DEREVERBERATION AND NOISE REDUCTION TECHNIQUES BASED ON ACOUSTIC MULTI-CHANNEL EQUALIZATION

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

von
Ina Kodrasi
geboren am 06. April 1987
in Elbasan (Albanien)

Ina Kodrasi: *Dereverberation and Noise Reduction Techniques based on Acoustic Multi-Channel Equalization*

ERSTGUTACHTER:
Prof. Dr. ir. Simon Doclo

WEITERE GUTACHTER:
Prof. Dr.-Ing. Timo Gerkmann
Dr. Patrick A. Naylor

TAG DER DISPUTATION:
14. Dezember 2015

ACKNOWLEDGMENTS

This thesis has been written at the Signal Processing Group in the Department of Medical Physics and Acoustics of the Carl von Ossietzky Universität Oldenburg in Oldenburg, Germany. I would like to take the opportunity to thank the many people who contributed to the completion of this work.

First, I would like to express my sincere gratitude to my supervisor Simon Doclo for his continuous support, his numerous ideas and suggestions, the freedom to pursue my scientific interests, and the invaluable guidance when this pursuit faltered.

Furthermore, I would like to thank Patrick Naylor and Timo Gerkmann for participating in my thesis committee and for showing much interest in my work, as well as Steven van de Par for chairing the thesis committee.

Special thanks go to all my current and former colleagues at the Signal Processing Group for providing a friendly and pleasant work environment, for numerous interesting and fruitful discussions, and for many enjoyable evenings at home and at conferences. In particular, I would like to thank Daniel Marquardt for the endless (non)scientific discussions and his intuitive explanations of signal processing and life.

I would also like to thank Stefan Goetze, Benjamin Cauchi, and Andreas Volgenandt from the Fraunhofer Institute for Digital Media and Technology for many interesting discussions, successful collaborations, and the support with measurements.

Last but not least, I would like to express my deep gratitude to my family and my friends for their continuous support and encouragement.

Për mamin edhe babin: pa mbështetjen dhe inkurajimin tuaj ky punim nuk do të kishte qenë i mundur.

Oldenburg, March 2016
Ina Kodrasi

ABSTRACT

In many hands-free speech communication applications such as teleconferencing or voice-controlled applications, the recorded microphone signals do not only contain the desired speech signal, but also attenuated and delayed copies of the desired speech signal due to reverberation as well as additive background noise. Reverberation and background noise cause a signal degradation which can impair speech intelligibility and decrease the performance for many signal processing techniques.

Acoustic multi-channel equalization techniques, which aim at inverting or reshaping the measured or estimated room impulse responses between the speech source and the microphone array, comprise an attractive approach to speech dereverberation since in theory perfect dereverberation can be achieved. However in practice, such techniques suffer from several drawbacks, such as uncontrolled perceptual effects, sensitivity to perturbations in the measured or estimated room impulse responses, and background noise amplification. The aim of this thesis is to tackle these drawbacks by designing perceptually advantageous and robust acoustic multi-channel equalization techniques for speech dereverberation as well as for joint dereverberation and noise reduction.

First, in order to control the perceptual speech quality, we propose the *perceptually advantageous* partial multi-channel equalization technique based on the multiple-input/output inverse theorem (PMINT), which aims not only at suppressing the late reflections but also at controlling the early reflections. Simulation results show that the proposed PMINT technique results in a better perceptual speech quality than state-of-the-art acoustic multi-channel equalization techniques, such as the multiple-input/output inverse theorem (MINT), channel shortening (CS), and relaxed multi-channel least-squares (RMCLS).

Second, in order to *increase the robustness* of all considered acoustic multi-channel equalization techniques against room impulse response perturbations, i.e., of the MINT, CS, RMCLS, and PMINT techniques, we propose several methods. On the one hand, we propose *signal-independent methods*, i.e., decreasing the reshaping filter length to improve the conditioning of the optimization criteria or incorporating (automatic) regularization to reduce the energy of distortions due to room impulse response perturbations. On the other hand, we propose a *signal-dependent method*, i.e., using a sparsity-promoting penalty function to promote sparsity in the output speech signal and reduce artifacts generated by non-robust techniques. All proposed methods are validated using instrumental performance measures and subjective listening tests, which show that the regularized and sparsity-promoting extensions of the PMINT technique yield the best dereverberation performance in comparison to the robust extensions of state-of-the-art acoustic multi-channel equalization techniques.

Finally, in order to achieve *joint dereverberation and noise reduction* we propose two techniques based on robust acoustic multi-channel equalization. The first technique, namely regularized PMINT for joint dereverberation and noise reduction (RP-DNR), can be seen as an extension of the regularized PMINT technique that explicitly takes the noise statistics into account. The second technique, namely multi-channel Wiener filter for joint dereverberation and noise reduction (MWF-DNR), in addition takes the speech statistics into account and uses the dereverberated output signal of the regularized PMINT technique as the reference signal for the multi-channel Wiener filter. In addition to the regularization parameter used in the regularized PMINT technique, a weighting parameter is introduced in the RP-DNR and MWF-DNR techniques to trade off between dereverberation and noise reduction. To determine the regularization and weighting parameters, we propose automatic non-intrusive procedures based on the L-hypersurface and the L-curve. Simulation results show that the RP-DNR technique maintains the high dereverberation performance of the regularized PMINT technique while improving the noise reduction performance. Furthermore, simulation results show that the MWF-DNR technique yields a significantly better noise reduction performance than the RP-DNR technique at the expense of a worse dereverberation performance, depending on the amount of estimation errors in the speech correlation matrix.

ZUSAMMENFASSUNG

Die Mikrofonsignale vieler Freichsprechanwendungen, beispielsweise im Bereich Telekonferenz oder Sprachsteuerung, beinhalten nicht nur das gewünschte Sprachsignal, sondern auch Nachhall sowie additives Hintergrundrauschen. Dieser Nachhall und das Hintergrundrauschen führen zu Signalverschlechterungen, welche die Sprachverständlichkeit sowie die Leistungsfähigkeit verschiedener Algorithmen zur Signalverarbeitung beeinträchtigen.

Ein attraktiver Ansatz Mikrofonsignale zu enthallen stellen mehrkanalige akustische Entzerrungsverfahren dar, da diese, zumindest theoretisch, eine perfekte Enthallung erreichen können. Diese Verfahren zielen auf eine Invertierung oder Angleichung der gemessenen oder geschätzten Raumimpulsantworten ab. In der praktischen Anwendung dieser Verfahren ergeben sich jedoch verschiedene Probleme. Dazu zählen beispielsweise die Erzeugung unkontrollierter wahrnehmbarer Artefakte, eine hohe Anfälligkeit gegenüber Abweichungen der gemessenen oder geschätzten Raumimpulsantworten, sowie eine Verstärkung des Hintergrundrauschens. Der Schwerpunkt der folgenden Arbeit stellt das Lösen dieser Probleme dar. Dazu werden verschiedene akustische Mehrkanal-Entzerrungsverfahren sowohl zur Sprachenthallung als auch zur gleichzeitigen Enthallung und Rauschunterdrückung entwickelt, die sich robust verhalten sowie wahrnehmungsbasierte Vorteile bieten.

Um die wahrgenommene Sprachqualität zu steuern, wird zunächst ein unter den Gesichtspunkten der Wahrnehmung vorteilhafter partieller mehrkanaliger Entzerrungsansatz vorgeschlagen (bezeichnet als PMINT), der auf dem so-genannten „multiple-input/output inverse theorem“ (MINT) basiert. Dieser Ansatz versucht gleichzeitig die späten Raumreflexionen zu unterdrücken und die frühen Raumreflexionen zu kontrollieren. Simulationsergebnisse zeigen, dass der vorgeschlagene PMINT-Ansatz eine bessere Sprachqualität als bisherige Mehrkanal-Entzerrungsverfahren liefert, die auf MINT, dem sogenannten „channel shortening“ (CS) oder dem sogenannten „relaxed multi-channel least-squares“ (RMCLS) Ansatz basieren.

Um die Robustheit der betrachteten akustischen Mehrkanal-Entzerrungsverfahren, beispielsweise MINT, PMINT, CS oder RMCLS, gegenüber Abweichungen der Raumimpulsantworten zu verbessern, werden verschiedene Ansätze vorgestellt. Zu diesem Zweck werden zunächst signalunabhängige Methoden betrachtet. Dazu zählen eine Verringerung der Länge des Angleichungsfilters um die Konditionierung des Optimierungskriteriums zu verbessern sowie die Einbeziehung einer (automatischen) Regularisierung um die Signalverzerrungen zu reduzieren. Anschließend werden signalabhängige Methoden betrachtet. Dazu zählen die Verwendung einer Straffunktion, die dünnbesetzte Ausgangssprachsignale fördert und dadurch Artefakte verringert, die durch Algorithmen mit geringer Robustheit erzeugt werden.

Die vorgeschlagenen signalunabhängigen und signalabhängigen Methoden werden mittels geeigneter Leistungsmaße sowie subjektiver Hörtests evaluiert. Die Ergebnisse zeigen, dass der PMINT-Ansatz die beste Enthallungsleistung im Vergleich zu den robusten Erweiterungen bisheriger akustischer Mehrkanal-Entzerrungsverfahren bietet, wenn der PMINT-Ansatz mit der vorgeschlagenen Straffunktion und Regularisierung erweitert wird.

Um kombinierte Enthallung und Störgeräuschunterdrückung zu erreichen, werden basierend auf robusten akustischen Mehrkanal-Entzerrungsverfahren zwei Algorithmen vorgestellt. Der erste Algorithmus, welcher als regularisiertes PMINT zur gleichzeitigen Enthallung und Störgeräuschunterdrückung (RP-DNR) bezeichnet wird, kann als eine Erweiterung des regularisiertem PMINT Algorithmus betrachtet werden, welche die Statistik des Störgeräusches explizit berücksichtigt. Der zweite Algorithmus, bezeichnet als mehrkanaliges Wiener Filter zur gleichzeitigen Enthallung und Störgeräuschunterdrückung (MWF-DNR), berücksichtigt zusätzlich die Statistik des Sprachsignals und verwendet das enthallte Ausgangssignal des regularisierten PMINT Algorithmus als Referenzsignal für das mehrkanaligen Wiener Filter. Zusätzlich zu dem Regularisierungsparameter, welcher im regularisierten PMINT verwendet wird, wird für die RP-DNR und MWF-DNR Algorithmen ein Gewichtungsparameter eingeführt, welcher eine Gewichtung zwischen Enthallung und Störgeräuschunterdrückung ermöglicht. Zur Bestimmung beider Parameter werden automatische Verfahren basierend auf L-Hyperflächen und L-Kurven vorgestellt. Die Simulationsergebnisse zeigen, dass der RP-DNR Algorithmus denselben Grad an Enthallung wie der regularisierte PMINT Algorithmus erreicht und gleichzeitig die Störgeräuschunterdrückung verbessert. Zusätzlich zeigen die Ergebnisse, dass der MWF-DNR Algorithmus, abhängig von dem Grad der Schätzfehler in der Sprachkorrelationsmatrix, eine signifikant bessere Störgeräuschunterdrückung auf Kosten einer schlechteren Enthallungsleistung aufweist.

GLOSSARY

Acronyms and abbreviations

ADMM	alternating direction method of multipliers
AIR	acoustic impulse response
BSI	blind system identification
CAPZ	common-acoustical-poles-and-zeros
CD	cepstral distance
CS	channel shortening
DRR	direct-to-reverberant ratio
DTFT	discrete-time Fourier transform
EDC	energy decay curve
EIR	equalized impulse response
FIR	finite impulse response
fwSSNR	frequency-weighted segmental signal-to-noise ratio
GSC	generalized sidelobe canceller
IDTFT	inverse discrete-time Fourier transform
IIR	infinite impulse response
ISTFT	inverse short-time Fourier transform
L-CS	channel shortening using a shorter reshaping filter length
L-MINT	multiple-input/output inverse theorem using a shorter reshaping filter length
L-PMINT	partial multi-channel equalization based on the multiple-input/output inverse theorem using a shorter reshaping filter length
L-RMCLS	relaxed multi-channel least-squares using a shorter reshaping filter length
MINT	multiple-input/output inverse theorem
MVDR	minimum variance distortionless response
MUSHRA	multiple stimuli with hidden reference and anchor
MWF	multi-channel Wiener filter

MWF-DNR	multi-channel Wiener filter for joint dereverberation and noise reduction
NPM	normalized projection misalignment
PESQ	perceptual evaluation of speech quality
PMINT	partial multi-channel equalization based on the multiple-input/output inverse theorem
PSD	power spectral density
R-CS	regularized channel shortening
RIR	room impulse response
RMCLS	relaxed multi-channel least-squares
R-MINT	regularized multiple-input/output inverse theorem
RP-DNR	regularized partial multi-channel equalization based on the multiple-input/output inverse theorem for joint dereverberation and noise reduction
R-PMINT	regularized partial multi-channel equalization based on the multiple-input/output inverse theorem
R-RMCLS	regularized relaxed multi-channel least-squares
SCLS	single-channel least-squares
S-CS	sparsity-promoting channel shortening
SIR	speech-to-interference ratio
S-MINT	sparsity-promoting multiple-input/output inverse theorem
SNR	signal-to-noise ratio
S-PMINT	sparsity-promoting partial multi-channel equalization based on the multiple-input/output inverse theorem
S-RMCLS	sparsity-promoting relaxed multi-channel least-squares
SRNR	signal-to-reverberation-and-noise ratio
SSI	supervised system identification
STFT	short-time Fourier transform

Mathematical notation

a	scalar a
$\mathbf{a}$	vector $\mathbf{a}$
L_a	length of vector $\mathbf{a}$
$\mathbf{A}$	matrix $\mathbf{A}$
$\hat{a}$	estimate of scalar a
$\hat{\mathbf{a}}$	estimate of vector $\mathbf{a}$

$\hat{\mathbf{A}}$	estimate of matrix $\mathbf{A}$
a^*	complex conjugate of scalar a
$\mathbf{a}^T$	transpose of vector $\mathbf{a}$
$\mathbf{A}^T$	transpose of matrix $\mathbf{A}$
$\mathbf{A}^H$	conjugate transpose of matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	inverse of matrix $\mathbf{A}$
$\mathbf{A}^+$	pseudo-inverse of matrix $\mathbf{A}$
$a(i)$	i-th element of vector $\mathbf{a}$
$\mathbf{a}^{(k)}$	value of vector $\mathbf{a}$ at iteration index k
$\sigma_{\mathbf{A}}(i)$	i-th singular value of matrix $\mathbf{A}$
$\chi_{\mathbf{A}}$	condition number of matrix $\mathbf{A}$
$\text{diag}\{\mathbf{a}\}$	square diagonal matrix with vector $\mathbf{a}$ on the diagonal

$x(n)$	discrete-time sequence at discrete-time index n
$X(\omega)$	discrete-time Fourier transform of $x(n)$ at angular frequency ω
$X(t, f)$	short-time Fourier transform of $x(n)$ at time frame index t and frequency bin index f
$P_x(\omega)$	power spectral density of $x(n)$
$\mathbf{R}_{\mathbf{x}}(n)$	auto-correlation matrix of vector $\mathbf{x}(n)$
$\mathbf{R}_{\mathbf{xy}}(n)$	cross-correlation matrix of vectors $\mathbf{x}(n)$ and $\mathbf{y}(n)$

$*$	convolution operator
$\mathcal{E}$	expected value operator
$\mathbf{\Psi}$	short-time Fourier transform operator
$\mathbf{\Psi}^H$	inverse short-time Fourier transform operator
$\lceil \cdot \rceil$	ceiling operator
$\{\cdot\}'$	first-order derivative
$\{\cdot\}''$	second-order derivative
$\lvert \cdot \rvert$	magnitude
$\lVert \cdot \rVert_0$	l_0-norm
$\lVert \cdot \rVert_1$	l_1-norm
$\lVert \cdot \rVert_2$	l_2-norm

Fixed symbols

n	discrete-time index
ω	angular frequency
t	time frame index
f	frequency bin index
m	microphone index
M	number of microphones
T_{60}	reverberation time
f_s	sampling frequency
L_e	desired window length in number of samples
L_d	desired window length in ms

$h_m(n)$	room impulse response between the source and the m-th microphone
$h_{e,m}(n)$	direct path and early reflections of the room impulse response between the source and the m-th microphone
$h_{r,m}(n)$	late reflections of the room impulse response between the source and the m-th microphone
$w_m(n)$	filter applied to the m-th microphone
$c(n)$	equalized impulse response
$s(n)$	clean speech signal
$x_m(n)$	reverberant speech component in the m-th microphone signal
$x_{e,m}(n)$	early reverberation component in the m-th microphone signal
$x_{r,m}(n)$	late reverberation component in the m-th microphone signal
$v_m(n)$	noise component in the m-th microphone signal
$y_m(n)$	m-th microphone signal
$z(n)$	output signal
$z_x(n)$	speech component in the output signal
$z_v(n)$	noise component in the output signal
$z_{e,x}(n)$	early reverberation component in the output signal
$z_{r,x}(n)$	late reverberation component in the output signal

$\mathbf{h}_m$	m-th room impulse response vector
$\hat{\mathbf{h}}_m$	m-th perturbed room impulse response vector
$\mathbf{e}_m$	perturbation of the m-th room impulse response vector

$\mathbf{w}_m$	m-th filter vector
$\mathbf{x}_m(n)$	m-th reverberant speech component vector
$\mathbf{v}_m(n)$	m-th noise component vector
$\mathbf{y}_m(n)$	m-th received signal vector
$\mathbf{h}$	stacked room impulse response vector
$\mathbf{w}$	stacked filter vector
$\mathbf{c}$	equalized impulse response vector
$\hat{\mathbf{c}}$	perturbed equalized impulse response vector
$\mathbf{c}_t$	target equalized impulse response vector
$\mathbf{s}(n)$	clean speech vector
$\mathbf{x}(n)$	stacked reverberant speech component vector
$\mathbf{v}(n)$	stacked noise component vector
$\mathbf{y}(n)$	stacked microphone signal vector
$\mathbf{z}(n)$	output signal vector
$\mathbf{H}_m$	convolution matrix of the m-th room impulse response
$\mathbf{H}$	stacked multi-channel convolution matrix
$\hat{\mathbf{H}}$	perturbed stacked multi-channel convolution matrix
$\mathbf{E}$	stacked multi-channel convolution matrix of the perturbations
$\mathbf{W}$	least-squares weighting matrix
$\mathbf{W}_{\mathrm{R}}$	relaxed multi-channel least-squares weighting matrix
$\mathbf{W}_{\mathrm{d}}$	channel shortening desired weighting matrix
$\mathbf{W}_{\mathrm{u}}$	channel shortening undesired weighting matrix
$\mathbf{I}$	identity matrix

κ	curvature of a parametric surface
ϵ_c	least-squares dereverberation error energy
ϵ_r	channel shortening dereverberation error energy
ϵ_e	distortion energy
ϵ_s	sparsity measure
ϵ_x	speech distortion
ϵ_v	output noise power
δ	regularization parameter for regularized techniques
η	weighting parameter for sparsity-promoting techniques
ρ	penalty parameter for alternating direction method of multipliers
μ	weighting parameter for joint dereverberation and noise reduction techniques

L_{o}	optimal reshaping filter length
δ_{o}	optimal regularization parameter
δ_{a}	automatic regularization parameter
η_{o}	optimal weighting parameter
ρ_{o}	optimal penalty parameter

J_{LS}	least-squares cost function
J_{M}	multiple-input/output inverse theorem cost function
J_{CS}	channel shortening cost function
J_{R}	relaxed multi-channel least-squares cost function
J_{P}	partial multi-channel equalization based on the multiple-input/output inverse theorem cost function
$J_{\mathrm{R\text{-}LS}}$	regularized least-squares cost function
$J_{\mathrm{R\text{-}M}}$	regularized multiple-input/output inverse theorem cost function
$J_{\mathrm{R\text{-}CS}}$	regularized channel shortening cost function
$J_{\mathrm{R\text{-}R}}$	regularized relaxed multi-channel least-squares cost function
$J_{\mathrm{R\text{-}P}}$	regularized partial multi-channel equalization based on the multiple-input/output inverse theorem cost function
$J_{\mathrm{S\text{-}LS}}$	sparsity-promoting least-squares cost function
$J_{\mathrm{S\text{-}CS}}$	sparsity-promoting channel shortening cost function
$J_{\mathrm{RP\text{-}DNR}}$	regularized partial multi-channel equalization based on the multiple-input/output inverse theorem for joint dereverberation and noise reduction cost function
$J_{\mathrm{MWF\text{-}DNR}}$	multi-channel Wiener filter for joint dereverberation and noise reduction cost function
f_{sp}	sparsity-promoting penalty function
f_{sp}^{0}	l_0-norm sparsity-promoting penalty function
f_{sp}^{1}	l_1-norm sparsity-promoting penalty function
$f_{\mathrm{sp}}^{w,1}$	weighted l_1-norm sparsity-promoting penalty function
$\mathcal{L}_{\mathrm{S\text{-}LS}}$	augmented Lagrangian for sparsity-promoting least-squares optimization
$\mathcal{L}_{\mathrm{S\text{-}CS}}$	augmented Lagrangian for sparsity-promoting channel shortening optimization

$\mathbf{w}_{\mathrm{LS}}$	least-squares reshaping filter
$\mathbf{w}_{\mathrm{M}}$	multiple-input/output inverse theorem reshaping filter

$\mathbf{w}_{\text{CS}}$	channel shortening reshaping filter
$\mathbf{w}_{\text{R}}$	relaxed multi-channel least-squares reshaping filter
$\mathbf{w}_{\text{P}}$	partial multi-channel equalization based on the multiple-input/output inverse theorem reshaping filter
$\mathbf{w}_{\text{R-LS}}$	regularized least-squares reshaping filter
$\mathbf{w}_{\text{R-M}}$	regularized multiple-input/output inverse theorem reshaping filter
$\mathbf{w}_{\text{R-R}}$	regularized relaxed multi-channel least-squares reshaping filter
$\mathbf{w}_{\text{R-P}}$	regularized partial multi-channel equalization based on the multiple-input/output inverse theorem reshaping filter
$\mathbf{w}_{\text{S-LS}}$	sparsity-promoting least-squares reshaping filter
$\mathbf{w}_{\text{S-M}}$	sparsity-promoting multiple-input/output inverse theorem reshaping filter
$\mathbf{w}_{\text{S-R}}$	sparsity-promoting relaxed multi-channel least-squares reshaping filter
$\mathbf{w}_{\text{S-P}}$	sparsity-promoting partial multi-channel equalization based on the multiple-input/output inverse theorem reshaping filter
$\mathbf{w}_{\text{RP-DNR}}$	regularized partial multi-channel equalization based on the multiple-input/output inverse theorem for joint dereverberation and noise reduction filter
$\mathbf{w}_{\text{MWF-DNR}}$	multi-channel Wiener filter for joint dereverberation and noise reduction

CONTENTS

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Reverberation in an enclosure . . . 2
- 1.3 Overview of speech enhancement techniques . . . 4
- 1.4 Acoustic channel equalization . . . 9
- 1.5 Outline of the thesis and main contributions . . . 12

2 Problem Formulation and Instrumental Performance Measures 17
- 2.1 Problem formulation . . . 17
- 2.2 Room impulse response perturbations . . . 24
- 2.3 Instrumental performance measures . . . 27
- 2.4 Summary . . . 30

3 Partial Multi-Channel Equalization based on the Multiple-Input/Output Inverse Theorem 33
- 3.1 Acoustic multi-channel equalization techniques . . . 34
- 3.2 Partial multi-channel equalization based on MINT . . . 38
- 3.3 Generalized framework for least-squares equalization . . . 39
- 3.4 Simulations . . . 42
- 3.5 Summary . . . 48

4 Acoustic Multi-Channel Equalization Using Shorter Reshaping Filters 49
- 4.1 Reshaping filter length in least-squares equalization techniques . . . 50
- 4.2 Reshaping filter length in the channel shortening technique . . . 54
- 4.3 Simulations . . . 56
- 4.4 Summary . . . 65

5 Regularized Acoustic Multi-Channel Equalization 67
- 5.1 Incorporating regularization in acoustic multi-channel equalization . . . 68
- 5.2 Regularized acoustic multi-channel equalization reshaping filters . . . 70
- 5.3 Automatic regularization parameter . . . 73
- 5.4 Non-applicability of the automatic procedure to the regularized channel shortening technique . . . 76
- 5.5 Simulations . . . 77
- 5.6 Summary . . . 88

6 Sparsity-Promoting Acoustic Multi-Channel Equalization 89
- 6.1 Sparsity of speech signals . . . 90
- 6.2 Incorporating sparsity-promoting penalty functions in acoustic multi-channel equalization . . . 91
- 6.3 Sparsity-promoting acoustic multi-channel equalization reshaping filters 95
- 6.4 Simulations . . . 100
- 6.5 Summary . . . 114

7 Objective and Subjective Evaluation of Robust Acoustic Multi-Channel Equalization Techniques 115
7.1 Acoustic systems and algorithmic settings 116
7.2 Objective evaluation 117
7.3 Subjective evaluation 119
7.4 Summary 124
8 Joint Dereverberation and Noise Reduction based on Robust Acoustic Multi-Channel Equalization 127
8.1 Joint dereverberation and noise reduction techniques 128
8.2 Insights on the RP-DNR and MWF-DNR techniques 132
8.3 Automatic regularization and weighting parameters 135
8.4 Simulations 137
8.5 Summary 146
9 Conclusion and Further Research 149
9.1 Conclusion 149
9.2 Suggestions for further research 153
A Interlacing Inequalities for Shorter Reshaping Filters in Least-Squares Equalization Techniques 155
B Frequency Domain One- and Two-Stage Techniques for Joint Dereverberation and Noise Reduction 157
B.1 Two-stage technique for joint dereverberation and noise reduction . . 158
B.2 One-stage technique for joint dereverberation and noise reduction . . 161
B.3 Analytical comparison of the one-stage and two-stage techniques . . 162

BIBLIOGRAPHY 165

LIST OF FIGURES

Fig. 1.1 Schematic illustration of a typical acoustic scenario in hands-free speech communication applications. 2
Fig. 1.2 Schematic illustration of reverberation in an enclosure. 3
Fig. 1.3 An exemplary room impulse response (reverberation time $T_{60} \approx$ 450 ms and direct-to-reverberant ratio DRR = 0 dB. 4
Fig. 1.4 Schematic illustration of single-channel spectral enhancement for joint dereverberation and noise reduction. $y(n)$ denotes the received microphone signal and $z(n)$ denotes the output speech signal. 6
Fig. 1.5 Schematic illustration of the Generalized Sidelobe Canceller structure for multi-channel speech enhancement. $y_m(n)$ denotes the m-th received microphone signal, $m = 1,\ 2,\ \ldots, M$, with M the number of microphones, and $z(n)$ denotes the output speech signal. 7
Fig. 1.6 Schematic illustration of two-stage techniques for joint dereverberation and noise reduction. $y_m(n)$ denotes the m-th received microphone signal, $m = 1,\ 2,\ \ldots, M$, with M the number of microphones, and $z(n)$ denotes the output speech signal. 8
Fig. 1.7 Schematic illustration of acoustic multi-channel equalization techniques for dereverberation. $y_m(n)$ denotes the m-th received microphone signal, $m = 1,\ 2,\ \ldots, M$, with M the number of microphones, and $z(n)$ denotes the output speech signal. 9
Fig. 1.8 Schematic overview of the thesis. 14
Fig. 2.1 Schematic illustration of a typical time domain multi-channel speech enhancement system. 19
Fig. 3.1 Exemplary equalized impulse response when the true room impulse responses are known obtained using the (a) MINT technique, (b) CS technique, (c) RMCLS technique, and (d) PMINT technique. The delay is set to $\tau = 90$, corresponding to 11.25 ms, and the desired window length is set to $L_e = 400$, corresponding to 50 ms. The considered acoustic system is the same as in Section 3.4.1. 39
Fig. 3.2 The true room impulse response between the speech source and the first microphone. 43
Fig. 3.3 Performance of the CS, RMCLS, and PMINT techniques for known true RIRs in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD. 45
Fig. 3.4 Performance of the MINT, CS, RMCLS, and PMINT techniques in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d =$ 50 ms, (c) ΔPESQ, and (d) ΔCD (averaged over several NPMs). . . 47
Fig. 4.1 Schematic illustration of the construction of the $p_s \times q_s$–dimensional sub-matrix $\mathbf{W}_s\hat{\mathbf{H}}_s$ from the $p_t \times q_t$–dimensional matrix $\mathbf{W}_t\hat{\mathbf{H}}_t$. . . . 52

Fig. 4.2 Singular values of an exemplary matrix $\mathbf{W}_t\hat{\mathbf{H}}_t$ ($L_t = 1200$) and of two sub-matrices $\mathbf{W}_s\hat{\mathbf{H}}_s$ ($L_s = 800$ and $L_s = 500$) for the (a) PMINT technique (i.e., $\mathbf{W}_t = \mathbf{I}$) and (b) RMCLS technique (i.e., $\mathbf{W}_t = \mathbf{W}_{\text{R}}$ and $L_d = 50$ ms). The largest and smallest non-zero singular values of each matrix are explicitly denoted. The considered acoustic system is the same as in Section 3.4.1. 53
Fig. 4.3 Performance of the MINT and L-MINT techniques in terms of (a) ΔDRR, (b) EDC, (c) ΔPESQ, and (d) ΔCD (NPM = −33 dB). 58
Fig. 4.4 Performance of the CS and L-CS techniques in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD (NPM = −33 dB). 60
Fig. 4.5 Performance of the RMCLS and L-RMCLS techniques in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD (NPM = −33 dB). 61
Fig. 4.6 Performance of the PMINT and L-PMINT techniques in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD (NPM = −33 dB). 63
Fig. 4.7 Performance of the L-MINT, L-CS, L-RMCLS, and L-PMINT techniques in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD (averaged over several NPM values). 64
Fig. 4.8 Spectrograms of the (a) reference signal, (b) reverberant microphone signal, (c) output speech signal obtained using the CS technique ($L_t = 1200$), and (d) output speech signal obtained using the L-CS technique ($L_s = 800$) ($L_d = 50$ ms and NPM = −33 dB). 66
Fig. 5.1 Exemplary L-curve obtained using the regularized PMINT technique with the regularization parameter values ranging from 10^{-9} to 10^{-1}. The considered acoustic system is the same as in Section 3.4.1 . . . 75
Fig. 5.2 Performance of the MINT and the optimally regularized MINT techniques in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD (NPM = −33 dB). 80
Fig. 5.3 Performance of the CS and the optimally regularized CS techniques in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d =$ 50 ms, (c) ΔPESQ, and (d) ΔCD (NPM = −33 dB). 81
Fig. 5.4 Performance of the RMCLS and the optimally regularized RMCLS techniques in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD (NPM = −33 dB). . 83
Fig. 5.5 Performance of the PMINT and the optimally regularized PMINT techniques in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD (NPM = −33 dB). . 84
Fig. 5.6 Performance of the optimally regularized MINT, CS, RMCLS, and PMINT techniques in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD (averaged over several NPM values). 85

Fig. 5.7 Performance of the regularized PMINT technique using the optimal and the automatic regularization parameters δ_{o} and δ_{a} in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD (averaged over several NPM values). . . . 87
Fig. 6.1 Exemplary spectrograms of (a) clean speech and (b) reverberant speech. The STFT is computed using a 32 ms Hamming window with 50 % overlap between successive frames ($T_{60} \approx 610$ ms). 91
Fig. 6.2 Exemplary illustration of the proximal mappings for the l_0-norm, l_1-norm, and weighted l_1-norm ($\frac{\eta}{\rho} = 1$ and $u(i) = 2$). 99
Fig. 6.3 Performance of the MINT and sparsity-promoting MINT techniques with different penalty functions in terms of (a) ΔDRR, (b) EDC, (c) ΔPESQ, and (d) ΔCD (NPM = -33 dB). 103
Fig. 6.4 Performance of the CS and sparsity-promoting CS techniques with different penalty functions in terms of (a) ΔDRR, (b) EDC, (c) ΔPESQ, and (d) ΔCD (NPM = -33 dB). 105
Fig. 6.5 Performance of the RMCLS and sparsity-promoting RMCLS techniques with different penalty functions in terms of (a) ΔDRR, (b) EDC, (c) ΔPESQ, and (d) ΔCD (NPM = -33 dB). 107
Fig. 6.6 Performance of the PMINT and sparsity-promoting PMINT techniques with different penalty functions in terms of (a) ΔDRR, (b) EDC, (c) ΔPESQ, and (d) ΔCD (NPM = -33 dB). 108
Fig. 6.7 Spectrograms of the (a) clean speech signal and (b) reverberant microphone signal for the considered acoustic system ($T_{60} \approx 360$ ms). . 110
Fig. 6.8 Exemplary spectrograms of the output speech signal obtained using the (a) RMCLS technique, (b) l_0-norm sparsity-promoting RMCLS technique, (c) l_1-norm sparsity-promoting RMCLS technique, and (d) weighted l_1-norm sparsity-promoting RMCLS technique (NPM = -33 dB). 111
Fig. 6.9 Performance of the weighted l_1-norm sparsity-promoting MINT, CS, RMCLS, and PMINT techniques in terms of (a) ΔDRR, (b) EDC, (c) ΔPESQ, and (d) ΔCD (averaged over several NPM values). . . . 113
Fig. 7.1 The DRR improvement obtained using the robust extensions of the RMCLS and PMINT techniques. 118
Fig. 7.2 Performance of the robust extensions of the RMCLS and PMINT techniques in terms of (a) ΔPESQ and (b) ΔCD. 119
Fig. 7.3 MUSHRA scores for the anchor, reverberant microphone signal $x_1(n)$, and output speech signals obtained using the robust extensions of the RMCLS and PMINT techniques for (a) acoustic system 1 and $\mathrm{NPM}_1 = -33$ dB (S_1-NPM_1), (b) acoustic system 2 and $\mathrm{NPM}_1 = -33$ dB (S_2-NPM_1), (c) acoustic system 1 and $\mathrm{NPM}_2 = -15$ dB (S_1-NPM_2), and (d) acoustic system 2 and $\mathrm{NPM}_2 = -15$ dB (S_2-NPM_2). The scores of the hidden reference, close to 100 with small variance, are not displayed. On each box, the central mark is the median, the edges of the box are the 25-th and 75-th percentiles, and the whiskers extend to 1.5 times the interquartile range from the median. 120

Fig. 8.1 Exemplary parametric surface of the output noise power ϵ_v versus the dereverberation error energy ϵ_c and the distortion energy ϵ_e for the RP-DNR technique, with the regularization and weighting parameters δ and μ ranging from 10^{-7} to 10. 136
Fig. 8.2 Exemplary parametric plot of the output noise power ϵ_v versus the speech distortion ϵ_x for the MWF-DNR technique, with the weighting parameter μ ranging from 10^{-7} to 10. 137
Fig. 8.3 Performance of the RP-DNR technique for different regularization and weighting parameters δ and μ in terms of (a) ΔDRR and (b) ψ_{NR}. The circles denote the automatically determined regularization and weighting parameters (NPM = -33 dB, SIR = 0 dB). 140
Fig. 8.4 Performance of the MWF-DNR technique for different regularization and weighting parameters δ and μ in terms of (a) ΔDRR and (b) ψ_{NR}. The circles denote the automatically determined regularization and weighting parameters (NPM = -33 dB, SIR = 0 dB). 141
Fig. 8.5 Performance of the automatically parametrized RP-DNR and MWF-DNR techniques in terms of (a) ΔDRR, (b) ΔPESQ, (c) ψ_{NR}, (d) ΔSRNR, and (e) ΔfwSSNR (averaged over several NPM values, perfectly estimated correlation matrices). 144
Fig. 8.6 Performance of the automatically parametrized RP-DNR and MWF-DNR techniques in terms of (a) ΔDRR, (b) ΔPESQ, (c) ψ_{NR}, (d) ΔSRNR, and (e) ΔfwSSNR (averaged over several NPM values, erroneously estimated correlation matrices). 146
Fig. B.1 Acoustic system configuration for the two-stage beamforming technique for joint dereverberation and noise reduction. 158
Fig. B.2 Acoustic system configuration for the one-stage reformulation of the two-stage beamforming technique for joint dereverberation and noise reduction. 161

LIST OF TABLES

Table 3.1 Rank of the coefficient and augmented matrix for least-squares equalization techniques. 40
Table 4.1 Notation for different reshaping filter lengths and the corresponding least-squares matrices. 51
Table 4.2 Condition number of an exemplary matrix $\mathbf{W}_t\hat{\mathbf{H}}_t$ ($L_t = 1200$) and of two sub-matrices $\mathbf{W}_s\hat{\mathbf{H}}_s$ ($L_s = 800$ and $L_s = 500$) for the PMINT technique (i.e., $\mathbf{W}_t = \mathbf{I}$) and the RMCLS technique (i.e., $\mathbf{W}_t = \mathbf{W}_{\text{R}}$ and $L_d = 50$ ms). The considered acoustic system is the same as in Section 3.4.1. 54
Table 4.3 Maximum generalized eigenvalue for channel shortening exemplary matrices $\hat{\mathbf{D}}_s$ and $\hat{\mathbf{U}}_s$ constructed $L_s = 800$ and $L_s = 500$. The considered acoustic system is the same as in Section 3.4.1 and the desired window length is $L_d = 50$ ms. 56
Table 4.4 Optimal reshaping filter length for the L-MINT technique for several desired window lengths (NPM = -33 dB). 58
Table 4.5 Optimal reshaping filter length for the L-CS technique for several desired window lengths (NPM = -33 dB). 60
Table 4.6 Optimal reshaping filter length for the L-RMCLS technique for several desired window lengths (NPM = -33 dB). 61
Table 4.7 Optimal reshaping filter length for the L-PMINT technique for several desired window lengths (NPM = -33 dB). 63
Table 5.1 Regularized least-squares cost function for different regularized least-squares techniques. 71
Table 5.2 Regularized least-squares reshaping filter for different regularized least-squares techniques. 71
Table 5.3 Optimal regularization parameter for the regularized MINT technique for several desired window lengths (NPM = -33 dB). 80
Table 5.4 Optimal regularization parameter for the regularized CS technique for several desired window lengths (NPM = -33 dB). 81
Table 5.5 Optimal regularization parameter for the regularized RMCLS technique for several desired window lengths (NPM = -33 dB). 83
Table 5.6 Optimal regularization parameter for the regularized PMINT technique for several desired window lengths (NPM = -33 dB). 84
Table 6.1 Sparsity-promoting least-squares reshaping filter update rules for different sparsity-promoting least-squares techniques. 97
Table 6.2 Optimal parameters for the sparsity-promoting MINT technique with different penalty functions (NPM = -33 dB). 103
Table 6.3 Optimal parameters for the sparsity-promoting CS technique with different penalty functions (NPM = -33 dB). 105

Table 6.4 Optimal parameters for the sparsity-promoting RMCLS technique with different penalty functions (NPM = −33 dB). 107
Table 6.5 Optimal parameters for the sparsity-promoting PMINT technique with different penalty functions (NPM = −33 dB). 108
Table 7.1 Characteristics of the considered acoustic systems. 116
Table 7.2 ANOVA results for the different considered acoustic scenarios. . . . 122
Table 7.3 Overview of the student's t-test results for acoustic system 1 and $\text{NPM}_1 = -33$ dB (S_1-NPM_1). The ticks represent a statistically significant difference, i.e., $p < 0.05$, and the crosses represent no statistically significant difference, i.e., $p \geq 0.05$. 123
Table 7.4 Overview of the student's t-test results for acoustic system 2 and $\text{NPM}_1 = -33$ dB (S_2-NPM_1). The ticks represent a statistically significant difference, i.e., $p < 0.05$, and the crosses represent no statistically significant difference, i.e., $p \geq 0.05$. 123
Table 7.5 Overview of the student's t-test results for acoustic system 1 and $\text{NPM}_2 = -15$ dB (S_1-NPM_2). The ticks represent a statistically significant difference, i.e., $p < 0.05$, and the crosses represent no statistically significant difference, i.e., $p \geq 0.05$. 123
Table 7.6 Overview of the student's t-test results for acoustic system 2 and $\text{NPM}_2 = -15$ dB (S_2-NPM_2). The ticks represent a statistically significant difference, i.e., $p < 0.05$, and the crosses represent no statistically significant difference, i.e., $p \geq 0.05$. 124
Table 8.1 Performance of the PMINT technique, automatically regularized R-PMINT technique, and automatically parametrized RP-DNR and MWF-DNR techniques (averaged over several NPM values; SIR = 0 dB). For each performance measure, the best performance is highlighted. 142
Table 8.2 Performance of the PMINT technique, automatically regularized R-PMINT technique, and automatically parametrized RP-DNR and MWF-DNR techniques (averaged over several NPM values; SIR = 5 dB). For each performance measure, the best performance is highlighted. 143

1

INTRODUCTION

1.1 Motivation

The rapid rise of powerful portable smart devices such as smartphones, tablets, and smartwatches, has resulted in a vastly expanding market for hands-free speech communication interfaces. Such interfaces are being deployed in a wide range of applications such as assisted living technologies, car interior communication systems, and consumer electronics devices. Furthermore, computer hardware and networks have been dramatically evolving in the last years, giving rise to multimedia applications such as hands-free teleconferencing, which allows people who are geographically dispersed to hold conferences by sending and receiving audio and video data over networks. The work presented in this thesis is motivated by the continuously and rapidly growing demand for high-quality hands-free communication in such a wide range of applications.

In hands-free communication, speech is acquired by a single microphone or multiple microphones placed at a distance from the speaker in an adverse acoustic environment. As is schematically illustrated in Fig. 1.1 for a multi-microphone setup, the received microphone signals contain not only the desired speech signal, but also other interferences such as reverberation and background noise [1].

Reverberation arises whenever sound is produced in an enclosed space and the acoustic waves coming from the sound source are reflected by the walls and other surrounding objects. While carefully controlled and moderate reverberation may be desirable [2], severe reverberation yields a degradation in speech intelligibility [3–5]. Moreover, since reverberation alters the characteristics of the speech signal, the performance of acoustic source localization techniques and automatic speech recognition systems rapidly degrades with increasing reverberation levels [6–9].

Background noise arises, e.g., due to other speakers, passing traffic, or electronic appliances. When its level is comparable or larger than the speech level, listening comfort and speech intelligibility are significantly degraded [3, 4]. Furthermore, the performance of acoustic source localization techniques and automatic speech recognition systems also rapidly degrades with increasing background noise levels [1].

Therefore, reverberation and background noise cause a signal degradation which can impair speech intelligibility and which decrease the performance for many signal processing techniques. Noise reduction techniques have been widely investigated in the

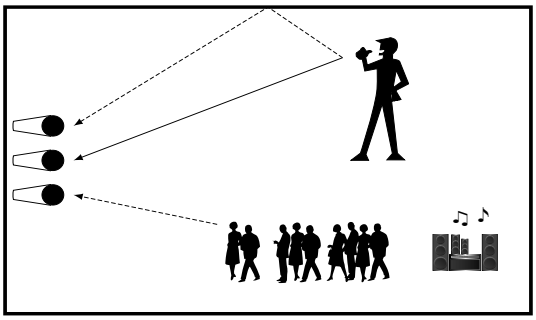


Fig. 1.1: Schematic illustration of a typical acoustic scenario in hands-free speech communication applications.

literature, and several significant contributions and robust solutions have already been proposed [10–19]. On the other hand, dereverberation as well as joint dereverberation and noise reduction has received much less attention until recently [20]. The main difference between reverberation and noise is that reverberation depends on the desired speech signal whereas noise is usually assumed to be independent from the desired speech signal. As a result, the wide range of effective noise reduction techniques that have been developed so far cannot be readily applied to speech dereverberation. Furthermore, modeling the convolutive nature of reverberation is significantly more complex than modeling the additive combination of the speech and noise signals, posing a challenge in developing effective dereverberation techniques, both from a theoretical perspective as well as from a computational complexity and numerical precision perspective. Although a significant progress has been made in the last years in the field of dereverberation as indicated by the large number of contributions in a recent international challenge [21], robust and perceptually advantageous solutions for dereverberation and for joint dereverberation and noise reduction remain to be established.

The objective of this thesis is to investigate, develop, and evaluate **robust** and **perceptually advantageous** speech **dereverberation** techniques based on **acoustic multi-channel equalization**, as well as to effectively integrate them with noise reduction in order to achieve **joint dereverberation and noise reduction**.

1.2 Reverberation in an enclosure

Since dereverberation is the central topic of this thesis, in this section some insights on the qualitative and quantitative aspects of reverberation are provided.

Reverberation is the collection of reflected sounds from walls and objects within an enclosure. As is schematically illustrated in Fig. 1.2, the recorded microphone signal in a reverberant environment consists of the direct path signal and multiple delayed and attenuated versions, referred to as reverberation. Reverberation can be divided into two components: early reverberation and late reverberation.

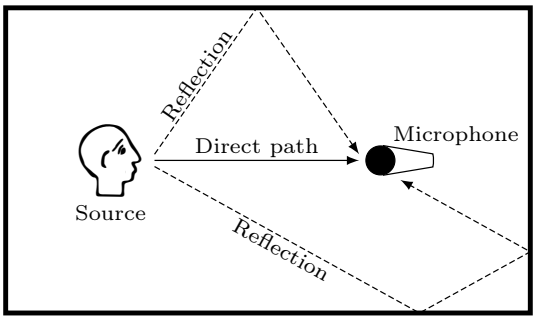


Fig. 1.2: Schematic illustration of reverberation in an enclosure.

Early reverberation refers to the reflected signals arriving at the microphone immediately after the direct path signal, typically considered to be within 1–50 ms [22]. As the source-microphone geometry changes, also the early reverberation changes, providing insights about the volume of the enclosure and about the position of the source within the enclosure [23–25]. Psychoacoustically, early reverberation is perceived to reinforce the direct path signal and has been shown to have a positive effect on speech intelligibility [5,22,26,27]. However, early reverberation also causes coloration, which can degrade the quality of the recorded speech signal [22].

Late reverberation refers to the reflected signals arriving at the microphone after the early reverberation, typically considered to be 50 ms after the direct path signal [22]. Late reverberation is the main cause of speech intelligibility degradation [3–5], particularly for non-native speakers [3] and for the hearing-impaired [4, 28].

The acoustic path between the source and the microphone can be described by the *acoustic impulse response* (AIR), which can, e.g., be measured by exciting the acoustic environment with an impulsive sound. While AIR is used to refer to an acoustic impulse response in general, the acoustic context is commonly limited to be within a room. Hence, in the following we will refer to the impulse response as a *room impulse response* (RIR). Similarly to the recorded microphone signal, the RIR can be divided into three components: i) *direct path*, ii) *early reflections*, and iii) *late reflections*. Fig. 1.3 shows an example of a measured RIR, indicating the direct path, early reflections, and late reflections. As can be observed, the direct path propagation from the source to the microphone initially yields a period of zero amplitude (or nearly-zero due to the discrete sampling of the RIR), which is referred to as the RIR delay. The delay depends on the source-microphone distance, the sampling frequency, and the speed of sound. Furthermore, the early reflections of an RIR are typically considered to be well-separated and distinct impulses with a large amplitude, whereas the late reflections have a significantly smaller amplitude and a diffuse-like nature.

Room impulse responses can be modeled using *all-zero*, *all-pole*, or *pole-zero* models [29–33]. The most commonly used all-zero model, i.e., the finite impulse response (FIR) filter model, can achieve a high degree of accuracy, with the drawback

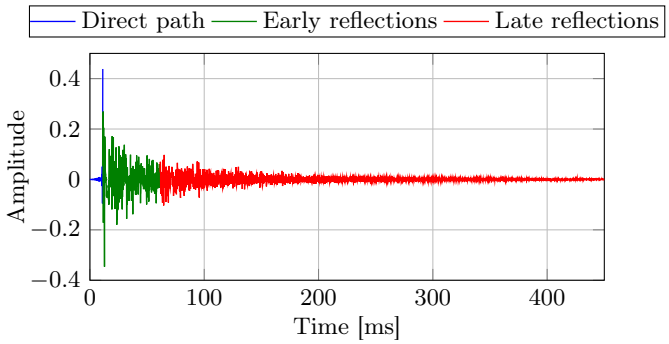


Fig. 1.3: An exemplary room impulse response (reverberation time $T_{60} \approx 450$ ms and direct-to-reverberant ratio DRR = 0 dB.

that the model order is high for long RIRs [29]. It should be noted that since RIRs are typically several thousand samples long (depending on the acoustic environment), the likelihood that RIRs share near-common zeroes tends to be high [34]. In order to decrease the model order, the all-pole model, i.e., the infinite impulse response (IIR) filter model, has also been investigated in [30]. However, since acoustic systems generally are mixed-phase systems [35–37], a stable all-pole model can only model the minimum-phase component of the acoustic system. In order to model all components, pole-zero models such as the common-acoustical-poles-and-zeros (CAPZ) model have been investigated [31,32]. The estimation of the model parameters in the CAPZ model however requires non-linear optimization procedures, which might lead to instability issues and convergence to local minima. Instead of exactly modeling the RIR, also statistical descriptions of the RIR characteristics in terms of quantities such as source-microphone distance or volume of the enclosure have been investigated [38–40].

Reverberation in an enclosure is generally quantified using the *reverberation time*. The reverberation time T_{60}, originally introduced by W. C. Sabine [22], is defined as the time taken by the reverberant energy to decay by 60 dB once the direct path signal has been interrupted. Since the reverberation time is invariant to changes in the source-microphone geometry whereas the RIR highly depends on the source-microphone geometry, a commonly-used quantity which reflects the spatial dependence of the RIR is the *direct-to-reverberant ratio* (DRR), defined as the ratio of the energy of the direct path of the RIR to the energy of the reflections [20].

1.3 Overview of speech enhancement techniques

With the continuously growing demand for high-quality hands-free communication, speech enhancement techniques aiming at dereverberation and at joint dereverberation and noise reduction have become indispensable. In the last decades, several single- as well as multi-channel techniques have been proposed, with multi-channel

techniques being generally preferred since they enable to exploit both the spectro-temporal and the spatial characteristics of the received microphone signals.

Dereverberation techniques can be divided into many categories, such as single- and multi-channel techniques, complete and partial dereverberation techniques, or blind and non-blind techniques. We categorize existing dereverberation techniques into

i) spectral enhancement (single- and multi-channel techniques),
ii) beamforming and multi-channel Wiener filtering (multi-channel techniques),
iii) blind probabilistic modeling-based (single- and multi-channel techniques), and
iv) acoustic channel equalization (single- and multi-channel techniques).

This categorization is not meant to be exclusive and techniques from different categories share common assumptions, models, and processing. In this section a coarse overview of several spectral enhancement, beamforming and multi-channel Wiener filtering, and probabilistic modeling-based techniques is provided. This overview is by no means self-contained covering all speech enhancement techniques proposed in the past decades, but it provides a general description of the different approaches that have been proposed to achieve dereverberation as well as joint dereverberation and noise reduction. A summary of other existing speech enhancement techniques can be found, e.g., in [20,40]. Since acoustic channel equalization is the central topic of this thesis, a more detailed overview of techniques belonging to this category is provided in Section 1.4.

i) Spectral enhancement

Spectral enhancement techniques traditionally refer to single-channel techniques and have been widely investigated for several decades for the enhancement of noisy speech signals [41–43]. One of the first proposed spectral enhancement techniques is spectral subtraction [44, 45], which subtracts an estimate of the noise magnitude from the noisy speech magnitude. In order to improve the quality of the processed speech signal, i.e., reduce speech distortion or residual noise, several modifications to the traditional spectral subtraction technique have been proposed, such as over-subtracting estimates of the noise spectrum and spectral flooring [46], non-linear spectral subtraction [47, 48], or the incorporation of psychoacoustically motivated over-subtraction parameters [49]. A significant drawback of spectral subtraction is the random variation of the estimated noise spectrum, often giving rise to disturbing artifacts known as musical noise.

A theoretically more solid and advanced approach to spectral enhancement is the derivation of statistically optimal clean speech estimators, which are based on statistical models for the speech and noise signals and a perceptually relevant distortion measure to be minimized. A wide range of estimators have been investigated for the enhancement of noisy speech, differing in the statistical models used for the speech and noise signals and in the distortion measure they minimize, e.g., in [50–56]. These techniques achieve noise reduction by applying a (typically) real-valued time-frequency-dependent gain function to the time-frequency representation of the noisy

speech signal. The gain function is computed based on an estimate of the speech and noise power spectral density (PSD).

Spectral enhancement techniques can be adapted to achieve dereverberation, as long as an estimate of the reverberation PSD can be obtained. The first single-channel dereverberation technique based on spectral enhancement was proposed in [57]. This technique relies on statistical room acoustics and models the room impulse response as an independent and identically distributed white Gaussian noise sequence with an exponentially decaying variance [39]. Based on this model, an estimate of the late reverberation PSD is obtained. This technique was extended to multiple microphones in [58, 59], where it is experimentally validated that by exploiting multiple microphones, a better dereverberation performance can be obtained. Statistically optimal estimators for dereverberation have also been investigated [40], and several single- and multi-channel techniques for estimating the late reverberation PSD using statistical room acoustics have been proposed, e.g., in [40, 60–62].

Spectral enhancement techniques have also been extended to achieve joint dereverberation and noise reduction, e.g., in [40, 63]. Fig. 1.4 presents a schematic representation of single-channel spectral enhancement for joint dereverberation and noise reduction. As illustrated, first a time-frequency representation of the received reverberant and noisy signal is computed using an analysis filter bank. Based on estimates of the noise and the late reverberation PSDs, a spectral gain function is designed and applied to the received signal spectrum. The time-domain output speech signal is then obtained by using a synthesis filter bank.

Spectral enhancement techniques are amongst the most computationally efficient techniques suitable for real-time processing, however, particular care must be taken when implementing them such that the arising speech distortion and musical noise are controlled. Furthermore, by applying a (typically) real-valued gain function, these techniques are not able to perfectly recover the dereverberated and denoised speech signal.

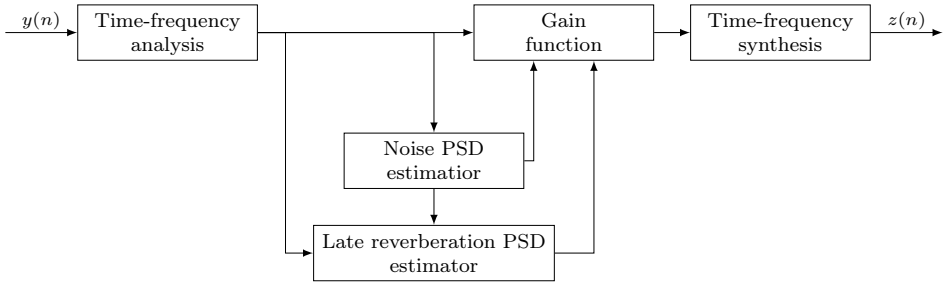


Fig. 1.4: Schematic illustration of single-channel spectral enhancement for joint dereverberation and noise reduction. $y(n)$ denotes the received microphone signal and $z(n)$ denotes the output speech signal.

ii) Beamforming and multi-channel Wiener filtering

Multi-channel beamforming techniques for noise reduction have been extensively investigated in the literature, aiming to design filters which spatially focus on the speech source and suppress the background noise not coming from the same direction as the speech source [19,64–69]. Typical fixed beamforming techniques include delay-and-sum beamforming [64], differential microphone arrays [65, 66], minimum variance distortionless response (MVDR) beamformers [68,69], and frequency-invariant beamformers [67]. To increase the noise reduction performance of fixed beamformers, also adaptive beamforming techniques have been widely investigated, typically implemented in a Generalized Sidelobe Canceller (GSC) structure [70–72]. As is schematically illustrated in Fig. 1.5, the GSC consists of a fixed beamformer, creating a so-called desired reference signal, a blocking matrix, creating a so-called interference reference signal, and an adaptive filter aiming to suppress the residual interference in the desired reference signal. Fixed and adaptive beamforming techniques have been originally investigated for the enhancement of noisy speech. While the dereverberation performance of fixed beamformers is limited, adaptive beamforming techniques cannot be straightforwardly used for dereverberation.

Nevertheless, GSC-like structures have also been tailored to enhance reverberant and noisy speech [73–79]. In [73] the delay-and-sum beamformer has been used as a fixed beamformer to generate the desired reference signal. In addition, the delay-and-subtract beamformer has been used as a blocking matrix to cancel the direct speech component and generate a reference interference signal (containing reverberation and noise). Alternatively, it has been proposed in [74,75] to compute the blocking matrix using blind source separation, aiming to cancel both the direct and early reverberation component. A slightly different approach is taken in [76–78], where it is assumed that the spatial coherence matrix of the late reverberation can be modeled as a scaled diffuse sound field, which holds for frequencies above the Schroeder frequency [38]. Based on this model, a maximum likelihood estimate of the late reverberation PSD at the output of the blocking matrix is derived. The

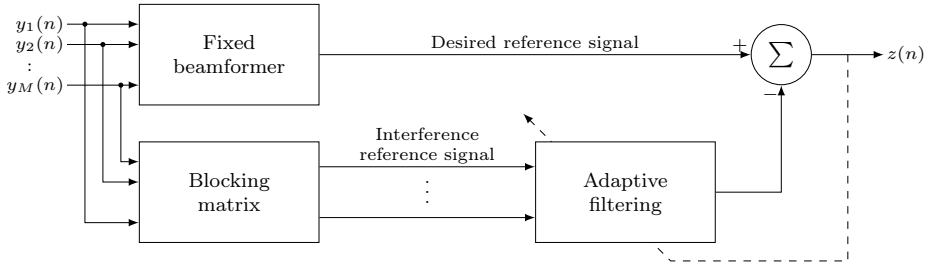


Fig. 1.5: Schematic illustration of the Generalized Sidelobe Canceller structure for multi-channel speech enhancement. $y_m(n)$ denotes the m-th received microphone signal, $m = 1, 2, \ldots, M$, with M the number of microphones, and $z(n)$ denotes the output speech signal.

estimated late reverberation PSD is used in a spectral filter to suppress the late reverberant power at the output of a fixed MVDR beamformer.

Other techniques to joint dereverberation and noise reduction directly combine fixed beamformers and the previously described single-channel spectral filters [63, 80], as is schematically illustrated in Fig. 1.6. Such techniques consist of two stages, where in the first stage a fixed beamformer, e.g., delay-and-sum or MVDR beamformer, is used to suppress some reverberation and background noise, and in the second stage a single-channel spectral postfilter is applied to suppress the residual reverberation and noise at the beamformer output.

A very related class of multi-channel speech enhancement techniques is based on the multi-channel Wiener filter (MWF) [81–87], which can be decomposed into an MVDR beamformer and a single-channel spectral postfilter. The MWF computes the minimum mean square error estimate of a reference signal. The estimation of several reference signals has been considered, e.g., the clean speech signal [81,82,87], the reverberant speech component at an arbitrarily chosen microphone [83, 85, 86], or a spatially pre-processed reference speech signal [84].

Due to possible microphone mismatches and estimation errors in the direction-of-arrival of the speech source, GSC-based techniques can cause distortion of the desired speech signal. Furthermore, due to the spectral filtering involved in MWF-based techniques, a trade-off arises between speech distortion and reverberation and noise reduction, which should be carefully controlled. Similarly as the spectral enhancement techniques, also beamforming and MWF-based techniques are not able to perfectly recover the dereverberated and denoised speech signal.

iii) Blind probabilistic modeling-based

A recent category of dereverberation techniques are blind probabilistic modeling-based techniques [88–98], which use statistical models to represent the unknown clean speech signal and the unknown room impulse responses. Using the received microphone signals, the parameters of the assumed statistical models for the clean speech signal and for the room impulse responses are then blindly estimated.

Blind probabilistic modeling-based techniques generally operate in the frequency-domain and model the acoustic transfer function as an auto-regressive process [91, 94, 95]. Alternatively, the acoustic transfer function has been modeled using the convolutive transfer approximation [93, 97, 98]. In addition, the clean speech spectral coefficients have been modeled using a Gaussian distribution [91], a Laplacian

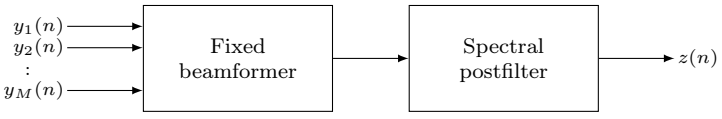


Fig. 1.6: Schematic illustration of two-stage techniques for joint dereverberation and noise reduction. $y_m(n)$ denotes the m-th received microphone signal, $m = 1,\ 2,\ \ldots, M$, with M the number of microphones, and $z(n)$ denotes the output speech signal.

distribution [95], or an all-pole model [94]. In [96, 98] it has been shown that by modeling the clean speech spectral coefficients using sparse circular priors, a better dereverberation performance can be achieved.

Furthermore, by also modeling the noise spectral coefficients, blind probabilistic modeling-based techniques have been successfully extended to achieve joint dereverberation and noise reduction [99, 100].

Clearly, these techniques fundamentally rely on realistic statistical models to describe the underlying speech process and acoustic transfer functions, hence, the choice of these models is of crucial importance to the dereverberation and noise reduction performance. Moreover, such techniques typically employ iterative procedures to blindly estimate the model parameters, which can in general be computationally complex. Similarly as the previously presented techniques, also blind probabilistic modeling-based techniques are not able to perfectly recover the dereverberated and denoised speech signal.

1.4 Acoustic channel equalization

In general the previously presented dereverberation techniques offer the potential to achieve a good dereverberation performance. However, they can never achieve a perfect dereverberation performance, i.e., exactly recover the clean speech signal or the early reverberation component. On the other hand, acoustic channel equalization techniques, which aim at inverting or reshaping the RIRs between the speech source and the microphones, can in theory achieve perfect dereverberation performance. Motivated by this potential, this thesis deals with acoustic channel equalization for dereverberation as well as for joint dereverberation and noise reduction.

Speech dereverberation using acoustic channel equalization can be considered to be a two-stage approach. As is schematically illustrated in Fig. 1.7 for a multi-channel scenario, in the first stage the RIRs are measured or estimated using supervised system identification (SSI) methods [101] or blind system identification (BSI) methods [102–107]. Using the measured or estimated RIRs, in the second stage equal-

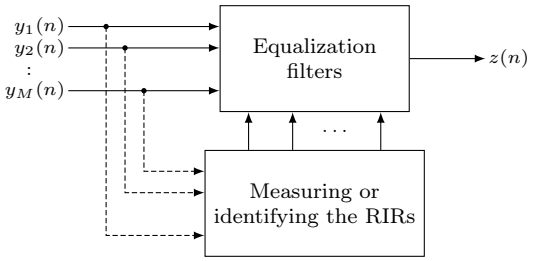


Fig. 1.7: Schematic illustration of acoustic multi-channel equalization techniques for dereverberation. $y_m(n)$ denotes the m-th received microphone signal, $m = 1, 2, \ldots, M$, with M the number of microphones, and $z(n)$ denotes the output speech signal.

ization filters aiming to invert or partially reshape the RIRs are applied to the microphone signals, such that the output speech signal is equal to the clean speech signal or to the early reverberation component in one of the microphone signals.

As already mentioned, acoustic channel equalization comprises an attractive approach to speech dereverberation, since in theory perfect dereverberation can be achieved. As will be described in the following, in practice however, such techniques suffer from a number of drawbacks, i.e.,

i) the measured or estimated RIRs are generally perturbed from the true RIRs due to temperature variations [108], spatial mismatch in the source-microphone geometry [109], or due to the sensitivity of SSI or BSI methods to interfering noise [110, 111]. Acoustic channel equalization techniques can be very sensitive to such perturbations, failing to achieve dereverberation and leading to additional distortions in the output speech signal,

ii) the so-called partial equalization techniques, which aim at partially reshaping the RIRs such that only the late reverberation is suppressed, may lead to undesired perceptual effects, and

iii) acoustic channel equalization techniques design dereverberation filters without taking the presence of the background noise into account, which may lead to noise amplification.

The aim of this thesis is to deal with all these drawbacks by designing **robust** and **perceptually advantageous acoustic multi-channel equalization techniques** for speech **dereverberation** as well as for **joint dereverberation and noise reduction**.

In the past decades, several equalization techniques have been investigated, which can be classified into single-channel and multi-channel techniques, fixed and adaptive versions, time-domain and frequency-domain implementations, and complete equalization techniques aiming to recover the clean speech signal as well as partial equalization techniques aiming to recover the early reverberation component in one of the microphone signals. In the following, an overview of several proposed equalization techniques is presented. For the sake of clarity, we categorize acoustic channel equalization techniques into single-channel and multi-channel techniques.

Single-channel equalization

Single-channel equalization techniques aim at designing an inverse filter, such that the effect of a single RIR is inverted and the (possibly delayed) clean speech signal is recovered. Although this may seem straightforward, designing an inverse filter is quite challenging in practice. Since acoustic transfer functions are generally mixed-phase systems [35–37], a stable and causal inverse filter does not exist. As a result, approximate time-domain inverse filtering techniques such as single-channel least-squares (SCLS) [112] and homomorphic inverse filtering [112, 113] have been investigated. In the SCLS technique, the least-squares error between the response of the system, i.e., the convolution of the RIR and the filter, and a desired response is minimized. In homomorphic inverse filtering, the RIR is first decomposed into

a minimum-phase component and a maximum-phase component. An exact time-domain inverse filter is designed for the minimum-phase component, whereas the maximum-phase component is only approximately inverted using truncation. In a comparative study between the SCLS and homomorphic inverse filtering techniques in [112], it has been concluded that the SCLS technique yields a better dereverberation performance. However, this technique still results in several drawbacks in practice. In order to achieve a good dereverberation performance, the SCLS inverse filter needs to be several thousand samples long, resulting in a computationally complex and often infeasible inverse filter design [114]. Furthermore, independently of the inverse filter length, the least-squares error between the response of the system and the desired response remains larger than 0, since a perfect inverse filter cannot be designed. Most importantly, the SCLS technique is sensitive to RIR perturbations, failing to achieve dereverberation and resulting in annoying distortions even in the presence of moderate RIR perturbation levels [37, 114]. In [37] we have proposed a novel single-channel equalization technique which i) resolves the computational complexity issues associated with the inverse filter design by operating in the frequency-domain, ii) alleviates the stability issues by incorporating regularization, and iii) partly suppresses the pre-echoes in the output speech signal which arise due to the acausality of the inverse filter by using a single-channel spectral enhancement scheme based on [115, 116]. Nevertheless, this technique still remains quite sensitive to RIR perturbations.

Multi-channel equalization

Although the RIRs are non-minimum phase, it has been shown that using multiple microphones it is possible to perfectly invert an acoustic system [36, 117]. Under the condition that the RIRs do not share any common zeros, perfect dereverberation can be achieved based on the time-domain multiple-input/output inverse theorem (MINT) [36]. Unlike the single-channel case, the length of the dereverberation filters is similar to the length of the RIRs, resulting in a computationally feasible inverse filter design. However, these inverse filters have been shown to be very sensitive to even moderate RIR perturbation levels [109, 114].

