SPARSE MULTI-CHANNEL LINEAR PREDICTION FOR BLIND SPEECH DEREVERBERATION

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

von
Ante Jukić
geboren am 26. April 1986
in Imotski (Kroatien)

Ante Jukić: *Sparse Multi-Channel Linear Prediction for Blind Speech Dereverberation*

ERSTGUTACHTER:
Prof. Dr. ir. Simon Doclo, *University of Oldenburg, Germany*

WEITERE GUTACHTER:
Prof. Dr.-Ing. Timo Gerkmann, *University of Hamburg, Germany*
Prof. Dr. Sharon Gannot, *Bar-Ilan University, Israel*

TAG DER DISPUTATION:
30. Oktober 2017

ACKNOWLEDGMENTS

This thesis has been written at the Signal Processing Group at the Department of Medical Physics and Acoustics of the University of Oldenburg. I would like to take this opportunity to express my gratitude for the help and support of many people who have contributed to its completion.

First, I want to thank my supervisor Simon Doclo for giving me the opportunity to write this thesis and providing me with valuable guidance and support through the years. I am deeply grateful that he gave me the freedom to pursue my interests while providing numerous ideas, and he has helped me to improve my work in many different ways. I would also like to thank Timo Gerkmann for interesting discussions and influential advice. I am grateful to Sharon Gannot for his interest in my work and his constructive and insightful feedback. Furthermore, I want to thank Steven van de Par for participating in my thesis committee and his valuable comments.

This research was mainly funded by the DREAMS project and being involved in this project was a great professional and personal experience. It has given me the opportunity to continuously meet with talented researchers and great people, and I have benefited from the many interesting and fun discussions. I am indebted to Toon van Waterschoot for co-supervising my work and always giving perceptive feedback, and to his whole team at KU Leuven for making my stay in their lab a very memorable experience. Also, many thanks to Aldona for helping with a plethora of practical arrangements.

I am particularly grateful to all current and former members of the Signal Processing Group. They have supported me when I first moved to Germany and we have had a lot of fun over the past years. I would especially like to thank Henning, Daniel, Ina, and Ali, for their continuous help and, even more, for all the laughs we had over the countless coffee breaks, football evenings and flight layovers.

I want to thank all of my colleagues at the Department of Medical Physics and Acoustics, especially Katja and Renata for helping with all departmental matters. Also, I would like to thank Stefan Goetze and Benjamin Cauchi from Fraunhofer Institute for Digital Media and Technology for the discussions and collaboration.

Most importantly, I would like to thank my family for their unlimited trust, encouragement and the everlasting support.

Oldenburg, November 2017
Ante Jukić

ABSTRACT

In many speech communication applications, such as hands-free telephony and hearing aids, the microphones are located at a distance from the speaker. Therefore, in addition to the desired speech signal, the microphone signals typically contain undesired reverberation and noise, caused by acoustic reflections and undesired sound sources. Since these disturbances tend to degrade the quality of speech communication, decrease speech intelligibility and negatively affect speech recognition, efficient dereverberation and denoising methods are required.

This thesis deals with blind dereverberation methods, not requiring any knowledge about the room impulse responses between the speaker and the microphones. More specifically, we propose a general framework for blind speech dereverberation based on multi-channel linear prediction (MCLP) and exploiting sparsity of the speech signal in the time-frequency domain.

Firstly, we consider the noiseless case and propose batch speech dereverberation methods based on the subband MCLP-based signal model and a general sparse prior for the desired speech signal coefficients in the time-frequency domain. We propose a single-output method using a variational representation of the sparse prior that promotes sparsity of the desired speech signal coefficients across time and estimate the prediction filter by maximizing the likelihood function using an iteratively reweighted least squares algorithm. We analytically show that the proposed method generalizes the conventional MCLP-based dereverberation method based on a time-varying Gaussian model. We also show that the proposed method can be formulated as the minimization of the non-convex ℓ_p-norm of the desired speech signal coefficients. Furthermore, we propose a multiple-output extension using a group sparse cost function, which promotes sparsity over time and takes into account grouping of the coefficients across the channels. We use the non-convex mixed $\ell_{p,2}$-norm as the cost function and derive the corresponding iteratively reweighted least squares algorithm. Simulations results show that the proposed sparse MCLP-based methods with a non-convex sparsity-promoting cost function result in a better dereverberation performance than the conventional MCLP-based methods based on the time-varying Gaussian model.

Secondly, we consider adaptive speech dereverberation methods which are suitable for online processing and dynamic scenarios, e.g., when the desired speaker or the microphones are moving. We first extend the proposed batch sparse MCLP-based dereverberation methods to adaptive methods and derive the corresponding recursive least squares algorithm. Since these adaptive dereverberation methods may lead to overestimation of the undesired speech signal and hence distortions of the desired speech signal, we propose to constrain the power of the estimated undesired speech signal, leading to constrained sparse MCLP for adaptive speech dereverberation. Moreover, we propose to reduce the computational complexity of the adaptive methods by using a diagonal approximation. Simulation results show that the proposed constrained sparse MCLP increases the robustness to the selection of the forgetting factor and the filter length, which can be especially advantageous in dynamic scenarios when the filters need to adapt quickly or the optimal parameters are not known.

Thirdly, we propose a general framework for speech dereverberation that includes the subband MCLP-based signal model in the time-frequency domain and the wideband MCLP-based signal model in the time domain. More specifically, we formulate different optimization problems for speech dereverberation using either the wideband or the subband signal model with a sparse analysis or synthesis prior for the desired speech signal coefficients in the time-frequency domain. Moreover, we propose to incorporate the structure of the speech signal by exploiting neighborhood or low-rank structure in the time-frequency domain. Simulation results show that all proposed formulations result in a high dereverberation performance, with the wideband signal model with analysis sparsity leading to the best result. Furthermore, it is shown that incorporating the speech structure enables to further improve the performance of the proposed methods.

Finally, we consider the influence of additive noise and propose a framework for sparsity-based joint dereverberation and denoising using the subband signal model. More specifically, we propose a denoising method using a mixed-norm as the sparsity-promoting cost function and imposing a bound for the noise energy. Furthermore, we propose a joint dereverberation and denoising method by including the noise term in the MCLP-based signal model and imposing a bound for the noise energy. Simulation results show that the proposed joint method results in a significantly improved performance for low SNRs compared to the group sparse MCLP-based dereverberation method. Furthermore, simulation results show that a good performance can also be obtained using a two-stage procedure, combining sparse MCLP-based dereverberation with sparsity-based denoising.

ZUSAMMENFASSUNG

Bei vielen Anwendungen der Sprachkommunikation, wie z.B. bei der freihändigen Telefonie und bei Hörgeräten, befinden sich die Mikrofone in einer gewissen Entfernung zu dem Sprecher. Daher enthalten die Mikrofonsignale neben der gewünschten Sprache auch aus unerwünschten Nachhall, welcher durch akustische Reflektionen erzeugt wird, sowie aus unerwünschte Störgeräusche. Da diese Störungen dazu häufig die Sprachqualität und Sprachverständlichkeit reduzieren und einen negativen Effekt auf automatische Spracherkennung haben, sind Methoden zur Enthallung und Störgeräuschunterdrückung notwendig.

Diese Thesis betrachtet Methoden zur blinden Enthallung, d.h. es wird kein Wissen über die Raumimpulsantworten zwischen dem Sprecher und den Mikrofonen vorausgesetzt. Genauer stellen wir ein generalisiertes Framework zur blinden Enthallung vor, welches auf mehrkanaliger linearen Vorhersage (engl. multi-channel linear prediction, MCLP) basiert und die Spärlichkeit des Sprachsignals im Zeit-Frequenz-Bereich ausnutzt.

Als erstes nehmen wir an, dass kein Störgeräusch vorhanden ist und stellen Methoden zur batchverarbeitenden Enthallung, basierend auf einem Teilband MCLP-Signalmodell und der Spärlichkeitsannahme der Sprachkoeffizienten im Zeit-Frequenz-Bereich vor. Wir stellen eine Methode mit einem Ausgangskanal vor, bei der wir eine Variationsrepräsentation der Spärlichkeitsannahme der Sprachkoeffizienten über die Zeit benutzen und bei der wir das Vorhersagefilter schätzen indem die Likelihood Funktion mittels eines iterativ neu gewichteten Kleinste-Quadrate Algorithmus maximiert wird. Wir zeigen analytisch, dass die hier vorgestellte Methode die konventionelle auf MCLP basierende Methode zur Enthallung generalisiert, welche auf einem zeitvariantem Gaussmodell basiert. Wir zeigen außerdem, dass die vorgestellte Methode als Minimierung der nicht-konvexen ℓ_p-Norm der Sprachkoeffizienten formuliert werden kann. Des Weiteren stellen wir eine Erweiterung mit mehrkanaligem Ausgang vor, bei der wir eine "group sparse" Kostenfunktion verwenden, welche die Spärlichkeit der Sprachkoeffizienten über die Zeit einführt und die die kanalübergreifende Gruppierung der Koeffizienten in Betracht zieht. Wir benutzen die nicht-konvexe, gemischte $\ell_{p,2}$-Norm als Kostenfunktion und leiten den dazugehörigen iterativ neu gewichteten Kleinste-Quadrate Algorithmus her. Simulationen zeigen, dass die hier vorgestellte MCLP-Methode in einer besseren Enthallungs-Performance resultiert, als die konventionelle, auf zeitvariantem Gaussmodell basierende MCLP-Methode.

Als zweites betrachten wir adaptive Methoden zur Enthallung von Sprache, welche für die Echtzeitanwendungen sowie dynamische Szenarien geeignet sind, z.B. wenn sich der gewünschte Sprecher oder die Mikrofone bewegen. Dabei erweitern wir

zuerst die vorgestellte batchverarbeitende, auf Spärlichkeitsannahme und MCLP basierende Enthallungs-Methode hin zu adaptiven Methoden und leiten den rekursiven Kleinste-Quadrate Algorithmus her. Da die resultierenden adaptiven Methoden dazu die Störgeräusche überschätzen könnten und somit zu Sprachverzerrungen führen, schlagen wir eine Begrenzung der geschätzten Leistung der Störgeräusche vor, was zur begrenzten, auf Spärlichkeit basierenden MCLP-Methode für adaptive Enthallung von Sprache führt. Des Weiteren schlagen wir vor die rechnerische Komplexität der adaptiven Methoden zu reduzieren, indem eine diagonale Approximation verwendet wird. Simulationen zeigen, dass die vorgestellte Methode die Robustheit gegenüber der Auswahl der Gedächtnisfaktoren und der Filterlänge erhöht, was gerade in dynamischen Szenarien vorteilhaft ist, in denen sich die Filter schnell adaptieren müssen oder wenn die optimalen Parameter nicht bekannt sind.

Als drittes stellen wir ein generelles Framework zur Enthallung von Sprache vor, das das auf Teilband-MCLP basierende Signalmodell im Zeit-Frequenz-Bereich und das auf Breitband-MCLP basierende Signalmodell im Zeitbereich enthält. Genauer formulieren wir verschiedene Optimierungsprobleme zur Enthallung von Sprache indem entweder das Breitband- oder das Teilband-Signalmodell mit einer spärliche Analysen- oder Synthesenannahme für die Sprachkoeffizienten im Zeit-Frequenz-Bereich verwendet wird. Außerdem schlagen wir vor die Struktur des Sprachsignals mit einzubeziehen, indem die Nachbarschafts- oder Niedrigrank-Struktur im Zeit-Frequenz-Bereich ausgenutzt wird. Simulationen zeigen, dass alle vorgestellten Formulierungen in einer hohen Enthallungs-Performance resultieren, wobei das Breitband-Signalmodell mit Analysenspärlichkeit zur besten Performance führt. Des Weiteren wird gezeigt, dass die Berücksichtigung der Strukturen im Sprachsignal zu einer weiteren Verbesserung der Performance der vorgestellten Methoden führt.

Zuletzt ziehen wir den Einfluss von additivem Störgeräusch in Betracht und stellen ein Framework zur gleichzeitigen Enthallung und Störgeräuschunterdrückung vor, welches auf der Spärlichkeitsannahme basiert und das Teilband Signalmodell verwendet. Genauer schlagen wir eine Methode zur Störgeräuschunterdrückung vor, bei der eine gemischte Norm als Kostenfunktion für die Spärlichkeitseinführung benutzt wird und die Leistung des Störgeräusches begrenzt wird. Außerdem stellen wir eine Methode zur gleichzeitigen Enthallung und Störgeräuschunterdrückung vor, bei der der Term des Störgeräusches in das MCLP-basierende Signalmodell inkludiert wird und bei der ebenfalls die Leistung des Störgeräusches begrenzt wird. Simulationen zeigen, dass die vorgeschlagene, gemeinsame Methode in einer signifikant besseren Performance für niedrige SNRs resultiert, verglichen zur group sparse MCLP-basierenden Methode zur Enthallung. Außerdem zeigen die Simulationen, dass eine gute Performance erreicht werden kann, wenn die zweistufige Prozedur benutzt wird, verglichen mit der MCLP-basierenden Methode zur Enthallung mit auf Spärlichkeitsannahme basierender Störgeräuschunterdrückung.

GLOSSARY

Acronyms and abbreviations

ADA	adaptive
ADMM	alternating direction method of multipliers
ATF	acoustic transfer function
BSI	blind system identification
cADA	constrained adaptive
CAPZ	common acoustical poles and zeros
CD	cepstral distance
CGG	complex generalized Gaussian
DNN	deep neural network
DOA	direction of arrival
DRR	direct-to-reverberant ratio
EDC	energy decay curve
EM	expectation maximization
EVD	eigenvalue decomposition
FIR	finite impulse response
FISTA	fast iterative shrinkage/thresholding algorithm
fwsSNR	frequency-weighted segmental signal-to-noise ratio
GWPE	generalized weighted prediction error
HOS	higher-order statistics
IIR	infinite impulse response
IRL1	iteratively reweighted ℓ_1-norm
IRLS	iteratively reweighted least squares
IRS	inverse repeated sequence
ISTA	iterative shrinkage/thresholding algorithm
ISTFT	inverse short-time Fourier transform
LASSO	least absolute shrinkage and selection operator
LCMV	linearly constrained minimum variance
LLR	log-likelihood ratio

LMS	least mean squares
LP	linear prediction
LS	least-squares
MC	multi-channel
MCLP	multi-channel linear prediction
MDCT	modified discrete cosine transform
MIMO	multiple-input multiple-output
MINT	multiple-input/output inverse theorem
ML	maximum likelihood
MLS	maximum length sequence
MVDR	minimum variance distortionless response
MWF	multi-channel Wiener filter
NLMS	normalized least mean squares
NMF	nonnegative matrix factorization
PD	positive definite
PESQ	perceptual evaluation of speech quality
PSD	power spectral density
RETF	relative early transfer function
RIR	room impulse response
rIRLS	regularized iteratively reweighted least squares
RLS	recursive least squares
RSNR	reverberant signal-to-noise ratio
SMCLP	sparse multi-channel linear prediction
SNR	signal-to-noise ratio
SOS	second-order statistics
SSI	supervised system identification
SRMR	speech-to-reverberation modulation ratio
STFT	short-time Fourier transform
TF	time-frequency
TVG	time-varying Gaussian
WPE	weighted prediction error

Notation

x	scalar x
$\mathbf{x}$	vector $\mathbf{x}$
L_x	length of vector $\mathbf{x}$
$\mathbf{X}$	matrix $\mathbf{X}$
$\mathrm{diag}(\mathbf{x})$	diagonal matrix with $\mathbf{x}$ on the diagonal
$\tilde{\mathbf{X}}$	convolution matrix
x^*	complex conjugate of the scalar x
$\mathbf{x}^\mathsf{T}$	transpose of the vector $\mathbf{x}$
$\mathbf{x}^\mathsf{H}$	conjugate transpose of the vector $\mathbf{x}$
$\mathbf{X}^\mathsf{T}$	transpose of the matrix $\mathbf{X}$
$\mathbf{X}^\mathsf{H}$	conjugate transpose of the matrix $\mathbf{X}$
$\mathbf{X}^{-1}$	inverse of the matrix $\mathbf{X}$
$\mathbf{X}^{-\mathsf{H}}$	conjugate transpose of the inverse of the matrix $\mathbf{X}$
$\hat{x}$	estimated value of the scalar x
$\hat{\mathbf{x}}$	estimated value of the vector $\mathbf{x}$
$\hat{\mathbf{X}}$	estimated value of the matrix $\mathbf{X}$
$\hat{\mathbf{x}}^i$	estimated value of the vector $\mathbf{x}$ at the i-th iteration
$\hat{\mathbf{X}}^i$	estimated value of the matrix $\mathbf{X}$ at the i-th iteration
$\mathbb{R}$	the set of real numbers
$\bar{\mathbb{R}}$	the extended set of real numbers, i.e., $\mathbb{R} \cup \{-\infty, +\infty\}$
$\mathbb{C}$	the set of complex numbers

$*$	convolution operator
$\lceil \cdot \rceil$	ceiling operator
$\delta(.)$	Kronecker delta function
$\mathbf{\Psi}^\mathsf{H}$	time-frequency analysis operator, e.g., short-time Fourier transform
$\mathbf{\Psi}$	time-frequency synthesis operator, e.g., inverse short-time Fourier transform
$\mathrm{p}(.)$	prior distribution
$\mathcal{L}(.)$	likelihood function
$\mathcal{E}\{.\}$	mathematical expectation
$\mathcal{N}_\mathbb{C}(.;\mu_z,\lambda_z)$	complex Gaussian distribution with mean μ_z and variance λ_z
$\psi(.)$	scaling function for variational representation of a sparse prior
$\{\cdot\}'$	derivative

$\Gamma(.)$	gamma function
$\lvert \cdot \rvert$	absolute value
$\lVert \cdot \rVert_p$	l_p-norm
$\lVert \cdot \rVert_{\mathbf{w},1}$	weighted l_1-norm
$\lVert \cdot \rVert_{\mathbf{w},2}$	weighted l_2-norm
$\lVert \cdot \rVert_{p,q}$	$l_{p,q}$-norm
$\lVert \cdot \rVert_{p,q;\boldsymbol{\Phi}}$	$l_{p,q;\boldsymbol{\Phi}}$-norm
$\lVert \cdot \rVert_F$	Frobenius norm of a matrix
$L_\rho(.)$	augmented Lagrangian with parameter ρ
ρ	penalty parameter for the augmented Lagrangian
$\mathrm{prox}_P^\rho(.)$	proximal operator of the function $P(.)$ with parameter ρ

t	discrete-time index
T	number of time-domain samples
n	time frame index
N	number of time frames
k	subband index
K	number of subbands
m	microphone index
M	number of microphones
T_{60}	reverberation time
f_s	sampling frequency
$L_{\underline{h}}$	length of the time-domain RIR $\underline{h}$
$L_{\underline{h}}^{\mathrm{inv}}$	length of the inverse filter $\underline{h}_m^{\mathrm{inv}}$ in the time domain
$L_{\underline{g}}$	length of the prediction filter $\underline{g}$ in the time domain
$\underline{\tau}$	prediction delay in the time domain
$L_{\underline{h}}$	length of the convolutive ATF $\underline{h}$ in the subband domain
L_g	length of the prediction filter g in the subband domain
τ	prediction delay in the subband domain
i	reweighting iteration index
I	number of reweighting iterations
j	ADMM iteration index
J	number of ADMM iterations
$\underline{\boldsymbol{\mu}}$	dual variable vector in the time domain for the ADMM algorithm
$\boldsymbol{\mu}$	dual variable vector in the subband domain for the ADMM algorithm

$\mathbf{M}$	dual variable matrix in the subband domain for the ADMM algorithm

$\underline{s}(t)$	clean speech signal in the time domain
$\underline{y}_m(t)$	m-th microphone signal in the time domain
$\underline{x}_m(t)$	reverberant speech signal at the m-th microphone in the time domain
$\underline{v}_m(t)$	additive noise signal at the m-th microphone in the time domain
$\underline{d}_m(t)$	desired speech signal at the m-th microphone in the time domain
$\underline{h}_m(t)$	room impulse response between the source and the m-th microphone in the time domain
$\underline{h}_m^{\mathrm{inv}}(t)$	inverse filter for the m-th microphone in the time domain
$\underline{g}_{m',m}(t)$	prediction filter in the time domain, relating the m'-th and the m-th microphone

$\underline{\mathbf{y}}_m$	vector of the m-th microphone signal in the time domain
$\underline{\mathbf{x}}_m$	vector of the reverberant speech signal at the m-th microphone in the time domain
$\underline{\mathbf{v}}_m$	vector of the additive noise signal at the m-th microphone in the time domain
$\underline{\mathbf{d}}_m$	vector of the desired speech signal at the m-th microphone in the time domain
$\underline{\mathbf{g}}_{m',m}$	vector of the prediction filter in the time domain, relating the m'-th and the m-th microphone
$\underline{\mathbf{g}}_m$	vector of the multi-channel prediction filter for the m-th channel in the time domain
$\underline{\mathbf{Y}}$	matrix of the M-channel microphone signals in the time domain
$\underline{\mathbf{X}}$	matrix of the M-channel reverberant speech signal in the time domain
$\underline{\mathbf{V}}$	matrix of the M-channel additive noise signal in the time domain
$\underline{\mathbf{D}}$	matrix of the M-channel desired speech signal in the time domain
$\underline{\mathbf{G}}$	matrix of the multiple-input multiple-output prediction filter in the time domain

