewhat later. Again, differences across the different complexity variants of the ISCLP Kalman filter are now even less pronounced.

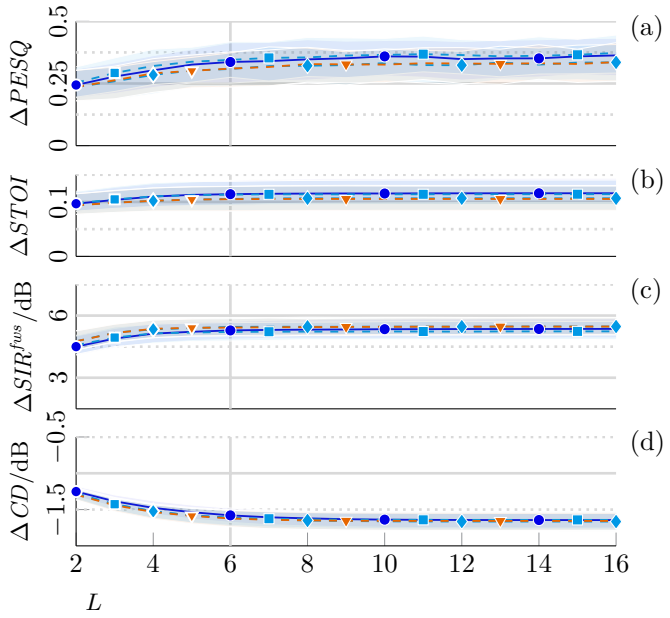


Figure 5.8: (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus L with respect to the reference microphone signal for the $\mathcal{O}(L^2M^2)$-cost [—●—], the $\mathcal{O}(LM^2)$-cost [- ■ -], the $\mathcal{O}(L^2M)$-cost [- ♦ -], and the $\mathcal{O}(LM)$-cost [- ▼ -] ISCLP Kalman filter at $SNR = 25\,\text{dB}$ if interfering speech is present.

5.4.2.3 Summary

We summarize the presented simulation results by noting that all three simplified variants of the ISCLP Kalman filter, i.e. the $\mathcal{O}(LM^2)$-cost, the $\mathcal{O}(L^2M)$-cost, and the $\mathcal{O}(LM)$-cost variant, perform nearly as well as the original $\mathcal{O}(L^2M^2)$-cost variant, thereby permitting far more favourable trade-offs between complexity and performance. The assumptions underlying these simplifications, cf. Sec. 5.3.2, are hence not critical, i.e. they do not form the bottleneck in the modeling and estimation chain. Instead, performance limitations may mostly be caused by other modeling deficiencies, such as deficiencies in the microphone signal model and the ISCLP state-space model, as well as parameter estimation errors, cf. Sec. 5.2.

5.5 Conclusion

The ISCLP Kalman filter requires $\mathcal{O}(L^2M^2)$ multiplications per frequency bin in its original formulation, with L the number of filter coefficients and M the number of channels. In this chapter, we have proposed low-complexity variants of the ISCLP Kalman filter, which have been obtained by simplification of the ISCLP Kalman filter update equations. Hereat, we have enforced the state estimation error correlation matrix to assume sparse structures corresponding to the negligence of either temporal, spatial, or all cross-correlations. The proposed simplifications lead to $\mathcal{O}(LM^2)$-cost, $\mathcal{O}(L^2M)$-cost, and $\mathcal{O}(LM)$-cost variants of the ISCLP Kalman filter, respectively, which may equivalently be represented by multiple Kalman filters sharing the same error signal and error signal PSD. Simulation results indicate that the simplified variants of the ISCLP Kalman filter perform nearly as well as the original variant, thereby permitting far more favourable trade-offs between complexity and performance.

Chapter 6

Conclusion

In this thesis, we have compared, proposed and evaluated existing and novel approaches in the fields of speech enhancement. We have taken account of practical challenges arising from typical applications. These challenges have led to the following technical requirements, which guided the design of the proposed approaches: the comprehensive enhancement of target speech in adverse acoustic conditions characterized by reverberation, interfering speech and background noise; the independence of in practice hardly reliable channel knowledge on the target source, the exploitation of spatial and temporal target-parameter knowledge and further its acquisition in a multi-source scenario; online processing in dynamic acoustic scenarios; and moderate computational complexity.

Specifically, we have analyzed and compared two existing speech enhancement architectures, namely the GSC and the MCLP architecture, in terms of their potential for dereverberation and noise reduction. We have proposed a comprehensive speech enhancement approach, namely the ISCLP Kalman filter, which integrates the GSC and MCLP and exploits spatial and temporal target-parameter knowledge. Further, we have proposed a parameter estimation approach for multi-source scenarios, namely square root-based multi-source early PSD estimation and RETF updating, which is employed in the ISCLP Kalman filter. Both the proposed speech enhancement approach and the proposed parameter estimation approach operate online in dynamic acoustic scenarios. In addition, we have proposed complexity-reduced variants of the ISCLP Kalman filter, which permit far more favourable trade-offs between computational complexity and performance.

The remainder of this conclusion is organized as follows. In Sec. 6.1, we review the results of the above mentioned contributions. In Sec. 6.2, we give suggestions for future research. In Sec. 6.3, we discuss industrial valorization.

6.1 Summary

The summary is organized per chapter.

Comparative Analysis of the GSC and MCLP

The presently most popular approach to dereverberation, namely MCLP, is based on a convolutive reverberation model, but does not require channel knowledge on the target source. However, MCLP is a mere dereverberation approach. Beamforming on the other hand, e.g., by means of the GSC, is a

well-established approach to interfering speech cancellation and noise reduction. In **Ch. 2**, as a preparatory step towards comprehensive speech enhancement, we therefore have analyzed and compared the MCLP and GSC architectures in terms of their potential for dereverberation and noise reduction. In the signal model, we have considered a reverberant target-speech component, coherent (i.e. point-source) noise and incoherent noise components.

The major difference between both architectures lies in the pre-processing of the microphone signals in the LP and the SC filter paths. Where MCLP applies a simple delay, the GSC instead applies a BM, which exploits spatial target-parameter knowledge and hence allows to distinguish the target speech source from other point sources. We have shown that both architectures are theoretically able to perform complete dereverberation if incoherent noise is absent. Due to the exploitation of spatial knowledge in the BM, in the absence of incoherent noise, coherent noise is theoretically completely canceled in the GSC, while it is merely dereverberated in MCLP. For both the GSC and MCLP, unbiased filter estimates are obtained if temporal target-parameter knowledge is incorporated in the cost function for filter estimation.

In order to confirm the theory and to assess the practical relevance of the theoretical findings, we have carried out TD simulations using perfect spatial knowledge in form of early target-speech RIRs, resulting in complete blocking of the early speech source image, and TFD domain simulations using deficient spatial knowledge in form of RETF estimates, resulting in incomplete blocking. The simulation results confirm the theoretical findings in case of complete blocking. In case of incomplete blocking, however, the GSC performs worse as compared to MCLP in terms of dereverberation.

The ISCLP Kalman Filter

In **Ch. 3**, based on the comparison of the MCLP and the GSC architecture in Ch. 2 and the success of MCLP-and-beamforming cascades in literature, we have proposed to integrate the GSC and MCLP into a parallel architecture referred to as ISCLP. In the signal model, we have considered several reverberant speech components, whereof some are to be dereverberated and others to be canceled, as well as a diffuse (e.g., babble) noise component to be suppressed.

Specifically, the ISCLP architecture consists of three signal paths: a reference path employing an MF, an SC path, composed of a BM and an SC filter, and a LP path, composed of a delay and an LP filter. The SC and the LP filter operate on different microphone signal frames. While the MF, the BM and the SC filter are multiplicative, the LP filter is convolutive. The MF and the BM perform spatial pre-processing, serving unconstrained estimation of the SC

filter, while the delay may analogously be considered as temporal pre-processing, serving unconstrained estimation of the LP filter. We have estimated the SC and LP filters jointly and online by means of a single Kalman filter. We further have proposed a spectral Wiener gain post-processor, relating to the Kalman filter's posterior state estimate. The ISCLP Kalman filter requires spatial and temporal target-parameter knowledge, which is obtained by means of the proposed parameter estimation approach for multi-source scenarios, i.e. by square root-based multi-source early PSD estimation and RETF updating. Further implementational aspects such as spatio-temporal target component leakage as well as process equation parameter tuning and initialization have been discussed.

The presented ISCLP Kalman filter has been benchmarked against two state-of-the-art approaches in presence and absence of interfering speech, namely first a pair of alternating Kalman filters respectively performing dereverberation and noise reduction, and second an MCLP+GSC Kalman filter cascade. With M the number of microphones, the ISCLP Kalman filter is roughly M^2 times less expensive than both reference algorithms. Nonetheless, simulation results indicate better or similar performance as compared to the original or modified version of the alternating Kalman filters, and better performance as compared to the MCLP+GSC Kalman filter cascade.

Multi-Source Early PSD Estimation and RETF Update

The ISCLP Kalman filter in Ch. 3 requires spatial and temporal target-parameter knowledge. In **Ch. 4**, we have proposed an appropriate online parameter estimation approach, namely square root-based multi-source early PSD estimation and RETF updating. In the signal model, we have considered multiple sources in a reverberant environment.