Several techniques have been proposed to improve the robustness of the MINT technique as well as to further increase the computational efficiency of the inverse filter design. With the aim of reducing the energy of the inverse filters, and hence, increasing the robustness of the MINT technique against RIR perturbations, the incorporation of regularization has been investigated in [108]. With the aim of designing approximate inverse filters which are less sensitive to near-common zeroes, in [118] adaptive time-domain multi-channel equalization techniques have been proposed, where the inverse filters are iteratively computed using a similar cost function as for the MINT technique. Such techniques however suffer from slow convergence, and methods to improve the convergence rate have been investigated in [119, 120]. In order to further increase the computational efficiency of the inverse filter design, frequency-domain multi-channel inverse filtering techniques have been proposed in [121, 122], where in [122] the Karhunen-Loéve transform domain has been

considered. Furthermore, in [123] multi-channel inversion using decimated and oversampled subbands has been investigated, where the RIRs are decomposed into equivalent subband filters prior to inversion. Such approaches improve the robustness of the MINT technique against RIR perturbations and result in a computationally more efficient filter design procedure. However, acoustic multi-channel equalization using the MINT technique or its adaptive and subband versions nevertheless remains rather sensitive to RIR perturbations.

Another possibility to increase the robustness against RIR perturbations is to relax the constraints on the inverse filter design by using so-called partial equalization techniques. Since early reflections tend to improve speech intelligibility [5,22,26,27] and late reflections are the major cause of speech intelligibility degradation [3–5], the objective of such techniques is to relax the constraints on the filter design by suppressing only the late reflections.

The first partial equalization technique proposed in the context of speech dereverberation is the channel shortening (CS) technique [124,125]. The CS technique uses an energy-based optimization criterion, designing a reshaping filter which maximizes the energy of the direct path and early reflections of the response of the system, while minimizing the energy of the late reflections. Instead of using an energy ratio optimization criterion, the relaxed multi-channel least-squares (RMCLS) technique proposed in [111,125] uses a weighted least-squares optimization criterion to achieve partial equalization. The RMCLS technique aims at setting the late reflections of the response of the system to zero, while not imposing any constraints on the early reflections. As is experimentally validated in [111, 125], relaxing the constraints on the reshaping filter design by aiming at suppressing only the late reflections can yield a significant increase in robustness against RIR perturbations. However, by not imposing any constraints on the remaining early reflections, the CS and RMCLS techniques may lead to undesired perceptual effects [126].

Other than sensitivity to RIR perturbations and uncontrolled perceptual effects, another drawback of acoustic multi-channel equalization techniques is their sensitivity to background noise. In contrast to the previously presented spectral enhancement, beamforming and multi-channel Wiener filtering, and blind probabilistic modeling-based techniques which have been successfully extended to achieve joint dereverberation and noise reduction, joint dereverberation and noise reduction based on acoustic multi-channel equalization remains an unexplored area. More importantly, acoustic multi-channel equalization filters have traditionally been designed without taking the presence of the background noise into account, which often results in a significant noise amplification.

1.5 Outline of the thesis and main contributions

Motivated by the potential to achieve a perfect dereverberation performance, this thesis deals with acoustic multi-channel equalization techniques. As already mentioned, these techniques suffer from several drawbacks in practice, i.e., i) sensitivity to RIR perturbations, ii) uncontrolled perceptual effects, and iii) background noise amplification. In this thesis all these problems will be dealt with by proposing **ro-**

bust and **perceptually advantageous acoustic multi-channel equalization techniques** for speech **dereverberation** as well as for **joint dereverberation and noise reduction**.

The main contributions of this thesis are threefold. First, we have **developed a perceptually advantageous** partial acoustic multi-channel equalization technique which aims at not only suppressing the late reflections but also at controlling the early reflections. Second, we have **derived several signal-independent and signal-dependent methods to increase the robustness** of equalization techniques against RIR perturbations, i.e., by **decreasing the reshaping filter length**, by **using (automatic) regularization**, and by **incorporating sparsity-promoting penalty functions**. We have **extensively evaluated** all proposed techniques using instrumental performance measures and the most promising techniques are compared in a subjective listening test. Third, we have **effectively integrated the dereverberation and noise reduction tasks** using robust acoustic multi-channel equalization.

In the remainder of this section a chapter by chapter overview of this thesis is presented, summarizing the main contributions. Additionally, references to the publications that have been produced in the context of this thesis are provided. A schematic overview of the thesis is given in Fig. 1.8.

In **Chapter 2** we describe the general signal processing aspects associated with multi-channel speech enhancement systems. We present time-domain and frequency-domain signal models and mathematically formulate the objective of dereverberation, noise reduction, as well as joint dereverberation and noise reduction. Furthermore, we provide an overview of the typically arising perturbations for measured or estimated RIRs. Finally, we present the instrumental performance measures used in this thesis to assess the dereverberation, noise reduction, as well as joint dereverberation and noise reduction performance.

In **Chapter 3** we propose a least-squares **perceptually advantageous** acoustic multi-channel equalization technique, referred to as the PMINT technique, which aims at suppressing the late reflections while preserving the early reflections. Furthermore, we derive a generalized framework for least-squares equalization techniques, i.e., for the MINT, RMCLS, and PMINT techniques. In addition, we present simulation results to evaluate the dereverberation performance of the proposed PMINT technique and state-of-the-art acoustic multi-channel equalization techniques. These results illustrate the importance of preserving the early reflections to improve the perceptual speech quality as well as the necessity to further increase the robustness of all considered acoustic multi-channel equalization techniques against RIR perturbations. Publications related to this chapter are [126–129].

In **Chapters 4, 5**, and **6** we propose several signal-independent and signal-dependent methods to increase the robustness of the PMINT technique and state-of-the-art acoustic multi-channel equalization techniques, i.e., MINT, CS, and RMCLS, against RIR perturbations.

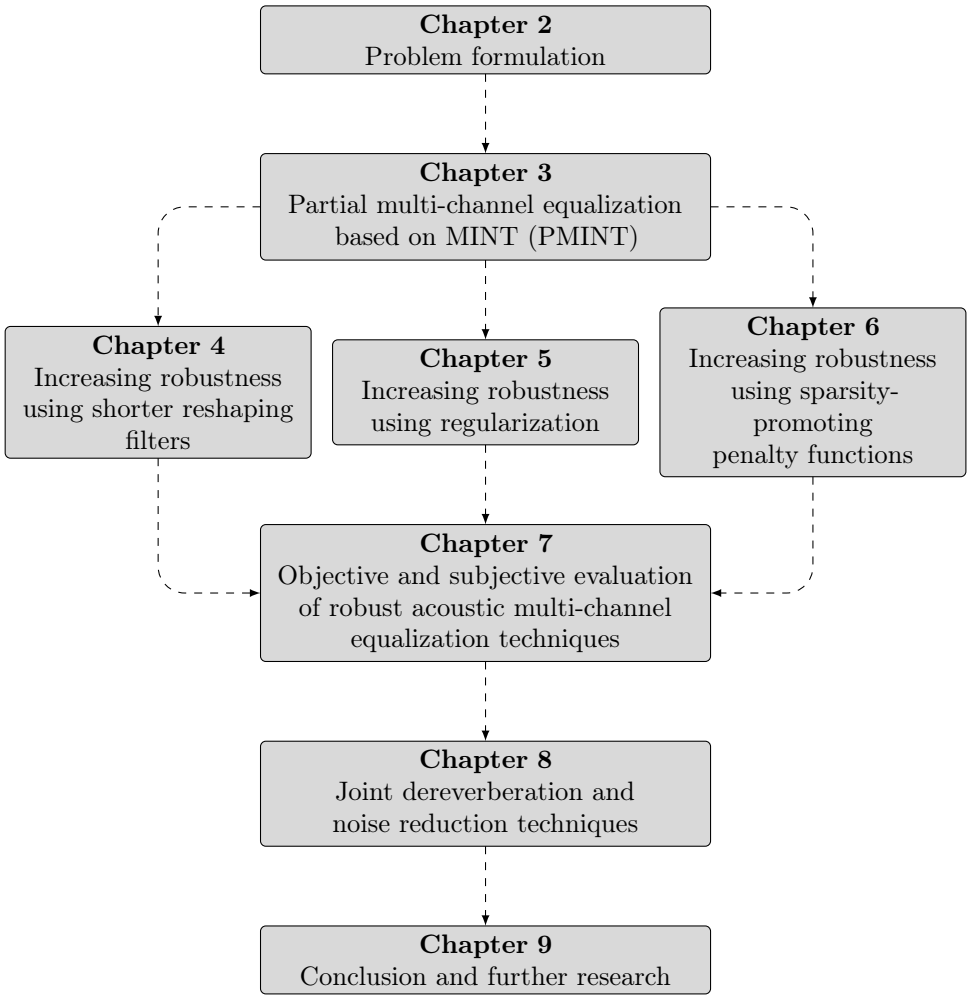


Fig. 1.8: Schematic overview of the thesis.

- In **Chapter 4** we propose a signal-independent method to increase the robustness of equalization techniques by **decreasing the reshaping filter length**, aiming to improve the conditioning of the optimization criteria. We derive analytical expressions showing that using shorter reshaping filters increases the robustness of the MINT, CS, RMCLS, and PMINT techniques against RIR perturbations, which are validated using simulation results. The publication related to this chapter is [130].

- In **Chapter 5** we propose another signal-independent method to increase the robustness of equalization techniques by **using regularization**, aiming to reduce the energy of distortions due to RIR perturbations. Furthermore, we propose and investigate an automatic non-intrusive procedure for determining

the regularization parameter. Simulation results validate that incorporating regularization significantly increases the dereverberation performance of the MINT, CS, RMCLS, and PMINT techniques in the presence of RIR perturbations. Moreover, it is shown that the proposed non-intrusive procedure for determining an automatic regularization parameter yields a similar performance as the optimal intrusively determined regularization parameter. Publications related to this chapter are [126, 127, 131, 132].

- In **Chapter 6** we propose a signal-dependent method to increase the robustness of equalization techniques by **using sparsity-promoting penalty functions**, aiming to promote sparsity in the output speech signal and reduce artifacts generated by non-robust techniques. We investigate several sparsity-promoting penalty functions and present insights on the advantages of using such penalty functions for speech dereverberation. Furthermore, we establish iterative optimization procedures for computing the sparsity-promoting dereverberation filters. Simulation results validate that incorporating sparsity-promoting penalty functions enables to increase the robustness of the MINT, CS, RMCLS, and PMINT techniques against RIR perturbations. Publications related to this chapter are [133, 134].

In **Chapter 7** we compare the performance of all proposed robust extensions of acoustic multi-channel equalization techniques (Chapters 4, 5, and 6) for several acoustic scenarios, both using instrumental performance measures as well as using subjective listening tests. Instrumental performance measures show that the regularized and sparsity-promoting RMCLS and PMINT techniques achieve the highest dereverberation performance. The subjective listening test shows that the robust extensions of the PMINT technique are generally preferred over the robust extensions of the RMCLS technique, with the sparsity-promoting PMINT technique yielding the best perceptual speech quality for low RIR perturbation levels and the regularized PMINT technique yielding the best perceptual speech quality for high RIR perturbation levels. The publication related to this chapter is [135].

Based on the previously developed robust acoustic multi-channel equalization techniques, in **Chapter 8** we propose two techniques for **joint dereverberation and noise reduction**. The first technique directly extends the regularized equalization techniques by incorporating the noise statistics into the reshaping filter design. In addition to the regularization parameter, a weighting parameter is introduced which enables to trade off between dereverberation and noise reduction. The second technique is based on the MWF, where the dereverberated reference signal for the MWF is generated using regularized equalization techniques. Also for the MWF-based technique, an additional weighting parameter is introduced which now enables to trade off between speech distortion and noise reduction. In addition, we propose automatic non-intrusive procedures for determining the regularization and weighting parameters for both techniques. Extensive simulations validate the importance of incorporating the noise statistics into the reshaping filter design to

achieve joint dereverberation and noise reduction. Publications related to this chapter are [136–138].

Chapter 9 summarizes the main contributions of the thesis and presents suggestions for further research.

2

PROBLEM FORMULATION AND INSTRUMENTAL PERFORMANCE MEASURES

In this chapter the general notation, the time and frequency domain signal models, and the instrumental performance measures used in the remainder of the thesis are presented.

Section 2.1 describes the general time and frequency domain models for speech signals recorded in a reverberant and noisy environment. Furthermore, the typical multi-channel speech enhancement configuration is presented and the objective of dereverberation, noise reduction, as well as joint dereverberation and noise reduction is mathematically formulated. Section 2.2 provides an overview of the typical perturbations that arise when measuring or estimating room impulse responses (RIR). Furthermore, a description of how RIR perturbations will be simulated in the remainder of this thesis is provided. Section 2.3 presents the instrumental performance measures used to assess the dereverberation performance, the noise reduction performance, as well as the joint dereverberation and noise reduction performance of the techniques proposed in this thesis.

2.1 Problem formulation

2.1.1 *Time domain signal model*

Consider the acoustic scenario depicted in Fig. 1.1, consisting of a single speech source, M microphones, and background noise. Each microphone signal $y_m(n)$, $m = 1, \ldots, M$, at discrete time index n, consists of a filtered version of the clean speech signal $s(n)$ and a noise component $v_m(n)$, i.e.,

$$y_m(n) = h_m(n) * s(n) + v_m(n) = x_m(n) + v_m(n), \tag{2.1}$$

where $h_m(n)$ denotes the room impulse response (RIR) between the speech source and the m-th microphone, $x_m(n)$ denotes the reverberant speech component at the m-th microphone, and $*$ denotes convolution. As described in Section 1.2, the RIR

$h_m(n)$ consists of a direct path and early reflections component $h_{e,m}(n)$ and a late reflections component $h_{r,m}(n)$, i.e.,

$$h_m(n) = h_{e,m}(n) + h_{r,m}(n). \tag{2.2}$$

Using (2.2), the received microphone signal in (2.1) can be written as

$$y_m(n) = \underbrace{h_{e,m}(n) * s(n)}_{x_{e,m}(n)} + \underbrace{h_{r,m}(n) * s(n)}_{x_{r,m}(n)} + v_m(n), \tag{2.3}$$

with $x_{e,m}(n)$ the early reverberation speech component and $x_{r,m}(n)$ the late reverberation speech component at the m-th microphone.

In the remainder of this thesis it is assumed that the RIRs are time-invariant. This is a commonly used assumption in several dereverberation techniques and is appropriate for acoustic scenarios where the source-microphone geometry does not rapidly vary and the environmental conditions are fixed. It should be noted however that there are many applications where the source-microphone geometry rapidly changes, e.g., due to a moving talker, or the environmental conditions change, e.g., due to the opening and closing of doors and windows. Nevertheless, acoustic multi-channel equalization techniques can in principle be extended to equalize time-varying RIRs, as long as the time-varying RIRs can be accurately estimated and tracked by system identification techniques (which is currently not possible by state-of-the-art system identification techniques).

Using a finite impulse response (FIR) filter model, the m-th RIR $\mathbf{h}_m$ can be written as

$$\mathbf{h}_m = [h_m(0)\ h_m(1)\ \ldots\ h_m(L_h - 1)]^T, \tag{2.4}$$

with L_h the RIR length. Using (2.4), the convolution in (2.1) is computed as

$$x_m(n) = \sum_{i=0}^{L_h-1} h_m(i)s(n-i). \tag{2.5}$$

The noise component $v_m(n)$ in (2.1) is assumed to be uncorrelated with the speech component $x_m(n)$ and can consist of directional interferences (e.g., fans, electronic appliances, or other speakers), spatially diffuse noise which can be commonly found in an office or a car environment [66, 139], and microphone self-noise. The noise may not only be detrimental to speech intelligibility, but it may also severely affect the performance of supervised system identification (SSI) and blind system identification (BSI) methods, as will be shown in Section 2.2.

2.1.2 *Time domain speech enhancement*

Fig. 2.1 depicts the typical filter-and-sum structure used for multi-channel speech enhancement. Applying filters $w_m(n)$ to the received microphone signals, the output speech signal $z(n)$ is given by

$$z(n) = \sum_{m=1}^{M} y_m(n) * w_m(n). \tag{2.6}$$

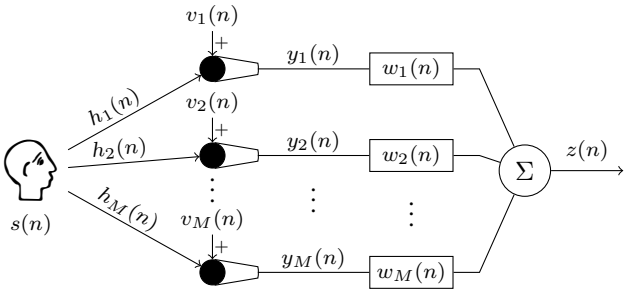


Fig. 2.1: Schematic illustration of a typical time domain multi-channel speech enhancement system.

Using (2.1), the output speech signal can be written as

$$z(n) = \underbrace{\sum_{m=1}^{M} x_m(n) * w_m(n)}_{z_x(n)} + \underbrace{\sum_{m=1}^{M} v_m(n) * w_m(n)}_{z_v(n)}, \tag{2.7}$$

with $z_x(n)$ the output speech component and $z_v(n)$ the output noise component. Similarly, the early reverberation output speech component $z_{e,x}(n)$ and the late reverberation output speech component $z_{r,x}(n)$ are defined as

$$z_{e,x}(n) = \sum_{m=1}^{M} x_{e,m}(n) * w_m(n), \tag{2.8}$$

$$z_{r,x}(n) = \sum_{m=1}^{M} x_{r,m}(n) * w_m(n). \tag{2.9}$$

Writing the speech component $x_m(n)$ in (2.7) in terms of the clean speech signal and the RIR, the output speech component $z_x(n)$ can also be expressed as

$$z_x(n) = \sum_{m=1}^{M} s(n) * h_m(n) * w_m(n) \tag{2.10}$$

$$= s(n) * \underbrace{\sum_{m=1}^{M} h_m(n) * w_m(n)}_{c(n)}, \tag{2.11}$$

where $c(n)$ is referred to as the *equalized impulse response* (EIR) between the clean speech signal $s(n)$ and the output speech component $z_x(n)$. The equalized impulse response $c(n)$ can be used to describe the dereverberation performance of the speech enhancement system. For example, complete dereverberation is achieved if $c(n)$ is equal to a (possibly delayed) impulse, whereas partial dereverberation is achieved if $c(n)$ contains only early reflections.

Because FIR filters are inherently stable, in this thesis we will consider time-invariant FIR filters $\mathbf{w}_m$ of length L_w, i.e.,

$$\mathbf{w}_m = [w_m(0)\ w_m(1)\ \ldots\ w_m(L_w-1)]^T. \tag{2.12}$$

Considering the L_w–dimensional received signal vector $\mathbf{y}_m(n)$, i.e.,

$$\mathbf{y}_m(n) = [y_m(n)\ y_m(n-1)\ \ldots\ y_m(n-L_w+1)]^T, \tag{2.13}$$

and the ML_w–dimensional stacked filter vector $\mathbf{w}$ and stacked received signal vector $\mathbf{y}(n)$, i.e.,

$$\mathbf{w} = [\mathbf{w}_1^T\ \mathbf{w}_2^T\ \ldots\ \mathbf{w}_M^T]^T, \tag{2.14}$$

$$\mathbf{y}(n) = [\mathbf{y}_1^T(n)\ \mathbf{y}_2^T(n)\ \ldots\ \mathbf{y}_M^T(n)]^T, \tag{2.15}$$

the output speech signal $z(n)$ can be expressed as

$$z(n) = \sum_{m=1}^{M} \mathbf{w}_m^T \mathbf{y}_m(n) = \mathbf{w}^T \mathbf{y}(n). \tag{2.16}$$

Defining the stacked reverberant speech vector $\mathbf{x}(n)$ and the stacked noise vector $\mathbf{v}(n)$ similarly as in (2.13) and (2.15), i.e.,

$$\mathbf{x}_m(n) = [x_m(n)\ x_m(n-1)\ \ldots\ x_m(n-L_w+1)]^T, \tag{2.17}$$

$$\mathbf{x}(n) = [\mathbf{x}_1^T(n)\ \mathbf{x}_2^T(n)\ \ldots\ \mathbf{x}_M^T(n)]^T, \tag{2.18}$$

and

$$\mathbf{v}_m(n) = [v_m(n)\ v_m(n-1)\ \ldots\ v_m(n-L_w+1)]^T, \tag{2.19}$$

$$\mathbf{v}(n) = [\mathbf{v}_1^T(n)\ \mathbf{v}_2^T(n)\ \ldots\ \mathbf{v}_M^T(n)]^T, \tag{2.20}$$

the output speech signal $z(n)$ can be written as

$$z(n) = \sum_{m=1}^{M} \mathbf{w}_m^T \mathbf{x}_m(n) + \sum_{m=1}^{M} \mathbf{w}_m^T \mathbf{v}_m(n) \tag{2.21}$$

$$= \underbrace{\mathbf{w}^T \mathbf{x}(n)}_{z_x(n)} + \underbrace{\mathbf{w}^T \mathbf{v}(n)}_{z_v(n)}. \tag{2.22}$$

Using the clean speech vector $\mathbf{s}(n)$ of length $L_c = L_h + L_w - 1$, i.e.,

$$\mathbf{s}(n) = [s(n)\ s(n-1)\ \ldots\ s(n-L_c-1)]^T, \tag{2.23}$$

and the $L_c \times ML_w$–dimensional multi-channel convolution matrix $\mathbf{H}$, i.e., $\mathbf{H} = [\mathbf{H}_1\ \mathbf{H}_2\ \dots\ \mathbf{H}_M]$, with

$$\mathbf{H}_m = \begin{bmatrix} h_m(0) & 0 & \cdots & 0 \\ h_m(1) & h_m(0) & \ddots & \vdots \\ \vdots & h_m(1) & \ddots & 0 \\ h_m(L_h-1) & \vdots & \ddots & h_m(0) \\ 0 & h_m(L_h-1) & \ddots & h_m(1) \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & h_m(L_h-1) \end{bmatrix}, \tag{2.24}$$

the stacked reverberant speech vector $\mathbf{x}(n)$ can also be expressed as

$$\mathbf{x}(n) = \mathbf{H}^T\mathbf{s}(n). \tag{2.25}$$

Substituting (2.25) in (2.22), the output speech signal can be further written as

$$z(n) = \underbrace{\mathbf{w}^T\mathbf{H}^T}_{\mathbf{c}^T}\mathbf{s}(n) + \mathbf{w}^T\mathbf{v}(n), \tag{2.26}$$

with $\mathbf{c} = [c(0)\ c(1)\ \dots\ c(L_c-1)]^T$ the equalized impulse response in vector notation, i.e.,

$$\mathbf{c} = \mathbf{H}\mathbf{w}. \tag{2.27}$$

The filter $\mathbf{w}$ can now be designed based on different objectives, i.e., i) aiming only at dereverberation, ii) aiming only at noise reduction, or iii) aiming at joint dereverberation and noise reduction.

i) The objective of *complete dereverberation* techniques is to design an inverse filter $\mathbf{w}$ such that the equalized impulse response $\mathbf{c}$ is equal to a (possibly delayed) impulse and the output speech component $z_x(n)$ is equal to the (possibly delayed) clean speech signal $s(n)$. The objective of *partial dereverberation* techniques is to design a reshaping filter $\mathbf{w}$ such that the equalized impulse response $\mathbf{c}$ contains only early reflections and the output speech component $z_x(n)$ contains only early reverberation. It should be noted that designing filters which aim only at speech dereverberation and do not take the background noise into account can result in noise amplification (cf. Section 8.4.3).

ii) The objective of *noise reduction* techniques is to design a filter $\mathbf{w}$ such that the power of the output noise component $z_v(n)$ is minimized (or ideally set to 0), while taking into account speech distortion. Several effective multi-microphone noise reduction techniques have been developed over the past decades [10–19], which however may not dereverberate the output speech signal.

iii) The objective of *joint dereverberation and noise reduction* techniques is to design a filter $\mathbf{w}$ which yields a dereverberated equalized impulse response $\mathbf{c}$ as well as minimizes the power of the output noise component $z_v(n)$.

While the noise reduction task has been extensively investigated, dereverberation and the effective integration of dereverberation and noise reduction has received much less attention until recently. In this thesis, Chapters 3-7 focus on developing robust and perceptually advantageous acoustic multi-channel equalization techniques for speech dereverberation, while Chapter 8 focuses on the effective integration of the developed dereverberation techniques with noise reduction.

2.1.3 *Frequency domain representation*

Most techniques discussed in this thesis will be based on the time domain signal model presented in Sections 2.1.1 and 2.1.2. However, in Chapter 6 and Appendix B also a frequency domain signal model will be used.

For completeness, in the following we briefly review the definitions of the discrete-time Fourier transform (DTFT), inverse discrete-time Fourier transform (IDTFT), short-time Fourier transform (STFT), and inverse short-time Fourier transform (ISTFT), cf. e.g., [140, 141].

Discrete-time Fourier transform: The DTFT $B(\omega)$ of the discrete-time signal $b(n)$ is defined as

$$B(\omega) = \sum_{n=-\infty}^{\infty} b(n)e^{-jn\omega}, \tag{2.28}$$

with ω the angular frequency ($-\pi < \omega \leq \pi$) and j the imaginary unit, i.e., $j^2 = -1$.

Inverse discrete-time Fourier transform: The discrete-time signal $b(n)$ can be recovered from $B(\omega)$ using the IDTFT, defined as

$$b(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} B(\omega)e^{jn\omega}d\omega. \tag{2.29}$$

Short-time Fourier transform: In practice, the continuous spectrum in (2.28) is approximated using the STFT, where frames of the discrete-time signal $b(n)$ are weighted by an analysis window $w_{\text{STFT}}(n)$ and transformed as

$$B(t,f) = \sum_{n=0}^{N-1} w_{\text{STFT}}(n)b(tR+n)e^{\frac{-j2\pi fn}{N}}, \tag{2.30}$$

with t the time frame index, f the frequency bin index, N the frame size, and R the frame shift.

Inverse short-time Fourier transform: The ISTFT is defined as

$$b(n) = \sum_{t}\sum_{f=0}^{N-1} w_{\text{ISTFT}}(n-tR)B(t,f)e^{\frac{j2\pi f(n-tR)}{N}}, \tag{2.31}$$

with $w_{\text{ISTFT}}(n)$ a synthesis window such that the perfect overlap-add constraint is satisfied.

Using the DTFT, the signal model in (2.1) can be written in the frequency domain as

$$Y_m(\omega) = H_m(\omega)S(\omega) + V_m(\omega) = X_m(\omega) + V_m(\omega), \tag{2.32}$$

with $Y_m(\omega)$, $H_m(\omega)$, $S(\omega)$, $V_m(\omega)$, and $X_m(\omega)$ the DTFTs of $y_m(n)$, $h_m(n)$, $s(n)$, $v_m(n)$, and $x_m(n)$, respectively. Defining the M–dimensional vector $\mathbf{y}(\omega)$ as

$$\mathbf{y}(\omega) = [Y_1(\omega)\ Y_2(\omega)\ \ldots\ Y_M(\omega)]^T, \tag{2.33}$$

and defining the M–dimensional vectors $\mathbf{h}(\omega)$, $\mathbf{x}(\omega)$, and $\mathbf{v}(\omega)$ similarly as in (2.33), the frequency domain signal model in (2.32) can be written in vector notation as

$$\mathbf{y}(\omega) = \mathbf{h}(\omega)S(\omega) + \mathbf{v}(\omega) = \mathbf{x}(\omega) + \mathbf{v}(\omega). \tag{2.34}$$

Furthermore, using the M–dimensional filter vector $\mathbf{w}(\omega)$, i.e.,

$$\mathbf{w}(\omega) = [W_1(\omega)\ W_2(\omega)\ \ldots\ W_M(\omega)]^T, \tag{2.35}$$

with $W_m(\omega)$ the DTFT of $w_m(n)$, the output speech signal $Z(\omega)$ can be expressed as

$$Z(\omega) = \mathbf{w}^T(\omega)\mathbf{y}(\omega) = \underbrace{\mathbf{w}^T(\omega)\mathbf{h}(\omega)}_{C(\omega)} S(\omega) + \mathbf{w}^T(\omega)\mathbf{v}(\omega), \tag{2.36}$$

with $C(\omega)$ the DTFT of the equalized impulse response. For convenience, (2.36) is typically written as [142]

$$Z(\omega) = \mathbf{w}^H(\omega)\mathbf{y}(\omega) = \sum_{m=1}^{M} W_m^*(\omega)Y_m(\omega), \tag{2.37}$$

with $\{\cdot\}^*$ denoting the complex conjugate and $\{\cdot\}^H$ denoting the complex conjugate transpose.

In practical speech enhancement systems operating in the frequency domain, the continuous frequency spectrum in (2.37) is approximated using the short-time Fourier transform. Using the filter STFT coefficients $W_m(f)$ and the microphone signal STFT coefficients $Y_m(t,f)$ defined similarly as in (2.30), the output speech signal STFT coefficients $Z(t,f)$ are given by

$$Z(t,f) = \sum_{m=1}^{M} W_m^*(f)Y_m(t,f) = \mathbf{w}^H(f)\mathbf{y}(t,f), \tag{2.38}$$

with $\mathbf{w}(f)$ and $\mathbf{y}(t,f)$ the M–dimensional vectors of the filter and microphone signal STFT coefficients.[1]

1 Note that time-invariant filters are assumed in this thesis, hence, the filter STFT coefficients $W_m(f)$ are independent of the time frame index t.

2.2 Room impulse response perturbations

As described in Section 1.5, one of the objectives of this thesis is to derive acoustic multi-channel equalization techniques which are robust against RIR perturbations. In this section we will provide some insights on the reasons why such perturbations arise when estimating or measuring RIRs. We distinguish between perturbations arising due to the sensitivity of SSI or BSI methods to interfering noise, and perturbations arising due to spatial mismatch in the source-microphone geometry.

2.2.1 *Supervised system identification (SSI)*

In order to avoid cumbersome notation, *note that only in this section* the vectors $\mathbf{y}_m(n)$, $\mathbf{x}_m(n)$, and $\mathbf{v}_m(n)$ will refer to L_y–dimensional vectors instead of the L_w–dimensional vectors defined in Section 2.1 (where it is typically assumed that $L_y \gg L_w$), i.e.,

$$\mathbf{y}_m(n) = [y_m(n)\ y_m(n-1)\ \ldots\ y_m(n-L_y+1)]^T, \tag{2.39}$$

and $\mathbf{x}_m(n)$ and $\mathbf{v}_m(n)$ defined similarly.

Supervised system identification refers to estimating the RIRs using knowledge of both the clean speech signal $s(n)$ and the microphone signal $y_m(n)$. When the clean speech signal is available, the estimation of the RIRs can be done individually for each RIR $\mathbf{h}_m$ using a standard least-squares approach [101].

The m-th received signal vector $\mathbf{y}_m(n)$ can be expressed as

$$\mathbf{y}_m(n) = \mathbf{S}(n)\mathbf{h}_m + \mathbf{v}_m(n), \tag{2.40}$$

with $\mathbf{y}_m(n)$, $\mathbf{x}_m(n)$, $\mathbf{v}_m(n)$ defined as in (2.39), $\mathbf{h}_m$ defined as in (2.4), and $\mathbf{S}(n)$ the $L_y \times L_h$–dimensional matrix

$$\mathbf{S}(n) = \begin{bmatrix} s(n) & s(n-1) & \cdots & s(n-L_h+1) \\ s(n-1) & s(n-2) & \cdots & s(n-L_h) \\ \vdots & \vdots & \ddots & \vdots \\ s(n-L_y+1) & s(n-L_y) & \cdots & s(n-L_h-L_y+2) \end{bmatrix}. \tag{2.41}$$

Using (2.40), the RIR $\mathbf{h}_m$ can be computed by minimizing the least-squares cost function [101, 111]

$$J = \|\mathbf{S}(n)\mathbf{h}_m - \mathbf{y}_m(n)\|_2^2, \tag{2.42}$$

resulting in the least-squares estimate[2]

$$\hat{\mathbf{h}}_m = \left[\mathbf{S}^T(n)\mathbf{S}(n)\right]^{-1}\mathbf{S}^T(n)\mathbf{y}_m(n). \tag{2.43}$$

2 Note that in order for the matrix $\mathbf{S}^T(n)\mathbf{S}(n)$ to be invertible, the matrix $\mathbf{S}(n)$ should be a full column-rank matrix.

The least-squares estimate in (2.43) can be expressed as

$$\hat{\mathbf{h}}_m = \left[\mathbf{S}^T(n)\mathbf{S}(n)\right]^{-1}\mathbf{S}^T(n)\mathbf{x}_m(n) + \left[\mathbf{S}^T(n)\mathbf{S}(n)\right]^{-1}\mathbf{S}^T(n)\mathbf{v}_m(n) \tag{2.44}$$

$$= \mathbf{h}_m + \underbrace{\left[\mathbf{S}^T(n)\mathbf{S}(n)\right]^{-1}\mathbf{S}^T(n)\mathbf{v}_m(n)}_{\mathbf{e}_m}, \tag{2.45}$$

which shows that the background noise $\mathbf{v}_m(n)$ causes the least-squares RIR estimate $\hat{\mathbf{h}}_m$ to differ from the true RIR $\mathbf{h}_m$ by $\mathbf{e}_m = \left[\mathbf{S}^T(n)\mathbf{S}(n)\right]^{-1}\mathbf{S}^T(n)\mathbf{v}_m(n)$. Hence, the background noise $\mathbf{v}_m(n)$ affects the performance of least-squares supervised system identification methods, particularly at low input signal-to-noise ratios and when the number of the available data samples L_y is small. Several approaches to increase the robustness of SSI methods against noise have been investigated, e.g., in [143,144]. In addition, based on different assumptions about the matrix $\mathbf{S}(n)$ and the noise $\mathbf{v}_m(n)$, models for the perturbations $\mathbf{e}_m$ arising from supervised system identification methods can be derived. For example, in [111] it is assumed that the noise is uncorrelated microphone self-noise and the perturbations $\mathbf{e}_m$ are modeled as a spatially white Gaussian noise sequence with a long-term average speech spectrum.

2.2.2 *Blind system identification (BSI)*

In order to avoid cumbersome notation, *note that only in this section* the vector $\mathbf{y}_m(n)$ will refer to an L_h–dimensional vector instead of the L_w–dimensional vector defined in (2.13), i.e.,

$$\mathbf{y}_m(n) = [y_m(n)\ y_m(n-1)\ \ldots\ y_m(n-L_h+1)]^T. \tag{2.46}$$

Commonly used methods for blind multi-channel system identification are based on second-order statistics [102–107]. Such methods can in theory identify the RIRs up to a scaling factor if the RIRs do not share any common zeros and the clean speech signal has a full-rank covariance matrix [102]. One of the first methods proposed for BSI in the context of acoustic channels is the multi-channel least mean squares method [102]. This method is based on the so-called cross-relation error $\mathrm{C_r}(n)$, defined as

$$\mathrm{C_r}(n) = \sum_{i=1}^{M-1}\sum_{m=i+1}^{M}\left[\mathbf{y}_i^T(n)\hat{\mathbf{h}}_m - \mathbf{y}_m^T(n)\hat{\mathbf{h}}_i\right]^2, \tag{2.47}$$

with $\mathbf{y}_m(n)$ defined in (2.46) and $\hat{\mathbf{h}}_m = [\hat{h}_m(0)\ \hat{h}_m(1)\ \ldots\ \hat{h}_m(L_h-1)]^T$ denoting the estimate of the m-th RIR $\mathbf{h}_m$. The estimate $\hat{\mathbf{h}} = [\hat{\mathbf{h}}_1\ \hat{\mathbf{h}}_2\ \ldots\ \hat{\mathbf{h}}_M]^T$ of the stacked RIR vector $\mathbf{h} = [\mathbf{h}_1\ \mathbf{h}_2\ \ldots\ \mathbf{h}_M]^T$ is then computed by minimizing the expected value of the cross-relation error in (2.47), subject to the constraint that the trivial solution, i.e., $\hat{\mathbf{h}} = \mathbf{0}$ is avoided. This minimization problem is equivalent to solving

$$\mathbf{R}(n)\hat{\mathbf{h}} = \mathbf{0}, \tag{2.48}$$

with the $ML_h \times ML_h$–dimensional matrix $\mathbf{R}(n)$ defined as

$$\mathbf{R}(n) = \begin{bmatrix} \sum_{m\neq 1} \mathbf{R}_{\mathbf{y}_m}(n) & -\mathbf{R}_{\mathbf{y}_2\mathbf{y}_1}(n) & \cdots & -\mathbf{R}_{\mathbf{y}_M\mathbf{y}_1}(n) \\ -\mathbf{R}_{\mathbf{y}_1\mathbf{y}_2}(n) & \sum_{m\neq 2} \mathbf{R}_{\mathbf{y}_m}(n) & \cdots & -\mathbf{R}_{\mathbf{y}_M\mathbf{y}_2}(n) \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{R}_{\mathbf{y}_1\mathbf{y}_M}(n) & -\mathbf{R}_{\mathbf{y}_2\mathbf{y}_M}(n) & \cdots & \sum_{m\neq M} \mathbf{R}_{\mathbf{y}_m}(n) \end{bmatrix}, \tag{2.49}$$

where $\mathbf{R}_{\mathbf{y}_m}(n)$ denotes the $L_h \times L_h$–dimensional auto-correlation matrix of the m-th received signal vector and $\mathbf{R}_{\mathbf{y}_m\mathbf{y}_i}(n)$ denotes the $L_h \times L_h$–dimensional cross-correlation matrix between the m-th and i-th received signal vectors. The expression in (2.48) represents the basic system of equations that different second-order statistics-based BSI methods aim to solve in a robust and efficient manner. In the absence of background noise, the null space of the matrix $\mathbf{R}(n)$ is of dimension 1 and is spanned by the stacked true RIR vector $\mathbf{h}$. In this case, the stacked RIR vector $\mathbf{h}$ can be perfectly estimated using the eigenvalue decomposition of the matrix $\mathbf{R}(n)$ and identifying $\mathbf{h}$ as the eigenvector corresponding to the eigenvalue which is equal to 0. To avoid the computation of the eigenvalue decomposition, more efficient adaptive methods can be employed [102, 103]. However, in the presence of background noise this property does not hold anymore and adaptive BSI methods are known to misconverge [110]. While more robust BSI methods addressing the misconvergence problem in the presence of background noise have been proposed, e.g., in [104–107], the sensitivity and misconvergence of BSI methods remains an issue, often yielding RIR estimates $\hat{\mathbf{h}}_m$ that differ from the true RIRs $\mathbf{h}_m$, i.e.,

$$\hat{\mathbf{h}}_m = \mathbf{h}_m + \mathbf{e}_m, \tag{2.50}$$

with $\mathbf{e}_m = [e_m(0)\ e_m(1)\ \ldots\ e_m(L_h-1)]^T$ the vector of RIR perturbations for the m-th channel. The perturbations $e_m(n)$ highly depend on the acoustic system (i.e., the actual RIRs and background noise level) and the BSI method used to estimate the RIRs. However, models have been developed to systematically describe these perturbations. For example, in [145] a statistical model has been developed to characterize the perturbations $e_m(n)$ when using the robust multi-channel frequency domain least mean squares method proposed in [106]. When no assumption about the used BSI method is made, the perturbations are generally assumed to be a spatially white Gaussian noise sequence proportional to the true RIR coefficients [146, 147], i.e.,

$$e_m(n) = v_m^w(n) h_m(n), \tag{2.51}$$

with $v_m^w(n)$ an uncorrelated white Gaussian noise sequence with zero mean and a variance that depends on the energy of the RIR perturbations.

2.2.3 *Spatial mismatch*

Even when the RIRs can be correctly estimated or exactly measured, in the case of a spatial mismatch, e.g., due to the source moving slightly or due to the microphone array being slightly displaced, the RIRs used in equalization techniques for designing

dereverberation filters can differ largely from the true RIRs. In [109] the robustness of equalization filters with respect to changing source or microphone positions has been theoretically analyzed. A frequency-dependent error term for the degradation arising in the equalized impulse response due to spatial mismatch has been derived, showing that even small changes in the source-microphone geometry of a few tenths of the acoustic wavelength can cause large degradation in the equalized impulse response. This analysis has been used in [148,149] to derive a model that characterizes the perturbations arising due to spatial mismatch. For details on the perturbations arising due to spatial mismatch, the reader is referred to [109,148,149].

Summarizing, when estimating or measuring RIRs, perturbations are likely to arise due to the sensitivity of SSI and BSI methods to interfering noise, as well as due to possible spatial mismatch in the source-microphone geometry. In this thesis we will present several methods to increase the robustness of acoustic multi-channel equalization techniques against these RIR perturbations. Increasing the robustness of SSI or BSI methods will not be considered in this thesis.

In the presented simulation results, RIR perturbations will be simulated by adding scaled white Gaussian noise to the true RIRs (cf. (2.51)) using the procedure proposed in [147]. However, it should be noted that the proposed techniques are not particularly tailored towards any specific type of perturbation, and hence, they can also be used when other kinds of RIR perturbations occur.

To quantify the level of RIR perturbations, we will use the normalized projection misalignment (NPM) measure [150], defined as

$$\mathrm{NPM} = 10\log_{10}\frac{\left\|\mathbf{h}_m - \frac{\mathbf{h}_m^T\hat{\mathbf{h}}_m}{\hat{\mathbf{h}}_m^T\hat{\mathbf{h}}_m}\hat{\mathbf{h}}_m\right\|_2^2}{\|\mathbf{h}_m\|_2^2}. \tag{2.52}$$

The normalization factor $\frac{\mathbf{h}_m^T\hat{\mathbf{h}}_m}{\hat{\mathbf{h}}_m^T\hat{\mathbf{h}}_m}$ in (2.52) is used such that the accuracy of the RIR estimate is evaluated independently of any scaling of the RIR (since a scaling factor is not important for the dereverberation performance of acoustic multi-channel equalization techniques). In [107], the reported NPM values achieved by state-of-the-art BSI methods (for relatively short RIRs) in the presence of additive noise range between -10 dB and -20 dB.

2.3 Instrumental performance measures

In this section we present the instrumental performance measures used to assess the dereverberation performance, the noise reduction performance, as well as the joint dereverberation and noise reduction performance of the techniques proposed in this thesis.

2.3.1 *Dereverberation*

Several instrumental performance measures for evaluating the dereverberation performance of speech enhancement techniques have been proposed (cf. [20] and the references therein). When the effect of a dereverberation technique can be represented by a linear filter and the equalized impulse response $c(n)$ can be computed, *channel-based measures* operating on the equalized impulse response can be used. When the effect of a dereverberation technique cannot be represented by a linear filter and the equalized impulse response $c(n)$ cannot be computed, *signal-based measures* operating on the output speech signal are typically used. Since acoustic multi-channel equalization techniques design and apply linear filters and the equalized impulse response $c(n)$ can be directly computed, channel-based measures will be used to evaluate the reverberant energy suppression and the reverberant energy decay rate. However, since channel-based measures do not necessarily correlate well with the perceptual quality of the output speech signal [151, 152], in addition we will also consider signal-based performance measures.

Channel-based measures

The reverberant energy suppression is evaluated using the *direct-to-reverberant ratio* (DRR) improvement [20] between the equalized impulse response $c(n)$ and the input RIR, generally chosen to be the first RIR $h_1(n)$. The DRR improvement, i.e., ΔDRR, is defined as

$$\Delta\text{DRR} = \text{oDRR} - \text{iDRR}, \tag{2.53}$$

with

$$\text{oDRR} = 10\log_{10}\frac{\sum_{n=0}^{n_d-1} c^2(n)}{\sum_{n=n_d}^{L_c-1} c^2(n)}, \qquad \text{iDRR} = 10\log_{10}\frac{\sum_{n=0}^{n_d-1} h_1^2(n)}{\sum_{n=n_d}^{L_h-1} h_1^2(n)}, \tag{2.54}$$

where the first n_d samples of the EIR and RIR represent the direct-path propagation and the remaining samples represent reflections. Although the DRR improvement exactly describes the reverberant energy suppression, it cannot be solely used to evaluate the dereverberation performance of dereverberation techniques, since it does not provide any insights on the reverberant energy decay rate. As an extreme scenario, consider the case where the reverberant samples of the equalized impulse response are time-reversed, i.e., the n_d-th sample becomes the L_c-the sample, the (n_d+1)-th sample becomes the (L_c-1)-th sample, and so on. Although the DRR of the original and time-reversed equalized impulse responses are the same, the decay rate patterns will be very different.

To evaluate the reverberant energy decay rate, the *energy decay curve* (EDC) [20] of the EIR $c(n)$ is compared to the energy decay curve of the input RIR, generally

chosen to be the first RIR $h_1(n)$. The EDC of the equalized impulse response is computed as

$$\mathrm{EDC}_{\mathbf{c}}(n) = \frac{1}{\|\mathbf{c}\|_2^2} \sum_{i=n}^{L_c-1} c^2(i), \quad n = 0,\ 1,\ \ldots,\ L_c - 1, \tag{2.55}$$

and the EDC of the first RIR is computed as

$$\mathrm{EDC}_{\mathbf{h}_1}(n) = \frac{1}{\|\mathbf{h}_1\|_2^2} \sum_{i=n}^{L_h-1} h_1^2(i), \quad n = 0,\ 1,\ \ldots,\ L_h - 1. \tag{2.56}$$

Signal-based measures

In [151, 152] it has been shown that instrumental performance measures relying on auditory models, such as the *perceptual evaluation of speech quality* (PESQ) measure [153], exhibit the highest correlation with subjective listening tests when evaluating the perceptual quality of dereverberation techniques. Furthermore, in [152] it has also been shown that the *cepstral distance* (CD) measure [154], which estimates the log-spectral distance between two spectra, exhibits a high correlation with subjective listening tests when evaluating the perceived amount of reverberation for a wide range of state-of-the-art dereverberation techniques.

In this thesis, both the PESQ and CD measures will be used as signal-based performance measures. Both measures are intrusive measures, comparing the output speech signal with a reference desired signal (usually the original clean speech signal $s(n)$ or the early reverberation speech component in the first microphone $x_{e,1}(n)$). The improvement in PESQ, i.e., ΔPESQ, is computed as the difference between the PESQ score of the output speech component $z_x(n)$ and the PESQ score of the reverberant microphone signal $x_1(n)$. Note that the PESQ score is limited in the range from 1 to 4.5, with a PESQ score of 1 indicating the lowest perceptual speech quality and a PESQ score of 4.5 indicating the highest quality. Similarly, the improvement in CD, i.e., ΔCD, is computed as the difference between the CD of the output speech component $z_x(n)$ and the CD of the reverberant microphone signal $x_1(n)$. Note that the CD measure is limited in the range from 0 dB to 10 dB as in [155], with a CD of 0 dB indicating that the desired and enhanced spectra are the same and a CD of 10 dB indicating a large difference between the desired and enhanced spectra. Hence, a higher PESQ score and a lower cepstral distance indicate an improvement in perceptual speech quality.

2.3.2 *Noise reduction*

Although this thesis mainly deals with acoustic multi-channel equalization techniques for speech dereverberation, for which the previously discussed instrumental performance measures can be used, in Chapter 8 joint dereverberation and noise reduction techniques are proposed. In order to evaluate the noise reduction perfor-

mance of these techniques, the broadband *noise reduction factor* ψ_{NR} [18] is used, defined as

$$\psi_{\text{NR}} = 10\log_{10}\frac{\mathcal{E}\{v_1^2(n)\}}{\mathcal{E}\{z_v^2(n)\}}, \tag{2.57}$$

with $v_1(n)$ the noise component in the first microphone and $z_v(n)$ the output noise component defined in (2.7).

2.3.3 *Joint dereverberation and noise reduction*

The joint dereverberation and noise reduction performance of the techniques discussed in Chapter 8 is evaluated in terms of the broadband *signal-to-reverberation-and-noise ratio* (SRNR) measure. The SRNR improvement, i.e., ΔSRNR, between the output speech signal $z(n)$ and the input signal, generally chosen to be the first microphone signal $y_1(n)$, is defined as

$$\Delta\text{SRNR} = \text{oSRNR} - \text{iSRNR}, \tag{2.58}$$

with

$$\text{oSRNR} = 10\log_{10}\frac{\mathcal{E}\{z_{e,x}^2(n)\}}{\mathcal{E}\{z_{r,x}^2(n)\} + \mathcal{E}\{z_v^2(n)\}}, \tag{2.59}$$

$$\text{iSRNR} = 10\log_{10}\frac{\mathcal{E}\{x_{e,1}^2(n)\}}{\mathcal{E}\{x_{r,1}^2(n)\} + \mathcal{E}\{v_1^2(n)\}}, \tag{2.60}$$

where $x_{e,1}(n)$ and $x_{r,1}(n)$ denote the early and late reverberation speech components in the first microphone signal and $z_{e,x}(n)$ and $z_{r,x}(n)$ denote the early and late reverberation output speech components, defined in (2.8) and (2.9).

In addition, in order to evaluate the overall perceptual quality of the dereverberated and denoised signal, the frequency-weighted segmental signal-to-noise-ratio (fwSSNR) measure [155] is used. Similarly as the PESQ and CD measures, the fwSSNR measure is also an intrusive measure comparing the output speech signal with a reference desired signal (usually the original clean speech signal $s(n)$ or the early reverberation speech component in the first microphone $x_{e,1}(n)$). The fwSSNR improvement, i.e., ΔfwSSNR, is computed as the difference between the fwSSNR of the output speech signal $z(n)$ and the fwSSNR of the first microphone signal $y_1(n)$.

2.4 Summary

In this chapter, the signal model for speech signals recorded in a reverberant and noisy environment has been presented. The typical configuration for multi-channel speech enhancement has been described in the time and frequency domain. In addition, the objective of dereverberation, noise reduction, as well as of joint dereverberation and noise reduction has been mathematically formulated. Furthermore, we discussed some fundamental assumptions about the RIRs and the enhancement filters, i.e., the RIRs and the enhancement filters are modeled as time-invariant FIR filters and the RIRs do not share any common zeroes.

Since the remainder of this thesis primarily deals with increasing the robustness of acoustic multi-channel equalization techniques against RIR perturbations, the reasons why these perturbations arise when estimating or measuring RIRs have been briefly discussed. We have distinguished between perturbations arising due to the sensitivity of SSI or BSI methods to interfering noise, as well as perturbations arising due to spatial mismatch in the source-microphone geometry. Furthermore, we have presented the model used in this thesis to simulate RIR perturbations, i.e., the RIR perturbations are modeled as spatially white Gaussian noise sequence proportional to the true RIR coefficients.

Finally, we have presented the instrumental performance measures used to assess the dereverberation, noise reduction, as well as the joint dereverberation and noise reduction performance of speech enhancement techniques. The dereverberation performance is evaluated using the DRR and EDC channel-based measures as well as the PESQ and CD signal-based measures. The noise reduction performance is evaluated using the noise reduction factor. Finally, the joint dereverberation and noise reduction performance is evaluated using the SRNR and fwSSNR measures.

3

PARTIAL MULTI-CHANNEL EQUALIZATION BASED ON THE MULTIPLE-INPUT/OUTPUT INVERSE THEOREM

In this chapter acoustic multi-channel equalization techniques for speech dereverberation are discussed. Assuming that estimates of the room impulse responses (RIRs) between the speech source and the microphones are available, these techniques design equalization filters aiming to reshape the available RIRs such that complete or partial dereverberation is achieved.

A widely known acoustic multi-channel equalization technique that aims at complete dereverberation is the multiple-input/output inverse theorem (MINT) technique. However, the MINT technique is known to be highly sensitive to RIR perturbations. In order to increase the robustness against RIR perturbations, partial multi-channel equalization techniques such as channel shortening (CS) and relaxed multi-channel least-squares (RMCLS) have been proposed, which aim to partially equalize the RIRs by suppressing only the late reflections. It has been experimentally validated that partial equalization techniques can yield a significant increase in robustness compared to complete equalization using the MINT technique. However, by not imposing any constraints on the early part of the equalized impulse

This chapter is partly based on:

[126] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1879–1890, Sep. 2013.

[127] I. Kodrasi and S. Doclo, "Robust partial multichannel equalization techniques for speech dereverberation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Kyoto, Japan, Mar. 2012, pp. 537–540.

[128] I. Kodrasi, S. Goetze, and S. Doclo, "A perceptually constrained channel shortening technique for speech dereverberation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 2013, pp. 151–155.

[129] I. Kodrasi, S. Goetze, and S. Doclo, "Regularized subspace-based acoustic multichannel equalization for speech dereverberation," in *Proc. European Signal Processing Conference*, Marrakech, Morocco, Sep. 2013.

response (EIR), the CS and RMCLS techniques may lead to undesired perceptual effects. In this chapter, we hence propose a perceptually advantageous partial acoustic multi-channel equalization technique. Furthermore, a generalized framework for least-squares equalization techniques is established.

In Section 3.1, the optimization criteria of state-of-the-art acoustic multi-channel equalization techniques (MINT, CS, and RMCLS) are mathematically formulated and the resulting reshaping filters are derived. In Section 3.2, we propose a perceptually advantageous partial multi-channel equalization technique based on MINT (PMINT), which aims to both suppress the late reflections in the equalized impulse response as well as directly control the early reflections. In Section 3.3, a generalized framework for all considered least-squares acoustic multi-channel equalization techniques, i.e., MINT, RMCLS, and PMINT, is established, which enables to analyze the properties of the resulting reshaping filters. Based on this analysis it is shown that all considered least-squares equalization techniques yield reshaping filters which lie in the subspace spanned by the solutions maximizing the channel shortening optimization criterion. Using instrumental performance measures, simulation results in Section 3.4 illustrate the importance of controlling the early reflections in the equalized impulse response in order to preserve the perceptual speech quality. Furthermore, it is shown that all considered equalization techniques, i.e., MINT, CS, RMCLS, and PMINT, are sensitive to RIR perturbations, either yielding a low dereverberation performance or entirely failing to achieve dereverberation. Several methods to increase the robustness of the considered acoustic multi-channel equalization techniques against RIR perturbations will be proposed in Chapters 4, 5, and 6.

3.1 Acoustic multi-channel equalization techniques

Acoustic multi-channel equalization techniques typically disregard the presence of background noise and design reshaping filters aiming only at speech dereverberation. Assuming that $\mathbf{v}(n) = \mathbf{0}$ in (2.26), the output speech signal is given by

$$z(n) = \underbrace{\mathbf{w}^T\mathbf{H}^T}_{\mathbf{c}^T}\mathbf{s}(n), \tag{3.1}$$

with $\mathbf{w}$ the ML_w–dimensional reshaping filter vector, cf. (2.14), $\mathbf{H}$ the $L_c \times ML_w$–dimensional multi-channel convolution matrix of the true RIRs, cf. (2.24), $\mathbf{s}(n)$ the L_c–dimensional clean speech vector, cf. (2.23), and $\mathbf{c}$ the L_c–dimensional equalized impulse response between the clean speech signal and the output speech signal which is equal to

$$\mathbf{c} = \mathbf{H}\mathbf{w}. \tag{3.2}$$

Acoustic multi-channel equalization techniques aim at speech dereverberation by designing reshaping filters based on different design objectives for the equalized impulse response. However, as discussed in Section 2.2, the available (measured or estimated) RIRs $\hat{h}_m$ typically differ from the true RIRs $h_m(n)$, i.e.,

$$\hat{h}_m(n) = h_m(n) + e_m(n), \tag{3.3}$$

with $e_m(n)$ the perturbations arising due to the sensitivity of supervised or blind system identification methods to interfering noise [101, 110], or due to spatial mismatch [109]. Hence, instead of using the true convolution matrix $\mathbf{H}$, acoustic multi-channel equalization techniques typically design reshaping filters using the perturbed convolution matrix $\hat{\mathbf{H}}$, constructed from the perturbed RIRs $\hat{h}_m(n)$, i.e.,

$$\hat{\mathbf{H}} = \mathbf{H} + \mathbf{E}, \tag{3.4}$$

with $\mathbf{E}$ the multi-channel convolution matrix of the RIR perturbations. Designing equalization filters to reshape the *estimated* equalized impulse response $\hat{\mathbf{c}}$, with

$$\hat{\mathbf{c}} = \hat{\mathbf{H}}\mathbf{w}, \tag{3.5}$$

does not necessarily result in a correctly reshaped *true* equalized impulse response $\mathbf{c}$. Hence, acoustic multi-channel equalization techniques that are robust against RIR perturbations are required in practice. In the following, the design objectives of existing equalization techniques, namely MINT [36], CS [124,125], and RMCLS [111, 125], are reviewed.

Multiple-input/output inverse theorem (MINT)

The objective of the MINT technique is to invert the acoustic system up to a delay τ, such that the output speech signal is a delayed version of the clean speech signal. To this purpose, an inverse filter $\mathbf{w}$ is designed such that

$$\hat{\mathbf{H}}\mathbf{w} = \mathbf{d}, \tag{3.6}$$

with $\mathbf{d}$ the target equalized impulse response defined as a delayed impulse, i.e.,

$$\mathbf{d} = [\underbrace{0 \ \dots \ 0}_{\tau} \ 1 \ 0 \ \dots \ 0]^T. \tag{3.7}$$

The delay τ is typically incorporated to relax the causality constraints on the inverse filter design [156]. The MINT filter is then computed by minimizing the least-squares cost function

$$J_{\text{M}} = \|\hat{\mathbf{H}}\mathbf{w} - \mathbf{d}\|_2^2. \tag{3.8}$$

As shown in [36], assuming that

- the perturbed RIRs do not share any common zeros in the z-plane, and
- $L_w \geq \left\lceil \frac{L_h - 1}{M - 1} \right\rceil$, where $\lceil \cdot \rceil$ denotes the ceiling operation,

the MINT filter that minimizes the least-squares cost function in (3.8) is equal to

$$\mathbf{w}_{\text{M}} = \hat{\mathbf{H}}^{+}\mathbf{d}, \tag{3.9}$$

with $\{\cdot\}^+$ denoting the Moore-Penrose pseudo-inverse. Since the perturbed multi-channel convolution matrix is assumed to be a full row-rank matrix [157], its pseudo-inverse can be computed as

$$\hat{\mathbf{H}}^{+} = \hat{\mathbf{H}}^T(\hat{\mathbf{H}}\hat{\mathbf{H}}^T)^{-1}. \tag{3.10}$$

When the true RIRs are available, i.e., $\hat{\mathbf{H}} = \mathbf{H}$, the equalized impulse response $\mathbf{c}$ is equal to the target equalized impulse response $\mathbf{d}$ and the MINT technique achieves perfect acoustic system inversion, i.e.,

$$\mathbf{c} = \mathbf{H}\mathbf{w}_{\text{M}} = \mathbf{d}. \tag{3.11}$$

However, in the presence of RIR perturbations, i.e., $\hat{\mathbf{H}} \neq \mathbf{H}$, the equalized impulse response

$$\mathbf{c} = \mathbf{H}\mathbf{w}_{\text{M}} = \mathbf{H}\hat{\mathbf{H}}^{+}\mathbf{d}, \tag{3.12}$$

not only differs from the target response $\mathbf{d}$, but usually yields large distortions in the output speech signal [111, 125, 126].

Partial multi-channel equalization techniques

Whereas the MINT technique is very sensitive to RIR perturbations, partial multi-channel equalization techniques such as CS and RMCLS have been shown to be more robust against RIR perturbations [111, 124, 125]. These techniques aim at suppressing only the late reflections, while imposing no constraints on the early reflections, which however may lead to undesired perceptual effects as shown in Section 3.4.

Channel shortening (CS): The CS technique has been extensively investigated in the context of digital communication applications [158] and in the past decade it has been applied to acoustic channel equalization in [124, 125]. The shortening of the acoustic channel is achieved by designing a reshaping filter which maximizes the energy in the first samples of the equalized impulse response, corresponding to the direct path and early reflections, while minimizing the energy in the remaining samples, corresponding to the late reflections. This optimization problem can be expressed as the maximization of the generalized Rayleigh quotient

$$J_{\text{CS}} = \frac{\|\mathbf{W}_{\text{d}}\hat{\mathbf{c}}\|_2^2}{\|\mathbf{W}_{\text{u}}\hat{\mathbf{c}}\|_2^2} = \frac{\|\mathbf{W}_{\text{d}}\hat{\mathbf{H}}\mathbf{w}\|_2^2}{\|\mathbf{W}_{\text{u}}\hat{\mathbf{H}}\mathbf{w}\|_2^2} = \frac{\mathbf{w}^T\hat{\mathbf{H}}^T\mathbf{W}_{\text{d}}^T\mathbf{W}_{\text{d}}\hat{\mathbf{H}}\mathbf{w}}{\mathbf{w}^T\hat{\mathbf{H}}^T\mathbf{W}_{\text{u}}^T\mathbf{W}_{\text{u}}\hat{\mathbf{H}}\mathbf{w}}, \tag{3.13}$$

with $\mathbf{W}_{\text{d}}$ and $\mathbf{W}_{\text{u}}$ being $L_c \times L_c$–dimensional diagonal weighting matrices of the desired and undesired part of the equalized impulse response, i.e.,

$$\mathbf{W}_{\text{d}} = \text{diag}\{[\underbrace{0 \ \ldots \ 0}_{\tau}\underbrace{1 \ \ldots \ 1}_{L_e}0 \ \ldots 0]\}, \tag{3.14}$$

$$\mathbf{W}_{\text{u}} = \text{diag}\{[\underbrace{1 \ \ldots \ 1}_{\tau}\underbrace{0 \ \ldots \ 0}_{L_e}1 \ \ldots 1]\}, \tag{3.15}$$

where L_e denotes the length of the desired direct path and early reflections in number of samples (typically set to correspond to at most 50 ms). Defining the $ML_w \times ML_w$–dimensional positive semi-definite matrices $\hat{\mathbf{D}}$ and $\hat{\mathbf{U}}$ as

$$\hat{\mathbf{D}} = \hat{\mathbf{H}}^T\mathbf{W}_{\text{d}}^T\mathbf{W}_{\text{d}}\hat{\mathbf{H}}, \tag{3.16}$$

$$\hat{\mathbf{U}} = \hat{\mathbf{H}}^T\mathbf{W}_{\text{u}}^T\mathbf{W}_{\text{u}}\hat{\mathbf{H}}, \tag{3.17}$$

the CS optimization problem in (3.13) can be expressed as the maximization of the generalized Rayleigh quotient

$$J_{\text{CS}} = \frac{\mathbf{w}^T \hat{\mathbf{D}} \mathbf{w}}{\mathbf{w}^T \hat{\mathbf{U}} \mathbf{w}}. \tag{3.18}$$

Maximizing (3.18) is equivalent to solving the generalized eigenvalue problem

$$\hat{\mathbf{D}} \mathbf{w} = \lambda \hat{\mathbf{U}} \mathbf{w}, \tag{3.19}$$

where the optimal reshaping filter $\mathbf{w}_{\text{CS}}$ is the generalized eigenvector corresponding to the largest generalized eigenvalue $\lambda_{\max}$, i.e.,

$$\hat{\mathbf{D}} \mathbf{w}_{\text{CS}} = \lambda_{\max} \hat{\mathbf{U}} \mathbf{w}_{\text{CS}}. \tag{3.20}$$

It should be noted that designing the reshaping filter using such an energy-based optimization criterion imposes no other, e.g., perceptually relevant, constraints on the early reflections of the equalized impulse response, which may lead to undesired perceptual effects (cf. Section 3.4.2). Furthermore, when the used reshaping filter length is $L_w \geq \left\lceil \frac{L_h - 1}{M - 1} \right\rceil$, multiple solutions to (3.20) exist (cf. Section 3.3), where each of these solutions will lead to a perceptually different output speech signal [125]. Out of these multiple solutions, in [125] it has been proposed to use the generalized eigenvector yielding the minimum l_2-norm equalized impulse response, since it has been observed that this eigenvector yields the best perceptual speech quality. However, it should be noted that this selection criterion was based on informal listening tests for perfectly known RIRs, which is generally not the case in practice.

Relaxed multi-channel least-squares (RMCLS): The RMCLS technique achieves partial channel equalization by introducing a diagonal weighting matrix $\mathbf{W}_{\text{R}}$ in the least-squares cost function in (3.8), i.e., the RMCLS cost function is defined as

$$J_{\text{R}} = \|\mathbf{W}_{\text{R}} (\hat{\mathbf{H}} \mathbf{w} - \mathbf{d})\|_2^2, \tag{3.21}$$

with $\mathbf{W}_{\text{R}}$ equal to

$$\mathbf{W}_{\text{R}} = \text{diag}\{[\underbrace{1 \ \ldots \ 1}_{\tau} \ \underbrace{1 \, 0 \ \ldots \ 0}_{L_e} 1 \ \ldots \ 1]\}. \tag{3.22}$$

By using the weighting matrix $\mathbf{W}_{\text{R}}$, the RMCLS cost function in (3.21) aims at setting the samples of the EIR corresponding to the delay τ and to the late reflections equal to $\mathbf{0}$, while the early reflections are not in any way constrained. Similarly to the MINT filter in (3.9), the RMCLS filter minimizing (3.21) can be computed as

$$\mathbf{w}_{\text{R}} = (\mathbf{W}_{\text{R}} \hat{\mathbf{H}})^{+} (\mathbf{W}_{\text{R}} \mathbf{d}). \tag{3.23}$$

It has been shown in [111,125,126] that relaxing the constraints on the reshaping filter design by using the weighting matrix $\mathbf{W}_{\text{R}}$ yields an increase in robustness against RIR perturbations in terms of suppression of the late reflections. However, similarly as the CS technique, by not imposing any constraints on the early reflections in the equalized impulse response, the RMCLS technique may lead to undesired perceptual effects (cf. Section 3.4.2).

3.2 Partial multi-channel equalization based on the multiple-input/output inverse theorem (PMINT)

In order to directly control the perceptual quality of the output speech signal by controlling both the early and late reflections of the equalized impulse response, we propose the PMINT technique, where the first part (i.e., the direct path and early reflections) of one of the available RIRs is used as the target equalized impulse response in (3.6), i.e.,

$$\hat{\mathbf{H}}\mathbf{w} = \hat{\mathbf{h}}_{e,p}, \tag{3.24}$$

with

$$\hat{\mathbf{h}}_{e,p} = [\underbrace{0 \ \dots \ 0}_{\tau} \ \underbrace{\hat{h}_p(0) \ \dots \ \hat{h}_p(L_e-1)}_{L_e} \ 0 \dots 0]^T, \tag{3.25}$$

and $p \in \{1, \ \dots, \ M\}$. Without loss of generality, also other target equalized impulse responses could be used instead of (3.25), as long as they are perceptually close to the true RIRs. Similarly to (3.8), the least-squares cost function to be minimized in the PMINT technique is defined as

$$J_{\text{P}} = \|\hat{\mathbf{H}}\mathbf{w} - \hat{\mathbf{h}}_{e,p}\|_2^2. \tag{3.26}$$

Assuming that the same conditions as for the MINT technique are satisfied (cf. Section 3.1), the reshaping filter minimizing the PMINT cost function in (3.26) can be computed as

$$\mathbf{w}_{\text{P}} = \hat{\mathbf{H}}^{+}\hat{\mathbf{h}}_{e,p}. \tag{3.27}$$

As an illustrative example, Fig. 3.1 depicts the equalized impulse responses obtained using the MINT, CS, RMCLS, and PMINT techniques when the true RIRs are known, i.e., $\hat{\mathbf{H}} = \mathbf{H}$. The delay τ is set to correspond to 11.25 ms and the length of the direct path and early reflections L_e is set to correspond to 50 ms. When the true RIRs are known, all acoustic multi-channel equalization techniques are able to achieve perfect suppression of the late reflections, hence, the depicted equalized impulse responses have been cut after 80 ms and the late reflections (which are zero for all considered techniques) have not been shown. As expected, the MINT technique (Fig. 3.1a) yields a delayed impulse. Furthermore, as can be observed in Figs. 3.1b and 3.1c, the CS and RMCLS techniques yield early reflections that do not exhibit the naturally decaying pattern of a room impulse response, typically resulting in an unnatural coloration and decreased perceptual quality of the output speech signal. By using the direct path and early reflections of one of the available RIRs as the target response, the equalized impulse response resulting from the PMINT technique shown in Fig. 3.1d resembles a typical RIR and preserves the perceptual speech quality of the output speech signal (cf. Section 3.4.2).