$\underline{\tilde{\mathbf{X}}}_{m,\underline{\tau}}$	convolution matrix of the reverberant speech signal at the m-th microphone delayed with $\underline{\tau}$ samples in the time domain
$\underline{\tilde{\mathbf{X}}}_{\underline{\tau}}$	multi-channel convolution matrix of the M-channel reverberant speech signal delayed with $\underline{\tau}$ samples in the time domain

$s(k,n)$	clean speech signal in the time-frequency domain
$y_m(k,n)$	m-th microphone signal in the time-frequency domain
$x_m(k,n)$	reverberant speech signal at the m-th microphone in the time-frequency domain
$v_m(k,n)$	additive noise signal at the m-th microphone in the time-frequency domain
$h_m(k,n)$	acoustic transfer function between the source and the m-th microphone in the time-frequency domain
$d_m(k,n)$	desired speech signal at the m-th microphone in the time-frequency domain
$g_{m',m}(k,n)$	prediction filter in the time-frequency domain, relating the m'-th and the m-th microphone

$\mathbf{y}_m(k)$	vector of the m-th microphone signal in the k-th subband
$\mathbf{x}_m(k)$	vector of the reverberant speech signal at the m-th microphone in the k-th subband
$\mathbf{v}_m(k)$	vector of the additive noise signal at the m-th microphone in the k-th subband
$\mathbf{d}_m(k)$	vector of the desired speech signal at the m-th microphone in the k-th subband
$\mathbf{g}_{m',m}(k)$	vector of the prediction filter for the m-th microphone in the k-th subband
$\mathbf{g}_m(k)$	vector of the multi-channel prediction filter for the m-th channel in the k-th subband
$\mathbf{Y}(k)$	matrix of the M-channel microphone signal in the k-th subband
$\mathbf{X}(k)$	matrix of the M-channel reverberant speech signal in the k-th subband
$\mathbf{V}(k)$	matrix of the M-channel additive noise signal in the k-th subband
$\mathbf{D}(k)$	matrix of the M-channel desired speech signal in the k-th subband
$\mathbf{G}(k)$	matrix of the multiple-input multiple-output prediction filter in the k-th subband

$\tilde{\mathbf{X}}_{m,\tau}(k)$	convolution matrix of the reverberant signal at the m-th microphone delayed with τ coefficients in the k-th subband
$\tilde{\mathbf{X}}_{\tau}(k)$	multi-channel convolution matrix of the M-channel reverberant signal delayed with τ coefficients in the k-th subband

$\mathbf{y}(k,n)$	M-channel microphone signal in the time-frequency domain
$\mathbf{x}(k,n)$	M-channel reverberant speech signal in the time-frequency domain
$\mathbf{v}(k,n)$	M-channel additive noise signal in the time-frequency domain
$\mathbf{d}(k,n)$	M-channel desired speech signal at the n-th time frame in the subband domain
$\mathbf{G}(k,n)$	multiple-input multiple-output prediction filter in the time-frequency domain
$\tilde{\mathbf{x}}_{\tau}(k,n)$	buffer of the M-channel signal $\mathbf{x}(k,n)$ in the time-frequency domain
$\boldsymbol{\sigma}_d^2(k,n)$	M-channel power spectral density of the desired speech signal in the time-frequency domain
$\boldsymbol{\sigma}_r^2(k,n)$	M-channel power spectral density of the reverberant speech signal in the time-frequency domain
$\boldsymbol{\sigma}_u^2(k,n)$	M-channel power spectral density of the undesired speech signal in the time-frequency domain
β	noise bound

CONTENTS

1 Introduction 1
1.1 Motivation . . . 1
1.2 Reverberation . . . 2
1.3 Overview of speech dereverberation methods . . . 4
1.4 Outline of the thesis and main contributions . . . 12
2 Signal models and instrumental performance measures 17
2.1 Signal models . . . 17
2.2 Sparsity of speech signals . . . 24
2.3 Instrumental performance measures . . . 25
2.4 Summary . . . 28
3 Sparse multi-channel linear prediction for speech dereverberation 31
3.1 Problem formulation . . . 32
3.2 Conventional MCLP-based dereverberation using TVG model . . . 33
3.3 MCLP-based dereverberation using a general sparse prior . . . 36
3.4 Reformulation as ℓ_p-norm minimization . . . 41
3.5 Simulations . . . 45
3.6 Summary . . . 56
4 Group sparse multi-channel linear prediction for multiple-input multiple-output dereverberation 59
4.1 Problem formulation . . . 60
4.2 Group-sparse modeling . . . 60
4.3 MIMO dereverberation using a group-sparse penalty . . . 62
4.4 Simulations . . . 66
4.5 Summary . . . 69
5 Constrained sparse multi-channel linear prediction for adaptive speech dereverberation 71
5.1 Problem formulation . . . 72
5.2 Adaptive group sparse MCLP . . . 73
5.3 Constrained adaptive group sparse MCLP . . . 75
5.4 Complexity reduction using diagonal approximation . . . 78
5.5 Simulations . . . 78
5.6 Summary . . . 88
6 General framework for sparsity-based speech dereverberation 91
6.1 Problem formulation . . . 92
6.2 Analysis and synthesis sparsity . . . 92
6.3 Wideband model and analysis sparsity . . . 93
6.4 Wideband model and synthesis sparsity . . . 95
6.5 Subband model . . . 96
6.6 Sparsity-promoting cost function . . . 97

6.7 Extension to multiple outputs 99
6.8 Relation to existing methods 101
6.9 Simulations 102
6.10 Summary 112
7 Sparsity-based multi-channel dereverberation and denoising 115
7.1 Problem formulation 116
7.2 Noise model 116
7.3 Sparsity-based denoising 117
7.4 Joint dereverberation and denoising 118
7.5 Sparsity-promoting cost function 121
7.6 Simulations 122
7.7 Summary 129
8 Conclusion and Further Research 131
8.1 Conclusion 131
8.2 Further research directions 134
A Variational representation of a sparse prior 137
A.1 Convex variational representation 137
A.2 Variance estimation 139
B Optimization 141
B.1 Iteratively reweighted methods for non-convex minimization 141
B.2 Proximal operator 144
B.3 Alternating direction method of multipliers 149
B.4 Proximal operator of the composition of an analysis operator and a convex function 150
B.5 Iterative shrinkage/thresholding algorithm 150
C Joint denoising and dereverberation 153

BIBLIOGRAPHY 157

LIST OF FIGURES

Fig. 1.1 An illustration of reverberation in a room with a single speech source and a microphone array. 3

Fig. 1.2 An example of a room impulse response consisting of the direct path, early reflections and late reverberation (reverberation time $T_{60} \approx 700$ ms). 5

Fig. 1.3 A block scheme of a typical spectral enhancement system. $y(t)$ denotes the observed microphone signal and $\hat{d}(t)$ denotes the estimated desired speech signal. 6

Fig. 1.4 A block scheme of a typical system combining spatial and spectral filtering. $y_m(t)$ denotes the m-th observed microphone signal, $m \in \{1, \ldots, M\}$, and $\hat{d}(t)$ denotes the estimated desired speech signal. . 8

Fig. 1.5 A block scheme of a typical indirect inverse filtering system. $y_m(t)$ denotes the m-th observed microphone signal, $m \in \{1, \ldots, M\}$, and $\hat{d}(t)$ denotes the estimated desired speech signal. 9

Fig. 1.6 A block scheme of a typical direct inverse filtering system. $y_m(t)$ denotes the m-th observed microphone signal, $m \in \{1, \ldots, M\}$, and $\hat{d}(t)$ denotes the estimated desired speech signal. 11

Fig. 1.7 Structure of the thesis. 14

Fig. 2.1 The considered multi-channel system with a single static speech source and M microphones in a reverberant and noisy environment. 18

Fig. 2.2 Influence of reverberation on sparsity of a speech signal in the STFT domain illustrated using a spectrogram of (a) anechoic speech signal, and (b) reverberant speech signal ($T_{60} \approx 700$ ms). 26

Fig. 2.3 Influence of reverberation on sparsity of a speech signal in the STFT domain illustrated using a histogram of the magnitude spectrum of anechoic speech and reverberant speech ($T_{60} \approx 700$ ms): (a) full histogram, and (b) detail for small magnitudes. 27

Fig. 3.1 A block scheme of an MCLP-based dereverberation system with the first microphone selected as the reference, i.e., ref = 1. 34

Fig. 3.2 Logarithm of the CGG prior p(.) in (3.18) for different values of the shape parameter p with the scale parameter ζ selected such that the variance is 1. Note that the plot only shows values on the real axis (i.e., the imaginary part of z is 0), and the prior is circular. . . 39

Fig. 3.3 Performance of the IRLS-p method with $L_g = 5$ for different initializations and number of iterations in terms of ΔfwsSNR (left) and ΔPESQ (right). Initialization is denoted with the label in the parentheses. 48

Fig. 3.4 Performance of the IRLS-p method with $L_g = 20$ for different initializations and number of iterations in terms of ΔfwsSNR (left) and ΔPESQ (right). Initialization is denoted with the label in the parentheses. 48
Fig. 3.5 Performance of the IRLS-p method with $L_g = 40$ for different initializations and number of iterations in terms of ΔfwsSNR (left) and ΔPESQ (right). Initialization is denoted with the label in the parentheses. 49
Fig. 3.6 Performance of the rIRLS-p method with $L_g = 5$ for different initializations and number of iterations in terms of ΔfwsSNR (left) and ΔPESQ (right). Initialization is denoted with the label in parentheses. 50
Fig. 3.7 Performance of the rIRLS-p method with $L_g = 20$ for different initializations and number of iterations in terms of ΔfwsSNR (left) and ΔPESQ (right). Initialization is denoted with the label in parentheses. 50
Fig. 3.8 Performance of the rIRLS-p method with $L_g = 40$ for different initializations and number of iterations in terms of ΔfwsSNR (left) and ΔPESQ (right). Initialization is denoted with the label in parentheses. 51
Fig. 3.9 Performance of the IRLS-p method with $M = 1$ for different values of the filter length L_g in terms of ΔfwsSNR (left) and ΔPESQ (right). 52
Fig. 3.10 Performance of the IRLS-p method with $M = 2$ for different values of the filter length L_g in terms of ΔfwsSNR (left) and ΔPESQ (right). 52
Fig. 3.11 Performance of the IRLS-p method with $M = 4$ for different values of the filter length L_g in terms of ΔfwsSNR (left) and ΔPESQ (right). 53
Fig. 3.12 Spectrograms of the microphone signal, direct speech signal and the desired speech signal estimated using IRLS-p with $p = 0.5$ and using $M \in \{1, 2, 4\}$ microphones. 54
Fig. 3.13 Performance of the IRLS-p method with $T_{60} \approx 250$ ms for different values of the filter length L_g in terms of ΔfwsSNR (left) and ΔPESQ (right). 56
Fig. 3.14 Performance of the IRLS-p method with $T_{60} \approx 600$ ms for different values of the filter length L_g in terms of ΔfwsSNR (left) and ΔPESQ (right). 56
Fig. 3.15 Performance of the IRLS-p method with $T_{60} \approx 700$ ms for different values of the filter length L_g in terms of ΔfwsSNR (left) and ΔPESQ (right). 57
Fig. 4.1 A block scheme of an MCLP-based MIMO dereverberation system. 61
Fig. 4.2 Illustration of computation of the mixed norm $\|\mathbf{D}\|_{p,2;\mathbf{\Phi}}$. 62
Fig. 4.3 Performance of the gIRLS-p method with $M = 2$ and $L_g = 25$ in terms of ΔfwsSNR and ΔPESQ. The correlation matrix $\mathbf{\Phi}$ was fixed ($\mathbf{\Phi} = \mathbf{I}$) or estimated using (4.12) ($\mathbf{\Phi} = \text{est}$). 67
Fig. 4.4 Performance of the gIRLS-p method with $M = 4$ and $L_g = 10$ in terms of ΔfwsSNR and ΔPESQ. The correlation matrix $\mathbf{\Phi}$ was fixed ($\mathbf{\Phi} = \mathbf{I}$) or estimated using (4.12) ($\mathbf{\Phi} = \text{est}$). 68

Fig. 4.5 Performance of the gIRLS-p method with $M = 2$ and $L_g = 25$ in terms of ΔfwsSNR (left) and ΔPESQ (right) for different SNRs. The correlation matrix $\mathbf{\Phi}$ was either fixed (top) or estimated using (4.12) (bottom). 69

Fig. 4.6 Performance of the gIRLS-p method with $M = 4$ and $L_g = 10$ in terms of ΔfwsSNR (left) and ΔPESQ (right) for different SNRs. The correlation matrix $\mathbf{\Phi}$ was either fixed (top) or estimated using (4.12) (bottom). 70

Fig. 5.1 Performance of the adaptive methods without diagonal approximation (ADA, cADA-u, cADA-z) with $M = 2$ and $L_g = 25$ in terms of ΔfwsSNR (left) and ΔPESQ (right) for different values of the forgetting factor γ and shape parameter p. 81

Fig. 5.2 Performance of the adaptive methods with diagonal approximation (d-ADA, d-cADA-u, d-cADA-z) with $M = 2$ and $L_g = 25$ in terms of ΔfwsSNR (left) and ΔPESQ (right) for different values of the forgetting factor γ and shape parameter p. 82

Fig. 5.3 Performance of the adaptive methods without the diagonal approximation (ADA, cADA-u, cADA-z) with $M = 4$ and $L_g = 10$ in terms of ΔfwsSNR (left) and ΔPESQ (right) for different values of the forgetting factor γ and shape parameter p. 83

Fig. 5.4 Performance of the adaptive methods with diagonal approximation (d-ADA, d-cADA-u, d-cADA-z) with $M = 4$ and $L_g = 10$ in terms of ΔfwsSNR (left) and ΔPESQ (right) for different values of the forgetting factor γ and shape parameter p. 85

Fig. 5.5 Performance of the adaptive methods without diagonal approximation (ADA, cADA-u, cADA-z) with $p = 0$ for different values of the forgetting factor γ and filter length L_g. 87

Fig. 5.6 Smoothed fwsSNR for a moving speaker vs. time for the microphone signal and the output signal obtained using ADA, cADA-u and cADA-z with $\gamma = 0.99$ (top) and $\gamma = 0.85$ (bottom). The speaker starts walking around 8 s. 88

Fig. 5.7 Spectrograms of the microphone signal (top) and the output signal obtained using ADA, cADA-u and cADA-z with $\gamma = 0.85$. The speaker starts walking around 8 s. 89

Fig. 6.1 Computation of the weight $\hat{w}(k, n)$ for the TF coefficient marked with a black square (■): (a) locally computed weight, (b) weight computed using a neighborhood with dimension 3 across time frames and subbands, and (c) weight computed using an NMF-based low-rank approximation with rank equal to 3. 100

Fig. 6.2 Performance of the ADMM and IRLS-based methods with $M = 2$ and $L_g = 25$ using local weights and $I = 1$ reweighting iteration in terms of ΔfwsSNR (left) and ΔPESQ (right), using the weighted ℓ_1-norm (top) and the weighted ℓ_2-norm (bottom). 105

Fig. 6.3 Performance of the ADMM and IRLS-based methods with $M = 2$ and $L_g = 25$ using local weights and $I = 20$ reweighting iterations in terms of ΔfwsSNR (left) and ΔPESQ (right), using the weighted ℓ_1-norm (top) and the weighted ℓ_2-norm (bottom). 106
Fig. 6.4 Performance of the ADMM-WL1 and IRLS-based methods for $p = 0$ and $p = 0.5$ with $M = 2$ and $L_g = 25$ using neighborhood weights and local weights (loc) with $I = 20$ reweighting iterations in terms of ΔfwsSNR (top) and ΔPESQ (bottom). 107
Fig. 6.5 Performance of the ADMM-WL1 and IRLS-based methods for $p = 0$ and $p = 0.5$ with $M = 2$ and $L_g = 25$ using NMF weights and the local weights (loc) with $I = 20$ reweighting iterations in terms of ΔfwsSNR (top) and ΔPESQ (bottom). 108
Fig. 6.6 Performance of the wideband analysis (WBA), wideband synthesis (WBS) and subband (SB) ADMM-WL1 methods using local weights with $I = 20$ reweighting iterations in terms of ΔfwsSNR (left) and ΔPESQ (right). 109
Fig. 6.7 Performance of the wideband analysis (WBA), wideband synthesis (WBS) and subband (SB) ADMM-WL1 methods using neighborhood weights with $I = 20$ reweighting iterations in terms of ΔfwsSNR (left) and ΔPESQ (right). 110
Fig. 6.8 Performance of the wideband analysis (WBA), wideband synthesis (WBS) and subband (SB) ADMM-WL1 methods using NMF weights with $I = 20$ reweighting iterations in terms of ΔfwsSNR (left) and ΔPESQ (right). 110
Fig. 6.9 Spectrograms of the microphone signal, direct speech signal and the desired speech signal estimated using $\mathcal{M}_1$ and $\mathcal{M}_6$ (showing frequencies up to 4 kHz). 113
Fig. 7.1 The energy of the spatially whitened noise $\|\mathbf{V}\boldsymbol{\Phi}_V^{-\mathsf{T}/2}\|_F^2$ and its expected value in each subband. The lines for all noise types are virtually identical and equal to the expected value. 123
Fig. 7.2 Performance of the proposed ADMM-based denoising method in terms of ΔfwsSNR (top) and ΔPESQ (bottom) for different input SNRs and values of the penalty parameter ρ. 125
Fig. 7.3 Performance of the proposed ADMM-based joint dereverberation and denoising method in terms of ΔfwsSNR (top) and ΔPESQ (bottom) for different input SNRs and values of the penalty parameter ρ. 126
Fig. 7.4 Performance of the dereverberation-only (DER), denoising (DEN), and joint dereverberation and denoising (JNT and $\mathrm{JNT}_{\mathrm{up}}$) methods in terms of ΔfwsSNR (top) and ΔPESQ (bottom) for different input SNRs. 128
Fig. 7.5 Performance of the the dereverberation-only (DER), joint dereverberation and denoising (JNT), two-stage dereverberation and denoising (DER+DEN), and two-stage denoising and dereverberation (DEN+DER) methods in terms of ΔfwsSNR (top) and ΔPESQ (bottom) for different input SNRs. 130

Fig. A.1 An example of a variational approximation of a sparse prior: the true sparse CGG prior with $p = 0.5$ (solid line), and scaled Gaussians at points $t \in \{0.1, 1, 4\}$ (dotted lines). 138
Fig. B.1 Quadratic upper bounds for a non-convex scalar cost function: the true cost function with $p = 0.5$ (solid line), and quadratic approximations at points $t^i \in \{0.1, 1, 4\}$ (dotted lines). 143
Fig. B.2 Upper bounds for a non-convex scalar cost function: the true cost function with $p = 0.5$ (solid line), and ℓ_1-norm approximations at points $t^i \in \{0.1, 1, 4\}$ (dotted lines). 144
Fig. B.3 Proximal operators (left) and the corresponding cost functions (right) on the real axis with $G_{\min} = \{0, 0.2\}$ and $\rho = w = 10$. For complex-valued inputs the proximal operator and the penalty function are extended using circular symmetry. 147

LIST OF TABLES

Table 5.1 Real-time factors of the adaptive algorithms. 84
Table 5.2 Overall PESQ values for the microphone signal and the output signal obtained using ADA, cADA-u and cADA-z for a moving speaker with different values of the forgetting factor. 87
Table 6.1 Summary of the results obtained with the selected variants of sparse MCLP with different signal models, cost functions and iterative optimization algorithms in terms of ΔfwsSNR and ΔPESQ and real-time factors. 112

1

INTRODUCTION

1.1 Motivation

Speech is an effective means of communicating information and emotions between humans and can also be conveniently used in human-computer interfaces to provide a natural communicational channel [1–3]. Recent advances in computer hardware have resulted in computationally powerful mobile and portable electronic devices, which are nowadays commonly used for communication and as personal assistants. Due to increasingly complex computer interfaces there has been a resurged interest in speech-based human-computer interfaces, which can serve as a natural and flexible means of interaction with various devices. Currently, speech communication is highly important in various applications, such as hands-free telephony, assistive listening devices, speech-based interfaces for computers and entertainment systems in homes, workplaces and public venues. These applications require high-quality speech communication which needs to take into account the specific requirements imposed by the user, the environment and the application itself. The increased use of speech communication in complex and diverse acoustic environments has resulted in an increasing interest in the topic of speech signal processing [4–6]. In particular, this thesis is motivated by the increased number of applications employing far-field hands-free speech communication in reverberant environments.

Microphones placed at a distance from the speaker are commonly used, e.g., in hands-free telephony, speech-controlled devices or hearing aids. While this gives the user a relatively high degree of flexibility, microphones placed at a distance in, e.g., an office or a living room, are very likely to capture a speech signal corrupted with various undesired disturbances in addition to the desired speech signal, such as reverberation and noise [4, 7]. Reverberation is naturally caused by reflections of the sound waves against surfaces and objects within the room. Although a moderate amount of reverberation can be beneficial, strong reverberation is typically detrimental to speech communication, resulting in a decreased speech quality and automatic speech recognition performance [8–11]. Additionally, the additive noise caused by, e.g., other speakers or sound sources, can further reduce the effectiveness of speech communication [8–10].