Given RETF estimates, state-of-the-art approaches to early PSD estimation conventionally minimize the approximation error defined with respect to an estimate of the early correlation matrix, which we have referred to as conventional MP. Instead, we here have factorized the early correlation matrix model and minimize the approximation error defined with respect to an estimate of the early-correlation-matrix square root, which we have referred to as the square-root MP. The square-root MP seeks a unitary matrix and the square roots of the early PSDs up to an arbitrary complex argument, and therewith constitutes a generalization of the orthogonal Procrustes problem. As opposed to the conventional MP, non-negative inequality constraints are not required in the square-root MP. The square-root MP may be solved iteratively, requiring one SVD per iteration. The estimated unitary matrix and early PSD square

roots further allow to recursively update the RETF estimate, which is not inherently possible in the conventional approach. The respectively required estimates of the early correlation matrix and the early-correlation-matrix square root have been obtained from an estimate of the microphone signal correlation matrix by means of the GEVD. Hereat, in order to compensate for inevitable recursive averaging, we have restored non-stationarities by desmoothing the generalized eigenvalues.

The proposed approach has been evaluated in two kinds of simulations. In the first kind, the data is generated based on the microphone signal correlation matrix model and assumed geometric and physical properties, excluding modeling errors from the evaluation. This is referred to as model-based-data case. In the second kind, the data is generated from recorded speech and measured RIRs, creating a more practical setup. This is referred to as acoustic-data case. In both cases, the simulation results indicate better performance of the square-root MP as compared to the conventional MP. As shown in model-based-data case, the square-root MP can be solved in only one iteration if initialized accordingly. As shown in the acoustic-data case, both the square-root MP and the conventional MP suffer somewhat from residual late reverberation in the early-correlation-matrix estimate.

Low-Complexity ISCLP Kalman Filters

The ISCLP Kalman filter in Ch. 3 requires $\mathcal{O}(L^2M^2)$ multiplications per frequency bin in its original formulation, with L the number of (multi-channel) filter coefficients and M the number of channels. In **Ch. 5**, we have therefore proposed low-complexity variants of the ISCLP Kalman filter.

The low-complexity variants of the ISCLP Kalman filter have been obtained by simplification of the ISCLP Kalman filter update equations. Here, we have enforced the state estimation error correlation matrix to assume sparse structures corresponding to the negligence of either temporal, spatial, or all cross-correlations. The proposed simplifications lead to $\mathcal{O}(LM^2)$-cost, $\mathcal{O}(L^2M)$-cost, and $\mathcal{O}(LM)$-cost variants of the ISCLP Kalman filter, respectively, which may equivalently be represented by multiple Kalman filters sharing the same error signal and error signal PSD.

Simulation results based on the same acoustic scenarios as for the original ISCLP Kalman filter in Ch. 3 indicate that the simplified ISCLP Kalman filter variants perform nearly as well as the original variant, thereby permitting far more favourable trade-offs between complexity and performance.

6.2 Suggestions for Future Research

The presented work leaves room for future research in several directions. We outline a few examples as follows:

- In (hybrid) MCLP-based speech enhancement approaches [90–96, 98–103, 111–116], as outlined in Sec. 1.2.2.1, two main paradigms exist regarding temporal target-parameter knowledge. In iterative batch processing approaches [90, 91, 93–96, 112–115], temporal target-parameter knowledge is acquired in alternation with the LP filter estimate and hence is solely based on intermediate enhanced-signal batches. In online processing approaches [92, 98–103, 111, 116] instead, in each frame, temporal target-parameter knowledge is acquired a priori and hence is typically independent of previous enhanced-signal frames [92, 98, 100–103, 111]. While some exemplary comparisons between batch and online processing have been done in MCLP-based dereverberation [101, 103], a systematic comparison and evaluation of these two paradigms, i.e. intermediate versus a-priori temporal target-parameter estimation, has yet not been done in literature. For the comprehensive speech enhancement task, such a comparison may be done within the ISCLP architecture proposed in Sec. 3.3.1. In this context, the role of spatio-temporal target-component leakage, cf. Sec. 3.4.1, may be of particular interest.
- The ISCLP Kalman filter proposed in Ch. 3 exploits a-priori temporal target-parameter knowledge, which is obtained independently by means of square root-based multi-source early PSD estimation and RETF updating as proposed in Ch. 4. Instead of this independent acquisition, one may consider to couple the speech enhancement and target-parameter estimation tasks. While it may be rather obvious to estimate a-priori temporal target-parameter knowledge in the current frame based on previous enhanced-signal frames as in [99, 116], the comparison of the original [116] and the modified alternating Kalman filters in Sec. 3.5.5.1 indicates that this approach leads to reduced performance as compared to independent estimation due to the highly time-varying statistics of the early target-speech source image. Alternatively, however, the output of the ISCLP cancellation filter may be used to (re-)estimate the comparably slowly time-varying coherence of undesired sound components, in particular the coherence of late reverberation and background noise. This coherence estimate may in turn be used for parameter estimation instead of solely relying on the diffuse field assumption, cf. Sec. 4.5, and hence potentially improve performance. Note that this coherence

estimation approach requires that instead of the MF output, each microphone signal is enhanced separately.

- The ISCLP Kalman filter proposed in Ch. 3 produces a single enhanced output signal only. Some applications however rely on multiple output channels, such that an extension may become necessary. In BSS [94, 135, 136] for instance, one is interested in obtaining an estimate of each of the N early (speech) source images, which requires N output channels. A corresponding extension of the ISCLP architecture proposed in Sec. 3.3.1 and the ISCLP state-space model and Kalman filter in Sec. 3.3.2 is rather straightforward. In hearing aids, the wearer perceptually benefits from the preservation of binaural cues such as inter-aural time and level differences (ITDs and ILDs) or the coherence of undesired sound components, which requires two output channels. While several binaural-cue preseveration approaches based on additive models for noise and late reverberation exist [42, 195–198] and also multiple-output MCLP architectues have been proposed [94, 95, 97], the use of convolutive reverberation models as in MCLP and ISCLP specifically for binaural cue preservation has yet not been explored.

- In [115], it has recently been proposed to unify an MCLP+MVDR cascade, which is shown to result in a parallel architecture of an MVDR beamformer and an LP filter. The MVDR beamformer and LP filter coefficients are jointly estimated by means of batch processing. Although motivated differently, the architecture proposed in [115] is obviously closely related to the ISCLP architecture proposed in Sec. 3.3.1, which integrates the GSC (i.e. an unconstrained variant of the MVDR beamformer) and MCLP. A further investigation on the relations and the differences between both may provide deeper understanding of the architectures at hand and be fertile in the development of further modifications and extensions.

- In the multi-source early PSD estimation and RETF updating approach proposed in Ch. 4, the RETFs are updated independently in each frequency bin, cf. Sec. 4.4.3. However, we may argue that the early source images are dominated by their direct components, and hence the RETFs should relate to the corresponding free-field transfer functions, which are solely defined by DoAs and imply predetermined phase relations between microphones as a function of frequency [144–146]. Hence, more robust RETF estimates may be obtained by penalizing deviation from potential free-field transfer functions, which requires optimization across frequency bins.

- In the multi-source early PSD estimation and RETF updating approach proposed in Ch. 4, the early correlation matrix estimate contains residual

late reverberation, which has been shown to somewhat affect the multi-source early PSD estimation, cf. Sec 4.6.2.4. In case of speech, in order to reduce the effects of residual late reverberation, we may exploit sparse representations of the early speech source image in the TFD [96,97,100,199] namely by introducing a sparsity-promoting penalty term in the square-root MP in Sec. 4.4.2. Promoting sparsity may further help to improve the source-component separation, as in the majority of time-frequency tiles at most one speech source is active. The latter characteristic may also be exploited in acoustic scenarios with more speech sources than microphones, which so far have not been considered.

- In the multi-source early PSD estimation and RETF updating approach proposed in Ch. 4, in order to compensate for the inevitable recursive averaging in the estimation of the microphone signal correlation matrix, we have proposed to restore non-stationarities by subspace-based desmoothing, cf. Sec. 4.5. Precisely, desmoothing is perfomed by filtering of the (generalized) eigenvalues, where the eigenvalue filter is defined as the inverse of the recursive averaging filter. While the proposed approach has been shown to be effective in practice, its derivation has been based on an intuitive interpretation rather than analytic algebraic arguments. An profound algebraic derivation may yield further insight into the assets and limitations of the proposed approach as well as a solid base for further developments. An evaluation of the presumably greatly reduced dependency of resultant early PSD estimates on the forgetting factor used in recursive averaging may be highly relevant. We may further assess the effect of desmoothing on the performance of other speech enhancement approaches exploiting temporal target-parameter knowledge, e.g., the MWF [42, 61–68, 129, 130].

- In the multi-source early PSD estimation and RETF updating approach proposed in Ch. 4, the GEVD of the recursively averaged microphone signal correlation matrix estimate and the diffuse coherence matrix is computed independently in each frame, cf. Sec. 4.5. In order to reduce the computational complexity of the approach, the GEVD may instead be estimated recursively by means of the power method [182, 183]. In addition, if the GEVD is estimated recursively, the recursive sorting [153] of generalized eigenvalue-eigenvector pairs, which is otherwise required for subspace-based desmoothing, becomes redundant.

- The combination of the ISCLP Kalman filter and square root-based multi-source early PSD estimation and RETF updating as proposed in Ch. 3 and Ch. 4 has so far been tested only on synthesized microphone signals generated from measured RIRs [26], dry speech signals [25], and artifically diffused babble noise [27, 167], which enabled us to perform detailed

objective evaluations by means of intrusive performance measures. In addition, evaluation based on recorded microphone signals and subjective listening tests may be performed. Subjective quality and intelligibility scores may e.g., be obtained by means of multi-stimulus tests with hidden reference and anchor (MUSHRA) [200] and by measuring the speech reception threshold (SRT) [15], respectively. Further, the impact on ASR [31, 45–49] performance may be evaluated, e.g., based on the evaluation frameworks defined in the ASpIRE [47], the REVERB [48], and the CHiME [49] challenges.