However, given the similarity of the MINT and PMINT filters in (3.9) and (3.27), the PMINT technique is expected to inherit the sensitivity of the MINT technique to RIR perturbations, as will also be experimentally validated in Section 3.4.3. Increasing the robustness of the PMINT technique against RIR perturbations will be discussed in the following chapters.

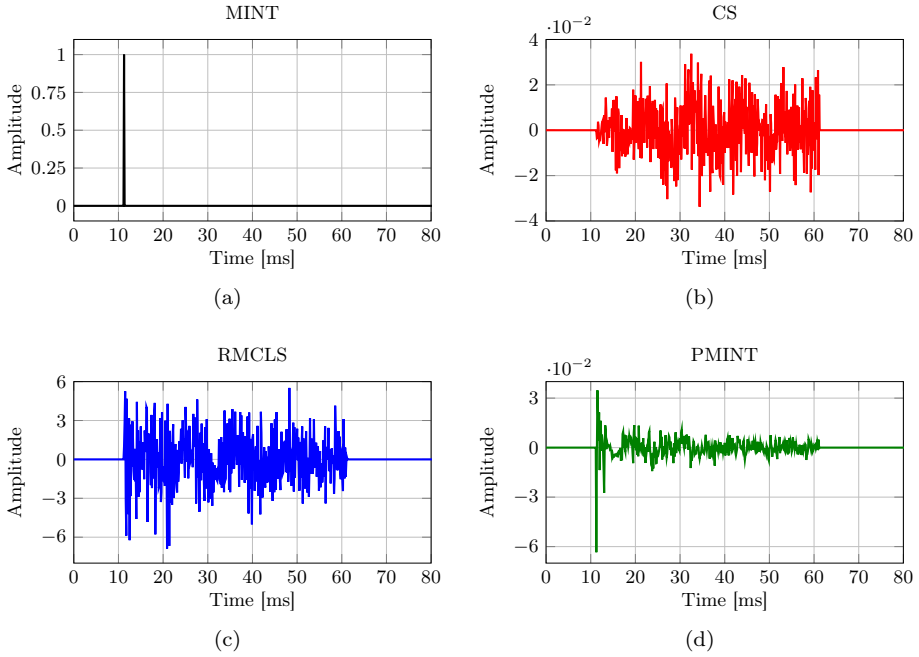


Fig. 3.1: Exemplary equalized impulse response when the true room impulse responses are known obtained using the (a) MINT technique, (b) CS technique, (c) RMCLS technique, and (d) PMINT technique. The delay is set to $\tau = 90$, corresponding to 11.25 ms, and the desired window length is set to $L_e = 400$, corresponding to 50 ms. The considered acoustic system is the same as in Section 3.4.1.

3.3 Generalized framework for least-squares acoustic multi-channel equalization techniques and their relation to channel shortening

In this section, a generalized framework for all previously presented least-squares equalization techniques, i.e., MINT, RMCLS, and PMINT, is established. A theoretical analysis based on the Rouché-Capelli theorem is provided to determine the properties, i.e., existence and uniqueness, of the solution(s) for each least-squares cost function. Using this analysis it is shown that all considered least-squares equalization techniques yield reshaping filters which lie in the subspace spanned by the multiple channel shortening solutions.

Least-squares equalization techniques

The objective of all presented least-squares equalization techniques in Sections 3.1 and 3.2, i.e., MINT, RMCLS, and PMINT, can be expressed in a unified manner as

$$\mathbf{W}\hat{\mathbf{H}}\mathbf{w} = \mathbf{W}\mathbf{c}_t, \tag{3.28}$$

with $\mathbf{W}$ a diagonal weighting matrix and $\mathbf{c}_t$ the target equalized impulse response. Depending on the definition of the weighting matrix $\mathbf{W}$ and the target equalized impulse response $\mathbf{c}_t$, the objective of all considered least-squares techniques can be derived. Using $\mathbf{W} = \mathbf{I}$, with $\mathbf{I}$ the $L_c \times L_c$–dimensional identity matrix, and $\mathbf{c}_t = \mathbf{d}$, the objective of the MINT technique is derived. Using $\mathbf{W} = \mathbf{W}_{\text{R}}$ and $\mathbf{c}_t = \mathbf{d}$, the objective of the RMCLS technique is derived. Finally, using $\mathbf{W} = \mathbf{I}$ and $\mathbf{c}_t = \hat{\mathbf{h}}_{e,p}$, the objective of the PMINT technique is derived. Hence, the cost function of all considered least-squares techniques can be represented in a unified manner as

$$J_{\text{LS}} = \|\mathbf{W}(\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t)\|_2^2, \tag{3.29}$$

and the least-squares filter minimizing (3.29) can be computed as

$$\mathbf{w}_{\text{LS}} = (\mathbf{W}\hat{\mathbf{H}})^{+}(\mathbf{W}\mathbf{c}_t). \tag{3.30}$$

In the following, the Rouché-Capelli theorem is used to establish the existence and uniqueness of solutions to (3.29) for the different definitions of the weighting matrix $\mathbf{W}$ and the target equalized impulse response $\mathbf{c}_t$.

Rouché-Capelli theorem [159]: *Consider the system of equations* $\mathbf{A}\mathbf{q} = \mathbf{b}$, *where the matrix* $\mathbf{A}$ *has dimensions* $u \times v$. *Such a system has a solution if and only if the rank of the coefficient matrix* $\mathbf{A}$ *is equal to the rank of the augmented matrix* $[\mathbf{A}|\mathbf{b}]$. *If a solution exists and* $\text{rank}(\mathbf{A}) = v$, *this solution is unique, otherwise there are an infinite number of solutions.*

Assuming that the reshaping filter length is $L_w \geq \left\lceil \frac{L_h - 1}{M - 1} \right\rceil$ and the matrix $\hat{\mathbf{H}}$ is a full row-rank matrix, Table 3.1 summarizes the rank of the coefficient and augmented matrix for each considered least-squares technique.[1] Since for all considered definitions of $\mathbf{W}$ and $\mathbf{c}_t$ the rank of the coefficient matrix is equal to the rank of the augmented matrix, the system of equations in (3.28) is always solvable.

For the MINT and PMINT techniques, we need to distinguish among the following two cases:

Table 3.1: Rank of the coefficient and augmented matrix for least-squares equalization techniques.

Technique	$\mathbf{A}\mathbf{q} = \mathbf{b}$	$\text{rank}(\mathbf{A})$	$\text{rank}([\mathbf{A}\|\mathbf{b}])$
MINT	$\hat{\mathbf{H}}\mathbf{w} = \mathbf{d}$	L_c	L_c
RMCLS	$\mathbf{W}_{\text{R}}\hat{\mathbf{H}}\mathbf{w} = \mathbf{W}_{\text{R}}\mathbf{d}$	$L_c - L_e + 1$	$L_c - L_e + 1$
PMINT	$\hat{\mathbf{H}}\mathbf{w} = \hat{\mathbf{h}}_{e,p}$	L_c	L_c

1 Note that while the reshaping filter length $L_w \geq \left\lceil \frac{L_h - 1}{M - 1} \right\rceil$ will be mostly considered in this thesis, in Chapter 4 we will also consider the reshaping filter length $L_w < \left\lceil \frac{L_h - 1}{M - 1} \right\rceil$.

i) if $L_c = ML_w$, i.e., $\hat{\mathbf{H}}$ is a square matrix, and hence $\text{rank}(\hat{\mathbf{H}}) = L_c = ML_w$, there is a unique solution to (3.28);
ii) otherwise if $L_c < ML_w$, i.e., $\hat{\mathbf{H}}$ is a fat matrix, there are an infinite number of solutions and the reshaping filter in (3.30) is the minimum-norm solution [160].

For the RMCLS technique there is always an infinite number of solutions since the number of columns in $\mathbf{W}_{\text{R}}\hat{\mathbf{H}}$ is always greater than its rank, i.e., $ML_w > L_c - L_e + 1$. The reshaping filter in (3.30) is the minimum-norm solution [160].

Since the system of equations in (3.28) is solvable for all considered least-squares equalization techniques, the reshaping filter $\mathbf{w}_{\text{LS}}$ results in an estimated equalized impulse response $\hat{\mathbf{c}}_{\text{LS}} = \hat{\mathbf{H}}\mathbf{w}_{\text{LS}}$ with non-zero samples in the direct path and early reflections and zero samples in the late reflections, i.e.,

$$\hat{\mathbf{c}}_{\text{LS}} = [\underbrace{0 \ \dots \ 0}_{\tau} \ \underbrace{\hat{c}_{\text{LS}}(0) \ \hat{c}_{\text{LS}}(1) \ \dots \ \hat{c}_{\text{LS}}(L_e-1)}_{L_e} \ 0 \ \dots \ 0]^T. \tag{3.31}$$

Note that for the MINT technique $\hat{c}_{\text{LS}}(0) \neq 0$, whereas all remaining samples in the estimated equalized impulse response are equal to 0.

Relation to the channel shortening technique

As shown in [125], when the used reshaping filter length is $L_w \geq \left\lceil \frac{L_h - 1}{M-1} \right\rceil$, the CS maximization problem in (3.18) can be reformulated as finding a reshaping filter $\mathbf{w}$ which belongs to the null space of $\hat{\mathbf{U}}$ but does not belong to the null space of $\hat{\mathbf{D}}$, i.e.,

$$\begin{cases} \mathbf{w}^T\hat{\mathbf{D}}\mathbf{w} \neq 0 \\ \mathbf{w}^T\hat{\mathbf{U}}\mathbf{w} = 0 \end{cases}, \tag{3.32}$$

such that the generalized Rayleigh quotient in (3.18) is maximized to ∞. The system of equations in (3.32) can be rewritten as

$$\begin{cases} \mathbf{w}^T(\hat{\mathbf{D}} + \hat{\mathbf{U}})\mathbf{w} \neq 0 \\ \mathbf{w}^T\hat{\mathbf{U}}\mathbf{w} = 0 \end{cases}. \tag{3.33}$$

The matrix $\hat{\mathbf{D}} + \hat{\mathbf{U}}$ is equal to

$$\hat{\mathbf{D}} + \hat{\mathbf{U}} = \hat{\mathbf{H}}^T\mathbf{W}_{\text{d}}^T\mathbf{W}_{\text{d}}\hat{\mathbf{H}} + \hat{\mathbf{H}}^T\mathbf{W}_{\text{u}}^T\mathbf{W}_{\text{u}}\hat{\mathbf{H}} \tag{3.34}$$

$$= \hat{\mathbf{H}}^T(\mathbf{W}_{\text{d}}^T\mathbf{W}_{\text{d}} + \mathbf{W}_{\text{u}}^T\mathbf{W}_{\text{u}})\hat{\mathbf{H}} \tag{3.35}$$

$$= \hat{\mathbf{H}}^T\hat{\mathbf{H}}, \tag{3.36}$$

which is a positive semi-definite matrix. Since the convolution matrix $\hat{\mathbf{H}}$ is assumed to be a full row-rank matrix with $\text{rank}(\hat{\mathbf{H}}) = L_c$, also $\text{rank}(\hat{\mathbf{D}} + \hat{\mathbf{U}}) = \text{rank}(\hat{\mathbf{H}}^T\hat{\mathbf{H}}) = L_c$. Exploiting the relationship between the rank and the dimension of the null space of a matrix [160], the dimension of the null space of $\hat{\mathbf{D}} + \hat{\mathbf{U}}$ is equal to

$$\dim[\text{null space}(\hat{\mathbf{D}} + \hat{\mathbf{U}})] = ML_w - L_c. \tag{3.37}$$

In addition, since $\text{rank}(\hat{\mathbf{U}}) = \text{rank}(\hat{\mathbf{H}}) - L_e = L_c - L_e$, the dimension of the null space of $\hat{\mathbf{U}}$ is equal to

$$\dim[\text{null space}(\hat{\mathbf{U}})] = ML_w - (L_c - L_e). \tag{3.38}$$

Since the matrices $\hat{\mathbf{D}}+\hat{\mathbf{U}}$, $\hat{\mathbf{D}}$, and $\hat{\mathbf{U}}$ are positive semi-definite matrices, the vectors belonging to the null space of $\hat{\mathbf{D}}+\hat{\mathbf{U}}$ also belong to the null space of $\hat{\mathbf{D}}$ and to the null space of $\hat{\mathbf{U}}$, i.e.,

$$\mathbf{w}^T(\hat{\mathbf{D}}+\hat{\mathbf{U}})\mathbf{w} = 0 \Rightarrow \mathbf{w}^T\hat{\mathbf{D}}\mathbf{w} = -\mathbf{w}^T\hat{\mathbf{U}}\mathbf{w} \Rightarrow \mathbf{w}^T\hat{\mathbf{D}}\mathbf{w} = 0 \text{ and } \mathbf{w}^T\hat{\mathbf{U}}\mathbf{w} = 0. \tag{3.39}$$

Since every vector in the null space of $\hat{\mathbf{D}}+\hat{\mathbf{U}}$ also belongs to the null space of $\hat{\mathbf{U}}$, there must be $[ML_w - (L_c - L_e)] - [ML_w - L_c] = L_e$ linearly independent vectors that belong to the null space of $\hat{\mathbf{U}}$ but do not belong to the null space of $\hat{\mathbf{D}}+\hat{\mathbf{U}}$. Hence, the number of linearly independent vectors satisfying (3.33) and therefore maximizing the generalized Rayleigh quotient in (3.18) to ∞ is equal to L_e.

Applying the desired and undesired weighting matrices of the CS technique (cf. (3.14) and (3.15)) to the least-squares estimated equalized impulse response $\hat{\mathbf{c}}_{\text{LS}}$ in (3.31) yields

$$\mathbf{W}_{\text{d}}\hat{\mathbf{c}}_{\text{LS}} \neq \mathbf{0} \quad \text{and} \quad \mathbf{W}_{\text{u}}\hat{\mathbf{c}}_{\text{LS}} = \mathbf{0}. \tag{3.40}$$

Based on (3.40), it can be said that the least-squares reshaping filter $\mathbf{w}_{\text{LS}}$ satisfies the system of equations in (3.33), i.e.,

$$\begin{cases} \mathbf{w}_{\text{LS}}^T(\hat{\mathbf{D}}+\hat{\mathbf{U}})\mathbf{w}_{\text{LS}} \neq 0 \\ \mathbf{w}_{\text{LS}}^T\hat{\mathbf{U}}\mathbf{w}_{\text{LS}} = 0 \end{cases}. \tag{3.41}$$

Therefore, all solutions $\mathbf{w}_{\text{LS}}$ of the least-squares equalization techniques lie in the subspace spanned by the channel shortening solutions, i.e.,

$$\mathbf{w}_{\text{LS}} = \mathbf{S}_{\text{CS}}\boldsymbol{\alpha}, \tag{3.42}$$

with $\mathbf{S}_{\text{CS}} = [\mathbf{w}_{\text{CS}}^1\ \mathbf{w}_{\text{CS}}^2\ \ldots\ \mathbf{w}_{\text{CS}}^{L_e}]$ the $ML_w \times L_e$–dimensional matrix whose columns are the L_e channel shortening solutions and $\boldsymbol{\alpha}$ an L_e–dimensional linear combination vector. Since all least-squares reshaping filters lie in the subspace spanned by the channel shortening solutions, it can be said that this subspace offers the potential to achieve a high dereverberation performance, both in terms of reverberant energy suppression and perceptual speech quality preservation. We have investigated the use of this subspace to derive more robust and perceptually advantageous reshaping filters in [128, 129], where the multiple CS solutions are combined to achieve the least-squares optimization criteria. While all methods proposed in this thesis to increase the robustness of equalization techniques can be directly incorporated into the techniques proposed in [128, 129], in this thesis we have limited the discussion to the basic channel shortening and least-squares techniques.

3.4 Simulations

In this section, the dereverberation performance of the acoustic multi-channel equalization techniques presented in Sections 3.1 and 3.2 is investigated. In Section 3.4.1

the considered acoustic system and the used algorithmic settings are introduced. Section 3.4.2 investigates the performance of the partial acoustic multi-channel equalization techniques, i.e., CS, RMCLS, and PMINT, when the true RIRs are known. Section 3.4.3 investigates the performance of all considered acoustic multi-channel equalization techniques, i.e., MINT, CS, RMCLS, and PMINT, in the presence of RIR perturbations.

3.4.1 *Acoustic system and algorithmic settings*

We have considered an acoustic scenario with a single speech source and $M = 4$ omni-directional microphones. The source-microphone distance is 3 m and the distance between the microphones is 5 cm. Room impulse responses from the MARDY database [161] have been used, where the room reverberation time is $T_{60} \approx 450$ ms and the direct-to-reverberant ratio is DRR $= 0$ dB. The RIRs have been measured using the swept-sine technique [162] and the length of the RIRs has been set to $L_h = 3600$ at a sampling frequency $f_s = 8$ kHz. For illustration, Fig. 3.2 depicts the RIR between the source and the first microphone. To generate the reverberant signals, 10 sentences (approximately 17 s long) from the HINT database [163] have been convolved with the measured RIRs.

In order to simulate RIR perturbations, the measured RIRs are perturbed by adding scaled white noise as described in Section 2.2. The considered normalized projection misalignment (NPM) values between the true and the perturbed RIRs are (cf. (2.52))

$$\text{NPM} \in \{-33 \text{ dB}, \ -27 \text{ dB}, \ -21 \text{ dB}, \ -15 \text{ dB}\}. \tag{3.43}$$

For all considered acoustic multi-channel equalization techniques, the used reshaping filter length is $L_w = \left\lceil \frac{L_h - 1}{M-1} \right\rceil = 1200$, the delay is arbitrarily set to $\tau = 90$, and the performance for several desired window lengths L_d ranging from 10 ms to 50 ms is investigated, i.e.,

$$L_d = \frac{L_e \times 1000}{f_s} \in \{10 \text{ ms}, \ 20 \text{ ms}, \ 30 \text{ ms}, \ 40 \text{ ms}, \ 50 \text{ ms}\}. \tag{3.44}$$

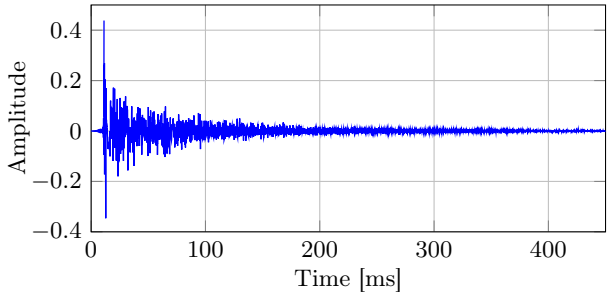


Fig. 3.2: The true room impulse response between the speech source and the first microphone.

The target equalized impulse response for the PMINT technique in (3.25) is set to the direct path and the early reflections of the perturbed RIR of the first microphone, i.e., $\hat{\mathbf{h}}_{e,1}$. Furthermore, the CS reshaping filter is selected as the generalized eigenvector yielding the smallest l_2-norm estimated equalized impulse response as proposed in [125].

Using the instrumental performance measures described in Section 2.3, the dereverberation performance is evaluated in terms of the reverberant energy suppression and the perceptual speech quality improvement. The reverberant energy suppression is evaluated using the direct-to-reverberant ratio improvement (ΔDRR) between the equalized impulse response $\mathbf{c}$ and the true RIR $\mathbf{h}_1$ (cf. (2.53)), as well as the energy decay curve (EDC) of the equalized impulse response $\mathbf{c}$ (cf. (2.55)). The improvement in perceptual speech quality is evaluated using the improvement in PESQ [153] (ΔPESQ) and in cepstral distance [154] (ΔCD) between the output speech signal $z(n)$ and the reverberant microphone signal $x_1(n)$. The reference signal employed for the PESQ and cepstral distance measures is $x_{e,1}(n) = s(n) * h_{e,1}(n)$, i.e., the clean speech signal convolved with the direct path and early reflections of the first RIR (which changes as the desired window length L_d changes).

3.4.2 *Performance of partial acoustic multi-channel equalization techniques when the true RIRs are known*

In this section, the performance of the considered partial equalization techniques, i.e., CS, RMCLS, and PMINT, is investigated when the true RIRs are known.[2] Based on the theoretical discussion in Section 3.3, all considered techniques perfectly achieve their design objective when the true RIRs are known, i.e., perfect suppression of the late reflections. However, the aim of this section is to provide more insights on the importance of preserving the direct path and early reflections for perceptual speech quality preservation as well as to provide a baseline of what perfect dereverberation performance means in terms of measures such as ΔDRR, EDC, ΔPESQ, and ΔCD.

To evaluate the reverberant energy suppression, Fig. 3.3a depicts the direct-to-reverberant ratio improvement for the CS, RMCLS, and PMINT techniques. It can be observed that the PMINT technique achieves the highest ΔDRR for all considered desired window lengths, outperforming the CS and RMCLS techniques. By not controlling the early reflections in the equalized impulse response, the CS and RMCLS techniques seem to introduce additional energy in the first L_e samples (cf. Figs. 3.1b and 3.1c), which is accounted for as reverberant energy in the DRR computation, hence, decreasing the resulting ΔDRR. Furthermore, as the desired window length increases, the CS and RMCLS techniques worsen the DRR in comparison to the reverberant microphone signal, whereas the PMINT technique achieves an improvement for all considered desired window lengths.

2 Note that the MINT technique is not considered in this simulation, since it perfectly recovers the (delayed) clean speech signal when the true RIRs are known.

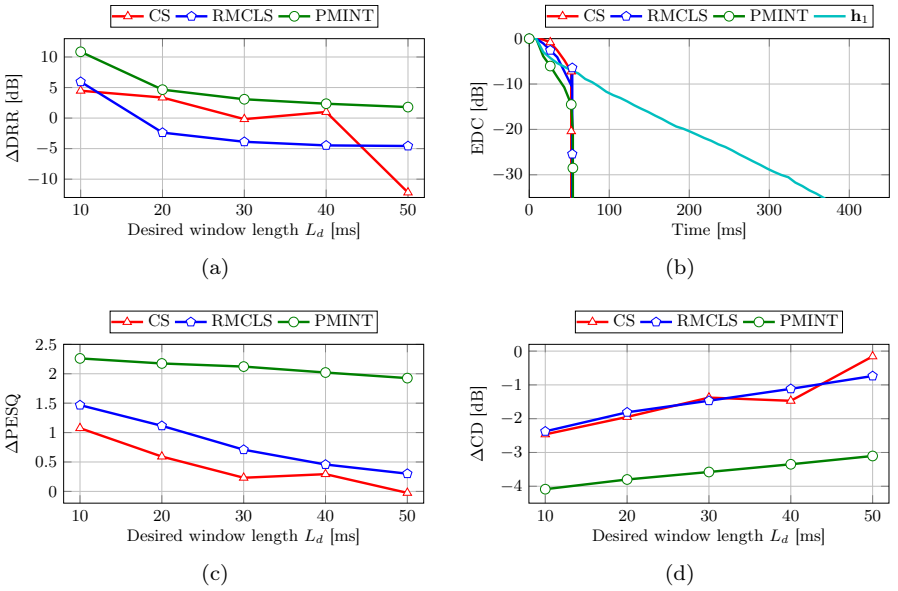


Fig. 3.3: Performance of the CS, RMCLS, and PMINT techniques for known true RIRs in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD.

To evaluate the decay rate of the reverberant energy, Fig. 3.3b depicts the energy decay curve of the true RIR $\mathbf{h}_1$ and the energy decay curve of the equalized impulse response $\mathbf{c}$ obtained using the CS, RMCLS, and PMINT techniques for the desired window length $L_d = 50$ ms. Since all acoustic multi-channel equalization techniques perfectly suppress the late reflections when the true RIRs are known, they all result in the reverberant energy decaying to $-\infty$ after 50 ms.

To evaluate the perceptual speech quality, Fig. 3.3c illustrates the PESQ score improvement for all considered partial multi-channel equalization techniques. Since the PMINT technique perfectly recovers the reference signal when the true RIRs are known, the performance it achieves represents the upper boundary of the achievable performance in terms of ΔPESQ. As expected, the ΔPESQ values for the PMINT technique decrease as the desired window length increases, since the microphone signal becomes more similar to the reference signal. Furthermore, the CS and RMCLS techniques achieve a worse PESQ score improvement compared to the PMINT technique, particularly for increasing desired window lengths. As the desired window length increases, more early reflections are left uncontrolled in the CS and RMCLS techniques, causing as a result the drop in perceptual speech quality.

Similar conclusions can be derived based on the cepstral distance improvement depicted in Fig. 3.3d, where again the performance of the PMINT technique represents the upper boundary of the achievable performance in terms of ΔCD, whereas the CS and RMCLS techniques result in a significantly worse performance.

Summarizing, preserving the early reflections as in the proposed PMINT technique is advantageous in order to achieve a high dereverberation performance, both in terms of reverberant energy suppression and perceptual speech quality improvement. However, one should realize that the results presented above have been obtained for perfectly known RIRs, which is typically not the case in practice. In the presence of RIR perturbations, the sensitivity of equalization techniques to RIR perturbations needs to be taken into account.

3.4.3 *Performance of acoustic multi-channel equalization techniques in the presence of RIR perturbations*

In this section, the performance of all considered acoustic multi-channel equalization techniques, i.e., MINT, CS, RMCLS, and PMINT, is investigated in the presence of RIR perturbations for the NPM values given in (3.43). The presented performance measures are averaged over all considered NPM values.

To evaluate the reverberant energy suppression, Fig. 3.4a depicts the direct-to-reverberant ratio improvement for the MINT, CS, RMCLS, and PMINT techniques. It should be noted that the MINT technique is independent of the desired window length L_d, however, in order to be able to compare the performance of the MINT technique to partial equalization techniques which depend on the desired window length L_d, the performance of the MINT technique is presented on the same plot. It can be observed that by relaxing the constraints on the reshaping filter design, the RMCLS technique is the most robust technique, yielding the highest ΔDRR for all considered desired window lengths. However, the RMCLS technique only slightly improves the DRR in comparison to the true RIR $\mathbf{h}_1$ for the desired window length $L_d = 10$ ms. For increasing desired window lengths, negative ΔDRR values are obtained. Furthermore, as expected, the MINT technique fails to achieve dereverberation and decreases the DRR in comparison to the true RIR $\mathbf{h}_1$ by approximately 17 dB. Moreover, the PMINT technique seems to inherit the sensitivity of the MINT technique to RIR perturbations, also failing to achieve dereverberation for all considered desired window lengths. Although the CS technique relaxes the constraints on the reshaping filter design by using an energy-based optimization criterion, it also fails to achieve dereverberation in this scenario. It should however be noted that out of the multiple CS solutions, reshaping filters yielding a significantly better ΔDRR than the one resulting in the minimum l_2-norm equalized impulse response (depicted here) could be found. As mentioned in Section 3.1, selecting the reshaping filter as the one yielding the minimum l_2-norm equalized impulse response was proposed in [125] based on observations for perfectly known RIRs. Clearly, different selection criteria should be investigated in the presence of RIR perturbations.

To evaluate the decay rate of the reverberant energy, Fig. 3.4b depicts the energy decay curve of the true RIR $\mathbf{h}_1$ and the energy decay curve of the equalized impulse response $\mathbf{c}$ obtained using the MINT, CS, RMCLS, and PMINT techniques for the desired window length $L_d = 50$ ms. Similar conclusions as for the ΔDRR analysis can be derived, i.e., the RMCLS technique is the most robust technique yielding

a faster decay rate of the reverberant energy in comparison to the true RIR $\mathbf{h}_1$, whereas the MINT, CS, and PMINT techniques introduce additional distortions in the output speech signal.

To evaluate the perceptual speech quality improvement, Figs. 3.4c and 3.4d illustrate the PESQ score and cepstral distance improvement achieved by the MINT, CS, RMCLS, and PMINT techniques. Similarly as for the reverberant energy suppression analysis, it can be observed that while the MINT, CS, and PMINT techniques significantly decrease the perceptual speech quality in comparison to the reverberant microphone signal $x_1(n)$, the RMCLS technique is the most robust technique. However, it can be observed that the RMCLS technique does not significantly improve the perceptual speech quality, yielding a similar or slightly better PESQ score and cepstral distance than the reverberant microphone signal $x_1(n)$.

Summarizing these simulation results, it can be said that the MINT, CS, and PMINT techniques are very sensitive to RIR perturbations and typically fail to achieve dereverberation. The RMCLS technique is more robust, however, its dereverberation performance in terms of direct-to-reverberant-ratio and perceptual speech quality improvement is not satisfactory. Hence, in the following chapters, several methods to increase the robustness of all considered acoustic multi-channel equalization techniques will be proposed.

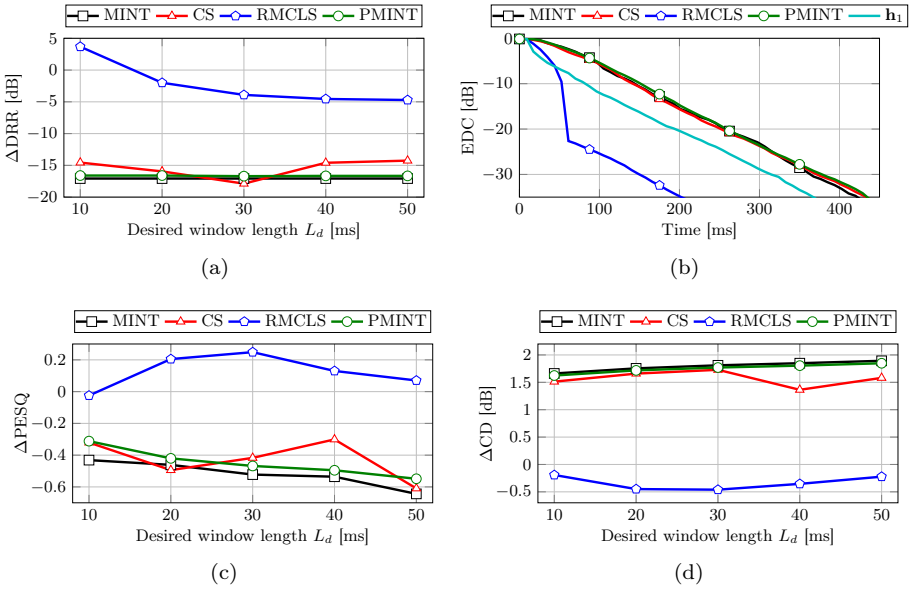


Fig. 3.4: Performance of the MINT, CS, RMCLS, and PMINT techniques in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD (averaged over several NPMs).

3.5 Summary

In this chapter an overview of existing acoustic multi-channel equalization techniques, i.e., MINT, CS, and RMCLS, has been provided and a novel partial multi-channel equalization technique based on MINT, i.e., PMINT, has been proposed. The proposed PMINT technique aims to simultaneously suppress the late reflections in the equalized impulse response and directly control the early reflections in order to preserve the perceptual speech quality.

In addition, we have established a generalized framework for least-squares acoustic multi-channel equalization techniques, i.e., for the MINT, RMCLS, and PMINT techniques. Within this framework, we showed that least-squares equalization techniques yield reshaping filters that lie in the subspace spanned by the multiple solutions maximizing the channel shortening cost function.

Simulation results have shown that if the true RIRs are known, all techniques yield a perfect reverberant tail suppression, with the PMINT technique yielding the highest perceptual speech quality. Furthermore, it has been shown that in the presence of RIR perturbations, the MINT, CS, and PMINT techniques fail to achieve dereverberation and introduce additional distortions in the output speech signal, whereas the RMCLS technique is more robust against RIR perturbations. Nevertheless, the dereverberation performance of the RMCLS technique in the presence of RIR perturbations is not satisfactory. Increasing the robustness of all considered acoustic multi-channel equalization techniques against RIR perturbations will be discussed in the following chapters.

4

ACOUSTIC MULTI-CHANNEL EQUALIZATION USING SHORTER RESHAPING FILTERS

As shown in Chapter 3, acoustic multi-channel equalization techniques are based on least-squares and generalized eigenvalue optimization criteria, which are also used in a wide range of other applications, such as data fitting and modeling using differential equations. Over the last decades, the sensitivity of least-squares and generalized eigenvalue solutions has become a well investigated topic in linear algebra. It has been shown that the sensitivity of least-squares solutions to perturbations can be evaluated by the so-called condition number of the matrix being inverted. Furthermore, it has been shown that for generalized eigenvalue solutions, infinite generalized eigenvalues are more sensitive to perturbations than finite generalized eigenvalues.

As shown in Chapter 3, when the true room impulse responses (RIRs) are known, acoustic multi-channel equalization techniques can achieve perfect dereverberation when the reshaping filter is long enough. However, since in practice the available RIRs usually differ from the true RIRs, this choice of the reshaping filter length may not be optimal. In this chapter we propose to increase the robustness of acoustic multi-channel equalization techniques by using shorter reshaping filters, such that better conditioned optimization criteria are obtained.

In Section 4.1 we derive a mathematical link between the reshaping filter length and the condition number of the (weighted) multi-channel convolution matrix, hence, the sensitivity of the least-squares equalization techniques to RIR perturbations. We show that shorter reshaping filters than conventionally used yield a smaller condition number, i.e., a higher robustness against RIR perturbations. Furthermore, in Section 4.2 it is analytically shown that shorter reshaping filters in the channel shortening technique are also more robust against RIR perturbations, since they

This chapter is partly based on:

[130] I. Kodrasi and S. Doclo, "The effect of inverse filter length on the robustness of acoustic multichannel equalization," in *Proc. European Signal Processing Conference*, Bucharest, Romania, Aug. 2012.

result in a better conditioned optimization criterion with finite generalized eigenvalues. Using instrumental performance measures, simulation results in Section 4.3 validate that decreasing the reshaping filter length in all considered acoustic multi-channel equalization techniques, i.e., MINT, CS, RMCLS, and PMINT, increases the robustness against RIR perturbations. Furthermore, it is shown that out of all considered techniques, the RMCLS and PMINT techniques using shorter reshaping filters result in the highest dereverberation performance, both in terms of reverberant energy suppression and perceptual speech quality improvement. Clearly, a positive by-product of using shorter reshaping filters is the reduction in the computational complexity of the filter design.

4.1 Reshaping filter length in least-squares equalization techniques

As described in Section 3.3, the least-squares reshaping filters can be computed as

$$\mathbf{w}_{\text{LS}} = (\mathbf{W}\hat{\mathbf{H}})^{+}(\mathbf{W}\mathbf{c}_t), \tag{4.1}$$

with $\mathbf{W}$ a $L_c \times L_c$–dimensional technique-dependent weighting matrix, cf. Section 3.3, $\hat{\mathbf{H}}$ the $L_c \times ML_w$–dimensional perturbed multi-channel convolution matrix, cf. (3.4), and $\mathbf{c}_t$ the L_c–dimensional target equalized impulse response, cf. Section 3.3. In the following, the Wedin theorem is used to evaluate the sensitivity of the least-squares reshaping filters $\mathbf{w}_{\text{LS}}$ to perturbations in the (weighted) multi-channel convolution matrix $\mathbf{W}\hat{\mathbf{H}}$.

Wedin theorem [164]: *Consider the system of equations* $\mathbf{A}\mathbf{q} = \mathbf{b}$*, where the matrix* $\mathbf{A}$ *has dimensions* $u \times v$ *and rank* $r \leq \min\{u, v\}$*. Let* $\mathbf{A}$ *be perturbed to* $\mathbf{A} + \Delta\mathbf{A}$*. The pseudo-inverse solution* $\mathbf{q} = \mathbf{A}^{+}\mathbf{b}$ *is then perturbed to* $\mathbf{q} + \Delta\mathbf{q} = (\mathbf{A}+\Delta\mathbf{A})^{+}\mathbf{b}$*, where* $\Delta\mathbf{q}$ *is the deviation between the true and the perturbed solution. The condition number* $\chi_{\mathbf{A}}$ *of the matrix* $\mathbf{A}$ *is defined as*

$$\chi_{\mathbf{A}} = \frac{\|\mathbf{A}\|_2}{\|\mathbf{A}^{+}\|_2} = \frac{\sigma_{\mathbf{A}}(1)}{\sigma_{\mathbf{A}}(r)}, \tag{4.2}$$

with $\sigma_{\mathbf{A}}(i)$ *the* i*-th singular value of the matrix* $\mathbf{A}$*, ordered such that* $\sigma_{\mathbf{A}}(1) \geq \sigma_{\mathbf{A}}(2) \geq \cdots \geq \sigma_{\mathbf{A}}(r) > 0$*. Using* $\chi_{\mathbf{A}}$ *and defining the variable* ξ *as*

$$\xi = \frac{\|\Delta\mathbf{A}\|_2}{\|\mathbf{A}\|_2}, \tag{4.3}$$

the norm of the deviation between the true and the perturbed solution is bounded by

$$\|\Delta\mathbf{q}\|_2 \leq \frac{\chi_{\mathbf{A}}\xi\|\mathbf{q}\|_2}{1 - \chi_{\mathbf{A}}\xi} + \|(\mathbf{A}\mathbf{A}^{T})^{+}\mathbf{b}\|_2\|\mathbf{A}\|_2, \tag{4.4}$$

where it is assumed that $\chi_{\mathbf{A}}\xi < 1$*.*

The inequality in (4.4) shows that a large condition number $\chi_{\mathbf{A}}$ results in a high sensitivity of least-squares solutions to perturbations in the data [164]. Hence, as

a measure of the sensitivity of the least-squares reshaping filter in (4.1) to perturbations in $\mathbf{W}\hat{\mathbf{H}}$, the condition number of $\mathbf{W}\hat{\mathbf{H}}$ can be used. In the following, it is shown that by using shorter reshaping filters than conventionally used, the condition number of $\mathbf{W}\hat{\mathbf{H}}$ can be decreased.

For clarity of presentation, in the following the notation summarized in Table 4.1 will be used. Reshaping filters in acoustic multi-channel equalization techniques are typically designed using the filter length L_t, i.e., based on the $p_t \times q_t$–dimensional matrix $\mathbf{W}_t\hat{\mathbf{H}}_t$ with fewer or the same number of rows than columns, i.e., $p_t \leq q_t$. Furthermore, the rank of the matrix $\mathbf{W}_t\hat{\mathbf{H}}_t$ is $r_t \leq p_t$. However, reshaping filters can also be designed using a shorter filter length $L_s < L_t$, i.e., based on the $p_s \times q_s$–dimensional matrix $\mathbf{W}_s\hat{\mathbf{H}}_s$. Considering that $L_s < \left\lceil \frac{L_h-1}{M-1} \right\rceil$, which implies that $L_s < \frac{L_h-1}{M-1}$, the matrix $\mathbf{W}_s\hat{\mathbf{H}}_s$ is a tall matrix with fewer columns than rows, i.e., $q_s < p_s$, since

$$(M-1)L_s < L_h - 1 \Rightarrow \underbrace{ML_s}_{q_s} < \underbrace{L_h + L_s - 1}_{p_s}. \tag{4.5}$$

Furthermore, the matrix $\mathbf{W}_s\hat{\mathbf{H}}_s$ is a full column-rank matrix, i.e., $\text{rank}(\mathbf{W}_s\hat{\mathbf{H}}_s) = r_s = q_s$. As is schematically illustrated in Fig. 4.1, the matrix $\mathbf{W}_s\hat{\mathbf{H}}_s$ is a submatrix of $\mathbf{W}_t\hat{\mathbf{H}}_t$, constructed by deleting $L_t - L_s$ rows and $M(L_t - L_s)$ columns

Table 4.1: Notation for different reshaping filter lengths and the corresponding least-squares matrices.

Variable	Denotes
$L_t = \left\lceil \frac{L_h-1}{M-1} \right\rceil$	Reshaping filter length conventionally used in acoustic multi-channel equalization techniques
$\mathbf{W}_t\hat{\mathbf{H}}_t$	Least-squares matrix when the used reshaping filter length is L_t
$p_t = L_h + L_t - 1$	Number of rows in $\mathbf{W}_t\hat{\mathbf{H}}_t$
$q_t = ML_t$	Number of columns in $\mathbf{W}_t\hat{\mathbf{H}}_t$
$r_t \leq p_t$	Rank of $\mathbf{W}_t\hat{\mathbf{H}}_t$
$L_s < L_t$	Reshaping filter length that is smaller than L_t
$\mathbf{W}_s\hat{\mathbf{H}}_s$	Least-squares matrix when the used reshaping filter length is L_s
$p_s = L_h + L_s - 1$	Number of rows in $\mathbf{W}_s\hat{\mathbf{H}}_s$
$q_s = ML_s$	Number of columns in $\mathbf{W}_s\hat{\mathbf{H}}_s$
$r_s = q_s$	Rank of $\mathbf{W}_s\hat{\mathbf{H}}_s$

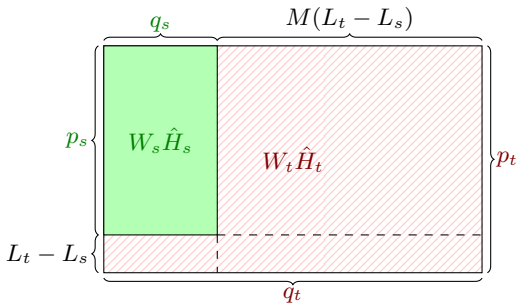


Fig. 4.1: Schematic illustration of the construction of the $p_s \times q_s$–dimensional sub-matrix $\mathbf{W}_s\hat{\mathbf{H}}_s$ from the $p_t \times q_t$–dimensional matrix $\mathbf{W}_t\hat{\mathbf{H}}_t$.

from $\mathbf{W}_t\hat{\mathbf{H}}_t$. Aiming at establishing a relation between the condition numbers of the matrices $\mathbf{W}_s\hat{\mathbf{H}}_s$ and $\mathbf{W}_t\hat{\mathbf{H}}_t$, with

$$\chi_{\mathbf{W}_s\hat{\mathbf{H}}_s} = \frac{\sigma_{\mathbf{W}_s\hat{\mathbf{H}}_s}(1)}{\sigma_{\mathbf{W}_s\hat{\mathbf{H}}_s}(r_s)}, \tag{4.6}$$

$$\chi_{\mathbf{W}_t\hat{\mathbf{H}}_t} = \frac{\sigma_{\mathbf{W}_t\hat{\mathbf{H}}_t}(1)}{\sigma_{\mathbf{W}_t\hat{\mathbf{H}}_t}(r_t)}, \tag{4.7}$$

we consider the following interlacing inequalities between the singular values of a matrix and its sub-matrices.

Interlacing inequalities [165]: *Given a matrix* $\mathbf{A}$ *of dimensions* $u \times v$ *and a sub-matrix* $\mathbf{B}$ *obtained by deleting* l *rows and/or* l *columns from* $\mathbf{A}$*, the singular values of* $\mathbf{A}$ *and* $\mathbf{B}$ *interlace as*

$$\sigma_{\mathbf{A}}(i) \geq \sigma_{\mathbf{B}}(i) \geq \sigma_{\mathbf{A}}(i+l) \qquad i = 1, \ldots, \min\{u-l, v-l\}. \tag{4.8}$$

Based on the interlacing inequalities in (4.8), in Appendix A we have derived the following inequalities relating the largest and smallest non-zero singular values of $\mathbf{W}_t\hat{\mathbf{H}}_t$ and $\mathbf{W}_s\hat{\mathbf{H}}_s$:

$$\sigma_{\mathbf{W}_t\hat{\mathbf{H}}_t}(1) \geq \sigma_{\mathbf{W}_s\hat{\mathbf{H}}_s}(1), \tag{4.9}$$

$$\sigma_{\mathbf{W}_s\hat{\mathbf{H}}_s}(r_s) \geq \sigma_{\mathbf{W}_t\hat{\mathbf{H}}_t}(r_t). \tag{4.10}$$

It readily follows from (4.9) and (4.10) that the condition number of $\mathbf{W}_s\hat{\mathbf{H}}_s$ is smaller or equal than the condition number of $\mathbf{W}_t\hat{\mathbf{H}}_t$, i.e.,

$$\chi_{\mathbf{W}_s\hat{\mathbf{H}}_s} = \frac{\sigma_{\mathbf{W}_s\hat{\mathbf{H}}_s}(1)}{\sigma_{\mathbf{W}_s\hat{\mathbf{H}}_s}(r_s)} \leq \frac{\sigma_{\mathbf{W}_t\hat{\mathbf{H}}_t}(1)}{\sigma_{\mathbf{W}_t\hat{\mathbf{H}}_t}(r_t)} = \chi_{\mathbf{W}_t\hat{\mathbf{H}}_t}. \tag{4.11}$$

Therefore, using a shorter reshaping filter than conventionally used in least-squares equalization techniques can result (and based on simulation results it generally does) in a lower condition number of the matrix being inverted.[1]

Fig. 4.2 depicts the singular values of an exemplary matrix $\mathbf{W}_t\hat{\mathbf{H}}_t$ for the PMINT technique (i.e., $\mathbf{W}_t = \mathbf{I}$) and for the RMCLS technique (i.e., $\mathbf{W}_t = \mathbf{W}_{\mathrm{R}}$), constructed using the reshaping filter length $L_t = 1200$. The used acoustic system is the same as the one described in Section 3.4.1, with $M = 4$ microphones and $L_h = 3600$. Furthermore, the singular values of two sub-matrices $\mathbf{W}_s\hat{\mathbf{H}}_s$, constructed using

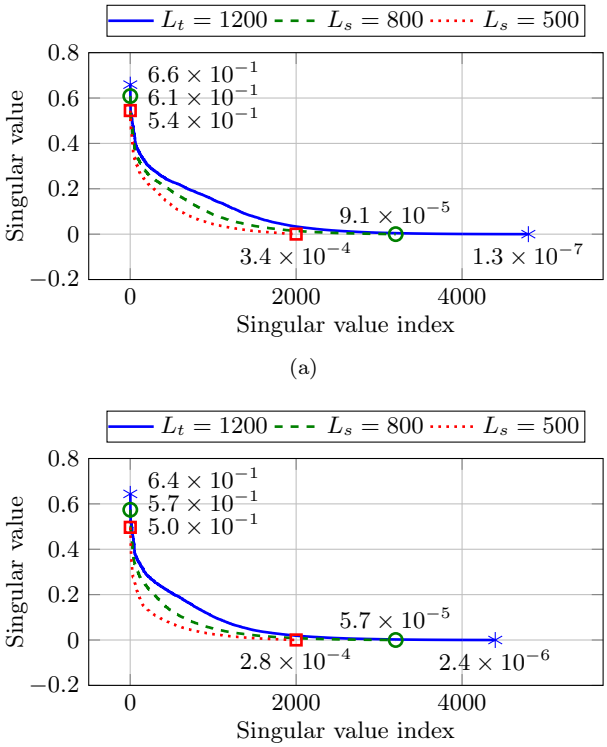


Fig. 4.2: Singular values of an exemplary matrix $\mathbf{W}_t\hat{\mathbf{H}}_t$ ($L_t = 1200$) and of two sub-matrices $\mathbf{W}_s\hat{\mathbf{H}}_s$ ($L_s = 800$ and $L_s = 500$) for the (a) PMINT technique (i.e., $\mathbf{W}_t = \mathbf{I}$) and (b) RMCLS technique (i.e., $\mathbf{W}_t = \mathbf{W}_{\mathrm{R}}$ and $L_d = 50$ ms). The largest and smallest non-zero singular values of each matrix are explicitly denoted. The considered acoustic system is the same as in Section 3.4.1.

1 It should be noted that decreasing the reshaping filter length to improve the conditioning of the least-squares optimization criterion differs from the truncated singular value decomposition approach we proposed in [127], where the singular values of the (weighted) multi-channel convolution matrix smaller than a given threshold are directly set equal to zero.

$L_s = 800$ and $L_s = 500$, are also depicted. The largest and smallest non-zero singular values of each matrix are marked in order to illustrate the inequalities presented in (4.9) and (4.10). Using these singular values, the computed condition numbers of the different matrices are presented in Table 4.2, where it can be observed that using a shorter reshaping filter than conventionally used decreases the condition number of the (weighted) multi-channel convolution matrix for the least-squares equalization techniques.

Using a shorter reshaping filter can be considered as a method of regularizing the least-squares solution, yielding a trade-off between dereverberation accuracy when the true RIRs are available and robustness in the presence of RIR perturbations. Designing shorter reshaping filters is not only desirable to increase the robustness of least-squares equalization techniques against RIR perturbations, but also because of the lower computational complexity of the filter design.

4.2 Reshaping filter length in the channel shortening technique

The analysis presented in Section 4.1 relating the reshaping filter length to the condition number of $\mathbf{W}\hat{\mathbf{H}}$, and hence, to the robustness of the reshaping filters against RIR perturbations, can be applied only to the least-squares equalization techniques, i.e., MINT, RMCLS, and PMINT. In the following we show that decreasing the reshaping filter length is however also advantageous to increase the robustness of the channel shortening technique.

As presented in Section 3.1, the channel shortening optimization criterion is a generalized eigenvalue problem, and the perturbation theory for eigenvalue problems is a well investigated topic in linear algebra [160, 166]. As shown in (3.20), the channel shortening reshaping filter can be computed as the generalized eigenvector corresponding to the maximum generalized eigenvalue $\lambda_{\max}$ of the generalized eigenvalue problem

$$\hat{\mathbf{D}}\mathbf{w}_{\mathrm{CS}} = \lambda_{\max}\hat{\mathbf{U}}\mathbf{w}_{\mathrm{CS}}. \tag{4.12}$$

Furthermore, it was shown in Section 3.3 that when the used reshaping filter length is $L_w \geq \left\lceil \frac{L_h-1}{M-1} \right\rceil$, the maximum generalized eigenvalue in (4.12) is $\lambda_{\max} = \infty$.

Table 4.2: Condition number of an exemplary matrix $\mathbf{W}_t\hat{\mathbf{H}}_t$ ($L_t = 1200$) and of two submatrices $\mathbf{W}_s\hat{\mathbf{H}}_s$ ($L_s = 800$ and $L_s = 500$) for the PMINT technique (i.e., $\mathbf{W}_t = \mathbf{I}$) and the RMCLS technique (i.e., $\mathbf{W}_t = \mathbf{W}_{\mathrm{R}}$ and $L_d = 50$ ms). The considered acoustic system is the same as in Section 3.4.1.

Reshaping filter length	PMINT	RMCLS
$L_t = 1200$	$\chi_{\mathbf{W}_t\hat{\mathbf{H}}_t} = 5 \times 10^6$	$\chi_{\mathbf{W}_t\hat{\mathbf{H}}_t} = 3 \times 10^5$
$L_s = 800$	$\chi_{\mathbf{W}_s\hat{\mathbf{H}}_s} = 7 \times 10^3$	$\chi_{\mathbf{W}_s\hat{\mathbf{H}}_s} = 1 \times 10^4$
$L_s = 500$	$\chi_{\mathbf{W}_s\hat{\mathbf{H}}_s} = 2 \times 10^3$	$\chi_{\mathbf{W}_s\hat{\mathbf{H}}_s} = 2 \times 10^3$

It has been shown in [166] that infinite generalized eigenvalues arising from the common null spaces of the involved matrices (i.e., $\hat{\mathbf{D}}$ and $\hat{\mathbf{U}}$) are extremely sensitive to small perturbations in the data. Hence, the generalized eigenvalue problem in (4.12) is highly ill-conditioned and the generalized eigenvector $\mathbf{w}_{\text{CS}}$ associated with $\lambda_{\max} = \infty$ is very sensitive to perturbations in $\hat{\mathbf{D}}$ and $\hat{\mathbf{U}}$. As will be shown in the following, decreasing the reshaping filter length in the channel shortening technique yields a better conditioned generalized eigenvalue optimization criterion with $\lambda_{\max} < \infty$, which significantly decreases the sensitivity of the reshaping filter to perturbations in $\hat{\mathbf{D}}$ and $\hat{\mathbf{U}}$.

Using a shorter reshaping filter of length L_s, the $q_s \times q_s$–dimensional matrices $\hat{\mathbf{D}}_s$ and $\hat{\mathbf{U}}_s$ are constructed similarly as in (3.16) and (3.17), i.e.,

$$\hat{\mathbf{D}}_s = \hat{\mathbf{H}}_s^T \mathbf{W}_{d_s}^T \mathbf{W}_{d_s} \hat{\mathbf{H}}_s, \tag{4.13}$$

$$\hat{\mathbf{U}}_s = \hat{\mathbf{H}}_s^T \mathbf{W}_{u_s}^T \mathbf{W}_{u_s} \hat{\mathbf{H}}_s, \tag{4.14}$$

with $\hat{\mathbf{H}}_s$ the $p_s \times q_s$–dimensional multi-channel convolution matrix and $\mathbf{W}_{d_s}$ and $\mathbf{W}_{u_s}$ the $p_s \times p_s$–dimensional desired and undesired weighting matrices defined as in (3.14) and (3.15). As described in Section 4.1, the multi-channel convolution matrix $\hat{\mathbf{H}}_s$ is a full column-rank matrix of rank q_s, hence the matrices $\hat{\mathbf{D}}_s + \hat{\mathbf{U}}_s$ and $\hat{\mathbf{U}}_s$ are also full column-rank matrices. Following similar arguments as in Section 3.3, since

$$\text{rank}(\hat{\mathbf{D}}_s + \hat{\mathbf{U}}_s) = q_s, \tag{4.15}$$

$$\text{rank}(\hat{\mathbf{U}}_s) = q_s, \tag{4.16}$$

the null space of the matrices $\hat{\mathbf{D}}_s + \hat{\mathbf{U}}_s$ and $\hat{\mathbf{U}}_s$ is empty, i.e.,

$$\dim[\text{null space}(\hat{\mathbf{D}}_s + \hat{\mathbf{U}}_s)] = 0, \tag{4.17}$$

$$\dim[\text{null space}(\hat{\mathbf{U}}_s)] = 0. \tag{4.18}$$

Therefore, the generalized eigenvalue problem

$$\hat{\mathbf{D}}_s \mathbf{w}_{\text{CS}} = \lambda_{\max_s} \hat{\mathbf{U}}_s \mathbf{w}_{\text{CS}} \tag{4.19}$$

does not have infinite generalized eigenvalues, i.e., $\lambda_{\max_s} < \infty$.

Table 4.3 presents the maximum generalized eigenvalue for exemplary matrices $\hat{\mathbf{D}}_s$ and $\hat{\mathbf{U}}_s$ constructed using the shorter reshaping filter lengths $L_s = 800$ and $L_s = 500$. The used acoustic system is the same as the one described in Section 3.4.1, with $M = 4$ microphones and $L_h = 3600$. It can be observed that decreasing the reshaping filter length in the channel shortening technique results in finite generalized eigenvalues.

Hence, decreasing the reshaping filter length in the channel shortening technique improves the conditioning of the optimization criterion, yielding generalized eigenvectors that are less sensitive to perturbations in $\hat{\mathbf{D}}_s$ and $\hat{\mathbf{U}}_s$. It should be noted that when the used reshaping filter length is $L_s < \left\lceil \frac{L_h - 1}{M - 1} \right\rceil$, the least-squares equalization techniques do not yield reshaping filters that satisfy the channel shortening optimization criterion, i.e., the analysis in Section 3.3 does not apply.

Table 4.3: Maximum generalized eigenvalue for channel shortening exemplary matrices $\hat{\mathbf{D}}_s$ and $\hat{\mathbf{U}}_s$ constructed $L_s = 800$ and $L_s = 500$. The considered acoustic system is the same as in Section 3.4.1 and the desired window length is $L_d = 50$ ms.

Reshaping filter length	$\lambda_{\max_s}$
$L_s = 800$	1×10^5
$L_s = 500$	7×10^3

4.3 Simulations

In this section, we investigate the dereverberation performance of all considered equalization techniques when using shorter reshaping filters than conventionally used. In Section 4.3.1 the considered acoustic system and the used algorithmic settings are introduced. Section 4.3.2 investigates the robustness increase of acoustic multi-channel equalization techniques when using shorter reshaping filters. In Section 4.3.3 the performance of the robust extensions of the considered techniques is extensively compared.

4.3.1 *Acoustic system and algorithmic settings*

We have considered the same acoustic scenario as in Section 3.4.1, i.e., a single speech source and $M = 4$ omni-directional microphones. The source-microphone distance is 3 m and the distance between the microphones is 5 cm. Room impulse responses from the MARDY database [161] have been used, where the room reverberation time is $T_{60} \approx 450$ ms and the direct-to-reverberant ratio is DRR $= 0$ dB. The RIRs have been measured using the swept-sine technique [162] and the length of the RIRs has been set to $L_h = 3600$ at a sampling frequency $f_s = 8$ kHz.

Similarly as in Section 3.4, in order to simulate RIR perturbations, the measured RIRs are perturbed by adding scaled white noise as described in Section 2.2. The considered normalized projection misalignment (NPM) values between the true and the perturbed RIRs are (cf. (2.52))

$$\text{NPM} \in \{-33 \text{ dB}, \ -27 \text{ dB}, \ -21 \text{ dB}, \ -15 \text{ dB}\}. \tag{4.20}$$

For all considered acoustic multi-channel equalization techniques, the conventionally used reshaping filter length is $L_t = \left\lceil \frac{L_h - 1}{M - 1} \right\rceil = 1200$, the delay is set to $\tau = 90$, and the performance for several desired window lengths L_d ranging from 10 ms to 50 ms is investigated, i.e.,

$$L_d \in \{10 \text{ ms}, \ 20 \text{ ms}, \ 30 \text{ ms}, \ 40 \text{ ms}, \ 50 \text{ ms}\}. \tag{4.21}$$

The target equalized impulse response for the PMINT technique is set to the direct path and the early reflections of the perturbed RIR of the first microphone, i.e., $\hat{\mathbf{h}}_{e,1}$.

Furthermore, when the used reshaping filter length is L_t, the channel shortening reshaping filter is selected as the generalized eigenvector yielding the minimum l_2-norm estimated equalized impulse response as proposed in [125].

Using the instrumental performance measures described in Section 2.3, the dereverberation performance is evaluated in terms of the reverberant energy suppression and the perceptual speech quality improvement. The reverberant energy suppression is evaluated using the direct-to-reverberant ratio improvement (ΔDRR) between the equalized impulse response $\mathbf{c}$ and the true RIR $\mathbf{h}_1$ (cf. (2.53)), as well as the energy decay curve (EDC) of the equalized impulse response $\mathbf{c}$ (cf. (2.55)). The improvement in perceptual speech quality is evaluated using the improvement in PESQ [153] (ΔPESQ) and in cepstral distance [154] (ΔCD) between the output speech signal $z(n)$ and the reverberant microphone signal $x_1(n)$. The reference signal employed for the PESQ and cepstral distance measures is $x_{e,1}(n) = s(n) * h_{e,1}(n)$, i.e., the clean speech signal convolved with the direct path and early reflections of the first RIR (which changes as the desired window length L_d changes).

In order to evaluate the effectiveness of using shorter reshaping filters for the considered acoustic multi-channel equalization techniques, we investigate the performance for several filter lengths L_s, i.e.,

$$L_s \in \{500,\ 600,\ \ldots,\ 1100,\ 1200\}. \tag{4.22}$$

The optimal filter length L_o is determined to be the filter length yielding the highest perceptual speech quality in terms of the PESQ score. It should be noted that the computation of the PESQ score for determining the optimal filter length is an intrusive procedure that is not applicable in practice, since knowledge of the clean speech signal and of the true RIRs is required to compute the reference signal and the equalized impulse response $\mathbf{c} = \mathbf{H}\mathbf{w}$. In addition, it should be noted that although the MINT technique is independent of the desired window length L_d, we determine an optimal filter length in the MINT technique for each desired window length L_d by changing the reference signal in the PESQ computation, such that the MINT technique can be compared to partial equalization techniques (which are dependent on the desired window length L_d).

4.3.2 *Robustness increase of acoustic multi-channel equalization techniques when using shorter reshaping filters*

In this section, the performance of all considered acoustic multi-channel equalization techniques when using the conventional reshaping filter length L_t is compared to the performance when using the optimal shorter reshaping filter length L_o. We consider an exemplary scenario with NPM $= -33$ dB. For the sake of clarity, we will refer to the equalization techniques using the shorter reshaping filter length L_o as L-MINT, L-CS, L-RMCLS, and L-PMINT.

Robustness increase of the MINT technique

Fig. 4.3 depicts the performance of the MINT and L-MINT techniques in terms of ΔDRR, EDC, ΔPESQ, and ΔCD. For completeness, the used optimal reshaping filter lengths are presented in Table 4.4. Since acoustic system inversion based on the MINT technique is very sensitive to RIR perturbations, the optimal reshaping filter length as illustrated in Table 4.4 is small, i.e., $L_o = 500$.

As shown by the ΔDRR values depicted in Fig. 4.3a, using the L-MINT technique significantly increases the ΔDRR in comparison to using the MINT technique. While the conventionally used reshaping filter length worsens the DRR in comparison to the true RIR $\mathbf{h}_1$, using a shorter reshaping filter yields an improvement of approximately 5 dB.

To evaluate the reverberant energy decay rate, Fig. 4.3b shows the energy decay curve of the true RIR $\mathbf{h}_1$ and the energy decay curve of the equalized impulse re-

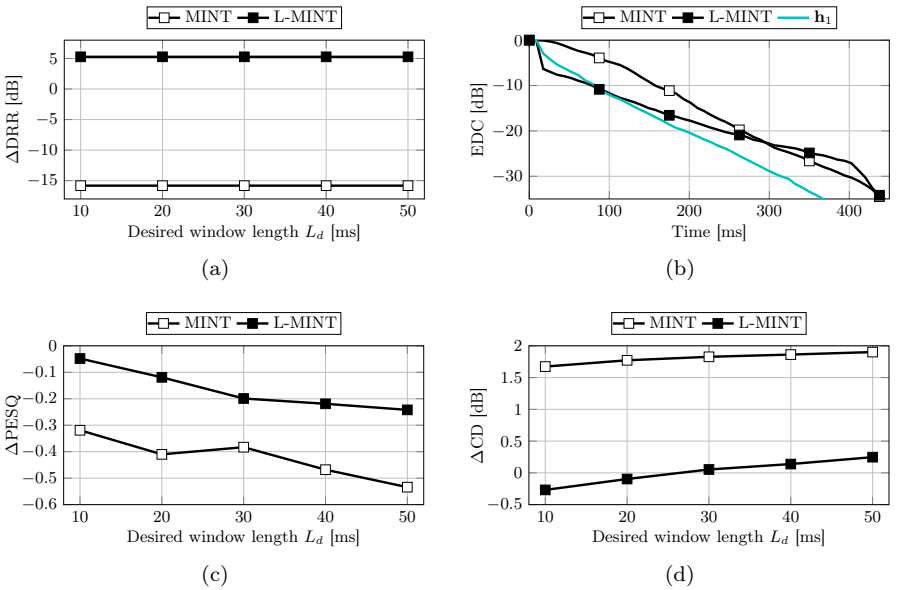


Fig. 4.3: Performance of the MINT and L-MINT techniques in terms of (a) ΔDRR, (b) EDC, (c) ΔPESQ, and (d) ΔCD (NPM = −33 dB).

Table 4.4: Optimal reshaping filter length for the L-MINT technique for several desired window lengths (NPM = −33 dB).

Desired window length L_d [ms]	10	20	30	40	50
Optimal reshaping filter length L_o	500	500	500	500	500

sponse $\mathbf{c}$ obtained using the MINT ($L_t = 1200$) and L-MINT ($L_s = 500$) techniques. It can be observed that while the reverberant energy decays faster when using the L-MINT technique, it is nevertheless generally decaying at a slower rate than the reverberant energy in the true RIR $\mathbf{h}_1$.

To evaluate the perceptual speech quality, Figs. 4.3c and 4.3d depict the ΔPESQ and ΔCD values achieved by the MINT and L-MINT techniques. It can be observed that using the L-MINT technique yields a better overall perceptual speech quality than using the MINT technique. However, as can be seen by the negative ΔPESQ values and the approximately zero ΔCD values, the overall perceptual speech quality is still not improved in comparison to the reverberant microphone signal $x_1(n)$.

Therefore as expected from the theoretical analysis in Section 4.1, these simulation results validate that using a shorter reshaping filter than conventionally used in the least-squares MINT technique is advantageous to increase the robustness against RIR perturbations. However, acoustic system inversion using the L-MINT technique nevertheless remains sensitive even to moderate RIR perturbation levels, yielding a slow reverberant energy decay rate and a worse overall perceptual speech quality than the reverberant microphone signal. Hence, using a shorter reshaping filter is not sufficient to make the MINT technique robust against RIR perturbations.

Robustness increase of the CS technique

Fig. 4.4 depicts the performance of the CS and L-CS techniques in terms of ΔDRR, EDC, ΔPESQ, and ΔCD. The used optimal reshaping filter lengths are presented in Table 4.5. Since the optimal filter length depends on the acoustic system, the desired window length, and the level of RIR perturbations, no general statement about the value of the optimal filter length can be made. However, it can be observed that the optimal reshaping filter lengths for the CS technique are smaller than the conventionally used filter length L_t and larger than the optimal reshaping filter lengths for the MINT technique, since by design partial channel equalization is typically more robust than channel inversion using the MINT technique.[2]

As shown in Fig. 4.4a, using a shorter reshaping filter than conventionally used in the CS technique significantly increases the ΔDRR values, particularly for short desired window lengths L_d. As the desired window length increases, the ΔDRR obtained using the L-CS technique decreases, since additional energy is introduced which is accounted for as reverberant energy in the DRR calculation (cf. (2.54)).

To evaluate the reverberant energy decay rate, Fig. 4.4b depicts the energy decay curve of the true RIR $\mathbf{h}_1$ and the energy decay curve of the equalized impulse response $\mathbf{c}$ obtained using the CS ($L_t = 1200$) and L-CS ($L_s = 800$) techniques for the desired window length $L_d = 50$ ms. As expected, the CS technique fails to

2 The better performance of the CS technique in comparison to the MINT technique is not apparent here. However, as described in Section 3.4.3, selecting the generalized eigenvector as the one yielding the minimum l_2-norm equalized impulse response for the CS technique does not yield the best performance in the presence of RIR perturbations, i.e., significantly better performing generalized eigenvectors can be found.