In order to reduce the detrimental influence of reverberation and noise on speech communication, effective dereverberation and denoising methods need to be employed [12].

Speech denoising has been addressed in many contributions over the last decades, and a number of single- and multi-channel denoising methods have been proposed [4,6,13–21]. Typically, it is assumed that the additive noise is uncorrelated or independent from the desired speech signal, which can be used to design the corresponding estimators of the desired speech signal. Many proposed denoising methods can provide a significant benefit when reverberation is relatively low and the noise is dominant, e.g., when the speaker-microphone distance is small or the environment is not highly reverberant. However, reverberation is typically highly correlated with the desired speech signal, since it is a filtered version of the clean speech signal, and it is widely recognized that dereverberation is still a greater challenge than denoising [5, 22]. Hence, speech dereverberation has become a very active research topic more recently [5, 22, 23], which could be attributed to the emergence of novel applications employing distant microphones, but also to the increased computational capabilities of the deployed devices enabling to tackle the temporal dependencies induced by reverberation.

1.2 Reverberation

In this section, we consider a scenario with a single speech source and multiple microphones in a reverberant room, as illustrated in Fig. 1.1. The signal captured by the microphones consists of a superposition of the direct speech signal and a number of reflections [5]. The direct speech signal is equal to the clean speech source signal, up to an attenuation and a propagation delay. The reflections of the speech signal against the boundaries of the environment and possible objects constitute reverberation. These reflections are essentially delayed and attenuated images of the clean speech source signal, where the attenuation and the delay depend on the acoustic properties of the reflecting surfaces in the environment and the equivalent path between the source and the microphone. Reflections arriving from different directions create a perceptual impression of spaciousness and lead to an impression of the speaker positioned at a large distance from the microphone [5]. In some applications, reverberation can be used to shape the perception of space and enhance the listening experience, e.g., of music. However, this multi-path propagation of sound, if not controlled, can often have detrimental effects on the efficiency of speech communication [5].

When considering the influence of reverberation on the speech signal captured in a room, reverberation is typically decomposed into early reflections and late reverberation:

- Early reflections arrive at the microphone immediately after the direct speech signal. The considered time window corresponding to the early reflections is typically in the order of tens of milliseconds, with a typical value of 50 ms [5, 24]. These reflections commonly cause spectral coloration of the speech signal [5, 24]. However, they can also be beneficial for speech intelli-

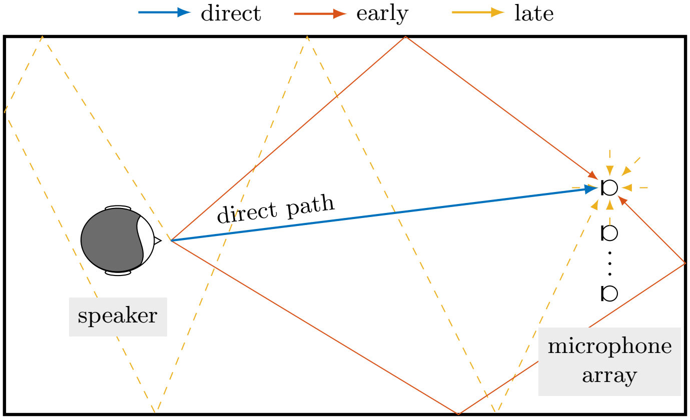


Fig. 1.1: An illustration of reverberation in a room with a single speech source and a microphone array.

gibility, since the early reflections can be perceptually integrated with the direct speech signal, hence increasing the effective signal-to-noise ratio of the desired speech signal [11,25–28]. Furthermore, the spatial information present in the individual early reflections can be used to infer the geometry of the room and perform source localization [29–32], or exploited to improve spatial filtering [33–35].

- Late reverberation arrives at the microphone after the early reflections. As opposed to the early reflections, which consists of spatially distinct reflections, late reverberation typically arrives at the microphone approximately uniformly from all directions [5]. It is widely accepted that the late reverberation in general has a detrimental effect both on the performance of automatic speech recognition systems [10,22] and speech quality and intelligibility [8,9,11,36,37], with the effect augmented by age, hearing impairment, and for non-native speakers [9,38–40]. This is especially noticeable in scenarios with microphones placed at a relatively large distance from the speaker, when the energy of late reverberation becomes comparable to or larger than the energy of the direct speech signal and early reflections.

In the case of a static acoustic scenario, the multi-path sound propagation between the source and the microphone can be characterized using the corresponding impulse response. The acoustic impulse response between the source and the microphone in a room is commonly referred to as the room impulse response (RIR) [5]. The structure of an RIR is directly related to the previously discussed structure of reverberation, i.e., an RIR can be decomposed into the direct path, early reflections and late reverberation, as indicated in Fig. 1.2. The direct path of the RIR corresponds to the attenuation of the direct speech signal and the propagation delay between the speech source and the microphone. The early reflections of the RIR typically consist of well-defined impulses with a relatively large amplitude [5], while the late reverberation consists of many decaying and densely-spaced impulses, which are often modeled as being randomly distributed [5]. As the RIR depends both on the room properties and the positions of the source and the microphone, it can change significantly even for small perturbations of the source or the microphone positions or due to thermal fluctuation [5,24,41]. Different RIR models have been considered in the literature, depending on the application and the imposed constraints. A commonly used representation of an RIR, also used in this thesis, is the finite impulse response (FIR) model. However, alternative models, such as the infinite impulse response (IIR) model, the common acoustical poles and zeros (CAPZ) model as well as orthonormal basis functions have been considered in the literature [42–44].

Methods for measuring or estimating RIRs can be classified into two main categories: supervised system identification (SSI) and blind system identification (BSI). SSI methods estimate the RIR using a known test source signal and the captured microphone signal, e.g., using maximum length sequences (MLS) [45,46], inverse repeated sequences (IRS) [47], time-stretched pulses [48] or sine-sweeps [49,50]. Alternatively, if the source signal and the captured microphone signal are available, the RIR can be estimated, e.g., using least-squares (LS) regression [51–53]. On the other hand, BSI methods aim to estimate the RIR using only the captured microphone signals, and are typically based on second-order statistics (SOS) [54,55] or higher-order statistics (HOS) [56]. However, in many cases the estimated RIRs differ significantly from the actual RIRs, due to possible spatial mismatch, thermal fluctuations, and sensitivity of currently available system identification methods to additive noise.

The amount of reverberation in a room can be specified using the reverberation time [5,57]. The reverberation time T_{60} is the time required for the reverberant energy to decay 60 dB after the sound source has been deactivated. Unlike the RIR, the reverberation time depends only on the properties of the room, i.e., the room geometry and the surface reflectivity [5]. In addition to the reverberation time T_{60}, the amount of reverberation for a certain source and microphone position can be specified using the direct-to-reverberant ratio (DRR). The DRR is defined as the ratio of the energy of the direct path component of the RIR and the energy of the reverberant component of the RIR. Another relevant measure of reverberation is the clarity index, defined as the ratio of the energy of the direct and early component of the RIR and the energy of the late component of the RIR [5].

1.3 Overview of speech dereverberation methods

The aim of speech dereverberation is to obtain an estimate of the desired speech signal, i.e., the direct speech signal with or without early reflections, by processing the reverberant and noisy microphone signals. Depending on the application, a suitable speech dereverberation method can be designed aiming to improve speech quality, speech intelligibility or the performance of an automatic speech recognition system. In the following section, we present an overview of the existing methods proposed in the literature.

Given the variety of speech dereverberation methods that have been proposed in the literature, different classifications of the proposed methods are possible. For example,

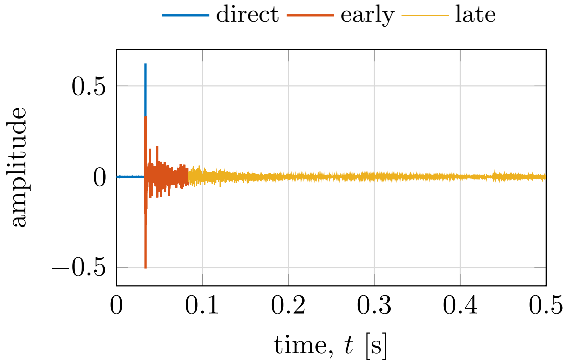


Fig. 1.2: An example of a room impulse response consisting of the direct path, early reflections and late reverberation (reverberation time $T_{60} \approx 700$ ms).

dereverberation methods could be classified into single-channel and multi-channel methods, based on the number of microphone signals which are used. Alternatively, dereverberation methods could be classified into blind and non-blind methods, depending on whether some knowledge about the RIRs is assumed to be available or not. Similarly, dereverberation methods could be classified based on the employed signal model into reverberation cancellation, reverberation suppression, and direct methods [58].

According to the type of processing, speech dereverberation methods can be classified into (*i*) spectral enhancement, (*ii*) combined spatial and spectral filtering, (*iii*) indirect inverse filtering, and (*iv*) direct inverse filtering. Spectral enhancement methods typically perform dereverberation by applying a real-valued gain to the spectral coefficients of the microphone signal. Spatial filtering methods typically perform dereverberation by combining signals from multiple microphones to exploit spatial information. Indirect inverse filtering methods typically perform dereverberation by using estimated or measured transfer functions between the speech source and the microphones to design the inverse filters. Direct inverse filtering methods typically perform dereverberation by exploiting properties of the speech signal to design the inverse filters. In the following, we give an overview of different speech dereverberation methods in these categories. A more detailed analysis of many dereverberation methods can be found in, e.g., in [5,6,52,53,58,59].

1.3.1 *Spectral enhancement*

Spectral enhancement methods typically exploit spectro-temporal information to enhance a single-channel speech signal. The spectral coefficients of the desired speech signal are commonly estimated by applying a (real-valued) gain to the spectral coefficients of the microphone signal, aiming to suppress the undesired disturbance in the captured signal. A block scheme of a typical spectral enhancement system is depicted in Fig. 1.3.

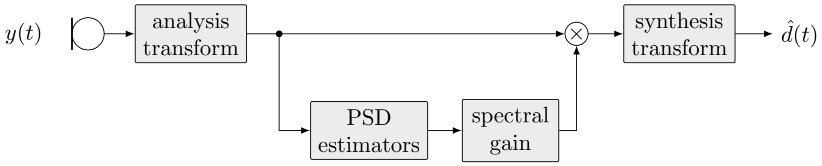


Fig. 1.3: A block scheme of a typical spectral enhancement system. $y(t)$ denotes the observed microphone signal and $\hat{d}(t)$ denotes the estimated desired speech signal.

Spectral enhancement has been traditionally employed for denoising, and many denoising methods have been proposed in the literature [16,19,60,61], typically requiring an estimate of the noise power spectral density (PSD) [62–65] . Initially, spectral enhancement was based on spectral subtraction [66], but several modifications of the basic spectral subtraction have been proposed to reduce speech distortions and musical noise [67]. More recently, optimal gain functions have been derived based on statistical models for the speech and the noise coefficients [19,68–73].

Speech dereverberation using spectral enhancement typically uses the same gain function as for denoising, with the late reverberant PSD used instead of the noise PSD. Spectral enhancement-based speech dereverberation has been originally proposed in [74], where an estimate of the late reverberant PSD has been obtained using a temporal exponential decay model for the RIR [75] and consequently used to compute the gain function. In [76], multiple microphone signals have been used to obtain a spatially averaged amplitude spectrum for estimating the late reverberant PSD. Statistically optimal gain functions and late reverberant PSD estimators based on a statistical model of the RIR have been presented in [59,77,78]. Alternatively, Bayesian dereverberation of power spectrograms based on an autoregressive reverberation model has been considered in [79], while joint temporal and spectral modeling based on non-negative models has been considered in [80]. Spectral enhancement for joint denoising and dereverberation based on a statistical model for the late reverberant PSD has been considered in [59,81,82]. Similarly, probabilistic models for speech and noise have been used to derive joint estimators for denoising and dereverberation in [83–85]. In [86], the late reverberant PSD has been estimated by linear prediction, by assuming a sparse predictor, and a sparse prior for the speech PSD has been used in [87].

More recently, several spectral enhancement methods based on neural network-based models have been proposed. While early contributions used shallow networks [88], recent contributions have been based on deep neural networks (DNNs) [89]. Although a large amount of data and computational resources are usually required for training the neural networks, processing using a trained model requires much less resources and can be used in real-time applications [58, 90]. DNNs are typically trained to either directly estimate the desired speech coefficients or a gain function [91]. Different DNN structures have been employed for spectral enhancement, such as the autoencoder and the recurrent neural networks [92–95]. DNNs for joint dereverberation and denoising have also been considered in [90, 96–98].

Due to its simplicity, spectral enhancement is usually computationally inexpensive and suitable for many applications. However, there is typically a tradeoff between speech distortion and undesired signal suppression, and perfect dereverberation can in general not be achieved using spectral enhancement.

1.3.2 *Combined spatial and spectral filtering*

In general, a combination of spatial and spectral filtering can be used to exploit both spatial and spectro-temporal information for speech enhancement.

Spatial filtering, also referred to as beamforming, has been commonly employed for denoising when multiple microphones are available [20, 99–103]. In beamforming, the multi-channel input signals are linearly filtered and summed in such a way that the desired speech signal is preserved in the output signal, while the background noise is reduced. Beamformers can be signal independent, i.e., exploiting only the geometry of the microphone array, or signal dependent, i.e., exploiting the statistics of the microphone signals. A classical signal-independent beamformer is the delay-and-sum beamformer, which applies delays to the microphone signals in order to align the desired signal from a certain direction in all microphones before summing them, thereby suppressing the incoherent disturbance [99]. Classical signal-dependent beamformers are the minimum variance distortionless response (MVDR) beamformer [100, 104] and the more general linearly constrained minimum variance (LCMV) beamformer [105]. While beamforming can be very effective for denoising in the presence of directional interferences, it is typically less effective for diffuse disturbances such as reverberation, especially when employing a small number of microphones [5, 106].

Different combinations of spatial and spectral filtering have been proposed to reduce reverberation. In [107], the late reverberant PSD for each channel has been estimated using long-term linear prediction (LP) and spectral enhancement has been used to reduce reverberation in each channel, followed by a delay-and-sum beamformer. In [108], blind source separation has been used to obtain an estimate of the noise and late reverberant PSD, which is used for dereverberation by performing spectral enhancement of the averaged microphone signals. A two-stage beamformer has been proposed in [109], where a signal-independent superdirective beamformer has been used to reduce reverberation, followed by a signal-dependent beamformer to reduce the residual noise.

Several methods employing a beamformer followed by spectral enhancement have been used to improve the dereverberation performance of the beamformer, with a block scheme of a typical system depicted in Fig. 1.4. In [82, 110–112], a fixed beamformer has been combined with a stand-alone single-channel spectral enhancement for suppressing residual noise and reverberation. An estimator of the late reverberant PSD based on a temporal exponential decay RIR model has been used in [82, 110], while an estimator based on acoustic equalization has been used in [112]. Assuming that late reverberation is isotropic, maximum likelihood (ML) estimators for the late reverberant PSD have been proposed in [113–115], which is subsequently used to reduce reverberation using spectral enhancement at the output

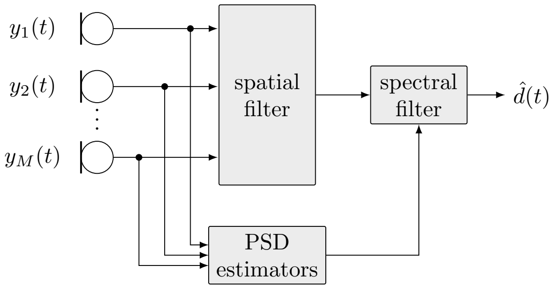


Fig. 1.4: A block scheme of a typical system combining spatial and spectral filtering. $y_m(t)$ denotes the m-th observed microphone signal, $m \in \{1, \ldots, M\}$, and $\hat{d}(t)$ denotes the estimated desired speech signal.

of an MVDR beamformer. Similarly, the late reverberant PSD has been estimated using an eigenvalue decomposition (EVD) in [116], assuming that the late reverberation is isotropic. Note that the considered combinations of an MVDR beamformer and a single-channel spectral enhancement based on a Wiener filter can be seen as variants of the multi-channel Wiener filter (MWF) [103, 117, 118]. Since the design of the MVDR beamformer is based on an estimate of the direction of arrival (DOA) or the relative early transfer functions (RETF) of the desired speech source, the performance may suffer in the presence of estimation errors [116, 119].

In addition to the DOA of the speech source, i.e., of the direct path signal, the DOAs of the early reflections have been exploited in the LCMV beamformer in [33]. Similarly, different acoustic rake receivers, also exploiting early reflections, have been proposed in [34]. A joint dereverberation and denoising method has been proposed in [120], where estimation of both the acoustic transfer functions (ATFs) and the source signal has been formulated in a probabilistic framework. Recently, several beamforming methods based on neural networks have been proposed for multi-channel speech enhancement for automatic speech recognition [121, 122], with the spatial and spectral filtering typically integrated and trained together with the speech recognizer.

Methods combining spatial and spectral filtering are typically robust and computationally efficient. However, dereverberation performance of beamforming is typically limited and its combination with spectral filtering results in an inherent tradeoff between speech distortion and undesired signal suppression in the output signal. Therefore, these methods are typically unable to perfectly recover the desired speech signal.

1.3.3 *Indirect inverse filtering*

Indirect inverse filtering typically consists of two steps, as depicted in Fig 1.5. Firstly, the RIRs or the ATFs between the speech source and the microphones need to be identified, i.e., the RIRs or the ATFs between the source and the microphones are estimated using supervised or blind system identification. Secondly, the estimates

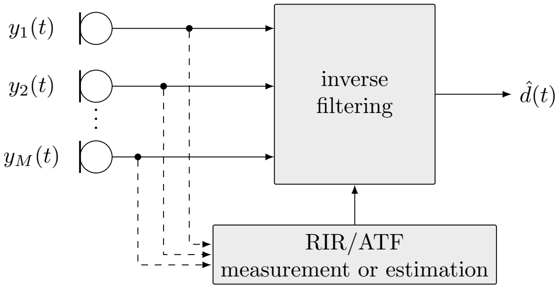


Fig. 1.5: A block scheme of a typical indirect inverse filtering system. $y_m(t)$ denotes the m-th observed microphone signal, $m \in \{1, \dots, M\}$, and $\hat{d}(t)$ denotes the estimated desired speech signal.

are used to perform inverse filtering, i.e., acoustic channel equalization, to reshape the transfer function in such a way that the influence of reverberation is either reduced or completely removed.

Assuming that multiple microphones are available and the RIRs between the source and the microphones do not share common zeros, perfect dereverberation can be achieved using inverse filtering based on the multiple-input/output inverse theorem (MINT) [123]. While this is theoretically appealing, numerous studies have demonstrated the sensitivity of the inverse filters to RIR mismatch [5, 52, 53, 124, 125]. In practice, there is typically a significant difference between the estimated or measured RIRs and the actual RIRs, due to either spatial mismatch or inaccuracy of system identification. Therefore, a large body of research has been focused on increasing the robustness of acoustic channel equalization, assuming that estimated RIRs are available.

Inversion in subbands has been considered in [126–128], resulting in a reduced computational complexity and increased robustness. In [129, 130], robustness has been increased by using channel shortening for partial equalization, i.e., by not aiming at perfect dereverberation but preserving the early reflections in the output signal. In [130, 131], a robust design of inverse filters for partial equalization based on weighted least squares has been proposed. In [132], a robust partial equalization based on MINT has been proposed, improving the perceptual quality of partial equalization. Robustness of several partial equalization methods with respect to the mismatch of the RIRs has been increased by using shorter filter lengths [133], by signal-independent regularization [125, 134], and by signal-dependent regularization using sparsity of the output speech signal [135, 136]. In [137], joint dereverberation and denoising based on multi-channel equalization has been proposed, by taking into account the second-order statistics of the noise and the speech in the filter design.

As an alternative to the two-step indirect inverse filtering, estimation of the transfer functions and the desired speech signal has been formulated jointly. A joint estimation of the subband convolutive ATFs, i.e., the subband analogues of the RIRs, and the desired speech coefficients has been proposed in [138], which iterates between the ATF identification and subband inverse filtering-based desired speech estimation. The likelihood function for the parameters is obtained using Gaussian modeling, and the parameters are estimated using expectation-maximization (EM), with a Kalman smoother used to solve a structured least-squares problem. An online extension of [138] has been proposed in [139], using a recursive EM scheme and the Kalman filter. A similar signal model has been used in [140], where we proposed an iterative algorithm for estimating the subband convolutive ATFs and the speech coefficients by using a sparse model for the speech coefficients and a Bayesian-based cost function. However, since these methods perform indirect inverse filtering using the current estimate of the convolutive ATFs, they can also be sensitive when the estimation errors are large.

Inverse filtering can also be used for single-channel dereverberation, although perfect dereverberation is in general not possible since the RIRs are generally mixed-phased, and a causal stable inverse filter does not exist [123]. In [141], a least-squares regression between the obtained and the desired response has been used to design an approximate inverse filter. However, a very long inverse filter is required for satisfactory performance, resulting in a high complexity, and sensitivity to RIR mismatch remains a problem [142]. In [143], computational complexity has been reduced by inverse filtering in the frequency domain and using spectral enhancement to reduce the artifacts.