6.3 Industrial Valorization

As outlined in Sec. 1.2.1, speech enhancement has many applications ranging from digital telephony over hearing devices to voice-driven human-to-machine communication in smart devices, and hence bears a tremendous economic potential. In 2018, the number of mobile telephony subscriptions was already greater than the global population, with nearly the whole world population living within the range of a mobile-cellular network signal [201]. Global smart phone sales reached $522 billion USD in 2018 [202]. On the VoIP market, 204.8 billion corporate consumer users accounting for $86.20 billion USD in global revenues are predicted for 2020 [203]. The global market for hearing devices, which is driven by the rising aging population, is expected to reach a value of $12.1 billion USD by 2025 despite too low penetration rates in particular in developing countries [204, 205]. The revenues on the global speech and voice recognition market are estimated to reach $31.82 billion USD by 2025, corresponding to a compound annual growth rate (CAGR) of 17.2% during the forecast period [206]. A large number of well-known and less well-known companies operate in these domains, such as Samsung, Apple, Huawei, Microsoft, Google, Amazon, NXP Semiconductors, Phonak, Oticon, Starkey, Widex, Cochlear, and many more.

In the design of the approaches proposed in Ch. 3 to Ch. 5, technical requirements arising from practical applications have been taken into account in order to lower the threshold for industrial valorization on the above mentioned markets. However, regardless of the actual application, a number of further steps has to be taken in order to transfer the presented research into a product.

The parameter estimation approach proposed in Ch. 4 requires initial spatial knowledge, and so has to be extended for or integrated with initial RETF or DoA estimation. Further, sensor and quantization noise is always present in practical applications and thus needs to be considered in the signal model and the parameter estimation strategy.

The speech enhancement and parameter estimation approaches proposed in Ch. 3 to Ch. 5 have to be cascaded or integrated with other processing blocks, e.g., with AEC in telephony, AFC and non-linear compression in hearing devices, or ASR in smart devices. The order of cascading may be critical to the overall functionality, while integration, e.g., by sharing information, joint tuning, or joint optimization across processing blocks, may improve the performance. In applications such as video-conferencing, also audio-visual modality integration may be beneficial. Visual data may, e.g., support both initial acquisition and updating of spatial target-parameter knowledge in the parameter estimation approach proposed in Ch. 3.

Due to finite precision in practical applications, in particular on platforms with fixed-point arithmetic, all implementations need to be numerically robust. The standard formulation of the Kalman filter for instance is known to be numerically unstable, and so numerically better-behaved QR decomposition-based formulations [187, 188] may instead be used to implement the ISCLP Kalman filter and its low-complexity variants proposed in Ch. 3 and Ch. 5.

The resulting software needs to be migrated from machine-independent higher level programming languages typically used during algorithmic development such as MATLAB [152–154] or Python to lower level programming languages such as C++ and C or even machine-specific assembly languages. Apart from the signal processing layer, the newly developed processing blocks also have to be integrated on a software-architectural layer. Test scripts have to be newly written or extended to ensure that the previously developed software functions as expected. Depending on the business model, software libraries may be sold as a stand-alone product or in combination with dedicated hardware either to original equipment manufacturers (OEMs) or directly to the end user. During the product's life cycle, the software needs to be supported and maintained.

Throughout the entire product development process, the performance in terms of improvements in speech quality and intelligibility or ASR performance needs to be continuously (re-)evaluated.

Appendix A

Appendix to Chapter 2

A.1 GSC Enhancement in Absence of Incoherent Noise – Reformulation

A.1.1 GSC Filter Output

Analogously to (2.16)–(2.17), let $\mathbf{C}_e \in \mathbb{C}^{(NL_s \times NL_s)}$ and its counterpart $\mathbf{C}_\ell$ be defined by

$$\mathbf{C}_{\mathrm{e}} = \begin{pmatrix} \mathbf{I}^{d\times d} & \mathbf{0}^{d\times(NL_s-d)} \\ \mathbf{0}^{(NL_s-d)\times d} & \mathbf{0}^{d\times(NL_s-d)} \end{pmatrix}, \tag{A.1}$$

$$\mathbf{C}_\ell = \mathbf{I}^{NL_s\times NL_s} - \mathbf{C}_{\mathrm{e}}, \tag{A.2}$$

and let $\mathbf{\Psi}'_{\tilde{s}} \in \mathbb{C}^{NL_s-d\times NL_s-d}$ be the submatrix of $\mathbf{\Psi}_{\tilde{s}}$ spanning its last $NL_s - d$ rows and columns, such that

$$\mathbf{C}_\ell \mathbf{\Psi}_{\tilde{s}} \mathbf{C}_\ell = \begin{pmatrix} \mathbf{0}^{d\times d} & \mathbf{0}^{d\times(NL_s-d)} \\ \mathbf{0}^{(NL_s-d)\times d} & \mathbf{\Psi}' \end{pmatrix}. \tag{A.3}$$

With $\mathbf{HB}$ as given in (2.27), the expression $((\mathbf{HB})^H \mathbf{\Psi}_{\tilde{s}} \mathbf{HB})^P$ in (2.54) can then be written as

$$\left((\mathbf{HB})^H \mathbf{\Psi}_{\tilde{s}} \mathbf{HB}\right)^P = (\mathbf{H}_B^H \mathbf{\Psi}' \mathbf{H}_B)^P$$

$$= \mathbf{H}_B^P \boldsymbol{\Psi}_{\tilde{s}}'^{-1} \mathbf{H}_B^{+T}. \tag{A.4}$$

Inserting (2.27) and (A.4) in (2.54) while noting that $\mathbf{H}_B \mathbf{H}_B^P = \mathbf{I}$ since $\mathbf{H}_B$ is assumed to have full row rank, and further using (A.3) and Lemma 1 from Appendix A.2, we obtain

$$\begin{aligned} z(l) &= \Big(\mathbf{HB}\big((\mathbf{HB})^H \boldsymbol{\Psi}_{\tilde{s}} \mathbf{HB}\big)^P (\mathbf{HB})^H \boldsymbol{\Psi}_{\tilde{s}} \mathbf{Hg}\Big)^H \mathbf{s}(l) \\ &= \left(\begin{pmatrix} \mathbf{0}^{d\times d} & \mathbf{0}^{d\times(NL_s-d)} \\ \mathbf{0}^{(NL_s-d)\times d} & \boldsymbol{\Psi}_{\tilde{s}}'^{-1} \end{pmatrix} \boldsymbol{\Psi}_{\tilde{s}} \mathbf{Hg} \right)^H \mathbf{s}(l) \\ &= \big((\mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}} \mathbf{C}_\ell)^P \boldsymbol{\Psi}_{\tilde{s}} \mathbf{Hg}\big)^H \mathbf{s}(l). \end{aligned} \tag{A.5}$$

With $\mathbf{C}_e + \mathbf{C}_\ell = \mathbf{I}$, the matrix $\boldsymbol{\Psi}_{\tilde{s}}$ may be written as

$$\begin{aligned} \boldsymbol{\Psi}_{\tilde{s}} &= \mathbf{C}_e \boldsymbol{\Psi}_{\tilde{s}} + \mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}} \\ &= \mathbf{C}_e \boldsymbol{\Psi}_{\tilde{s}} + \mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}} \mathbf{C}_\ell + \mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}} \mathbf{C}_e. \end{aligned} \tag{A.6}$$

Substituting (A.6) in (A.5), we find $(\mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}} \mathbf{C}_\ell)^P (\mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}} \mathbf{C}_\ell) = \mathbf{C}_\ell$ from (A.3), while $(\mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}} \mathbf{C}_\ell)^P \mathbf{C}_e \boldsymbol{\Psi}_{\tilde{s}} = \mathbf{0}$, such that $(\mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}} \mathbf{C}_\ell)^P \boldsymbol{\Psi}_{\tilde{s}}$ in (A.5) becomes

$$(\mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}} \mathbf{C}_\ell)^P \boldsymbol{\Psi}_{\tilde{s}} = \mathbf{C}_\ell + (\mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}} \mathbf{C}_\ell)^P \mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}} \mathbf{C}_e. \tag{A.7}$$

Using (A.1)–(A.2), (2.16)–(2.17), and Lemma 1 from Appendix A.2, the term $(\mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}} \mathbf{C}_\ell)^P \mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}} \mathbf{C}_e$ in (A.7) takes the form

$$\begin{aligned} &(\mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}} \mathbf{C}_\ell)^P \mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}} \mathbf{C}_e \\ &= \begin{pmatrix} (\mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}_1} \mathbf{C}_\ell)^P \mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}_1} \mathbf{C}_e & \mathbf{0}^{L_s\times(N-1)L_s} \\ \mathbf{0}^{(N-1)L_s\times L_s} & \mathbf{0}^{(N-1)L_s\times(N-1)L_s} \end{pmatrix}. \end{aligned} \tag{A.8}$$

Inserting (A.7) in (A.5), multiplying out, using (A.1)–(A.2) and (2.14)–(2.17), it can be shown that

$$\begin{aligned} z(l) &= (\mathbf{C}_\ell \mathbf{Hg})^H \mathbf{s}(l) + \big((\mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}} \mathbf{C}_\ell)^P \mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}} \mathbf{C}_e \mathbf{Hg}\big)^H \mathbf{s}(l) \\ &= \sum_{n=2}^{N} q_{s_n}(l) + q_{s_1|\ell}(l) \\ &\quad + \big((\mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}_1} \mathbf{C}_\ell)^P \mathbf{C}_\ell \boldsymbol{\Psi}_{\tilde{s}_1} \mathbf{C}_e \mathbf{H}_1 \mathbf{g}\big)^H \mathbf{s}_1(l). \end{aligned} \tag{A.9}$$

With Lemma 1 from Appendix A.2 and $\mathbf{C}_{d|\ell}$ defined in (2.19), $(\mathbf{C}_\ell \mathbf{\Psi}_{\tilde{s}_1} \mathbf{C}_\ell)^P$ can be written as

$$(\mathbf{C}_\ell \mathbf{\Psi}_{\tilde{s}_1} \mathbf{C}_\ell)^P = \mathbf{C}_{d|\ell}{}^H (\mathbf{C}_{d|\ell} \mathbf{\Psi}_{\tilde{s}_1} \mathbf{C}_{d|\ell}{}^H) \mathbf{C}_{d|\ell}, \tag{A.10}$$

Substituting (A.10) into (A.9) and extracting $\mathbf{C}_{d|\ell}{}^H$ on the left, we obtain (2.55)–(2.56).