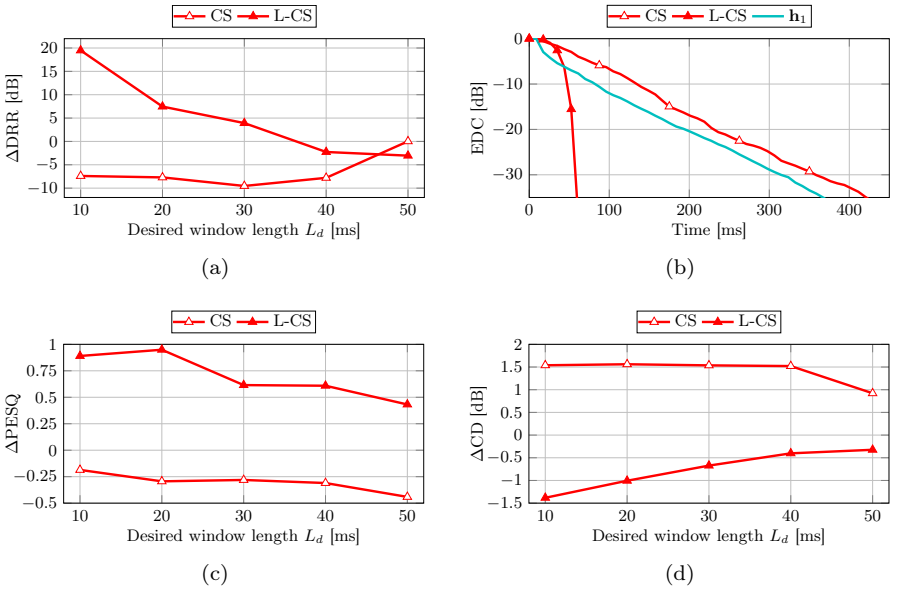


Fig. 4.4: Performance of the CS and L-CS techniques in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD (NPM = −33 dB).

Table 4.5: Optimal reshaping filter length for the L-CS technique for several desired window lengths (NPM = −33 dB).

Desired window length L_d [ms]	10	20	30	40	50
Optimal reshaping filter length L_o	900	1000	1100	700	800

achieve dereverberation, whereas the L-CS technique yields a significant increase in robustness against RIR perturbations, with the reverberant energy decaying at a much faster rate than the reverberant energy in the true RIR $\mathbf{h}_1$. The high direct-to-reverberant ratio improvement and the faster reverberant energy decay rate achieved by the L-CS technique is also reflected in the overall perceptual speech quality improvement as measured by ΔPESQ and ΔCD, depicted in Figs. 4.4c and 4.4d. It can be observed that while the CS technique fails to improve the perceptual speech quality for all considered desired window lengths, the L-CS technique results in a significant improvement in comparison to the reverberant microphone signal $x_1(n)$. As expected, the ΔPESQ values for the L-CS technique decrease with increasing desired window length and the ΔCD values increase, since more early reflections are left uncontrolled.

Therefore as expected from the theoretical analysis in Section 4.2, these simulation results demonstrate that using a shorter reshaping filter than conventionally used in

the CS technique is advantageous and significantly increases the robustness against RIR perturbations.

Robustness increase of the RMCLS technique

Fig. 4.5 depicts the performance of the RMCLS and L-RMCLS techniques in terms of ΔDRR, EDC, ΔPESQ, and ΔCD. The used optimal reshaping filter lengths are presented in Table 4.6, where it can be observed that for all considered desired window lengths, the optimal reshaping filter length L_s is smaller than the conventionally used reshaping filter length L_t.

Since the RMCLS technique is significantly more robust than the MINT and CS techniques (cf. Section 3.4.3), it can be observed in Fig. 4.5a that the DRR improvement obtained when a shorter filter is used in the RMCLS technique is less than the

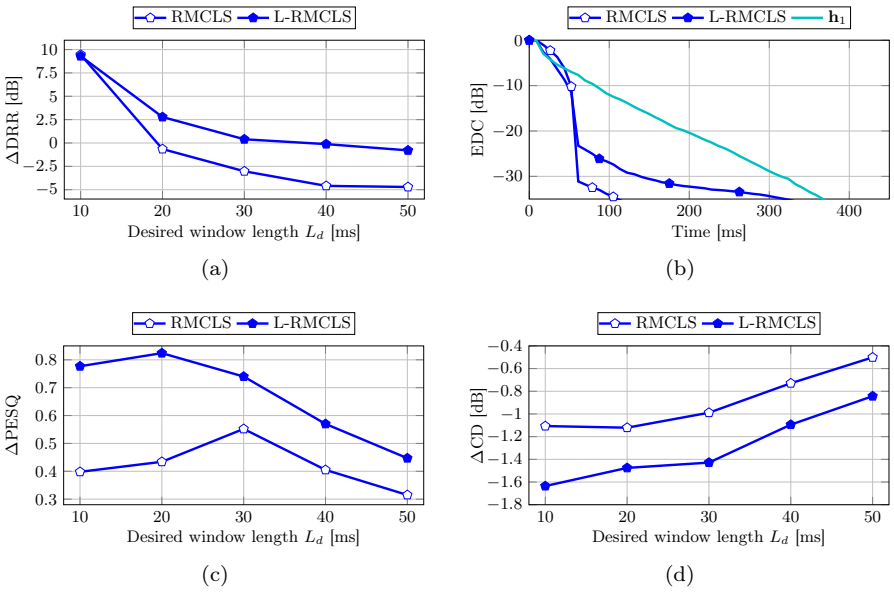


Fig. 4.5: Performance of the RMCLS and L-RMCLS techniques in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD (NPM = −33 dB).

Table 4.6: Optimal reshaping filter length for the L-RMCLS technique for several desired window lengths (NPM = −33 dB).

Desired window length L_d [ms]	10	20	30	40	50
Optimal reshaping filter length L_o	900	600	700	600	600

improvement obtained when a shorter filter is used in the MINT or CS techniques. Nevertheless, for increasing desired window lengths, an improvement in ΔDRR of up to 5 dB is obtained when using the L-RMCLS technique instead of the RMCLS technique.

To evaluate the reverberant energy decay rate, Fig. 4.5b depicts the energy decay curve of the true RIR $\mathbf{h}_1$ and the energy decay curve of the equalized impulse response $\mathbf{c}$ obtained using the RMCLS ($L_t = 1200$) and L-RMCLS ($L_s = 600$) techniques for the desired window length $L_d = 50$ ms. It can be observed that for the L-RMCLS technique the reverberant energy decays at a slower rate than for the RMCLS technique. This can be explained by the fact that the optimal reshaping filter length is being chosen as the one yielding the highest PESQ score. Since the RMCLS technique using the conventional filter length L_t yields a fast reverberant energy decay rate but not a high perceptual speech quality improvement, a shorter reshaping filter which yields a better perceptual speech quality results in a slower reverberant energy decay rate. The latter is illustrated in Figs. 4.5c and 4.5d, which show that the L-RMCLS technique yields a better overall perceptual speech quality than the RMCLS technique, significantly improving the obtained ΔPESQ and ΔCD values for all desired window lengths L_d.

Therefore as expected from the theoretical analysis in Section 4.1, these simulation results demonstrate that using a shorter reshaping filter than conventionally used in the weighted least-squares RMCLS technique is advantageous and increases the robustness against RIR perturbations.

Robustness increase of the PMINT technique

Fig. 4.6 depicts the performance of the PMINT and L-PMINT techniques in terms of ΔDRR, EDC, ΔPESQ, and ΔCD. The used optimal reshaping filter lengths are presented in Table 4.7, where it can be observed that for all considered desired window lengths, the optimal reshaping filter length L_s is smaller than the conventionally used reshaping filter length L_t.

Similarly as for the MINT and CS techniques, it can be observed in Figs. 4.6a and 4.6b that using a shorter reshaping filter than conventionally used in the PMINT technique results in a significant improvement in reverberant energy suppression, both in terms of a higher direct-to-reverberant ratio improvement and a faster decay rate of the reverberant energy.

Furthermore, the ΔPESQ and ΔCD values depicted in Figs. 4.6c and 4.6d confirm that using the PMINT technique fails to improve the overall perceptual speech quality in comparison to the reverberant microphone signal, whereas using the L-PMINT technique yields a significantly better performance, with an improvement of approximately 0.7 in PESQ score and 1.5 dB in cepstral distance for all considered desired window lengths.

Hence, as expected from the theoretical analysis in Section 4.1, decreasing the reshaping filter length in the least-squares PMINT technique results in a significant increase in robustness against RIR perturbations, both in terms of reverberant energy suppression and perceptual speech quality improvement.

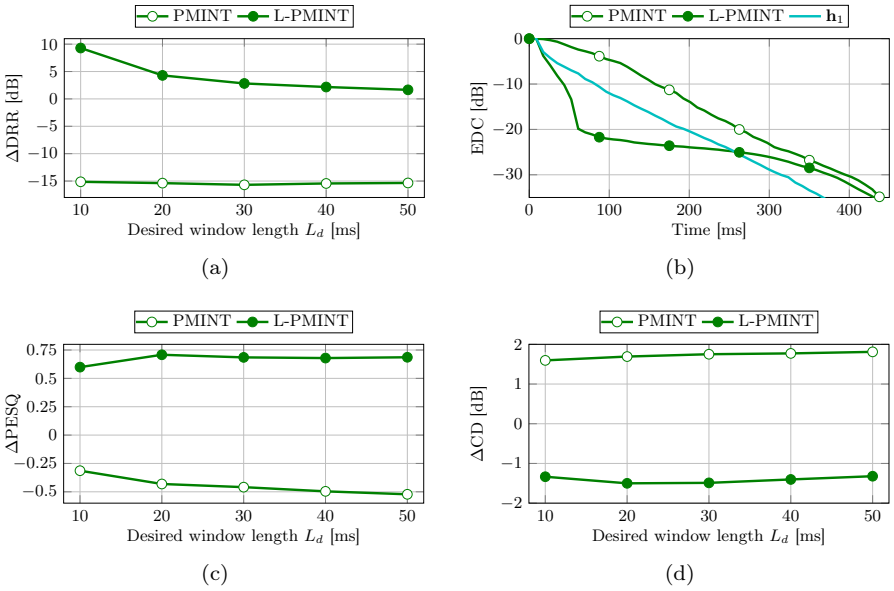


Fig. 4.6: Performance of the PMINT and L-PMINT techniques in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD (NPM = −33 dB).

Table 4.7: Optimal reshaping filter length for the L-PMINT technique for several desired window lengths (NPM = −33 dB).

Desired window length L_d [ms]	10	20	30	40	50
Optimal reshaping filter length L_o	600	600	600	600	600

4.3.3 *Comparison of robust acoustic multi-channel equalization techniques*

The simulation results in Section 4.3.2 have shown that for all considered equalization techniques, shorter reshaping filters than conventionally used yield a significant increase in robustness in the presence of RIR perturbations. In this section, the performance of acoustic multi-channel equalization techniques using optimal shorter reshaping filters is extensively compared for all considered NPM values in (4.20). Similarly as in Section 3.4.3, the presented performance measures are averaged over all considered NPM values.

To compare the reverberant energy suppression, Figs. 4.7a and 4.7b depict the DRR improvement and the energy decay curve obtained by all equalization techniques when shorter reshaping filters are used. It can be observed in Fig. 4.7a that for short desired window lengths (i.e., $L_d = 10$ ms), partial equalization techniques achieve

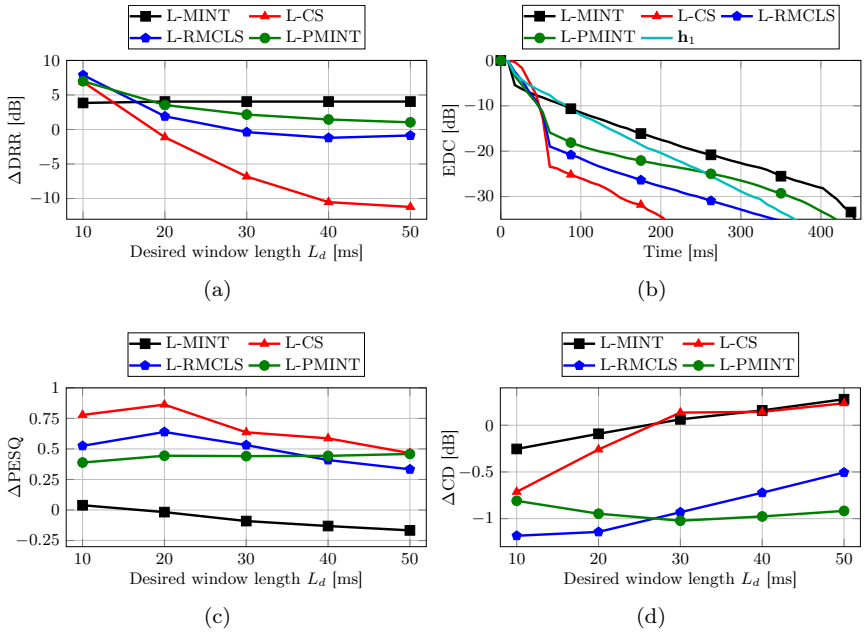


Fig. 4.7: Performance of the L-MINT, L-CS, L-RMCLS, and L-PMINT techniques in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD (averaged over several NPM values).

the highest ΔDRR and outperform the L-MINT technique. As the desired window length increases, due to the additional energy introduced in the early reflections by partial equalization techniques, the L-MINT technique results in a higher ΔDRR. However, as can be seen from the energy decay curves presented in Fig. 4.7b, the L-MINT technique yields the slowest decay rate of the reverberant energy, whereas the partial equalization techniques are significantly more robust. Among the partial equalization techniques, it can be observed that due to its energy-based optimization criterion, the L-CS technique results in the worst DRR improvement but the fastest decay rate of the late reverberant energy. On the other hand, the L-RMCLS and L-PMINT techniques result in a good performance in terms of both performance measures, with the L-PMINT technique yielding a better DRR improvement but a slower reverberant energy decay rate than the L-RMCLS technique.

To compare the overall perceptual speech quality improvement, Figs. 4.7c and 4.7d depict the PESQ score and cepstral distance improvement obtained by all considered techniques. As expected from the previously presented simulation results, the L-MINT technique fails to improve the perceptual speech quality in terms of both perceptual measures. Furthermore, among the partial acoustic multi-channel equalization techniques, the L-CS technique yields the highest improvement in terms of PESQ but the lowest improvement in terms of cepstral distance, which is a somewhat contradictory result.

To better understand the difference arising between these two perceptual measures for the L-CS technique, Fig. 4.8 depicts the spectrograms of the reference signal $x_{e,1}(n)$, reverberant microphone signal $x_1(n)$, output speech signal $z(n)$ obtained using the CS technique ($L_t = 1200$), and output speech signal $z(n)$ obtained using the L-CS technique ($L_s = 800$) for an exemplary scenario with NPM $= -33$ dB and $L_d = 50$ ms. While the L-CS technique suppresses the reverberant energy, it can be observed in Fig. 4.8d that also frequency components of the desired speech signal are suppressed. Furthermore, an audible high-energy tone appears at approximately 1.7 kHz. These distortions are clearly not reflected in the high PESQ score improvement achieved by the L-CS technique but are reflected in the low cepstral distance improvement. Hence, it can be said that the L-CS technique achieves reverberant energy suppression but introduces additional audible artifacts and distortions in the output speech signal.

On the other hand, the L-RMCLS and L-PMINT techniques yield a significantly better performance in terms of both perceptual quality measures (cf. Figs. 4.7c and 4.7d), with the L-RMCLS technique yielding a better perceptual speech quality for short desired window lengths and the L-PMINT technique yielding a better perceptual speech quality for longer desired window lengths.

Summarizing these simulations results, out of the considered robust extensions of acoustic multi-channel equalization techniques, the L-RMCLS and L-PMINT techniques result in the highest dereverberation performance, both in terms of reverberant energy suppression and perceptual speech quality improvement.

4.4 Summary

In this chapter we have investigated the effect of the reshaping filter length on the robustness of acoustic multi-channel equalization techniques against RIR perturbations. The condition number of the (weighted) multi-channel convolution matrix has been used to evaluate the sensitivity of the least-squares equalization techniques to RIR perturbations. We have analytically shown that using shorter reshaping filters than conventionally used results in a lower condition number, and hence, in an increased robustness against RIR perturbations. Furthermore, we have also analytically shown that shorter reshaping filters in the channel shortening technique are less sensitive to RIR perturbations, since they result in a better conditioned optimization criterion with finite generalized eigenvalues.

The presented simulation results have validated that decreasing the reshaping filter length in all considered acoustic multi-channel equalization techniques yields a significantly better dereverberation performance in the presence of RIR perturbations. Furthermore, it has been shown that out of the considered techniques, the RMCLS and PMINT techniques using shorter reshaping filters result in the best dereverberation performance, both in terms of reverberant energy suppression and perceptual speech quality improvement. The advantage of building upon the RMCLS technique to increase the robustness against RIR perturbations lies in its relaxed optimization criterion, whereas the advantage of building upon the PMINT technique lies in its direct control of the early reflections of the equalized impulse response.

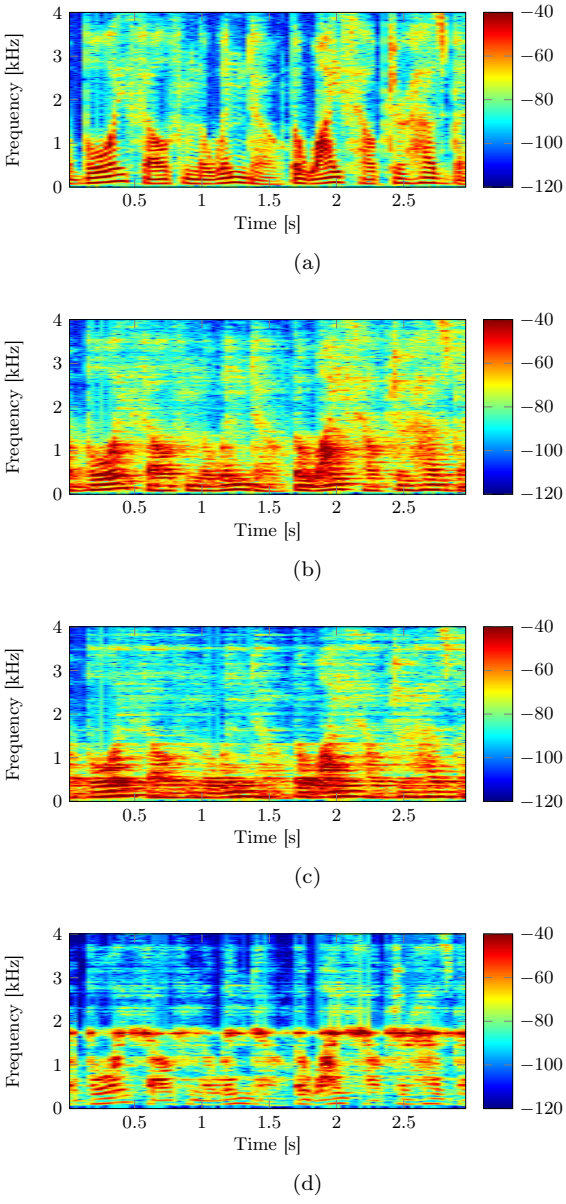


Fig. 4.8: Spectrograms of the (a) reference signal, (b) reverberant microphone signal, (c) output speech signal obtained using the CS technique ($L_t = 1200$), and (d) output speech signal obtained using the L-CS technique ($L_s = 800$) ($L_d = 50$ ms and NPM $= -33$ dB).

It should however be noted that the optimal reshaping filter length has been determined intrusively. An automatic non-intrusive procedure for determining the reshaping filter length remains a topic for future investigation.

5

REGULARIZED ACOUSTIC MULTI-CHANNEL EQUALIZATION

As shown in Chapter 3, although acoustic multi-channel equalization techniques can in theory achieve perfect dereverberation, in practice they are sensitive to room impulse response (RIR) perturbations. Whereas in Chapter 4 we have shown that the robustness of acoustic multi-channel equalization techniques against RIR perturbations can be increased by using shorter reshaping filters, in this chapter we propose to increase the robustness by incorporating the energy of distortions arising due to RIR perturbations in the different optimization criteria, such that this energy is reduced. The distortion energy term is scaled by a regularization parameter, which enables to trade off between the dereverberation error energy and the distortion energy. In general, the optimal regularization parameter yielding the best performance needs to be determined intrusively (i.e., using knowledge of the true RIRs), limiting the practical applicability of the regularized equalization techniques. Therefore, in this chapter we also propose and investigate an automatic non-intrusive procedure for determining the regularization parameter based on the L-curve.

Section 5.1 establishes the general framework for incorporating regularization in acoustic multi-channel equalization techniques, i.e., in the MINT, CS, RMCLS, and PMINT techniques. The regularized least-squares and channel shortening reshaping

This chapter is partly based on:

[126] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1879–1890, Sep. 2013.

[127] I. Kodrasi and S. Doclo, "Robust partial multichannel equalization techniques for speech dereverberation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Kyoto, Japan, Mar. 2012, pp. 537–540.

[131] I. Kodrasi and S. Doclo, "Increasing the robustness of acoustic multichannel equalization by means of regularization," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Aachen, Germany, Sep. 2012, pp. 161–164.

[132] I. Kodrasi, S. Goetze, and S. Doclo, "Non-intrusive regularization for least-squares multichannel equalization for speech dereverberation," in *Proc. IEEE Convention of Electrical and Electronics Engineers in Israel*, Eilat, Israel, Nov. 2012.

filters are discussed in Section 5.2. In Section 5.3 it is proposed to automatically determine the regularization parameter as the point of maximum curvature of the parametric plot of the distortion energy versus the dereverberation error energy. The proposed automatic procedure is only applicable to the regularized least-squares techniques. Insights on why this procedure cannot be used for the regularized channel shortening technique are provided in Section 5.4. By means of instrumental performance measures, simulation results in Section 5.5 show that incorporating regularization in the considered acoustic multi-channel equalization techniques significantly increases the robustness against RIR perturbations. It is shown that the intrusively regularized PMINT technique outperforms all considered intrusively regularized multi-channel equalization techniques in terms of overall perceptual speech quality improvement. Finally, it is shown that the automatically determined non-intrusive regularization parameter in regularized PMINT leads to a similar performance as the intrusively determined optimal regularization parameter, making the regularized PMINT technique a robust, perceptually advantageous, and practically applicable multi-channel equalization technique for speech dereverberation.

5.1 Incorporating regularization in acoustic multi-channel equalization

As discussed in Section 2.1.2, acoustic multi-channel equalization techniques typically disregard the presence of background noise and design reshaping filters aiming only at speech dereverberation. Assuming that $\mathbf{v}(n) = \mathbf{0}$ in (2.26), the output signal of the speech enhancement system is given by

$$z(n) = \underbrace{\mathbf{w}^T\mathbf{H}^T}_{\mathbf{c}^T}\mathbf{s}(n), \tag{5.1}$$

with $\mathbf{w}$ the ML_w–dimensional reshaping filter vector, cf. (2.14), $\mathbf{H}$ the $L_c \times ML_w$–dimensional multi-channel convolution matrix of the true RIRs, cf. (2.24), $\mathbf{s}(n)$ the L_c–dimensional clean speech vector, cf. (2.23), and $\mathbf{c}$ the L_c–dimensional equalized impulse response between the clean speech signal and the output speech signal, cf. (2.27). Since the multi-channel convolution matrix of the true RIRs is typically not available, acoustic multi-channel equalization techniques design reshaping filters using the perturbed convolution matrix

$$\hat{\mathbf{H}} = \mathbf{H} + \mathbf{E}, \tag{5.2}$$

with $\mathbf{E}$ the convolution matrix of the RIR perturbations. When reshaping filters are designed using the perturbed convolution matrix $\hat{\mathbf{H}}$, the *true* equalized impulse response can be written as

$$\mathbf{c} = \mathbf{H}\mathbf{w} = (\hat{\mathbf{H}} - \mathbf{E})\mathbf{w} = \hat{\mathbf{H}}\mathbf{w} - \mathbf{E}\mathbf{w}. \tag{5.3}$$

When the reshaping filter length is $L_w \geq \left\lceil \frac{L_h - 1}{M - 1} \right\rceil$, equalization techniques achieve their design objectives for the perturbed full row-rank matrix $\hat{\mathbf{H}}$ (cf. Section 3.3). Hence, when using the least-squares or the channel shortening reshaping filters derived in Chapter 3, the first term in (5.3) represents the (weighted) *target* equalized

impulse response cf. (3.28), whereas the second term represents distortions due to RIR perturbations. Considering the RIR perturbations to be random fluctuations from the true RIRs, the mean distortion energy in the *true* equalized impulse response is given by

$$\mathcal{E}\{\|\mathbf{E}\mathbf{w}\|_2^2\} = \mathbf{w}^T\mathcal{E}\{\mathbf{E}^T\mathbf{E}\}\mathbf{w}, \tag{5.4}$$

with $\mathcal{E}$ the expected value operator. In order to reduce the mean distortion energy in the *true* equalized impulse response and thereby increase the robustness of acoustic multi-channel equalization techniques, in this chapter we propose to add the mean distortion energy term in (5.4) to the least-squares and channel shortening cost functions J_{LS} and J_{CS} defined in (3.29) and (3.18).

The matrix $\mathcal{E}\{\mathbf{E}^T\mathbf{E}\}$ in (5.4) obviously depends on the energy and the type of RIR perturbations, e.g., perturbations arising due to microphone position deviations [109], or perturbations arising due to supervised or blind system identification methods [101,110]. While models have been developed to characterize different types of perturbations as described in Section 2.2, the exact matrix $\mathcal{E}\{\mathbf{E}^T\mathbf{E}\}$ is typically not known in practice. To account for inaccuracies in modeling $\mathcal{E}\{\mathbf{E}^T\mathbf{E}\}$, we propose to introduce a parameter δ and use

$$\mathcal{E}\{\mathbf{E}^T\mathbf{E}\} = \delta\mathbf{R}_e, \tag{5.5}$$

with $\mathbf{R}_e$ constructed based on a perturbation model (e.g., as proposed in [145, 148]) and assumed to be a full-rank matrix. When no knowledge about the type of perturbations is available, they can be assumed to be spatially and temporally white, i.e., $\mathbf{R}_e = \mathbf{I}$, with $\mathbf{I}$ being the $ML_w \times ML_w$–dimensional identity matrix.

Incorporating the term in (5.5) with $\mathbf{R}_e = \mathbf{I}$ in the MINT technique has already been investigated in [108], where it has been experimentally validated that a significant robustness increase against RIR perturbations can be obtained. In this chapter, we investigate the effectiveness of incorporating this term to increase the robustness of partial multi-channel equalization techniques, i.e., the CS, RMCLS, and PMINT techniques.

Incorporating the term in (5.5) in the least-squares cost function J_{LS} in (3.29) yields

$$J_{\text{R-LS}} = J_{\text{LS}} + \delta\mathbf{w}^T\mathbf{R}_e\mathbf{w} \tag{5.6}$$

$$= \underbrace{\|\mathbf{W}(\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t)\|_2^2}_{\epsilon_c} + \delta\,\underbrace{\mathbf{w}^T\mathbf{R}_e\mathbf{w}}_{\epsilon_e}, \tag{5.7}$$

where ϵ_c denotes the dereverberation error energy for the least-squares techniques, ϵ_e denotes the distortion energy due to RIR perturbations, and the parameter δ can be viewed as a regularization parameter providing a trade-off between both terms. Hence, the cost function in (5.7) is referred to as the regularized least-squares cost function.

In order to incorporate the distortion energy in the channel shortening technique, the generalized Rayleigh quotient *maximization* problem in (3.18) is first reformulated

in terms of a generalized Rayleigh quotient *minimization* problem, such that the channel shortening cost function to be minimized can be written as

$$J_{\text{CS}}^{\min} = \frac{\mathbf{w}^T\hat{\mathbf{U}}\mathbf{w}}{\mathbf{w}^T\hat{\mathbf{D}}\mathbf{w}}. \tag{5.8}$$

Using (5.8), the proposed regularized channel shortening cost function $J_{\text{R-CS}}$ can be written as

$$J_{\text{R-CS}} = J_{\text{CS}}^{\min} + \delta\mathbf{w}^T\mathbf{R}_e\mathbf{w} \tag{5.9}$$

$$= \underbrace{\frac{\mathbf{w}^T\hat{\mathbf{U}}\mathbf{w}}{\mathbf{w}^T\hat{\mathbf{D}}\mathbf{w}}}_{\epsilon_r} + \delta\underbrace{\mathbf{w}^T\mathbf{R}_e\mathbf{w}}_{\epsilon_e}, \tag{5.10}$$

where ϵ_r denotes the dereverberation error energy for the channel shortening technique.

5.2 Regularized acoustic multi-channel equalization reshaping filters

In this section, the regularized least-squares and channel shortening reshaping filters will be derived and discussed. Furthermore, analytical insights about the impact of the regularization parameter in the regularized least-squares reshaping filter will be provided.

Regularized least-squares reshaping filter

To compute the regularized least-squares reshaping filter minimizing (5.7), the gradient of the cost function $J_{\text{R-LS}}$ with respect to $\mathbf{w}$ is set to $\mathbf{0}$, i.e.,

$$\frac{\partial J_{\text{R-LS}}}{\partial \mathbf{w}} = 2(\mathbf{W}\hat{\mathbf{H}})^T(\mathbf{W}\hat{\mathbf{H}})\mathbf{w} - 2(\mathbf{W}\hat{\mathbf{H}})^T(\mathbf{W}\mathbf{c}_t) + 2\delta\mathbf{R}_e\mathbf{w} = \mathbf{0}, \tag{5.11}$$

yielding the regularized least-squares reshaping filter

$$\mathbf{w}_{\text{R-LS}} = [(\mathbf{W}\hat{\mathbf{H}})^T(\mathbf{W}\hat{\mathbf{H}}) + \delta\mathbf{R}_e]^{-1}(\mathbf{W}\hat{\mathbf{H}})^T(\mathbf{W}\mathbf{c}_t). \tag{5.12}$$

Since the matrix $\mathbf{R}_e$ is assumed to be a full-rank matrix, the matrix $[(\mathbf{W}\hat{\mathbf{H}})^T(\mathbf{W}\hat{\mathbf{H}}) + \delta\mathbf{R}_e]$ is an invertible matrix. For completeness, Tables 5.1 and 5.2 summarize the regularized least-squares cost functions and reshaping filters for the considered regularized least-squares techniques, i.e., for the different definitions of the weighting matrix $\mathbf{W}$ and the target equalized impulse response $\mathbf{c}_t$ discussed in Section 3.3.

As the regularization parameter δ approaches 0, i.e., disregarding the RIR perturbations, the regularized least-squares reshaping filter in (5.12) is equal to the minimum-norm least-squares reshaping filter in (3.30), i.e.,

$$\lim_{\delta\to 0} \mathbf{w}_{\text{R-LS}} = \mathbf{w}_{\text{LS}}. \tag{5.13}$$

Table 5.1: Regularized least-squares cost function for different regularized least-squares techniques.

Technique	Cost function
R-MINT	$J_{\text{R-M}} = \|\hat{\mathbf{H}}\mathbf{w} - \mathbf{d}\|_2^2 + \delta \mathbf{w}^T \mathbf{R}_e \mathbf{w}$
R-RMCLS	$J_{\text{R-R}} = \|\mathbf{W}_{\text{R}}(\hat{\mathbf{H}}\mathbf{w} - \mathbf{d})\|_2^2 + \delta \mathbf{w}^T \mathbf{R}_e \mathbf{w}$
R-PMINT	$J_{\text{R-P}} = \|\hat{\mathbf{H}}\mathbf{w} - \hat{\mathbf{h}}_{e,p}\|_2^2 + \delta \mathbf{w}^T \mathbf{R}_e \mathbf{w}$

Table 5.2: Regularized least-squares reshaping filter for different regularized least-squares techniques.

Technique	Reshaping filter
R-MINT	$\mathbf{w}_{\text{R-M}} = (\hat{\mathbf{H}}^T\hat{\mathbf{H}} + \delta \mathbf{R}_e)^{-1}\hat{\mathbf{H}}^T\mathbf{d}$
R-RMCLS	$\mathbf{w}_{\text{R-R}} = [(\mathbf{W}_{\text{R}}\hat{\mathbf{H}})^T(\mathbf{W}_{\text{R}}\hat{\mathbf{H}}) + \delta \mathbf{R}_e]^{-1}(\mathbf{W}_{\text{R}}\hat{\mathbf{H}})^T(\mathbf{W}_{\text{R}}\mathbf{d})$
R-PMINT	$\mathbf{w}_{\text{R-P}} = (\hat{\mathbf{H}}^T\hat{\mathbf{H}} + \delta \mathbf{R}_e)^{-1}\hat{\mathbf{H}}^T\hat{\mathbf{h}}_{e,p}$

While the limit in (5.13) is rather intuitive, it is not straightforward to directly deduce it by comparing the regularized least-squares and the least-squares reshaping filters in (5.12) and (3.30), since the matrix $(\mathbf{W}\hat{\mathbf{H}})^T(\mathbf{W}\hat{\mathbf{H}})$ is not invertible for $L_w \geq \left\lceil \frac{L_h - 1}{M - 1} \right\rceil$. In the following, the equality in (5.13) is analytically derived and the presented derivations are further used in Section 5.3 to provide a better understanding of the influence of the regularization parameter in the regularized least-squares techniques.

Consider the joint diagonalization [167] of the positive (semi-)definite symmetric matrices $(\mathbf{W}\hat{\mathbf{H}})^T(\mathbf{W}\hat{\mathbf{H}})$ and $\mathbf{R}_e$, i.e.,

$$\begin{cases} (\mathbf{W}\hat{\mathbf{H}})^T(\mathbf{W}\hat{\mathbf{H}}) = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T, & (5.14) \\ \mathbf{R}_e = \mathbf{Q}\mathbf{\Gamma}\mathbf{Q}^T, & (5.15) \end{cases}$$

with $\mathbf{Q}$ being an $ML_w \times ML_w$–dimensional invertible but not necessarily orthogonal matrix and $\mathbf{\Lambda}$ and $\mathbf{\Gamma}$ being $ML_w \times ML_w$–dimensional diagonal matrices, i.e.,

$$\mathbf{\Lambda} = \text{diag}\{[\lambda(1)\ \lambda(2)\ \ldots\ \lambda(r)\ 0\ \ldots\ 0]\}, \quad (5.16)$$
$$\mathbf{\Gamma} = \text{diag}\{[\gamma(1)\ \gamma(2)\ \ldots\ \gamma(ML_w)]\}, \quad (5.17)$$

where

$$\lambda(1) \geq \lambda(2) \geq \cdots \geq \lambda(r) > 0, \quad (5.18)$$
$$\gamma(1) \geq \gamma(2) \geq \cdots \geq \gamma(ML_w) > 0, \quad (5.19)$$

and

$$r = \text{rank}[(\mathbf{W}\hat{\mathbf{H}})^T(\mathbf{W}\hat{\mathbf{H}})] = \text{rank}(\mathbf{W}\hat{\mathbf{H}}). \tag{5.20}$$

Using (5.14), the matrix $\mathbf{W}\hat{\mathbf{H}}$ can be expressed as

$$\mathbf{W}\hat{\mathbf{H}} = \sqrt{\mathbf{\Lambda}}\mathbf{Q}^T, \tag{5.21}$$

with $\sqrt{\mathbf{\Lambda}} = \text{diag}\{[\sqrt{\lambda(1)}\ \sqrt{\lambda(2)}\ \ldots\ \sqrt{\lambda(r)}\ 0\ \ldots\ 0]\}$. Using (5.14), (5.15), and (5.21), the regularized least-squares reshaping filter in (5.12) can be written as

$$\mathbf{w}_{\text{R-LS}} = (\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T + \delta\mathbf{Q}\mathbf{\Gamma}\mathbf{Q}^T)^{-1}\mathbf{Q}\sqrt{\mathbf{\Lambda}}(\mathbf{W}\mathbf{c}_t) \tag{5.22}$$

$$= [\mathbf{Q}(\mathbf{\Lambda} + \delta\mathbf{\Gamma})\mathbf{Q}^T]^{-1}\mathbf{Q}\sqrt{\mathbf{\Lambda}}(\mathbf{W}\mathbf{c}_t) \tag{5.23}$$

$$= \mathbf{Q}^{-T}(\mathbf{\Lambda} + \delta\mathbf{\Gamma})^{-1}\sqrt{\mathbf{\Lambda}}(\mathbf{W}\mathbf{c}_t). \tag{5.24}$$

Expressing the matrix/vector product in (5.24) as a summation, the regularized least-squares reshaping filter can be further expressed as

$$\mathbf{w}_{\text{R-LS}} = \sum_{i=1}^{r} \frac{\sqrt{\lambda(i)}(\mathbf{W}\mathbf{c}_t)_i}{\lambda(i) + \delta\gamma(i)}\bar{\mathbf{q}}_i, \tag{5.25}$$

where $\bar{\mathbf{q}}_i$ denotes the transpose of the i-th row of $\mathbf{Q}^{-1}$ and $(\mathbf{W}\mathbf{c}_t)_i$ denotes the i-th element of the vector $\mathbf{W}\mathbf{c}_t$.

Similarly, using (5.21), the least-squares reshaping filter in (3.30) can be expressed as

$$\mathbf{w}_{\text{LS}} = (\sqrt{\mathbf{\Lambda}}\mathbf{Q}^T)^+(\mathbf{W}\mathbf{c}_t) \tag{5.26}$$

$$= \mathbf{Q}^{-T}(\sqrt{\mathbf{\Lambda}})^+(\mathbf{W}\mathbf{c}_t) \tag{5.27}$$

$$= \sum_{i=1}^{r} \frac{(\mathbf{W}\mathbf{c}_t)_i}{\sqrt{\lambda(i)}}\bar{\mathbf{q}}_i. \tag{5.28}$$

Comparing (5.25) and (5.28), it is now clear that the regularized least-squares solution is equal to the minimum-norm least-squares solution as the regularization parameter δ approaches 0, i.e.,

$$\lim_{\delta\to 0} \mathbf{w}_{\text{R-LS}} = \sum_{i=1}^{r} \lim_{\delta\to 0} \frac{\sqrt{\lambda(i)}(\mathbf{W}\mathbf{c}_t)_i}{\lambda(i) + \delta\gamma(i)}\bar{\mathbf{q}}_i = \sum_{i=1}^{r} \frac{(\mathbf{W}\mathbf{c}_t)_i}{\sqrt{\lambda(i)}}\bar{\mathbf{q}}_i = \mathbf{w}_{\text{LS}}. \tag{5.29}$$

Therefore as expected, when disregarding the RIR perturbations by using a small value for the regularization parameter δ, the regularized least-squares techniques result in a similar performance as the least-squares techniques, i.e., they typically fail to achieve dereverberation.

Regularized channel shortening reshaping filter

Unfortunately, no analytical solution minimizing the regularized channel shortening cost function in (5.10) is available. Hence, we have resorted to an iterative optimization procedure to minimize this non-linear cost function, for which we have used the

MATLAB function *fminunc* [168]. In order to improve the numerical robustness and the convergence speed of the optimization procedure, the gradient and the Hessian of the regularized channel shortening cost function, i.e.,

$$\frac{\partial J_{\text{R-CS}}}{\partial \mathbf{w}} = 2\left[\frac{(\mathbf{w}^T\hat{\mathbf{D}}\mathbf{w})\hat{\mathbf{U}}\mathbf{w} - (\mathbf{w}^T\hat{\mathbf{U}}\mathbf{w})\hat{\mathbf{D}}\mathbf{w}}{(\mathbf{w}^T\hat{\mathbf{D}}\mathbf{w})^2} + \delta\mathbf{R}_e\mathbf{w}\right], \tag{5.30}$$

$$\begin{aligned}\frac{\partial^2 J_{\text{R-CS}}}{\partial \mathbf{w}^2} = 2\Bigg[&\frac{(\mathbf{w}^T\hat{\mathbf{D}}\mathbf{w})\hat{\mathbf{U}} - (\mathbf{w}^T\hat{\mathbf{U}}\mathbf{w})\hat{\mathbf{D}} + 2(\hat{\mathbf{U}}\mathbf{w}\mathbf{w}^T\hat{\mathbf{D}} - \hat{\mathbf{D}}\mathbf{w}\mathbf{w}^T\hat{\mathbf{U}})}{(\mathbf{w}^T\hat{\mathbf{D}}\mathbf{w})^2} \\ &- 4\frac{[(\mathbf{w}^T\hat{\mathbf{D}}\mathbf{w})\hat{\mathbf{U}}\mathbf{w} - (\mathbf{w}^T\hat{\mathbf{U}}\mathbf{w})\hat{\mathbf{D}}\mathbf{w}]\mathbf{w}^T\hat{\mathbf{D}}}{(\mathbf{w}^T\hat{\mathbf{D}}\mathbf{w})^3} + \delta\mathbf{R}_e\Bigg],\end{aligned} \tag{5.31}$$

can be provided. There is however no guarantee that the optimization procedure will converge to the global minimum of the regularized channel shortening cost function instead of a local minimum. Nevertheless, the simulation results in Section 5.5 show that also when using a numerical optimization procedure to compute the regularized channel shortening reshaping filter, a significant increase in robustness in the presence of RIR perturbations can be obtained.

5.3 Automatic procedure for determining the regularization parameter in regularized least-squares techniques

Increasing the regularization parameter δ in the regularized least-squares cost function in (5.7) or the regularized channel shortening cost function in (5.10) on the one hand is supposed to yield a lower distortion energy ϵ_e, but on the other hand is supposed to yield a higher dereverberation error energy ϵ_c or ϵ_r, i.e., a larger deviation between the (weighted) *target* equalized impulse response and the (weighted) *estimated* equalized impulse response. The actual dereverberation performance, i.e., the deviation of the (weighted) *target* equalized impulse response from the (weighted) *true* equalized impulse response, clearly depends on both terms (cf. (5.3)), i.e., ideally both terms should be equal to zero. Hence, due to the arising trade-off between the dereverberation error energy and the distortion energy, the use of an optimal regularization parameter is important. However, the optimal value of the regularization parameter δ yielding the best performance depends on many factors such as the acoustic system, the RIR perturbations, and the used equalization technique. While in simulations the optimal regularization parameter can be intrusively determined, i.e., exploiting the known true RIRs (cf. Section 5.5.1), an automatic non-intrusive procedure is required in practice. In this section, such an automatic procedure for determining the regularization parameter in regularized least-squares techniques is proposed.

The dereverberation error energy ϵ_c and the distortion energy ϵ_e for the regularized least-squares reshaping filter are given by

$$\epsilon_c = \|\mathbf{W}(\hat{\mathbf{H}}\mathbf{w}_{\text{R-LS}} - \mathbf{c}_t)\|_2^2 \tag{5.32}$$

$$= \mathbf{w}_{\text{R-LS}}^T(\mathbf{W}\hat{\mathbf{H}})^T(\mathbf{W}\hat{\mathbf{H}})\mathbf{w}_{\text{R-LS}} - 2\mathbf{w}_{\text{R-LS}}^T(\mathbf{W}\hat{\mathbf{H}})^T(\mathbf{W}\mathbf{c}_t) + (\mathbf{W}\mathbf{c}_t)^T(\mathbf{W}\mathbf{c}_t), \tag{5.33}$$

and

$$\epsilon_e = \mathbf{w}_{\text{R-LS}}^T \mathbf{R}_e \mathbf{w}_{\text{R-LS}}. \tag{5.34}$$

Using the joint diagonalization from (5.14) and the regularized least-squares reshaping filter from (5.24), the first term of the dereverberation error energy in (5.33) can be expressed as

$$\mathbf{w}_{\text{R-LS}}^T (\mathbf{W}\hat{\mathbf{H}})^T (\mathbf{W}\hat{\mathbf{H}}) \mathbf{w}_{\text{R-LS}} = (\mathbf{W}\mathbf{c}_t)^T \mathbf{\Lambda}^2 (\mathbf{\Lambda} + \delta\mathbf{\Gamma})^{-2} (\mathbf{W}\mathbf{c}_t). \tag{5.35}$$

Similarly, using (5.21) and (5.24), the second term of the dereverberation error energy in (5.33) can be expressed as

$$-2\mathbf{w}_{\text{R-LS}}^T (\mathbf{W}\hat{\mathbf{H}})^T (\mathbf{W}\mathbf{c}_t) = -2(\mathbf{W}\mathbf{c}_t)^T \mathbf{\Lambda} (\mathbf{\Lambda} + \delta\mathbf{\Gamma})^{-1} (\mathbf{W}\mathbf{c}_t). \tag{5.36}$$

Substituting (5.35) and (5.36) in (5.33), the dereverberation error energy can be written as

$$\epsilon_c = (\mathbf{W}\mathbf{c}_t)^T [\mathbf{\Lambda}^2 (\mathbf{\Lambda} + \delta\mathbf{\Gamma})^{-2} - 2\mathbf{\Lambda} (\mathbf{\Lambda} + \delta\mathbf{\Gamma})^{-1} + \mathbf{I}] (\mathbf{W}\mathbf{c}_t) \tag{5.37}$$

$$= \sum_{i=1}^{r} \frac{\delta^2 \gamma^2(i) (\mathbf{W}\mathbf{c}_t)_i^2}{[\lambda(i) + \delta\gamma^2(i)]^2}. \tag{5.38}$$

Similarly, using the joint diagonalization in (5.15) and the regularized least-squares reshaping filter in (5.24), the distortion energy can be expressed as

$$\epsilon_e = (\mathbf{W}\mathbf{c}_t)^T \mathbf{\Lambda}^2 \mathbf{\Gamma} (\mathbf{\Lambda} + \delta\mathbf{\Gamma})^{-1} (\mathbf{W}\mathbf{c}_t) \tag{5.39}$$

$$= \sum_{i=1}^{r} \frac{\lambda^2(i) \gamma(i) (\mathbf{W}\mathbf{c}_t)_i^2}{\lambda(i) + \delta\gamma(i)}. \tag{5.40}$$

It is now straightforward to see that increasing the regularization parameter δ increases the dereverberation error energy ϵ_c in (5.38) but decreases the distortion energy ϵ_e in (5.40). Furthermore, as expected it can be observed that as the regularization parameter approaches 0, the dereverberation error energy in (5.38) also approaches 0 whereas the distortion energy is equal to $\sum_{i=1}^{r} \lambda(i)\gamma(i)(\mathbf{W}\mathbf{c}_t)_i$. An appropriate regularization parameter should hence incorporate knowledge about both ϵ_c and ϵ_e, such that both terms are small.

In order to automatically compute a regularization parameter for regularized least-squares problems, it has been proposed in [169,170] to use a parametric plot of the trade-off quantities for several values of the regularization parameter. Because of the arising trade-off, this plot has an L-shape with the corner (i.e., the point of maximum curvature) located where the regularized least-squares solution changes from being dominated by over-regularization to being dominated by under-regularization. We therefore propose to automatically determine the regularization parameter in regularized least-squares techniques as the one maximizing the curvature of the parametric plot of the distortion energy ϵ_e versus the dereverberation error energy ϵ_c. As is experimentally validated in Section 5.5, such a regularization parameter also leads to a nearly-optimal performance compared to an intrusively determined regularization parameter.

Fig. 5.1 depicts an exemplary L-curve obtained using the regularized PMINT technique with regularization parameter values ranging from 10^{-9} to 10^{-1}. The used acoustic system is the same as the one described in Section 3.4.1, with $M = 4$ microphones and $L_h = 3600$. As illustrated in this figure, increasing the value of the regularization parameter δ decreases the distortion energy ϵ_e but increases the dereverberation error energy ϵ_c. At the point of maximum curvature, i.e., $\delta = 10^{-5}$ in the depicted example, both ϵ_c and ϵ_e are small.

The curvature κ of the parametric plot of the distortion energy versus the dereverberation error energy can be analytically computed as [171]

$$\kappa = \frac{\epsilon_c' \epsilon_e'' - \epsilon_c'' \epsilon_e'}{(\epsilon_c' + \epsilon_e')^{\frac{3}{2}}}, \tag{5.41}$$

with $\{\cdot\}'$ and $\{\cdot\}''$ denoting the first- and second-order derivatives with respect to δ. The computation of the first- and second-order derivatives using the expressions for ϵ_c and ϵ_e in (5.38) and (5.40) yields

$$\epsilon_c' = \sum_{i=1}^{r} \frac{2\delta\lambda(i)\gamma^2(i)(\mathbf{W}\mathbf{c}_t)_i^2}{[\lambda(i) + \delta\gamma(i)]^3}, \tag{5.42}$$

$$\epsilon_c'' = \sum_{i=1}^{r} \frac{2\lambda^2(i)\gamma^2(i)(\mathbf{W}\mathbf{c}_t)_i^2}{[\lambda(i) + \delta\gamma(i)]^4}, \tag{5.43}$$

$$\epsilon_e' = \sum_{i=1}^{r} \frac{-\lambda^2(i)\gamma^2(i)(\mathbf{W}\mathbf{c}_t)_i^2}{[\lambda(i) + \delta\gamma(i)]^2}, \tag{5.44}$$

$$\epsilon_e'' = \sum_{i=1}^{r} \frac{2\lambda^2(i)\gamma^3(i)(\mathbf{W}\mathbf{c}_t)_i^2}{[\lambda(i) + \delta\gamma^3(i)]^3}. \tag{5.45}$$

Substituting these first- and second-order derivatives in the analytical curvature expression in (5.41), the curvature κ can be analytically expressed as a function of the regularization parameter δ. Hence, a one-dimensional iterative optimization

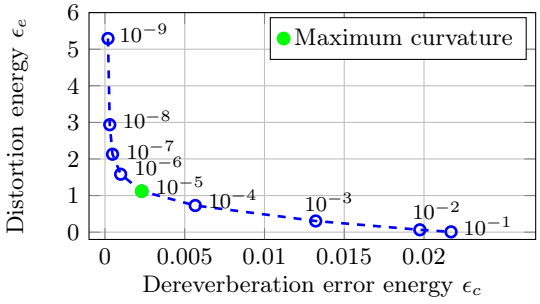


Fig. 5.1: Exemplary L-curve obtained using the regularized PMINT technique with the regularization parameter values ranging from 10^{-9} to 10^{-1}. The considered acoustic system is the same as in Section 3.4.1

procedure can then be used to determine the parameter that maximizes the curvature. However, as can be seen from the presented first- and second-order derivatives in (5.42) to (5.45), analytically computing the curvature involves the computation of the joint diagonalization of the (typically large-dimensional) matrices $(\mathbf{W}\hat{\mathbf{H}})^T(\mathbf{W}\hat{\mathbf{H}})$ and $\mathbf{R}_e$, and the manipulation of the diagonal elements $\lambda(i)$ and $\gamma(i)$. Using an iterative optimization procedure to maximize the curvature not only results in a high computational complexity, but is also prone to numerical errors. Unfortunately, standard numerical procedures to compute the diagonal elements $\lambda(i)$ and $\gamma(i)$, particularly the ones close to 0, do not seem to exhibit sufficient numerical accuracy.

Therefore, in this work the triangle method, which is a numerically robust geometric procedure, is used to compute the point of maximum curvature [172]. In this method the regularized least-squares filter is computed for a discrete set of regularization parameters δ and the discrete L-curve is generated. Once the discrete L-curve is generated, the points corresponding to the maximum distortion energy and the maximum dereverberation error energy (i.e., $\delta = 10^{-9}$ and $\delta = 10^{-1}$ in the depicted exemplary L-curve in Fig. 5.1) are considered to be fixed vertexes of a triangle. Different triangles are then formed with the third vertex being any of the remaining points on the discrete L-curve. For each formed triangle it is first determined whether this part of the curve has an L-shape or an inverted L-shape. If an L-shape is found, the angle at the corresponding vertex is computed. If this angle is sufficiently small (to guarantee that the curve is sufficiently sharp at that point), then the vertex is a candidate to be the point of maximum curvature, otherwise it is disregarded. Out of all possible candidate points, the vertex yielding the smallest angle is selected as the point of maximum curvature (i.e., $\delta = 10^{-5}$ in the depicted exemplary L-curve in Fig. 5.1).

5.4 Non-applicability of the automatic procedure to the regularized channel shortening technique

The automatic procedure for determining the regularization parameter proposed in Section 5.3 implicitly relies on the fact that the dereverberation error energy and the distortion energy are strictly monotonically increasing and strictly monotonically decreasing functions of the regularization parameter δ. Only in this case it is meaningful to discuss the arising trade-off and to compute the automatic regularization parameter as the point of maximum curvature of the parametric plot of the distortion energy versus the dereverberation error energy. For regularized least-squares techniques, it can be shown that the dereverberation error energy ϵ_c and the distortion energy ϵ_e are strictly monotonically increasing and strictly monotonically decreasing functions of the regularization parameter δ, since the first-order derivatives derived in (5.42) and (5.44) are positive and negative, i.e.,

$$\epsilon_c' > 0, \tag{5.46}$$

$$\epsilon_e' < 0. \tag{5.47}$$

Since for the regularized channel shortening technique no analytical solution can be found, and furthermore, there is no guarantee that the iterative optimization procedure converges to the global minimum instead of a local one, it cannot be proven that the iterative optimization procedure yields dereverberation error and distortion energies ϵ_r and ϵ_e that are strictly monotonically increasing and decreasing functions of the regularization parameter δ. Hence, an automatic procedure for determining the regularization parameter in the regularized channel shortening technique remains a topic for future investigation.

5.5 Simulations

In this section, we investigate the dereverberation performance of all considered equalization techniques when incorporating regularization. In Section 5.5.1 the considered acoustic system and the used algorithmic settings are introduced. Section 5.5.2 investigates the robustness increase of the considered acoustic multi-channel equalization techniques when incorporating an optimal intrusively determined regularization parameter. In Section 5.5.3 the performance of the intrusively regularized acoustic multi-channel equalization techniques is extensively compared. Finally, in Section 5.5.4 the performance of the automatic non-intrusively regularized PMINT technique is compared to the performance of the optimal intrusively regularized counterpart.

5.5.1 *Acoustic system and algorithmic settings*

We have considered the same acoustic system as in Section 3.4.1, i.e., a single speech source and $M = 4$ omni-directional microphones. The source-microphone distance is 3 m and the distance between the microphones is 5 cm. Room impulse responses from the MARDY database [161] have been used, where the room reverberation time is $T_{60} \approx 450$ ms and the direct-to-reverberant ratio is DRR $= 0$ dB. The RIRs have been measured using the swept-sine technique [162] and the length of the RIRs has been set to $L_h = 3600$ at a sampling frequency $f_s = 8$ kHz.

Similarly as in Section 3.4, in order to simulate RIR perturbations, the measured RIRs are perturbed by adding scaled white noise as described in Section 2.2. The considered normalized projection misalignment (NPM) values between the true and the perturbed RIRs are (cf. (2.52))

$$\text{NPM} \in \{-33 \text{ dB}, -27 \text{ dB}, -21 \text{ dB}, -15 \text{ dB}\}. \tag{5.48}$$

For all considered techniques the reshaping filter length is $L_w = \left\lceil \frac{L_h - 1}{M - 1} \right\rceil = 1200$, the delay is set to $\tau = 90$, and the performance for several desired window lengths L_d ranging from 10 ms to 50 ms is investigated, i.e.,

$$L_d \in \{10 \text{ ms}, 20 \text{ ms}, 30 \text{ ms}, 40 \text{ ms}, 50 \text{ ms}\}. \tag{5.49}$$

The target equalized impulse response for the PMINT and regularized PMINT techniques is set to the direct path and early reflections of the perturbed RIR of

the first microphone, i.e., $\hat{\mathbf{h}}_{e,1}$. Furthermore, the channel shortening reshaping filter is selected as the generalized eigenvector yielding the minimum l_2-norm estimated equalized impulse response as proposed in [125].

Using the instrumental performance measures described in Section 2.3, the dereverberation performance is evaluated in terms of the reverberant energy suppression and the perceptual speech quality improvement. The reverberant energy suppression is evaluated using the direct-to-reverberant ratio improvement (ΔDRR) between the equalized impulse response $\mathbf{c}$ and the true RIR $\mathbf{h}_1$ (cf. (2.53)), as well as the energy decay curve (EDC) of the equalized impulse response $\mathbf{c}$ (cf. (2.55)). The improvement in perceptual speech quality is evaluated using the improvement in PESQ [153] (ΔPESQ) and in cepstral distance [154] (ΔCD) between the output speech signal $z(n)$ and the reverberant microphone signal $x_1(n)$. The reference signal employed for the PESQ and cepstral distance measures is $x_{e,1}(n) = s(n) * h_{e,1}(n)$, i.e., the clean speech signal convolved with the direct path and early reflections of the first RIR (which changes as the desired window length L_d changes).

In order to evaluate the effectiveness of incorporating regularization in the considered equalization techniques, we investigate the performance for several regularization parameters δ, i.e.,

$$\delta = \{10^{-7},\ 10^{-6},\ \ldots,\ 10^{-1},\ 1,\ 3,\ 5,\ 7,\ 10\}. \tag{5.50}$$

For the perturbation matrix in (5.5) we have assumed that $\mathbf{R}_e = \mathbf{I}$. Similarly as in Section 4.3 for determining the optimal reshaping filter length, the optimal regularization parameter δ_{o} has been determined as the parameter yielding the highest perceptual speech quality in terms of the PESQ score. It should be noted that the computation of the PESQ score for determining the optimal regularization parameter is an intrusive procedure which is not applicable in practice, since knowledge of the clean speech signal and of the true RIRs is required to compute the reference signal and the equalized impulse response $\mathbf{c} = \mathbf{H}\mathbf{w}$. In Section 5.5.4, the performance when using the automatic non-intrusive procedure for determining the regularization parameter proposed in Section 5.3 will be investigated. In addition, it should be noted that although the MINT technique is independent of the desired window length L_d, we determine an optimal regularization parameter in the MINT technique for each desired window length L_d by changing the reference signal in the PESQ computation, such that the MINT technique can be compared to partial equalization techniques (which are dependent on the desired window length L_d).

Since for the regularized channel shortening technique an iterative optimization procedure should be used, a termination criterion needs to be imposed. In our implementation, the termination criterion is either the number of iterations exceeding 100 or the relative change in the solution norm dropping below 10^{-5}. Furthermore, since it cannot be guaranteed that the iterative optimization procedure proposed in Section 5.2 converges to the global minimum of the regularized channel shortening cost function, the initialization of this procedure may influence the resulting reshaping filter. We have hence investigated three different initializations $\mathbf{w}_{\text{init}}$, i.e.,

i) $\mathbf{w}_{\text{init}} = [1\ 0\ \ldots\ 0]^T$, i.e., the filter yielding the first microphone signal,
ii) $\mathbf{w}_{\text{init}}$ is randomly initialized with normally distributed coefficients,
iii) $\mathbf{w}_{\text{init}} = \mathbf{w}_{\text{CS}}$, i.e., the channel shortening reshaping filter ($\delta = 0$) is used to initialize the iterative optimization procedure.

In all simulations we observed a significant difference in performance for the different initializations of the iterative optimization procedure. Therefore, it appears that the iterative optimization procedure for computing the regularized channel shortening reshaping filter typically converges to local minima. However, using the first initialization, i.e., $\mathbf{w}_{\text{init}} = [1\ 0\ \ldots\ 0]^T$, always resulted in the highest performance, hence, the following simulation results are generated using this initialization.

5.5.2 *Robustness increase of acoustic multi-channel equalization when incorporating regularization*

In this section, the performance of all considered (non-regularized) acoustic multi-channel equalization techniques is compared to the performance of their regularized extensions using the intrusively determined regularization parameter δ_{o}. We consider an exemplary scenario with NPM $= -33$ dB.

Robustness increase of the MINT technique

Fig. 5.2 depicts the performance of the MINT and the optimally regularized MINT techniques in terms of ΔDRR, EDC, ΔPESQ, and ΔCD. For completeness, the used optimal regularization parameters are presented in Table 5.3.

As shown by the ΔDRR values depicted in Fig. 5.2a, incorporating regularization in the MINT technique significantly improves the reverberant energy suppression. It can be observed that when the desired window length is between 10 ms and 40 ms, the DRR improves by approximately 15 dB in comparison to the true RIR $\mathbf{h}_1$, whereas when the desired window length is 50 ms, the DRR improves by approximately 5 dB. Although the regularized MINT technique is independent of the desired window length, this difference arises because of intrusively determining the regularization parameter as the one maximizing the PESQ score, with the reference signal $x_{e,1}(n)$ being different for each desired window length L_d. As the desired window length changes, also the reference signal for determining the optimal regularization parameter changes, which can result in a different regularization parameter for the regularized MINT technique, and hence, in a different reshaping filter.

To evaluate the decay rate of the reverberant energy, Fig. 5.2b depicts the energy decay curve of the true RIR $\mathbf{h}_1$ and the energy decay curve of the equalized impulse response $\mathbf{c}$ obtained using the MINT and the optimally regularized MINT techniques for the desired window length $L_d = 50$ ms. While as expected the MINT technique fails to achieve dereverberation and results in a slower decay rate of the reverberant energy than for the true RIR $\mathbf{h}_1$, the optimally regularized MINT technique yields a better performance, resulting in a slight improvement in comparison to $\mathbf{h}_1$.

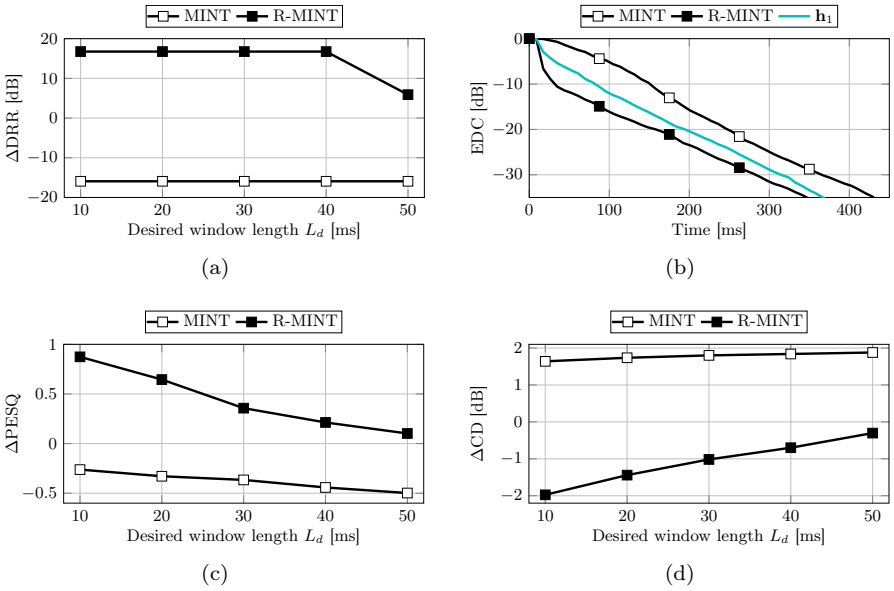


Fig. 5.2: Performance of the MINT and the optimally regularized MINT techniques in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD (NPM = −33 dB).

Table 5.3: Optimal regularization parameter for the regularized MINT technique for several desired window lengths (NPM = −33 dB).

Desired window length L_d [ms]	10	20	30	40	50
Optimal regularization parameter δ_o	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-1}

The improvement in direct-to-reverberant ratio and reverberant energy decay rate achieved by the optimally regularized MINT technique is also reflected in the significant overall perceptual speech quality improvement, as measured by the ΔPESQ and ΔCD values presented in Figs. 5.2c and 5.2d. It can be observed that while the MINT technique worsens the PESQ score and the cepstral distance in comparison to the reverberant microphone signal $x_1(n)$, the regularized MINT technique achieves a significantly better performance.

In summary, incorporating regularization in the MINT technique yields a significantly higher robustness against RIR perturbations, confirming the observations in [108]. Nevertheless, acoustic system inversion using the regularized MINT technique remains quite sensitive to RIR perturbations and does not yield a satisfactory performance in terms of the decay rate of the reverberant energy (as illustrated in Fig. 5.2b).

Robustness increase of the CS technique

Fig. 5.3 depicts the performance of the CS and the optimally regularized CS techniques in terms of ΔDRR, EDC, ΔPESQ, and ΔCD. The used optimal regularization parameters are presented in Table 5.4.

As illustrated in Fig. 5.3a, the optimally regularized CS technique achieves a significantly higher ΔDRR than the CS technique, particularly for short desired window lengths. As the desired window length increases, the ΔDRR obtained using the regularized CS technique decreases, since additional energy is introduced which is accounted for as reverberant energy in the DRR computation (cf. (2.54)).

To evaluate the decay rate of the reverberant energy, Fig. 5.3b depicts the energy decay curve of the true RIR $\mathbf{h}_1$ and the energy decay curve of the equalized impulse response $\mathbf{c}$ obtained using the CS and the optimally regularized CS techniques for

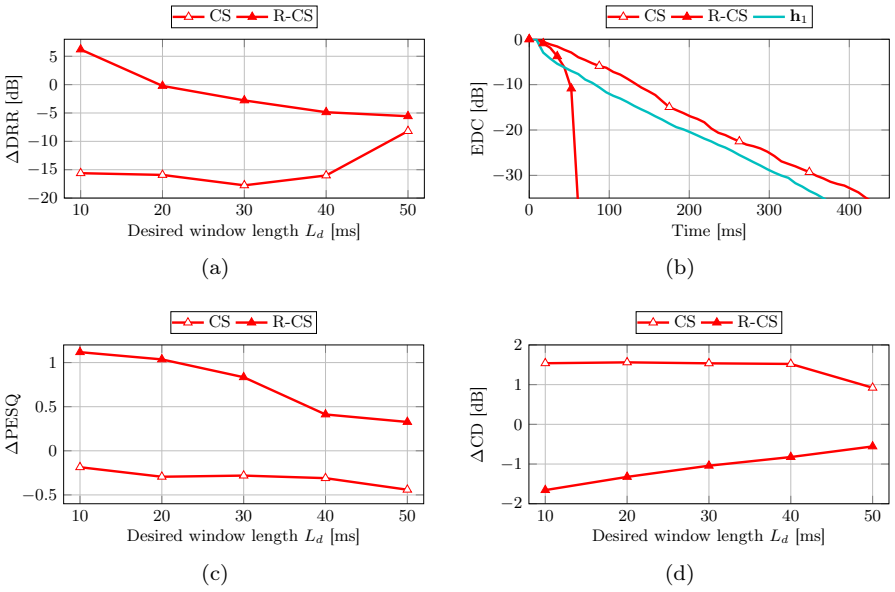


Fig. 5.3: Performance of the CS and the optimally regularized CS techniques in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD (NPM = −33 dB).

Table 5.4: Optimal regularization parameter for the regularized CS technique for several desired window lengths (NPM = −33 dB).

Desired window length L_d [ms]	10	20	30	40	50
Optimal regularization parameter δ_{o}	10^{-1}	10^{-4}	10^{-4}	10^{-1}	10^{-2}

the desired window length $L_d = 50$ ms. It can be observed that while the CS technique fails to achieve dereverberation, the regularized CS technique results in a significantly better performance, with the reverberant energy decaying at a much faster rate than the reverberant energy in the true RIR $\mathbf{h}_1$.

The improvement in direct-to-reverberant ratio and reverberant energy decay rate achieved by the optimally regularized CS technique is also reflected in the overall perceptual speech quality improvement as measured by the ΔPESQ and ΔCD values presented in Figs. 5.3c and 5.3d. It can be observed that the CS technique fails to improve the perceptual speech quality for all desired window lengths, whereas the regularized CS technique yields a significant improvement in comparison to the reverberant microphone signal $x_1(n)$. As expected, the ΔPESQ values for the regularized CS technique decrease with increasing desired window length whereas the ΔCD values increase, since more early reflections are left uncontrolled.