In general, methods based on inverse filtering typically exploit the complete information about the transfer function between the speech source and the microphones as captured by the RIRs. If multiple microphones are available, perfect dereverberation can be achieved using indirect inverse filtering if the RIRs are perfectly known (assuming the MINT conditions hold). However, when the measured or estimated RIRs are perturbed with respect to the actual RIRs, the performance suffers, even when using robust inverse filtering methods. Furthermore, time-domain inverse filtering methods are typically not suitable for real-time applications, and simultaneous estimation of RIRs and robust acoustic equalization remains a difficult problem, both due to the RIR fluctuations and the computational complexity of designing the time-domain inverse filters.

1.3.4 *Direct inverse filtering*

Direct inverse filtering methods aim to achieve speech dereverberation without requiring information about the transfer functions, such as RIRs or ATFs, between the speech source and the microphones. The inverse filters are applied directly on the microphone signals which are combined in such a way that the reverberation is removed or suppressed, with a typical system depicted in Fig. 1.6. As opposed to the indirect inverse filtering methods, the main advantage of the direct methods is that the inverse filters are typically designed by exploiting the properties of the speech signal and not using the transfer functions between the speech source and the microphones.

Most methods from this class are based on multi-channel linear prediction (MCLP) [144, 145]. More specifically, the multi-channel microphone signals are filtered and summed to obtain an estimate of the undesired reverberant speech com-

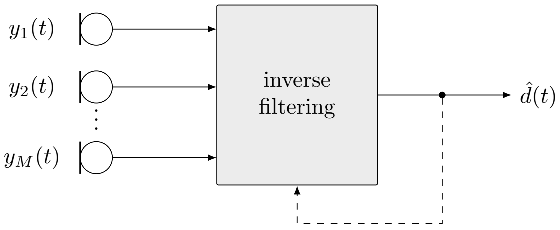


Fig. 1.6: A block scheme of a typical direct inverse filtering system. $y_m(t)$ denotes the m-th observed microphone signal, $m \in \{1, \dots, M\}$, and $\hat{d}(t)$ denotes the estimated desired speech signal.

ponent at the reference microphone, which can then be subtracted to remove reverberation at the reference microphone. Based on MINT, it can be shown that prediction filters that achieve perfect dereverberation do exist, and that they can be indirectly computed from the RIRs and the MINT-based inverse filters. Moreover, perfect dereverberation is possible for multiple sources, as long as the number of microphones is larger than the number of sources. While this ensures that such prediction filters in theory exist, estimating them blindly, i.e., without using the RIRs, can be a difficult task [146]. Since virtually no information about the transfer functions is used, effective solutions typically exploit some information about the desired speech signal [146, 147].

Initially, the prediction filters were estimated by minimizing the energy of the output signal, i.e., the prediction residual [107, 148–152]. However, since the energy minimization criterion has been used for temporally white signals, this typically leads to prediction filters that result in excessive equalization (whitening) of the speech signal [144, 145]. Several strategies for reducing this effect have been proposed in the literature [146]. In [148, 149], the microphone signals are first whitened using an estimated whitening filter, and the prediction filters are then estimated from the preprocessed microphone signals. In this way, the estimated prediction filters would predict mainly the undesired reverberant signal. In [150], the average speech characteristics are estimated from all microphones, and used to compensate for the excessive whitening at the output of the prediction filters. In [107], in addition to pre-whitening, a prediction delay has been introduced to preserve the short-term speech correlation and to estimate only the late reverberation using MCLP. In [152], time-varying speech characteristics and time-invariant prediction filters have been jointly estimated, thereby obtaining the prediction filters which do not perform excessive whitening, since they capture only the properties of the acoustic channel.

A somewhat different approach has been used in [153, 154]. More specifically, ML estimation of the prediction filters has been formulated using a time-varying Gaussian model for the desired speech signal. A pre-trained dictionary-based speech model has been used to prevent excessive whitening in [153], while pre-whitening and a prediction delay have been used with a simplified speech model in [154]. An efficient subband variant of the latter approach, referred to as the weighted prediction error (WPE) method, has been proposed in [154, 155]. More specifically, the WPE method has been formulated using a delayed MCLP-based signal model and a locally Gaussian model for the coefficients of the desired speech signal in each subband. An iterative optimization procedure has been derived for ML estimation of the prediction filters.

Several methods extending WPE have been proposed in the literature [156–172], e.g., by combining WPE with spatial and spectral filtering. An extension to multiple-output speech dereverberation has been proposed in [161]. The proposed method has been derived using a cost function that minimizes the inter-frame dependence of the desired speech signal coefficients in each subband, resulting in an iterative optimization algorithm for estimating the prediction filters. A single-output adaptive variant of WPE based on recursive least squares has been proposed in [156], suitable for dynamic acoustic scenarios. Similarly, a multiple-output adaptive variant has been proposed in [163]. Similarly, adaptive variants using Kalman filter-based subband processing have been proposed in [169, 170]. In [166], a probabilistically formulated combination of inverse filtering, beamforming and spectral enhancement has been proposed, using a locally Gaussian model for the desired speech coefficients and a probabilistic model for time-varying ATFs. The unknown parameters have been iteratively estimated using an expectation-maximization algorithm in each subband. Furthermore, a combination of dereverberation and source separation has been proposed in [157, 160, 173], where the prediction filters and demixing matrices are estimated jointly. Combined dereverberation and denoising has been considered by joint estimation of the prediction filters and the denoised signal in a probabilistic framework in [158, 159, 165, 167, 168], while combinations of WPE with independent spatial and spectral filtering have been considered in [163, 164].

Another relevant method for dereverberation and source separation using direct inverse filtering has been proposed in [174] by exploiting non-gaussianity, non-whiteness and non-stationarity of the speech signals. While allowing general source models and different applications [175], this approach is typically computationally complex.

In general, direct inverse filtering methods typically exploit properties of the speech signal to design the dereverberation filters. Theoretically, using inverse filtering ensures that perfect dereverberation is possible when multiple microphones are available. Furthermore, by exploiting the speech signal properties, these methods alleviate the need for information about the transfer function between the speech source and the microphones, as opposed to the indirect inverse filtering methods. This is a large advantage over the indirect filtering methods, e.g., in scenarios when the transfer functions are unknown, varying over time, or cannot be measured or estimated.

1.4 Outline of the thesis and main contributions

This thesis deals with the problem of blind speech dereverberation by developing a class of direct inverse filtering methods, i.e., inverse filtering methods not requiring estimated or measured RIRs or ATFs. More specifically, we propose a general framework for blind speech dereverberation using the MCLP-based signal model for the reverberant speech and exploiting sparsity of the speech signal in the time-frequency (TF) domain.

The main contributions of this thesis can be summarized as follows. Firstly, we propose a probabilistic and deterministic formulation of sparsity-promoting MCLP in the subband domain, generalizing existing single- and multi-output MCLP-based dereverberation methods. Secondly, we propose a constrained sparse MCLP formulation for adaptive speech dereverberation, increasing the robustness of the existing adaptive methods. Thirdly, we propose a general framework for speech dereverberation based on sparse MCLP by using either a wideband or a subband signal model, exploiting sparsity of the speech signal in the TF domain and incorporating additional structure of the speech signal. Finally, we propose a sparsity-based dereverberation and denoising method in a joint framework. A structured overview of the thesis is given in Fig. 1.7.

In Chapter 2, we present wideband and subband signal models used for the multichannel reverberant signal. We introduce the notion of sparsity of a signal and discuss sparsity of the speech signals in the TF domain, demonstrating the influence of reverberation on TF sparsity. Furthermore, we define the instrumental measures used to evaluate the performance of speech dereverberation methods.

In Chapter 3, we consider the noiseless case and propose a single-output batch method for blind speech dereverberation based on sparse multi-channel linear prediction using the subband signal model. We formulate the estimation of the prediction filter using a sparse prior for the TF coefficients of the speech signal, and present a general algorithm based on a variational representation of the sparse prior. We show that the conventional MCLP-based dereverberation method is included as a special case of the proposed method. The content of this chapter is related to the work published in [176–178].

In Chapter 4, we extend the single-output method from Chapter 3 to a multiple-output blind speech dereverberation method based on group sparse multi-channel linear prediction. We formulate the optimization problem using a cost function which promotes sparsity across time and takes into account grouping of the coefficients across the microphones, generalizing the conventional single- and multiple-output dereverberation methods. The content of this chapter is related to the work published in [179].

In Chapter 5, we extend the batch methods from Chapters 3 and 4 to an adaptive blind speech dereverberation method based on constrained sparse multi-channel linear prediction. The proposed adaptive method may in some cases lead to distortions due to overestimation of the undesired speech signal. In order to prevent excessive cancellation of the desired speech signal, we use an estimate of the late reverberant PSD to constrain the estimated undesired speech signal, and thereby increase the robustness of the adaptive dereverberation method. Furthermore, we propose a diagonal approximation for reducing the computational complexity. The content of this chapter is related to the work published in [180, 181].

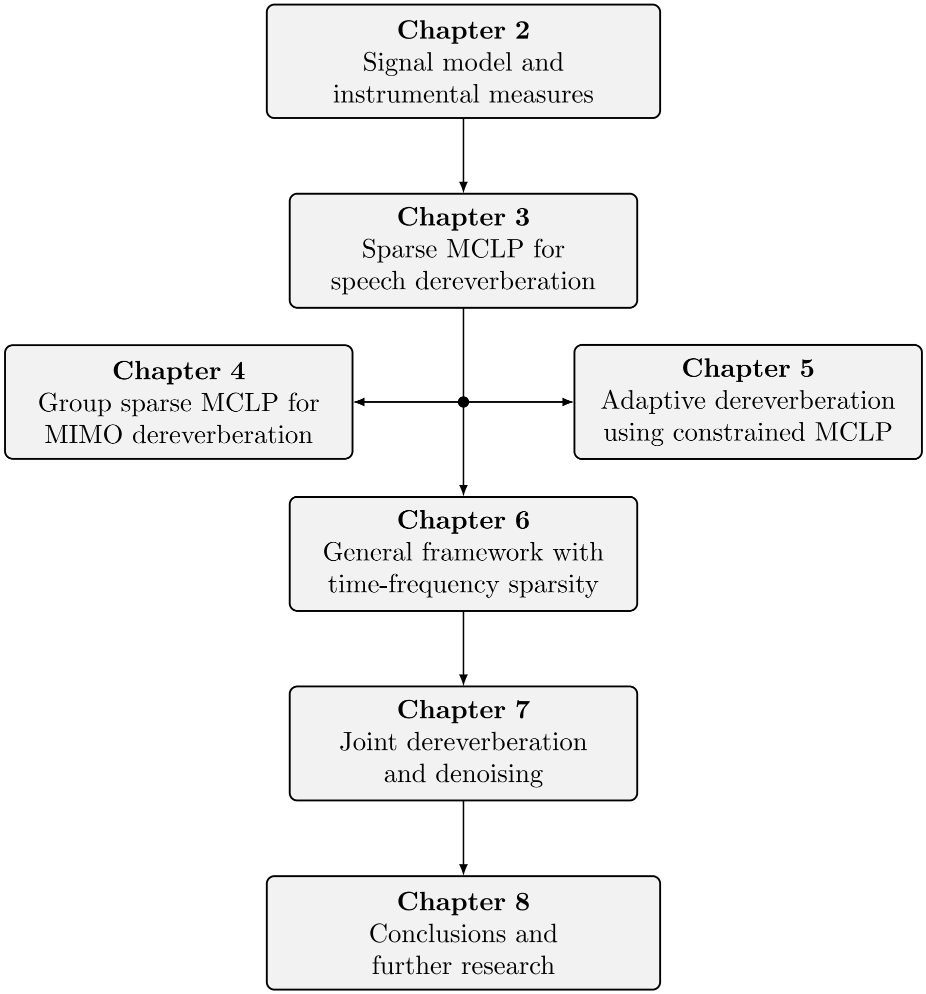

Chapter 2
Signal model and instrumental measures
Chapter 3
Sparse MCLP for speech dereverberation
Chapter 4
Group sparse MCLP for MIMO dereverberation
Chapter 5
Adaptive dereverberation using constrained MCLP
Chapter 6
General framework with time-frequency sparsity
Chapter 7
Joint dereverberation and denoising
Chapter 8
Conclusions and further research

Fig. 1.7: Structure of the thesis.

Whereas in the previous chapters only the subband signal model was considered, in Chapter 6 we propose a general framework for speech dereverberation using MCLP-based signal models and exploiting sparsity in the TF domain. We formulate optimization problems using either the wideband or the subband signal model, with a sparsity-promoting cost function and a general TF analysis operator. We investigate different cost functions with and without exploiting the TF structure of the speech signal. The content of this chapter is related to the work published in [182, 183].

Whereas in the previous chapters only the noiseless case was considered, in Chapter 7 we propose a method for joint dereverberation and denoising in the subband domain. We formulate optimization problems for denoising and joint dereverberation and denoising by exploiting sparsity of the speech signal and imposing a bound for the energy of the noise component in the signal model, assuming that the noise correlation matrix is known.

In Chapter 8, we summarize the main contributions of the thesis and discuss possible topics for further research, i.e., extensions of the presented methods and possible applications.

2

SIGNAL MODELS AND INSTRUMENTAL PERFORMANCE MEASURES

In this chapter, we present wideband and subband signal models, discuss sparsity of speech signals and define the instrumental measures used to evaluate the performance of speech dereverberation methods.

In Section 2.1 we present the wideband signal model for a speech signal captured in a reverberant enclosure, based on multi-channel linear prediction in the time domain. Furthermore, we present a subband signal model, which is an approximation of the wideband signal model with independent modeling applied in each subband. In Section 2.2 we introduce the notion of sparsity of a signal and briefly discuss the influence of reverberation on TF sparsity of speech signals. In Section 2.3 we present the instrumental performance measures used to evaluate the performance of the speech enhancement methods in the remainder of the thesis.

2.1 Signal models

2.1.1 *Wideband signal model*

We consider an acoustic scenario where a single static speech source in a reverberant and noisy environment is captured by M microphones as given in Figure 2.1. Let $\underline{s}(t)$ denote the anechoic speech signal in the time domain, with t denoting the discrete-time index. The time-domain signal $\underline{y}_m(t)$ captured at the m-th microphone, $m \in \{1, \ldots, M\}$, can be modeled in the time domain as

$$\underline{y}_m(t) = \underline{x}_m(t) + \underline{v}_m(t), \tag{2.1}$$

where $\underline{x}_m(t)$ is the reverberant speech signal observed at the m-th microphone and $\underline{v}_m(t)$ is the additive noise signal observed at the m-th microphone. The reverberant speech signal $\underline{x}_m(t)$, without the additive noise signal $\underline{v}_m(t)$, can be modeled in the time domain as

$$\underline{x}_m(t) = \sum_{l=0}^{L_{\underline{h}}-1} \underline{h}_m(l)\underline{s}(t - t_m - l) = \underline{h}_m(t) * \underline{s}(t - t_m), \tag{2.2}$$

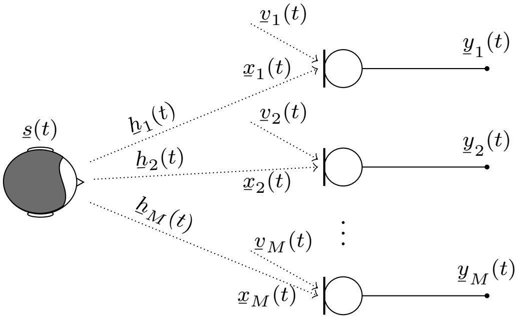


Fig. 2.1: The considered multi-channel system with a single static speech source and M microphones in a reverberant and noisy environment.

where $\underline{h}_m(t)$ denotes the FIR filter with length $L_{\underline{h}}$ representing the RIR between the speech source and the m-th microphone without the direct path delay, t_m is the delay of the direct path signal, and $*$ denotes the convolution operator. The reverberant speech signal $\underline{x}_m(t)$ can be further decomposed into a desired speech signal $\underline{d}_m(t)$ and an undesired speech signal $\underline{u}_m(t)$ as

$$\underline{x}_m(t) = \underbrace{\sum_{l=0}^{\underline{\tau}-1} \underline{h}_m(l)\underline{s}(t-t_m-l)}_{\underline{d}_m(t)} + \underbrace{\sum_{l=0}^{L_{\underline{h}}-\underline{\tau}-1} \underline{h}_m(l+\underline{\tau})\underline{s}(t-t_m-\underline{\tau}-l)}_{\underline{u}_m(t)}. \tag{2.3}$$

The desired speech signal $\underline{d}_m(t)$ is defined as the convolution of the delayed anechoic speech signal $\underline{s}(t)$ with the early part of the m-th RIR $\underline{h}_m$, corresponding to the first $\underline{\tau}$ samples. Therefore, the desired speech signal $\underline{d}_m(t)$ consists of the direct speech signal and the early reflections. The undesired speech signal $\underline{u}_m(t)$ consists of the late reverberation and is defined as the convolution of the delayed anechoic speech signal $\underline{s}(t)$ with the late part of the m-th RIR $\underline{h}_m$, corresponding to all samples after $\underline{\tau}$.

When multiple microphones are available, i.e., $M > 1$, the anechoic speech signal $\underline{s}(t)$ can be obtained from the reverberant speech signals $\underline{x}_m(t)$ under certain conditions. More specifically, assuming that the RIRs $\underline{h}_m$ do not share any common zeros in the z-plane, the multiple-input/output inverse theorem (MINT) [123] states that there exists a set of inverse filters $\underline{h}_m^{\text{inv}}$, $m \in \{1, \dots, M\}$, such that the anechoic speech signal $\underline{s}(t)$ can be recovered up to a delay t_m by filtering and summing the reverberant microphone signals $\underline{x}_m(t)$, i.e.,

$$\underline{s}(t-t_m) = \sum_{m=1}^{M} \sum_{l=0}^{L_{\underline{h}}^{\text{inv}}-1} \underline{h}_m^{\text{inv}}(l)\underline{x}_m(t-l), \tag{2.4}$$

with $L_{\underline{h}}^{\text{inv}} \geq \left\lceil \frac{L_{\underline{h}}-1}{M-1} \right\rceil$ the length of the inverse filters, where $\lceil . \rceil$ denotes the ceiling operator. Using the expression for the anechoic speech signal $\underline{s}(t)$ in (2.4), the undesired speech signal $\underline{u}_m(t)$ in (2.3) can be expressed as

$$\underline{u}_m(t) = \sum_{m'=1}^{M} \sum_{l_1=0}^{L_{\underline{h}}-\underline{\tau}-1} \sum_{l_2=0}^{L_{\underline{h}}^{\text{inv}}-1} \underline{h}_m(l_1+\underline{\tau})\underline{h}_{m'}^{\text{inv}}(l_2)\underline{x}_{m'}\left(t-\underline{\tau}-(l_1+l_2)\right). \tag{2.5}$$

The expression for the undesired speech signal in (2.5) can be rewritten as

$$\underline{u}_m(t) = \sum_{m'=1}^{M} \sum_{l=0}^{L_{\underline{g}}-1} \underline{x}_{m'}(t-\underline{\tau}-l)\underline{g}_{m',m}(l), \tag{2.6}$$

i.e., the undesired speech signal $\underline{u}_m(t)$ at the m-th microphone can be expressed as the sum of filtered delayed reverberant speech signals on all microphones. The filter coefficients $\underline{g}_{m',m}(t)$ depend on the RIRs $\underline{h}_m$ and the corresponding inverse filters $\underline{h}_m^{\text{inv}}$, and can be written as

$$\underline{g}_{m',m}(t) = \sum_{l_1=0}^{L_{\underline{h}}-\underline{\tau}-1} \sum_{l_2=0}^{L_{\underline{h}}^{\text{inv}}-1} \underline{h}_m(l_1+\underline{\tau})\underline{h}_{m'}^{\text{inv}}(l_2)\delta\left(l_1+l_2-t\right), \tag{2.7}$$

where $\delta(.)$ is the Kronecker delta function, resulting in $L_{\underline{g}} = L_{\underline{h}} + L_{\underline{h}}^{\text{inv}} - \underline{\tau} - 1$ coefficients for each filter $\underline{g}_{m',m}(t)$, i.e., the length of the prediction filter obtained using (2.7) satisfies the following relation

$$L_{\underline{g}} \geq L_{\underline{h}} + \left\lceil \frac{L_{\underline{h}}-1}{M-1} \right\rceil - \underline{\tau} - 1. \tag{2.8}$$

Using the filters $\underline{g}_{m,m'}$ from (2.7), the undesired speech signal can be compactly written as

$$\underline{u}_m(t) = \sum_{m'=1}^{M} \underline{g}_{m',m}(t) * \underline{x}_{m'}(t-\underline{\tau}), \tag{2.9}$$

such that the signal model for the reverberant speech signal in (2.3) can be rewritten as

$$\underline{x}_m(t) = \underline{d}_m(t) + \sum_{m'=1}^{M} \underline{g}_{m',m}(t) * \underline{x}_{m'}(t-\underline{\tau}). \tag{2.10}$$

The filters $\underline{g}_{m',m}$ are typically referred to as the prediction filters, since they can be used to predict the undesired speech signal $\underline{u}_m(t)$ at the m-th microphone. The signal model in (2.10) is referred to as the multi-channel linear prediction (MCLP) model, since it is based on predicting the undesired speech signal $\underline{u}_m(t)$ at the current time by linear filtering of the delayed reverberant speech signal $\underline{x}_{m'}(t-\underline{\tau})$ at all microphones. The signal model for the desired speech signal in (2.3) implies that perfect dereverberation up to a scaling and a direct path delay can be achieved when $\underline{\tau} = 1$, i.e., the desired speech signal is equal to the anechoic speech signal scaled with $\underline{h}_m(0)$ and delayed for t_m samples. When $\underline{\tau} > 1$, some early reflections are preserved in the desired speech signal, i.e., the desired speech signal is equal to the anechoic speech signal convolved with the first $\underline{\tau}$ samples of the RIR $\underline{h}_m$ and delayed for t_m.