A.1.2 GSC Bias

With Lemma 1 in Appendix A.2, the first term in (2.56), $(\mathbf{C}_{d|\ell} \mathbf{\Psi}_{\tilde{s}_1} \mathbf{C}_{d|\ell}{}^H)^P$, can be written as

$$(\mathbf{C}_{d|\ell} \mathbf{\Psi}_{\tilde{s}_1} \mathbf{C}_{d|\ell}{}^H)^P = \begin{pmatrix} \mathbf{\Psi}'^{-1}_{\tilde{s}_1} & \mathbf{0}^{(L_s-d)\times d} \\ \mathbf{0}^{d\times(L_s-d)} & \mathbf{0}^{d\times d} \end{pmatrix}, \tag{A.11}$$

where $\mathbf{\Psi}'_{\tilde{s}_1} \in \mathbb{C}^{(L_s-d)\times(L_s-d)}$ is the submatrix of $\mathbf{\Psi}_{\tilde{s}_1}$ spanning the last $L_s - d$ rows and columns, matching the first term in (2.48) for $n = 1$. Note that for $d = L_b$, $\mathbf{\Psi}'^{-1}_{\tilde{s}_1} \in \mathbb{C}^{(L_s-d)\times(L_s-d)}$ in (A.11) and $\mathbf{\Psi}^{-1}_{\tilde{s}_1} \in \mathbb{C}^{L_s\times L_s}$ in (2.48) also correspond in terms of dimensions: inserting (2.26) into (2.9) yields $L_s - d = L_h + L_w - 1$ in the GSC, while inserting (2.21) into (2.9) yields $L_s = L_h + L_w - 1$ in MCLP. Finally, since the second term in (2.56), $\mathbf{C}_{d|\ell} \mathbf{\Psi}_{\tilde{s}_1} \mathbf{C}_e \mathbf{H}_1 \mathbf{g}$, is equivalent to the second term in (2.48), both expressions (2.56) and (2.48) yield the same bias component.

A.2 The Pesudoinverse of Block-Diagonal Matrices

Lemma 1. *The pseudoinverse* $\mathbf{A}^P$ *of a block-diagonal matrix* $\mathbf{A}$ *defined by the blocks* $\mathbf{A}_n$, $n = 1 \dots N$, *on its diagonal is given by a block-diagonal matrix composed of the pseudoinverses* $\mathbf{A}_n^P$ *of the individual blocks, i.e.*

$$\textit{if} \qquad \mathbf{A} = \text{Blkdiag}\left[\mathbf{A}_1,\ \dots,\ \mathbf{A}_N\right],$$

$$\textit{then} \qquad \mathbf{A}^P = \text{Blkdiag}\left[\mathbf{A}_1^P,\ \dots,\ \mathbf{A}_N^P\right].$$

This lemma can be proven easily by verifying the four criteria defining the pseudoinverse $\mathbf{A}^P$ of the matrix $\mathbf{A}$, i.e. $\mathbf{A}\mathbf{A}^P\mathbf{A} = \mathbf{A}$, $\mathbf{A}^P\mathbf{A}\mathbf{A}^P = \mathbf{A}^P$, $(\mathbf{A}\mathbf{A}^P)^H = \mathbf{A}\mathbf{A}^P$, and $(\mathbf{A}^P\mathbf{A})^H = \mathbf{A}^P\mathbf{A}$. It is further important to note that the pseudoinverse of a zero matrix is equal to its transpose.

Appendix B

Appendix to Chapter 4

B.1 The Orthogonal Procrustes Problem – Reformulation

We note that the Frobenius norm $\left\|\mathbf{E}_{sq}(\boldsymbol{\Omega}, \hat{\boldsymbol{\varphi}}_s^{1/2|(i-1)})\right\|_F^2$ in (4.35) may be expressed by means of the trace operator as

$$\left\|\mathbf{E}_{sq}(\boldsymbol{\Omega}, \hat{\boldsymbol{\varphi}}_s^{1/2|(i-1)})\right\|_F^2 = \mathrm{tr}\big[\mathbf{E}_{sq}(\boldsymbol{\Omega}, \hat{\boldsymbol{\varphi}}_s^{1/2|(i-1)})\mathbf{E}_{sq}^H(\boldsymbol{\Omega}, \hat{\boldsymbol{\varphi}}_s^{1/2|(i-1)})\big]. \tag{B.1}$$

Using (B.1) with $\mathbf{E}_{sq}(\boldsymbol{\Omega}, \hat{\boldsymbol{\varphi}}_s^{1/2|(i-1)}) = \hat{\boldsymbol{\Psi}}_{x_e}^{1/2}\boldsymbol{\Omega} - \hat{\mathbf{H}}\,\mathrm{Diag}[\hat{\boldsymbol{\varphi}}_s^{1/2|(i-1)}]$ according to (4.32), we reformulate (4.35) as

$$\begin{aligned}
\hat{\boldsymbol{\Omega}}^{(i)} &= \arg\min_{\boldsymbol{\Omega}} \left\|\mathbf{E}_{sq}(\boldsymbol{\Omega}, \hat{\boldsymbol{\varphi}}_s^{1/2|(i-1)})\right\|_F^2 \\
&\qquad \text{s.t.} \quad \boldsymbol{\Omega}\boldsymbol{\Omega}^H = \mathbf{I} \\
&= \arg\min_{\boldsymbol{\Omega}} \; -\mathrm{tr}\big[\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}\boldsymbol{\Omega}\,\mathrm{Diag}[\hat{\boldsymbol{\varphi}}_s^{H/2|(i-1)}]\hat{\mathbf{H}}^H\big] \\
&\qquad\qquad -\mathrm{tr}\big[\hat{\mathbf{H}}\mathrm{Diag}[\hat{\boldsymbol{\varphi}}_s^{1/2|(i-1)}]\boldsymbol{\Omega}^H\boldsymbol{\Psi}_{x_e}^{H/2}\big] \\
&\qquad \text{s.t.} \quad \boldsymbol{\Omega}\boldsymbol{\Omega}^H = \mathbf{I} \\
&= \arg\max_{\boldsymbol{\Omega}} \Re\Big\{\mathrm{tr}\big[\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}\boldsymbol{\Omega}\,\mathrm{Diag}[\hat{\boldsymbol{\varphi}}_s^{H/2|(i-1)}]\hat{\mathbf{H}}^H\big]\Big\} \\
&\qquad \text{s.t.} \quad \boldsymbol{\Omega}\boldsymbol{\Omega}^H = \mathbf{I},
\end{aligned} \tag{B.2}$$

where constant terms have been dropped in the first transition. Inserting $\mathrm{tr}\big[\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}\boldsymbol{\Omega}\,\mathrm{Diag}[\hat{\boldsymbol{\varphi}}_s^{H/2|(i-1)}]\hat{\mathbf{H}}^H\big] = \mathrm{tr}\big[\boldsymbol{\Omega}\,\mathrm{Diag}[\hat{\boldsymbol{\varphi}}_s^{H/2|(i-1)}]\hat{\mathbf{H}}^H\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}\big]$ in (B.2), we finally obtain (4.37)–(4.38).

B.2 The Orthogonal Procrustes Problem – Solution

Inserting the SVD $\mathbf{C}^{H|(i-1)} = \mathbf{U}_L\boldsymbol{\Sigma}\mathbf{U}_R^H$ according to (4.39) into (4.37), we obtain

$$\begin{aligned}\hat{\boldsymbol{\Omega}}^{(i)} &= \arg\max_{\boldsymbol{\Omega}} \Re\big\{\mathrm{tr}[\boldsymbol{\Omega}\mathbf{C}_{sq}^{(i-1)}]\big\} \\ &\qquad \text{s.t.} \quad \boldsymbol{\Omega}\boldsymbol{\Omega}^H = \mathbf{I}, \\ &= \arg\max_{\boldsymbol{\Omega}} \Re\big\{\mathrm{tr}[\boldsymbol{\Omega}\mathbf{U}_R\boldsymbol{\Sigma}\mathbf{U}_L^H]\big\} \\ &\qquad \text{s.t.} \quad \boldsymbol{\Omega}\boldsymbol{\Omega}^H = \mathbf{I}. \end{aligned} \tag{B.3}$$

With $\mathrm{tr}[\boldsymbol{\Omega}\mathbf{U}_R\boldsymbol{\Sigma}\mathbf{U}_L^H] = \mathrm{tr}[\mathbf{U}_L^H\boldsymbol{\Omega}\mathbf{U}_R\boldsymbol{\Sigma}]$ and $\mathbf{U}_L^H\boldsymbol{\Omega}\mathbf{U}_R$ being unitary (as it is a product of unitary matrices), the trace in (B.3) is maximized if $\mathbf{U}_L^H\boldsymbol{\Omega}\mathbf{U}_R = \mathbf{I}$, and hence we obtain $\hat{\boldsymbol{\Omega}}^{(i)} = \mathbf{U}_L\mathbf{U}_R^H$.