In summary, incorporating regularization in the CS technique significantly increases the robustness in the presence of RIR perturbations, resulting in a high reverberant energy suppression and perceptual speech quality improvement.

Robustness increase of the RMCLS technique

Fig. 5.4 depicts the performance of the RMCLS and the optimally regularized RMCLS techniques in terms of ΔDRR, EDC, ΔPESQ, and ΔCD. The used optimal regularization parameters are presented in Table 5.5.

Since the RMCLS technique is significantly more robust than the MINT and CS techniques, it can be observed in Fig. 5.4a that the DRR improvement obtained when incorporating regularization in the RMCLS technique is less than the improvement obtained when incorporating regularization in the MINT or CS techniques. Nevertheless, an improvement in ΔDRR of up to 8 dB is obtained when regularization is incorporated in the RMCLS technique. Most importantly, the optimally regularized RMCLS technique improves the DRR in comparison to the true RIR $\mathbf{h}_1$ for all considered desired window lengths, which is not the case for the RMCLS technique.

To evaluate the decay rate of the reverberant energy, Fig. 5.4b depicts the energy decay curve of the true RIR $\mathbf{h}_1$ and the energy decay curve of the equalized impulse response $\mathbf{c}$ obtained using the RMCLS and the optimally regularized RMCLS techniques for the desired window length $L_d = 50$ ms. It can be observed that the optimally regularized RMCLS technique yields a significantly faster decay rate of the reverberant energy than in the true RIR $\mathbf{h}_1$, with the performance being similar to the performance of the RMCLS technique.

Furthermore, as depicted in Figs. 5.4c and 5.4d, the regularized RMCLS technique yields a significantly better perceptual speech quality than the RMCLS technique, improving the obtained ΔPESQ and ΔCD values for all desired window lengths.

In summary, incorporating regularization in the RMCLS technique further increases the robustness in the presence of RIR perturbations, resulting in a high reverberant energy suppression and a particularly significant improvement in perceptual speech quality.

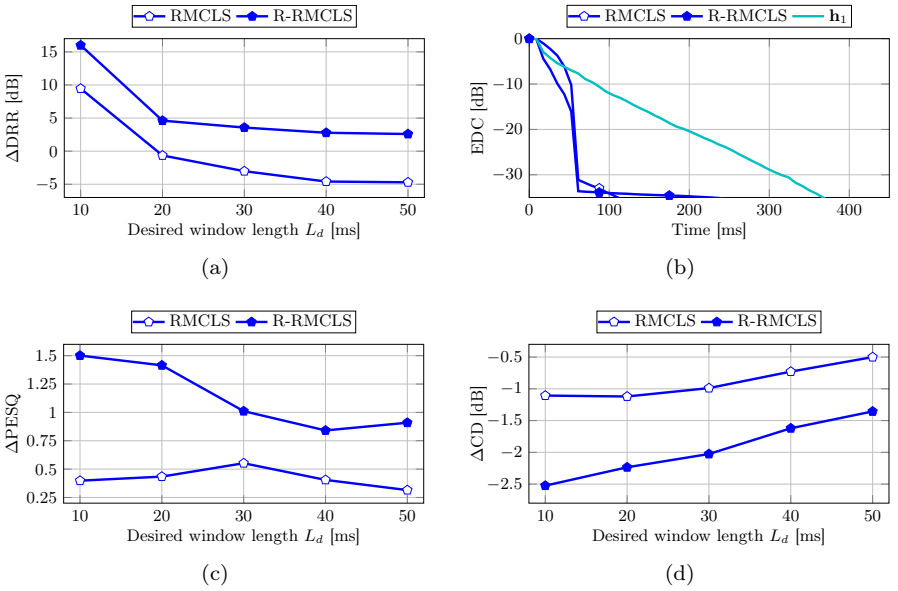


Fig. 5.4: Performance of the RMCLS and the optimally regularized RMCLS techniques in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD (NPM = −33 dB).

Table 5.5: Optimal regularization parameter for the regularized RMCLS technique for several desired window lengths (NPM = −33 dB).

Desired window length L_d [ms]	10	20	30	40	50
Optimal regularization parameter δ_o	10^{-5}	10^{-5}	10^{-5}	10^{-4}	10^{-4}

Robustness increase of the PMINT technique

Fig. 5.5 depicts the performance of the PMINT and the optimally regularized PMINT techniques in terms of ΔDRR, EDC, ΔPESQ, and ΔCD. The used optimal regularization parameters are presented in Table 5.6.

Similarly as for the MINT and CS techniques, it can be observed in Figs. 5.5a and 5.5b that incorporating regularization in the PMINT technique results in a significantly higher DRR and a faster reverberant energy decay rate. Fig. 5.5a shows that while the PMINT technique worsens the DRR in comparison to the true RIR $\mathbf{h}_1$, the optimally regularized PMINT technique yields a high improvement, in particular for short desired window lengths. As the desired window length increases, additional energy is introduced in the equalized impulse response, which is accounted for as reverberant energy in the DRR computation (cf. (2.54)). Furthermore, Fig. 5.5b

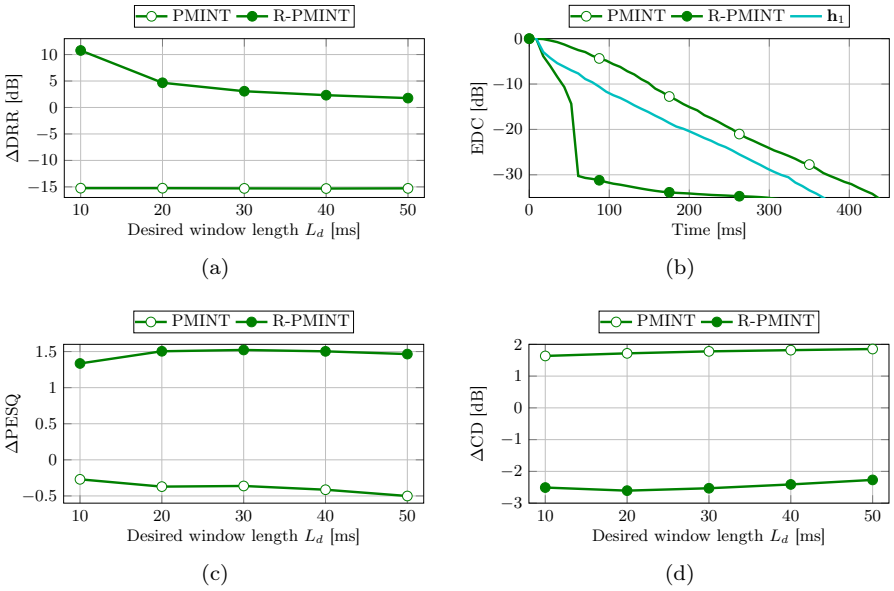


Fig. 5.5: Performance of the PMINT and the optimally regularized PMINT techniques in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD (NPM = −33 dB).

Table 5.6: Optimal regularization parameter for the regularized PMINT technique for several desired window lengths (NPM = −33 dB).

Desired window length L_d [ms]	10	20	30	40	50
Optimal regularization parameter δ_o	10^{-5}	10^{-5}	10^{-5}	10^{-5}	10^{-5}

shows that while the PMINT technique results in a slower decay rate of the reverberant energy than in the true RIR $\mathbf{h}_1$, the regularized PMINT technique yields a significantly better performance.

Similarly as for the other equalization techniques, incorporating regularization in the PMINT technique improves the perceptual speech quality, as shown by the ΔPESQ and ΔCD values depicted in Figs. 5.5c and 5.5d. It can be observed that while the PMINT technique fails to improve the perceptual speech quality in comparison to the reverberant microphone signal $x_1(n)$, the optimally regularized PMINT technique yields a significantly better performance, with an improvement of approximately 1.5 in PESQ and 2.5 dB in cepstral distance for all considered desired window lengths.

In summary, incorporating regularization in the PMINT technique results in a significant increase in robustness against RIR perturbations, both in terms of reverberant energy suppression and perceptual speech quality improvement.

5.5.3 *Comparison of robust acoustic multi-channel equalization techniques*

The simulation results in Section 5.5.2 have shown that incorporating regularization in all considered acoustic multi-channel equalization techniques yields a significant increase in robustness in the presence of RIR perturbations. In this section, the performance of the optimally regularized equalization techniques is extensively compared for the NPM values in (5.48). Similarly as in Section 3.4.3, the presented performance measures are averaged over all considered NPM values.

To compare the reverberant energy suppression, Figs. 5.6a and 5.6b depict the DRR improvement and the energy decay curve obtained by the regularized techniques. It can be observed that the optimally regularized MINT technique typically achieves the highest DRR improvement (only not for $L_d = 10$ ms). This is to be expected since the regularized MINT technique aims at optimizing the DRR by using a delayed impulse as the target equalized impulse response. However, since inverting an acoustic system is not as robust as partially reshaping it (cf. Section 3.4.3), the decay rate of the reverberant energy for the optimally regularized MINT tech-

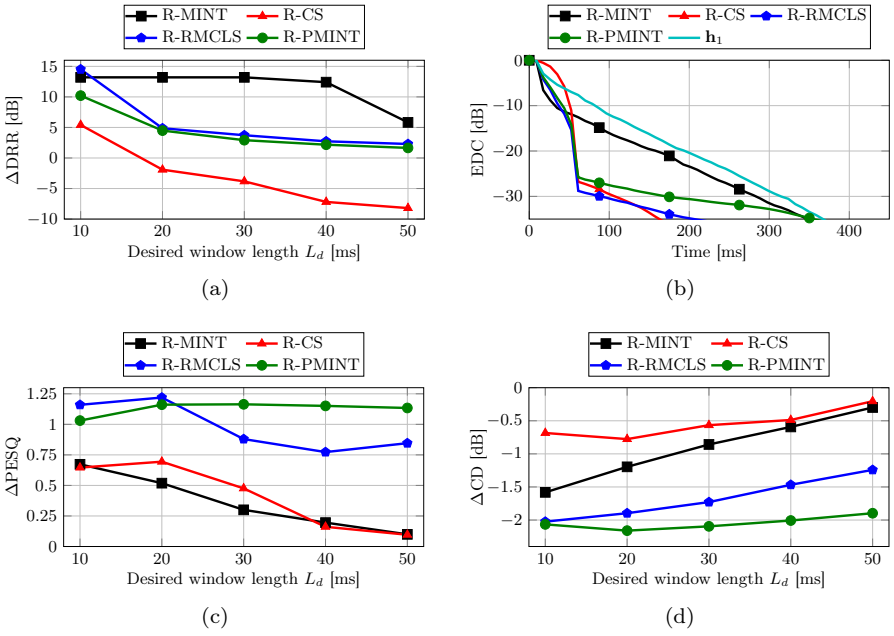


Fig. 5.6: Performance of the optimally regularized MINT, CS, RMCLS, and PMINT techniques in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD (averaged over several NPM values).

nique is significantly slower than for the optimally regularized partial multi-channel equalization techniques. Among the regularized partial multi-channel equalization techniques, the regularized CS technique yields the worst DRR improvement but the fastest reverberant energy decay rate. On the other hand, the regularized RMCLS and PMINT techniques result in a similarly high performance in terms of both measures, with the regularized RMCLS technique yielding a slightly faster reverberant energy decay rate.

To compare the overall perceptual speech quality, Figs. 5.6c and 5.6d depict the PESQ score and cepstral distance improvement obtained by the considered optimally regularized techniques. It is important to note that all regularized techniques improve the perceptual speech quality for all desired window lengths when compared to the reverberant microphone signal $x_1(n)$. It can be observed that the regularized MINT and CS techniques result in the lowest PESQ score and cepstral distance improvement. On the other hand, the regularized RMCLS and regularized PMINT techniques result in a significantly better perceptual speech quality improvement. It can be observed that for the short desired window lengths of 10 ms and 20 ms, the perceptual speech quality improvement obtained by both techniques is similar. However, as the desired window length increases, the regularized PMINT technique yields a higher perceptual speech quality, since more early reflections are left uncontrolled in the regularized RMCLS technique.

In summary, these simulation results demonstrate that the optimally regularized PMINT technique is a robust and perceptually advantageous equalization technique, typically outperforming the other considered techniques in terms of perceptual speech quality. The high performance improvement obtained for the PMINT technique when regularization is incorporated can be explained by the significantly larger reverberant tail suppression. The remaining advantage lies in the direct control of the early reflections.

5.5.4 *Automatic regularization parameter in the regularized PMINT technique*

The simulation results in Section 5.5.3 have shown that the regularized PMINT technique yields a high dereverberation performance in the presence of RIR perturbations. However, the regularization parameter has been determined intrusively exploiting the known clean speech signal and the known true RIRs, which is inapplicable in practice. In this section we investigate the performance difference for the regularized PMINT technique when using the non-intrusive and practically applicable procedure for determining the regularization parameter δ_{a} proposed in Section 5.3 instead of the optimal intrusively determined regularization parameter δ_{o}. Similarly as before, the NPM values in (5.48) are considered and the presented performance measures are averaged over all considered NPM values.

For the automatic procedure, the regularized PMINT reshaping filter is computed for the set of regularization parameters in (5.50). The dereverberation error energy and the distortion energy for each regularization parameter are then computed using (5.33) and (5.34). As described in Section 5.3 the discrete L-curve is constructed

and the regularization parameter δ_{a} corresponding to the point of maximum curvature is determined using the triangle method [172].

To compare the reverberant energy suppression, Figs. 5.7a and 5.7b depict the DRR improvement and the energy decay curve obtained by the regularized PMINT technique with regularization parameters δ_{o} and δ_{a}. As illustrated, hardly any difference can be observed in these performance measures when using the optimal or the automatic regularization parameter. As shown in Fig. 5.7a, the automatically determined and practically applicable regularization parameter δ_{a} yields a high improvement in direct-to-reverberant ratio, particularly for short desired window lengths. Furthermore, as shown in Fig. 5.7b, this parameter also results in a significantly faster decay rate of the reverberant energy than in the true RIR $\mathbf{h}_1$.

To compare the perceptual speech quality, Figs. 5.7c and 5.7d depict the PESQ score and cepstral distance improvement obtained by the regularized PMINT technique with regularization parameters δ_{o} and δ_{a}. It can be observed that the automatically determined regularization parameter δ_{a} yields a very similar perceptual speech quality improvement as using the intrusively determined regularization parameter δ_{o} for longer desired window lengths. For short desired window lengths, using δ_{a} results in a loss of approximately less than 0.5 in PESQ score and approximately less than 0.5 dB in cepstral distance. It should be noted that the proposed automatic procedure for determining the regularization parameter is solely based on the

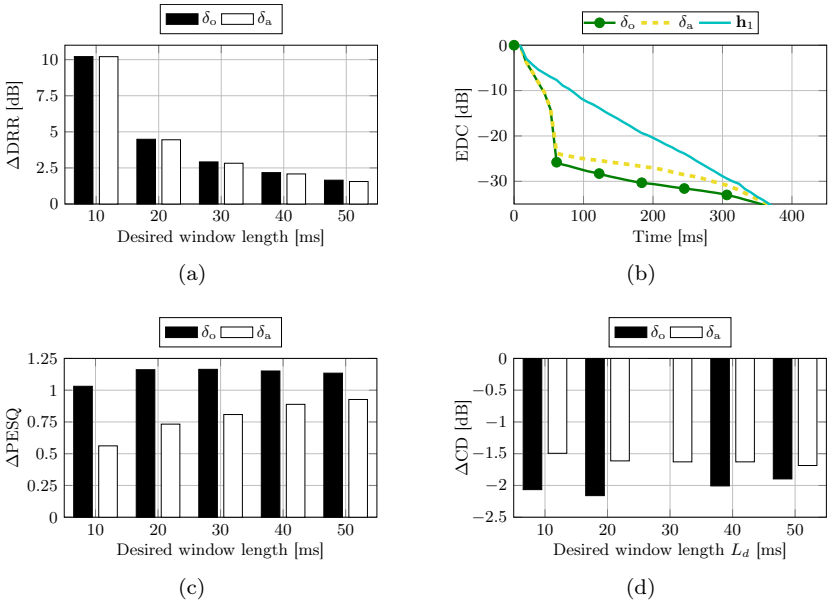


Fig. 5.7: Performance of the regularized PMINT technique using the optimal and the automatic regularization parameters δ_{o} and δ_{a} in terms of (a) ΔDRR, (b) EDC for the desired window length $L_d = 50$ ms, (c) ΔPESQ, and (d) ΔCD (averaged over several NPM values).

dereverberation error and distortion energy considerations, without directly taking into account perceptual criteria. Hence, achieving such a similar perceptual speech quality using this parameter as compared to the intrusively determined parameter which maximizes the PESQ score can be considered quite a good result.

Summarizing, the presented results show that incorporating the automatic regularization parameter in the regularized PMINT technique leads to a nearly-optimal performance, making the regularized PMINT technique not only a robust and perceptually advantageous equalization technique, but practically applicable as well.

5.6 Summary

In this chapter we have proposed to increase the robustness of acoustic multi-channel equalization techniques by incorporating regularization, such that the energy of distortions due to RIR perturbations is controlled. In addition, we have proposed an automatic non-intrusive procedure for determining the regularization parameter based on the L-curve.

Extensive simulation results have shown that incorporating regularization in all considered acoustic multi-channel equalization techniques, i.e., MINT, CS, RMCLS, and PMINT, yields a significantly better dereverberation performance in the presence of RIR perturbations. It has been shown that out of the considered regularized techniques, the regularized RMCLS and PMINT techniques result in the highest dereverberation performance, both in terms of reverberant energy suppression and perceptual speech quality improvement. The regularized RMCLS technique yields a slightly better reverberant energy suppression, whereas the regularized PMINT technique results in a higher perceptual speech quality improvement. The advantage of building upon the RMCLS technique to increase the robustness against RIR perturbations lies in its relaxed optimization criterion, whereas the advantage of building upon the PMINT technique lies in its direct control of the early reflections. Furthermore, it has been experimentally validated that the automatic non-intrusive regularization parameter in the regularized PMINT technique leads to a similar performance as the intrusively determined optimal regularization parameter, making the regularized PMINT technique a robust, perceptually advantageous, and practically applicable multi-channel equalization technique for speech dereverberation.

6

SPARSITY-PROMOTING ACOUSTIC MULTI-CHANNEL EQUALIZATION

In Chapters 4 and 5 we have proposed to increase the robustness of acoustic multi-channel equalization techniques against room impulse response (RIR) perturbations either by decreasing the reshaping filter length or by incorporating regularization, which are both signal-independent methods. In this chapter, we propose a signal-dependent method to increase the robustness by enforcing the output speech signal to exhibit characteristics of a clean speech signal. While in principle any well-defined measure that distinguishes clean and reverberant speech can be used, we propose to exploit the observation that clean speech is more sparse than reverberant speech in the time-frequency domain. Based on this observation, the presented least-squares and channel shortening cost functions are extended with a sparsity-promoting penalty function, which aims at obtaining a reshaping filter that sparsifies the time-frequency representation of the resulting output speech signal. Similarly as for the distortion energy in Chapter 5, the sparsity-promoting penalty function is scaled by a weighting parameter, which enables to trade off between dereverberation error energy and sparsity of the resulting output speech signal.

Section 6.1 discusses the sparse nature of clean speech signals in the short-time Fourier transform (STFT) domain and the effects of reverberation on the statistics of the STFT coefficients. In Section 6.2 the general framework for incorporating a sparsity-promoting penalty function in acoustic multi-channel equalization techniques is established. Furthermore, several commonly used sparsity-promoting penalty functions (l_0-norm, l_1-norm, and weighted l_1-norm) are introduced, and insights on the advantages of using these penalty functions for speech dereverberation

This chapter is partly based on:

[133] I. Kodrasi, A. Jukić, and S. Doclo, "Robust sparsity-promoting acoustic multi-channel equalization for speech dereverberation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Shanghai, China, Mar. 2016.

[134] I. Kodrasi and S. Doclo, "Robust acoustic multi-channel equalization for speech dereverberation using signal-dependent penalty functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016, manuscript submitted for publication.

are provided. The iterative algorithms used to compute the sparsity-promoting least-squares and channel shortening reshaping filters are presented in Section 6.3. By means of instrumental performance measures, simulation results in Section 6.4 show that incorporating a sparsity-promoting penalty function in the considered acoustic multi-channel equalization techniques, i.e., MINT, CS, RMCLS, and PMINT, increases the robustness against RIR perturbations. Furthermore, it is shown that out of the considered sparsity-promoting penalty functions, the weighted l_1-norm is advantageous in order to preserve the spectro-temporal structure of speech signals and to achieve speech dereverberation. Finally, it is experimentally validated that the weighted l_1-norm sparsity-promoting RMCLS and PMINT techniques are computationally tractable and robust techniques, yielding a high dereverberation performance, both in terms of reverberant energy suppression and perceptual speech quality improvement.

6.1 Sparsity of speech signals

Clean speech signals are known to be sparse in the STFT domain [173]. This means that speech is present only in some time frames, and in these time frames only some frequency bins have significant energy while many frequency bins have (nearly) no energy. Sparsity in the STFT domain arises due to pauses between phonemes and due to formant transitions in voiced sounds. Supported by empirical observations, e.g., in [174–177], it is therefore widely accepted that the clean speech STFT coefficients can be statistically modeled using sparse priors. Furthermore, empirical observations, e.g., in [173, 175], have shown that when clean speech is corrupted by reverberation (and noise), the STFT coefficients of the mixture are less sparse than the STFT coefficients of the clean speech signal. To illustrate the fact that clean speech is more sparse than reverberant speech in the STFT domain, Figs. 6.1a and 6.1b depict exemplary spectrograms of clean and reverberant speech signals. Due to the spectro-temporal smearing effect of reverberation, the speech pauses between phonemes and the formant transitions in vowels are filled by reverberant energy, making the reverberant spectrogram less sparse than the clean speech spectrogram.

Exploiting the sparse nature of clean speech signals has proven to be beneficial in many speech enhancement techniques, such as in under-determined blind source separation [173, 178–180], adaptive beamforming [175], and single- or multi-channel dereverberation [95, 96, 181, 182]. In the context of dereverberation, in [181] a probabilistic modeling-based single-channel technique has been proposed, where dereverberation is achieved using an iterative algorithm to compute a filter that yields a sparse output speech signal in the STFT domain. In [182] a blind multi-channel speech dereverberation technique has been investigated, which jointly estimates the clean speech signal and the RIRs, exploiting the sparse nature of the clean speech STFT coefficients and a statistical reverberation model for the RIRs. Furthermore, in [96] it has been shown that the dereverberation performance of the conventional probabilistic modeling-based multi-channel linear prediction technique [91] can be improved by modeling the dereverberated speech signal using a sparse prior.

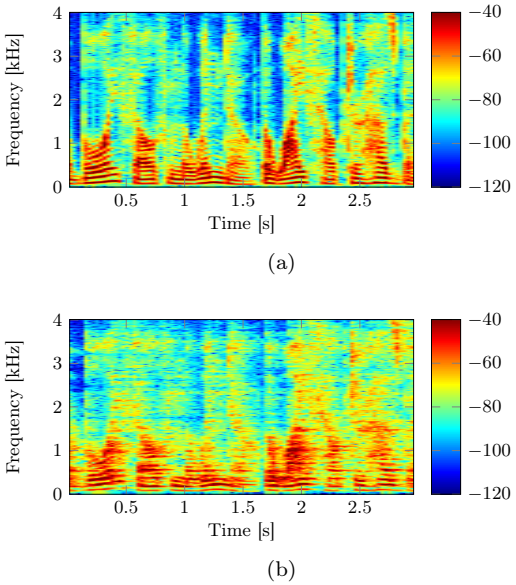


Fig. 6.1: Exemplary spectrograms of (a) clean speech and (b) reverberant speech. The STFT is computed using a 32 ms Hamming window with 50 % overlap between successive frames ($T_{60} \approx 610$ ms).

Given the successful exploitation of the sparse nature of clean speech in several speech enhancement techniques, in this chapter we propose a signal-dependent method to increase the robustness of acoustic multi-channel equalization techniques, taking into account the sparsity of the output speech signal STFT coefficients in the reshaping filter design.

6.2 Incorporating sparsity-promoting penalty functions in acoustic multi-channel equalization

In this section, we discuss how sparsity-promoting penalty functions can be incorporated in acoustic multi-channel equalization techniques. Section 6.2.1 establishes the general framework for extending the equalization cost functions with a sparsity-promoting penalty function. In addition, Section 6.2.2 introduces several commonly used sparsity-promoting penalty functions and provides insights on the advantages of using these penalty functions for speech dereverberation.

6.2.1 *Sparsity-promoting acoustic multi-channel equalization cost functions*

As discussed in Section 2.1.2, acoustic multi-channel equalization techniques typically disregard the presence of background noise and design reshaping filters aiming

only at speech dereverberation. Assuming that $\mathbf{v}(n) = \mathbf{0}$ in (2.26), the output signal of the speech enhancement system is given by

$$z(n) = \mathbf{w}^T\mathbf{x}(n) = \underbrace{\mathbf{w}^T\mathbf{H}^T}_{\mathbf{c}^T}\mathbf{s}(n), \tag{6.1}$$

with $\mathbf{w}$ the ML_w–dimensional reshaping filter vector, cf. (2.14), $\mathbf{x}(n)$ the ML_w-dimensional vector of the microphone signals, cf. (2.17), $\mathbf{H}$ the $L_c \times ML_w$–dimensional multi-channel convolution matrix of the true RIRs, cf. (2.24), $\mathbf{s}(n)$ the L_c-dimensional clean speech vector, cf. (2.23), and $\mathbf{c}$ the L_c–dimensional equalized impulse response between the clean speech signal and the output speech signal, cf. (2.27). In order to incorporate the output speech signal in the reshaping filter design, we consider the L_z–dimensional output signal vector $\mathbf{z}(n)$, i.e.,

$$\mathbf{z}(n) = [z(n)\ z(n-1)\ \ldots\ z(n-L_z+1)]^T. \tag{6.2}$$

Based on (6.1), the output signal vector can be written as

$$\mathbf{z}(n) = \mathbf{X}(n)\mathbf{w}, \tag{6.3}$$

where $\mathbf{X}(n)$ is the $L_z \times ML_w$–dimensional multi-channel convolution matrix of the microphone signals, i.e.,

$$\mathbf{X}(n) = [\mathbf{X}_1(n)\ \mathbf{X}_2(n)\ \ldots\ \mathbf{X}_M(n)], \tag{6.4}$$

with the $L_z \times L_w$–dimensional convolution matrix $\mathbf{X}_m(n)$ equal to

$$\mathbf{X}_m(n) = \begin{bmatrix} x_m(n) & x_m(n-1) & \cdots & x_m(n-L_w+1) \\ x_m(n-1) & x_m(n-2) & \cdots & x_m(n-L_w) \\ \vdots & \vdots & \ddots & \vdots \\ x_m(n-L_z+1) & x_m(n-L_z) & \cdots & x_m(n-L_w-L_z+2) \end{bmatrix}. \tag{6.5}$$

For notational convenience, the time index n is omitted when possible in the remainder of this chapter.

Since clean speech can be considered to be more sparse than reverberant speech in the time-frequency domain [173], the STFT is used to obtain a time-frequency representation of the output speech signal. The STFT coefficients of the output speech signal are computed as (cf. Section 2.1.3)

$$Z(t,f) = \sum_{n=0}^{N-1} w_{\text{STFT}}(n) z(tR+n) e^{\frac{-j2\pi fn}{N}}, \tag{6.6}$$

with $t = 0,\ 1,\ \ldots,\ T-1$, the time frame index and T the total number of time frames, $f = 0,\ 1,\ \ldots,\ N-1$, the frequency bin index and N the frame size, $w_{\text{STFT}}(n)$ the STFT analysis window, and R the frame shift. Similarly as in [179, 180], we define the STFT operator $\mathbf{\Psi} \in \mathcal{C}^{L_{\tilde{z}} \times L_z}$, which transforms the L_z–dimensional time domain vector $\mathbf{z}$ into the $L_{\tilde{z}}$–dimensional time-frequency domain vector $\tilde{\mathbf{z}}$ consisting

of all STFT coefficients $Z(t,f)$ (i.e., $\tilde{\mathbf{z}}$ denotes the stacked vector of all columns of the spectrogram of $\mathbf{z}$), i.e.,

$$\tilde{\mathbf{z}} = \mathbf{\Psi}\mathbf{z} = \mathbf{\Psi}\mathbf{X}\mathbf{w}, \tag{6.7}$$

with

$$\tilde{\mathbf{z}} = [Z(0,0)\ \ldots\ Z(0,N-1)\ \ldots\ Z(T-1,0)\ \ldots\ Z(T-1,N-1)]^T \tag{6.8}$$

$$= [\tilde{Z}(0)\ \ldots\ \tilde{Z}(L_{\tilde{z}}-1)]^T, \tag{6.9}$$

and $L_{\tilde{z}} = T \times N$ being the total number of STFT coefficients. Using a tight STFT analysis window $w_{\text{STFT}}(n)$, i.e., the same window can be used as a synthesis window such that the perfect overlap-add constraint is satisfied, the inverse short-time Fourier transform (ISTFT) operator $\mathbf{\Psi}^H \in \mathcal{C}^{L_z \times L_{\tilde{z}}}$ is such that

$$\mathbf{\Psi}^H\mathbf{\Psi} = \mathbf{I}, \tag{6.10}$$

with $\mathbf{I}$ the $L_z \times L_z$–dimensional identity matrix.

In order to sparsify the STFT coefficients of the output speech signal, we propose to incorporate a sparsity-promoting penalty function $f_{\text{sp}}(\tilde{\mathbf{z}})$ in the least-squares and channel shortening cost functions defined in (3.29) and (3.18). As it will be experimentally validated in Section 6.4.2, incorporating sparsity-promoting penalty functions increases the robustness of equalization techniques against RIR perturbations.

The proposed sparsity-promoting least-squares cost function $J_{\text{S-LS}}$ is defined as

$$J_{\text{S-LS}} = J_{\text{LS}} + \eta f_{\text{sp}}(\tilde{\mathbf{z}}) \tag{6.11}$$

$$= \underbrace{\|\mathbf{W}(\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t)\|_2^2}_{\epsilon_c} + \eta \underbrace{f_{\text{sp}}(\mathbf{\Psi}\mathbf{X}\mathbf{w})}_{\epsilon_s}, \tag{6.12}$$

where ϵ_c denotes the least-squares dereverberation error energy as in (5.7), ϵ_s denotes the sparsity measure of the STFT representation of the output speech signal, and η is a weighting parameter providing a trade-off between the two terms.

Furthermore, in order to incorporate a sparsity-promoting penalty function in the channel shortening cost function, the channel shortening *maximization* problem in (3.18) is first reformulated as a generalized Rayleigh quotient *minimization* problem as in (5.8). The proposed sparsity-promoting channel shortening cost function $J_{\text{S-CS}}$ is then defined as

$$J_{\text{S-CS}} = J_{\text{CS}}^{\min} + \eta f_{\text{sp}}(\tilde{\mathbf{z}}) \tag{6.13}$$

$$= \underbrace{\frac{\mathbf{w}^T\hat{\mathbf{U}}\mathbf{w}}{\mathbf{w}^T\hat{\mathbf{D}}\mathbf{w}}}_{\epsilon_r} + \eta \underbrace{f_{\text{sp}}(\mathbf{\Psi}\mathbf{X}\mathbf{w})}_{\epsilon_s}, \tag{6.14}$$

where ϵ_r denotes the channel shortening dereverberation error energy as in (5.10).

In the following section, several sparsity-promoting penalty functions $f_{\text{sp}}(\tilde{\mathbf{z}})$ are presented and discussed.

6.2.2 *Sparsity-promoting penalty functions*

In the presence of reverberation, the pauses between phonemes and the formant transitions in voiced sounds, i.e., the (nearly) zero energy STFT coefficients of the clean speech signal, are filled with reverberant energy (cf. Figs. 6.1a and 6.1b). In addition, in the presence of RIR perturbations, non-robust acoustic multi-channel equalization techniques introduce additional distortions (i.e., additional non-zero STFT coefficients) in the output speech signal (cf., e.g., Fig. 4.8c). In order to increase the robustness of acoustic multi-channel equalization techniques against RIR perturbations and obtain an output speech signal that better resembles a clean speech signal, one possibility is to minimize the number of non-zero coefficients in $\tilde{\mathbf{z}}$, which can be achieved by using an l_0-norm[1] penalty function, i.e.,

$$f_{\text{sp}}^{0}(\tilde{\mathbf{z}}) = \|\tilde{\mathbf{z}}\|_0 = |i : \tilde{z}(i) \neq 0|. \tag{6.15}$$

However, the l_0-norm in (6.15) is non-convex and it is well known that optimization problems with non-convex penalty functions are typically hard (if not impossible) to solve, particularly for large scale problems [183]. In addition, iterative methods proposed to solve such optimization problems are not guaranteed to converge to the global minimum, but only to a local minimum [184].

A common alternative to the l_0-norm is the l_1-norm, i.e.,

$$f_{\text{sp}}^{1}(\tilde{\mathbf{z}}) = \|\tilde{\mathbf{z}}\|_1 = \sum_{i=0}^{L_{\tilde{z}}-1} |\tilde{z}(i)|, \tag{6.16}$$

which differs from the l_0-norm by penalizing larger coefficients of $\tilde{\mathbf{z}}$ more than smaller coefficients. The l_1-norm can be viewed as a convex relaxation of the l_0-norm, and efficient methods have been proposed to solve optimization problems with l_1-norm penalty functions [185, 186]. Furthermore, it has been shown that under certain conditions, replacing the l_0-norm by the l_1-norm provides the solution to the original l_0-norm optimization problem [187,188]. In practice however, the l_1-norm relaxation is very often used when these conditions are not satisfied, typically resulting in a solution which does not optimize the original l_0-norm optimization problem, but nevertheless provides smaller l_0-norm values.

To counteract the magnitude dependency of the l_1-norm, i.e., to better mimic the l_0-norm, the weighted l_1-norm penalty function has been proposed [189], which selectively penalizes the coefficients of $\tilde{\mathbf{z}}$ based on a weighting vector $\mathbf{u}$, i.e.,

$$f_{\text{sp}}^{w,1}(\tilde{\mathbf{z}}) = \|\text{diag}\{\mathbf{u}\}\tilde{\mathbf{z}}\|_1 = \sum_{i=0}^{L_{\tilde{z}}-1} |u(i)\tilde{z}(i)|, \tag{6.17}$$

with $\mathbf{u}$ an $L_{\tilde{z}}$–dimensional vector of positive scalar weights, i.e., $u(i) > 0$, $i = 0,\ 1,\ \ldots,\ L_{\tilde{z}} - 1$. Using $u(i) = 1$, $i = 0,\ 1,\ \ldots,\ L_{\tilde{z}} - 1$, yields the standard l_1-norm

1 Note that the l_0-norm is not a norm in the mathematical sense, since it does not satisfy all the norm properties.

penalty function in (6.16). To counteract the magnitude dependency of the l_1-norm and to promote the same sparsity structure in the output signal which is present in the desired signal, it has been proposed in [189] to select the weighting vector such that it has small values on the non-zero locations of the desired signal and significantly larger values elsewhere. Replacing the l_1-norm by a weighted l_1-norm has been shown to enhance sparsity and improve the performance in sparse signal recovery applications [189, 190]. In the context of speech dereverberation, it would hence be desirable to select the weights to be inversely proportional to the magnitude of the STFT coefficients of the clean speech signal, i.e.,

$$u(i) = \frac{1}{|\tilde{s}(i)|}, \quad i = 0,\ 1,\ \ldots,\ L_{\tilde{z}} - 1, \tag{6.18}$$

with $\tilde{s}(i)$ the STFT coefficients of the clean speech signal computed as in (6.7). Using these weights results in penalizing more, and hence, decreasing more the coefficients of $\tilde{\mathbf{z}}$ corresponding to the small STFT coefficients of the clean speech signal. However, since the clean speech signal is not available, we propose to use any of the reverberant microphone signals and compute the weights as

$$u(i) = \frac{1}{|\tilde{x}_p(i)| + \zeta}, \quad i = 0,\ 1,\ \ldots,\ L_{\tilde{z}} - 1, \tag{6.19}$$

where $\tilde{x}_p(i)$ are the STFT coefficients of the p-th microphone signal computed as in (6.7) with $p \in \{1,\ \ldots,\ M\}$ and $\zeta > 0$ is a small positive scalar to avoid division by 0. As is experimentally validated in Section 6.4, incorporating the weighted l_1-norm penalty function in (6.17) using the weights in (6.19) yields a better dereverberation performance than incorporating the l_0- or l_1-norm penalty functions. As is experimentally validated in Section 6.4.3, the advantage of using the weighted l_1-norm instead of the l_0- or l_1-norm lies in the fact that appropriate weights as in (6.19) ensure that the spectro-temporal structure of a typical speech signal is preserved.

6.3 Sparsity-promoting acoustic multi-channel equalization reshaping filters

Since no closed-form solution is available for the cost functions in (6.12) and (6.14) which incorporate sparsity-promoting penalty functions, iterative optimization algorithms are required, which can be computationally expensive. As a result, there has been much research in the development of efficient iterative optimization algorithms for solving convex as well as non-convex optimization problems. We have chosen to use the alternating direction method of multipliers (ADMM), since it is a well-suited algorithm for solving large-scale optimization problems of the form (6.12) and (6.14) [191,192]. The ADMM algorithm was originally proposed in [193] and has been successfully applied to a large number of statistical problems such as sparse signal recovery [194] and image processing [195, 196]. As described in [191], the ADMM algorithm can be considered to be a decomposition-coordination procedure, in which the optimization of simple local sub-problems is coordinated to optimize a more complex global problem.

6.3.1 *Sparsity-promoting least-squares reshaping filter*

Within the ADMM framework, the minimization of the sparsity-promoting least-squares cost function in (6.12) is reformulated as

$$\min_{\mathbf{w}} \left[\|\mathbf{W}(\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t)\|_2^2 + \eta f_{\text{sp}}(\tilde{\mathbf{a}}) \right] \text{ subject to } \boldsymbol{\Psi}\mathbf{X}\mathbf{w} = \tilde{\mathbf{a}}, \tag{6.20}$$

with $\tilde{\mathbf{a}}$ an auxiliary variable which is introduced such that the optimization problem in (6.12) can be split into simpler sub-problems. In order to solve (6.20), the augmented Lagrangian is formed, i.e.,

$$\mathcal{L}_{\text{S-LS}}(\mathbf{w}, \tilde{\mathbf{a}}, \boldsymbol{\gamma}) = \|\mathbf{W}(\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t)\|_2^2 + \eta f_{\text{sp}}(\tilde{\mathbf{a}}) + \boldsymbol{\gamma}^H(\boldsymbol{\Psi}\mathbf{X}\mathbf{w} - \tilde{\mathbf{a}}) + \frac{\rho}{2}\|\boldsymbol{\Psi}\mathbf{X}\mathbf{w} - \tilde{\mathbf{a}}\|_2^2, \tag{6.21}$$

with $\boldsymbol{\gamma}$ the $L_{\tilde{z}}$–dimensional vector of Lagrangian multipliers and $\rho > 0$ a penalty parameter. As the penalty parameter approaches ∞, i.e., $\rho \to \infty$, it is ensured that the variable $\boldsymbol{\Psi}\mathbf{X}\mathbf{w}$ converges to the auxiliary variable $\tilde{\mathbf{a}}$. In each iteration of the ADMM algorithm, the values of the variables $\mathbf{w}$ and $\tilde{\mathbf{a}}$ are updated by minimizing the augmented Lagrangian in (6.21) with respect to $\mathbf{w}$ and $\tilde{\mathbf{a}}$. The advantage of using the ADMM algorithm is that the minimization over the variables $\mathbf{w}$ and $\tilde{\mathbf{a}}$ is done in an alternating manner, which allows the problem to be easily decomposed and solved.

Using the scaled dual variable

$$\boldsymbol{\lambda} = \frac{\boldsymbol{\gamma}}{\rho}, \tag{6.22}$$

the augmented Lagrangian in (6.21) is equal to

$$\mathcal{L}_{\text{S-LS}}(\mathbf{w}, \tilde{\mathbf{a}}, \boldsymbol{\lambda}) = \|\mathbf{W}(\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t)\|_2^2 + \eta f_{\text{sp}}(\tilde{\mathbf{a}}) + \frac{\rho}{2}\|\boldsymbol{\Psi}\mathbf{X}\mathbf{w} - \tilde{\mathbf{a}} + \boldsymbol{\lambda}\|_2^2 - \frac{\rho}{2}\|\boldsymbol{\lambda}\|_2^2. \tag{6.23}$$

While (6.21) and (6.23) are equivalent, the augmented Lagrangian in (6.23) is often used within the ADMM framework for convenience, since it results in shorter expressions for the ADMM update rules. Using (6.23), the ADMM update rules for the sparsity-promoting least-squares techniques can be expressed as [191]

$$\mathbf{w}^{(k+1)} = \arg\min_{\mathbf{w}} \left[\|\mathbf{W}(\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t)\|_2^2 + \frac{\rho}{2}\|\boldsymbol{\Psi}\mathbf{X}\mathbf{w} - \tilde{\mathbf{a}}^{(k)} + \boldsymbol{\lambda}^{(k)}\|_2^2 \right], \tag{6.24}$$

$$\tilde{\mathbf{a}}^{(k+1)} = \arg\min_{\tilde{\mathbf{a}}} \left[\eta f_{\text{sp}}(\tilde{\mathbf{a}}) + \frac{\rho}{2}\|\boldsymbol{\Psi}\mathbf{X}\mathbf{w}^{(k+1)} - \tilde{\mathbf{a}} + \boldsymbol{\lambda}^{(k)}\|_2^2 \right], \tag{6.25}$$

$$\boldsymbol{\lambda}^{(k+1)} = \boldsymbol{\lambda}^{(k)} + \boldsymbol{\Psi}\mathbf{X}\mathbf{w}^{(k+1)} - \tilde{\mathbf{a}}^{(k+1)}, \tag{6.26}$$

where $\{\cdot\}^{(k)}$ denotes the variable in the k-th iteration and the update rule in (6.26) is the dual variable update rule used for coordination [191]. Hence, the original minimization problem in (6.12) is decomposed into simpler sub-problems which are solved in an alternating fashion using the update rules in (6.24), (6.25), and (6.26) until a convergence criterion is satisfied or a maximum number of iterations is exceeded.

Sparsity-promoting filter update rule

In order to derive the update rule for the least-squares reshaping filter in (6.24), the gradient of the cost function with respect to $\mathbf{w}$ is set to $\mathbf{0}$, i.e.,

$$2\hat{\mathbf{H}}^T\mathbf{W}^T\mathbf{W}\hat{\mathbf{H}}\mathbf{w} + \rho\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\hat{\mathbf{H}}^T\mathbf{W}^T\mathbf{W}\mathbf{c}_t - \rho\mathbf{X}^T\boldsymbol{\Psi}^H(\tilde{\mathbf{a}}^{(k)} - \boldsymbol{\lambda}^{(k)}) = 0, \quad (6.27)$$

yielding the sparsity-promoting reshaping filter update rule

$$\mathbf{w}_{\text{S-LS}}^{(k+1)} = (\underbrace{2\hat{\mathbf{H}}^T\mathbf{W}^T\mathbf{W}\hat{\mathbf{H}} + \rho\mathbf{X}^T\mathbf{X}}_{\mathbf{C}})^{-1}[\underbrace{2\hat{\mathbf{H}}^T\mathbf{W}^T\mathbf{W}\mathbf{c}_t}_{\mathbf{b}_1} + \rho\underbrace{\mathbf{X}^T\boldsymbol{\Psi}^H(\tilde{\mathbf{a}}^{(k)} - \boldsymbol{\lambda}^{(k)})}_{\mathbf{b}_2^{(k)}}] \quad (6.28)$$

$$= \mathbf{C}^{-1}[\mathbf{b}_1 + \rho\mathbf{b}_2^{(k)}], \quad (6.29)$$

where the variables $\mathbf{C}$, $\mathbf{b}_1$, and $\mathbf{b}_2$ are introduced to simplify the notation and to show that only the variable $\mathbf{b}_2$ is iteration-dependent. Although (6.29) requires a matrix inversion in each iteration of the ADMM algorithm, the matrix $\mathbf{C}$ remains fixed throughout the optimization procedure, such that the filter update can be efficiently computed by, e.g., storing the LU factorization of $\mathbf{C}$ and using forward and backward substitution in each iteration. Similarly, also the vector $\mathbf{b}_1$ remains fixed throughout the optimization procedure, such that it needs to be computed only once.

For completeness, Table 6.1 summarizes the sparsity-promoting least-squares filter update rules for the considered sparsity-promoting least-squares techniques, i.e., for the different definitions of the weighting matrix $\mathbf{W}$ and the target equalized impulse response $\mathbf{c}_t$ introduced in Section 3.3.

Auxiliary variable update rule

The update rule for the auxiliary variable in (6.25) depends on the used sparsity-promoting penalty function. To simplify the notation, we define the variable

$$\mathbf{b}^{(k)} = \boldsymbol{\Psi}\mathbf{X}\mathbf{w}^{(k+1)} + \boldsymbol{\Lambda}^{(k)}. \quad (6.30)$$

Table 6.1: Sparsity-promoting least-squares reshaping filter update rules for different sparsity-promoting least-squares techniques.

Sparsity-promoting technique	Filter update rule
S-MINT	$\mathbf{w}_{\text{S-M}}^{(k+1)}=(2\hat{\mathbf{H}}^T\hat{\mathbf{H}} + \rho\mathbf{X}^T\mathbf{X})^{-1}\left[2\hat{\mathbf{H}}^T\mathbf{d} + \rho\mathbf{X}^T\boldsymbol{\Psi}^H(\tilde{\mathbf{a}}^{(k)} - \boldsymbol{\lambda}^{(k)})\right]$
S-RMCLS	$\mathbf{w}_{\text{S-R}}^{(k+1)}=(2\hat{\mathbf{H}}^T\mathbf{W}_{\text{R}}^T\mathbf{W}_{\text{R}}\hat{\mathbf{H}}+\rho\mathbf{X}^T\mathbf{X})^{-1}\left[2\hat{\mathbf{H}}^T\mathbf{W}_{\text{R}}^T\mathbf{W}_{\text{R}}\mathbf{d}+\rho\mathbf{X}^T\boldsymbol{\Psi}^H(\tilde{\mathbf{a}}^{(k)}-\boldsymbol{\lambda}^{(k)})\right]$
S-PMINT	$\mathbf{w}_{\text{S-P}}^{(k+1)}=(2\hat{\mathbf{H}}^T\hat{\mathbf{H}} + \rho\mathbf{X}^T\mathbf{X})^{-1}\left[2\hat{\mathbf{H}}^T\hat{\mathbf{h}}_{e,p} + \rho\mathbf{X}^T\boldsymbol{\Psi}^H(\tilde{\mathbf{a}}^{(k)} - \boldsymbol{\lambda}^{(k)})\right]$

Substituting (6.30) in (6.25), the auxiliary variable update rule can be written as

$$\tilde{\mathbf{a}}^{(k+1)} = \arg\min_{\tilde{\mathbf{a}}} \left[\eta f_{\text{sp}}(\tilde{\mathbf{a}}) + \frac{\rho}{2} \|\mathbf{b}^{(k)} - \tilde{\mathbf{a}}\|_2^2 \right]. \tag{6.31}$$

The solution of the optimization problem in (6.31) is the proximal mapping of the sparsity-promoting penalty function [197]. The proximal mapping for the l_0-, l_1-, and weighted l_1-norm penalty functions exists in closed form [185, 197], which enables to efficiently compute the auxiliary variable update rule in each iteration of the ADMM algorithm. The proximal mapping for these penalty functions is presented in the following.

- f_{sp}^{0} *(l_0-norm)*
  The proximal mapping for the l_0-norm penalty function is the component-wise hard-thresholding map, i.e.,

$$\tilde{a}(i)^{(k+1)} = \begin{cases} 0 & \text{if} \quad |b(i)^{(k)}| \leq \frac{\eta}{\rho}, \\ b(i)^{(k)} & \text{otherwise.} \end{cases} \tag{6.32}$$

  Hard-thresholding uses a non-linear operator to reduce the l_0-norm of $\tilde{\mathbf{a}}$ in each iteration of the ADMM algorithm by changing all but the largest elements (i.e., larger than $\frac{\eta}{\rho}$) to 0.
- f_{sp}^{1} *(l_1-norm)*
  The proximal mapping for the l_1-norm penalty function is the component-wise soft-thresholding map, i.e.,

$$\tilde{a}(i)^{(k+1)} = \begin{cases} 0 & \text{if} \quad |b(i)^{(k)}| \leq \frac{\eta}{\rho}, \\ \left\{ |b(i)^{(k)}| - \frac{\eta}{\rho} \right\} \frac{b(i)^{(k)}}{|b(i)^{(k)}|} & \text{otherwise.} \end{cases} \tag{6.33}$$

  Soft-thresholding reduces the l_1-norm of $\tilde{\mathbf{a}}$ in each iteration of the ADMM algorithm by subtracting $\frac{\eta}{\rho}$ from the absolute value of every element of $\tilde{\mathbf{a}}$.
- $f_{sp}^{w,1}$ *(weighted l_1-norm)*
  The proximal mapping for the weighted l_1-norm penalty function is similar to the soft-thresholding in (6.33), with the only difference consisting in the multiplication of the soft threshold with the weights in $\mathbf{u}$, i.e.,

$$\tilde{a}(i)^{(k+1)} = \begin{cases} 0 & \text{if} \quad |b(i)^{(k)}| \leq \frac{\eta}{\rho} u(i), \\ \left\{ |b(i)^{(k)}| - \frac{\eta}{\rho} u(i) \right\} \frac{b(i)^{(k)}}{|b(i)^{(k)}|} & \text{otherwise.} \end{cases} \tag{6.34}$$

  Hence, weighted soft-thresholding reduces the l_1-norm of $\tilde{\mathbf{a}}$ in each iteration of the ADMM algorithm by subtracting $\frac{\eta}{\rho} u(i)$ from the absolute value of every element of $\tilde{\mathbf{a}}$.

Fig. 6.2 provides a schematic illustration of the difference between hard-thresholding, soft-thresholding, and weighted soft-thresholding for exemplary values $\frac{\eta}{\rho} = 1$ and $u(i) = 2$.

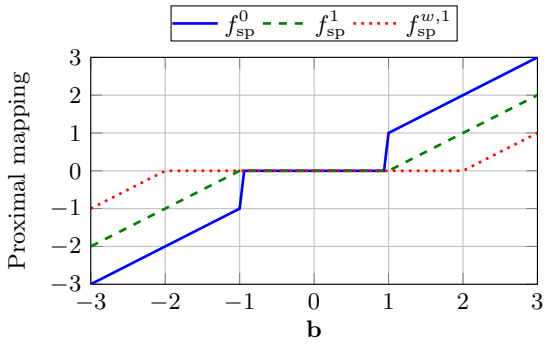


Fig. 6.2: Exemplary illustration of the proximal mappings for the l_0-norm, l_1-norm, and weighted l_1-norm ($\frac{\eta}{\rho} = 1$ and $u(i) = 2$).

Summarizing, using the reshaping filter update rule in (6.28), the auxiliary variable update rule in (6.32), (6.33), or (6.34) depending on the used sparsity-promoting penalty function, and the dual variable update rule in (6.26) until a termination criterion is satisfied, the sparsity-promoting least-squares reshaping filter can be computed. The initialization and termination criterion used for the ADMM algorithm will be discussed in Section 6.4.1.

6.3.2 *Sparsity-promoting channel shortening reshaping filter*

Similarly as for the sparsity-promoting least-squares cost function, the minimization of the sparsity-promoting channel shortening cost function in (6.14) is reformulated within the ADMM framework as

$$\min_{\mathbf{w}} \left[\frac{\mathbf{w}^T \hat{\mathbf{U}} \mathbf{w}}{\mathbf{w}^T \hat{\mathbf{D}} \mathbf{w}} + \eta f_{\text{sp}}(\tilde{\mathbf{a}}) \right] \quad \text{subject to} \quad \mathbf{\Psi X w} = \tilde{\mathbf{a}}, \tag{6.35}$$

with $\tilde{\mathbf{a}}$ an auxiliary variable. The augmented Lagrangian of (6.35) now is equal to

$$\mathcal{L}_{\text{S-CS}}(\mathbf{w}, \tilde{\mathbf{a}}, \boldsymbol{\gamma}) = \frac{\mathbf{w}^T \hat{\mathbf{U}} \mathbf{w}}{\mathbf{w}^T \hat{\mathbf{D}} \mathbf{w}} + \eta f_{\text{sp}}(\tilde{\mathbf{a}}) + \boldsymbol{\gamma}^H (\mathbf{\Psi X w} - \tilde{\mathbf{a}}) + \frac{\rho}{2} \|\mathbf{\Psi X w} - \tilde{\mathbf{a}}\|_2^2, \tag{6.36}$$

with $\boldsymbol{\gamma}$ the $L_{\tilde{z}}$–dimensional vector of Lagrangian multipliers and $\rho > 0$ a penalty parameter. Rewriting the augmented Lagrangian in (6.36) in terms of the scaled dual variable

$$\boldsymbol{\lambda} = \frac{\boldsymbol{\gamma}}{\rho} \tag{6.37}$$

yields

$$\mathcal{L}_{\text{S-CS}}(\mathbf{w}, \tilde{\mathbf{a}}, \boldsymbol{\lambda}) = \frac{\mathbf{w}^T \hat{\mathbf{U}} \mathbf{w}}{\mathbf{w}^T \hat{\mathbf{D}} \mathbf{w}} + \eta f_{\text{sp}}(\tilde{\mathbf{a}}) + \frac{\rho}{2} \|\mathbf{\Psi X w} - \tilde{\mathbf{a}} + \boldsymbol{\lambda}\|_2^2 - \frac{\rho}{2} \|\boldsymbol{\lambda}\|_2^2. \tag{6.38}$$

Based on (6.38) the update rules for the sparsity-promoting channel shortening technique are given by:

$$\mathbf{w}^{(k+1)} = \arg\min_{\mathbf{w}} \left[\frac{\mathbf{w}^T \hat{\mathbf{U}} \mathbf{w}}{\mathbf{w}^T \hat{\mathbf{D}} \mathbf{w}} + \frac{\rho}{2} \| \mathbf{\Psi} \mathbf{X} \mathbf{w} - \tilde{\mathbf{a}}^{(k)} + \mathbf{\Lambda}^{(k)} \|_2^2 \right], \tag{6.39}$$

$$\tilde{\mathbf{a}}^{(k+1)} = \arg\min_{\tilde{\mathbf{a}}} \left[\eta f_{\mathrm{sp}}(\tilde{\mathbf{a}}) + \frac{\rho}{2} \| \mathbf{\Psi} \mathbf{X} \mathbf{w}^{(k+1)} - \tilde{\mathbf{a}} + \mathbf{\Lambda}^{(k)} \|_2^2 \right], \tag{6.40}$$

$$\mathbf{\Lambda}^{(k+1)} = \mathbf{\Lambda}^{(k)} + \mathbf{\Psi} \mathbf{X} \mathbf{w}^{(k+1)} - \tilde{\mathbf{a}}^{(k+1)}. \tag{6.41}$$

As can be observed, the update rule for the auxiliary variable in (6.40) and the dual variable in (6.41) are the same as the update rule for the auxiliary variable in (6.25) and the dual variable in (6.26). The only difference between the update rules for the sparsity-promoting least-squares and channel shortening techniques consists in the reshaping filter update rule in (6.24) and (6.39). Unfortunately, no analytical solution minimizing (6.39) is available and therefore we have resorted to an iterative optimization procedure for minimizing this non-linear cost function, for which we have used the MATLAB function *fminunc* [168]. In order to improve the numerical robustness and the convergence speed, the gradient of (6.39) with respect to $\mathbf{w}$, i.e.,

$$2 \frac{\hat{\mathbf{U}} \mathbf{w} (\mathbf{w}^T \hat{\mathbf{D}} \mathbf{w}) - 2 (\mathbf{w}^T \hat{\mathbf{U}} \mathbf{w}) \hat{\mathbf{D}} \mathbf{w}}{(\mathbf{w}^T \hat{\mathbf{D}} \mathbf{w})^2} + \rho \mathbf{X}^T \mathbf{X} \mathbf{w} - 2 \rho \mathbf{X}^T \mathbf{\Psi}^H [\tilde{\mathbf{a}}^{(k)} - \mathbf{\Lambda}^{(k)}], \tag{6.42}$$

can be provided to the iterative optimization procedure. It should be noted that using an iterative optimization procedure to compute the filter update within each iteration of the ADMM algorithm results in a large computational complexity. Reducing the computational complexity for computing the sparsity-promoting channel shortening reshaping filter remains a topic for future investigation.

6.4 Simulations

In this section, we investigate the dereverberation performance of the sparsity-promoting least-squares and channel shortening equalization techniques. In Section 6.4.1 the considered acoustic system and the used algorithmic settings are introduced. In Section 6.4.2 the robustness increase of the considered acoustic multi-channel equalization techniques when incorporating different sparsity-promoting penalty functions is investigated. In Section 6.4.4 the performance of the weighted l_1-norm sparsity-promoting acoustic multi-channel equalization techniques is extensively compared.

6.4.1 *Acoustic system and algorithmic settings*

We have considered an acoustic system with a single speech source and $M = 4$ omni-directional microphones. The source-microphone distance is 2 m and the distance between the microphones is 4 cm. The room reverberation time is $T_{60} \approx 360$ ms and the direct-to-reverberant ratio is DRR = 5 dB [198]. The RIRs have been measured

using the swept-sine technique [162] and the length of the RIRs has been set to $L_h = 2880$ at a sampling frequency $f_s = 8$ kHz. To generate the reverberant signals, 10 sentences (approximately 17 s long) from the HINT database [163] have been convolved with the measured RIRs.[2]

Similarly as in Section 3.4, in order to simulate RIR perturbations, the measured RIRs are perturbed by adding scaled white noise as described in Section 2.2. The considered normalized projection misalignment (NPM) values between the true and the perturbed RIRs are (cf. (2.52))

$$\text{NPM} \in \{-33 \text{ dB}, -27 \text{ dB}, -21 \text{ dB}, -15 \text{ dB}\}. \tag{6.43}$$

For all considered techniques, the reshaping filter length is $L_w = \left\lceil \frac{L_h - 1}{M - 1} \right\rceil = 960$, the delay is set to $\tau = 90$, and the desired window length is $L_d = 10$ ms.[3] The target equalized impulse response for the PMINT and the sparsity-promoting PMINT techniques is set to the direct path and early reflections of the perturbed RIR of the first microphone, i.e., $\hat{\mathbf{h}}_{e,1}$. Furthermore, the channel shortening reshaping filter is selected as the generalized eigenvector yielding the minimum l_2-norm estimated equalized impulse response as proposed in [125].

In order to reduce the computational complexity of the reshaping filter design, the sparsity-promoting reshaping filters are computed using only the first 2 sentences of the microphone signals (approximately 3 s long). However, the complete output speech signal has been used for the evaluation. The STFT is computed using a 32 ms Hamming window with 50 % overlap between successive frames. The total number of time frames is $T = 208$, the frame size is $N = 256$, and hence, $L_{\tilde{z}} = T \times N = 26832$. For the weighted l_1-norm penalty function, the weighting vector $\mathbf{u}$ in (6.19) is generated using the first microphone signal, i.e., $p = 1$, and $\zeta = 10^{-8}$. Furthermore, for the ADMM algorithm a termination criterion needs to be imposed. In our implementation, the termination criterion is either the number of iterations exceeding a given maximum number of iterations or the relative change in the solution norm dropping below a given tolerance, i.e.,

$$k + 1 > k_{\max} \quad \text{or} \quad \frac{\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|_2}{\|\mathbf{w}^{(k)}\|_2} < \epsilon_{\mathbf{w}}, \tag{6.44}$$

with $k_{\max} = 150$ and $\epsilon_{\mathbf{w}} = 10^{-3}$.

Furthermore, since using the l_0-norm penalty function yields a non-convex optimization problem, the initialization of the ADMM algorithm may influence the resulting reshaping filter. Hence, we have investigated three different initializations of the filter $\mathbf{w}$, i.e.,

2 The considered acoustic system is different than in the previous chapters, with lower reverberation time, and hence, shorter RIRs. Shorter RIRs have been used in these simulation results due to the high computational complexity of the sparsity-promoting channel shortening technique. In Chapter 7, the performance of the sparsity-promoting RMCLS and PMINT techniques is investigated for higher reverberation times.

3 The desired window length has been limited to one value due to the high computational complexity of the sparsity-promoting channel shortening technique.

i) $\mathbf{w}^{(1)} = [1\ 0\ \ldots\ 0]^T$, i.e., the filter yielding the first microphone signal,
ii) $\mathbf{w}^{(1)}$ is randomly initialized with normally distributed coefficients,
iii) $\mathbf{w}^{(1)} = \mathbf{w}_{\text{LS}}$ or $\mathbf{w}^{(1)} = \mathbf{w}_{\text{CS}}$, i.e., initialization with the according reshaping filter resulting from the least-squares or channel shortening equalization techniques.

In all simulations we observed that using the first filter initialization, i.e., $\mathbf{w}^{(1)} = [1\ 0 \ldots 0]^T$, results in the best performance, and hence, the following simulation results have been generated using this filter initialization.

Using the instrumental performance measures described in Section 2.3, the dereverberation performance is evaluated in terms of the reverberant energy suppression and the perceptual speech quality improvement. The reverberant energy suppression is evaluated using the direct-to-reverberant ratio improvement (ΔDRR) between the equalized impulse response $\mathbf{c}$ and the true RIR $\mathbf{h}_1$ (cf. (2.53)), as well as the energy decay curve (EDC) of the equalized impulse response $\mathbf{c}$ (cf. (2.55)). The improvement in perceptual speech quality is evaluated using the improvement in PESQ [153] (ΔPESQ) and in cepstral distance [154] (ΔCD) between the output speech signal $z(n)$ and the reverberant microphone signal $x_1(n)$. The reference signal employed for the PESQ and cepstral distance measures is $x_{e,1}(n) = s(n) * h_{e,1}(n)$, i.e., the clean speech signal convolved with the direct path and early reflections of the first RIR.

In order to evaluate the effectiveness of incorporating a sparsity-promoting penalty function for increasing the robustness of acoustic multi-channel equalization techniques, we investigate the performance for several weighting and penalty parameters η and ρ, i.e.,

$$\eta = \{10^{-7},\ 10^{-6},\ 10^{-5},\ 10^{-4}\}, \tag{6.45}$$

$$\rho = \{10^{-7},\ 10^{-6},\ \ldots,\ 10^{-1}\}. \tag{6.46}$$

Similarly as in Chapters 4 and 5 for determining the optimal reshaping filter length and the optimal regularization parameter, in this chapter the optimal weighting and penalty parameters η_{o} and ρ_{o} are determined as the ones yielding the highest perceptual speech quality in terms of the PESQ score. It should be noted that the computation of the PESQ score for determining the optimal regularization parameter is an intrusive procedure which is not applicable in practice, since knowledge of the clean speech signal and the true RIRs is required to compute the reference signal and the equalized impulse response $\mathbf{c} = \mathbf{H}\mathbf{w}$.

6.4.2 *Robustness increase of acoustic multi-channel equalization techniques when incorporating sparsity-promoting penalty functions*

In this section, the performance of all considered acoustic multi-channel equalization techniques is compared to the performance of their sparsity-promoting counterparts with different penalty functions. An exemplary scenario with NPM = -33 dB is considered.

Robustness increase of the MINT technique

Fig. 6.3 depicts the performance of the MINT and sparsity-promoting MINT techniques with different penalty functions in terms of ΔDRR, EDC, ΔPESQ, and ΔCD. For completeness, the used optimal weighting and penalty parameters are presented in Table 6.2.

The ΔDRR values presented in Fig. 6.3a show that, as expected, the MINT technique fails to suppress the reverberant energy, worsening the DRR by approximately 20 dB in comparison to the true RIR $\mathbf{h}_1$. Furthermore, it can be observed that the l_0- and l_1-norm sparsity-promoting MINT techniques fail to achieve dereverberation and result in a lower DRR than the true RIR $\mathbf{h}_1$. Since acoustic system inversion using the MINT technique is very sensitive to RIR perturbations, incorporating an l_0- or l_1-norm penalty function which does not necessarily preserve the spectro-

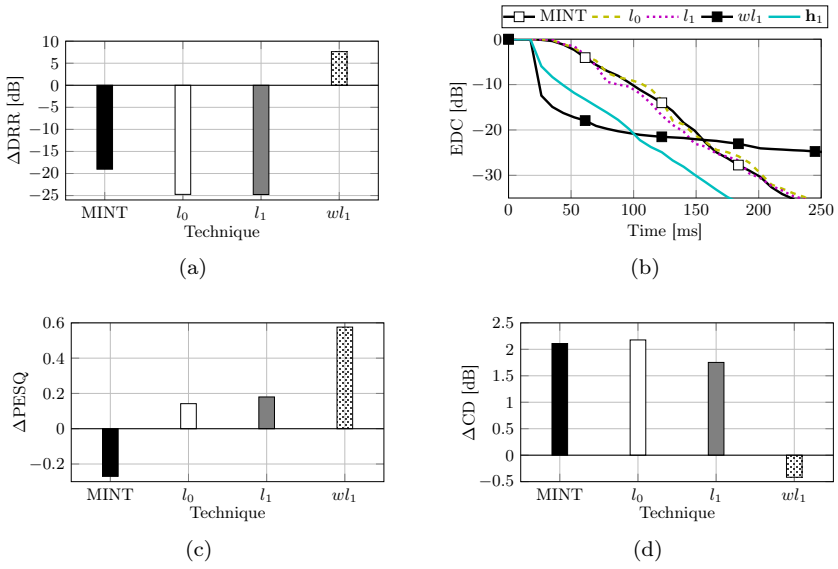


Fig. 6.3: Performance of the MINT and sparsity-promoting MINT techniques with different penalty functions in terms of (a) ΔDRR, (b) EDC, (c) ΔPESQ, and (d) ΔCD (NPM = −33 dB).

Table 6.2: Optimal parameters for the sparsity-promoting MINT technique with different penalty functions (NPM = −33 dB).

Parameter	l_0-norm	l_1-norm	wl_1-norm
η_o	10^{-4}	10^{-5}	10^{-4}
ρ_o	10^{-6}	10^{-5}	10^{-1}

temporal structure of the speech signal is not sufficient to increase the robustness of the MINT technique. On the other hand, by sparsifying the STFT representation of the output speech signal and preserving its spectro-temporal structure using the weighted l_1-norm sparsity-promoting MINT technique, the reverberant energy suppression significantly increases.

To evaluate the reverberant energy decay rate, Fig. 6.3b depicts the energy decay curve of the true RIR $\mathbf{h}_1$ and the energy decay curve of the equalized impulse response $\mathbf{c}$ obtained using the MINT and the sparsity-promoting MINT techniques with different penalty functions. As expected, the MINT technique fails to achieve dereverberation and results in a slower decay rate of the reverberant energy than in the true RIR $\mathbf{h}_1$. Furthermore, also the l_0- and l_1-norm sparsity-promoting MINT techniques fail to achieve dereverberation, yielding a similar decay rate of the reverberant energy as the MINT technique. On the other hand, using the weighted l_1-norm sparsity-promoting MINT technique increases the robustness and slightly improves the decay rate of the reverberant energy in comparison to the true RIR $\mathbf{h}_1$.