If the true RIRs $\underline{h}_m$ and the corresponding inverse filters $\underline{h}_m^{\text{inv}}$ are perfectly known, the prediction filters $\underline{g}_{m',m}$ can be computed using the expression in (2.7), which can then be used to compute the undesired speech signal $\underline{u}(t)$. However, the algorithms presented in this thesis do not assume any knowledge of the RIRs $\underline{h}_m$ and aim to estimate the prediction filters and the desired speech signal without exploiting the RIRs, i.e., the expression in (2.7) only serves to ensure that such filters in theory exist.

Assuming that a batch of T time-domain samples is available, the observed reverberant speech signal and the desired speech signal in the time domain can be represented in vector form as

$$\begin{aligned} \underline{\mathbf{x}}_m &= [\underline{x}_m(1), \underline{x}_m(2), \ldots, \underline{x}_m(T)]^{\mathsf{T}} \\ \underline{\mathbf{d}}_m &= [\underline{d}_m(1), \underline{d}_m(2), \ldots, \underline{d}_m(T)]^{\mathsf{T}}, \end{aligned} \tag{2.11}$$

with $.^{\mathsf{T}}$ denoting the transpose operator. The MCLP signal model in (2.10) can then be written in vector form as

$$\underline{\mathbf{x}}_m = \underline{\mathbf{d}}_m + \sum_{m'=1}^{M} \tilde{\underline{\mathbf{X}}}_{m',\underline{\tau}} \underline{\mathbf{g}}_{m',m}, \tag{2.12}$$

where $\tilde{\underline{\mathbf{X}}}_{m',\underline{\tau}} \in \mathbb{R}^{T \times L_{\underline{g}}}$ is a convolution matrix of the time-domain signal $\underline{x}_{m'}(t)$ delayed for $\underline{\tau}$ samples, i.e.,

$$\tilde{\underline{\mathbf{X}}}_{m',\underline{\tau}} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x_{m'}(1) & 0 & \ddots & \vdots \\ x_{m'}(2) & x_{m'}(1) & \ddots & \vdots \\ \vdots & x_{m'}(2) & \ddots & 0 \\ \vdots & \vdots & \ddots & x_{m'}(1) \\ \vdots & \vdots & \ddots & \vdots \\ x_{m'}(T-\underline{\tau}) & \cdots & \cdots & x_{m'}(T-\underline{\tau}-L_{\underline{g}}+1) \end{bmatrix}, \tag{2.13}$$

and $\underline{\mathbf{g}}_{m',m} \in \mathbb{R}^{L_{\underline{g}}}$ is the vector of the prediction filter relating the m'-th microphone to the m-th microphone, i.e.,

$$\underline{\mathbf{g}}_{m',m} = \left[\underline{g}_{m',m}(0), \underline{g}_{m',m}(1), \ldots, \underline{g}_{m',m}(L_{\underline{g}} - 1)\right]^{\mathsf{T}}. \tag{2.14}$$

By defining the multi-channel convolution matrix $\tilde{\underline{\mathbf{X}}}_{\underline{\mathcal{T}}} \in \mathbb{R}^{T \times ML_{\underline{g}}}$ and the multi-channel prediction filter $\underline{\mathbf{g}}_m \in \mathbb{R}^{ML_{\underline{g}}}$ as

$$\begin{aligned} \tilde{\underline{\mathbf{X}}}_{\underline{\mathcal{T}}} &= \left[\tilde{\underline{\mathbf{X}}}_{1,\underline{\mathcal{T}}}, \ldots, \tilde{\underline{\mathbf{X}}}_{M,\underline{\mathcal{T}}}\right], \\ \underline{\mathbf{g}}_m &= \left[\underline{\mathbf{g}}_{1,m}^{\mathsf{T}}, \ldots, \underline{\mathbf{g}}_{M,m}^{\mathsf{T}}\right]^{\mathsf{T}}, \end{aligned} \tag{2.15}$$

the MCLP signal model for the m-th channel in (2.12) can be written as

$$\underline{\mathbf{x}}_m = \underline{\mathbf{d}}_m + \tilde{\underline{\mathbf{X}}}_{\underline{\mathcal{T}}} \underline{\mathbf{g}}_m. \tag{2.16}$$

The corresponding MCLP signal model for all M channels can be written in matrix form as

$$\underline{\mathbf{X}} = \underline{\mathbf{D}} + \tilde{\underline{\mathbf{X}}}_{\underline{\mathcal{T}}} \underline{\mathbf{G}}, \tag{2.17}$$

where the multi-channel reverberant speech matrix $\underline{\mathbf{X}} \in \mathbb{R}^{T \times M}$, the multi-channel desired speech matrix $\underline{\mathbf{D}} \in \mathbb{R}^{T \times M}$, and the multiple-input multiple-output (MIMO) prediction filter $\underline{\mathbf{G}} \in \mathbb{R}^{ML_{\underline{g}} \times M}$ are defined as

$$\begin{aligned} \underline{\mathbf{X}} &= [\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_M] \\ \underline{\mathbf{D}} &= [\underline{\mathbf{d}}_1, \ldots, \underline{\mathbf{d}}_M] \\ \underline{\mathbf{G}} &= \left[\underline{\mathbf{g}}_1, \ldots, \underline{\mathbf{g}}_M\right]. \end{aligned} \tag{2.18}$$

Finally, the MCLP-based signal model for the complete observed microphone signal can be written as

$$\underline{\mathbf{Y}} = \underline{\mathbf{D}} + \tilde{\underline{\mathbf{X}}}_{\underline{\mathcal{T}}} \underline{\mathbf{G}} + \underline{\mathbf{V}}, \tag{2.19}$$

where $\underline{\mathbf{Y}} \in \mathbb{R}^{T \times M}$ is the multi-channel microphone signal matrix, and $\underline{\mathbf{V}} \in \mathbb{R}^{T \times M}$ is the multi-channel noise matrix.

2.1.2 *Subband signal model*

While the wideband signal model from the previous section holds perfectly when the MINT conditions are fulfilled, the length of the prediction filter in (2.8) is approximately proportional to the length of the time-domain RIRs which can be prohibitively large, especially for large reverberation times.

In order to reduce the length of the filters, a similar model can be used in the subband domain, e.g., in the short-time Fourier transform (STFT) domain [154,155] or the polyphase filter bank domain [158]. Additionally, certain properties of the speech signal can be more naturally modeled and exploited in the TF domain, e.g., sparsity or TF structure. Without loss of generality, in this thesis we will consider the subband model in the STFT domain.

Let $s(k,n) \in \mathbb{C}$ denote a coefficient of the anechoic speech signal in the STFT domain, with $k \in \{1,\dots,K\}$ denoting the subband index and $n \in \{1,\dots,N\}$ denoting the time frame index. The signal observed at the m-th microphone, $m \in \{1,\dots,M\}$, can be modeled in the STFT domain as

$$y_m(k,n) = x_m(k,n) + v_m(k,n), \tag{2.20}$$

where $x_m(k,n) \in \mathbb{C}$ is the STFT coefficient of the reverberant speech signal observed at the m-th microphone and $v_m(k,n) \in \mathbb{C}$ is the STFT coefficient of the noise signal observed at the m-th microphone. The time-domain signal model in (2.2) can be approximated in the STFT domain using the convolutive transfer function approximation [184–186], i.e., the reverberant speech signal $x_m(k,n)$ can be modeled using a subband model in the STFT domain as

$$x_m(k,n) = \sum_{l=0}^{L_h-1} h_m(k,l)s(k,n-n_m-l) = h_m(k,n) * s(k,n-n_m), \tag{2.21}$$

where $h_m(k,n)$ represents the (convolutive) ATF between the speech source and the m-th microphone with length L_h time frames, n_m is the delay of the direct path, and $*$ denotes the convolution operator operating across time frames. The subband model in (2.21) is practically very interesting because the time-domain convolution in (2.2) is divided into a set of convolutions across time frames in the TF domain in (2.21). This subband convolution model has been successfully used in various acoustical signal processing applications [77, 139, 155, 184–186]. The advantage of the subband model in (2.21) is that the convolutive ATFs in the TF domain are much shorter than the RIRs in the time domain, i.e., $L_h \ll L_{\underline{h}}$. Consequently, the subband model can be used to significantly reduce the computational complexity due to shorter ATFs and possibility of independent processing in each subband.

As in the previous section, the reverberant speech signal $x_m(k,n)$ can be decomposed into a desired speech signal $d_m(k,n)$ and an undesired speech signal $u_m(k,n)$ as

$$x_m(k,n) = \underbrace{\sum_{l=0}^{\tau-1} h_m(k,l)s(k,n-n_m-l)}_{d_m(k,n)} + \underbrace{\sum_{l=0}^{L_h-\tau-1} h_m(k,l+\tau)s(k,n-n_m-\tau-l)}_{u_m(k,n)}. \tag{2.22}$$

The desired speech signal $d_m(k,n)$ consists of the direct signal and early reflections, corresponding to the first τ coefficients of the ATF $h_m(k,n)$.

Assuming that a batch of N time frames is available, the subband model in (2.21) can be represented in the MCLP form similarly as for the wideband model in (2.12) as

$$\mathbf{x}_m(k) = \mathbf{d}_m(k) + \sum_{m'=1}^{M} \tilde{\mathbf{X}}_{m',\tau}(k)\mathbf{g}_{m',m}(k), \tag{2.23}$$

where the vectors $\mathbf{x}_m(k)$ and $\mathbf{d}_m(k)$ are defined as

$$\begin{aligned} \mathbf{x}_m(k) &= \left[x_m(k,1), x_m(k,2), \dots, x_m(k,N)\right]^{\mathsf{T}}, \\ \mathbf{d}_m(k) &= \left[d_m(k,1), d_m(k,2), \dots, d_m(k,N)\right]^{\mathsf{T}}. \end{aligned} \tag{2.24}$$

The matrix $\tilde{\mathbf{X}}_{m',\tau}(k) \in \mathbb{C}^{N \times L_g}$ is a convolution matrix of $x_{m'}(k,n)$ delayed for τ time frames in the k-th subband, i.e.,

$$\tilde{\mathbf{X}}_{m',\tau}(k) = \begin{bmatrix} 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x_{m'}(k,1) & 0 & \ddots & \vdots \\ x_{m'}(k,2) & x_{m'}(k,1) & \ddots & \vdots \\ \vdots & x_{m'}(k,2) & \ddots & 0 \\ \vdots & \vdots & \ddots & x_{m'}(k,1) \\ \vdots & \vdots & \ddots & \vdots \\ x_{m'}(k,N-\tau) & \dots & \dots & x_{m'}(k,N-\tau-L_g+1) \end{bmatrix}, \tag{2.25}$$

and $\mathbf{g}_{m',m}(k) \in \mathbb{C}^{L_g}$ is the vector form of the prediction filter relating the m'-th channel to m-th channel in the k-th subband, i.e.,

$$\mathbf{g}_{m',m}(k) = \left[g_{m',m}(k,0), g_{m',m}(k,1), \dots, g_{m',m}(k,L_g-1)\right]^{\mathsf{T}}. \tag{2.26}$$

Defining the multi-channel convolution matrix $\tilde{\mathbf{X}}_\tau(k) \in \mathbb{C}^{N \times ML_g}$ and the multi-channel prediction filter $\mathbf{g}_m(k) \in \mathbb{C}^{ML_g}$ as

$$\begin{aligned} \tilde{\mathbf{X}}_\tau(k) &= \left[\tilde{\mathbf{X}}_{1,\tau}(k), \dots, \tilde{\mathbf{X}}_{M,\tau}(k)\right] \\ \mathbf{g}_m(k) &= \left[\mathbf{g}_{1,m}^{\mathsf{T}}(k), \dots, \mathbf{g}_{M,m}^{\mathsf{T}}(k)\right]^{\mathsf{T}}, \end{aligned} \tag{2.27}$$

the MCLP signal model for the m-th channel in the k-th subband can be written compactly in a vector form as

$$\mathbf{x}_m(k) = \mathbf{d}_m(k) + \tilde{\mathbf{X}}_\tau(k)\mathbf{g}_m(k). \tag{2.28}$$

Similarly as for the wideband model in Section 2.1.1, if the true ATFs $h_m(k,n)$ and the corresponding inverse ATFs $h_m^{\mathrm{inv}}(k,n)$ in the k-th subband are perfectly known, the prediction filters can be computed similarly as in (2.7). However, the algorithms presented in this thesis do not assume any knowledge of the ATFs $h_m(k,n)$ and aim to estimate the prediction filters and the desired speech signal without exploiting the ATFs.

The signal model for the desired speech signal in (2.22) implies that when $\tau = 1$ the desired speech signal is equal to the delayed anechoic speech signal coefficient scaled with the complex-valued early ATF signal $h_m(k,0)$. Since the scaling depends on the subband index k, some coloration will typically be present in the desired speech signal. When $\tau > 1$, additional early reflections are also preserved in the desired speech signal.

The corresponding MCLP signal model for all M channels can be written in matrix form as

$$\mathbf{X}(k) = \mathbf{D}(k) + \tilde{\mathbf{X}}_\tau(k)\mathbf{G}(k), \tag{2.29}$$

where the multi-channel reverberant speech matrix $\mathbf{X}(k) \in \mathbb{C}^{N \times M}$, the multi-channel desired speech matrix $\mathbf{D}(k) \in \mathbb{C}^{N \times M}$, and the MIMO prediction filter $\mathbf{G}(k) \in \mathbb{C}^{ML_g \times M}$ are defined as

$$\begin{aligned}\mathbf{X}(k) &= \left[\mathbf{x}_1(k), \ldots, \mathbf{x}_M(k)\right], \\ \mathbf{D}(k) &= \left[\mathbf{d}_1(k), \ldots, \mathbf{d}_M(k)\right], \\ \mathbf{G}(k) &= \left[\mathbf{g}_1(k), \ldots, \mathbf{g}_M(k)\right].\end{aligned} \tag{2.30}$$

Finally, the MCLP-based signal model for the complete observed microphone signal in the k-th subband can be written as

$$\mathbf{Y}(k) = \mathbf{D}(k) + \tilde{\mathbf{X}}_\tau(k)\mathbf{G}(k) + \mathbf{V}(k), \tag{2.31}$$

where $\mathbf{Y}(k) \in \mathbb{C}^{N \times M}$ is the multi-channel microphone signal matrix in the k-th subband, and $\mathbf{V}(k) \in \mathbb{C}^{N \times M}$ is the multi-channel noise matrix in the k-th subband.

2.2 Sparsity of speech signals

In this section we discuss sparsity of speech signals in the TF domain, its application in speech signal processing and the influence of reverberation on TF sparsity.

In general, sparsity of a vector is related to the magnitude of its elements, e.g., a vector with many elements equal to zero is referred to as a sparse vector. In other words, only a relatively small number of elements in a sparse vector have a large magnitude. Note that the notion of sparsity can be extended to matrices and other signals in general. Over the past two decades, sparse regularization has become widely utilized to regularize different ill-posed problems in signal and image processing and machine learning [187].

In the time domain speech signals typically do not exhibit a very high level of sparsity. However, in the TF domain they typically have a sparse representation, since the energy of most speech signals is dominantly contained in a relatively small number of TF bins [188–190]. The observed sparsity of speech signals in the TF domain can be attributed to the combined effects of their spectral shape, harmonic structure and speech pauses. In general, sparsity of audio signals can be related to the sound production system and its resonances, and the temporal activity of the production mechanism, which together result in harmonic structures over frequencies and varying temporal structures [191]. In addition to the widely used STFT, examples of other relevant TF transforms in the context of speech signal processing include the modified discrete cosine transform (MDCT), or the non-stationary Gabor transform [192].

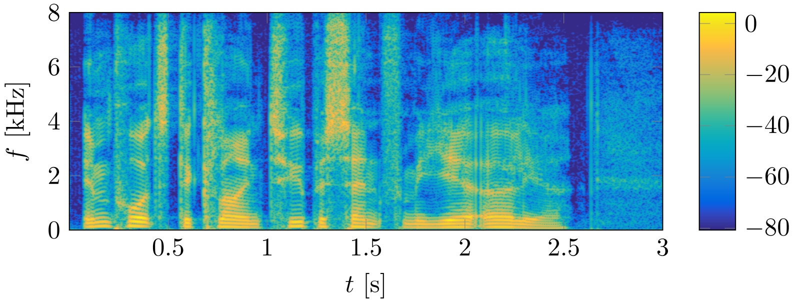
In the context of speech and audio processing, sparsity in the TF domain has been exploited in many different applications. A classical example is source separation [188, 189, 193–196], where it is typically assumed that the sources are approximately mutually disjoint in the TF domain, i.e., at each TF point only a single source is dominant. Using this assumption, the mixing matrix can be estimated and used to recover the source signals. Another example is sparsity-based denoising, where assuming that the speech signal has a sparse TF representation, the desired speech signal cam be recovered from the noisy observation [197–199]. Similarly, sparsity has been extensively exploited in other applications, such as beamforming [200, 201], declipping and inpainting [202–205], and coding of speech and audio signals [191, 206, 207].

Reverberation may have a significant influence on the sparsity of speech signals in the TF domain. More specifically, the temporal smearing due to reverberation influences the distribution of the speech energy in the TF domain. This results in a decreased number of TF coefficients with very low energy, which effectively reduces the sparsity of the speech signal captured at the microphone inside a reverberant enclosure [189]. The influence of reverberation on the sparsity of the speech signal in the STFT domain is illustrated in Fig. 2.2, which depicts spectrograms of anechoic speech signal and reverberant speech signal in a reverberant enclosure with the reverberation time of $T_{60} \approx 700$ ms. By comparing the spectrograms of the clean and the reverberant speech signal, it is clear that the reverberant signal exhibits a lower level of sparsity. More specifically, smearing of the speech energy across speech pauses between the phonemes results in a reduced sparsity in the STFT domain. The distribution of the magnitude of the STFT coefficients of the corresponding anechoic and reverberant speech signals is depicted in Fig. 2.3. From these histograms it is evident that the number of TF coefficients that are very close to zero is reduced while the number of non-zero coefficients is increased in the presence of reverberation. In the remainder of this thesis, this difference in TF sparsity between the anechoic speech signal and the observed reverberant microphone signal will be exploited in an MCLP-based framework to achieve speech dereverberation.

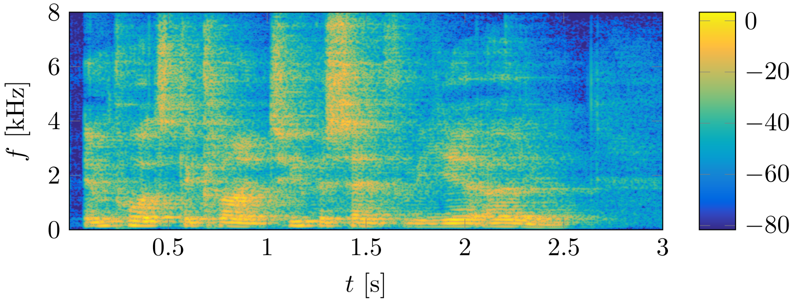
2.3 Instrumental performance measures

In this section, we present several instrumental performance measures used to evaluate the performance of the speech enhancement algorithms in the remainder of the thesis.

Different instrumental measures have been proposed to evaluate speech dereverberation performance [5]. Typically, these instrumental measures are classified into



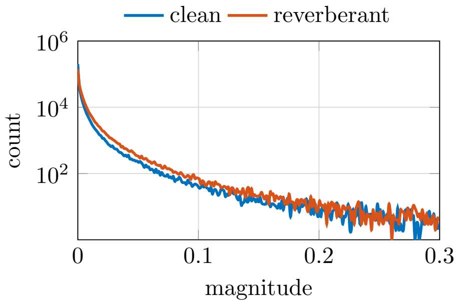
(a) Anechoic speech signal



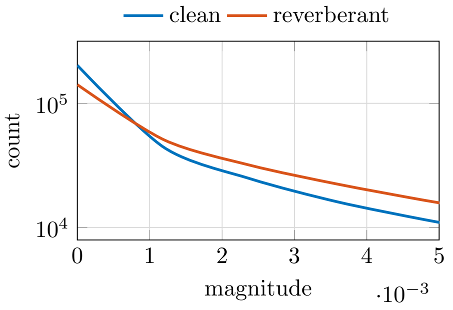
(b) Reverberant speech signal

Fig. 2.2: Influence of reverberation on sparsity of a speech signal in the STFT domain illustrated using a spectrogram of (a) anechoic speech signal, and (b) reverberant speech signal ($T_{60} \approx 700$ ms).

channel-based measures and signal-based measures. Channel-based measures are computed using the impulse response between the source signal and the received or processed signal, e.g., the RIR or the equalized impulse response [5, 52, 53]. Typical examples of channel-based measures are the direct-to-reverberant ratio (DRR) and the energy-decay curve (EDC). These measures are especially useful for speech dereverberation based on equalization, where dereverberation is performed by shaping the equalized impulse response [5, 52, 53]. On the other hand, signal-based measures are computed directly using the speech signal under investigation [5, 208]. Therefore, signal-based measures are applicable in a wider range of scenarios, e.g., when the impulse responses are not available, and provide a reasonable means of objective evaluation [5]. These measures can be further divided into intrusive measures, requiring a reference signal, and non-intrusive measures, not requiring a reference signal. Intrusive measures are generally based on some measure of discrepancy between the speech signal under investigation and the reference signal, which is typically the anechoic speech signal or the direct signal with early reflections. Examples of such measures are the signal-to-noise ratio (SNR)-based measures, cepstral distance (CD) [209], the log-likelihood ratio (LLR) [209], and the perceptual evaluation of speech quality (PESQ) measure [210, 211]. Non-intrusive measures typically operate



(a) Histogram



(b) Detail of the histogram

Fig. 2.3: Influence of reverberation on sparsity of a speech signal in the STFT domain illustrated using a histogram of the magnitude spectrum of anechoic speech and reverberant speech ($T_{60} \approx 700$ ms): (a) full histogram, and (b) detail for small magnitudes.

by first extracting relevant features from the speech signal under investigation and passing them through a prediction model, which can be an analytical function or a pre-trained model. Examples of such measures are the speech-to-reverberation modulation ratio (SRMR) measure [212] and its improved variant [213], non-intrusive measures based on pre-trained machine-learning models [214], and others [215].