Acknowledgements

This research work was carried out at the ESAT Laboratory of KU Leuven, in the frame of KU Leuven Research Council CoE PFV/10/002 (OPTEC), KU Leuven internal funds C2-16-00449 'Distributed Digital Signal Processing for Ad-hoc Wireless Local Area Audio Networking', Impulse Fund IMP/14/037; IWT O&O Project nr. 150611 'Proof-of-concept of a Rationed Architecture for Vehicle Entertainment and NVH Next-generation Acoustics (RAVENNA)'; VLAIO O&O Project no. HBC.2017.0358 'SPOTT - Tomorrow's Scalable and PersOnalised advertising Technnology, Today'; EU FP7-PEOPLE Marie Curie Initial Training Network 'Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS)', funded by the European Commission under Grant Agreement no. 316969. The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program/ERC Consolidator Grant: SONORA (no. 773268). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.

Bibliography

[1] R. H. Bolt and A. D. MacDonald, "Theory of speech masking by reverberation," *J. Acoust. Soc. Amer.*, vol. 21, no. 6, pp. 577–580, Nov. 1949.

[2] J. P. Moncur and D. Dirks, "Binaural and monaural speech intelligibility in reverberation," *J. Speech Hearing Research*, vol. 10, no. 2, pp. 186–195, Jan. 1967.

[3] R. Plomp, "Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound source (speech or noise)," *Acta Acustic.*, vol. 34, no. 4, pp. 200–211, Feb. 1976.

[4] A. K. Nábělek and D. Mason, "Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms," *J. Speech, Lang., Hearing Research*, vol. 24, no. 3, pp. 375–383, Sep. 1981.

[5] A. K. Nábělek, T. R Letowski, and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification," *J. Acoust. Soc. Amer.*, vol. 86, no. 4, pp. 1259–1265, Oct. 1989.

[6] A. W. Bronkhorst and R. Plomp, "Binaural speech intelligibility in noise for hearing-impaired listeners," *J. Acoust. Soc. Amer.*, vol. 86, no. 4, pp. 1374–1383, Oct. 1989.

[7] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *J. Acoust. Soc. Amer.*, vol. 88, no. 4, pp. 1725–1736, Oct. 1990.

[8] Y. Takata and A. K. Nábělek, "English consonant recognition in noise and in reverberation by Japanese and American listeners," *J. Acoust. Soc. Amer.*, vol. 88, no. 2, pp. 663–666, Aug. 1990.

[9] K. L. Payton, R. M. Uchanski, and L. D. Braida, "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing," *J. Acoust. Soc. Amer.*, vol. 95, no. 3, pp. 1581–1592, Mar. 1994.

[10] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. 5th Intl. Conf. Spoken Lang. Process.*, Sydney, Australia, Nov. 1998.

[11] J. S. Bradley, R. D. Reich, and S. G. Norcross, "On the combined effects of signal-to-noise ratio and room acoustics on speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 106, no. 4, pp. 1820–1828, June 1999.

[12] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustic.*, vol. 86, no. 1, pp. 117–128, Jan. 2000.

[13] S. R. Bistafa and J. S. Bradley, "Reverberation time and maximum background-noise level for classrooms from a comparative study of speech intelligibility metrics," *J. Acoust. Soc. Amer.*, vol. 107, no. 2, pp. 861–875, Jan. 2000.

[14] B. Libbey and P. H. Rogers, "The effect of overlap-masking on binaural reverberant word intelligibility," *J. Acoust. Soc. Amer.*, vol. 116, no. 5, pp. 3141–3151, Nov. 2004.

[15] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Amer.*, vol. 120, no. 1, pp. 331–342, Apr. 2006.

[16] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[17] K. Kokkinakis and P. C. Loizou, "Evaluation of objective measures for quality assessment of reverberant speech," in *Proc. 2011 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2011)*, Prague, Czech Republic, May 2011, pp. 2420–2423.

[18] H. Kuttruff, *Room Acoustics*, SPON Press, 5th edition, 2014.

[19] W. C. Sabine, *Collected Papers on Acoustics*, Peninsula Publishing, 1993.

[20] X. Huang, A. Acero, and H.j Hon, *Spoken language processing: A guide to theory, algorithm, and system development*, Prentice Hall, 2001.

[21] K. Honda, "Physiological processes of speech production," in *Springer handbook of speech processing*, pp. 7–26. Springer, 2008.

[22] J. P. A. Lochner and J. F. Burger, "The subjective masking of short time delayed echoes by their primary sounds and their contribution to the intelligibility of speech," *Acta Acust.*, vol. 8, no. 1, pp. 1–10, Jan. 1958.

[23] A. J. Watkins and N. J. Holt, "Effects of a complex reflection on vowel identification," *Acta Acust.*, vol. 86, no. 3, pp. 532–542, May 2000.

[24] M. Hodgson and E. Nosal, "Effect of noise and occupancy on optimal reverberation times for speech intelligibility in classrooms," *J. Acoust. Soc. Amer.*, vol. 111, no. 2, pp. 931–939, Feb. 2002.

[25] Bang and Olufsen, "Music for Archimedes," Compact Disc B&O, 1992.

[26] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. 2014 Intl. Workshop Acoustic Signal Enhancement (IWAENC 2014)*, Antibes – Juan les Pins, France, Sept. 2014, pp. 313–317.

[27] Auditec, "Auditory tests (revised)," Compact Disc Auditec, 1997.

[28] S. Doclo, *Multi-microphone noise reduction and dereverberation techniques for speech applications*, Ph.D. thesis, KU Leuven, Belgium, 2003.

[29] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*, Springer, 2005.

[30] P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2007.

[31] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*, Springer, 2010.

[32] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[33] H. Fletcher and R. H. Galt, "The perception of speech and its relation to telephony," *J. Acoust. Soc. Amer.*, vol. 22, no. 2, pp. 89–151, Mar. 1950.

[34] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment*, Wiley, 2006.

[35] R. Martin, U. Heute, and C. Antweiler, *Advances in digital speech transmission*, Wiley, 2008.

[36] W. C. Chu, "Speech coding algorithms," *Foundation and evolution of standardized coders*, 2003.

[37] J. M. Kates and M. R. Weiss, "A comparison of hearing-aid array-processing techniques," *J. Acoust. Soc. Amer.*, vol. 99, no. 5, pp. 3138–3148, May 1996.

[38] A. Spriet, *Adaptive filtering techniques for noise reduction and acoustic feedback cancellation in hearing aids*, Ph.D. thesis, KU Leuven, Belgium, 2004.

[39] 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. Applied Signal Process.*, vol. 2005, pp. 2915–2929, Dec. 2005.

[40] H. Dillon, *Hearing aids*, Hodder Arnold, 2008.

[41] J. M. Kates, *Digital hearing aids*, Plural publishing, 2008.

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

[43] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[44] G. Bernardi, *Design and Evaluation of Feedback Control Algorithms for Implantable Hearing Devices*, Ph.D. thesis, KU Leuven, Belgium, 2018.

[45] M. Wölfel and J. W. McDonough, *Distant speech recognition*, Wiley, 2009.

[46] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114–126, Nov. 2012.

[47] M. Harper, "The automatic speech recognition in reverberant environments (ASpIRE) challenge," in *Proc. 2015 IEEE Workshop Autom. Speech Recog., Underst. (ASRU)*, Dec. 2015, pp. 547–554.

[48] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 7, pp. 1–19, Dec. 2016.

[49] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. 19th Annual Conf. Intl. Speech Comm. Assoc. (INTERSPEECH 2018)*, Hyderabad, India, Sep. 2018, pp. 1561–1565.

[50] P. S. R. Diniz, *Adaptive filtering*, Springer, 1997.

[51] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 4th edition, 2002.

[52] D. Simon, *Optimal state estimation: Kalman, H-infinity, and nonlinear approaches*, Wiley, 2006.

[53] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in network and acoustic echo cancellation*, Springer, 2001.

[54] T. van Waterschoot, *Design and evaluation of digital signal processing algorithms for acoustic feedback and echo cancellation*, Ph.D. thesis, KU Leuven, Belgium, 2009.

[55] T. van Waterschoot and M. Moonen, "Fifty years of acoustic feedback control: State of the art and future challenges," *Proc. IEEE*, vol. 99, no. 2, pp. 288–327, Feb. 2011.

[56] K. Lebart, J. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acust.*, vol. 87, no. 3, pp. 359–366, May 2001.

[57] E. A. P. Habets and S. Gannot, "Dual-microphone speech dereverberation using a reference signal," in *Proc. 2007 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2007)*, Honolulu, HI, USA, Apr. 2007, pp. 901–904.

[58] E. A. P. Habets, *Single- and multi-microphone speech dereverberation using spectral enhancement*, Ph.D. thesis, Technische Universiteit Eindhoven, The Netherlands, 2007.

[59] O. Thiergart, G. Del Galdo, and E. A. P. Habets, "On the spatial coherence in mixed sound fields and its application to signal-to-diffuse ratio estimation," *J Acoust. Soc. of Amer.*, vol. 132, no. 4, pp. 2337–2346, Oct. 2012.

[60] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 1006–1018, June 2015.

[61] S. Braun and E. A. P. Habets, "A multichannel diffuse power estimator for dereverberation in the presence of multiple sources," *EURASIP J. Audio, Speech, Music Process.*, vol. 2015, no. 34, pp. 1–14, Dec. 2015.

[62] O. Schwartz, S. Gannot, and E. A. P. Habets, "Joint maximum likelihood estimation of late reverberant and speech power spectral density in noisy environments," in *Proc. 2016 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2016)*, Shanghai, China, Mar. 2016, pp. 151–155.

[63] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1599–1612, Sep. 2016.

[64] O. Schwartz, S. Gannot, and E. A. P. Habets, "Joint estimation of late reverberant and speech power spectral densities in noisy environments using Frobenius norm," in *Proc. 24th European Signal Process. Conf. (EUSIPCO 2016)*, Budapest, Hungary, Aug. 2016, pp. 1123–1127.