The higher reverberant energy suppression achieved by the weighted l_1-norm sparsity-promoting MINT technique is also reflected in the higher perceptual speech quality improvement, as shown by the ΔPESQ and ΔCD values depicted in Figs. 6.3c and 6.3d.

Summarizing, since acoustic system inversion using the MINT technique is very sensitive to RIR perturbations, using the l_0- or l_1-norm sparsity-promoting penalty functions does not suffice to achieve dereverberation. On the other hand, incorporating the weighted l_1-norm penalty function in the MINT technique yields a significant increase in robustness in the presence of RIR perturbations. However, acoustic system inversion using the weighted l_1-norm sparsity-promoting MINT technique nevertheless remains sensitive to RIR perturbations and does not yield a satisfactory dereverberation performance in terms of the decay rate of the reverberant energy (as illustrated in Fig. 6.3b).

Robustness increase of the CS technique

Fig. 6.4 depicts the performance of the CS and sparsity-promoting CS techniques with different penalty functions in terms of ΔDRR, EDC, ΔPESQ, and ΔCD. For completeness, the used optimal weighting and penalty parameters are presented in Table 6.3.

The ΔDRR values presented in Fig. 6.4a show that all proposed sparsity-promoting penalty functions are able to increase the robustness of the CS technique against RIR perturbations and yield a large DRR improvement in comparison to the true RIR $\mathbf{h}_1$. Hence, unlike for the MINT technique, the incorporation of an l_0- or l_1-norm penalty function improves the robustness for the CS technique. This can be explained by the fact that the CS technique, aiming at partial equalization, is in principle more

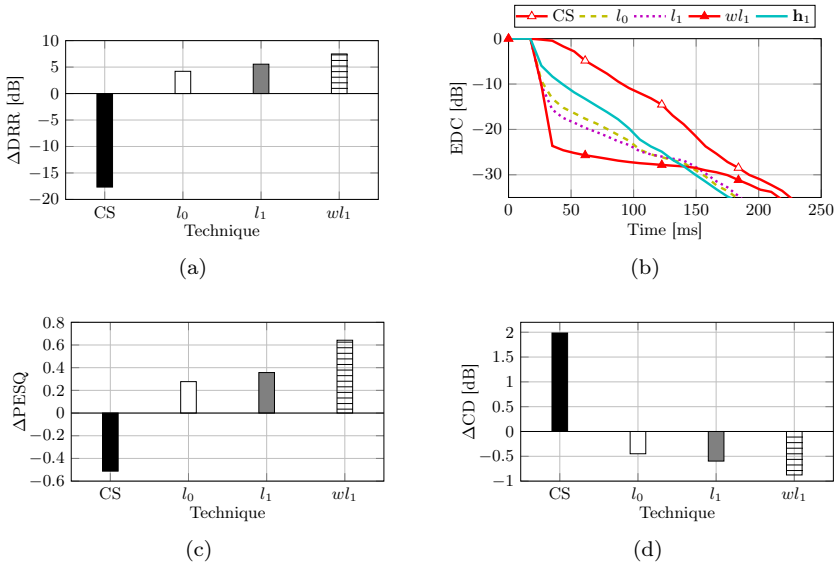


Fig. 6.4: Performance of the CS and sparsity-promoting CS techniques with different penalty functions in terms of (a) ΔDRR, (b) EDC, (c) ΔPESQ, and (d) ΔCD (NPM = −33 dB).

Table 6.3: Optimal parameters for the sparsity-promoting CS technique with different penalty functions (NPM = −33 dB).

Parameter	l_0-norm	l_1-norm	wl_1-norm
η_o	10^{-7}	10^{-4}	10^{-5}
ρ_o	10^{-1}	10^{-1}	10^{-1}

robust than the MINT technique, aiming at complete equalization.[4] Furthermore, it can be observed that out of the proposed penalty functions, the weighted l_1-norm penalty function yields the best ΔDRR.

To evaluate the reverberant energy decay rate, Fig. 6.4b depicts the energy decay curve of the true RIR $\mathbf{h}_1$ and the energy decay curve of the equalized impulse response $\mathbf{c}$ obtained using the CS and the sparsity-promoting CS technique with different penalty functions. It can be observed that any sparsity-promoting penalty

4 The better performance of the CS technique in comparison to the MINT technique is not apparent here. However, as described in Section 3.4.3, selecting the generalized eigenvector as the one yielding the minimum l_2-norm equalized impulse response for the CS technique does not yield the best performance in the presence of RIR perturbations, i.e., significantly better performing generalized eigenvectors can be found.

function results in a faster decay rate of the reverberant energy than in $\mathbf{h}_1$. Furthermore, the weighted l_1-norm penalty function yields the best performance and results in the fastest reverberant energy decay rate.

The improvement in direct-to-reverberant ratio and reverberant energy decay rate achieved by the sparsity-promoting CS techniques is also reflected in the overall perceptual speech quality improvement as evaluated by the ΔPESQ and ΔCD measures depicted in Figs. 6.4c and 6.4d. It can be observed that the weighted l_1-norm penalty function results in the best performance in terms of both measures. The higher perceptual speech quality improvement achieved when using the weighted l_1-norm penalty function arises due to the better preservation of the spectro-temporal structure of the speech signal (cf. Section 6.4.3).

In summary, incorporating the weighted l_1-norm penalty function in the CS technique significantly improves the performance, both in terms of reverberant energy suppression and perceptual speech quality improvement. However, it should be noted that due to the iterative optimization procedure used in each iteration of the ADMM algorithm for computing the filter update in (6.39), the sparsity-promoting CS technique has a very large computational complexity.

Robustness increase of the RMCLS technique

Fig. 6.5 depicts the performance of the RMCLS and sparsity-promoting RMCLS techniques with different penalty functions in terms of ΔDRR, EDC, ΔPESQ, and ΔCD. The used optimal weighting and penalty parameters are presented in Table 6.4.

The ΔDRR values presented in Fig. 6.5a show that, similarly as for the CS technique, all proposed sparsity-promoting penalty functions increase the robustness of the RMCLS technique against RIR perturbations and yield a significantly larger ΔDRR than the RMCLS technique. Furthermore, it can be observed that the weighted l_1-norm penalty function results in a similar ΔDRR as the l_0-norm, whereas a slightly lower performance is obtained when using the l_1-norm.

To evaluate the decay rate of the reverberant energy, Fig. 6.5b depicts the energy decay curve of the true RIR $\mathbf{h}_1$ and the energy decay curve of the equalized impulse response $\mathbf{c}$ obtained using the RMCLS and the sparsity-promoting RMCLS techniques with different penalty functions. It can be observed that for all penalty functions, the sparsity-promoting RMCLS technique results in a slower decay rate of the reverberant energy than the RMCLS technique. This can be explained by the fact that the optimal weighting and penalty parameters are being chosen as the ones yielding the highest perceptual speech quality. Since the RMCLS technique yields a fast reverberant energy decay rate but a low perceptual speech quality (cf. Section 3.4.3), the incorporation of a sparsity-promoting penalty function results in a slightly slower reverberant energy decay rate but a better perceptual speech quality.

The latter is confirmed by the ΔPESQ values presented in Fig. 6.5c, which show that incorporating a sparsity-promoting penalty function in the RMCLS technique significantly improves the perceptual speech quality. Furthermore, it can be observed

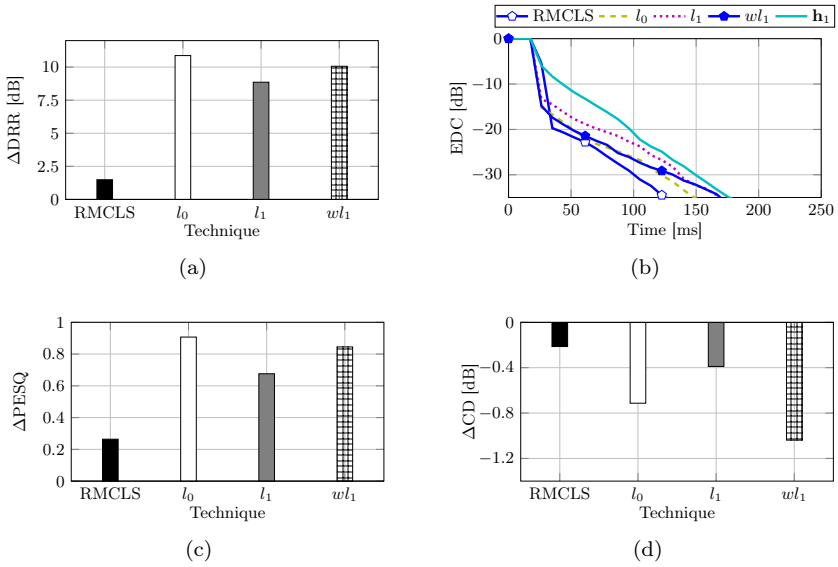


Fig. 6.5: Performance of the RMCLS and sparsity-promoting RMCLS techniques with different penalty functions in terms of (a) ΔDRR, (b) EDC, (c) ΔPESQ, and (d) ΔCD (NPM = −33 dB).

Table 6.4: Optimal parameters for the sparsity-promoting RMCLS technique with different penalty functions (NPM = −33 dB).

Parameter	l_0-norm	l_1-norm	wl_1-norm
η_o	10^{-4}	10^{-5}	10^{-6}
ρ_o	10^{-3}	10^{-4}	10^{-4}

that the l_0- and weighted l_1- norm penalty functions yield a similar perceptual speech quality improvement, whereas a lower performance is obtained when using the l_1-norm penalty function. In addition, the ΔCD values presented in Fig. 6.5d also show that incorporating a sparsity-promoting penalty function in the RMCLS technique improves the perceptual speech quality, with the weighted l_1-norm penalty function yielding the best performance.

In summary, incorporating sparsity-promoting penalty functions in the RMCLS technique significantly increases the robustness against RIR perturbations. Incorporating an l_0- or weighted l_1-norm penalty function yields a similar performance, outperforming the l_1-norm penalty function. Since the l_0-norm is non-convex and hence, there is no guarantee that the global minimum of the l_0-norm sparsity-promoting RMCLS cost function is reached, in the following we will only consider the weighted l_1-norm sparsity-promoting RMCLS technique.

Robustness increase of the PMINT technique

Fig. 6.6 depicts the performance of the PMINT and sparsity-promoting PMINT techniques with different penalty functions in terms of ΔDRR, EDC, ΔPESQ, and ΔCD. For completeness, the used optimal weighting and penalty parameters are presented in Table 6.5.

Similarly as for the CS and the RMCLS techniques, the ΔDRR values and the EDCs depicted in Figs. 6.6a and 6.6b show that all proposed penalty functions increase the robustness of the PMINT technique, yielding a larger ΔDRR and a faster reverberant energy decay rate. Moreover, it can be observed that using the weighted l_1-norm penalty function results in the largest DRR improvement and fastest decay rate of the reverberant energy.

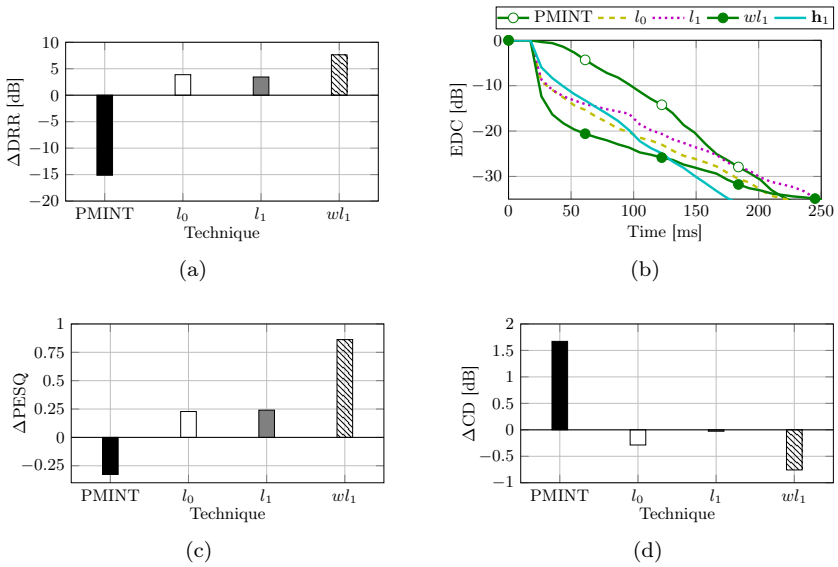


Fig. 6.6: Performance of the PMINT and sparsity-promoting PMINT techniques with different penalty functions in terms of (a) ΔDRR, (b) EDC, (c) ΔPESQ, and (d) ΔCD (NPM = −33 dB).

Table 6.5: Optimal parameters for the sparsity-promoting PMINT technique with different penalty functions (NPM = −33 dB).

Parameter	l_0-norm	l_1-norm	wl_1-norm
η_o	10^{-4}	10^{-5}	10^{-6}
ρ_o	10^{-2}	10^{-2}	10^{-3}

The higher reverberant energy suppression achieved by the weighted l_1-norm sparsity-promoting PMINT technique is also reflected in the higher perceptual speech quality improvement, as shown by the ΔPESQ and ΔCD values depicted in Figs. 6.6c and 6.6d.

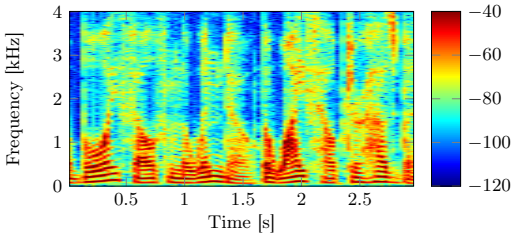
In summary, incorporating sparsity-promoting penalty functions in the PMINT technique increases the robustness against RIR perturbations, with the weighted l_1-norm outperforming the l_0- and l_1-norm penalty functions in terms of reverberant energy suppression and perceptual speech quality improvement.

It should be noted that the performance of the weighted l_1-norm sparsity-promoting techniques obviously depends on the used weighting vector $\mathbf{u}$. When no prior information is available about the sparsity structure of the desired signal, the iteratively re-weighted l_1-norm minimization technique [189] can be used, where the weights are updated in each iteration based on the magnitude of the solution from the previous iteration. For the relatively low reverberation time considered in the simulation results in this chapter, the weights in (6.19) are a rather good approximation of the sparsity structure of the clean speech signal, and hence, iteratively updating the weighting vector does not bring any significant performance improvements.

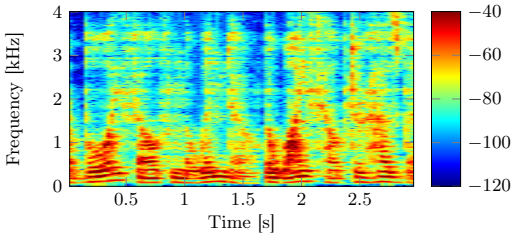
6.4.3 *Discussion*

The previously presented results have shown that incorporating a sparsity-promoting penalty function can significantly increase the robustness of acoustic multichannel equalization techniques against RIR perturbations. It has been validated that the weighted l_1-norm penalty function consistently outperforms the l_1-norm and the l_0-norm penalty functions in terms of reverberant energy suppression and perceptual speech quality improvement (except for the RMCLS technique, where a similar performance to the l_0-norm penalty function is obtained), where the l_1-norm and the l_0-norm penalty functions may even completely fail to achieve dereverberation (as for the MINT technique). These results can be even better illustrated by analyzing the spectrograms of the output speech signals for the different considered penalty functions.

Figs. 6.7a and 6.7b depict the spectrograms of the clean speech and reverberant speech signals for the considered acoustic system. Furthermore, Fig. 6.8 depicts exemplary spectrograms of the output speech signal obtained using the RMCLS and the sparsity-promoting RMCLS techniques with l_0-norm, l_1-norm, and weighted l_1-norm penalty functions.

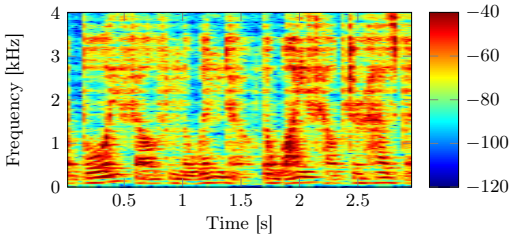


(a)

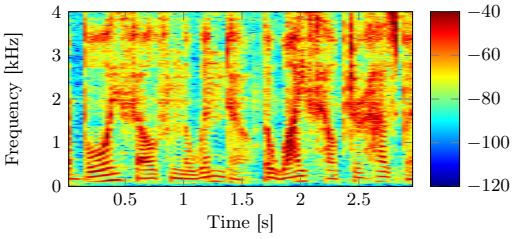


(b)

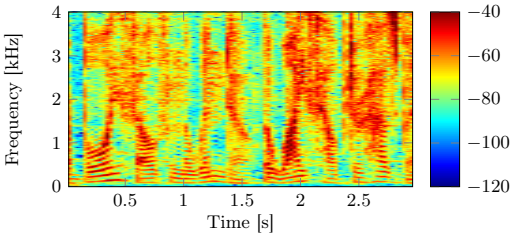
Fig. 6.7: Spectrograms of the (a) clean speech signal and (b) reverberant microphone signal for the considered acoustic system ($T_{60} \approx 360$ ms).



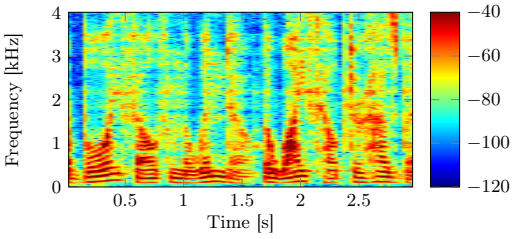
(a)



(b)



(c)



(d)

Fig. 6.8: Exemplary spectrograms of the output speech signal obtained using the (a) RMCLS technique, (b) l_0-norm sparsity-promoting RMCLS technique, (c) l_1-norm sparsity-promoting RMCLS technique, and (d) weighted l_1-norm sparsity-promoting RMCLS technique (NPM = −33 dB).

It can be observed in Fig. 6.8a that due to its sensitivity to RIR perturbations, using the RMCLS technique introduces energy in time-frequency bins where the clean speech signal, and even the reverberant speech signal, does not have any energy. As shown in Figs. 6.8b, 6.8c, and 6.8d, incorporating a sparsity-promoting penalty clearly yields a sparser spectrogram and effectively decreases the energy in these bins. However, Figs. 6.8b and 6.8c also show that while the l_0- and l_1-norm penalty functions yield a sparser spectrogram, particularly for the lower frequencies, the higher frequencies, above approximately 2 kHz, still contain a significant amount of energy. Since the l_0-norm is non-convex, the l_0-norm sparsity-promoting RMCLS reshaping filter does not necessarily correspond to the global minimum of the sparsity-promoting cost function. Furthermore, the l_1-norm sparsity-promoting RMCLS reshaping filter is magnitude-dependent and hence concentrates on sparsifying the time-frequency bins with the largest magnitude. On the other hand, as shown in Fig. 6.8d, the weighted l_1-norm penalty function mainly concentrates on sparsifying the time-frequency bins where the clean speech signal does not have any energy, hence, preserving the spectro-temporal structure of a typical clean speech signal and achieving dereverberation.

6.4.4 *Comparison of the weighted l_1-norm sparsity-promoting acoustic multi-channel equalization techniques*

The simulation results in Section 6.4.2 have shown that for all considered acoustic multi-channel equalization techniques, the weighted l_1-norm penalty function results in a significant increase in robustness against RIR perturbations. In this section the performance of all considered weighted l_1-norm sparsity-promoting techniques is extensively compared for the NPM values in (6.43). Similarly as in Section 3.4.3, the presented performance measures are averaged over all considered NPM values.

To compare the reverberant energy suppression, Figs. 6.9a and 6.9b depict the DRR improvement and the energy decay curve obtained by the weighted l_1-norm sparsity-promoting techniques. It can be observed that all techniques achieve a similar performance in terms of DRR improvement, with the sparsity-promoting RMCLS technique yielding the highest ΔDRR. However, the ΔDRR achieved by all techniques differs by at most 2 dB, which can be considered to be rather insignificant. On the other hand, the decay rate of the reverberant energy depicted in Fig. 6.9b shows more significant differences between the different sparsity-promoting techniques. Since complete equalization using MINT is very sensitive to RIR perturbations, the sparsity-promoting MINT technique yields the slowest reverberant energy decay rate. Due to its energy-based optimization criterion, the sparsity-promoting CS technique yields the highest performance, significantly improving the reverberant energy decay rate in comparison to the true RIR $\mathbf{h}_1$. However, also the sparsity-promoting RMCLS and PMINT techniques yield a faster decay rate of the reverberant energy in comparison to the true RIR $\mathbf{h}_1$.

To compare the perceptual speech quality, Figs. 6.9c and 6.9d depict the PESQ score and cepstral distance improvement obtained by all considered techniques. It can be observed in Fig. 6.9c that the ΔPESQ achieved by all techniques is similar,

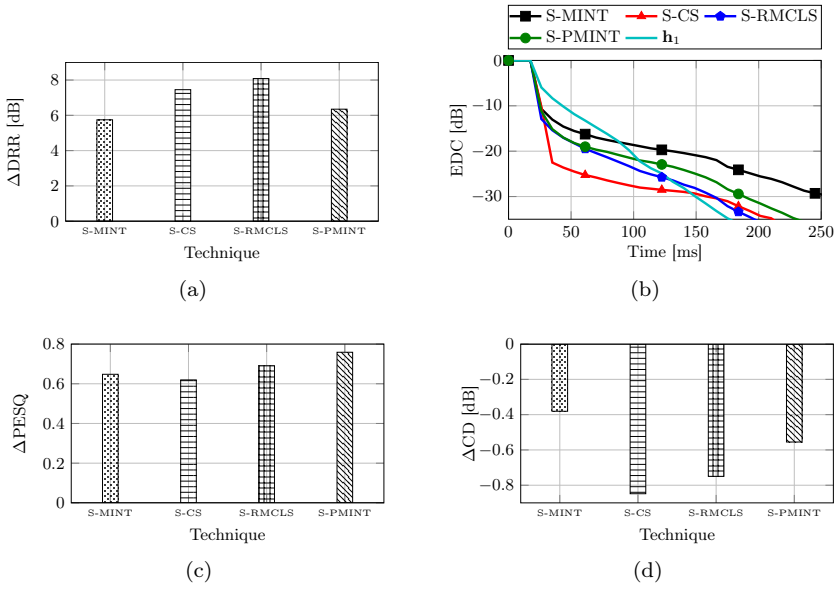


Fig. 6.9: Performance of the weighted l_1-norm sparsity-promoting MINT, CS, RMCLS, and PMINT techniques in terms of (a) ΔDRR, (b) EDC, (c) ΔPESQ, and (d) ΔCD (averaged over several NPM values).

with the sparsity-promoting PMINT technique yielding the best performance. This is to be expected since the sparsity-promoting PMINT technique aims at preserving the direct path and early reflections of the resulting equalized impulse response. Furthermore, Fig. 6.9d also shows that the ΔCD achieved by all techniques is similar, with the sparsity-promoting CS technique yielding the best performance. However, also the sparsity-promoting RMCLS and PMINT techniques result in a good performance, decreasing the cepstral distance in comparison to the reverberant microphone signal. As expected, the sparsity-promoting MINT technique yields the worst performance, nevertheless decreasing the cepstral distance in comparison to the reverberant microphone signal.

From these results it can be concluded that the sparsity-promoting partial equalization techniques yield a high performance in terms of reverberant energy suppression and perceptual speech quality improvement. While the weighted l_1-norm sparsity-promoting CS and RMCLS techniques yield the highest reverberant energy suppression in terms of DRR and EDC, the weighted l_1-norm sparsity-promoting PMINT technique results in the highest perceptual speech quality improvement in terms of PESQ. However, it should be realized that the computational complexity of the sparsity-promoting CS technique is impractically high.

It should be noted that the proposed sparsity-promoting techniques significantly increase the robustness against RIR perturbations only exploiting well known characteristics of clean speech signals, without relying on additional knowledge about

the structure of the RIR perturbations (as the regularized techniques proposed in Chapter 5), which can be considered to be a very good result.

6.5 Summary

In this chapter, we have proposed to increase the robustness of acoustic multi-channel equalization techniques against RIR perturbations using a sparsity-promoting penalty function to promote sparsity in the output speech signal and reduce artifacts generated by non-robust techniques. The least-squares and channel shortening cost functions have been extended with different sparsity-promoting penalty functions, for which iterative algorithms based on the alternating direction method of multipliers have been derived.

Extensive simulation results have shown that incorporating the weighted l_1-norm penalty function in all considered acoustic multi-channel equalization techniques, i.e., MINT, CS, RMCLS, and PMINT, yields a significantly better dereverberation performance in the presence of RIR perturbations. It has been shown that out of the considered techniques, the sparsity-promoting RMCLS and PMINT techniques are computationally tractable techniques which result in the highest dereverberation performance, both in terms of reverberant energy suppression and perceptual speech quality improvement.

It should however be noted that the optimal weighting and penalty parameters for incorporating a sparsity-promoting penalty function have been determined intrusively. An automatic non-intrusive procedure for determining these parameters remains a topic for future investigation.

7

OBJECTIVE AND SUBJECTIVE EVALUATION OF ROBUST ACOUSTIC MULTI-CHANNEL EQUALIZATION TECHNIQUES

In Chapters 4, 5, and 6 we have proposed several signal-independent and signal-dependent methods to increase the robustness of acoustic multi-channel equalization techniques against room impulse response (RIR) perturbations, i.e., using a shorter reshaping filter length to make the optimization criteria better conditioned, incorporating regularization to reduce the energy of distortions due to RIR perturbations, and incorporating a sparsity-promoting penalty function to sparsify the output speech signal and reduce artifacts generated by non-robust techniques. Simulation results using instrumental performance measures in Chapters 4, 5, and 6 have shown that all three methods are effective in increasing the robustness of the considered equalization techniques, i.e., MINT, CS, RMCLS, and PMINT. However, it has also been experimentally validated that complete equalization using the robust extensions of the MINT technique nevertheless remains sensitive to RIR perturbations and does not yield a satisfactory dereverberation performance. Furthermore, it has been experimentally validated that the robust extensions of the CS technique either yield a low perceptual speech quality (the CS technique using shorter reshaping filters and the regularized CS technique) or are computationally very complex (the sparsity-promoting CS technique). On the other hand, the robust extensions of the RMCLS and PMINT techniques appear to achieve a high dereverberation performance, both in terms of reverberant energy suppression and perceptual speech quality improvement. Aiming to determine the most effective method for increasing robustness and the most perceptually advantageous techniques, in this chapter we compare all robust extensions of the RMCLS and PMINT techniques using instrumental performance measures for several acoustic scenarios, i.e., for several acoustic

This chapter is partly based on:

[135] I. Kodrasi, B. Cauchi, S. Goetze, and S. Doclo, "Objective and subjective evaluation of robust acoustic multi-channel equalization," *Journal of the Audio Engineering Society*, 2016, manuscript submitted for publication.

systems and RIR perturbations levels. Since instrumental performance measures do not necessarily correlate well with human perception, we have also conducted a subjective listening test, evaluating the overall speech quality.

In Section 7.1 the considered acoustic systems and the used algorithmic settings are introduced. By means of instrumental performance measures, simulation results in Section 7.2 show that the regularized RMCLS and PMINT techniques yield the highest dereverberation performance, while for some acoustic scenarios, the performance of the sparsity-promoting RMCLS and PMINT techniques is also comparable. The listening test results in Section 7.3 show that the robust extensions of the PMINT technique typically yield a better perceptual speech quality than the robust extensions of the RMCLS technique. While the sparsity-promoting PMINT technique yields the best perceptual speech quality when the level of RIR perturbations is low, the regularized PMINT technique yields the best perceptual speech quality when the level of RIR perturbations is high.

7.1 Acoustic systems and algorithmic settings

We have considered two different acoustic systems with a single speech source and $M = 4$ omni-directional microphones.[1] For each acoustic system, Table 7.1 presents the room reverberation time T_{60}, the direct-to-reverberant ratio DRR, the source-microphone distance d_{sm}, the inter-microphone distance d_{im}, and the length of the room impulse responses L_h at a sampling frequency $f_s = 8$ kHz. To generate the reverberant signals, 10 sentences (approximately 17s) from the HINT database [163] have been convolved with the measured RIRs.

Similarly as in Section 3.4, in order to simulate RIR perturbations, the measured RIRs are perturbed by adding scaled white noise as described in Section 2.2. The considered normalized projection misalignment (NPM) values between the true and the perturbed RIRs are (cf. (2.52))

$$\text{NPM}_1 = -33 \text{ dB} \quad \text{and} \quad \text{NPM}_2 = -15 \text{ dB}. \tag{7.1}$$

The following robust extensions of the RMCLS and PMINT techniques are investigated:

Table 7.1: Characteristics of the considered acoustic systems.

Acoustic system	T_{60} [ms]	DRR [dB]	d_{sm} [m]	d_{im} [m]	L_h
S_1	450	0	3	0.05	3600
S_2	610	−2	2	0.04	4880

1 Note that the first considered acoustic system is the same as in Chapters 3, 4, and 5.

i) L-RMCLS: the RMCLS technique using a shorter reshaping filter length, cf. Section 4.1,
ii) R-RMCLS: the regularized RMCLS technique, cf. Section 5.2,
iii) S-RMCLS: the weighted l_1-norm sparsity-promoting RMCLS technique, cf. Section 6.3,
iv) L-PMINT: the PMINT technique using a shorter reshaping filter length, cf. Section 4.1,
v) R-PMINT: the regularized PMINT technique, cf. Section 5.2,
vi) S-PMINT: the weighted l_1-norm sparsity-promoting PMINT technique, cf. Section 6.3.

For all considered techniques, the conventionally used reshaping filter length is $L_t = \left\lceil \frac{L_h-1}{M-1} \right\rceil = 1200$ for the first acoustic system and $L_t = \left\lceil \frac{L_h-1}{M-1} \right\rceil = 1627$ for the second acoustic system. Furthermore, the delay is set to $\tau = 90$ and the performance for the desired window length $L_d = 10$ ms is investigated. The target equalized impulse response for the robust extensions of the PMINT technique is set to the direct path and the early reflections of the perturbed RIR of the first microphone, i.e., $\hat{\mathbf{h}}_{e,1}$.

The considered reshaping filter lengths L_s for the L-RMCLS and L-PMINT techniques, the considered regularization parameters δ for the R-RMCLS and R-PMINT techniques, and the considered weighting and penalty parameters η and ρ for the S-RMCLS and S-PMINT techniques are

$$L_s \in \{500,\ 600,\ \ldots,\ L_t\}, \tag{7.2}$$

$$\delta \in \{10^{-7},\ 10^{-6},\ \ldots,\ 10^{-1},\ 1,\ 3,\ 5,\ 7,\ 10\}, \tag{7.3}$$

$$\eta \in \{10^{-7},\ 10^{-6},\ 10^{-5},\ 10^{-4}\}, \tag{7.4}$$

$$\rho \in \{10^{-7},\ 10^{-6},\ \ldots,\ 10^{-1}\}. \tag{7.5}$$

Similarly as in Chapters 4, 5, and 6, the optimal reshaping filter length L_o for the L-RMCLS and L-PMINT techniques, the optimal regularization parameter δ_o for the R-RMCLS and R-PMINT techniques, and the optimal weighting and penalty parameters η_o and ρ_o for the S-RMCLS and S-PMINT techniques are determined intrusively as the ones maximizing the PESQ score.

7.2 Objective evaluation

Using the instrumental performance measures described in Section 2.3, the dereverberation performance is evaluated in terms of the reverberant energy suppression and the perceptual speech quality improvement. The reverberant energy suppression is evaluated using the direct-to-reverberant ratio improvement (ΔDRR) between the equalized impulse response $\mathbf{c}$ and the true RIR $\mathbf{h}_1$ (cf. (2.53)). The perceptual speech quality improvement is evaluated using the improvement in PESQ [153] (ΔPESQ) and in cepstral distance [154] (ΔCD) between the output speech signal $z(n)$ and the reverberant microphone signal $x_1(n)$. The reference signal employed for the PESQ and cepstral distance measures is $x_{e,1}(n) = s(n) * h_{e,1}(n)$, i.e., the clean speech signal convolved with the direct path and the early reflections of the first RIR.

Fig. 7.1 presents the DRR improvement for the considered acoustic systems and NPM values. The following conclusions can be drawn by comparing the presented DRR improvement values:

i) The robust extensions of the RMCLS technique generally yield a similar or higher DRR improvement than the robust extensions of the PMINT technique (except for the R-PMINT technique outperforming the R-RMCLS technique for S_1-NPM_2 and the L-PMINT technique outperforming the L-RMCLS technique for S_2-NPM_2).

ii) The R-RMCLS technique typically yields the highest DRR improvement for the considered scenarios (except for the scenario S_1-NPM_2 where the S-RMCLS technique yields the highest DRR improvement).

iii) The R-PMINT and S-PMINT techniques result in a similar DRR improvement.

iv) The L-RMCLS and L-PMINT techniques yield the lowest DRR improvement out of all proposed robust extensions This is not surprising since these techniques simply use a shorter reshaping filter, without explicitly taking into account the structure of the RIR perturbations or the characteristics of the output speech signal.

v) The performance of all considered techniques is generally higher for the first acoustic system than for the second acoustic system. This may be explained by the higher reverberation time of the second acoustic system, leading to a larger number of perturbed RIR taps to be reshaped, and hence, an increased sensitivity of all considered techniques to RIR perturbations.

In order to evaluate the perceptual speech quality obtained by the robust extensions of the RMCLS and PMINT techniques, Figs. 7.2a and 7.2b depict the ΔPESQ and ΔCD values for the considered scenarios. Both instrumental perceptual quality measures show that the regularized techniques yield a similar or higher perceptual speech quality than the other proposed robust extensions, with the R-RMCLS technique yielding a slightly higher PESQ score than the R-PMINT technique and the R-PMINT technique yielding a slightly lower cepstral distance than the R-RMCLS technique. However, for the second acoustic system (i.e., for the scenarios S_2-NPM_1

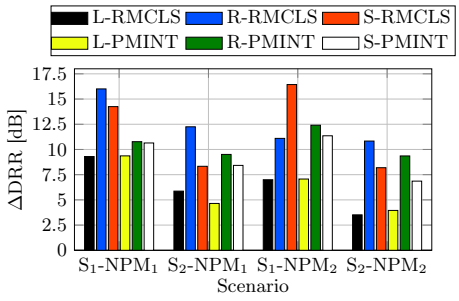


Fig. 7.1: The DRR improvement obtained using the robust extensions of the RMCLS and PMINT techniques.

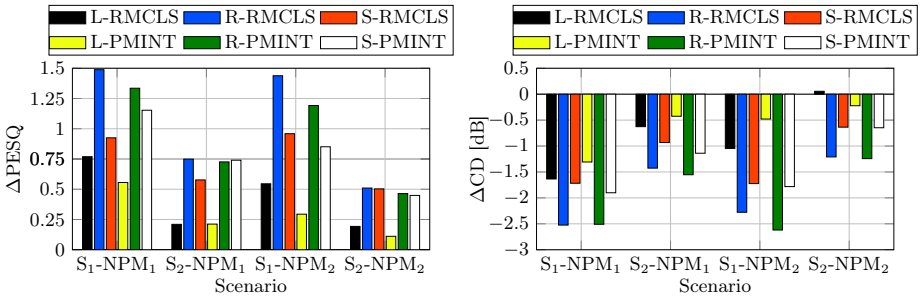


Fig. 7.2: Performance of the robust extensions of the RMCLS and PMINT techniques in terms of (a) ΔPESQ and (b) ΔCD.

and S_2-NPM_2), also the performance of the S-RMCLS and S-PMINT techniques is comparable to the performance of the R-RMCLS and R-PMINT techniques.

Summarizing these simulation results, based on instrumental performance measures it can be said that the regularized RMCLS and PMINT techniques yield the highest performance in terms of reverberant energy suppression and/or perceptual speech quality improvement. In addition, for certain scenarios, the sparsity-promoting RMCLS and PMINT techniques also appear to yield a comparable performance.

7.3 Subjective evaluation

Since instrumental performance measures do not necessarily correlate well with human perception, especially in the context of speech dereverberation, we have also compared the proposed robust extensions using a subjective listening test, evaluating the overall speech quality. The subjective evaluation is based on a multi stimulus test with hidden reference and anchor (MUSHRA) using the specifications in [199]. The same scenarios described in Section 7.1 are considered, i.e., 2 acoustic systems and 2 NPM values. The subjective evaluation is conducted for the reverberant microphone signal $x_1(n)$ and for the output speech signals obtained using the L-RMCLS, R-RMCLS, S-RMCLS, L-PMINT, R-PMINT, and S-PMINT techniques. In addition to these signals, a hidden reference and an anchor are presented to the subjects. The hidden reference is $x_{e,1}(n)$, i.e., the clean speech signal convolved with the direct path and the early reflections of the true RIR. The anchor is the low-pass filtered microphone signal $x_1(n)$ (cut-off frequency of 3 kHz). The signals are diotically presented to the subjects through headphones (Sennheiser HDA 200) at a sampling frequency $f_s = 8$ kHz (using an RME Fireface UFX sound card), with all signals normalized in amplitude.

A total of 13 self-reported normal-hearing subjects who are familiar with speech processing participated in the listening test. The subjects evaluated 2 sentences (approximately 4 s long) for each considered scenario in terms of the attribute “overall

speech quality" on a scale from 0 to 100. Prior to the actual measurements, the subjects were trained to familiarize themselves with the task and the signals under test. Furthermore, they could adjust the sound volume to a comfortable level. The order of presentation of signals and scenarios were randomized between all subjects.

The obtained MUSHRA scores are summarized in Fig. 7.3, where the scores for the different considered scenarios are individually plotted for clarity of presentation. For all considered scenarios, it can generally be observed that the rating variability between subjects (as shown by the whiskers in each boxplot) is rather large. This is commonly the case for subjective listening tests evaluating the overall speech quality achieved by dereverberation algorithms, cf. e.g., [63,200]. Since the artifacts and distortions produced by different techniques are quite different, these artifacts and distortions may be differently judged by different listeners and the perception of these distortions by different subjects is also rather different.

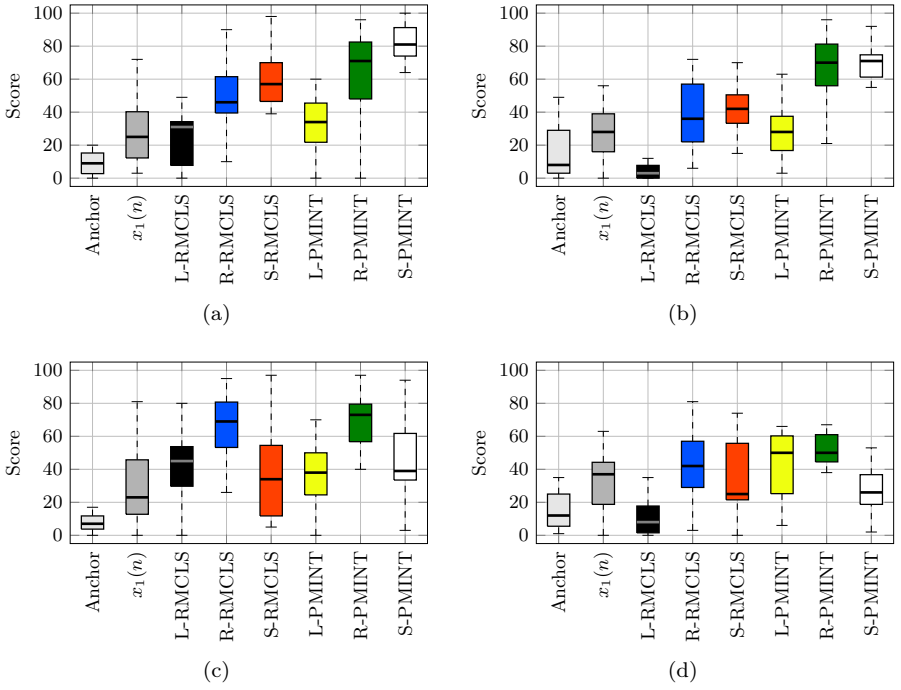


Fig. 7.3: MUSHRA scores for the anchor, reverberant microphone signal $x_1(n)$, and output speech signals obtained using the robust extensions of the RMCLS and PMINT techniques for (a) acoustic system 1 and $\mathrm{NPM}_1 = -33$ dB (S_1-NPM_1), (b) acoustic system 2 and $\mathrm{NPM}_1 = -33$ dB (S_2-NPM_1), (c) acoustic system 1 and $\mathrm{NPM}_2 = -15$ dB (S_1-NPM_2), and (d) acoustic system 2 and $\mathrm{NPM}_2 = -15$ dB (S_2-NPM_2). The scores of the hidden reference, close to 100 with small variance, are not displayed. On each box, the central mark is the median, the edges of the box are the 25-th and 75-th percentiles, and the whiskers extend to 1.5 times the interquartile range from the median.

For the moderate RIR perturbation level $\mathrm{NPM}_1 = -33$ dB, Figs. 7.3a and 7.3b show that the S-PMINT technique yields the highest perceptual speech quality. Furthermore, it can be observed that all proposed techniques typically improve the perceptual speech quality in comparison to the reverberant microphone signal (except for the L-RMCLS technique yielding a worse perceptual speech quality than the reverberant microphone signal for the second acoustic system and the L-PMINT technique yielding a similar perceptual speech quality as the reverberant microphone signal for the second acoustic system). In addition, the robust extensions of the PMINT technique result in a better perceptual speech quality than the robust extensions of the RMCLS technique. Finally, it can be observed that the sparsity-promoting techniques yield the best perceptual speech quality, whereas the techniques using a shorter reshaping filter length result in the worst perceptual speech quality.

As the RIR perturbation level increases to $\mathrm{NPM}_2 = -15$ dB, Figs. 7.3c and 7.3d show that the R-PMINT technique yields the best perceptual speech quality. In addition, it can be observed that while for the first acoustic system all techniques improve the perceptual speech quality in comparison to the reverberant microphone signal, for the second acoustic system, the L-RMCLS, S-RMCLS, and S-PMINT techniques yield a worse perceptual speech quality than the reverberant microphone signal. In addition, it can be seen that the robust extensions of the PMINT technique result in a similar or slightly better perceptual speech quality than the robust extensions of the RMCLS technique (except for the L-PMINT technique yielding a worse perceptual speech quality than the L-RMCLS technique for the first acoustic system). Moreover, it is shown that the regularized techniques yield the best perceptual speech quality, whereas the sparsity-promoting techniques generally result in the worst perceptual speech quality (except for the L-RMCLS technique yielding a worse perceptual speech quality than the S-RMCLS technique for the second acoustic system). Finally, it can be observed that the perceptual speech quality achieved by all techniques is better for the first acoustic system than for the second one. This may be explained by the higher reverberation time of the second acoustic system, leading to a larger number of perturbed RIR taps to be reshaped, and hence, an increased sensitivity of all considered techniques to RIR perturbations.

In summary, in all considered scenarios the trends remain similar, i.e., the robust extensions of the PMINT technique yield a similar or better perceptual speech quality than the robust extensions of the RMCLS technique. Furthermore, the sparsity-promoting PMINT technique results in the best perceptual speech quality for moderate RIR perturbation levels, whereas the regularized PMINT technique results in the best perceptual speech quality for high RIR perturbation levels.

To determine whether the previously discussed results are statistically significant, a statistical analysis is conducted. Since the data is normally distributed as shown by Shapiro-Wilk tests [201], a repeated measures analysis of variance (ANOVA) with the factor "technique" is performed for the different considered scenarios. As summarized in Table 7.2, the statistical analysis shows a significant influence of the factor "technique" for all considered scenarios. To determine the sources of significance, post-hoc tests (with Bonferroni-Holm corrections [202]) using student's t-test are separately conducted for each scenario. The obtained results are presented in

Table 7.2: ANOVA results for the different considered acoustic scenarios.

Scenario	ANOVA result
Acoustic system 1 and $\text{NPM}_1 = -33$ dB (S_1-NPM_1)	$F(7, 84) = 22.3$, $p < 0.001$
Acoustic system 2 and $\text{NPM}_1 = -33$ dB (S_2-NPM_1)	$F(7, 84) = 35.1$, $p < 0.001$
Acoustic system 1 and $\text{NPM}_2 = -15$ dB (S_1-NPM_2)	$F(7, 84) = 20.2$, $p < 0.001$
Acoustic system 2 and $\text{NPM}_2 = -15$ dB (S_2-NPM_2)	$F(7, 84) = 13.5$, $p < 0.001$

Tables 7.3-7.6, with the ticks representing a statistically significant difference, i.e., $p < 0.05$, and the crosses representing no statistically significant difference, i.e., $p \geq 0.05$. The presented results are obviously symmetric across the diagonal since such entries correspond to the same pair comparison.

Table 7.3 shows that for the first acoustic system and the moderate RIR perturbation level $\text{NPM}_1 = -33$ dB, only the S-PMINT technique yields a statistically significant improvement in comparison to the reverberant microphone signal. Furthermore, it can be observed that the S-PMINT technique is statistically significant better than most other techniques (except for the S-RMCLS and R-PMINT techniques). After the S-PMINT technique, the R-PMINT technique appears to yield the most statistically significant improvements in comparison to other techniques.

Table 7.4 shows that for the second acoustic system and the moderate RIR perturbation level $\text{NPM}_1 = -33$ dB, only the R-PMINT and S-PMINT techniques yield a statistically significant improvement in comparison to the reverberant microphone signal. Furthermore, it can be observed that the R-PMINT and S-PMINT techniques are statistically significant better than all other techniques. In addition, the robust extensions of the PMINT technique are statistically significant better than the robust extensions of the RMCLS technique, i.e., the L-PMINT technique is statistically significant better than the L-RMCLS technique, the R-PMINT technique is statistically significant better than the R-RMCLS technique, and the S-PMINT technique is statistically significant better than the S-RMCLS technique.

Table 7.5 shows that for the first acoustic system and the high RIR perturbation level $\text{NPM}_2 = -15$ dB, only the R-RMCLS and R-PMINT techniques are statistically significant better than the reverberant microphone signal. Furthermore, it can be observed that the R-RMCLS and R-PMINT techniques are statistically significant better than all other techniques.

Finally, Table 7.6 shows that for the second acoustic system and the high RIR perturbation level $\text{NPM}_2 = -15$ dB, only the R-PMINT technique is statistically significant better than the reverberant microphone signal. Furthermore, the R-PMINT technique is also statistically significant better than the L-RMCLS and S-PMINT techniques. In addition, the R-RMCLS, S-RMCLS, and L-PMINT techniques appear to be similar and only statistically significant better than the L-RMCLS technique.

Table 7.3: Overview of the student's t-test results for acoustic system 1 and $\mathrm{NPM}_1 = -33$ dB (S_1-NPM_1). The ticks represent a statistically significant difference, i.e., $p < 0.05$, and the crosses represent no statistically significant difference, i.e., $p \geq 0.05$.

	Anchor	$x_1(n)$	L-RMCLS	R-RMCLS	S-RMCLS	L-PMINT	R-PMINT	S-PMINT
Anchor		✘	✘	✔	✔	✔	✔	✔
$x_1(n)$	✘		✘	✘	✘	✘	✘	✔
L-RMCLS	✘	✘		✘	✘	✘	✔	✔
R-RMCLS	✔	✘	✘		✘	✘	✘	✔
S-RMCLS	✔	✘	✘	✘		✔	✘	✘
L-PMINT	✔	✘	✘	✘	✔		✔	✔
R-PMINT	✔	✘	✔	✘	✘	✔		✘
S-PMINT	✔	✔	✔	✔	✘	✔	✘	

Table 7.4: Overview of the student's t-test results for acoustic system 2 and $\mathrm{NPM}_1 = -33$ dB (S_2-NPM_1). The ticks represent a statistically significant difference, i.e., $p < 0.05$, and the crosses represent no statistically significant difference, i.e., $p \geq 0.05$.

	Anchor	$x_1(n)$	L-RMCLS	R-RMCLS	S-RMCLS	L-PMINT	R-PMINT	S-PMINT
Anchor		✘	✘	✘	✔	✘	✔	✔
$x_1(n)$	✘		✔	✘	✘	✘	✔	✔
L-RMCLS	✘	✔		✔	✔	✔	✔	✔
R-RMCLS	✘	✘	✔		✘	✘	✔	✔
S-RMCLS	✔	✘	✔	✘		✘	✔	✔
L-PMINT	✘	✘	✔	✘	✘		✔	✔
R-PMINT	✔	✔	✔	✔	✔	✔		✘
S-PMINT	✔	✔	✔	✔	✔	✔	✘	

Table 7.5: Overview of the student's t-test results for acoustic system 1 and $\mathrm{NPM}_2 = -15$ dB (S_1-NPM_2). The ticks represent a statistically significant difference, i.e., $p < 0.05$, and the crosses represent no statistically significant difference, i.e., $p \geq 0.05$.

	Anchor	$x_1(n)$	L-RMCLS	R-RMCLS	S-RMCLS	L-PMINT	R-PMINT	S-PMINT
Anchor		✘	✔	✔	✘	✘	✔	✔
$x_1(n)$	✘		✘	✔	✘	✘	✔	✘
L-RMCLS	✔	✘		✔	✘	✘	✔	✘
R-RMCLS	✔	✔	✔		✔	✔	✘	✔
S-RMCLS	✘	✘	✘	✔		✘	✔	✘
L-PMINT	✘	✘	✘	✔	✘		✔	✘
R-PMINT	✔	✔	✔	✘	✔	✔		✔
S-PMINT	✔	✘	✘	✔	✘	✘	✔	

Table 7.6: Overview of the student's t-test results for acoustic system 2 and $\text{NPM}_2 = -15$ dB (S_2-NPM_2). The ticks represent a statistically significant difference, i.e., $p < 0.05$, and the crosses represent no statistically significant difference, i.e., $p \geq 0.05$.

	Anchor	$x_1(n)$	L-RMCLS	R-RMCLS	S-RMCLS	L-PMINT	R-PMINT	S-PMINT
Anchor		✔	✘	✔	✔	✔	✔	✘
$x_1(n)$	✔		✔	✘	✘	✘	✔	✘
L-RMCLS	✘	✔		✔	✔	✔	✔	✘
R-RMCLS	✔	✘	✔		✘	✘	✘	✘
S-RMCLS	✔	✘	✔	✘		✘	✘	✘
L-PMINT	✔	✘	✔	✘	✘		✘	✘
R-PMINT	✔	✔	✔	✘	✘	✘		✔
S-PMINT	✘	✘	✘	✘	✘	✘	✔	

In summary, these results confirm the benefit of using the robust extensions of the PMINT technique to improve the perceptual speech quality over the reverberant microphone signal $x_1(n)$, with the S-PMINT or R-PMINT techniques the only techniques that yield a statistically significant improvement over the reverberant microphone signal for all considered scenarios.

Although a full correlation analysis between objective and subjective results is beyond the scope of this thesis, it can be said that the objective evaluation based on instrumental performance measures in Section 7.2 provided a good indication about the performance of the different techniques, i.e., it indicated that the regularized and sparsity-promoting techniques generally outperform the techniques using a shorter reshaping filter length. Furthermore, the objective evaluation results showed that the performance of the considered techniques for the first acoustic system is typically better than for the second acoustic system. Similar conclusions were also deduced from the subjective listening test results. However, the perceptual advantage of the S-PMINT technique over the R-PMINT technique for low RIR perturbation levels as well as the fact that not all techniques improve the perceptual speech quality in comparison to the reverberant microphone signal could have not been directly deduced from the objective evaluation. Therefore, it should be noted that while objective performance measures are a valuable tool when designing speech dereverberation techniques, the impact of acoustic multi-channel equalization techniques can only be truly assessed using subjective listening tests.

7.4 Summary

The objective of this chapter was to determine the most effective method for increasing the robustness of acoustic multi-channel equalization techniques against RIR perturbations as well as to determine the most perceptually advantageous technique. To this end, we conducted an objective and subjective evaluation of all robust extensions of the RMCLS and PMINT techniques proposed in the previous chapters for different scenarios, i.e., for different acoustic systems and NPM values.

Objective evaluation results based on instrumental performance measures showed that the regularized RMCLS and the regularized PMINT techniques typically yield the best reverberant energy suppression and/or perceptual speech quality improvement. Furthermore, they showed that the sparsity-promoting techniques yield a comparable performance for some scenarios, whereas the techniques using a shorter reshaping filter length yield the lowest performance improvement.

These trends were to a certain extent confirmed by the subjective evaluation results based on MUSHRA. In addition, the subjective listening test also showed that the robust extensions of the PMINT technique are generally preferred over the robust extensions of the RMCLS technique (although the statistical significance criterion was not always satisfied). Furthermore, it was shown that the sparsity-promoting or regularized PMINT techniques are the only techniques that yield a statistically significant improvement over the reverberant microphone signal for all considered scenarios. While the sparsity-promoting PMINT technique is the preferred technique when the level of RIR perturbations is low, the regularized PMINT technique is the preferred technique when the level of RIR perturbations is high. Given the robustness and perceptual advantages of the regularized PMINT technique as confirmed by the subjective listening test, in the following chapter dealing with joint dereverberation and noise reduction we will only consider the regularized PMINT technique.

It should be noted that although we did not conduct a formal correlation analysis between the objective and subjective evaluation results, it was observed that while objective performance measures provide useful insights on the performance of dereverberation techniques, the impact of acoustic multi-channel equalization techniques can only be truly assessed using subjective listening tests.

8

JOINT DEREVERBERATION AND NOISE REDUCTION BASED ON ROBUST ACOUSTIC MULTI-CHANNEL EQUALIZATION

As shown in Chapter 7, robust acoustic multi-channel equalization techniques such as the regularized partial multi-channel equalization technique based on the multiple-input/ output inverse theorem (R-PMINT), are able to achieve a high dereverberation performance in the presence of room impulse response perturbations. However, although the R-PMINT technique is able to achieve a high dereverberation performance, it may lead to amplification of the background noise since the actual noise statistics are not taken into account in the reshaping filter design. In this chapter we investigate the effective integration of the dereverberation and noise reduction tasks based on acoustic multi-channel equalization.

In Section 8.1 we propose two time domain techniques aiming at joint dereverberation and noise reduction based on acoustic multi-channel equalization. The first technique, namely R-PMINT for joint dereverberation and noise reduction (RP-DNR), extends the R-PMINT technique by explicitly taking the noise statistics into account. In addition to the regularization parameter used in the R-PMINT technique, the RP-DNR technique introduces another weighting parameter to trade off between dereverberation and noise reduction. The second technique, namely multi-channel Wiener filter (MWF) for joint dereverberation and noise reduction (MWF-DNR), takes both the speech and the noise statistics into account and uses the R-PMINT

This chapter is partly based on:

[136] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multichannel equalization," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Antibes, France, Sep. 2014, pp. 139–143.

[137] I. Kodrasi and S. Doclo, "Incorporating the noise statistics in acoustic multi-channel equalization," in *Proc. AES International Conference on Dereverberation and Reverberation of Audio, Music, and Speech*, Leuven, Belgium, Feb. 2016.

[138] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multi-channel equalization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 680–693, Apr. 2016.

filter to compute a dereverberated reference signal for the MWF. The MWF-DNR technique also introduces an additional weighting parameter, which now provides a trade-off between speech distortion and noise reduction. In Section 8.2, theoretical insights on the difference between the proposed RP-DNR and MWF-DNR techniques are provided. In Section 8.3, we propose automatic non-intrusive procedures for determining the regularization and weighting parameters in the RP-DNR and MWF-DNR techniques. By means of instrumental performance measures, simulation results in Section 8.4 demonstrate that the RP-DNR technique maintains the high dereverberation performance of the R-PMINT technique while improving the noise reduction performance. Furthermore, it is shown that the MWF-DNR technique yields a significantly better noise reduction performance than the RP-DNR technique, at the expense of a worse dereverberation performance.

8.1 Joint dereverberation and noise reduction techniques

As already mentioned in Section 1.1, other than reverberation, background noise is another commonly present source of interference in typical hands-free communication applications. Background noise arises, e.g., due to other speakers, passing traffic, or electronic appliances. For example, the interference of traffic noise with the use of a telephone on a busy street or the interference of a music source with holding a conference call is probably well known to everyone. When the level of the background noise is comparable or larger than the speech level, listening comfort and speech intelligibility are significantly degraded [3, 4]. Furthermore, the performance of acoustic source localization techniques and automatic speech recognition systems also rapidly degrades with increasing background noise levels [1].

When not neglecting the background noise, as done in the previous chapters, the output speech signal of the multi-channel speech enhancement system in Fig. 2.1 is given by

$$z(n) = \mathbf{w}^T\mathbf{x}(n) + \mathbf{w}^T\mathbf{v}(n) = \underbrace{\mathbf{w}^T\mathbf{H}^T}_{\mathbf{c}^T}\mathbf{s}(n) + \mathbf{w}^T\mathbf{v}(n), \tag{8.1}$$

with $\mathbf{w}$ the ML_w–dimensional reshaping filter vector, cf. (2.14), $\mathbf{x}(n)$ the ML_w-dimensional vector of the reverberant speech component, cf. (2.17), $\mathbf{v}(n)$ the ML_w-dimensional vector of the noise component, cf. (2.19), $\mathbf{H}$ the $L_c \times ML_w$–dimensional multi-channel convolution matrix of the true RIRs, cf. (2.24), $\mathbf{s}(n)$ the L_c–dimensional clean speech vector, cf. (2.23), and $\mathbf{c}$ the L_c–dimensional equalized impulse response between the clean speech signal and the output speech signal, cf. (2.27). The acoustic multi-channel equalization techniques discussed in the previous chapters perform speech dereverberation by designing a reshaping filter $\mathbf{w}$ that aims to minimize the error between the equalized impulse response (EIR) $\mathbf{c}$ in (8.1) and a target EIR $\mathbf{c}_t$. Since the background noise $\mathbf{v}(n)$ is completely disregarded, the output noise power is not controlled and may even be amplified compared to the noise power in the microphone signals (cf. Section 8.4.3).

In the following, the $ML_w \times ML_w$–dimensional correlation matrices of the reverberant speech component $\mathbf{x}(n)$, noise component $\mathbf{v}(n)$, and microphone signal $\mathbf{y}(n)$ are defined as

$$\mathbf{R}_{\mathbf{x}}(n) = \mathcal{E}\{\mathbf{x}(n)\mathbf{x}^T(n)\}, \tag{8.2}$$

$$\mathbf{R}_{\mathbf{v}}(n) = \mathcal{E}\{\mathbf{v}(n)\mathbf{v}^T(n)\}, \tag{8.3}$$

$$\mathbf{R}_{\mathbf{y}}(n) = \mathcal{E}\{\mathbf{y}(n)\mathbf{y}^T(n)\}, \tag{8.4}$$

with $\mathcal{E}$ denoting the expected value operator. Since the reverberant speech component $\mathbf{x}(n)$ can be written as (cf. (8.1))

$$\mathbf{x}(n) = \mathbf{H}^T\mathbf{s}(n), \tag{8.5}$$

the reverberant speech component correlation matrix $\mathbf{R}_{\mathbf{x}}(n)$ can also be expressed as

$$\mathbf{R}_{\mathbf{x}}(n) = \mathbf{H}^T\mathbf{R}_{\mathbf{s}}(n)\mathbf{H}, \tag{8.6}$$

with $\mathbf{R}_{\mathbf{s}}(n)$ the $L_c \times L_c$–dimensional clean speech correlation matrix, i.e.,

$$\mathbf{R}_{\mathbf{s}}(n) = \mathcal{E}\{\mathbf{s}(n)\mathbf{s}^T(n)\}. \tag{8.7}$$

Assuming that the speech and the noise components are uncorrelated, the microphone signal correlation matrix $\mathbf{R}_{\mathbf{y}}(n)$ can be written as

$$\mathbf{R}_{\mathbf{y}}(n) = \mathbf{R}_{\mathbf{x}}(n) + \mathbf{R}_{\mathbf{v}}(n). \tag{8.8}$$

In order to simultaneously achieve dereverberation and reduce the output noise power ϵ_v, with

$$\epsilon_v = \mathcal{E}\{[\mathbf{w}^T\mathbf{v}(n)]^2\} = \mathbf{w}^T\mathbf{R}_{\mathbf{v}}(n)\mathbf{w}, \tag{8.9}$$

in the following sections two novel techniques are proposed, namely R-PMINT for joint dereverberation and noise reduction, which takes the noise statistics into account, and multi-channel Wiener filter for joint dereverberation and noise reduction, which takes both the speech and the noise statistics into account. In principle, the proposed techniques can be used to extend any acoustic multi-channel equalization technique that is robust against RIR perturbations and yields a high dereverberation performance. Considering the high and robust dereverberation performance of the R-PMINT technique (as validated in Chapter 7), the techniques proposed in this chapter are discussed as extensions of the R-PMINT technique.

For conciseness, the time index n will be omitted when possible in the remainder of this chapter.

8.1.1 *R-PMINT for joint dereverberation and noise reduction (RP-DNR)*

As discussed in Section 5.2, the R-PMINT cost function and reshaping filter are given by (cf. Tables 5.1 and 5.2)

$$J_{\text{R-P}} = \underbrace{\|\hat{\mathbf{H}}\mathbf{w} - \hat{\mathbf{h}}_{e,p}\|_2^2}_{\epsilon_c} + \delta \underbrace{\mathbf{w}^T\mathbf{R}_e\mathbf{w}}_{\epsilon_e}, \tag{8.10}$$

$$\mathbf{w}_{\text{R-P}} = (\hat{\mathbf{H}}^T\hat{\mathbf{H}} + \delta\mathbf{R}_e)^{-1}\hat{\mathbf{H}}^T\hat{\mathbf{h}}_{e,p}, \tag{8.11}$$

with $\hat{\mathbf{H}}$ the $L_c \times ML_w$–dimensional multi-channel convolution matrix of the perturbed RIRs, cf. (3.4), $\hat{\mathbf{h}}_{e,p}$ the direct path and early reflections of the p-th perturbed RIR, cf. (3.25), $\mathbf{R}_e$ the matrix modeling the RIR perturbations, cf. (5.5), ϵ_c the dereverberation error energy, cf. (5.7), ϵ_e the distortion energy due to RIR perturbations, cf. (5.7), and δ the regularization parameter controlling the trade-off between these two terms.

Aiming at controlling the dereverberation error energy ϵ_c, the distortion energy ϵ_e, as well as the output noise power ϵ_v, we propose to extend the R-PMINT cost function in (8.10) by taking the actual noise statistics explicitly into account. The R-PMINT cost function for joint dereverberation and noise reduction (RP-DNR) is then defined as

$$J_{\text{RP-DNR}} = J_{\text{R-P}} + \mu\epsilon_v \tag{8.12}$$

$$= \underbrace{\|\hat{\mathbf{H}}\mathbf{w} - \hat{\mathbf{h}}_{e,p}\|_2^2}_{\epsilon_c} + \delta \underbrace{\mathbf{w}^T\mathbf{R}_\mathbf{e}\mathbf{w}}_{\epsilon_e} + \mu \underbrace{\mathbf{w}^T\mathbf{R}_\mathbf{v}\mathbf{w}}_{\epsilon_v}, \tag{8.13}$$

with δ a regularization parameter determining the weight given to the distortion energy and μ an additional weighting parameter determining the weight given to the output noise power. In order to compute the filter minimizing (8.13), the gradient of the RP-DNR cost function with respect to $\mathbf{w}$ is set equal to $\mathbf{0}$, i.e.,

$$\frac{\partial J_{\text{RP-DNR}}}{\partial \mathbf{w}} = 2\hat{\mathbf{H}}^T\hat{\mathbf{H}}\mathbf{w} - 2\hat{\mathbf{H}}^T\hat{\mathbf{h}}_{e,p} + 2\delta\mathbf{R}_e\mathbf{w} + \mu\mathbf{R}_\mathbf{v}\mathbf{w} = \mathbf{0}, \tag{8.14}$$

yielding the RP-DNR filter

$$\mathbf{w}_{\text{RP-DNR}} = (\hat{\mathbf{H}}^T\hat{\mathbf{H}} + \delta\mathbf{R}_\mathbf{e} + \mu\mathbf{R}_\mathbf{v})^{-1}\hat{\mathbf{H}}^T\hat{\mathbf{h}}_{e,p}. \tag{8.15}$$

Clearly, the dereverberation and noise reduction performance of the RP-DNR filter in (8.15) depend on the regularization and weighting parameters δ and μ. Increasing the regularization parameter δ results in a higher suppression of the distortion energy at the expense of a higher dereverberation error energy and a larger output noise power, whereas increasing the weighting parameter μ results in a better noise reduction performance at the expense of a worse dereverberation performance (which simultaneously depends on the dereverberation error energy and the distortion energy). While in simulations the optimal values for the parameters δ and μ can be intrusively determined, i.e., using knowledge of the true RIRs and of the true noise statistics, in practice an automatic non-intrusive procedure is required. In Section 8.3 a novel procedure based on the L-hypersurface is proposed for the joint automatic computation of both parameters.

8.1.2 *Multi-channel Wiener filter for joint dereverberation and noise reduction (MWF-DNR)*

The RP-DNR technique proposed in Section 8.1.1 aims at joint dereverberation and noise reduction by considering only the perturbed RIRs and the noise statistics. Taking also the speech statistics into account, we propose a second technique to

achieve joint dereverberation and noise reduction by minimizing the mean-square error between the output speech signal and a dereverberated reference signal s_{ref}, i.e.,

$$J = \mathcal{E}\{(\mathbf{w}^T\mathbf{y} - s_{\text{ref}})^2\}. \tag{8.16}$$

The cost function in (8.16) is the well known MWF cost function [81–87], where the reference signal for the MWF is the dereverberated speech signal. The estimation of several reference signals has been considered for the MWF, e.g., the clean speech signal [81,82,87], the reverberant speech component at an arbitrarily chosen microphone [83,85,86], or a spatially pre-processed reference speech signal [84]. For example, in [87] it has been proposed to use the frequency domain MWF to estimate the output of a superdirective beamformer, such that joint dereverberation and noise reduction is achieved. This technique will be discussed in more detail in Appendix B, where we will also highlight its differences compared with the technique proposed in this section.