In this thesis, we will use the frequency-weighted segmental signal-to-noise ratio (fwsSNR) [22,216] and the PESQ measure [210,211] to evaluate the performance of speech enhancement methods, which can correlate well with the perceived amount of reverberation and the perceptual quality of the speech signals. The fwsSNR measure exhibits a high correlation with the perceived amount of reverberation [22], and a relatively high correlation with the perceptual quality of speech signals [215, 217]. Although the PESQ measure was not designed for evaluating the perceptual quality of dereverberation algorithms, it is commonly used since it exhibits a relatively high correlation with the perceived amount of reverberation [215], and a high correlation with the perceptual quality of speech signals [52, 215, 217, 218].

The fwsSNR measure is computed as described in [23, 219]. More specifically, let $\hat{x}(k,n)$ denote the TF coefficients of the signal under investigation and $x(k,n)$ the TF coefficients of the reference signal. The fwsSNR is then computed as

$$\text{fwsSNR} = \frac{1}{N} \sum_{n=1}^{N} \text{clip}_{-10}^{35} \left(\text{fwsSNR}(n) \right), \tag{2.32}$$

where $\text{clip}_{-10}^{35}(.)$ is a clipping operator limiting the value of each time frame between -10 dB and 35 dB, i.e., $\text{clip}_{-10}^{35}(x) = \max\left(\min\left(x, 35\right), -10\right)$ and $\text{fwsSNR}(n)$ is the frequency-weighted segmental SNR computed for the n-th time frame as

$$\text{fwsSNR}(n) = \frac{10}{\sum_{k=1}^{K} f(k,n)} \sum_{k=1}^{K} f(k,n) \log_{10} \frac{|x(k,n)|^2}{\left(|x(k,n)| - |\hat{x}(k,n)|\right)^2} \tag{2.33}$$

where $f(k,n)$ is the weight for the k-th subband and n-th frame [219]. The weights are computed using the magnitude spectrum of the reference signal as $f(k,n) = |x(k,n)|^{0.2}$ [23, 219].

The PESQ measure is computed as described in the wideband extension of ITU P.862 [210, 211]. The obtained values correspond to the mean opinion score (MOS) assessing the quality of the speech signal under investigation. These values are always between 1 and 4.5, with 1 corresponding to a bad perceptual quality and 4.5 corresponding to an excellent perceptual quality. A detailed description of the mapping between the raw PESQ score, ranging between -0.5 and 4.5, and the PESQ MOS can be found in [211].

Since both fwsSNR and PESQ are intrusive measures, they require a reference signal which is typically selected as the clean speech signal observed at the reference microphone. In the remainder of the thesis, we report the change of the fwsSNR and PESQ values, ΔfwsSNR and ΔPESQ, computed as the difference between the measure obtained for the output signal and the measure obtained for the input signal. A positive value indicates an improvement relative to the input signal in terms of the corresponding measure, while a negative value indicates a deterioration relative to the input signal in terms of the corresponding measure.

2.4 Summary

In this chapter, we presented two signal models for the microphone signals of a speech source observed in a reverberant enclosure. More specifically, we focused on signal models based on multi-channel linear prediction, since these models will be employed in the remainder of the thesis. We defined the wideband signal model, employing time-domain RIRs, and the subband signal model, which is a widely used and efficient approximation of the wideband signal model using convolutive ATFs.

The dereverberation methods proposed in the remainder of the thesis will mainly exploit the sparsity of the speech signal. Therefore, we introduced the notion of sparsity, discussed sparsity of speech signals in the TF domain, and briefly reviewed its application in speech signal processing. Furthermore, we showed that reverberation decreases the sparsity of speech signal in the TF domain.

Finally, we presented the instrumental measures that will be used to evaluate the dereverberation performance, more specifically, signal-dependent intrusive fwsSNR and PESQ measures.

3

SPARSE MULTI-CHANNEL LINEAR PREDICTION FOR SPEECH DEREVERBERATION

In this chapter, we consider the noiseless case and present a single-output batch method for blind speech dereverberation based on sparse MCLP using the subband signal model introduced in the previous chapter.

A subband MCLP-based method for blind speech dereverberation based on variance-normalized delayed MCLP has been proposed in [154, 155], referred to as weighted prediction error (WPE). This method assumes an autoregressive model of the reverberation process, i.e., it is assumed that the reverberant speech signal at a certain time can be predicted from the previous samples of the reverberant microphone signals. The desired speech signal can then be estimated as the prediction error, i.e., speech dereverberation boils down to estimating the parameters of the MCLP model. An additional delay is typically introduced in the MCLP model in order to prevent distortion of the short-time correlation of the speech signal, thereby only suppressing late reverberation [107, 154]. Conventionally, the complex-valued STFT coefficients of the desired speech signal are modeled using a time-varying Gaussian (TVG) model, under the assumption that the STFT coefficients can be modeled locally (i.e., in each TF bin) using a complex Gaussian distribution with an unknown variance. Speech dereverberation using WPE is then performed by estimating the unknown parameters of the MCLP and TVG models in a ML sense.

This chapter is partly based on:

[176] A. Jukić, S. Doclo, "Speech dereverberation using weighted prediction error with Laplacian model of the desired signal," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 5172–5176.

[177] A. Jukić, T. van Waterschoot, T. Gerkmann, S. Doclo, "Speech dereverberation with multi-channel linear prediction and sparse priors for the desired signal," in *Proceedings of the Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Nancy, France, May 2014, pp. 23–26.

[178] A. Jukić, T. van Waterschoot, T. Gerkmann, S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 9, pp. 1509–1520, Sept. 2015.

In this chapter, we aim to provide a different and more general view on MCLP-based speech dereverberation in the STFT domain. Firstly, instead of assuming a TVG model we consider a general sparse prior for the desired speech signal and use ML estimation to estimate the parameters of the MCLP model [177]. The sparse prior is formulated using a variational representation that is based on a locally Gaussian model [220–222]. The used model for the desired speech signal can be interpreted as a TVG model with an additional hyperprior on the unknown variance. To derive a practical algorithm, we focus on sparse priors in the family of complex generalized Gaussian (CGG) distributions, resulting in an iterative sparse MCLP-based speech dereverberation method. In the proposed framework, we show that the conventional WPE method [154] can be considered as a special case based on a prior that strongly promotes sparsity of the estimated speech signal. Secondly, we reformulate the sparse MCLP-based method with CGG prior as an optimization problem minimizing the ℓ_p-norm of the desired speech signal. Furthermore, we show that the iterative algorithm for sparse MCLP with CGG prior is equivalent to an iteratively reweighted least-squares algorithm applied to ℓ_p-norm minimization [223], with WPE being a special case. In the experimental section we evaluate the performance of the conventional and the proposed methods for different acoustic scenarios using several instrumental speech quality measures. The obtained results show that the speech enhancement performance can be consistently improved using the proposed methods. While the improvements are mild, these come with no additional computational cost, and are consistent with the derived theoretical insights.

In Section 3.1 we mathematically formulate the considered problem of blind speech dereverberation using MCLP in the STFT domain. The conventional method for MCLP-based speech dereverberation based on a TVG model for the desired speech signal is presented in Section 3.2. The proposed method using a general sparse prior for the desired speech signal is presented in Section 3.3. In Section 3.4 both the conventional and the proposed methods are reformulated as a minimization of the ℓ_p-norm of the desired speech signal. The simulation results are presented in Section 3.5.

3.1 Problem formulation

We consider an acoustic scenario with a single static speech source captured by M microphones in a reverberant enclosure without the presence of additive noise. Given a batch of N time frames, using the subband signal model in (2.28) and assuming an arbitrarily chosen reference microphone ref $\in \{1, \dots, M\}$, the MCLP-based signal model for the reverberant microphone signal at the reference microphone in the k-th subband can be written as

$$\mathbf{x}_{\text{ref}}(k) = \mathbf{d}_{\text{ref}}(k) + \tilde{\mathbf{X}}_{\tau}(k)\mathbf{g}_{\text{ref}}(k), \tag{3.1}$$

where $\mathbf{x}_{\text{ref}}(k) \in \mathbb{C}^N$ is a vector of the STFT coefficients of the reference microphone signal, $\mathbf{d}_{\text{ref}}(k) \in \mathbb{C}^N$ is a vector of the STFT coefficients of the desired speech signal at the reference microphone, and $\tilde{\mathbf{X}}_{\tau}(k) \in \mathbb{C}^{N \times ML_g}$ and $\mathbf{g}_{\text{ref}}(k) \in \mathbb{C}^{ML_g}$ are the multi-channel convolution matrix and the multi-channel prediction filter defined in (2.27). The problem of blind speech dereverberation can now be formulated as the blind estimation of the desired speech signal $\mathbf{d}_{\text{ref}}(k)$ using only the reverberant observations $\mathbf{x}_m(k)$, i.e., without using the ATFs between the speech source and the microphones. As defined in (2.22), the desired speech signal $\mathbf{d}_{\text{ref}}(k)$ in (3.1) consists of the direct signal and early reflections, which are known to be possibly beneficial for speech intelligibility [27]. Using the signal model in (3.1) and given an estimate $\hat{\mathbf{g}}_{\text{ref}}(k)$ of the prediction filter, the desired speech signal in the k-th subband can be estimated as

$$\hat{\mathbf{d}}_{\text{ref}}(k) = \mathbf{x}_{\text{ref}}(k) - \tilde{\mathbf{X}}_{\tau}(k)\hat{\mathbf{g}}_{\text{ref}}(k) \tag{3.2}$$

with $\hat{(.)}$ denoting the estimated value. In this case, the desired speech signal $\mathbf{d}_{\text{ref}}$ can be interpreted as the prediction error of the delayed linear prediction model [154]. Dereverberation can be performed by estimating the multi-channel prediction filter $\hat{\mathbf{g}}_{\text{ref}}(k)$ for each subband k and applying (3.2). The enhanced time-domain signal is then obtained by performing the inverse STFT on the obtained STFT coefficients of the desired speech signal. Note that the MCLP signal is only valid if multiple microphones are available, but it can nevertheless also be used for single-channel dereverberation. A block scheme of an MCLP-based speech dereverberation system is depicted in Fig. 3.1. In the remainder of this chapter each subband will be processed independently and the index k will be omitted where possible for notational convenience.

The prediction delay τ in the signal model (3.1) should ensure that the direct speech signal in the reference microphone cannot be predicted using linear filtering in (3.2), i.e., that subtracting the predicted undesired speech signal does not destroy the short-time autocorrelation of the desired speech signal [107, 154]. If the inter-microphone distances are relatively small, as is commonly the case for many speech communication applications, the relative delays between the reference microphone and the other microphones are rather small, i.e., in the order of milliseconds, for all possible source positions. In this case, the required prediction delay only depends on the short-term autocorrelation of the speech signal, which is typically on the order of tens of milliseconds. A common practice for MCLP-based dereverberation is hence to set the prediction delay in the range of 30 to 40 ms [107, 154]. It has been shown in [154] that with a suitable prediction delay and given enough time frames, subtracting the undesired speech signal in (3.2) from the reference microphone signal does not change the direct component, while possibly altering the early reflections.

3.2 Conventional MCLP-based dereverberation using TVG model

Several MCLP-based speech dereverberation methods have been proposed using a TVG model for the desired signal [138, 154, 155, 162, 166]. More specifically, the desired signal coefficient $d_{\text{ref}}(k, n)$ in each TF bin is modeled as a zero-mean random variable by means of a circular complex Gaussian distribution with an unknown and

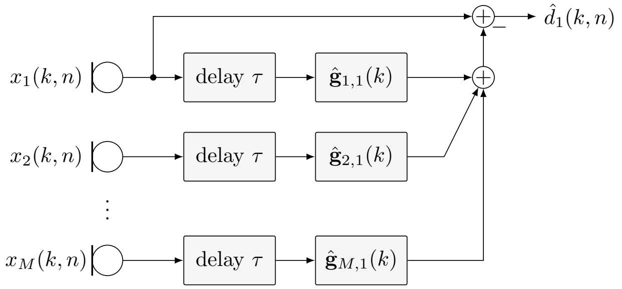


Fig. 3.1: A block scheme of an MCLP-based dereverberation system with the first microphone selected as the reference, i.e., ref = 1.

time-varying variance. The probability density function for the desired signal can then be written as

$$\mathcal{N}_{\mathbb{C}}\left(d_{\text{ref}}(k,n);0,\lambda(k,n)\right)=\frac{1}{\pi\lambda(k,n)}e^{-\frac{|d_{\text{ref}}(k,n)|^2}{\lambda(k,n)}}, \tag{3.3}$$

where the variance $\lambda(k,n)$ is considered to be an unknown parameter that needs to be estimated. The use of the TVG model is motivated by the fact that it can model any signal with a time-varying power spectrum [146,154,162]. Since the TVG model does not include any dependency across frequencies and since it is assumed that the STFT coefficients are independent across time, the likelihood function for all coefficients in a single subband (with the index k omitted) can be written as

$$\mathcal{L}\left(\mathbf{d}_{\text{ref}},\boldsymbol{\lambda}\right)=\mathcal{L}\left(\mathbf{g}_{\text{ref}},\boldsymbol{\lambda}\right)=\prod_{n=1}^{N}\mathcal{N}_{\mathbb{C}}\left(d_{\text{ref}}\left(n\right);0,\lambda(n)\right), \tag{3.4}$$

with $\boldsymbol{\lambda}=[\lambda(1),\ldots,\lambda(N)]^{\mathsf{T}}$ the vector of unknown variances and $\mathbf{g}_{\text{ref}}$ the prediction filter. Note that the desired speech signal $d_{\text{ref}}(n)$ in (3.4) depends on the prediction filter $\mathbf{g}_{\text{ref}}$ as in (3.2). The assumption that the coefficients of the desired speech signal are independent across time is a common simplification that has been successfully employed in dereverberation, but also in other speech enhancements methods [19]. The prediction filter $\mathbf{g}_{\text{ref}}$ and the variances $\boldsymbol{\lambda}$ are estimated by maximizing the likelihood in (3.4) with respect to the unknown parameters, i.e., minimizing the negative log-likelihood by solving the following optimization problem

$$\min_{\boldsymbol{\lambda}>0,\mathbf{g}_{\text{ref}}}\sum_{n=1}^{N}\frac{|d_{\text{ref}}(n)|^2}{\lambda(n)}+\log\pi\lambda(n). \tag{3.5}$$

Since the joint minimization of (3.5) with respect to the prediction filter $\mathbf{g}_{\text{ref}}$ and the variances $\boldsymbol{\lambda}$ can not be performed analytically, it was proposed in [154] to use an alternating optimization procedure. The original problem in (3.5) is split into two subproblems that can be solved more easily. The two subproblems are solved in an alternating fashion and the whole procedure is repeated iteratively, alternating between minimization with respect to $\boldsymbol{\lambda}$ and $\mathbf{g}_{\text{ref}}$. While this optimization procedure results in simple update rules, there is no guarantee that the iterative procedure will lead to the globally optimal solution (cf. Section 3.4).

ESTIMATION OF $\boldsymbol{\lambda}$: In the first step, the cost function in (3.5) is minimized with respect to the variances $\boldsymbol{\lambda}$, assuming that the prediction filter is fixed to the value $\hat{\mathbf{g}}_{\text{ref}}^{i-1}$ from the previous iteration[1]. The estimate $\hat{\mathbf{d}}_{\text{ref}}^{i-1}$ can then be calculated using (3.2) and the variance for the n-th time frame can be estimated as

$$\hat{\lambda}^i(n) = \arg\min_{\lambda(n)>0} \frac{\left|\hat{d}_{\text{ref}}^{i-1}(n)\right|^2}{\lambda(n)} + \log \pi\lambda(n). \tag{3.6}$$

The solution to this optimization problem is given as $\hat{\lambda}^i(n) = \left|\hat{d}_{\text{ref}}^{i-1}(n)\right|^2$, i.e.,

$$\hat{\boldsymbol{\lambda}}^i = \left|\hat{\mathbf{d}}_{\text{ref}}^{i-1}\right|^2, \tag{3.7}$$

where the absolute value and the power are applied element-wise on the elements of the vector. In practice, a small positive constant ε_{min} is added to the estimated variances to prevent division by zero.

ESTIMATION OF $\mathbf{g}_{\text{ref}}$: In the second step, the cost function in (3.5) is minimized with respect to the prediction filter $\mathbf{g}_{\text{ref}}$, assuming that the variances are fixed to the values $\hat{\boldsymbol{\lambda}}^i$ from the i-th iteration. A least-squares (LS) problem for estimating the prediction filter is formulated as

$$\hat{\mathbf{g}}_{\text{ref}}^i = \arg\min_{\mathbf{g}_{\text{ref}}} \sum_{n=1}^{N} \frac{\left|d_{\text{ref}}(n)\right|^2}{\hat{\lambda}^i(n)} = \arg\min_{\mathbf{g}_{\text{ref}}} \mathbf{d}_{\text{ref}}^{\mathsf{H}} \left(\hat{\boldsymbol{\Lambda}}^i\right)^{-1} \mathbf{d}_{\text{ref}}, \tag{3.8}$$

where $\hat{\boldsymbol{\Lambda}} = \operatorname{diag}\left(\hat{\boldsymbol{\lambda}}\right)$ is a diagonal matrix with $\hat{\boldsymbol{\lambda}}$ on its diagonal. By substituting (3.2) into (3.8) and assuming that the matrix $\tilde{\mathbf{X}}_\tau$ has a full column rank, an estimate $\hat{\mathbf{g}}_{\text{ref}}^i$ of the prediction filter can be computed as

$$\hat{\mathbf{g}}_{\text{ref}}^i = \left(\tilde{\mathbf{X}}_\tau^{\mathsf{H}} \left(\hat{\boldsymbol{\Lambda}}^i\right)^{-1} \tilde{\mathbf{X}}_\tau\right)^{-1} \tilde{\mathbf{X}}_\tau^{\mathsf{H}} \left(\hat{\boldsymbol{\Lambda}}^i\right)^{-1} \mathbf{x}_{\text{ref}}. \tag{3.9}$$

This alternating procedure is repeated until a convergence criterion is satisfied or a maximum number of iterations is exceeded. The iterative algorithm is typically initialized by setting the initial estimate of the prediction filters to be zero, which

---

[1] In the following $(.)^i$ denotes the estimated value in the i-th iteration.

is equivalent to setting the initial estimate of the desired speech signal to the reverberant microphone signal, i.e.,

$$\hat{\mathbf{d}}_{\mathrm{ref}}^{0} = \mathbf{x}_{\mathrm{ref}}. \tag{3.10}$$

Since the desired speech signal is estimated as the prediction error in the MCLP-based signal model, the presented method is often referred to as the weighted prediction error (WPE) method [154, 155]. The WPE method has been modified to include pre-trained log-spectral priors in [162], and instead of using a TVG model we have proposed to use a time-varying Laplacian model for the desired speech signal in [176].

3.3 MCLP-based dereverberation using a general sparse prior

As discussed in Section 2.2, anechoic clean speech signals are naturally sparse in the TF domain, and it is widely accepted that the STFT coefficients of speech signals can be well modeled using sparse priors. This holds both locally, by observing the STFT coefficients in a single TF bin [68, 224], as well as globally, when considering the distribution of the STFT coefficients in a single subband [190]. Although the real and imaginary parts of the complex-valued STFT coefficients are often assumed to be independent to simplify computations, it has been observed that the distribution of the complex-valued speech coefficients is actually approximately circular [225, 226]. In this section the desired speech signal coefficients in a single subband are modeled using a sparse circular prior, which is used to estimate the prediction filter in the MCLP model in (3.2). The proposed prior can be interpreted as a generalization of the TVG model (cf. Section 3.2), obtained by adding a hyperprior for the variances of the locally Gaussian model. A similar approach can be used with other local models, e.g., the locally Laplacian model in [176]. In Section 3.3.1 we present a convex representation of a general sparse prior, and use it for MCLP-based dereverberation in Section 3.3.2. In Section 3.3.3 we formulate dereverberation using a complex generalized Gaussian distribution, and relate the proposed method to the conventional method based on a TVG model in Section 3.3.4.