[65] I. Kodrasi and S. Doclo, "EVD-based multi-channel dereverberation of a moving speaker using different RETF estimation methods," in *Proc. 2017 IEEE Hands-free Speech Com. Mic. Arrays (HSCMA 2017)*, San Francisco, CA, USA, Mar. 2017, pp. 116–120.

[66] I. Kodrasi and S. Doclo, "Joint late reverberation and noise power spectral density estimation in a spatially homogeneous noise field," in *Proc. 2018 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2018)*, Calgary, AB, Canada, Apr. 2018, pp. 441–445.

[67] S. Braun, A. Kuklasiński, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE/ACM Trans. Audio, Speech, Lang. Process*, vol. 26, no. 6, pp. 1056–1071, June 2018.

[68] I. Kodrasi and S. Doclo, "Analysis of eigenvalue decomposition-based late reverberation power spectral density estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1102–1114, June 2018.

[69] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Robust joint estimation of multi-microphone signal model parameters," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 7, pp. 1136–1150, July 2019.

[70] M. Tammen, S. Doclo, and I. Kodrasi, "Joint estimation of RETF vector and power spectral densities for speech enhancement based on alternating least squares," in *Proc. 2019 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2019)*, Brighton, United Kingdom, May 2019, pp. 795–799.

[71] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[72] M. Kallinger and A. Mertins, "Multi-channel room impulse response shaping – a study," in *Proc. 2006 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2006)*, Toulouse, France, May 2006, pp. 101–104.

[73] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 34013, pp. 1–12, Dec. 2007.

[74] W. Zhang, E. A. P. Habets, and P. A. Naylor, "On the use of channel shortening in multichannel acoustic system equalization," in *Proc. 2010 Intl. Workshop Acoust. Echo Noise Control (IWAENC 2010)*, Tel Aviv, Isreal, Sep. 2010.

[75] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 9, pp. 1879–1890, Sep. 2013.

[76] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-domain single-channel inverse filtering for speech dereverberation: Theory and practice," in *Proc. 2014 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2014)*, Florence, Italy, May 2014, pp. 5177–5181.

[77] F. Lim, W. Zhang, E. A. P. Habets, and P. A. Naylor, "Robust multichannel dereverberation using relaxed multichannel least squares," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 9, pp. 1379–1390, Sep. 2014.

[78] I. Kodrasi and S. Doclo, "Signal-dependent penalty functions for robust acoustic multi-channel equalization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 7, pp. 1512–1525, July 2017.

[79] X. Li, L. Girin, S. Gannot, and R. Horaud, "Multichannel speech separation and enhancement using the convolutive transfer function," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 3, pp. 645–659, Mar. 2019.

[80] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP J. Adv. Signal Process.*, vol. 2003, no. 11, pp. 1074–1090, Oct. 2003.

[81] J. Benesty, J Chen, Y. Huang, and J. Dmochowski, "On microphone-array beamforming from a MIMO acoustic signal processing persepective," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1053–1063, Mar. 2007.

[82] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the MVDR beamformer in room acoustics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 158–170, Jan. 2010.

[83] E. A. P. Habets and J. Benesty, "A two-stage beamforming approach for noise reduction and dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 5, pp. 945–958, May 2013.

[84] O. Thiergart, M. Taseska, and E. A. P. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2182–2196, Dec 2014.

[85] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 2, pp. 240–251, Feb. 2015.

[86] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP J. Adv. Signal Process.*, vol. 2015, no. 61, pp. 1–12, July 2015.

[87] O. Schwartz, S. Gannot, and E. A. P. Habets, "Nested generalized sidelobe canceller for joint dereverberation and noise reduction," in *Proc. 2015 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2015)*, Brisbane, Australia, Apr. 2015, pp. 106–110.

[88] M. Triki and D. T. M. Slock, "Blind dereverberation of a single source based on multichannel linear prediction," in *Proc. 2005 Intl. Workshop Acoustic Echo Noise Control (IWAENC 2005)*, Eindhoven, Netherlands, Sep. 2005, pp. 173–176.

[89] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multichannel linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 430–440, Jan. 2007.

[90] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, "Blind speech dereverberation with multi-channel linear prediction based

on short time Fourier transform representation," in *Proc. 2008 Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2008)*, Las Vegas, NV, USA, Apr. 2008, pp. 85–88.

[91] T. Nakatani, B. H. Juang, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, "Speech dereverberation based on maximum-likelihood estimation with time-varying Gaussian source model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1512–1527, Nov. 2008.

[92] T. Yoshioka, H. Tachibana, T. Nakatani, and M. Miyoshi, "Adaptive dereverberation of speech signals with speaker-position change detection," in *Proc. 2009 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2009)*, Taipei, Taiwan, Apr. 2009, pp. 3733–3736.

[93] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, Aug. 2010.

[94] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 69–84, Jan. 2011.

[95] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, July 2012.

[96] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1509–1520, June 2015.

[97] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Group sparsity for MIMO speech dereverberation," in *IEEE Workshop Appl. Signal Process. Audio, Acoust. (WASPAA 2015)*, New Paltz, NY, USA, Oct. 2015, pp. 1–5.

[98] T. Dietzen, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, "Partitioned block frequency domain Kalman filter for multi-channel linear prediction based blind speech dereverberation," in *Proc. 2016 Intl. Workshop Acoustic Signal Enhancement (IWAENC 2016)*, Xi'An, China, Sep. 2016, pp. 1–5.

[99] S. Braun and E. A. P. Habets, "Online dereverberation for dynamic scenarios using a Kalman filter with an autoregressive model," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1741–1745, Dec. 2016.

[100] A. Jukić, T. van Waterschoot, and S. Doclo, "Adaptive speech dereverberation using constrained sparse multichannel linear prediction," *IEEE Signal Process. Lett.*, vol. 24, no. 1, pp. 101–105, Jan. 2017.

[101] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online WPE dereverberation," in *Proc. 18th Annual Conf. Intl. Speech Comm. Assoc. (INTERSPEECH 2018)*, Stockholm, Sweden, Aug. 2017, pp. 384–388.

[102] T. Dietzen, S. Doclo, A. Spriet, W. Tirry, M. Moonen, and T. van Waterschoot, "Low complexity Kalman filter for multi-channel linear prediction based blind speech dereverberation," in *Proc. 2017 IEEE Workshop Appl. Signal Process. Audio, Acoust. (WASPAA 2017)*, New Paltz, NY, USA, Oct. 2017.

[103] J. Heymann, L. Drude, R. Haeb-Umbach, K. Kinoshita, and T. Nakatani, "Frame-online DNN-WPE dereverberation," in *Proc. 2018 Intl. Workshop Acoustic Signal Enhancement (IWAENC 2018)*, Tokyo, Japan, Sep. 2018, pp. 466–470.

[104] Sharon Gannot and Marc Moonen, "On the application of the unscented Kalman filter to speech processing," in *Proc. 2003 Intl. Workshop Acoustic Echo Noise Control (IWAENC 2003)*, Kyoto, Japan, Sep. 2003, pp. 27–30.

[105] C. Evers and J. R. Hopgood, "Multichannel online blind speech dereverberation with marginalization of static observation parameters in a Rao-Blackwellized particle filter," *J. Signal Process. Systems*, vol. 63, no. 3, pp. 315–332, June 2011.

[106] D. Schmid, G. Enzner, S. Malik, D. Kolossa, and R. Martin, "Variational bayesian inference for multichannel dereverberation and noise reduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 8, pp. 1320–1335, Aug. 2014.

[107] B. Schwartz, S. Gannot, and E. A. P. Habets, "Online speech dereverberation using Kalman filter and EM algorithm," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 2, pp. 394–406, Feb 2015.

[108] F. Lim, M. R. P. Thomas, and P. A. Naylor, "MINTFormer: A spatially aware channel equalizer," in *IEEE Workshop Appl. Signal Process. Audio, Acoust. (WASPAA 2013)*, New Paltz, NY, USA, Oct. 2013, pp. 1–4.

[109] M. R. P. Thomas, I. J. Tashev, F. Lim, and P. A. Naylor, "Optimal beamforming as a time domain equalization problem with application to room acoustics," in *Proc. 2014 Intl. Workshop Acoustic Signal Enhancement (IWAENC 2014)*, Juan-les-Pins, France, Sep. 2014, pp. 75–79.

[110] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multi-channel equalization," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 24, no. 4, pp. 680–693, Apr. 2016.

[111] T. Yoshioka, "Dereverberation for reverberation-robust microphone arrays," in *Proc. 21st European Signal Process. Conf. (EUSIPCO 2013)*, Marrakech, Morocco, Sep. 2013, pp. 1–5.

[112] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, S. Araki, T. Hori, and T. Nakatani, "Strategies for distant speech recognition in reverberant environments," *EURASIP J. Adv. Signal Process.*, vol. 2015, no. 60, pp. 1–15, July 2015.

[113] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W.J. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. 2015 IEEE Workshop Autom. Speech Recog., Underst. (ASRU)*, Scottsdale, AZ, USA, Dec. 2015, pp. 436–443.

[114] L. Drude, C. Boeddeker, J. Heymann, R. Haeb-Umbach, K. Kinoshita, M. Delcroix, and T. Nakatani, "Integrating neural network based beamforming and weighted prediction error dereverberation," in *Proc. 19th Annual Conf. Intl. Speech Comm. Assoc. (INTERSPEECH 2018)*, Hyderabad, India, Sep. 2018, pp. 3043–3047.

[115] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Process. Lett.*, vol. 26, no. 6, pp. 903–907, June 2019.

[116] S. Braun and E. A. P. Habets, "Linear prediction-based online dereverberation and noise reduction using alternating Kalman filters," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 240–251, June 2018.