Considering the high and robust dereverberation performance of the R-PMINT technique, we propose to use the R-PMINT filter to generate the dereverberated reference signal in (8.16), i.e.,

$$s_{\text{ref}} = \mathbf{w}_{\text{R-P}}^T\mathbf{x} \approx \hat{\mathbf{h}}_{e,p}^T\mathbf{s}. \tag{8.17}$$

Assuming that the speech and the noise components are uncorrelated, the cost function in (8.16) can be decomposed as

$$J = \underbrace{\mathcal{E}\{(\mathbf{w}^T\mathbf{x} - \mathbf{w}_{\text{R-P}}^T\mathbf{x})^2\}}_{\epsilon_x} + \underbrace{\mathcal{E}\{(\mathbf{w}^T\mathbf{v})^2\}}_{\epsilon_v}, \tag{8.18}$$

with ϵ_x denoting the speech distortion, which now refers to the deviation of the output speech component from the dereverberated reference signal $\mathbf{w}_{\text{R-P}}^T\mathbf{x}$. Similarly as in the speech distortion weighted MWF [85], where a weighting parameter μ has been introduced to trade off between speech distortion and noise reduction, the cost function of the proposed multi-channel Wiener filter for joint dereverberation and noise reduction (MWF-DNR) is defined as

$$J_{\text{MWF-DNR}} = \underbrace{\mathcal{E}\{(\mathbf{w}^T\mathbf{x} - \mathbf{w}_{\text{R-P}}^T\mathbf{x})^2\}}_{\epsilon_x} + \mu\underbrace{\mathcal{E}\{(\mathbf{w}^T\mathbf{v})^2\}}_{\epsilon_v}. \tag{8.19}$$

In order to compute the filter minimizing (8.19), the gradient of the MWF-DNR cost function with respect to $\mathbf{w}$ is set equal to $\mathbf{0}$, i.e.,

$$\frac{\partial J_{\text{MWF-DNR}}}{\partial \mathbf{w}} = 2\mathbf{R}_{\mathbf{x}}\mathbf{w} - 2\mathbf{R}_{\mathbf{x}}\mathbf{w}_{\text{R-P}} + 2\mu\mathbf{R}_{\mathbf{v}}\mathbf{w} = \mathbf{0}, \tag{8.20}$$

yielding the MWF-DNR filter

$$\mathbf{w}_{\text{MWF-DNR}} = (\mathbf{R}_{\mathbf{x}} + \mu\mathbf{R}_{\mathbf{v}})^{-1}\mathbf{R}_{\mathbf{x}}\mathbf{w}_{\text{R-P}}. \tag{8.21}$$

Substituting the R-PMINT filter from (8.11) in (8.21), the MWF-DNR filter can be written as

$$\mathbf{w}_{\text{MWF-DNR}} = (\mathbf{R}_{\mathbf{x}} + \mu\mathbf{R}_{\mathbf{v}})^{-1}\mathbf{R}_{\mathbf{x}}(\hat{\mathbf{H}}^T\hat{\mathbf{H}} + \delta\mathbf{R}_{\mathbf{e}})^{-1}\hat{\mathbf{H}}^T\mathbf{c}_t, \tag{8.22}$$

which explicitly shows the dependency of the MWF-DNR filter on the regularization and weighting parameters δ and μ. Clearly, the dereverberation and noise reduction performance of the MWF-DNR filter in (8.22) depends on both parameters. The regularization parameter δ affects the dereverberation performance of the R-PMINT filter $\mathbf{w}_{\text{R-P}}$, hence, also the reference signal $\mathbf{w}_{\text{R-P}}^T\mathbf{x}$ for the MWF-DNR technique. The weighting parameter μ affects the speech distortion ϵ_x (hence, the dereverberation performance of the MWF-DNR filter) as well as the noise reduction performance. While in simulations the optimal values for the parameters δ and μ can be intrusively determined, i.e., using knowledge of the true RIRs and of the true speech and noise statistics, in practice an automatic non-intrusive procedure is required. In Section 8.3 we propose to automatically determine the regularization and weighting parameters δ and μ using two decoupled optimization procedures based on the L-curve method proposed in Section 5.3.

8.2 Insights on the RP-DNR and MWF-DNR techniques

As already mentioned, the performance of the RP-DNR and MWF-DNR techniques depends on the regularization and weighting parameters δ and μ. In this section, analytical insights on the RP-DNR and MWF-DNR techniques for several settings of the regularization and weighting parameters are provided. We distinguish between the following three cases:

i) both the regularization and the weighting parameters are different from 0, i.e., taking into account both the RIR perturbations and the background noise,
ii) the regularization parameter is different from 0 whereas the weighting parameter approaches 0, i.e., disregarding only the background noise, and
iii) both the regularization and the weighting parameter approach 0, i.e., disregarding both the RIR perturbations and the background noise.

Case i) $\delta \neq 0$ and $\mu \neq 0$. When taking into account both the RIR perturbations and the background noise, the main difference between the RP-DNR and MWF-DNR filters in (8.15) and (8.22) consists in the fact that the MWF-DNR filter uses the true reverberant speech component correlation matrix $\mathbf{R}_{\mathbf{x}}$, which implicitly depends on the true convolution matrix $\mathbf{H}$ and on the clean speech correlation matrix $\mathbf{R}_{\mathbf{s}}$, whereas the RP-DNR filter uses only the perturbed convolution matrix $\hat{\mathbf{H}}$. Substituting (8.6) in (8.22), the MWF-DNR filter can be written as

$$\mathbf{w}_{\text{MWF-DNR}} = (\mathbf{H}^T\mathbf{R}_{\mathbf{s}}\mathbf{H} + \mu\mathbf{R}_{\mathbf{v}})^{-1}\mathbf{H}^T\mathbf{R}_{\mathbf{s}}\mathbf{H}(\hat{\mathbf{H}}^T\hat{\mathbf{H}} + \delta\mathbf{R}_{\mathbf{e}})^{-1}\hat{\mathbf{H}}^T\hat{\mathbf{h}}_{e,p}. \tag{8.23}$$

As can be seen in (8.23), unlike the RP-DNR filter, the MWF-DNR filter indirectly incorporates knowledge of the true convolution matrix $\mathbf{H}$ and the clean speech correlation matrix $\mathbf{R}_{\mathbf{s}}$. It can be shown that only when assuming that i) the clean speech signal is uncorrelated, ii) the true RIRs are available, and iii) the regularization parameter δ approaches 0, i.e., $\delta \to 0$, the RP-DNR and MWF-DNR filters are equivalent.

First, assuming that the clean speech signal is uncorrelated, i.e., $\mathbf{R}_{\mathbf{s}} = \sigma_s^2\mathbf{I}$, with σ_s^2 the clean speech variance, the MWF-DNR filter is equal to

$$\mathbf{w}_{\text{MWF-DNR}} = (\mathbf{H}^T\mathbf{H} + \frac{\mu}{\sigma_s^2}\mathbf{R}_{\mathbf{v}})^{-1}\mathbf{H}^T\mathbf{H}(\hat{\mathbf{H}}^T\hat{\mathbf{H}} + \delta\mathbf{R}_{\mathbf{e}})^{-1}\hat{\mathbf{H}}^T\hat{\mathbf{h}}_{e,p}. \tag{8.24}$$

Hence, even for an uncorrelated clean speech signal (which is generally not the case in practice), the MWF-DNR filter in (8.24) differs from the RP-DNR filter in (8.15) by indirectly incorporating the true $\mathbf{H}^T\mathbf{H}$.

Second, assuming that the true RIRs are available, i.e., $\hat{\mathbf{H}} = \mathbf{H}$, the RP-DNR filter in (8.15) and the MWF-DNR filter in (8.24) can be written as

$$\mathbf{w}_{\text{RP-DNR}} = (\mathbf{H}^T\mathbf{H} + \delta\mathbf{R}_{\mathbf{e}} + \mu\mathbf{R}_{\mathbf{v}})^{-1}\mathbf{H}^T\mathbf{h}_{e,p}, \tag{8.25}$$

$$\mathbf{w}_{\text{MWF-DNR}} = (\mathbf{H}^T\mathbf{H} + \frac{\mu}{\sigma_s^2}\mathbf{R}_{\mathbf{v}})^{-1}\mathbf{H}^T\mathbf{H}(\mathbf{H}^T\mathbf{H} + \delta\mathbf{R}_{\mathbf{e}})^{-1}\mathbf{H}^T\mathbf{h}_{e,p}. \tag{8.26}$$

Finally, assuming that the regularization parameter δ approaches 0, i.e., $\delta \to 0$, the RP-DNR filter in (8.25) and the MWF-DNR filter in (8.26) can be written as

$$\mathbf{w}_{\text{RP-DNR}} = (\mathbf{H}^T\mathbf{H} + \mu\mathbf{R}_{\mathbf{v}})^{-1}\mathbf{H}^T\mathbf{h}_{e,p}, \tag{8.27}$$

$$\mathbf{w}_{\text{MWF-DNR}} = (\mathbf{H}^T\mathbf{H} + \frac{\mu}{\sigma_s^2}\mathbf{R}_{\mathbf{v}})^{-1}\mathbf{H}^T\mathbf{h}_{e,p}, \tag{8.28}$$

where (8.28) is derived from (8.26) using $\lim_{\delta\to 0}(\mathbf{H}^T\mathbf{H} + \delta\mathbf{R}_{\mathbf{e}})^{-1}\mathbf{H}^T\mathbf{h}_{e,p} = \mathbf{H}^+\mathbf{h}_{e,p}$ (cf. Section 5.2). Comparing (8.27) and (8.28), it can be observed that under the assumptions of an uncorrelated clean speech signal, knowledge of the true RIRs, and $\delta \to 0$, the RP-DNR and MWF-DNR filters are equivalent (up to the scaling of the weighting parameter μ by the clean speech variance σ_s^2). However, obviously in practice the clean speech signal is correlated, i.e., $\mathbf{R}_{\mathbf{s}} \neq \sigma_s^2\mathbf{I}$, and most importantly, the true RIRs are not known. Hence, as is experimentally validated in Section 8.4.4, by incorporating the true speech statistics $\mathbf{R}_{\mathbf{x}}$ in the MWF-DNR technique, the noise reduction and the overall joint dereverberation and noise reduction performance is significantly improved in comparison to the RP-DNR technique. The importance of incorporating the true correlation matrix $\mathbf{R}_{\mathbf{x}}$ is further validated in Section 8.4.4 by the performance degradation of the MWF-DNR technique in the presence of speech correlation matrix estimation errors.

Case ii) $\delta \neq 0$ *and* $\mu \to 0$. As the weighting parameter μ approaches 0, i.e., disregarding the background noise but taking into account the RIR perturbations, the RP-DNR filter in (8.15) is equal to the R-PMINT filter in (8.11), i.e.,

$$\lim_{\mu\to 0} \mathbf{w}_{\text{RP-DNR}} = \mathbf{w}_{\text{R-P}}. \tag{8.29}$$

Hence, using a small value for the weighting parameter μ in the RP-DNR technique will result in a similar performance as the R-PMINT technique.

Similarly, assuming a full-rank reverberant speech component correlation matrix $\mathbf{R}_{\mathbf{x}}$, as the weighting parameter μ approaches 0, the MWF-DNR filter in (8.22) is also equal to the R-PMINT filter in (8.11), i.e.,

$$\lim_{\mu\to 0} \mathbf{w}_{\text{MWF-DNR}} = \mathbf{w}_{\text{R-P}}. \tag{8.30}$$

However, the reverberant speech component correlation matrix $\mathbf{R}_{\mathbf{x}}$ in (8.6) is typically ill-conditioned, due to the commonly-occurring rank deficiency of the clean speech correlation matrix $\mathbf{R}_{\mathbf{s}}$ and due to the fact that the multi-channel convolution matrix $\mathbf{H}$ is a full row-rank matrix. For a rank-deficient $\mathbf{R}_{\mathbf{x}}$

$$\lim_{\mu\to 0} \mathbf{w}_{\text{MWF-DNR}} = \mathbf{R}_{\mathbf{x}}^{+}\mathbf{R}_{\mathbf{x}}\mathbf{w}_{\text{R-P}} \tag{8.31}$$

$$= [w_{\text{R-P}}(0)\ w_{\text{R-P}}(1)\ \ldots\ w_{\text{R-P}}(r-1)\ 0\ \ldots\ 0]^{T}, \tag{8.32}$$

with r the rank of the reverberant speech component correlation matrix $\mathbf{R}_{\mathbf{x}}$.

Hence, when disregarding the background noise but taking into account the RIR perturbations, the RP-DNR filter results in a similar performance as the R-PMINT filter, whereas the MWF-DNR filter yields a slightly different performance from the R-PMINT filter assuming $\mathbf{R}_{\mathbf{x}}$ is rank-deficient.

Case iii) $\delta \to 0$ and $\mu \to 0$. As discussed in Case ii), as the weighting parameter μ approaches 0, the RP-DNR filter in (8.15) is equal to the R-PMINT filter in (8.11). Furthermore, it was shown in Section 5.2 that as the regularization parameter δ approaches 0, the R-PMINT filter is equal to the PMINT filter. Therefore, as the regularization and weighting parameters δ and μ approach 0, i.e., disregarding the RIR perturbations and the background noise, the RP-DNR filter in (8.15) is equal to the PMINT filter in (3.27), i.e.,

$$\lim_{\substack{\delta\to 0\\ \mu\to 0}} \mathbf{w}_{\text{RP-DNR}} = \mathbf{w}_{\text{P}}. \tag{8.33}$$

Hence, using small values for the regularization and the weighting parameters in the RP-DNR technique will result in a similar performance as the PMINT technique, i.e., a high sensitivity to RIR perturbations and noise amplification.

Similarly, assuming a full-rank speech correlation matrix $\mathbf{R}_{\mathbf{x}}$, as the regularization and weighting parameters δ and μ approach 0, the MWF-DNR filter in (8.22) is also equal to the PMINT filter in (3.27), i.e.,

$$\lim_{\substack{\delta\to 0\\ \mu\to 0}} \mathbf{w}_{\text{MWF-DNR}} = \mathbf{w}_{\text{P}}, \tag{8.34}$$

whereas for a rank-deficient speech correlation matrix $\mathbf{R}_{\mathbf{x}}$ of rank r, the minimum l_2-norm MWF-DNR filter is equal to the first r coefficients of the PMINT filter, i.e.,

$$\lim_{\substack{\delta\to 0\\ \mu\to 0}} \mathbf{w}_{\text{MWF-DNR}} = \mathbf{R}_{\mathbf{x}}^{+}\mathbf{R}_{\mathbf{x}}\mathbf{w}_{\text{P}} \tag{8.35}$$

$$= [w_{\text{P}}(0)\ w_{\text{P}}(1)\ \ldots\ w_{\text{P}}(r-1)\ 0\ \ldots\ 0]^{T}. \tag{8.36}$$

Hence, when disregarding both the RIR perturbations and the background noise, the RP-DNR filter results in a similar performance as the PMINT filter, whereas the MWF-DNR filter yields a slightly different performance from the PMINT filter assuming $\mathbf{R}_{\mathbf{x}}$ is rank-deficient.

8.3 Automatic regularization and weighting parameters

The optimal value of the regularization and weighting parameters in the RP-DNR and MWF-DNR techniques depends on the acoustic system, the RIR perturbations, the background noise, as well as on what is more important for the considered application, i.e., dereverberation or noise reduction. While in simulations these parameters can be determined intrusively, i.e., using knowledge of the true RIRs and of the speech and noise statistics, in practice an automatic non-intrusive procedure is required. In Section 5.3 an automatic non-intrusive procedure based on the L-curve has been proposed for determining the regularization parameter in the R-PMINT technique. In Section 8.3.2 we extend this procedure to the automatic computation of the regularization and weighting parameters in the MWF-DNR technique. Furthermore, in Section 8.3.1 a novel procedure based on the L-hypersurface is proposed for the joint automatic computation of both parameters in the RP-DNR technique.

8.3.1 *Automatic regularization and weighting parameters in the RP-DNR technique*

Different regularization and weighting parameters δ and μ obviously result in different RP-DNR filters in (8.15), yielding different dereverberation error energy ϵ_c, distortion energy ϵ_e, and output noise power ϵ_v, with

$$\epsilon_c = \|\hat{\mathbf{H}}\mathbf{w}_{\text{RP-DNR}} - \mathbf{c}_t\|_2^2, \tag{8.37}$$

$$\epsilon_e = \mathbf{w}_{\text{RP-DNR}}^T \mathbf{R}_e \mathbf{w}_{\text{RP-DNR}}, \tag{8.38}$$

$$\epsilon_v = \mathbf{w}_{\text{RP-DNR}}^T \mathbf{R}_{\mathbf{v}} \mathbf{w}_{\text{RP-DNR}}. \tag{8.39}$$

Similarly as for the regularization parameter in the R-PMINT technique, appropriate parameters δ and μ in the RP-DNR technique should incorporate knowledge about the dereverberation error energy, the distortion energy, and the output noise power, such that all three terms are low. Motivated by the simplicity and the applicability of the L-curve for regularizing least-squares techniques [170], the so-called L-hypersurface has been proposed in [203] as a multi-parameter generalization of the L-curve. Hence, similarly to the L-curve procedure where the optimal parameter is computed as the point of maximum curvature, we propose to compute the regularization and weighting parameters δ and μ as the point of maximum Gaussian curvature of the L-hypersurface, obtained by plotting the output noise power ϵ_v versus the dereverberation error energy ϵ_c and the distortion energy ϵ_e for several parameters δ and μ.

Fig. 8.1 depicts an exemplary L-hypersurface obtained by plotting ϵ_v versus ϵ_c and ϵ_e for several regularization and weighting parameters δ and μ for the RP-DNR technique. The point of maximum Gaussian curvature of the L-hypersurface is also depicted. Although the Gaussian curvature of a surface can be analytically computed, numerical inaccuracies due to the manipulation of typically large-dimensional matrices can occur when maximizing the curvature [204] (cf. Section 5.3), such that a numerically stable algorithm is required. In this chapter, the minimum distance

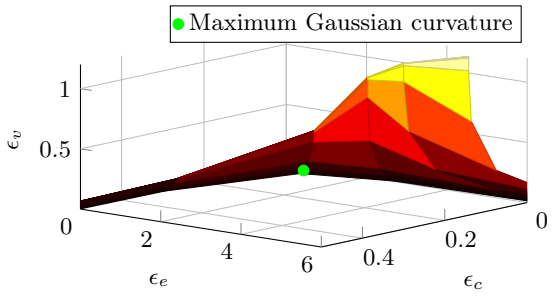


Fig. 8.1: Exemplary parametric surface of the output noise power ϵ_v versus the dereverberation error energy ϵ_c and the distortion energy ϵ_e for the RP-DNR technique, with the regularization and weighting parameters δ and μ ranging from 10^{-7} to 10.

method proposed in [204] will be used to compute the point of maximum Gaussian curvature.

8.3.2 *Automatic regularization and weighting parameter in the MWF-DNR technique*

Also for the MWF-DNR technique, different regularization and weighting parameters δ and μ result in different MWF-DNR filters in (8.22), yielding different dereverberation error energy ϵ_c, distortion energy ϵ_e, speech distortion ϵ_x, and output noise power ϵ_v. To automatically determine the regularization and weighting parameters δ and μ for the MWF-DNR technique, we propose to use two decoupled optimization procedures based on the L-curve as described in the following.

In order to obtain a dereverberated reference signal $\mathbf{w}_{\text{R-P}}^T\mathbf{x}$, first the parameter δ is automatically computed using the L-curve procedure proposed in Section 5.3 for determining the regularization parameter in the R-PMINT technique.
Second, for the determined regularization parameter δ, i.e., for a fixed filter $\mathbf{w}_{\text{R-P}}$, changing the weighting parameter μ in the MWF-DNR technique yields a different speech distortion ϵ_x and output noise power ϵ_v, i.e.,

$$\epsilon_x = \mathbf{w}_{\text{MWF-DNR}}^T\mathbf{R}_{\mathbf{x}}\mathbf{w}_{\text{MWF-DNR}} - 2\mathbf{w}_{\text{MWF-DNR}}^T\mathbf{R}_{\mathbf{x}}\mathbf{w}_{\text{R-P}} + \mathbf{w}_{\text{R-P}}^T\mathbf{R}_{\mathbf{x}}\mathbf{w}_{\text{R-P}}, \tag{8.40}$$

$$\epsilon_v = \mathbf{w}_{\text{MWF-DNR}}^T\mathbf{R}_{\mathbf{v}}\mathbf{w}_{\text{MWF-DNR}}, \tag{8.41}$$

with (8.40) derived by expanding $\epsilon_x = \mathcal{E}\{(\mathbf{w}^T\mathbf{x} - \mathbf{w}_{\text{R-P}}^T\mathbf{x})^2\}$ from (8.18). An appropriate weighting parameter μ in the MWF-DNR technique should incorporate knowledge about both the speech distortion and the output noise power, such that both terms are small. Fig. 8.2 depicts an exemplary parametric plot of the output noise power versus the speech distortion for a set of parameters μ. This parametric plot has an L-shape, with the point of maximum curvature, i.e., the corner of the L-curve, located where the MWF-DNR filter changes from being dominated by

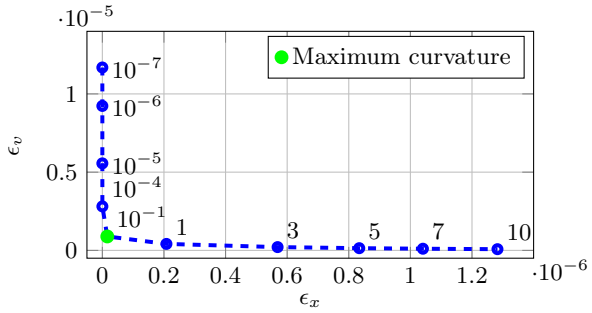


Fig. 8.2: Exemplary parametric plot of the output noise power ϵ_v versus the speech distortion ϵ_x for the MWF-DNR technique, with the weighting parameter μ ranging from 10^{-7} to 10.

large speech distortion to being dominated by large output noise power. Hence, we propose to compute the weighting parameter μ in the MWF-DNR technique as the point of maximum curvature of this parametric plot (i.e., $\mu = 10^{-1}$ in the depicted example). Similarly as for the computation of the regularization parameter in the R-PMINT technique, the curvature of the parametric plot of ϵ_v versus ϵ_x can be analytically computed and a one-dimensional optimization routine can be used to maximize the curvature. However, numerical problems can occur due to the manipulation of the typically large-dimensional matrices $\mathbf{R_x}$ and $\mathbf{R_v}$. Therefore, as in Chapter 5, in this chapter the numerically stable triangle method proposed in [172] will be used to determine the point of maximum curvature of the L-curve.

8.4 Simulations

In this section we investigate the dereverberation and noise reduction performance of the proposed RP-DNR and MWF-DNR techniques. In Section 8.4.1 the considered acoustic system and the used algorithmic settings are introduced. In Section 8.4.2 the influence of the regularization and weighting parameters on the performance of the RP-DNR and MWF-DNR techniques is investigated. In Section 8.4.3, the automatically parametrized RP-DNR and MWF-DNR techniques are compared to acoustic multi-channel equalization techniques, i.e., to the PMINT and the automatically regularized PMINT technique. Finally, in Section 8.4.4 the performance of the automatically parametrized RP-DNR and MWF-DNR techniques is extensively investigated for different noise levels, RIR perturbations, and correlation matrix estimation errors.

8.4.1 *Acoustic system and algorithmic settings*

We have considered an acoustic system with a single speech source and $M = 4$ omni-directional microphones. The speech source is placed in broadside direction at a distance of 2 m from the microphone array. The distance between the microphones

is 4 cm, the room reverberation time is $T_{60} \approx 610$ ms, and the direct-to-reverberant ratio is DRR $= -2$ dB [198]. The RIRs have been measured using the swept-sine technique [162] and the length of the RIRs has been set to $L_h = 4880$ at a sampling frequency $f_s = 8$ kHz.

To generate the reverberant speech components, 10 sentences from the HINT database [163] have been convolved with the measured RIRs. The noise consists of a directional interference and spatially diffuse noise. The interference is located in endfire direction at a distance of 2 m from the microphones. The spatially diffuse noise is simulated using [139]. The broadband input speech-to-diffuse noise ratio is set to 10 dB and the broadband input speech-to-interference ratio (SIR) is varied between -5 dB and 10 dB. The speech-plus-noise signal is approximately 17 s long and is preceded by a 7 s long noise-only signal, which is not taken into account during the evaluation.

Similarly as in Section 3.4, in order to simulate RIR perturbations, the measured RIRs are perturbed by adding scaled white noise as described in Section 2.2. The considered normalized projection misalignment (NPM) values between the true and the perturbed RIRs are (cf. (2.52))

$$\text{NPM} \in \{-33 \text{ dB}, -27 \text{ dB}, -21 \text{ dB}, -15 \text{ dB}\}. \tag{8.42}$$

For all considered techniques, the reshaping filter length is $L_w = \left\lceil \frac{L_h - 1}{M - 1} \right\rceil = 1672$, the delay is set to $\tau = 90$, and the performance for the desired window length $L_d = 10$ ms is investigated. Furthermore, the target equalized impulse response for the PMINT, R-PMINT, RP-DNR, and MWF-DNR techniques is set to the direct path and early reflections of the perturbed RIR of the first microphone, i.e., $\hat{\mathbf{h}}_{e,1}$. Similarly as in Section 5.5, for the distortion energy term in the R-PMINT and RP-DNR techniques we have assumed that $\mathbf{R}_e = \mathbf{I}$.

Furthermore, the speech and noise correlation matrices are computed as follows:

i) perfectly estimated from the speech and noise signals in order to evaluate the full potential of the proposed techniques by avoiding correlation matrix estimation errors (Sections 8.4.2 and 8.4.3), i.e.,

$$\mathbf{R}_{\mathbf{x}} = \frac{1}{L} \sum_{l=1}^{L} \mathbf{x}_l \mathbf{x}_l^T, \qquad \mathbf{R}_{\mathbf{v}} = \frac{1}{L} \sum_{l=1}^{L} \mathbf{v}_l \mathbf{v}_l^T, \tag{8.43}$$

with L denoting the number of available speech-plus-noise signal vectors.

ii) erroneously estimated as $\mathbf{R}_{\mathbf{x}} = \mathbf{R}_{\mathbf{y}} - \mathbf{R}_{\mathbf{v}}$, with $\mathbf{R}_{\mathbf{y}}$ estimated during the speech-plus-noise period and $\mathbf{R}_{\mathbf{v}}$ estimated during the noise-only period in order to achieve a more realistic evaluation of the proposed techniques (Section 8.4.4), i.e.,

$$\mathbf{R}_{\mathbf{y}} = \frac{1}{L} \sum_{l=1}^{L} \mathbf{y}_l \mathbf{y}_l^T, \qquad \mathbf{R}_{\mathbf{v}} = \frac{1}{L_v} \sum_{l=1}^{L_v} \mathbf{v}_l \mathbf{v}_l^T, \qquad \mathbf{R}_{\mathbf{x}} = \mathbf{R}_{\mathbf{y}} - \mathbf{R}_{\mathbf{v}}, \tag{8.44}$$

with L_v denoting the number of available noise-only signal vectors. Due to the fact that the speech and noise signals are not perfectly uncorrelated and

the noise is nonstationary, computing the speech correlation matrix as $\mathbf{R}_{\mathbf{x}} = \mathbf{R}_{\mathbf{y}} - \mathbf{R}_{\mathbf{v}}$ may not yield a positive semi-definite matrix, particularly at low input SIR. The estimated $\mathbf{R}_{\mathbf{x}}$ is therefore forced to be a positive semi-definite matrix by computing its eigenvalue decomposition and setting the negative eigenvalues equal to 0.

Using the instrumental performance measures described in Section 2.3, the dereverberation performance is evaluated in terms of the reverberant energy suppression and the perceptual speech quality improvement. The reverberant energy suppression is evaluated using the direct-to-reverberant ratio improvement (ΔDRR) between the equalized impulse response $\mathbf{c}$ and the true RIR $\mathbf{h}_1$ (cf. (2.53)). The improvement in perceptual speech quality is evaluated using the improvement in PESQ [153] (ΔPESQ) between the *output speech component* $z_x(n)$ and the *reverberant speech component* $x_1(n)$. The reference signal employed for the PESQ measure is $x_{e,1}(n) = s(n) * h_{e,1}(n)$, i.e., the clean speech signal convolved with the direct path and early reflections of the first RIR. Furthermore, the noise reduction performance is evaluated in terms of the noise reduction factor ψ_{NR}, (cf. (2.57)). The joint dereverberation and noise reduction performance is evaluated in terms of the improvement in signal-to-reverberation-and-noise ratio (ΔSRNR) between the output speech signal $z(n)$ and the first microphone signal $y_1(n)$ (cf. (2.58)). In order to evaluate the improvement in overall perceptual quality, we use the improvement in frequency-weighted segmental SNR (ΔfwSSNR) [155] between the output speech signal $z(n)$ and the first microphone signal $y_1(n)$, with $x_{e,1}(n)$ as the reference signal.

8.4.2 *Influence of the regularization and weighting parameters on the performance of the RP-DNR and MWF-DNR techniques*

In this section, the influence of the regularization and weighting parameters δ and μ on the performance of the RP-DNR and MWF-DNR techniques is investigated for an exemplary scenario of SIR = 0 dB and NPM = −33 dB. The considered regularization and weighting parameter values are

$$\delta \in \{10^{-7},\ 10^{-6},\ \ldots,\ 10^{-1},\ 1,\ 3,\ 5,\ 7,\ 10\}, \tag{8.45}$$

$$\mu \in \{10^{-7},\ 10^{-6},\ \ldots,\ 10^{-1},\ 1,\ 3,\ 5,\ 7,\ 10\}, \tag{8.46}$$

and the speech and noise correlation matrices are perfectly estimated from the speech and noise signals as in (8.43).

Figs. 8.3a and 8.3b depict the DRR improvement and the noise reduction factor for the RP-DNR technique. It can be observed that for small values of the regularization and weighting parameters δ and μ (e.g., $\delta = 10^{-7}$ and $\mu = 10^{-7}$), the dereverberation performance is high whereas the background noise is amplified. As expected, since the RIR perturbation level is relatively low, i.e., NPM = −33 dB, also the optimal value of the regularization parameter δ required for a high dereverberation performance is small (e.g., $\delta = 10^{-7}$). In addition, a small value of the

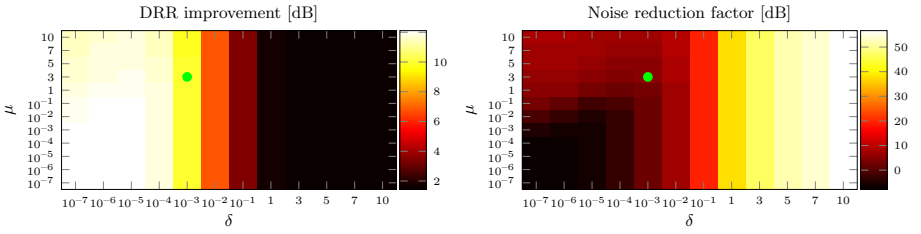


Fig. 8.3: Performance of the RP-DNR technique for different regularization and weighting parameters δ and μ in terms of (a) ΔDRR and (b) ψ_{NR}. The circles denote the automatically determined regularization and weighting parameters (NPM = -33 dB, SIR = 0 dB).

weighting parameter μ (e.g., $\mu = 10^{-7}$), i.e., (almost) disregarding the background noise, leads to noise amplification. Furthermore, it can be observed in Fig. 8.3a that for a fixed value of the weighting parameter μ (e.g., $\mu = 10^{-5}$), increasing the regularization parameter δ initially yields a slight increase in ΔDRR (not visible in Fig. 8.3a), however, as the regularization parameter δ is increased beyond 10^{-5}, the ΔDRR values decrease. This is to be expected since as already mentioned, for a relatively low RIR perturbation level, i.e., NPM = -33 dB, the optimal value of the regularization parameter δ required for a high dereverberation performance is small. Furthermore, it can be observed in Fig. 8.3b that for a fixed value of the weighting parameter μ (e.g., $\mu = 10^{-5}$), increasing the regularization parameter δ also increases the noise reduction factor. This can be explained by the fact that for increasing values of the regularization parameter δ the energy of the resulting RP-DNR filter decreases (since $\delta\mathbf{I}$ is used as the regularization term), which results in a smaller output noise power. As expected, for a fixed value of the regularization parameter δ (e.g., $\delta = 10^{-5}$), increasing the weighting parameter μ results in a trade-off between dereverberation and noise reduction performance, as can be seen by the decrease in DRR improvement and the increase in noise reduction factor. However, for large values of the regularization parameter δ (e.g., $\delta = 1$), increasing the weighting parameter μ hardly has any effect on the dereverberation or the noise reduction performance, since the resulting RP-DNR filter has very low energy.

For the considered scenario, the procedure proposed in Section 8.3.1 for automatically determining the regularization and weighting parameters based on the L-hypersurface yields $\delta = 10^{-3}$ and $\mu = 3$, which are denoted by the circles in Figs. 8.3a and 8.3b. While it is not possible to judge upon the optimality of a set of parameters, it can be observed that the automatic procedure yields parameters which result in a reasonable trade-off between dereverberation and noise reduction performance. This is also confirmed in Section 8.4.4 for other NPM and SIR values.

Figs. 8.4a and 8.4b depict the DRR improvement and the noise reduction factor for the MWF-DNR technique. Similarly as for the RP-DNR technique, it can be observed that for small values of the regularization and weighting parameters δ and

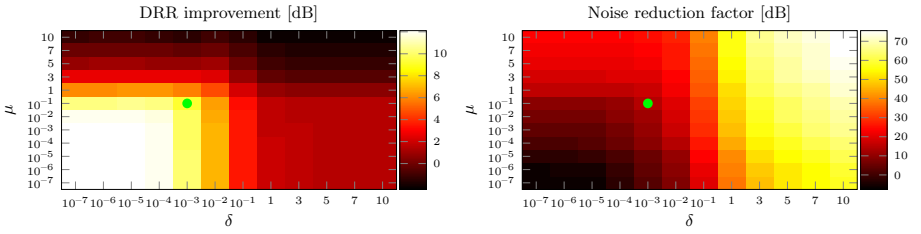


Fig. 8.4: Performance of the MWF-DNR technique for different regularization and weighting parameters δ and μ in terms of (a) ΔDRR and (b) ψ_{NR}. The circles denote the automatically determined regularization and weighting parameters (NPM = -33 dB, SIR = 0 dB).

μ (e.g., $\delta = 10^{-7}$ and $\mu = 10^{-7}$), the dereverberation performance is high whereas the background noise is amplified. Furthermore, it can be observed in Fig. 8.4a that for a fixed value of the weighting parameter μ (e.g., $\mu = 10^{-5}$), increasing the regularization parameter δ initially yields a slight increase in ΔDRR (not visible in Fig. 8.3a), however, as the regularization parameter δ is increased beyond 10^{-4}, the ΔDRR values decrease. This is to be expected since as already mentioned, for a relatively low RIR perturbation level, i.e., NPM = -33 dB, the optimal value of the regularization parameter δ required for a high dereverberation performance is small. Again similarly as for the RP-DNR technique, Fig. 8.4b shows that for a fixed value of the weighting parameter μ (e.g., $\mu = 10^{-5}$), increasing the regularization parameter δ also increases the noise reduction factor. Furthermore, as expected, for a fixed value of the regularization parameter δ (e.g., $\delta = 10^{-5}$), increasing the weighting parameter μ results in a trade-off between dereverberation and noise reduction performance, as illustrated by the decrease in DRR improvement and the increase in noise reduction factor.

For the considered example, the procedure proposed in Section 5.3 for automatically determining the regularization parameter δ in the R-PMINT technique yields $\delta = 10^{-3}$. Using this R-PMINT filter, the procedure proposed in Section 8.3.2 for automatically determining the weighting parameter μ in the MWF-DNR technique yields $\mu = 10^{-1}$. These parameter values are denoted by the circles in Figs. 8.4a and 8.4b. It can be observed that the two decoupled L-curve procedures for automatically determining the regularization and weighting parameters in the MWF-DNR technique yield parameters which result in a reasonable trade-off between dereverberation and noise reduction performance. This is also confirmed in Section 8.4.4 for other NPM and SIR values.

As shown by these simulation results, taking the RIR perturbations and the background noise into account by using appropriate regularization and weighting parameters is important to achieve joint dereverberation and noise reduction.

8.4.3 *Comparison of the automatically parametrized RP-DNR and MWF-DNR techniques to acoustic multi-channel equalization*

To further illustrate the importance of taking the noise statistics into account, in this section the performance of the automatically parametrized RP-DNR and MWF-DNR techniques is compared to the performance of the PMINT and the automatically regularized PMINT techniques, which do not take the actual noise statistics into account.[1] The considered input SIRs are

$$\text{SIR} \in \{0 \text{ dB}, 5 \text{ dB}\}, \tag{8.47}$$

and the presented performance measures for each SIR are averaged over the different NPM values in (8.42). Furthermore, the speech and noise correlation matrices for the RP-DNR and MWF-DNR techniques are perfectly estimated as in (8.43).

Tables 8.1 and 8.2 present the obtained ΔDRR, ΔPESQ, ψ_{NR}, ΔSRNR, and ΔfwSSNR values for SIR = 0 dB and SIR = 5 dB. As shown by the negative ΔDRR and ΔPESQ values, as expected, the PMINT technique fails to achieve dereverberation, introducing additional distortions in the output speech signal. On the other hand, by taking the RIR perturbations into account, the R-PMINT technique achieves a high reverberant energy suppression (ΔDRR = 9.37 dB) and perceptual speech quality improvement (ΔPESQ = 0.61). Furthermore, the proposed RP-DNR and MWF-DNR techniques achieve a very similar dereverberation performance as the R-PMINT technique. Although one would expect the dereverberation performance of the RP-DNR and MWF-DNR techniques to be worse than the dereverberation performance of the R-PMINT technique, the dereverberation performance is very similar. This occurs due to the automatic computation of the regularization

Table 8.1: Performance of the PMINT technique, automatically regularized R-PMINT technique, and automatically parametrized RP-DNR and MWF-DNR techniques (averaged over several NPM values; SIR = 0 dB). For each performance measure, the best performance is highlighted.

Measure	PMINT	R-PMINT	RP-DNR	MWF-DNR
ΔDRR [dB]	−10.26	**9.37**	9.33	9.23
ΔPESQ	−0.38	**0.61**	0.60	0.60
ψ_{NR} [dB]	−28.54	1.60	4.46	**11.79**
ΔSRNR [dB]	−12.06	2.13	3.78	**7.18**
ΔfwSSNR [dB]	−1.80	1.03	1.16	**2.82**

1 Note than in the regularized PMINT technique the matrix $\mathbf{R}_e$ can also be interpreted as a noise correlation matrix. However, we have assumed $\mathbf{R}_e = \mathbf{I}$, which does not correspond to the actual noise statistics.

Table 8.2: Performance of the PMINT technique, automatically regularized R-PMINT technique, and automatically parametrized RP-DNR and MWF-DNR techniques (averaged over several NPM values; SIR = 5 dB). For each performance measure, the best performance is highlighted.

Measure	PMINT	R-PMINT	RP-DNR	MWF-DNR
ΔDRR [dB]	−10.28	**9.37**	**9.37**	9.28
ΔPESQ	−0.38	**0.61**	**0.61**	**0.61**
ψ_{NR} [dB]	−27.98	1.47	2.25	**8.86**
ΔSRNR [dB]	−10.09	2.25	2.56	**4.98**
ΔfwSSNR [dB]	−2.76	1.49	1.57	**2.84**

parameter in the R-PMINT technique, which does not yield the optimal dereverberation performance one would obtain by intrusively determining the regularization parameter. Furthermore, as expected and as illustrated by the negative noise reduction factor, the PMINT technique leads to a very large noise amplification. Due to the decrease of the filter energy by incorporating a regularization parameter, the R-PMINT technique slightly reduces the noise by 1.60 dB for SIR = 0 dB and by 1.47 dB for SIR = 5 dB. However, by taking the actual noise statistics explicitly into account, the proposed RP-DNR technique improves the noise reduction factor to 4.46 dB for SIR = 0 dB and to 2.56 dB for SIR = 5 dB. By additionally taking the speech statistics into account the proposed MWF-DNR technique yields an even larger noise reduction factor of 11.79 dB for SIR = 0 dB and of 8.86 dB for SIR = 5 dB. The better joint dereverberation and noise reduction performance of the proposed RP-DNR and MWF-DNR techniques in comparison to acoustic multi-channel equalization techniques is also illustrated by the higher ΔSRNR and ΔfwSSNR values presented in Tables 8.1and 8.2, where the MWF-DNR technique outperforms the RP-DNR technique in terms of both instrumental measures.

Summarizing these results, it can be said that the noise statistics should be taken into account in order to avoid noise amplification and to achieve joint dereverberation and noise reduction. By additionally taking the speech statistics into account, an overall better performance can be achieved.

8.4.4 *Performance of the automatically parametrized RP-DNR and MWF-DNR techniques*

In this section the performance of the automatically parametrized RP-DNR and MWF-DNR techniques is extensively investigated for different noise levels, RIR perturbation levels, and correlation matrix estimation errors. The considered input SIRs are

$$\text{SIR} \in \{-5 \text{ dB}, -2.5 \text{ dB}, \ldots, 10 \text{ dB}\}, \tag{8.48}$$

and the presented performance measures for each SIR are averaged over the different NPM values in (8.42). The performance of the proposed RP-DNR and MWF-DNR techniques is investigated both for perfectly estimated correlation matrices, cf. (8.43), as well as for erroneously estimated correlation matrices, cf. (8.44).

Fig. 8.5 depicts the performance of the automatically parametrized RP-DNR and MWF-DNR techniques for perfectly estimated speech and noise correlation matrices. As shown by the ΔDRR and ΔPESQ values in Figs. 8.5a and 8.5b, the dereverber-

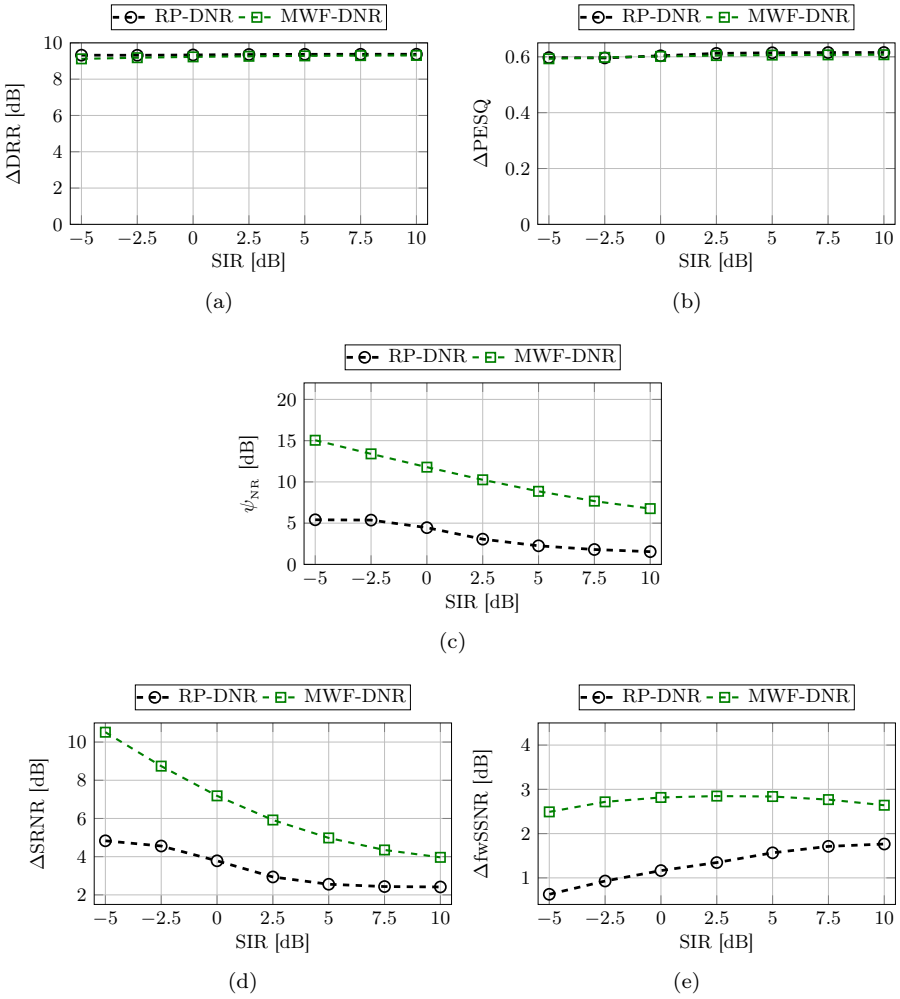


Fig. 8.5: Performance of the automatically parametrized RP-DNR and MWF-DNR techniques in terms of (a) ΔDRR, (b) ΔPESQ, (c) ψ_{NR}, (d) ΔSRNR, and (e) ΔfwSSNR (averaged over several NPM values, perfectly estimated correlation matrices).

ation performance of both techniques is very similar and almost independent of the SIR, with the RP-DNR technique yielding a slightly better performance at low input SIRs. However, as shown by the noise reduction factor in Fig. 8.5c, the MWF-DNR technique achieves a significantly better noise reduction performance, with the performance difference decreasing for increasing input SIR. The similar dereverberation performance but better noise reduction performance of the MWF-DNR technique is reflected in the higher ΔSRNR and ΔfwSSNR values achieved by the MWF-DNR technique, as depicted in Figs. 8.5d and 8.5e. Hence, it can be said that by taking the true speech statistics into account, the MWF-DNR technique outperforms the RP-DNR technique since it yields a similarly high dereverberation performance but a significantly better noise reduction performance.

Fig. 8.6 depicts the performance of the automatically parametrized RP-DNR and MWF-DNR techniques for erroneously estimated correlation matrices. Since the RP-DNR technique only requires the noise correlation matrix and since estimating $\mathbf{R}_{\mathbf{v}}$ from a long enough noise-only period (for the considered spatially stationary noise scenario) does not yield a significantly different estimate than for the previous simulation, the performance of the RP-DNR technique in terms of all performance measures is very similar as in Fig. 8.5. However, as shown in Figs. 8.6a and 8.6b, the dereverberation performance of the MWF-DNR technique is significantly lower than when using perfectly estimated correlation matrices. Due to the fact that the speech and noise signals are not perfectly uncorrelated, estimation errors occur in the estimate of the speech correlation matrix $\mathbf{R}_{\mathbf{x}} = \mathbf{R}_{\mathbf{y}} - \mathbf{R}_{\mathbf{v}}$, especially at low input SIRs. These estimation errors result in a worse dereverberated reference signal $\mathbf{R}_{\mathbf{x}}\mathbf{w}_{\text{R-P}}$ for the MWF-DNR technique, hence, significantly decreasing the dereverberation performance. However, the noise reduction performance of the MWF-DNR technique is still significantly better than the performance of the RP-DNR technique, as depicted in Fig. 8.6c. As depicted in Figs. 8.6d and 8.6e, the better noise reduction performance of the MWF-DNR technique also results in higher ΔSRNR and ΔfwSSNR values.

It should be noted that the noise reduction and the joint dereverberation and noise reduction performance of the MWF-DNR technique for erroneously estimated correlation matrices is better than for perfectly estimated correlation matrices (cf. Figs. 8.5 and 8.6), which may seem surprising at first. However, this can be explained by the automatic computation of the weighting parameter μ in the MWF-DNR technique, which for erroneously estimated correlation matrices yields a larger weighting parameter μ, hence a better noise reduction and a better joint dereverberation and noise reduction performance at the expense of a significantly worse dereverberation performance.

Summarizing, when the speech and noise correlation matrices can be accurately estimated, the MWF-DNR technique outperforms the RP-DNR technique since it yields a similarly high dereverberation performance at a significantly better noise reduction performance. However, when the correlation matrices are prone to estimation errors, the RP-DNR technique yields a significantly better dereverberation performance but a worse noise reduction performance than the MWF-DNR technique. Hence, the technique to be used should be chosen depending on what is more

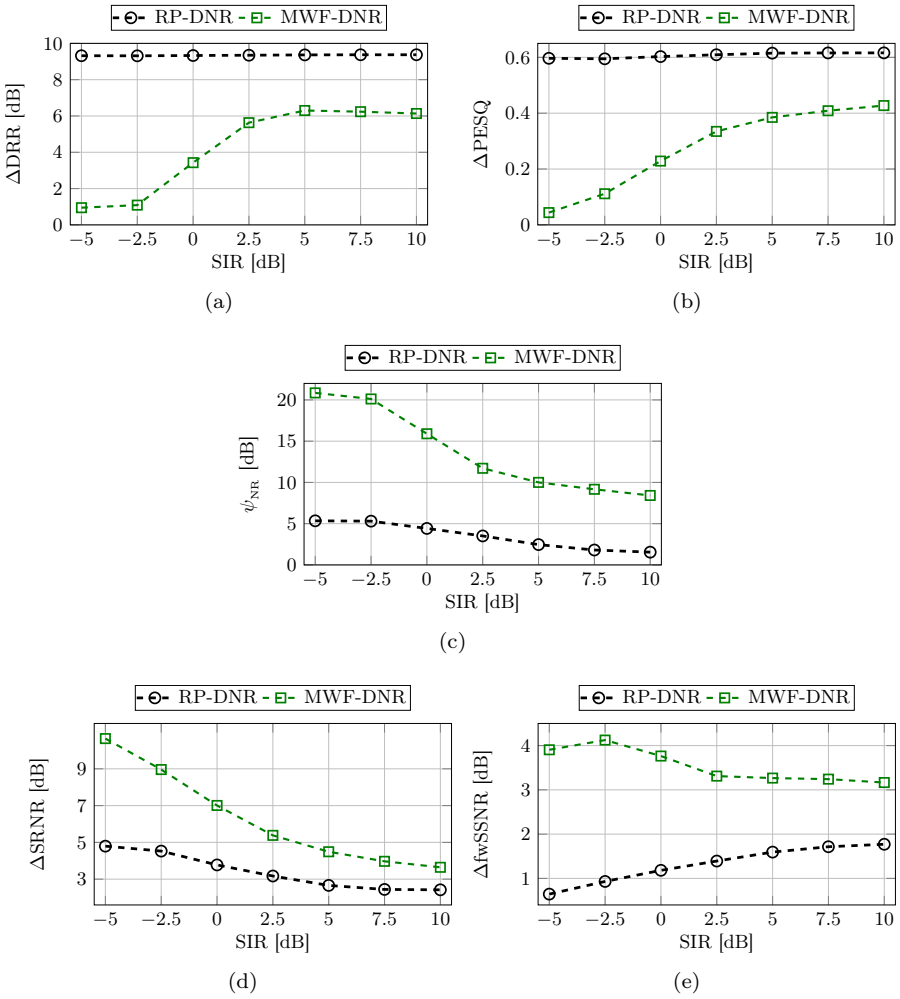


Fig. 8.6: Performance of the automatically parametrized RP-DNR and MWF-DNR techniques in terms of (a) ΔDRR, (b) ΔPESQ, (c) ψ_{NR}, (d) ΔSRNR, and (e) ΔfwSSNR (averaged over several NPM values, erroneously estimated correlation matrices).

important for the application under consideration, i.e., dereverberation or noise reduction performance.

8.5 Summary

In this chapter we have proposed two time domain techniques for joint dereverberation and noise reduction based on robust acoustic multi-channel equalization

techniques. The RP-DNR technique can be seen as an extension of the R-PMINT technique by explicitly taking the noise statistics into account. The MWF-DNR technique in addition takes the speech statistics into account and uses the dereverberated output speech signal of the R-PMINT technique as the reference signal for the MWF. In addition, we have proposed an automatic non-intrusive procedure based on the L-hypersurface for determining the regularization and weighting parameters in the RP-DNR technique, whereas two decoupled procedures based on the L-curve were used for automatically determining the regularization and weighting parameters in the MWF-DNR technique.

Extensive simulation results have shown that the RP-DNR technique maintains the high dereverberation performance of the R-PMINT technique while improving the noise reduction performance. Furthermore, it has been shown that the MWF-DNR technique yields a significantly better noise reduction performance than the RP-DNR technique at the expense of a worse dereverberation performance, depending on the amount of estimation errors in the speech correlation matrix.

9

CONCLUSION AND FURTHER RESEARCH

In this chapter we summarize the main contributions of this thesis and provide directions for further research.

9.1 Conclusion

In many hands-free speech communication applications such as teleconferencing or voice-controlled applications, the recorded microphone signals do not only contain the desired speech signal, but also attenuated and delayed copies of the desired speech signal due to reverberation as well as additive background noise. Reverberation and background noise cause a signal degradation which can impair speech intelligibility and which decreases the performance for many signal processing techniques.

The main objective of this thesis was to develop and optimally combine robust and perceptually advantageous speech dereverberation algorithms with noise reduction algorithms. Given that acoustic multi-channel equalization techniques for speech dereverberation offer the potential to achieve perfect dereverberation performance, these techniques were the central topic of this thesis. Since acoustic multi-channel equalization techniques require measured or estimated room impulse responses (RIRs) to be available, we investigated methods to increase the robustness of multi-channel equalization techniques against RIR perturbations. On the one hand, we proposed signal-independent methods, i.e., decreasing the reshaping filter length to improve the conditioning of the optimization criteria or incorporating regularization to reduce the energy of distortions due to RIR perturbations. On the other hand, we proposed a signal-dependent method, i.e., using a sparsity-promoting penalty function to sparsify the output speech signal and reduce artifacts generated by non-robust techniques. All proposed methods have been validated using instrumental performance measures and subjective listening tests. In addition, we proposed techniques to achieve joint dereverberation and noise reduction based on robust acoustic multi-channel equalization.

In Chapter 3 we reviewed state-of-the-art acoustic multi-channel equalization techniques, i.e., the multiple-input/output inverse theorem (MINT), channel shortening (CS), and relaxed multi-channel least-squares (RMCLS) techniques. In addition,

we proposed a least-squares perceptually advantageous equalization technique, i.e., the partial multi-channel equalization technique based on the multiple-input/output inverse theorem (PMINT), which aims to simultaneously preserve the perceptual speech quality and suppress the late reverberation. The PMINT technique preserves the perceptual speech quality by setting the early reflections of the target equalized impulse response equal to the early reflections of one of the available RIRs. In addition, similarly as in other state-of-the-art least-squares techniques, the late reverberation is suppressed by setting the late reflections of the target equalized impulse response equal to zero. Furthermore, we established a generalized framework for least-squares equalization techniques, i.e., MINT, RMCLS, and PMINT, which enabled to analyze the properties (existence and uniqueness) of the resulting reshaping filters. Based on this generalized framework, we analytically showed that least-squares techniques yield reshaping filters which lie in the subspace spanned by the multiple channel shortening solutions. Simulation results illustrated the importance of preserving the early reflections in order to improve the perceptual speech quality (e.g., for perfectly estimated RIRs with reverberation time $T_{60} \approx 450$ ms and the length of early reflections set to $L_d = 10$ ms, the RMCLS technique yields $\Delta\text{PESQ} = 1.5$, whereas the PMINT technique yields $\Delta\text{PESQ} = 2.3$). Furthermore, these results highlighted the necessity to increase the robustness of all considered acoustic multi-channel equalization techniques against RIR perturbations (e.g., for perturbed RIRs with $T_{60} \approx 450$ ms, $L_d = 10$ ms, and several perturbation levels, the RMCLS technique yields on average $\Delta\text{DRR} = 3.7$ dB and $\Delta\text{PESQ} = -0.1$, whereas the PMINT technique yields on average $\Delta\text{DRR} = -16.6$ dB and $\Delta\text{PESQ} = -0.3$).

Methods to increase the robustness of acoustic multi-channel equalization techniques against RIR perturbations were proposed in Chapters 4, 5, and 6.

In order to improve the conditioning of the optimization criteria, in Chapter 4 we proposed to increase the robustness of equalization techniques by decreasing the reshaping filter length. We derived a mathematical link between the reshaping filter length and the condition number of the (weighted) multi-channel convolution matrix, showing that shorter reshaping filters than conventionally used yield a smaller condition number, i.e., a higher robustness of the least-squares equalization techniques against RIR perturbations. Furthermore, we analytically showed that shorter reshaping filters in the channel shortening technique are also more robust against RIR perturbations, since they result in a better conditioned generalized eigenvalue optimization criterion with finite generalized eigenvalues. The presented simulation results validated the theoretical derivations, i.e., decreasing the reshaping filter length increases the robustness of the MINT, CS, RMCLS, and PMINT techniques, yielding a better dereverberation performance in the presence of RIR perturbations (e.g., for perturbed RIRs with $T_{60} \approx 450$ ms, $L_d = 10$ ms, and several perturbation levels, using shorter reshaping filters in RMCLS improves the ΔDRR by 4.2 dB and the ΔPESQ by 0.6, whereas using shorter reshaping filters in PMINT improves the ΔDRR by 23.4 dB and the ΔPESQ by 0.7). The advantage of using shorter reshaping filters is two-fold. First, the computational complexity of the re-

shaping filter design is decreased. Second, this is an effective method for increasing robustness which does not require any prior knowledge of the structure of the RIR perturbations. However, since this method does not incorporate any information about the RIR perturbations, clearly also its performance is rather limited.

To directly incorporate knowledge about the structure of the RIR perturbations, in Chapter 5 we proposed to increase the robustness of equalization techniques by using regularization such that the energy of distortions due to RIR perturbations is reduced. While the regularized least-squares reshaping filters were analytically derived, an iterative optimization procedure was used to compute the regularized channel shortening reshaping filter. Using the joint diagonalization of the (weighted) convolution matrix and of the matrix modeling the RIR perturbations, we analyzed the impact of regularization on the regularized least-squares reshaping filters. Furthermore, we proposed to automatically and non-intrusively determine the regularization parameter as the point of maximum curvature of the L-curve, obtained by plotting the distortion energy versus the dereverberation error energy for several regularization parameters. Although the curvature of the L-curve was analytically derived, we used the robust triangle method to maximize the curvature in order to avoid numerical inaccuracies. Simulation results showed that regularization significantly increases the dereverberation performance of the MINT, CS, RMCLS, and PMINT techniques (e.g., for perturbed RIRs with $T_{60} \approx 450$ ms, $L_d = 10$ ms, and several perturbation levels, incorporating intrusive regularization in RMCLS improves the ΔDRR by 10.8 dB and the ΔPESQ by 1.3, whereas incorporating intrusive regularization in PMINT improves the ΔDRR by 26.8 dB and the ΔPESQ by 1.3). Furthermore, the automatic non-intrusive procedure for determining the regularization parameter proved to be very effective, yielding a similar reverberant energy suppression and perceptual speech quality improvement as the intrusively determined regularization parameter. As a result, regularized equalization techniques can be considered to be robust and practically applicable equalization techniques for speech dereverberation.

While both methods proposed in Chapters 4 and 5 are signal-independent methods, in Chapter 6 we proposed a signal-dependent method, i.e., increase the robustness of equalization techniques by using a sparsity-promoting penalty function to sparsify the output speech signal and reduce artifacts generated by non-robust techniques. We extended the least-squares and channel shortening cost functions with different sparsity-promoting penalty functions, i.e., l_0-norm, l_1-norm, and the weighted l_1-norm. Furthermore, iterative algorithms based on the alternating direction method of multipliers were derived to compute the sparsity-promoting reshaping filters. Simulation results showed that incorporating the weighted l_1-norm sparsity-promoting penalty function significantly increases the robustness of the MINT, CS, RMCLS, and PMINT techniques against RIR perturbations (e.g., for perturbed RIRs with $T_{60} \approx 360$ ms, $L_d = 10$ ms, and the perturbation level NPM $= -33$ dB, incorporating the weighted l_1-norm penalty function in RMCLS improves the ΔDRR by 8.6 dB and the ΔPESQ by 0.5, whereas incorporating the weighted l_1-norm penalty function in PMINT improves the ΔDRR by 22.7 dB and

the ΔPESQ by 1.2). The advantage of sparsity-promoting equalization techniques lies in the fact that they exploit well-established characteristics of clean speech signals, without requiring prior information about the RIR perturbations. However, since the incorporation of sparsity-promoting penalty functions requires iterative algorithms for the reshaping filter design, these techniques are computationally more complex than the previously proposed techniques.

The simulation results in Chapters 4, 5, and 6 showed that the proposed robust extensions of the RMCLS and PMINT techniques outperform the proposed robust extensions of the MINT and CS techniques. The advantage of building upon the RMCLS technique lies in its relaxation of the constraints on the reshaping filter design, whereas the advantage of building upon the PMINT technique lies in its direct control of the early reflections. In order to determine the most effective method for increasing the robustness of acoustic multi-channel equalization techniques as well as to determine the most perceptually advantageous technique, in Chapter 7 we conducted a subjective evaluation of all robust extensions of the RMCLS and PMINT techniques for different scenarios, i.e., for different acoustic systems and RIR perturbation levels. The subjective listening test showed that the robust extensions of the PMINT technique are generally preferred over the robust extensions of the RMCLS technique. Furthermore, it was shown that the sparsity-promoting PMINT or the regularized PMINT techniques are the only techniques that yield a statistically significant improvement over the reverberant microphone signal for all considered scenarios, with the sparsity-promoting PMINT technique yielding the best perceptual speech quality for moderate RIR perturbation levels and the regularized PMINT technique yielding the best perceptual speech quality for high RIR perturbation levels.

Finally, in Chapter 8 we proposed two techniques for joint dereverberation and noise reduction, namely the regularized PMINT technique for joint dereverberation and noise reduction (RP-DNR) and the multi-channel Wiener filter (MWF) for joint dereverberation and noise reduction (MWF-DNR). The RP-DNR technique can be seen as an extension of the R-PMINT technique by explicitly taking the noise statistics into account. The MWF-DNR technique in addition takes the speech statistics into account and uses the dereverberated output signal of the R-PMINT technique as the reference signal for the MWF. In addition, we have proposed an automatic non-intrusive procedure based on the L-hypersurface for determining the regularization and weighting parameters in the RP-DNR technique, whereas two decoupled procedures based on the L-curve were used to automatically determine the regularization and weighting parameters in the MWF-DNR technique. Extensive simulation results have shown that the RP-DNR technique maintains the high dereverberation performance of the R-PMINT technique while improving the noise reduction performance (e.g., for perturbed RIRs with $T_{60} \approx 610$ ms, $L_d = 10$ ms, several perturbation levels, and a signal-to-interference ratio of 0 dB, the R-PMINT technique yields a ΔDRR of 9.4 dB and a noise reduction factor of 1.6 dB, whereas the RP-DNR technique yields a ΔDRR of 9.3 dB and a noise reduction factor of 4.5 dB). Furthermore, it has been shown that the MWF-DNR technique yields a

significantly better noise reduction performance than the RP-DNR technique at the expense of a worse dereverberation performance, depending on the amount of estimation errors in the speech correlation matrix (e.g., for perturbed RIRs with $T_{60} \approx 610$ ms, $L_d = 10$ ms, several perturbation levels, a signal-to-interference ratio of 0 dB, and erroneously estimated correlation matrices, the RP-DNR technique yields a ΔDRR of 9.4 dB and a noise reduction factor of 4.4 dB, whereas the MWF-DNR technique yields a ΔDRR of 3.4 dB and a noise reduction factor of 15.9 dB).

9.2 Suggestions for further research

In this thesis several time-domain methods have been proposed to increase the robustness of acoustic multi-channel equalization against RIR perturbations, i.e., i) decreasing the reshaping filter length, ii) incorporating regularization, and iii) incorporating sparsity-promoting penalty functions. As already mentioned, the proposed robust extensions of acoustic multi-channel equalization techniques have different advantages. Using a shorter reshaping filter length is effective in increasing the robustness without requiring any prior information about the RIR perturbation structure. If prior information about the RIR perturbation structure can be included, a better performance can be achieved. Hence, using regularization is more effective since knowledge about the RIR perturbation structure is incorporated. If this knowledge is not available, the regularized techniques would presumably not achieve such a high performance. Furthermore, using sparsity-promoting penalty functions is advantageous because well known characteristics of clean speech signals are exploited and no prior knowledge about the acoustic scenario is needed. Given the advantages of the individual techniques, investigating methods that optimally combine all three proposed techniques would be interesting and would offer the potential to increase the robustness of acoustic multi-channel equalization techniques even further.

The effectiveness of using shorter reshaping filters was validated by intrusively determining the optimal reshaping filter length. Similarly, the effectiveness of incorporating sparsity-promoting penalty functions was validated by intrusively determining the optimal weighting and penalty parameters. Intrusively determining these parameters is however not possible in practice, since knowledge of the true RIRs and of the clean speech signal is required. In future research it should be investigated how sensitive the performance of the proposed techniques is on these parameters. Furthermore, alternative procedures should be investigated, automatically determining the reshaping filter length and the weighting and penalty parameters using a non-intrusive approach. One possible approach would be to use non-intrusive dereverberation performance measures, such as a blind signal-based direct-to-reverberant ratio estimator [205–207] or the non-intrusive speech-to-reverberation modulation energy ratio measure [208]. Another alternative to automatically determine the reshaping filter length would be to adapt the L-curve method, e.g., using the condition number of the least-squares matrix and the dereverberation error energy as the trade-off quantities for the L-curve.

Furthermore, although the proposed robust acoustic multi-channel equalization techniques are computationally feasible techniques (particularly using shorter reshaping filters), the computational complexity nevertheless remains quite high (particularly incorporating sparsity-promoting penalty functions). It would be interesting to investigate methods that reduce the computational complexity of these techniques without degrading their performance and possibly even further increase their robustness. Instead of working in the time-domain, one approach would be to design the proposed reshaping filters using decimated and oversampled subbands as in [123], where the fullband RIRs are decomposed into equivalent subband filters prior to equalization.

Moreover, it would be useful to investigate and compare the performance of the different proposed techniques in the presence of more realistic RIR perturbations arising due to spatial mismatch or blind and supervised system identification methods. Although in the presented simulation results in this thesis it was validated that the regularized techniques generally outperform the other proposed robust techniques, we expect that this is not necessarily the case when the RIR perturbation structure cannot be well approximated.

Finally, in order to further improve the performance of acoustic multi-channel equalization in realistic acoustic scenarios, the robustness of system identification methods needs to be significantly improved, such that better RIR estimates can be delivered to acoustic multi-channel equalization techniques. If the robustness of system identification methods improves, one can exploit the full potential of acoustic multi-channel equalization techniques.

A

INTERLACING INEQUALITIES FOR SHORTER RESHAPING FILTERS IN LEAST-SQUARES EQUALIZATION TECHNIQUES

Aiming at establishing a relation between the condition numbers of the matrices $\mathbf{W}_s\hat{\mathbf{H}}_s$ and $\mathbf{W}_t\hat{\mathbf{H}}_t$, with

$$\chi_{\mathbf{W}_s\hat{\mathbf{H}}_s} = \frac{\sigma_{\mathbf{W}_s\hat{\mathbf{H}}_s}(1)}{\sigma_{\mathbf{W}_s\hat{\mathbf{H}}_s}(r_s)}, \tag{A.1}$$

$$\chi_{\mathbf{W}_t\hat{\mathbf{H}}_t} = \frac{\sigma_{\mathbf{W}_t\hat{\mathbf{H}}_t}(1)}{\sigma_{\mathbf{W}_t\hat{\mathbf{H}}_t}(r_t)}, \tag{A.2}$$

we consider the following interlacing inequalities between the singular values of a matrix and its sub-matrices.