3.3.1 *Convex representation of a sparse prior*

Intuitively, a prior is considered to be sparse when it is super-Gaussian, i.e., it exhibits a higher peak at the origin and heavier tails than the corresponding Gaussian prior. Here, we consider a general circular sparse prior for a complex-valued random variable z that can be represented as

$$\mathrm{p}(z) = e^{-f(|z|)}. \tag{3.11}$$

In general, p(.) can represent a proper sparse prior, e.g., a probability density, or an improper (non-integrable) sparse prior. Formally, it can be shown that when $f'(t)/t$ is decreasing on $t \in (0, \infty)$, with $f'(.)$ denoting the derivative of $f(.)$, the prior will be super-Gaussian, i.e., sparse [220]. In this case, $\mathrm{p}(z)$ can be conveniently represented as a maximization over scaled Gaussians with different variances, i.e.,

$$\mathrm{p}(z) = \max_{\lambda > 0} \mathcal{N}_{\mathbb{C}}\left(z; 0, \lambda\right) \psi\left(\lambda\right), \tag{3.12}$$

where $\psi(.)$ is a scaling function that can be interpreted as a hyperprior on the variance λ [220, 222]. This representation of a sparse prior is often referred to as the convex variational type due to its roots in convex analysis. Obviously, the scaling function $\psi(.)$ in (3.12) is related to $f(.)$ in (3.11), but the scaling function is typically not required explicitly in practical algorithms. For completeness, the form of the hyperprior $\psi(.)$ for a given sparse prior $\mathrm{p}(.)$ is given in Appendix A.

3.3.2 *Speech dereverberation using a general sparse prior*

We now propose to model the STFT coefficients of the desired speech signal using the circular sparse prior $\mathrm{p}\left(d_{\mathrm{ref}}(n)\right) = e^{-f(|d_{\mathrm{ref}}(n)|)}$, with its convex representation given as

$$\mathrm{p}\left(d_{\mathrm{ref}}(n)\right) = \max_{\lambda(n) > 0} \mathcal{N}_{\mathbb{C}}\left(d_{\mathrm{ref}}(n); 0, \lambda(n)\right) \psi\left(\lambda(n)\right). \tag{3.13}$$

This model can be interpreted as a generalization of the TVG model, with an additional hyperprior on the variance $\lambda(n)$ determined by the scaling function $\psi(.)$. Similarly as in the conventional method, the prediction filter $\mathbf{g}_{\mathrm{ref}}$ can be estimated by maximizing the likelihood formed using the model in (3.13). ML estimation of the prediction filter results in the following optimization problem

$$\max_{\mathbf{g}_{\mathrm{ref}}} \prod_{n=1}^{N} \mathrm{p}\left(d_{\mathrm{ref}}(n)\right) = \max_{\mathbf{g}_{\mathrm{ref}}} \prod_{n=1}^{N} \max_{\lambda(n) > 0} \mathcal{N}_{\mathbb{C}}\left(d_{\mathrm{ref}}(n); 0, \lambda(n)\right) \psi\left(\lambda(n)\right). \tag{3.14}$$

The optimization problem in (3.14) is a probabilistic formulation of sparse MCLP. Since maximizing the likelihood in (3.14) is equivalent to minimizing the negative log-likelihood with respect to the prediction filter $\mathbf{g}_{\mathrm{ref}}$ and the variances $\boldsymbol{\lambda}$, the optimization problem in (3.14) can be rewritten as

$$\min_{\boldsymbol{\lambda} > 0, \mathbf{g}_{\mathrm{ref}}} \sum_{n=1}^{N} \frac{|d_{\mathrm{ref}}(n)|^2}{\lambda(n)} + \log \pi \lambda(n) - \log \psi\big(\lambda(n)\big), \tag{3.15}$$

with $d_{\mathrm{ref}}(n)$ depending on $\mathbf{g}_{\mathrm{ref}}$ as in (3.2). When compared with the optimization problem in (3.5), the problem in (3.15) contains an additional term that depends on the scaling function $\psi(.)$. Proceeding similarly as in the previous section, the optimization can again be performed by applying an iterative optimization procedure which alternates between minimization with respect to $\boldsymbol{\lambda}$ and minimization with respect to $\mathbf{g}_{\mathrm{ref}}$.

ESTIMATION OF $\boldsymbol{\lambda}$: Assuming that the prediction filter is fixed to $\hat{\mathbf{g}}_{\text{ref}}^{i-1}$, the variance at the n-th time frame can be obtained by solving the following problem

$$\hat{\lambda}^i(n) = \arg \min_{\lambda(n)>0} \frac{\left|\hat{d}_{\text{ref}}^{i-1}(n)\right|^2}{\lambda(n)} + \log \pi\lambda(n) - \log \psi\left(\lambda(n)\right). \tag{3.16}$$

For the general sparse prior in (3.11), the solution to (3.16) is equal to

$$\hat{\lambda}^i(n) = \frac{2\left|\hat{d}_{\text{ref}}^{i-1}(n)\right|}{f'\left(\left|\hat{d}_{\text{ref}}^{i-1}(n)\right|\right)}, \tag{3.17}$$

with the details described in Appendix A.2. Note that although the optimization problem in (3.16) includes the scaling function $\psi(.)$, the optimal $\lambda(n)$ for this subproblem in (3.17) depends only on $f(.)$, so the scaling function $\psi(.)$ does not need to be explicitly known (cf. Appendix A.2).

ESTIMATION OF $\mathbf{g}$: Assuming that the variances $\boldsymbol{\lambda}$ are fixed to the value from the i-th iteration, the same LS problem is obtained as in the conventional method, with the solution given by (3.9).

3.3.3 *Complex generalized Gaussian prior*

As an example of a parametric sparse circular prior resulting in a practical algorithm, in the remainder of this chapter we will consider the complex generalized Gaussian (CGG) prior given as [227]

$$\mathrm{p}(z) = \frac{p}{2\pi\zeta\Gamma(2/p)} e^{-\frac{|z|^p}{\zeta^{p/2}}}, \tag{3.18}$$

with the scale parameter $\zeta > 0$, the shape parameter $p \in (0, 2]$, and $\Gamma(.)$ denoting the Gamma function. The circular Gaussian distribution is obtained by setting the shape parameter to $p = 2$, while smaller values of the shape parameter p result in more sparse priors, i.e., priors with a higher peak at zero and heavier tails. This is illustrated in the plot of the log-prior $\log \mathrm{p}(z)$ in Fig. 3.2. Since the CGG prior can be written in the form (3.11) with $f(.)$ given as

$$f(|z|) = \frac{|t|^p}{\zeta^{p/2}} - \log \frac{p}{2\pi\zeta\Gamma(2/p)}, \tag{3.19}$$

it can be represented using a convex representation in the form (3.12). Using (3.17) and (3.19), the optimal value of $\lambda(n)$ in the i-th iteration for a CGG prior for the desired signal can be written

$$\hat{\lambda}^i(n) = \frac{2\zeta^{p/2}}{p} \left|\hat{d}_{\text{ref}}^{i-1}(n)\right|^{2-p}. \tag{3.20}$$

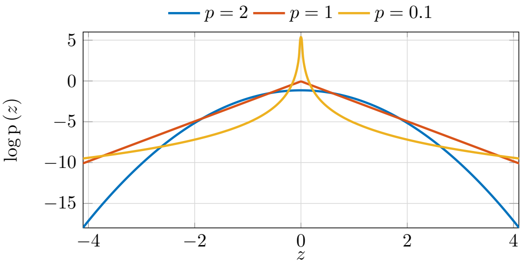


Fig. 3.2: Logarithm of the CGG prior p(.) in (3.18) for different values of the shape parameter p with the scale parameter ζ selected such that the variance is 1. Note that the plot only shows values on the real axis (i.e., the imaginary part of z is 0), and the prior is circular.

As can be seen in (3.20), the optimal $\hat{\lambda}^i(n)$ depends on the shape and scaling parameters of the CGG prior in (3.18). However, since the estimated prediction filter $\hat{\mathbf{g}}_{\text{ref}}$ computed using (3.9), and hence also the estimated desired speech signal $\hat{\mathbf{d}}_{\text{ref}}$ computed using (3.2), is invariant to a scaling of the variances $\boldsymbol{\lambda}$, the variance estimate in (3.20) can be simplified to

$$\hat{\lambda}^i(n) = \left|\hat{d}^{i-1}_{\text{ref}}(n)\right|^{2-p}, \tag{3.21}$$

which depends only on the shape parameter $p \in (0, 2]$ of the CGG prior. In practice, a small positive constant $\varepsilon_{\min}$ is added to the estimated variances to prevent division by zero, i.e.,

$$\hat{\boldsymbol{\lambda}}^i = \left(\left|\hat{\mathbf{d}}^{i-1}_{\text{ref}}\right|^2 + \varepsilon_{\min}\right)^{1-\frac{p}{2}}, \tag{3.22}$$

Finally, the obtained iterative optimization procedure is summarized in Alg. 1.

3.3.4 *Relation to the conventional method*

It can be observed that the variance update in (3.7) for the conventional method corresponds to setting the shape parameter $p = 0$ in the variance update for the proposed method in (3.21). By comparing the optimization problem in (3.5) with the proposed optimization problem in (3.15), it can be seen that the conventional WPE method is obtained by setting the scaling function $\psi(.)$ to a constant value in the proposed method. Hence, the prior for the desired signal in the conventional method, obtained in the proposed framework by setting the scaling function $\psi(.)$ in (3.12) to 1, is given by

$$\mathrm{p}\left(d_{\text{ref}}(n)\right) = \max_{\lambda(n)>0} \frac{e^{-\frac{|d_{\text{ref}}(n)|^2}{\lambda(n)}}}{\pi\lambda(n)} = \frac{e^{-1}}{\pi\left|d_{\text{ref}}(n)\right|^2} \propto \frac{1}{\left|d_{\text{ref}}(n)\right|^2} \tag{3.23}$$

Alg. 1 Iterative optimization algorithm for sparse MCLP with a CGG prior.

```
parameters: filter length $L_g$ and prediction delay $\tau$ in (3.1), shape parameter
    $p$ in (3.18), regularization parameter $\varepsilon_{\min}$, maximum number of iterations $I$,
    tolerance $\eta$
input: $M$-channel reverberant microphone signal coefficients $\mathbf{X}(k)$, $\forall k$
for each $k$ do
    $i \leftarrow 0$
    set $\hat{\mathbf{d}}_{\text{ref}}^{0}$                                          ▷ initialization
    repeat
        $i \leftarrow i + 1$
        $\hat{\boldsymbol{\Lambda}}^{i} \leftarrow \left( \left| \hat{\mathbf{d}}_{\text{ref}}^{i-1} \right|^{2} + \varepsilon_{\min} \right)^{1-\frac{p}{2}}$          ▷ (3.22)
        $\hat{\mathbf{g}}_{\text{ref}}^{i} \leftarrow \left( \tilde{\mathbf{X}}_{\tau}^{\mathsf{H}} \left( \hat{\boldsymbol{\Lambda}}^{i} \right)^{-1} \tilde{\mathbf{X}}_{\tau} \right)^{-1} \tilde{\mathbf{X}}_{\tau}^{\mathsf{H}} \left( \hat{\boldsymbol{\Lambda}}^{i} \right)^{-1} \mathbf{x}_{\text{ref}}$     ▷ (3.9)
        $\hat{\mathbf{d}}_{\text{ref}}^{i} \leftarrow \mathbf{x}_{\text{ref}} - \tilde{\mathbf{X}}_{\tau} \hat{\mathbf{g}}_{\text{ref}}^{i}$          ▷ (3.2)
    until $i = I$ or $\frac{\| \hat{\mathbf{d}}_{\text{ref}}^{i} - \hat{\mathbf{d}}_{\text{ref}}^{i-1} \|}{\| \hat{\mathbf{d}}_{\text{ref}}^{i-1} \|} < \eta$
end for
output: estimated desired signal coefficients $\hat{\mathbf{d}}_{\text{ref}}(k) = \hat{\mathbf{d}}_{\text{ref}}^{i}(k)$, $\forall k$
```

since the maximum is attained when $\lambda(n) = |d_{\text{ref}}(n)|^2$. The obtained prior can also be represented in the form (3.11) with

$$f(t) = \log t^2 + \text{const.} \tag{3.24}$$

Since the prior in (3.23) is not integrable, it is not a probability density and therefore it is an improper prior. More importantly, the prior in (3.23) used in the conventional method hence strongly favors values of the desired signal that are close to the origin, i.e., it is a strong sparse prior for the desired signal. This type of sparsity-promoting prior, resulting in a logarithmic penalty as in (3.24), has been previously used in various signal processing applications [221–223, 228]. This is an interesting relation, since the conventional WPE method was originally derived with the TVG model as the starting point, without explicitly enforcing sparsity on the estimated desired speech signal. However, this interpretation highlights the underlying role of the sparse prior (3.23) in estimating the desired speech signal, which turns out to be much more effective than minimizing the output energy, corresponding to a Gaussian assumption on the desired signal [154]. Although the flexibility of the TVG model to accurately represent any time-varying signal has been marked as its main advantage over the time-invariant Gaussian model, the actual reason for the success of the TVG model in MCLP-based dereverberation is its sparsity-promoting behavior when implemented in an ML estimation procedure.

3.4 Reformulation as ℓ_p-norm minimization

In this section, we aim to provide a better understanding of the cost functions underlying the proposed methods, and relating them to the general problem of sparse recovery. More specifically, we reformulate the conventional WPE and the proposed CGG-based methods for estimating the prediction filter $\hat{\mathbf{g}}_{\text{ref}}$ in terms of an ℓ_p-norm minimization problem. For a general prior p(.) and independent coefficients $d_{\text{ref}}(n)$ across time, the likelihood function of $\mathbf{g}_{\text{ref}}$ can be written as

$$\mathcal{L}\left(\mathbf{g}_{\text{ref}}\right) = \prod_{n=1}^{N} \mathrm{p}\left(d_{\text{ref}}(n)\right). \tag{3.25}$$

Given a sparse prior p(.) in the form (3.11), the ML estimate of the prediction filter $\hat{\mathbf{g}}_{\text{ref}}$ can hence be obtained by minimizing the negative log-likelihood, i.e.,

$$\hat{\mathbf{g}}_{\text{ref}} = \arg\min_{\mathbf{g}_{\text{ref}}} \sum_{n=1}^{N} f\left(\left|d_{\text{ref}}(n)\right|\right). \tag{3.26}$$

When p(.) is a CGG prior as in (3.18), this ML estimate can be obtained using (3.19) as a solution of the following problem

$$\begin{aligned} \min_{\mathbf{g}_{\text{ref}}} \quad & \left\|\mathbf{d}_{\text{ref}}\right\|_p^p \\ \text{subject to} \quad & \mathbf{d}_{\text{ref}} + \tilde{\mathbf{X}}_{\tau}\mathbf{g}_{\text{ref}} = \mathbf{x}_{\text{ref}}, \end{aligned} \tag{3.27}$$

where $\|.\|_p$ is the ℓ_p-norm[2] defined as

$$\|\mathbf{d}\|_p = \left(\sum_{n=1}^{N} |d(n)|^p\right)^{1/p} \tag{3.28}$$

The optimization problem in (3.27) is a deterministic formulation of sparse MCLP with an ℓ_p-norm cost function. For the conventional method with the prior p(.) given in (3.23), the ML estimate of the prediction filter is obtained using (3.24) by solving the following optimization problem

$$\begin{aligned} \min_{\mathbf{g}_{\text{ref}}} \quad & \sum_{n=1}^{N} \log\left|d_{\text{ref}}(n)\right| \\ \text{subject to} \quad & \mathbf{d}_{\text{ref}} + \tilde{\mathbf{X}}_{\tau}\mathbf{g}_{\text{ref}} = \mathbf{x}_{\text{ref}}. \end{aligned} \tag{3.29}$$

---

[2] Note that for $p < 1$ the ℓ_p-norm is non-convex and it is actually not a norm, e.g., it does not satisfy the triangle inequality.

This logarithmic cost function is often used in signal processing problems as an approximation of the ℓ_0-norm, counting the number of non-zero entries in a vector [223,228,229]. The ℓ_0-norm is related to the previously defined ℓ_p-norm through

$$\|\mathbf{d}\|_0 = \lim_{p \to 0} \sum_{n=1}^{N} |d(n)|^p \,, \tag{3.30}$$

and the logarithmic penalty is related to the ℓ_0-norm through [228]

$$\lim_{p \to 0} \frac{1}{p} \sum_{n=1}^{N} \left(|d(n)|^p - 1\right) = \sum_{n=1}^{N} \log |d(n)| \,. \tag{3.31}$$

Moreover, the set of local minima of the optimization problem in (3.29) corresponds to the set of local minima of the optimization problem [228]

$$\begin{aligned} \min_{\mathbf{g}_{\text{ref}}} \quad & \|\mathbf{d}_{\text{ref}}\|_0 \\ \text{subject to} \quad & \mathbf{d}_{\text{ref}} + \tilde{\mathbf{X}}_\tau \mathbf{g}_{\text{ref}} = \mathbf{x}_{\text{ref}} . \end{aligned} \tag{3.32}$$

Furthermore, by expressing the desired speech signal using (3.2) as

$$\mathbf{d}_{\text{ref}} = \boldsymbol{\Omega}\mathbf{u}, \tag{3.33}$$

with

$$\boldsymbol{\Omega} = \left[\mathbf{x}_{\text{ref}}, -\tilde{\mathbf{X}}_\tau\right], \quad \mathbf{u} = \left[1, \mathbf{g}_{\text{ref}}^{\mathsf{T}}\right]^{\mathsf{H}}, \tag{3.34}$$

where $\mathbf{u}$ is equivalent to the prediction filter $\mathbf{g}_{\text{ref}}$, the optimization problem (3.27) can be rewritten directly in terms of the prediction filter $\mathbf{u}$ as

$$\min_{\mathbf{u}} \|\boldsymbol{\Omega}\mathbf{u}\|_p^p \quad \text{subject to} \quad \mathbf{e}_1^{\mathsf{T}}\mathbf{u} = 1, \tag{3.35}$$

where $\mathbf{e}_1 = [1, 0, \ldots, 0]^{\mathsf{T}}$. Optimization problems in this form are addressed in the context of the cosparse analysis problem [230–232]. In that context, the matrix $\boldsymbol{\Omega}$ is the analysis matrix that transforms the unknown variable (i.e., the prediction filter $\mathbf{u}$) to the domain where sparsity is enforced (i.e., the desired signal coefficients $\mathbf{d}_{\text{ref}}$). By solving the problem in (3.35), an estimate of the prediction filter $\hat{\mathbf{u}}$ is computed that results in a sparse prediction error, i.e., the desired speech signal $\hat{\mathbf{d}}_{\text{ref}}$, with sparsity quantified by means of the ℓ_p-norm. A similar optimization problem was also considered in the context of sparse linear prediction in the time domain [207], applied for modeling and coding of speech signals.

The analytically derived sparsity-promoting cost function can be easily interpreted in the context of speech dereverberation. As demonstrated in Section 2.2, reverberation makes the observed microphone signal less sparse than the clean speech signal in the STFT domain. Therefore, on the one hand it is reasonable to enforce an estimate of the desired speech signal whose STFT coefficients are sparser than the STFT coefficients of the reverberant recording. On the other hand, the direct path and early reflections should be preserved in the estimated desired speech signal, which is enforced by the MCLP signal model with the prediction delay in (3.1), resulting in the optimization problem in (3.27).

In summary, both the conventional method as well as the proposed CGG-based method can be interpreted as iterative optimization procedures that aim to solve the sparse MCLP optimization problem in (3.27)/(3.35), corresponding to the proposed CGG-based method for values of the shape parameter $p \in (0, 2]$ and to the conventional method when $p = 0$.

3.4.1 *Iteratively reweighted LS for ℓ_p-norm minimization*

It should be noted that the optimization problem in (3.35) is non-convex for $p < 1$ and that iterative optimization procedures can in general converge only to a local minimum. However, even if reaching the global minimum cannot be guaranteed, employing a non-convex cost function can often result in a sparser estimated signal then when employing a convex cost function (e.g, for $p \geq 1$) [223]. Several optimization procedures for ℓ_p-norm minimization have been proposed in the literature. Typically, the original non-convex problem is transformed into a series of appropriate convex problems which are easy to solve. Here, we employ the iteratively reweighted LS (IRLS) algorithm for ℓ_p-norm minimization [223,228], and show that the obtained method is in some cases equivalent to the conventional method and the proposed method based on a CGG prior. More details about iteratively reweighted procedures for non-convex minimization are given in Appendix B.1.