[117] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, Apr. 2007.

[118] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 8, no. 3, pp. 267–281, May 2000.

[119] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. 2001 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2001)*, Salt Lake City, UT, USA, May 2001, pp. 3701–3704.

[120] Nikolay D Gaubitch, Patrick A Naylor, and Darren B Ward, "On the use of linear prediction for dereverberation of speech," in *Proc. 2003 Intl. Workshop Acoust. Echo Noise Control (IWAENC 2003)*, Kyoto, Japan, Sep. 2003, pp. 99–102.

[121] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proc. 2009 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2009)*, Taipei, Taiwan, Apr. 2009, pp. 45–48.

[122] Kshitiz Kumar, Rita Singh, Bhiksha Raj, and Richard Stern, "Gammatone sub-band magnitude-domain dereverberation for asr," in *Proc. 2011 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2011)*, Prague, Czech Republic, May 2011, pp. 4604–4607.

[123] N. Mohammadiha and S. Doclo, "Speech dereverberation using non-negative convolutive transfer function and spectro-temporal modeling," *IEEE/ACM Trans.Audio, Speech, Lang. Process.*, vol. 24, no. 2, pp. 276–289, Feb. 2016.

[124] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *Proc. 2014 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2014)*, Florence, Italy, May 2014, pp. 1759–1763.

[125] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 982–992, June 2015.

[126] Z. Zhang, L. Wang, A. Kai, T. Yamada, W. Li, and M. Iwahashi, "Deep neural network-based bottleneck feature and denoising autoencoder-based dereverberation for distant-talking speaker identification," *EURASIP J. Audio, Speech, Music Process.*, vol. 2015, no. 12, pp. 1–13, Dec. 2015.

[127] X. Xiao, S. Zhao, D. H. H. Nguyen, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "Speech dereverberation for enhancement and recognition

using dynamic features constrained deep neural networks and feature adaptation," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 4, pp. 1–18, Jan. 2016.

[128] B. Wu, K. Li, M. Yang, and C. H. Lee, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 102–111, Jan. 2017.

[129] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1218–1234, July 2006.

[130] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Comm.*, vol. 49, no. 7-8, pp. 636–656, July/Aug. 2007.

[131] J. D. Polack, *La transmission de l'énergie sonore dans les salles*, Ph.D. thesis, Université du Maine, Le Mans, France, 1988.

[132] F. Jacobsen and T. Roisin, "The coherence of reverberant sound fields," *J. Acoust. Soc. Amer.*, vol. 108, no. 1, pp. 204–210, July 2000.

[133] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer Handbook of Speech Processing*, pp. 945–978. Springer, 2007.

[134] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*, Springer, 2008.

[135] N. Madhu, *Acoustic Source Localization: Algorithms, Applications and Extensions to Source Separation*, Ph.D. thesis, Ruhr-Universität Bochum, Germany, 2009.

[136] Pierre Comon and Christian Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*, Academic Press, 2010.

[137] G. Xu, H. Liu, L. Tong, and T Kailath, "A least-squares approach to blind channel identification," *IEEE Trans. Signal Process.*, vol. 43, no. 12, pp. 2982–2993, Dec. 1995.

[138] Y. A. Huang and J. Benesty, "Adaptive multi-channel least mean square and newton algorithms for blind channel identification," *Signal Process.*, vol. 82, no. 8, pp. 1127–1138, Aug. 2002.

[139] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[140] M. A. Haque and M. K. Hasan, "Noise robust multichannel frequency-domain lms algorithms for blind channel identification," *IEEE Signal Process. Lett.*, vol. 15, pp. 305–308, Feb. 2008.

[141] S. Malik, D. Schmid, and G. Enzner, "A state-space cross-relation approach to adaptive blind SIMO system identification," *IEEE Signal Process. Lett.*, vol. 19, no. 8, pp. 511–514, Aug. 2012.

[142] M. R. P. Thomas, N. D. Gaubitch, E. A. P. Habets, and P. A. Naylor, "An insight into common filtering in noisy SIMO blind system identification," in *Proc. 2012 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2012)*, Kyoto, Japan, Mar. 2012, pp. 521–524.

[143] F. Lim and P. A. Naylor, "Statistical modelling of multichannel blind system identification errors," in *Proc. 2014 Intl. Workshop Acoustic Signal Enhancement (IWAENC 2014)*, Juan-les-Pins, France, Sep. 2014, pp. 119–123.

[144] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Ant. Prop.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[145] J. Scheuing and B. Yang, "Correlation-based TDOA-estimation for multiple sources in reverberant environments," in *Speech and Audio Processing in Adverse Environments*, pp. 381–416. Springer, 2008.

[146] Z. Chen, G. Gokeda, and Y. Yu, *Introduction to Direction-of-arrival Estimation*, Artech House, 2010.

[147] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[148] A. Bertrand and M. Moonen, "Distributed node-specific LCMV beamforming in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 233–246, Jan. 2012.

[149] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proc. 2015 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2015)*, Brisbane, QLD, Australia, Apr. 2015, pp. 544–548.

[150] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 4, pp. 785–799, Apr. 2014.

[151] Y. A. Huang, A. Luebs, J. Skoglund, and W. B. Kleijn, "Globally optimized least-squares post-filtering for microphone array speech enhancement," in *Proc. 2016 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2016)*, Mar. 2016, pp. 380–384.

[152] T. Dietzen, "GitHub repository: Integrated sidelobe cancellation and linear prediction Kalman filter for joint multi-microphone speech dereverberation, interfering speech cancellation, and noise reduction," `https://github.com/tdietzen/ISCLP-KF`, July 2019.

[153] T. Dietzen, "GitHub repository: square root-based multi-source early PSD estimation and recursive RETF update in reverberant environments by means of the orthogonal Procrustes problem," `https://github.com/tdietzen/SQRT-PSD-RETF`, July 2019.

[154] T. Dietzen, "GitHub repository: Low-complexity ISCLP Kalman filters," `https://github.com/tdietzen/ISCLP-KF/tree/LC-ISCLP-KF`, July 2019.

[155] T. Dietzen, N. Huleihel, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, "Speech dereverberation by data-dependent beamforming with signal pre-whitening," in *Proc. 23rd European Signal Process. Conf. (EUSIPCO 2015)*, Nice, France, Aug. 2015, pp. 2461–2465.

[156] T. Dietzen, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, "On the relation between data-dependent beamforming and multichannel linear prediction for dereverberation," in *Proc. AES 60th Intl. Conf. Dereverb. Reverb. Audio, Music, Speech (DREAMS)*, Leuven, Belgium, Jan. 2016, pp. 1–8.

[157] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Joint multi-microphone speech dereverberation and noise reduction using integrated sidelobe cancellation and linear prediction," in *Proc. 2018 Intl. Workshop Acoustic Signal Enhancement (IWAENC 2018)*, Tokyo, Japan, Sep. 2018, pp. 221–225.

[158] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Ant. Prop.*, vol. 1, no. 30, pp. 27–34, Jan. 1982.

[159] Carl D. Meyer, Ed., *Matrix analysis and applied linear algebra*, SIAM, 2000.

[160] Yi Hu and P C Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, 2008.

[161] E. De Sena, N. Antonello, M. Moonen, and T. van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 4, pp. 774–786, Apr. 2015.

[162] T. Dietzen, "Audio examples for IEEE/ACM TASLP 2018," `ftp://ftp.esat.kuleuven.be/pub/SISTA/tdietzen/reports/taslp18/audio`, Aug. 2018.

[163] T. Dietzen, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, "Comparative analysis of generalized sidelobe cancellation and multi-channel linear prediction for speech dereverberation and noise reduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 3, pp. 544–558, Mar. 2019.

[164] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Square root-based multi-source early PSD estimation and recursive RETF update in reverberant environments by means of the orthogonal Procrustes problem," ESAT-STADIUS Tech. Rep. TR 19-69, KU Leuven, Belgium, submitted for publication, June 2019.

[165] G. Enzner and P. Vary, "Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones," *Signal Process.*, vol. 86, no. 6, pp. 1140–1156, June 2006.

[166] N. Dal Degan and C. Prati, "Acoustic noise analysis and speech enhancement techniques for mobile radio applications," *Signal Process.*, vol. 15, pp. 43–56, July 1988.

[167] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Amer.*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.

[168] Intl. Telecommun. Union, "Perceptual evaluation of of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," in *ITU-T Recommendation P.862*, Feb. 2001.

[169] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[170] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[171] H. Q. Dam, S. Y. Low, H. H. Dam, and S. Nordholm, "Space constrained beamforming with source PSD updates," in *Proc. 2004 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2004)*, Montreal, QC, Canada, May 2004, pp. 93–96.

[172] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[173] R. Everson, "Orthogonal, but not orthonormal, Procrustes problems," Tech. Rep., Laboratory for Applied Mathematics, City University New York and Mount Sinai Medical School, NYC, USA, Mar. 1997.

[174] P. H. Schönemann, "A generalized solution of the orthogonal Procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, Mar. 1966.

[175] J. H. Manton, "Optimization algorithms exploiting unitary constraints," *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 635–650, Mar. 2002.

[176] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[177] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, Jan. 2014.

[178] N. Antonello, L. Stella, P. Patrinos, and T. van Waterschoot, "Proximal gradient algorithms: Applications in signal processing," ESAT-STADIUS Tech. Rep. TR 17-112, KU Leuven, Belgium, Jan. 2018.

[179] T. van Waterschoot, G. Rombouts, and M. Moonen, "Optimally regularized adaptive filtering algorithms for room acoustic signal enhancement," *Signal Process.*, vol. 88, no. 3, pp. 594–611, Mar. 2008.