Interlacing inequalities [165]: *Given a matrix* $\mathbf{A}$ *of dimensions* $u \times v$ *and a sub-matrix* $\mathbf{B}$ *obtained by deleting* l *rows and/or* l *columns from* $\mathbf{A}$*, the singular values of* $\mathbf{A}$ *and* $\mathbf{B}$ *interlace as*

$$\sigma_{\mathbf{A}}(i) \geq \sigma_{\mathbf{B}}(i) \geq \sigma_{\mathbf{A}}(i+l) \qquad i = 1, \ \ldots, \ \min\{u-l, v-l\}. \tag{A.3}$$

In order to construct the matrix $\mathbf{W}_s\hat{\mathbf{H}}_s$, we first create an intermediate $[p_t - (L_t - L_s)] \times [q_t - (L_t - L_s)]$–dimensional sub-matrix $\mathbf{T}$ by deleting $L_t - L_s$ rows and $L_t - L_s$ columns from $\mathbf{W}_t\hat{\mathbf{H}}_t$. The interlacing inequalities in (A.3) for the matrices $\mathbf{W}_t\hat{\mathbf{H}}_t$ and $\mathbf{T}$ can be written as

$$\sigma_{\mathbf{W}_t\hat{\mathbf{H}}_t}(i) \geq \sigma_{\mathbf{T}}(i) \geq \sigma_{\mathbf{W}_t\hat{\mathbf{H}}_t}[i + (L_t - L_s)], \ \ i = 1, \ \ldots, \ r_t - (L_t - L_s), \ \ldots, \ p_t - (L_t - L_s). \tag{A.4}$$

Using (A.4), the following inequalities between the singular values of the matrices $\mathbf{W}_t\hat{\mathbf{H}}_t$ and $\mathbf{T}$ hold:

$$\sigma_{\mathbf{W}_t\hat{\mathbf{H}}_t}(1) \geq \sigma_{\mathbf{T}}(1), \tag{A.5}$$

$$\sigma_{\mathbf{T}}[r_t - (L_t - L_s)] \geq \sigma_{\mathbf{W}_t\hat{\mathbf{H}}_t}(r_t). \tag{A.6}$$

In order to construct the matrix $\mathbf{W}_s\hat{\mathbf{H}}_s$, $(M-1)(L_t-L_s)$ columns are now deleted from the matrix $\mathbf{T}$. The interlacing inequalities in (A.3) for the matrices $\mathbf{T}$ and $\mathbf{W}_s\hat{\mathbf{H}}_s$ can be written as

$$\sigma_{\mathbf{T}}(i) \geq \sigma_{\mathbf{W}_s\hat{\mathbf{H}}_s}(i) \geq \sigma_{\mathbf{T}}[i+(M-1)(L_t-L_s)], \quad i=1,\ \ldots,\ r_s. \tag{A.7}$$

Using (A.7), the following inequalities between the singular values of the matrices $\mathbf{T}$ and $\mathbf{W}_s\hat{\mathbf{H}}_s$ hold:

$$\sigma_{\mathbf{T}}(1) \geq \sigma_{\mathbf{W}_s\hat{\mathbf{H}}_s}(1), \tag{A.8}$$

$$\sigma_{\mathbf{W}_s\hat{\mathbf{H}}_s}(r_s) \geq \sigma_{\mathbf{T}}[r_s+(M-1)(L_t-L_s)]. \tag{A.9}$$

The index of the singular value in the right hand side of (A.9) can be written as

$$r_s+(M-1)(L_t-L_s) = ML_s+(M-1)(L_t-L_s) = ML_t-(L_t-L_s) \geq r_t-(L_t-L_s), \tag{A.10}$$

with the inequality in (A.10) clearly holding since the number of columns in $\mathbf{W}_t\hat{\mathbf{H}}_t$ is greater or equal than its rank, i.e.,

$$q_t = ML_t \geq p_t \geq r_t. \tag{A.11}$$

Based on (A.10) and the fact that the singular values of a matrix are sorted in descending order, one can write

$$\sigma_{\mathbf{T}}[r_s+(M-1)(L_t-L_s)] \geq \sigma_{\mathbf{T}}[r_t-(L_t-L_s)]. \tag{A.12}$$

Using (A.12), the inequality in (A.9) can also be written as

$$\sigma_{\mathbf{W}_s\hat{\mathbf{H}}_s}(r_s) \geq \sigma_{\mathbf{T}}[r_t-(L_t-L_s)]. \tag{A.13}$$

Finally, combining (A.5), (A.6), (A.8), and (A.13) the following inequalities relating the largest and smallest non-zero singular values of $\mathbf{W}_t\hat{\mathbf{H}}_t$ and $\mathbf{W}_s\hat{\mathbf{H}}_s$ can be established:

$$\sigma_{\mathbf{W}_t\hat{\mathbf{H}}_t}(1) \geq \sigma_{\mathbf{W}_s\hat{\mathbf{H}}_s}(1), \tag{A.14}$$

$$\sigma_{\mathbf{W}_s\hat{\mathbf{H}}_s}(r_s) \geq \sigma_{\mathbf{W}_t\hat{\mathbf{H}}_t}(r_t). \tag{A.15}$$

It readily follows from (A.14) and (A.15) that the condition number of $\mathbf{W}_s\hat{\mathbf{H}}_s$ is smaller or equal than the condition number of $\mathbf{W}_t\hat{\mathbf{H}}_t$, i.e.,

$$\chi_{\mathbf{W}_s\hat{\mathbf{H}}_s} = \frac{\sigma_{\mathbf{W}_s\hat{\mathbf{H}}_s}(1)}{\sigma_{\mathbf{W}_s\hat{\mathbf{H}}_s}(r_s)} \leq \frac{\sigma_{\mathbf{W}_t\hat{\mathbf{H}}_t}(1)}{\sigma_{\mathbf{W}_t\hat{\mathbf{H}}_t}(r_t)} = \chi_{\mathbf{W}_t\hat{\mathbf{H}}_t}. \tag{A.16}$$

B

FREQUENCY DOMAIN ONE- AND TWO-STAGE TECHNIQUES FOR JOINT DEREVERBERATION AND NOISE REDUCTION

In [87], a frequency domain two-stage beamforming technique for joint dereverberation and noise reduction has been proposed. In the first stage, a superdirective beamformer is applied to generate a dereverberated signal, whereas in the second stage this dereverberated signal is used as a reference signal for a frequency domain MWF. Although this two-stage beamforming technique appears to be very similar to the MWF-DNR technique proposed in Section 8.1.2, three main differences arise, i.e.,

i) the MWF-DNR technique generates the dereverberated reference signal for the MWF using the R-PMINT filter, whereas the two-stage beamforming technique in [87] uses a superdirective beamformer,

ii) the MWF-DNR technique is a time domain technique, whereas the two-stage beamforming technique in [87] is a frequency domain technique, and

iii) the MWF-DNR technique is a one-stage technique applying a single filter to achieve joint dereverberation and noise reduction, whereas the beamforming technique in [87] is a two-stage technique applying two different filters.

Due to the substantial differences in i) and ii), a detailed performance comparison of the MWF-DNR technique and the two-stage beamforming technique is beyond the scope of this thesis. The difference in i) represents a substantial difference since designing acoustic multi-channel equalization filters in the frequency domain is of limited use in practice, since a set of optimal equalization filters in the frequency domain is only constrained to be stable, but not necessarily causal or finite [209, 210]. Furthermore, the difference in ii) also represents a substantial difference since acoustic multi-channel equalization techniques are non-blind techniques relying on measured or estimated RIRs to achieve dereverberation, whereas superdirective beamforming is a blind technique typically only requiring knowledge of the direction of arrival of the speech source. In this appendix, we will focus on the difference in iii), i.e., using a one-stage versus a two-stage implementation, which is relevant both for the time

domain MWF-DNR technique as well as for the frequency domain beamforming technique in [87].

The proposed one-stage MWF-DNR technique can be reformulated as a two-stage technique, where in the first stage dereverberation filters are applied to generate a dereverberated signal and in the second stage this dereverberated signal is used as the reference signal for the MWF. Similarly, the two-stage beamforming technique in [87] can be reformulated as a one-stage technique, where a single frequency domain filter is designed to achieve joint dereverberation and noise reduction. In this appendix we present the frequency domain technique proposed in [87] and show how this two-stage technique can be reformulated as a one-stage technique. Furthermore, we show that when the filter used to generate the dereverberated reference signal is invertible, the one-stage and the two-stage techniques are equivalent. However, when the filter used to generate the dereverberated reference signal is non-invertible, using a one-stage technique is advantageous and yields a higher narrowband output signal-to-noise ratio (SNR) than using a two-stage technique.

In Section B.1 the two-stage beamforming technique is presented, whereas in Section B.2 this technique is reformulated as a one-stage technique. In Section B.3 the narrowband output SNR of the one-stage and the two-stage techniques is analytically derived and compared.

B.1 Two-stage technique for joint dereverberation and noise reduction

As presented in Section 2.1.3, the time domain signal model in (2.1) can be written in the frequency domain as

$$Y_m(\omega) = \underbrace{S(\omega)H_m(\omega)}_{X_m(\omega)} + V_m(\omega), \tag{B.1}$$

where $Y_m(\omega)$, $S(\omega)$, $H_m(\omega)$, $X_m(\omega)$, and $V_m(\omega)$ denote the discrete-time Fourier transforms of $y_m(n)$, $s(n)$, $h_m(n)$, $x_m(n)$, and $v_m(n)$, respectively, at angular frequency ω.

In the two-stage beamforming technique in [87] depicted in Fig. B.1, first dereverberation filters $G_m(\omega)$ are applied to the received microphone signals followed by noise

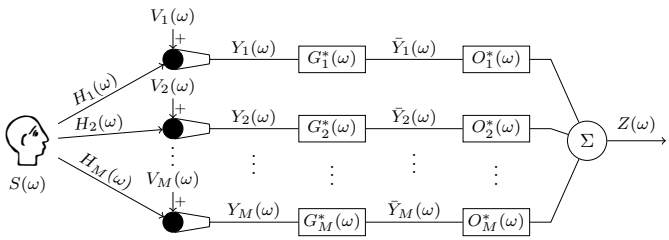


Fig. B.1: Acoustic system configuration for the two-stage beamforming technique for joint dereverberation and noise reduction.

reduction filters $O_m(\omega)$. Applying the dereverberation filters $G_m(\omega)$, the filtered microphone signals $\bar{Y}_m(\omega)$ after the first stage are given by

$$\bar{Y}_m(\omega) = G_m^*(\omega)Y_m(\omega) = \underbrace{G_m^*(\omega)X_m(\omega)}_{\bar{X}_m(\omega)} + \underbrace{G_m^*(\omega)V_m(\omega)}_{\bar{V}_m(\omega)}, \tag{B.2}$$

with $\bar{X}_m(\omega)$ and $\bar{V}_m(\omega)$ the filtered speech and noise components. The dereverberation filters $G_m(\omega)$ are designed such that the dereverberated reference signal $S_{\mathrm{r}}(\omega)$ is given by the sum of the filtered speech components, i.e.,

$$S_{\mathrm{r}}(\omega) = \sum_{m=1}^{M} \bar{X}_m(\omega) = \sum_{m=1}^{M} G_m^*(\omega)X_m(\omega). \tag{B.3}$$

While superdirective beamforming has been used in [87] to design the dereverberation filters $G_m(\omega)$, in principle any frequency domain dereverberation technique can be used to design these filters.

In the second-stage noise reduction filters $O_m(w)$ are applied, such that the output speech signal $Z(\omega)$ is given by the sum of the filtered microphone signals, i.e.,

$$Z(\omega) = \sum_{m=1}^{M} O_m^*(\omega)\bar{Y}_m(\omega) = \sum_{m=1}^{M} O_m^*(\omega)\bar{X}_m(\omega) + \sum_{m=1}^{M} O_m^*(\omega)\bar{V}_m(\omega). \tag{B.4}$$

In vector notation, the M–dimensional stacked vector of the microphone signals $\mathbf{y}(\omega)$ can be expressed as

$$\mathbf{y}(\omega) = \mathbf{x}(\omega) + \mathbf{v}(\omega), \tag{B.5}$$

with

$$\mathbf{y}(\omega) = [Y_1(\omega)\ Y_2(\omega)\ \ldots\ Y_M(\omega)]^T, \tag{B.6}$$

and $\mathbf{x}(\omega)$ and $\mathbf{v}(\omega)$ similarly defined. Using the diagonal matrix $\mathbf{G}(\omega)$ consisting of the dereverberation filter coefficients, i.e.,

$$\mathbf{G}(\omega) = \mathrm{diag}\{\mathbf{g}(\omega)\} = \begin{bmatrix} G_1(\omega) & 0 & \cdots & 0 \\ 0 & G_2(\omega) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & G_M(\omega) \end{bmatrix}, \tag{B.7}$$

with $\mathbf{g}(\omega) = [G_1(\omega)\ G_2(\omega)\ \ldots\ G_M(\omega)]^T$, the output signal vector of the dereverberation stage can be expressed as

$$\bar{\mathbf{y}}(\omega) = \mathbf{G}^H(\omega)\mathbf{y}(\omega) = \underbrace{\mathbf{G}^H(\omega)\mathbf{x}(\omega)}_{\bar{\mathbf{x}}(\omega)} + \underbrace{\mathbf{G}^H(\omega)\mathbf{v}(\omega)}_{\bar{\mathbf{v}}(\omega)}, \tag{B.8}$$

with $\bar{\mathbf{y}}(\omega)$, $\bar{\mathbf{x}}(\omega)$, and $\bar{\mathbf{v}}(\omega)$ the M–dimensional stacked vectors of the filtered microphone signals, filtered speech components, and filtered noise components defined

similarly as in (B.6). Furthermore, using the stacked M–dimensional noise reduction filter vector $\mathbf{o}(\omega)$, i.e.,

$$\mathbf{o}(\omega) = [O_1(\omega)\ O_2(\omega)\ \ldots\ O_M(\omega)]^T, \tag{B.9}$$

the output speech signal can be written in vector notation as

$$Z(\omega) = \mathbf{o}^H(\omega)\bar{\mathbf{y}}(\omega) = \mathbf{o}^H(\omega)\bar{\mathbf{x}}(\omega) + \mathbf{o}^H(\omega)\bar{\mathbf{v}}(\omega). \tag{B.10}$$

In [87], the MWF has been used to design the noise reduction filter vector $\mathbf{o}(\omega)$ such that the minimum mean-square error between the output speech signal $Z(\omega)$ and the dereverberated reference signal $S_{\mathrm{r}}(\omega)$ is minimized. The cost function of the frequency domain MWF used for the second stage is defined as

$$\mathcal{E}\{|Z(\omega) - S_{\mathrm{r}}(\omega)|^2\} = \mathcal{E}\{|\mathbf{o}^H(\omega)\bar{\mathbf{y}}(\omega) - S_{\mathrm{r}}(\omega)|^2\}, \tag{B.11}$$

where the reference signal $S_{\mathrm{r}}(\omega)$ can be expressed in vector notation as

$$S_{\mathrm{r}}(\omega) = \boldsymbol{\beta}^T\bar{\mathbf{x}}(\omega), \tag{B.12}$$

with $\boldsymbol{\beta} = [1\ 1\ \ldots\ 1]^T$. Assuming that the filtered speech and noise components $\bar{\mathbf{x}}(\omega)$ and $\bar{\mathbf{v}}(\omega)$ are uncorrelated and introducing a weighting parameter μ to trade off between speech distortion and noise reduction, the cost function of the speech-distortion weighting MWF used in the second stage to achieve noise reduction and estimate the dereverberated reference signal $S_{\mathrm{r}}(\omega)$ is defined as

$$J_{\mathrm{SII}}(\omega) = \mathcal{E}\{|\mathbf{o}^H(\omega)\bar{\mathbf{x}}(\omega) - \boldsymbol{\beta}^T\bar{\mathbf{x}}(\omega)|^2\} + \mu\mathcal{E}\{|\mathbf{o}^H(\omega)\bar{\mathbf{v}}(\omega)|^2\}. \tag{B.13}$$

The MWF $\mathbf{o}(\omega)$ minimizing (B.13) is then equal to

$$\mathbf{o}(\omega) = [\mathbf{R}_{\bar{\mathbf{x}}}(\omega) + \mu\mathbf{R}_{\bar{\mathbf{v}}}(\omega)]^{-1}\mathbf{R}_{\bar{\mathbf{x}}}(\omega)\boldsymbol{\beta}, \tag{B.14}$$

with $\mathbf{R}_{\bar{\mathbf{x}}}(\omega)$ and $\mathbf{R}_{\bar{\mathbf{v}}}(\omega)$ the correlation matrices of the speech and noise components after the first dereverberation stage, i.e.,

$$\mathbf{R}_{\bar{\mathbf{x}}}(\omega) = \mathcal{E}\{\bar{\mathbf{x}}(\omega)\bar{\mathbf{x}}^H(\omega)\}, \tag{B.15}$$

$$\mathbf{R}_{\bar{\mathbf{v}}}(\omega) = \mathcal{E}\{\bar{\mathbf{v}}(\omega)\bar{\mathbf{v}}^H(\omega)\}. \tag{B.16}$$

Summarizing, the two-stage beamforming technique proposed in [87] applies $\mathbf{G}(\omega)$ in the first stage to create a dereverberated reference signal and uses the MWF in (B.14) in the second stage to suppress the noise and estimate the dereverberated reference signal. The overall filter $\mathbf{w}_{\mathrm{II}}(\omega)$ applied to the received microphone signals $\mathbf{y}(\omega)$ in this two-stage technique is given by

$$\mathbf{w}_{\mathrm{II}}(\omega) = \mathbf{G}(\omega)\mathbf{o}(\omega) = \mathbf{G}(\omega)[\mathbf{R}_{\bar{\mathbf{x}}}(\omega) + \mu\mathbf{R}_{\bar{\mathbf{v}}}(\omega)]^{-1}\mathbf{R}_{\bar{\mathbf{x}}}(\omega)\boldsymbol{\beta}. \tag{B.17}$$

B.2 One-stage technique for joint dereverberation and noise reduction

Similarly as the time domain MWF-DNR technique proposed in Section 8.1.2, the two-stage frequency domain technique in Section B.1 can also be reformulated as a one-stage technique.

Fig. B.2 depicts a schematic representation of the one-stage counterpart of the technique discussed in Section B.1, where a single filter $\mathbf{w}(\omega)$ is applied to the received microphone signals, i.e.,

$$Z(\omega) = \mathbf{w}^H(\omega)\mathbf{y}(\omega) = \mathbf{w}^H(\omega)\mathbf{x}(\omega) + \mathbf{w}^H(\omega)\mathbf{v}(\omega), \tag{B.18}$$

with

$$\mathbf{w}(\omega) = [W_1(\omega)\ W_2(\omega)\ \ldots W_M(\omega)]^T. \tag{B.19}$$

In order to minimize the mean-square error between the output speech signal and the dereverberated reference signal in (B.12), the one-stage frequency domain MWF cost function is defined as

$$\mathcal{E}\{|Z(\omega) - S_{\text{r}}(\omega)|^2\} = \mathcal{E}\{|\mathbf{w}^H(\omega)\mathbf{y}(\omega) - S_{\text{r}}(\omega)|^2\}, \tag{B.20}$$

where the reference signal $S_{\text{r}}(\omega)$ can be expressed in terms of the reverberant signal component $\mathbf{x}(\omega)$ as

$$S_{\text{r}}(\omega) = \boldsymbol{\beta}^T\bar{\mathbf{x}}(\omega) = \underbrace{\boldsymbol{\beta}^T\mathbf{G}^H(\omega)}_{\mathbf{g}^H(\omega)}\mathbf{x}(\omega). \tag{B.21}$$

Similarly as before, assuming that the speech and noise components $\mathbf{x}(\omega)$ and $\mathbf{v}(\omega)$ are uncorrelated and introducing a weighting parameter μ, the speech distortion weighted MWF cost function for the one-stage frequency domain technique can be written as

$$J_{\text{I}}(\omega) = \mathcal{E}\{|\mathbf{w}^H(\omega)\mathbf{x}(\omega) - \mathbf{g}^H(\omega)\mathbf{x}(\omega)|^2\} + \mu\mathcal{E}\{|\mathbf{w}^H(\omega)\mathbf{v}(\omega)|^2\}. \tag{B.22}$$

Minimizing (B.22) yields the one-stage MWF

$$\mathbf{w}_{\text{I}}(\omega) = [\mathbf{R}_{\mathbf{x}}(\omega) + \mu\mathbf{R}_{\mathbf{v}}(\omega)]^{-1}\mathbf{R}_{\mathbf{x}}(\omega)\mathbf{g}(\omega), \tag{B.23}$$

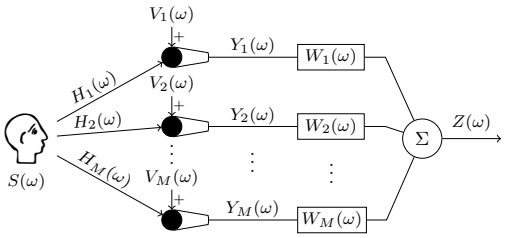


Fig. B.2: Acoustic system configuration for the one-stage reformulation of the two-stage beamforming technique for joint dereverberation and noise reduction.

where $\mathbf{R}_{\mathbf{x}}(\omega)$ and $\mathbf{R}_{\mathbf{v}}(\omega)$ are the received speech and noise correlation matrices, i.e.,

$$\mathbf{R}_{\mathbf{x}}(\omega) = \mathcal{E}\{\mathbf{x}(\omega)\mathbf{x}^H(\omega)\}, \tag{B.24}$$

$$\mathbf{R}_{\mathbf{v}}(\omega) = \mathcal{E}\{\mathbf{v}(\omega)\mathbf{v}^H(\omega)\}, \tag{B.25}$$

with $\mathbf{R}_{\mathbf{v}}(\omega)$ assumed to be a full-rank matrix. Note that the filtered speech and noise correlation matrices $\mathbf{R}_{\bar{\mathbf{x}}}(\omega)$ and $\mathbf{R}_{\bar{\mathbf{v}}}(\omega)$ in the two-stage technique in Section B.1 can be expressed in terms of the received speech and noise correlation matrices $\mathbf{R}_{\mathbf{x}}(\omega)$ and $\mathbf{R}_{\mathbf{v}}(\omega)$ as

$$\mathbf{R}_{\bar{\mathbf{x}}}(\omega) = \mathbf{G}^H(\omega)\mathbf{R}_{\mathbf{x}}(\omega)\mathbf{G}(\omega), \tag{B.26}$$

$$\mathbf{R}_{\bar{\mathbf{v}}}(\omega) = \mathbf{G}^H(\omega)\mathbf{R}_{\mathbf{v}}(\omega)\mathbf{G}(\omega). \tag{B.27}$$

B.3 Analytical comparison of the one-stage and two-stage techniques

In this section we analytically derive and compare the performance of the one-stage filter $\mathbf{w}_{\text{I}}(\omega)$ and the two-stage filter $\mathbf{w}_{\text{II}}(\omega)$ in terms of the narrowband output SNR (oSNR), with

$$\text{oSNR}(\omega) = \frac{\mathbf{w}^H(\omega)\mathbf{R}_{\mathbf{x}}(\omega)\mathbf{w}(\omega)}{\mathbf{w}^H(\omega)\mathbf{R}_{\mathbf{v}}(\omega)\mathbf{w}(\omega)}. \tag{B.28}$$

Narrowband output SNR of the one-stage filter

Assuming a single speech source (cf. (B.1)), the speech correlation matrix in (B.24) is a rank-1 matrix, i.e.,

$$\mathbf{R}_{\mathbf{x}}(\omega) = P_s(\omega)\mathbf{h}(\omega)\mathbf{h}^H(\omega), \tag{B.29}$$

with $P_s(\omega) = \mathcal{E}\{|S(\omega)|^2\}$ the power spectral density of the clean speech signal and $\mathbf{h}(\omega) = [H_1(\omega)\ H_2(\omega)\ \ldots\ H_M(\omega)]^T$ the vector of acoustic transfer functions between the speech source and the microphones. Using (B.29), the one-stage filter in (B.23) can be expressed as

$$\mathbf{w}_{\text{I}}(\omega) = [P_s(\omega)\mathbf{h}(\omega)\mathbf{h}^H(\omega) + \mu\mathbf{R}_{\mathbf{v}}(\omega)]^{-1}P_s(\omega)\mathbf{h}(\omega)\mathbf{h}^H(\omega)\mathbf{g}(\omega). \tag{B.30}$$

Using the matrix inversion lemma [211], the one-stage filter in (B.30) can be expressed as

$$\mathbf{w}_{\text{I}}(\omega) = \underbrace{\frac{P_s(\omega)\mathbf{h}^H(\omega)\mathbf{g}(\omega)}{\mu + P_s(\omega)\mathbf{h}^H(\omega)\mathbf{R}_{\mathbf{v}}^{-1}(\omega)\mathbf{h}(\omega)}}_{\Phi_{\text{I}}(\omega)}\mathbf{R}_{\mathbf{v}}^{-1}(\omega)\mathbf{h}(\omega), \tag{B.31}$$

with $\Phi_{\text{I}}(\omega)$ a complex-valued scalar. Substituting (B.31) in (B.28), the narrowband output SNR of the one-stage filter is equal to

$$\text{oSNR}_{\text{I}}(\omega) = \frac{\mathbf{w}_{\text{I}}^H(\omega)\mathbf{R}_{\mathbf{x}}(\omega)\mathbf{w}_{\text{I}}(\omega)}{\mathbf{w}_{\text{I}}^H(\omega)\mathbf{R}_{\mathbf{v}}(\omega)\mathbf{w}_{\text{I}}(\omega)} \tag{B.32}$$

$$= \frac{|\Phi_{\text{I}}(\omega)|^2\mathbf{h}^H(\omega)\mathbf{R}_{\mathbf{v}}^{-1}(\omega)P_s(\omega)\mathbf{h}(\omega)\mathbf{h}^H(\omega)\mathbf{R}_{\mathbf{v}}^{-1}(\omega)\mathbf{h}(\omega)}{|\Phi_{\text{I}}(\omega)|^2\mathbf{h}^H(\omega)\mathbf{R}_{\mathbf{v}}^{-1}(\omega)\mathbf{R}_{\mathbf{v}}(\omega)\mathbf{R}_{\mathbf{v}}^{-1}(\omega)\mathbf{h}(\omega)} \tag{B.33}$$

$$= P_s(\omega)\mathbf{h}^H(\omega)\mathbf{R}_{\mathbf{v}}^{-1}(\omega)\mathbf{h}(\omega). \tag{B.34}$$

As can be seen in (B.34), the narrowband output SNR of the one-stage filter is independent of the scalar $\Phi_{\text{I}}(\omega)$. Furthermore, the narrowband output SNR in (B.34) is the maximum generalized eigenvalue of the generalized eigenvalue problem

$$\mathbf{R}_{\mathbf{x}}(\omega)\mathbf{w}(\omega) = \lambda(\omega)\mathbf{R}_{\mathbf{v}}(\omega)\mathbf{w}(\omega). \tag{B.35}$$

Therefore, the one-stage filter $\mathbf{w}_{\text{I}}(\omega)$ is the generalized eigenvector associated with the maximum generalized eigenvalue.

Narrowband output SNR of the two-stage filter

In order to derive the narrowband output SNR of the two-stage filter, we first express the filtered speech correlation matrix $\mathbf{R}_{\bar{\mathbf{x}}}(\omega)$ in (B.26) as a rank-1 matrix. Substituting (B.29) in (B.26), $\mathbf{R}_{\bar{\mathbf{x}}}(\omega)$ can be expressed as

$$\mathbf{R}_{\bar{\mathbf{x}}}(\omega) = P_s(\omega)\underbrace{\mathbf{G}^H(\omega)\mathbf{h}(\omega)}_{\bar{\mathbf{h}}(\omega)}\underbrace{\mathbf{h}^H(\omega)\mathbf{G}(\omega)}_{\bar{\mathbf{h}}^H(\omega)}, \tag{B.36}$$

with $\bar{\mathbf{h}}(\omega)$ denoting the vector of acoustic transfer functions filtered by the dereverberation filter coefficients $G_m(\omega)$. Similarly as for the one-stage filter, using the matrix inversion lemma, the two-stage filter can be written as

$$\mathbf{w}_{\text{II}}(\omega) = \frac{P_s(\omega)\bar{\mathbf{h}}^H(\omega)\boldsymbol{\beta}}{\underbrace{\mu + P_s(\omega)\bar{\mathbf{h}}^H(\omega)\mathbf{R}_{\bar{\mathbf{v}}}^{-1}(\omega)\bar{\mathbf{h}}(\omega)}_{\Phi_{\text{II}}(\omega)}}\mathbf{G}(\omega)\mathbf{R}_{\bar{\mathbf{v}}}^{-1}(\omega)\bar{\mathbf{h}}(\omega), \tag{B.37}$$

with $\Phi_{\text{II}}(\omega)$ a complex-valued scalar. Substituting (B.27) in (B.37), the two-stage filter can be expressed as

$$\mathbf{w}_{\text{II}}(\omega) = \Phi_{\text{II}}(\omega)\ \mathbf{G}(\omega)[\mathbf{G}(\omega)^H\mathbf{R}_{\mathbf{v}}(\omega)\mathbf{G}(\omega)]^{-1}\bar{\mathbf{h}}(\omega). \tag{B.38}$$

Furthermore, using $\bar{\mathbf{h}}(\omega) = \mathbf{G}^H(\omega)\mathbf{h}(\omega)$ in (B.38), the two-stage filter can finally be written as

$$\mathbf{w}_{\text{II}}(\omega) = \Phi_{\text{II}}(\omega)\mathbf{G}(\omega)[\mathbf{G}^H(\omega)\mathbf{R}_{\mathbf{v}}(\omega)\mathbf{G}(\omega)]^{-1}\mathbf{G}^H(\omega)\mathbf{h}(\omega). \tag{B.39}$$

and the narrowband output SNR for the two-stage filter $\text{oSNR}_{\text{II}}(\omega)$ can be derived by substituting $\mathbf{w}_{\text{II}}(\omega)$ in (B.28).

Relation between the one- and two-stage filters and their narrowband output SNRs

For an invertible matrix $\mathbf{G}(\omega)$, the two-stage filter in (B.39) simplifies to

$$\mathbf{w}_{\text{II}}(\omega) = \Phi_{\text{II}}(\omega)\mathbf{R}_{\mathbf{v}}^{-1}(\omega)\mathbf{h}(\omega). \tag{B.40}$$

In addition, since $\Phi_{\text{II}}(\omega) = \Phi_{\text{I}}(\omega)$ for an invertible matrix $\mathbf{G}(\omega)$, the one-stage and two-stage filters are equivalent and the one-stage and two-stage narrowband output SNRs are equal, i.e.,

$$\text{oSNR}_{\text{I}}(\omega) = \text{oSNR}_{\text{II}}(\omega). \tag{B.41}$$

However, depending on the dereverberation filter $\mathbf{G}(\omega)$ used to generate the dereverberated reference signal, a stable inverse dereverberation filter $\mathbf{G}^{-1}(\omega)$ does not necessarily exist. Consider as an illustrative example using the matched filter to achieve dereverberation as in [81], i.e.,

$$\mathbf{g}(\omega) = \frac{\mathbf{h}(\omega)}{\|\mathbf{h}(\omega)\|_2^2}. \tag{B.42}$$

Since acoustic transfer functions are mixed phase functions [35–37], with zeros clustering near the unit circle [34], it is highly likely that the zeros of the acoustic transfer functions will cause some of the matched filter coefficients in (B.42) to be 0, i.e., $G_m(\omega) = 0$. Clearly, in such a case the filter matrix $\mathbf{G}(\omega)$ is not invertible, such that

$$\mathbf{w}_{\text{II}}(\omega) = \Phi_{\text{II}}(\omega)\mathbf{G}(\omega)[\mathbf{G}^H(\omega)\mathbf{R}_{\mathbf{v}}(\omega)\mathbf{G}(\omega)]^{-1}\mathbf{G}^H(\omega)\mathbf{h}(\omega) \tag{B.43}$$

$$\neq \Phi(\omega)\mathbf{w}_{\text{I}}(\omega), \tag{B.44}$$

with $\Phi(\omega)$ an arbitrary scaling constant.[1] Since $\mathbf{w}_{\text{I}}(\omega)$ is the generalized eigenvector yielding the maximum value of the generalized Rayleigh quotient in (B.28), i.e., the maximum narrowband output SNR, any other vector not equal to (a scaled version of) $\mathbf{w}_{\text{I}}(\omega)$ will result in a smaller value of the generalized Rayleigh quotient. Hence, for a non-invertible matrix $\mathbf{G}(\omega)$,

$$\text{oSNR}_{\text{II}}(\omega) < \text{oSNR}_{\text{I}}(\omega), \tag{B.45}$$

i.e., the narrowband output SNR of the two-stage technique is smaller than the narrowband output SNR of the one-stage technique. Intuitively, the inequality in (B.45) is to be expected. Using a non-invertible dereverberation filter in the first stage implies disregarding one or more microphone signals in a given frequency bin (since the microphone signal is multiplied by zero). Hence, the MWF in this case will operate on fewer microphones, decreasing the spatial diversity, and as a result yielding a lower narrowband output SNR.

Summarizing, it can be said that the two-stage frequency domain technique in [87] is equivalent to applying a one-stage frequency domain MWF if and only if the filter used to generate the reference signal is invertible. For a non-invertible filter, using a one-stage filter is more advantageous since it yields a higher narrowband output SNR than the two-stage filter.

1 Note that for a non-invertible matrix $\mathbf{G}(\omega)$, also the filtered noise correlation matrix $\mathbf{R}_{\tilde{\mathbf{v}}}(\omega)$ is non-invertible and in a practical implementation one would either use the pseudo-inverse or diagonal loading.

BIBLIOGRAPHY

[1] M. Omologo, P. Svaizer, and M. Matassoni, "Environmental conditions and acoustic transduction in hands-free speech recognition," *Speech Communication*, vol. 25, no. 1–3, pp. 75–95, Aug. 1998.

[2] F. A. Everest, *Master handbook of acoustics.* New York, USA: McGraw-Hill, 2001.

[3] Y. Takata and A. K. Nabelek, "English consonant recognition in noise and in reverberation by Japanese and American listeners," *Journal of the Acoustical Society of America*, vol. 88, no. 2, pp. 663–666, Aug. 1990.

[4] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 331–342, Jul. 2006.

[5] A. Warzybok, J. Rennies, T. Brand, S. Doclo, and B. Kollmeier, "Effects of spatial and temporal integration of a single early reflection on speech intelligibility," *Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. 269–282, Jan. 2013.

[6] B. Champagne, S. Bedard, and A. A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 2, pp. 148–152, Mar. 1996.

[7] A. Sehr, "Reverberation modeling for robust distant-talking speech recognition," Ph.D. dissertation, Friedrich-Alexander-Universität Erlangen-Nürenberg, Erlangen, Germany, Oct. 2009.

[8] R. Maas, E. A. P. Habets, A. Sehr, and W. Kellermann, "On the application of reverberation suppression to robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Kyoto, Japan, Mar. 2012, pp. 297–300.

[9] 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 Processing Magazine*, vol. 29, no. 6, pp. 114–126, Nov. 2012.

[10] S. L. Gay and J. Benesty, Eds., *Acoustic signal processing for telecommunication.* Dordrecht, The Netherlands: Kluwer Academic Publishers, 2000.

[11] M. Brandstein and D. Ward, Eds., *Microphone arrays: Signal processing techniques and applications.* Berlin, Germany: Springer, 2001.

[12] G. M. Davis, Ed., *Noise reduction in speech applications.* New York, USA: CRC Press, 2002.

[13] J. Benesty, S. Makino, and J. Chen, Eds., *Speech enhancement*. Berlin, Germany: Springer, 2005.

[14] J. Benesty, M. M. Sondhi, and Y. A. Huang, Eds., *Springer handbook of speech processing*. Berlin, Germany: Springer, 2007.

[15] P. C. Loizou, *Speech enhancement: Theory and practice*. New York, USA: CRC Press, 2007.

[16] S. V. Vaseghi, *Advanced digital signal processing and noise reduction*. Hoboken, USA: John Wiley & Sons, 2008.

[17] J. Benesty, J. Chen, and Y. A. Huang, *Microphone array signal processing*. Berlin, Germany: Springer, 2008.

[18] J. Benesty, J. Chen, Y. A. Huang, and I. Cohen, *Noise reduction in speech processing*. Berlin, Germany: Springer, 2009.

[19] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[20] P. A. Naylor and N. D. Gaubitch, Eds., *Speech dereverberation*. London, UK: Springer, 2010.

[21] "Reverberant voice enhancement and recognition benchmark (REVERB) challenge," Florence, Italy, May 2014.

[22] H. Kuttruff, *Room acoustics*. New York, USA: Taylor & Francis, 2000.

[23] I. Dokmanic, Y. M. Lu, and M. Vetterli, "Can one hear the shape of a room: The 2-D polygonal case," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011, pp. 321–324.

[24] F. Antonacci, J. Filos, M. R. P. Thomas, E. A. P. Habets, A. Sarti, P. A. Naylor, and S. Tubaro, "Inference of room geometry from acoustic impulse responses," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2683–2695, Dec. 2012.

[25] A. H. Moore, M. Brookes, and P. A. Naylor, "Room geometry estimation from a single channel acoustic impulse response," in *Proc. European Signal Processing Conference*, Marrakech, Morocco, Sep. 2013.

[26] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 3233–3244, Jun. 2003.

[27] I. Arweiler and J. M. Buchholz, "The influence of spectral characteristics of early reflections on speech intelligibility," *Journal of the Acoustical Society of America*, vol. 130, no. 2, pp. 996–1005, Aug. 2011.

[28] K. S. Helfer and L. A. Wilber, "Hearing loss, aging, and speech perception in reverberation and noise," *Journal of Sound and Vibration*, vol. 33, no. 1, pp. 149–155, Mar. 1990.

[29] J. N. Mourjopoulos, "Digital equalization of room acoustics," *Journal of Audio Engineering Society*, vol. 42, no. 11, pp. 884–900, Nov. 1994.

[30] J. Mourjopoulos and M. Paraskevas, "Pole and zero modeling of room transfer functions," *Journal of Sound and Vibration*, vol. 146, no. 1, pp. 281–302, Apr. 1991.

[31] Y. Haneda, S. Makino, and Y. Kaneda, "Modeling of a room transfer function using common acoustical poles," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, San Francisco, USA, Mar. 1992, pp. 213–216.

[32] ——, "Common acoustical pole and zero modeling of room transfer functions," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 320–328, Apr. 1994.

[33] G. Vairetti, T. Van Waterschoot, M. Moonen, M. Catrysse, and S. H. Jensen, "Sparse linear parametric modeling of room acoustics with orthonormal basis functions," in *Proc. European Signal Processing Conference*, Lisbon, Portugal, Sep. 2014.

[34] C. P. Hughesa and A. Nikeghbalia, "The zeros of random polynomials cluster uniformly near the unit circle," *Compositio Mathematica*, vol. 144, no. 3, pp. 734–746, May 2008.

[35] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *The Journal of the Acoustical Society of America*, vol. 66, no. 1, pp. 165–169, Feb. 1979.

[36] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[37] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-domain single-channel inverse filtering for speech dereverberation: theory and practice," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 5214–5218.

[38] M. R. Schroeder, "Statistical parameters of the frequency response curves of large rooms," *Journal of the Acoustical Society of America*, vol. 35, no. 5, pp. 299–306, May 1987.

[39] J. D. Polack, "La transmission de l'energie sonore dans les salles," Ph.D. dissertation, Université du Maine, Le Mans, France, Dec. 1988.

[40] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Universiteit Eindhoven, Eindhoven, The Netherlands, Jun. 2007.

[41] Y. Ephraim, H. Lev-Ari, and W. J. J. Roberts, "A brief survey of speech enhancement," in *The electronics handbook*, J. C. Whitaker, Ed. New York, USA: CRC Press, 2005.

[42] I. Cohen and S. Gannot, "Spectral enhancement methods," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. A. Huang, Eds. Berlin, Germany: Springer, 2008.

[43] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-domain based single-microphone noise reduction for speech enhancement - A survey of the state of the art*. San Rafael, USA: Morgan & Claypool Publishers, 2013, vol. 9,

no. 1.

[44] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[45] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[46] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Washington, USA, Apr. 1979, pp. 208–211.

[47] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), Hidden Markov models and the projection, for robust speech recognition in cars," *Speech Communication*, vol. 11, no. 2-3, pp. 215–228, Jul. 1992.

[48] S. Kamath and P. C. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, USA, May 2002, pp. 4164–4164.

[49] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 126–137, Mar. 1999.

[50] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[51] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[52] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 341–351, Sep. 2002.

[53] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 857–869, Sep. 2005.

[54] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 1110–1126, Jan. 2005.

[55] C. H. You, S. N. Koh, and S. Rahardja, "Beta-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 475–486, Jul. 2005.

[56] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1741–1752, Aug 2007.

[57] K. Lebart and J. M. Boucher, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoustica*, vol. 87, no. 3, pp. 359–366, May-Jun. 2001.

[58] E. A. P. Habets, "Multi-channel speech dereverberation based on a statistical model of late reverberation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, USA, Mar. 2005, pp. 173–176.

[59] ——, "Speech dereverberation based on a statistical model of late reverberation using a linear microphone array," in *Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Arrays*, Piscataway, USA, Mar. 2005.

[60] I. Tashev and D. Allred, "Reverberation reduction for improved speech recognition," in *Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Arrays*, Piscataway, USA, Mar. 2005.

[61] E. A. P. Habets, S. Gannot, and I. Cohen, "Speech dereverberation using backward estimation of the late reverberant spectral variance," in *Proc. IEEE Convention of Electrical and Electronics Engineers in Israel*, Eilat, Israel, Dec. 2008, pp. 384–388.

[62] ——, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–774, Sep. 2009.

[63] 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 Journal on Advances in Signal Processing*, vol. 2015, no. 1, 2015.

[64] B. D. V. Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE Acoustics, Speech, and Signal Processing Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[65] M. Buck, "Aspects of first-order differential microphone arrays in the presence of sensor imperfections," *European Transactions on Telecommunications*, vol. 13, no. 2, pp. 115–122, Jun. 2002.

[66] G. Elko, "Superdirectional microphone arrays," in *Acoustic signal processing for telecommunication*, S. L. Gay and J. Benesty, Eds. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2000.

[67] D. B. Ward, R. A. Kennedy, and R. C. Williamson, "Theory and design of broadband sensor arrays with frequency invariant far-field beam patterns," *The Journal of the Acoustical Society of America*, vol. 97, no. 2, pp. 91–95, Feb. 1995.

[68] H. Cox, R. M. Zeskind, and T. Kooij, "Practical supergain," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 3, pp. 393–398, Jun. 1986.

[69] K. U. Simmer, J. Bitzer, and C. Marro, "Superdirective microphone arrays," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Berlin, Germany: Springer, 2001.

[70] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*,

vol. 30, no. 1, pp. 27–34, Jan. 1982.

[71] S. Gannot, D. Burshtein, and E. E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[72] A. Krueger, E. Warsitz, and R. Haeb-Umbach, "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 206–219, Jan. 2011.

[73] E. A. P. Habets and S. Gannot, "Dual-microphone speech dereverberation using a reference signal," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, USA, Apr. 2007, pp. 901–904.

[74] A. Schwarz, K. Reindl, and W. Kellermann, "On blocking matrix-based dereverberation for automatic speech recognition," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Aachen, Germany, Sep. 2012.

[75] ——, "A two-channel reverberation suppression scheme based on blind signal separation and Wiener filtering," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Kyoto, Japan, Mar. 2012, pp. 113–116.

[76] S. Braun and E. A. P. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proc. European Signal Processing Conference*, Marrakech, Morocco, Sep. 2013.

[77] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids," in *Proc. European Signal Processing Conference*, Lisbon, Portugal, Sep. 2014.

[78] A. Kuklasiński, S. Doclo, T. Gerkmann, S. H. Jensen, and J. Jensen, "Multi-channel PSD estimators for speech dereverberation - a theoretical and experimental comparison," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Brisbane, Australia, Apr. 2015, pp. 91–95.

[79] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 240–251, Feb. 2015.

[80] S. Wisdom, T. Powers, L. Atlas, and J. Pitton, "Enhancement of reverberant and noisy speech by extending its coherence," in *Proc. REVERB challenge workshop*, Florence, Italy, May 2014.

[81] S. Doclo and M. Moonen, "Combined frequency-domain dereverberation and noise reduction technique for multi-microphone speech enhancement," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Darmstadt, Germany, Sep. 2001, pp. 31–34.

[82] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Berlin, Germany: Springer, 2001.

[83] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Transactions on Signal Processing*,

vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

[84] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004.

[85] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Communication*, vol. 49, no. 7-8, pp. 636–656, Jul. 2007.

[86] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. J. R. Liu, Eds. Hoboken, USA: John Wiley & Sons, 2010.

[87] E. A. P. Habets and J. Benesty, "A two-stage beamforming approach for noise reduction and dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 945–958, May 2013.

[88] J. R. Hopgood and P. J. W. Rayner, "A probabilistic framework for subband autoregressive models applied to room acoustics," in *Proc. IEEE Signal Processing Workshop on Statistical Signal Processing*, Singapore, Aug. 2001, pp. 492–495.

[89] ——, "Blind single channel deconvolution using nonstationary signal processing," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 476–488, Sep. 2003.

[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. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, USA, Mar. 2008, pp. 85–88.

[91] ——, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

[92] D. Schmid, G. Enzner, S. Malik, D. Kolossa, and R. Martin, "Variational Bayesian inference for multichannel dereverberation and noise reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 8, pp. 1320–1335, Aug. 2014.

[93] B. Schwartz, S. Gannot, and E. A. P. Habets, "Online speech dereverberation using Kalman filter and EM algorithm," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 394–406, Feb. 2015.

[94] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 69–84, Jan. 2011.

[95] A. Jukić and S. Doclo, "Speech dereverberation using weighted prediction error with Laplacian model of the desired signal," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 5172–5176.

[96] A. Jukić, T. Van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015.

[97] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, "Optimized speech dereverberation from probabilistic perspective for time varying acoustic transfer function," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1369–1380, Jul. 2013.

[98] A. Jukić, T. Van Waterschoot, T. Gerkmann, and S. Doclo, "Speech dereverberation with convolutive transfer function approximation using MAP and variational deconvolution approaches," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Antibes, France, Sep. 2014, pp. 50–54.

[99] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 231–246, Feb. 2009.

[100] N. Ito, S. Araki, and T. Nakatani, "Probabilistic integration of diffuse noise suppression and dereverberation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 5167–5171.

[101] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*. New Jersey, USA: Prentice Hall, 1993.

[102] Y. A. Huang and J. Benesty, "Adaptive multi-channel least mean square and Newton algorithms for blind channel identification," *Signal Processing*, vol. 82, no. 8, pp. 1127–1138, Aug. 2002.

[103] ——, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Transactions on Signal Processing*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[104] M. K. Hasan, N. M. Hossain, and P. A. Naylor, "Autocorrelation model-based identification method for ARMA systems in noise," *IEE Proceedings on Vision, Image and Signal Processing*, vol. 152, no. 5, pp. 520–526, Oct. 2005.

[105] N. D. Gaubitch, T. Hasan, and P. A. Naylor, "Noise robust adaptive blind channel identification using spectral constraints," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, May 2006.

[106] M. A. Haque and T. Hasan, "Noise robust multichannel frequency-domain LMS algorithms for blind channel identification," *IEEE Signal Processing Letters*, vol. 15, pp. 305–308, Feb. 2008.

[107] M. Hu, N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "Noise robust blind system identification algorithms based on a Rayleigh quotient cost function," in *Proc. European Signal Processing Conference*, Nice, France, Sep. 2015.

[108] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2007.

[109] B. D. Radlovic, R. C. Williamson, and R. A. Kennedy, "Equalization in an acoustic reverberant environment: robustness results," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 311–319, May 2000.

[110] K. Hasan and P. A. Naylor, "Analyzing effect of noise on LMS-type approaches to blind estimation of SIMO channels: robustness issue," in *Proc. European Signal Processing Conference*, Florence, Italy, Sep. 2006.

[111] F. Lim, W. Zhang, E. A. P. Habets, and P. A. Naylor, "Robust multichannel dereverberation using relaxed multichannel least squares," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1379–1390, Jun. 2014.

[112] J. Mourjopoulos, P. Clarkson, and J. Hammond, "A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Paris, France, May 1982, pp. 1858–1861.

[113] B. D. Radlovic and R. A. Kennedy, "Nonminimum-phase equalization and its subjective importance in room acoustics," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 728–737, Nov. 2000.

[114] P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Eindhoven, The Netherlands, Sep. 2005.

[115] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, USA, Apr. 2008, pp. 4897–4900.

[116] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[117] H. Hacihabibouglu and Z. Cvetkovic, "Multichannel dereverberation theorems and robustness issues," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 67–689, Feb. 2012.

[118] W. Zhang, A. W. H. Khong, and P. A. Naylor, "Adaptive inverse filtering of room acoustics," in *Proc. Asilomar Conference on Signals, Systems and Computers*, Oct. 2008, pp. 788–792.

[119] R. S. Rashobh and A. W. H. Khong, "A variable step-size multichannel equalization algorithm exploiting sparseness measure for room acoustics," in *IEEE International Symposium on Circuits and Systems*, May 2012, pp. 2753–2756.

[120] ——, "Adaptive multichannel equalization applied to room acoustics exploiting the sparsity of target response," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 5162–5166.

[121] ——, "A fast frequency-domain algorithm for equalizing acoustic impulse responses," *IEEE Signal Processing Letters*, vol. 19, no. 12, pp. 797–800, Dec. 2012.

[122] R. S. Rashobh, A. W. H. Khong, and D. Liu, "Multichannel equalization in the KLT and frequency domains with application to speech dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22,

no. 3, pp. 634–646, Mar. 2014.

[123] N. D. Gaubitch and P. A. Naylor, "Equalization of multichannel acoustic systems in oversampled subbands," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1061–1070, Aug. 2009.

[124] M. Kallinger and A. Mertins, "Multi-channel room impulse response shaping - a study," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, May 2006, pp. 101–104.

[125] W. Zhang, E. A. P. Habets, and P. A. Naylor, "On the use of channel shortening in multichannel acoustic system equalization," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Tel Aviv, Israel, Sep. 2010.

[126] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1879–1890, Sep. 2013.

[127] I. Kodrasi and S. Doclo, "Robust partial multichannel equalization techniques for speech dereverberation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Kyoto, Japan, Mar. 2012, pp. 537–540.

[128] I. Kodrasi, S. Goetze, and S. Doclo, "A perceptually constrained channel shortening technique for speech dereverberation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 2013, pp. 151–155.

[129] I. Kodrasi and S. Doclo, "Regularized subspace-based acoustic multichannel equalization for speech dereverberation," in *Proc. European Signal Processing Conference*, Marrakech, Morocco, Sep. 2013.

[130] ——, "The effect of inverse filter length on the robustness of acoustic multichannel equalization," in *Proc. European Signal Processing Conference*, Bucharest, Romania, Aug. 2012.

[131] I. Kodrasi, S. Goetze, and S. Doclo, "Increasing the robustness of acoustic multichannel equalization by means of regularization," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Aachen, Germany, Sep. 2012, pp. 161–164.

[132] ——, "Non-intrusive regularization for least-squares multichannel equalization for speech dereverberation," in *Proc. IEEE Convention of Electrical and Electronics Engineers in Israel*, Eilat, Israel, Nov. 2012.

[133] I. Kodrasi, A. Jukić, and S. Doclo, "Robust sparsity-promoting acoustic multichannel equalization for speech dereverberation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Shanghai, China, Mar. 2016.

[134] I. Kodrasi and S. Doclo, "Robust acoustic multi-channel equalization for speech dereverberation using signal-dependent penalty functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016, manuscript submitted for publication.

[135] I. Kodrasi, B. Cauchi, S. Goetze, and S. Doclo, "Objective and subjective evaluation of robust acoustic multi-channel equalization," *Journal of the Audio*

Engineering Society, 2016, manuscript submitted for publication.

[136] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multichannel equalization," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Antibes, France, Sep. 2014, pp. 139–143.

[137] ——, "Incorporating the noise statistics in acoustic multi-channel equalization," in *Proc. AES International Conference on Dereverberation and Reverberation of Audio, Music, and Speech*, Leuven, Belgium, Feb. 2016.

[138] ——, "Joint dereverberation and noise reduction based on acoustic multichannel equalization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 680–693, Apr. 2016.

[139] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.

[140] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-time processing of speech signals*. New York, USA: Macmillan Publishing Company, 1993.

[141] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing*. New Jersey, USA: Prentice Hall, 2009.

[142] B. V. Veen and K. M. Buckley, *Beamforming techniques for spatial filtering*. New York, USA: CRC Press, 1999.

[143] M. Fozunbal, T. Kalker, and R. W. Schafer, "Multi-channel echo control by model learning," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Seattle, USA, Sep. 2008.

[144] T. Koren, R. Talmon, and I. Cohen, "Supervised system identification based on local PCA models," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Kyoto, Japan, Mar. 2012, pp. 541–544.

[145] F. Lim and P. Naylor, "Statistical modelling of multichannel blind system identification errors," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Antibes, France, Sep. 2014, pp. 119–123.

[146] J. Cho, D. Morgan, and J. Benesty, "An objective technique for evaluating doubletalk detectors in acoustic echo cancelers," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 718–724, Nov. 1999.

[147] W. Zhang and P. A. Naylor, "An algorithm to generate representations of system identification errors," *Research Letters in Signal Processing*, vol. 2008, Jan. 2008.

[148] J. O. Jungmann, R. Mazur, M. Kallinger, M. Tiemin, and A. Mertins, "Combined acoustic MIMO channel crosstalk cancellation and room impulse response reshaping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1829–1842, Aug. 2012.

[149] J. O. Jungmann, R. Mazur, and A. Mertins, "Perturbation of room impulse responses and its application in robust listening room compensation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 2013, pp. 433–437.

[150] D. R. Morgan, J. Benesty, and M. M. Sondhi, "On the evaluation of estimated impulse responses," *IEEE Signal Processing Letters*, vol. 5, no. 7, pp. 174–176,

Jul. 1998.

[151] S. Goetze, E. Albertin, J. Rennies, E. A. P. Habets, and K. D. Kammeyer, "Speech quality assessment for listening-room compensation," in *Proc. AES International Conference on Sound Quality Evaluation*, Pitea, Sweden, Jun. 2010, pp. 11–20.

[152] S. Goetze, A. Warzybok, I. Kodrasi, J. O. Jungmann, B. Cauchi, J. Rennis, E. A. P. Habets, A. Mertins, T. Gerkmann, S. Doclo, and B. Kollmeier, "A study on speech quality and speech intelligibility measures for quality assessment of single-channel dereverberation algorithms," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Antibes, France, Sep. 2014, pp. 234–238.

[153] ITU-T, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs P.862*, International Telecommunications Union (ITU-T) Recommendation, Feb. 2001.

[154] S. Quackenbush, T. Barnwell, and M. Clements, *Objective measures of speech quality*. New Jersey, USA: Prentice-Hall, 1988.

[155] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[156] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Paris, France, Sep. 2006.

[157] G. Harikumar and Y. Bresler, "FIR perfect signal reconstruction from multiple convolutions: minimum deconvolver orders," *IEEE Transactions on Signal Processing*, vol. 46, no. 1, pp. 215–218, Jan. 1998.

[158] R. K. Martin, K. Vanbleu, M. Ding, G. Ysebaert, M. Milosevic, B. L. Evans, M. Moonen, and C. R. Johnson, "Unification and evaluation of equalization structures and design algorithms for discrete multitone modulation systems," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3880–3894, Oct. 2005.

[159] I. R. Shafarevich and A. O. Remizov, *Linear algebra and geometry*. Berlin, Germany: Springer, 2010.

[160] G. Golub and C. Van Loan, *Matrix Computations*. Baltimore, USA: The John Hopkins University Press, 1996.

[161] J. Wen, N. D. Gaubitch, E. A. P. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Paris, France, Sep. 2006.

[162] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proc. AES Convention*, Paris, France, Feb. 2000, pp. 18–22.

[163] M. Nilsson, S. D. Soli, and A. Sullivan, "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise,"

Journal of the Acoustical Society of America, vol. 95, no. 2, pp. 1085–1099, Feb. 1994.

[164] P. Wedin, "Perturbation theory for pseudo-inverses," *BIT Numerical Mathematics*, vol. 13, no. 2, pp. 217–232, Jun. 1973.

[165] R. A. Horn and C. R. Johnson, *Topics in matrix analysis*. Cambridge, United Kingdom: Cambridge University Press, 1999.

[166] G. W. Stewart, "Perturbation bounds for the definite generalized eigenvalue problem," *Linear Algebra and its Applications*, vol. 23, pp. 69–85, Feb. 1979.

[167] J.-F. Cardoso and A. Souloumiac, "Jacobi angles for simultaneous diagonalization," *SIAM Journal on Matrix Analysis and Applications*, vol. 17, no. 1, pp. 161–164, Jan. 1996.

[168] The Mathworks Inc. Matlab Optimization Toolbox User's Guide. [Online]. Available: http://www.mathworks.com/help/pdf_doc/optim/optim_tb.pdf

[169] P. C. Hansen, "Analysis of discrete ill-posed problems by means of the L-curve," *SIAM review*, vol. 34, no. 4, pp. 561–580, Dec. 1992.

[170] P. C. Hansen and D. P. O'Leary, "The use of the L-curve in the regularization of discrete ill-posed problems," *SIAM Journal on Scientific Computing*, vol. 14, no. 6, pp. 1487–1503, Nov. 1993.

[171] S. Sternberg, *Curvature in mathematics and physics*. New York, USA: Dover Pubications, 2012.

[172] J. L. Castellanos, S. Gómez, and V. Guerra, "The triangle method for finding the corner of the L-curve," *Applied Numerical Mathematics*, vol. 43, no. 4, pp. 359–373, Dec. 2002.

[173] S. Makino, S. Araki, and H. Sawada, "Underdetermined blind source separation using acoustic arrays," in *Handbook on array processing and sensor networks*, S. Haykin and K. J. R. Liu, Eds. Hoboken, USA: John Wiley & Sons, 2010.

[174] J. W. Shin, J.-H. Chang, and N. S. Kim, "Statistical modeling of speech signals based on generalized Gamma distribution," *IEEE Signal Processing Letters*, vol. 12, no. 3, pp. 258–261, Mar. 2005.

[175] K. Kumatani, J. McDonough, B. Rauch, D. Klakow, P. N. Garner, and L. Weifeng, "Beamforming with a maximum negentropy criterion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 994–1008, Jul. 2009.

[176] I. Tashev and A. Acero, "Statistical modeling of the speech signal," in *Proc. International Workshop on Acoustic, Echo, and Noise Control*, Tel Aviv, Israel, Sep. 2010.

[177] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," in *Proc. International Workshop on Acoustic, Echo, and Noise Control*, Tel Aviv, Israel, Sep. 2010.

[178] P. Bofill and M. Zibulevsky, "Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform," in *Proc. International Workshop on Independent Component Analysis and Blind Signal Separation*, Helsinki, Finland, Jun. 2000, pp. 87–92.

[179] M. Kowalski, E. Vincent, and R. Gribonval, "Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1818–1829, Sep. 2010.

[180] S. Arberet, P. Vandergheynst, R. E. Carrillo, J. Thiran, and Y. Wiaux, "Sparse reverberant audio source separation via reweighted analysis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 7, pp. 1391–1402, Mar. 2013.

[181] H. Kameoka, T. Nakatani, and T. T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Apr. 2009, pp. 45–48.

[182] T. Van Waterschoot, B. Defraene, M. Diehl, and M. Moonen, "Embedded optimization algorithms for multi-microphone dereverberation," in *Proc. European Signal Processing Conference*, Marrakech, Morocco, Sep. 2013.

[183] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM Journal on Computing*, vol. 24, no. 2, pp. 227–234, Apr. 1995.

[184] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York, USA: Springer, 2006.

[185] R. Chartrand, "Shrinkage mappings and their induced penalty functions," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 1026–1029.

[186] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, United Kingdom: Cambridge University Press, 2004.

[187] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[188] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[189] E. J. Candés, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted l_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 877–905, Oct. 2008.

[190] M. P. Friedlander, H. Mansour, R. Saab, and O. Yilmaz, "Recovering compressively sampled signals using partial support information," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 1122–1134, Feb. 2012.

[191] S. P. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[192] R. Chartrand, E. Y. Sidky, and X. Pan, "Nonconvex compressive sensing for X-ray CT: an algorithm comparison," in *Proc. Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, USA, Nov. 2013, pp. 665–669.

[193] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & Mathematics with Applications*, vol. 2, no. 1, pp. 17–40, May 1976.

[194] M. J. Fadili and J. L. Starck, "Monotone operator splitting for optimization problems in sparse recovery," in *Proc. IEEE International Conference on Image Processing*, San Diego, USA, Nov. 2009, pp. 1461–1464.

[195] Y. Wang, J. Yang, W. Yin, and Y. Zhang, "A new alternating minimization algorithm for total variation image reconstruction," *SIAM Journal on Imaging Sciences*, vol. 1, no. 3, pp. 248–272, Aug. 2008.

[196] T. Goldstein and S. Osher, "The split Bregman method for L1-regularized problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 323–343, Apr. 2009.

[197] N. Parikh and S. P. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, Jan. 2014.

[198] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Antibes, France, Sep. 2014, pp. 313–317.

[199] ITU-T, *Method for the subjective assessment of intermediate quality levels of coding systems*, International Telecommunications Union (ITU-T) Recommendation, Jan. 2003.

[200] A. Warzybok, I. Kodrasi, J. O. Jungmann, E. A. P. Habets, T. Gerkmann, A. Mertins, S. Doclo, B. Kollmeier, and S. Goetze, "Subjective speech quality and speech intelligibility evaluation of single-channel dereverberation algorithm," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Antibes, France, Sep. 2014, pp. 333–337.

[201] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, Dec. 1965.

[202] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[203] M. Belge, M. Kilmer, and E. L. Miller, "Simultaneous multiple regularization parameter selection by means of the L-hypersurface with applications to linear inverse problems posed in the wavelet transform domain," *SPIE 3459, Bayesian Inference for Inverse Problems*, vol. 328, Sep. 1998.

[204] ——, "Efficient determination of multiple regularization parameters in a generalized L-curve framework," *Inverse Problems*, vol. 18, no. 4, pp. 1161–1183, Jul. 2002.

[205] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda, "Estimating direct-to-reverberant energy ratio using D/R spatial correlation matrix model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2374–2384, Nov. 2011.

[206] O. Thiergart, M. Taseska, and E. A. P. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2182–2196, Dec. 2014.

[207] J. Eaton, A. Moore, P. A. Naylor, and J. Skoglund, "Direct-to-reverberant ratio estimation using a null-steered beamformer," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Brisbane, Australia,

Apr. 2015, pp. 46–50.

[208] T. H. Falk, Z. Chenxi, and C. Wai-Yip, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010.

[209] O. Kirkeby, P. A. Nelson, H. Hamada, and F. Orduna-Bustamante, "Fast deconvolution of multichannel systems using regularization," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 189–194, Mar. 1998.

[210] O. Kirkeby and P. A. Nelson, "Digital filter design for inversion problems in sound reproduction," *Journal of Audio Engineering Society*, vol. 47, no. 7/8, pp. 583–595, Jul. 1999.

[211] J. N. Higham, *Accuracy and Stability of Numerical Algorithms*. Philadelphia, USA: SIAM, 2002.

LIST OF PUBLICATIONS

Peer-reviewed Journal Papers

[1] I. Kodrasi and S. Doclo, "Robust acoustic multi-channel equalization for speech dereverberation using signal-dependent penalty functions,"*IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016, manuscript submitted for publication.

[2] I. Kodrasi, B. Cauchi, S. Goetze, and S. Doclo, "Objective and subjective evaluation of robust acoustic multi-channel equalization," *Journal of the Audio Engineering Society*, 2016, manuscript submitted for publication.

[3] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multi-channel equalization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 680–693, Apr. 2016.

[4] 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 Journal on Advances in Signal Processing*, vol. 2015, no. 1, 2015.

[5] I. Kodrasi, S. Goetze, and S. Doclo "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1879–1890, Sep. 2013.

Peer-reviewed Conference Papers

[1] I. Kodrasi, A. Jukić, and S. Doclo, "Robust sparsity-promoting acoustic multichannel equalization for speech dereverberation,", in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Shanghai, China, Mar. 2016.

[2] I. Kodrasi and S. Doclo, "Incorporating the noise statistics in acoustic multichannel equalization," in *Proc. 60-th Audio Engineering Society International Conference on Dereverberation and Reverberation of Audio, Music, and Speech*, Leuven, Belgium, Feb. 2016.

[3] I. Kodrasi, D. Marquardt, and S. Doclo, "Curvature-based optimization of the trade-off parameter in the speech distortion weighted multichannel Wiener filter," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Brisbane, Australia, Apr. 2015, pp. 315–319.

[4] A. Warzybok, I. Kodrasi, J. O. Jungmann, E. A. P. Habets, T. Gerkmann, A. Mertins, S. Doclo, B. Kollmeier, and S. Goetze, "Subjective speech qual-

ity and speech intelligibility evaluation of single-channel dereverberation algorithm," in *Proc. International Workshop on Acoustic Signal Enhancement*, Antibes, France, Sep. 2014, pp. 333–337.

[5] S. Goetze, A. Warzybok, I. Kodrasi, J. O. Jungmann, B. Cauchi, J. Rennies, E. A. P. Habets, A. Mertins, T. Gerkmann, S. Doclo, and B. Kollmeier, "A study on speech quality and speech intelligibility measures for quality assessment of single-channel dereverberation algorithms," in *Proc. International Workshop on Acoustic Signal Enhancement*, Antibes, France, Sep. 2014, pp. 234–238.

[6] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multichannel equalization," in *Proc. International Workshop on Acoustic Signal Enhancement*, Antibes, France, Sep. 2014, pp. 140–144.

[7] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-domain single-channel inverse filtering for speech dereverberation: theory and practice," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 5214–5218.

[8] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Joint dereverberation and noise reduction using beamforming and a single-channel speech enhancement scheme," in *Proc. REVERB Challenge Workshop*, Florence, Italy, May 2014.

[9] I. Kodrasi and S. Doclo, "Regularized subspace-based acoustic multichannel equalization for speech dereverberation," in *Proc. European Signal Processing Conference*, Marrakech, Morocco, Sep. 2013.

[10] I. Kodrasi, S. Goetze, and S. Doclo, "A perceptually constrained channel shortening technique for speech dereverberation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 2013, pp. 151–155.

[11] I. Kodrasi, S. Goetze, and S. Doclo, "Non-intrusive regularization for least-squares multichannel equalization for speech dereverberation," in *Proc. IEEE Convention of Electrical and Electronics Engineers in Israel*, Eilat, Israel, Nov. 2012.

[12] I. Kodrasi, S. Goetze, and S. Doclo, "Increasing the robustness of acoustic multichannel equalization by means of regularization," in *Proc. International Workshop on Acoustic Signal Enhancement*, Aachen, Germany, Sep. 2012, pp. 161–164.

[13] I. Kodrasi and S. Doclo, "The effect of inverse filter length on the robustness of acoustic multichannel equalization," in *Proc. European Signal Processing Conference*, Bucharest, Romania, Aug. 2012.

[14] I. Kodrasi and S. Doclo, "Robust partial multichannel equalization techniques for speech dereverberation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, Mar. 2012, pp. 537–540.

[15] I. Kodrasi, T. Rohdenburg, and S. Doclo, "Microphone position optimization for planar superdirective beamforming," in *Proc. IEEE International Confer-*

ence on Acoustics, Speech and Signal Processing, Prague, Czech Republic, May 2011, pp. 109–112.