[180] L. Ljung and T. Söderström, *Theory and practice of recursive identification*, MIT press, 1986.

[181] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Integrated sidelobe cancellation and linear prediction Kalman filter for joint multi-microphone dereverberation, interfering speech cancellation, and noise reduction," ESAT-STADIUS Tech. Rep. TR 19-70, KU Leuven, Belgium, submitted for publication, June 2019.

[182] G. H. Golub and C. F. Van Loan, *Matrix computations*, Johns Hopkins University Press, 2012.

[183] M. Tammen, I. Kodrasi, and S. Doclo, "Complexity reduction of eigenvalue decomposition-based diffuse power spectral density estimators using the power method," in *Proc. 2018 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2018)*, Calgary, AB, Canada, Apr. 2018, pp. 451–455.

[184] M. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5-2, pp. 845–856, Sep. 2005.

[185] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model," *EURASIP J. Adv. Signal Process.*, vol. 2005, no. 7, pp. 1110–1126, May 2005.

[186] G. Carayannis, D. Manolakis, and N. Kalouptsidis, "A fast sequential algorithm for least-squares filtering and prediction," *IEEE Trans. on Acoust., Speech, Signal Process.*, vol. 31, no. 6, pp. 1394–1402, Dec. 1983.

[187] J. G. Proakis, C. M. Radar, F. Ling, C. L. Nikias, M. Moonen, and I. K. Proudler, *Algorithms for statistical signal processing*, Prentice Hall, 2002.

[188] J. A. Apolinário, *QRD-RLS adaptive filtering*, Springer, 2009.

[189] G. Enzner, "Baysian inference model for applications of time-varying acoustic system identification," in *Proc. 18th European Signal Process. Conf. (EUSIPCO 2010)*, Aalborg, Denmark, Aug. 2010, pp. 1–5.

[190] F. Kuech, E. Mabande, and G. Enzner, "State-space architecture of the partitioned-block-based acoustic echo controller," in *Proc. 2014 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2014)*, Florence, Italy, July 2014, pp. 1295–1299.

[191] M. L. Valero, E. Mabande, and E. A. P. Habets, "A state-space partitioned-block adaptive filter for echo cancellation using inter-band correlations in the Kalman gain computation," in *Proc. 2015 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2015)*, Brisbane, Australia, Apr. 2015, pp. 599–603.

[192] G. Bernardi, T. van Waterschoot, J. Wouters, and M. Moonen, "Adaptive feedback cancellation using a partitioned-block frequency-domain Kalman filter approach with PEM-based signal prewhitening," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 9, pp. 1784–1798, Sep. 2017.

[193] D. P. Mandic, S. Kanna, and A. G. Constantinides, "On the intrinsic relationship between the Least Mean Square and Kalman filters [lecture notes]," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 117–122, Nov. 2015.

[194] Z. Chen, S. L. Gay, and S. Haykin, "Proportionate adaptation: New paradigms in adaptive filters," in *Least-Mean-Square Adaptive Filters*, pp. 293–334. Wiley, 2003.

[195] V. Hamacher, U. Kornagel, T. Lotter, and H. Puder, "Binaural signal processing in hearing aids: Technologies and algorithms," in *Advances in digital speech transmission*, pp. 401–429. Wiley, 2008.

[196] M. Jeub, M. Schafer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1732–1745, Sep. 2010.

[197] 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.*, vol. 19, no. 5, pp. 1368–1381, July 2011.

[198] D. Marquardt, *Development and evaluation of psychoacoustically motivated binaural noise reduction and cue preservation techniques*, Ph.D. thesis, Carl von Ossietzky Universität Oldenburg, 2016.

[199] T. van Waterschoot, B. Defraene, M. Diehl, and M. Moonen, "Embedded optimization algorithms for multi-microphone dereverberation," in *Proc. 21th European Signal Process. Conf. (EUSIPCO 2013)*, Marrakech, Morocco, Sep. 2013, pp. 1–5.

[200] Intl. Telecommun. Union, "Method for the subjective assessment of intermediate sound quality (MUSHRA)," in *ITU-T Recommendation BS.1534-3*, Oct. 2015.

[201] Intl. Telecommun. Union, *Measuring the Information Society Report Volume 1*, ITU Publications, 2018.

[202] Growth from Knowledge, "Global smartphone sales reached $522 billion in 2018," `https://www.gfk.com/insights/press-release/global-smartphone-sales-reached-522-billion-in-2018`, Feb. 2019.

[203] Future Market Insights, "VOIP services market: Global industry analysis and opportunity assessment 2015 - 2025," `https://www.futuremarketinsights.com/reports/global-voip-services-market`, to appear, Sep. 2019.

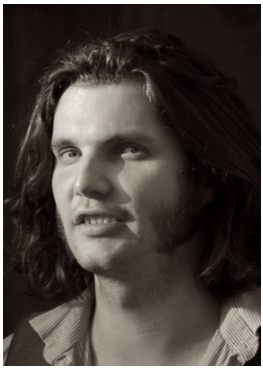
[204] Bizwit Research & Consulting LLP, "Global hearing aid market size study, by product (hearing aid devices and hearing implants) type of hearing loss (conductive hearing loss and sensorineural hearing loss), type of patient (adult and pediatric) and regional forecasts 2018

- 2025," https://www.wiseguyreports.com/reports/3741231-global-hearing-aid-market-size-study-by-product, Feb. 2019.

[205] World Health Organization, "Deafness and hearing loss," https://www.who.int/en/news-room/fact-sheets/detail/deafness-and-hearing-loss, Mar. 2019.

[206] Grand View Research, "Voice and speech recognition market size, share & trends analysis report, by function, by technology (AI, non-AI), by vertical (healthcare, BFSI, automotive), and segment forecasts, 2018 - 2025," https://www.grandviewresearch.com/industry-analysis/voice-recognition-market, Nov. 2018.

Curriculum Vitae



Thomas Dietzen received his Dipl.-Ing. degree from Kaiserslautern University, Germany, in 2011. Between 2012 and 2014, he was a research assistant at University of Heidelberg, Germany, and at Fraunhofer Institute for Integrated Circuits IIS, Germany. From 2014 to 2017, he has been a doctoral researcher at NXP Semiconductors Belgium NV, Belgium, in the frame of the FP7-PEOPLE Marie Curie ITN 'Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS)'. Currently, he pursues his PhD at KU Leuven, Belgium. His research is focused on signal processing algorithms for microphone array-based speech enhancement in adverse acoustic conditions, specifically on power-spectral-density estimation and spatio-temporal adaptive filtering for dereverberation, interfering speech cancellation and noise reduction. He has served as a reviewer for the IEEE/ACM Transactions on Audio, Speech, and Language Processing and the IEEE Signal Processing Letters.

List of Publications

International Journal Papers

1. **T. Dietzen**, S. Doclo, M. Moonen, and T. van Waterschoot, "Square root-based multi-source early PSD estimation and recursive RETF update in reverberant environments by means of the orthogonal Procrustes problem," ESAT-STADIUS Tech. Rep. TR 19-69, KU Leuven, Belgium, submitted for publication, June 2019.

2. **T. Dietzen**, S. Doclo, M. Moonen, and T. van Waterschoot, "Integrated sidelobe cancellation and linear prediction Kalman filter for joint multi-microphone dereverberation, interfering speech cancellation, and noise reduction," ESAT-STADIUS Tech. Rep. TR 19-70, KU Leuven, Belgium, submitted for publication, June 2019.

3. **T. Dietzen**, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, "Comparative analysis of generalized sidelobe cancellation and multi-channel linear prediction for speech dereverberation and noise reduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 3, pp. 544–558, Mar. 2019.

International Conference Papers

1. **T. Dietzen**, S. Doclo, M. Moonen, and T. van Waterschoot, "Joint multi-microphone speech dereverberation and noise reduction using integrated sidelobe cancellation and linear prediction," in *Proc. 2018 Intl. Workshop Acoustic Signal Enhancement (IWAENC 2018)*, Tokyo, Japan, Sep. 2018,

pp. 221–225.

2. **T. Dietzen**, S. Doclo, A. Spriet, W. Tirry, M. Moonen, and T. van Waterschoot, "Low complexity Kalman filter for multi-channel linear prediction based blind speech dereverberation," in *Proc. 2017 IEEE Workshop Appl. Signal Process. Audio, Acoust. (WASPAA 2017)*, New Paltz, NY, USA, Oct. 2017.

3. **T. Dietzen**, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, "Partitioned block frequency domain Kalman filter for multi-channel linear prediction based blind speech dereverberation," in *Proc. 2016 Intl. Workshop Acoustic Signal Enhancement (IWAENC 2016)*, Xi'An, China, Sep. 2016, pp. 1–5.

4. **T. Dietzen**, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, "On the relation between data-dependent beamforming and multichannel linear prediction for dereverberation," in *Proc. AES 60th Intl. Conf. Dereverb. Reverb. Audio, Music, Speech (DREAMS)*, Leuven, Belgium, Jan. 2016, pp. 1–8.

5. **T. Dietzen**, N. Huleihel, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, "Speech dereverberation by data-dependent beamforming with signal pre-whitening," in *Proc. 23rd European Signal Process. Conf. (EUSIPCO 2015)*, Nice, France, Aug. 2015, pp. 2461–2465.



FACULTY OF ENGINEERING TECHNOLOGY
DEPARTMENT OF ELECTRICAL ENGINEERING
STADIUS CENTER FOR DYNAMICAL SYSTEMS, SIGNAL PROCESSING AND DATA ANALYTICS
Kasteelpark Arenberg 10
B-3001 Leuven
TECHNOLOGY CLUSTER ELECTRICAL ENGINEERING
Andreas Vesaliusstraat 13
B-3000 Leuven
thomas.dietzen@esat.kuleuven.be