The basic idea in IRLS is to replace the ℓ_p-norm minimization problem in (3.35) with a series of squared ℓ_2-norm minimization subproblems [223,231]. Each ℓ_2-norm minimization subproblem can be easily solved, and the solution obtained in the current iteration is used to modify the subproblem in the next iteration. More specifically, the ℓ_p-norm cost function in (3.35) is replaced by a weighted ℓ_2-norm cost function in the i-th iteration as [223]

$$\hat{\mathbf{u}}^i = \arg\min_{\mathbf{u}} \mathbf{u}^{\mathsf{H}} \boldsymbol{\Omega}^{\mathsf{H}} \hat{\mathbf{W}}^i \boldsymbol{\Omega} \mathbf{u} \quad \text{subject to} \quad \mathbf{e}_1^{\mathsf{T}} \mathbf{u} = 1, \tag{3.36}$$

where $\hat{\mathbf{W}}^i = \operatorname{diag}\left(\hat{\mathbf{w}}^i\right)$ is a real-valued diagonal weighting matrix with the weight vector

$$\hat{\mathbf{w}}^i = \left[\hat{w}^i(1), \ldots, \hat{w}^i(N)\right]^{\mathsf{T}} \tag{3.37}$$

on the diagonal. The LS optimization problem in (3.36) has a closed-form solution for the prediction filter as

$$\hat{\mathbf{u}}^i = \left(\mathbf{e}_1^{\mathsf{T}} \left(\boldsymbol{\Omega}^{\mathsf{H}} \hat{\mathbf{W}}^i \boldsymbol{\Omega}\right)^{-1} \mathbf{e}_1\right)^{-1} \left(\boldsymbol{\Omega}^{\mathsf{H}} \hat{\mathbf{W}}^i \boldsymbol{\Omega}\right)^{-1} \mathbf{e}_1, \tag{3.38}$$

which is equivalent to estimating the prediction filter $\hat{\mathbf{g}}^i_{\text{ref}}$ in (3.9). The estimate of the desired signal in the i-th iteration is given using (3.33) as $\hat{\mathbf{d}}^i_{\text{ref}} = \boldsymbol{\Omega} \hat{\mathbf{u}}^i$, which is equivalent to the estimating the desired speech signal using (3.2).

The weights $\hat{w}(n)$ are updated in each iteration as

$$\hat{w}^i(n) = \left|\hat{d}_{\text{ref}}^{i-1}(n)\right|^{p-2}, \tag{3.39}$$

such that the convex cost function in (3.36) is a first-order approximation of the non-convex cost function in (3.35) (cf. Appendix B.1). By comparing the obtained update for the weights in (3.39) with the variance update in (3.21), it can be seen that the weights are equal to the inverse of the variances. The updates (3.38) and (3.39) result in an iterative procedure for minimizing (3.35). Intuitively, a large weight $\hat{w}(n)$ promotes the desired signal coefficient at the n-th frame to have a relatively small energy, corresponding to the sparsity-promoting behavior of the ℓ_p-norm. To avoid division by zero in (3.39), the optimization problem is typically regularized by adding a small positive value to the weights [223, 231], i.e.,

$$\hat{w}^i(n) = \left(\left|\hat{d}_{\text{ref}}^{i-1}(n)\right|^2 + \varepsilon^i\right)^{\frac{p}{2}-1}, \tag{3.40}$$

where the regularization parameter ε^i can in general be iteration dependent. When the role of the regularization parameter is just to avoid division by zero, the procedure is called unregularized IRLS [223], and it is equivalent to the iterative algorithm for the proposed sparse MCLP with CGG prior in Alg. 1. Setting the regularization parameter to a larger value can be used to make the linear system in (3.38) better conditioned. In practice, a common regularization strategy where the regularization parameter is initialized with a large value and then gradually decreased has been shown to be effective in avoiding local minima for $p < 1$ [223]. In this case the procedure is called regularized IRLS, and this regularization strategy can be related to Bayesian methods, with the regularization parameter having a similar role as the posterior variance of the corresponding coefficient [221, 228, 233]. A number of different strategies for updating the regularization parameter in iteratively reweighted algorithms have been investigated in [228].

The outline of the complete dereverberation method using the regularized IRLS algorithm (rIRLS-p) for the optimization problem in (3.35) in each subband k is given in Alg. 2. For each subband k the matrix $\mathbf{\Omega}$ is first normalized with the maximum magnitude of the STFT coefficients of the reference microphone signal $\mathbf{x}_{\text{ref}}$. In this way, the values of the regularization parameter can be set independently of the magnitudes of the coefficients in the given subband. The rIRLS-p algorithm for minimizing (3.35) is implemented similarly as in [223]. The updates (3.38) and (3.40) are iterated until the relative change of the ℓ_2-norm of the output is smaller than a tolerance or maximal number of iterations is exceeded. In that case, the regularization parameter ε^i is reduced by a factor 10, and the tolerance parameter is updated to $\sqrt{\varepsilon^i}/100$. Since $p < 1$ results in a non-convex problem in (3.35), initialization of the algorithm may influence the final estimate. More details on the initialization are given in Section 3.5.1.

Alg. 2 Regularized IRLS algorithm for sparse MCLP with ℓ_p-norm cost function (rIRLS-p). The parameter ε is initialized with a relatively large value and gradually reduced. $\|\mathbf{x}\|_\infty$ denotes the maximum absolute value of the elements in $\mathbf{x}$.

```
parameters: Filter length $L_g$ and prediction delay $\tau$ in (3.1), shape parameter $p$
    in (3.40), regularization parameters $\varepsilon_{\text{init}}, \varepsilon_{\text{min}}$, maximum number of reweighting
    iterations $I$
input: $M$-channel reverberant microphone signal coefficients $\mathbf{X}(k)$, $\forall k$
for each $k$ do
    $i \leftarrow 0$
    $\boldsymbol{\Omega} \leftarrow$ construct using (3.34), normalize $\boldsymbol{\Omega} \leftarrow \boldsymbol{\Omega}/\kappa$, with $\kappa = \|\mathbf{x}_{\text{ref}}\|_\infty$
    set $\hat{\mathbf{d}}^0_{\text{ref}}$, normalize $\hat{\mathbf{d}}^0_{\text{ref}} \leftarrow \hat{\mathbf{d}}^0_{\text{ref}}/\kappa$    ▷ initialization
    $\varepsilon^1 \leftarrow \varepsilon_{\text{init}}$
    repeat
        $i \leftarrow i + 1$
        $\hat{\mathbf{w}}^i \leftarrow \left(\left|\hat{\mathbf{d}}^{i-1}_{\text{ref}}\right|^2 + \varepsilon^i\right)^{\frac{p}{2}-1}$    ▷ (3.40)
        $\hat{\mathbf{u}}^i \leftarrow \left(\mathbf{e}_1^{\mathsf{T}} \left(\boldsymbol{\Omega}^{\mathsf{H}} \hat{\mathbf{W}}^i \boldsymbol{\Omega}\right)^{-1} \mathbf{e}_1\right)^{-1} \left(\boldsymbol{\Omega}^{\mathsf{H}} \hat{\mathbf{W}}^i \boldsymbol{\Omega}\right)^{-1} \mathbf{e}_1$    ▷ (3.38)
        $\hat{\mathbf{d}}^i_{\text{ref}} \leftarrow \boldsymbol{\Omega} \hat{\mathbf{u}}^i$    ▷ (3.33)
        if $\frac{\|\hat{\mathbf{d}}^i_{\text{ref}} - \hat{\mathbf{d}}^{i-1}_{\text{ref}}\|}{\|\hat{\mathbf{d}}^{i-1}_{\text{ref}}\|} < \frac{\sqrt{\varepsilon^i}}{100}$ then
            $\varepsilon^{i+1} \leftarrow \varepsilon^i/10$
        else
            $\varepsilon^{i+1} \leftarrow \varepsilon^i$
        end if
    until $i = I$ or $\varepsilon^i < \varepsilon_{\text{min}}$
    $\hat{\mathbf{d}}_{\text{ref}} \leftarrow \kappa \hat{\mathbf{d}}^i_{\text{ref}}$
end for
output: estimated desired signal coefficients $\hat{\mathbf{d}}_{\text{ref}}(k)$, $\forall k$
```

3.5 Simulations

In this section, the performance of the blind speech dereverberation methods based on sparse MCLP presented in Sections 3.3 and 3.4 is investigated. More specifically, we consider unregularized IRLS for sparse MCLP with CGG prior (IRLS-p) as presented in Alg. 1, with the conventional WPE method being a special case for $p = 0$ (cf. Section 3.2), and regularized IRLS for sparse MCLP (rIRLS-p) as presented in Alg. 2.

The considered acoustic scenario and the implementation details are outlined in Section 3.5.1. The influence of the initialization and the iteration-wise performance are investigated in Section 3.5.2. The influence of the filter length and the number of microphones is investigated in Section 3.5.3. The influence of the filter length in different acoustic scenarios is investigated in Section 3.5.4.

3.5.1 *Acoustic scenario and algorithmic setup*

We consider several acoustic scenarios from the REVERB challenge [22, 23] with a single speech source and omni-directional microphones placed at a distance of about 2 m from the source. The microphones are positioned on a circle with a radius of 10 cm, with a 45° angle between adjacent microphones and approximately 7.6 cm distance between adjacent microphones. In Sections 3.5.2 and 3.5.4 a scenario with $M = 2$ microphones is considered, while in Section 3.5.3 the number of microphones is set to $M \in \{1, 2, 4\}$. Measured RIRs from the REVERB challenge have been used [22,23], where the reverberation time is $T_{60} \approx \{250, 600, 700\}$ ms and the direct-to-reverberant ratio is DRR $\approx \{7, -2.4, 1.4\}$ dB. The RIRs have been measured using a maximum length sequence at a sampling frequency $f_s = 16$ kHz. The reverberant signals have been generated by convolving 10 speech samples (5 male and 5 female speakers) from the TIMIT database [234] with an average length of approximately 5.2 s with the measured RIRs.

The analysis and synthesis STFT is computed using a tight window based on a 64 ms Hamming window with a 16 ms window shift [235]. The prediction delay in (3.1) is set to $\tau = 2$ in all experiments. The IRLS-p method is implemented as in Alg. 1, with the conventional WPE corresponding to the case with $p = 0$. The variance estimate is regularized with the parameter $\varepsilon_{\text{min}} = 10^{-8}$ for all subbands k. The rIRLS-p method for minimizing (3.35) is implemented as in Alg. 2 with the initial value for the regularization parameter $\varepsilon_{\text{init}} = 0.1$ and the minimum value of the regularization parameter $\varepsilon_{\text{min}} = 10^{-8}$. The matrix $\mathbf{\Omega}$ is normalized with the maximum magnitude of the STFT coefficients of the reference microphone signal, ensuring both that the regularization parameter $\varepsilon_{\text{init}}$ is relatively large compared to the significant coefficients, and ε_{min} is always much smaller than the significant coefficients. Unless stated otherwise, the tolerance for the relative change of the ℓ_2-norm of the estimated desired signal is set to $\eta = 10^{-6}$ for the IRLS-p method. The same final tolerance applies for the rIRLS-p method since $\varepsilon_{\text{min}} = 10^{-8}$ corresponds to $\eta_{\text{min}} = 10^{-6}$ (cf. Alg. 2).

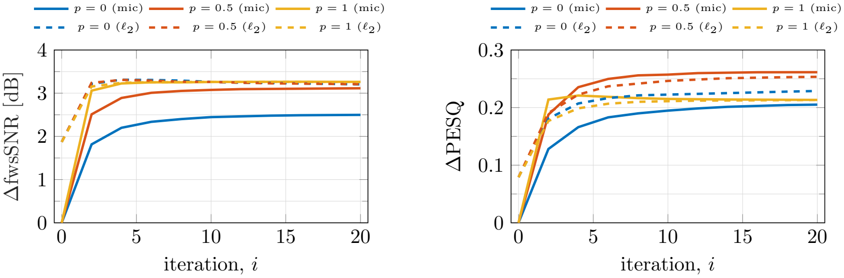
The dereverberation performance is evaluated in terms of the instrumental performance measures described in Section 2.3. The reverberation reduction performance is evaluated using the improvement in fwsSNR (ΔfwsSNR) [22, 216] between the processed output signal and the reverberant input signal. The perceptual speech quality is evaluated using the improvement in PESQ (ΔPESQ) [210, 211] between the processed output signal and the reverberant input signal. The reference signal used for the instrumental measures is the direct speech signal on the reference microphone, obtained by convolving the anechoic speech signal with the direct component of the corresponding RIR. The reported improvements of the instrumental measures are obtained by averaging over all 10 speech samples.

3.5.2 *Influence of initialization and iteration-wise performance*

In this section, we investigate the influence of the initialization and the number of iterations on the dereverberation performance of the considered methods for the scenario with $T_{60} \approx 700$ ms. Since the problem in (3.35) is non-convex for $p < 1$, the proposed methods may only converge to a local minimum, such that the final estimate of the desired speech signal typically depends on the initialization $\hat{\mathbf{d}}^0_{\text{ref}}$. On the one hand, in sparse recovery the reweighting methods are typically initialized using the least-squares solution $\hat{\mathbf{d}}_{\text{ref},\ell_2}$, i.e., by solving the optimization problem (3.35) for $p = 2$. On the other hand, iterative dereverberation methods are typically initialized using the reverberant microphone signal $\mathbf{x}_{\text{ref}}$. In this experiment, we evaluate the iteration-wise performance of IRLS-p and rIRLS-p for both initializations, i.e., $\hat{\mathbf{d}}^0_{\text{ref}} = \hat{\mathbf{d}}_{\text{ref},\ell_2}$ and $\hat{\mathbf{d}}^0_{\text{ref}} = \mathbf{x}_{\text{ref}}$. In all experiments the number of microphones is fixed to $M = 2$, the filter length is set to $L_g \in \{5, 20, 40\}$, i.e., a relatively short filter length ($L_g = 5$), a typically used filter length ($L_g = 20$), and a relatively large filter length ($L_g = 40$). The shape parameter is set to $p \in \{0, 0.5, 1\}$. Note that for $p = 1$ the optimization problem is convex and the initialization should not have an influence on the estimated value at convergence, although may affect the speed of convergence. The maximum number of reweighting iterations for IRLS-p is set to $I = 20$. The maximum number of reweighting iterations for rIRLS-p is set to $I = 200$, since it typically requires a larger number of iterations due to the reduction update for the regularization parameter. To prevent early termination of the algorithms, only the maximum number of iterations I is used as the stopping criterion in Alg. 1 and Alg. 2 for the simulations in this section.

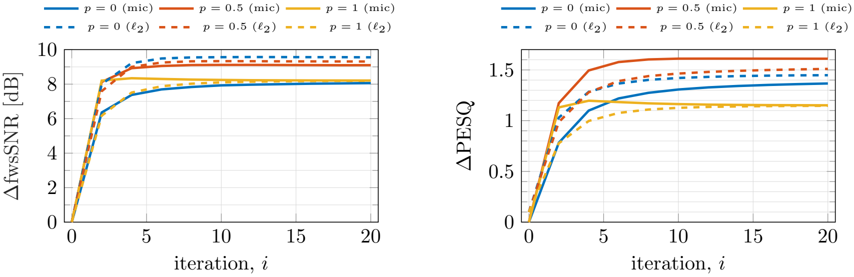
Improvements in terms of the considered performance measures obtained using the IRLS-p method are depicted in Figs. 3.3–3.5. Firstly, we consider the results for a relatively short filter length $L_g = 5$, depicted in Fig. 3.3. It can be observed that all methods converge after $I = 20$ iterations for all values of the shape parameter p and the type of initialization, and result in improvements in terms of ΔfwsSNR and ΔPESQ. Moreover, the largest relative improvements of the instrumental measures are obtained in the first few iterations. It can also be observed that the initialization has an influence on the performance, by comparing the obtained measures for $i = 0$, with the ℓ_2 initialization performing better than the initialization with the microphone signal. However, the performance difference between different initializations is typically reduced already after a single iteration (i.e., at $i = 1$). Furthermore, using $p = 0.5$ and $p = 1$ achieves the same performance at convergence independent of the initialization, while $p = 0$ performs somewhat worse when initialized with the microphone signal. Overall, the performance of the ℓ_2-initialized methods for different shape parameters p is virtually identical in terms of ΔfwsSNR, while $p = 0.5$ performs slightly better than the others in terms of ΔPESQ. The performance of the microphone-initialized methods with $p = 0.5$ and $p = 1$ is virtually identical in terms of ΔfwsSNR, while $p = 0.5$ performs slightly better than the others in terms of ΔPESQ.

Secondly, we consider the results for a typical filter length $L_g = 20$, depicted in Fig. 3.4. Again, it can be observed that all methods converge after $I = 20$ iterations for all values of the shape parameter p and types of initialization, and result in considerably larger improvements in terms of ΔfwsSNR and ΔPESQ than for $L_g = 5$. The largest improvements of the instrumental measures are obtained within a few iterations, and the initialization does not have a large influence on the initial performance. On the one hand, at convergence, initialization does not affect the



(a) ΔfwsSNR (b) ΔPESQ

Fig. 3.3: Performance of the IRLS-p method with $L_g = 5$ for different initializations and number of iterations in terms of ΔfwsSNR (left) and ΔPESQ (right). Initialization is denoted with the label in the parentheses.

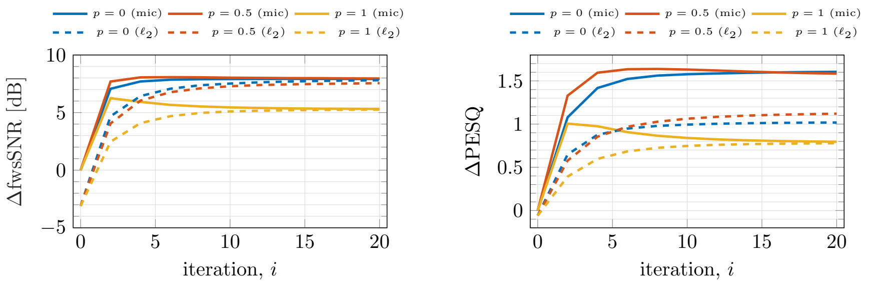


(a) ΔfwsSNR (b) ΔPESQ

Fig. 3.4: Performance of the IRLS-p method with $L_g = 20$ for different initializations and number of iterations in terms of ΔfwsSNR (left) and ΔPESQ (right). Initialization is denoted with the label in the parentheses.

methods with $p = 0.5$ and $p = 1$ in terms of ΔfwsSNR whereas ℓ_2 initialization results in a small deterioration of the performance for $p = 0.5$ in terms of ΔPESQ. On the other hand, for $p = 0$ at convergence the ℓ_2 initialization performs better in terms of ΔfwsSNR and slightly worse in terms of ΔPESQ. Overall, $p = 0$ with ℓ_2 initialization and $p = 0.5$ with both initializations achieve the best performance in terms of ΔfwsSNR, while $p = 0.5$ with microphone initialization achieves the best performance in terms of ΔPESQ.

Thirdly, we consider the results for a relatively long filter length $L_g = 40$, depicted in Fig. 3.5. As before, it can be observed that all methods converge after $I = 20$ iterations for all values of the shape parameter p and the type of initialization, and result in similar best-case improvements in terms of ΔfwsSNR and ΔPESQ as for $L_g = 20$. It can be observed that initialization has a significant influence on the initial performance in terms of ΔfwsSNR, even possibly resulting in deterioration with respect to the microphone signal for the ℓ_2 initialization, and the performance



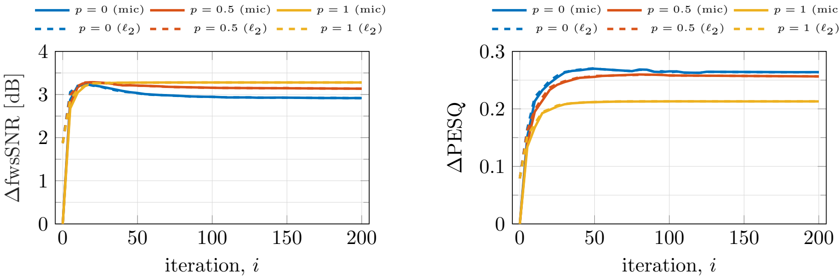
(a) ΔfwsSNR (b) ΔPESQ

Fig. 3.5: Performance of the IRLS-p method with $L_g = 40$ for different initializations and number of iterations in terms of ΔfwsSNR (left) and ΔPESQ (right). Initialization is denoted with the label in the parentheses.

at convergence in terms of ΔPESQ. Overall, $p = 0$ and $p = 0.5$ with microphone initialization achieve the best performance in terms of both ΔfwsSNR and ΔPESQ, with the ℓ_2-initialized methods performing worse, especially in terms of ΔPESQ.

In conclusion, the results in Figs. 3.3–3.5 indicate that the (unregularized) IRLS-p method can be substantially influenced by the used initialization. In many cases, initialization with the coefficients of the microphone signal is preferred to ℓ_2 initialization, especially for long filters (cf. Fig. 3.5). It should be noted that the ℓ_2-norm solution (i.e., the LS solution) is in general not very effective for dereverberation, since ℓ_2-norm minimization results in a minimum-energy estimate of the desired speech signal with typically many small but non-zero coefficients. This effect can especially be observed for $i = 0$ in Fig. 3.5, where the ℓ_2-norm solution results in deterioration of the speech quality when compared to the microphone signal. Therefore, in the following we will initialize the IRLS-p method with the coefficients of the reference microphone signal. In addition, the values of the shape parameter $p < 1$ typically perform better than $p = 1$ due to a stronger enforcement of sparsity of the desired signal coefficients (with the exception of relatively small differences in case of short filters, cf. Fig. 3.3a). Overall, $p = 0.5$ performs better than $p = 0$, as could be expected for the unregularized IRLS, since $p = 0.5$ results in a somewhat less aggressive cost function than $p = 0$, which can help to avoid local minima.

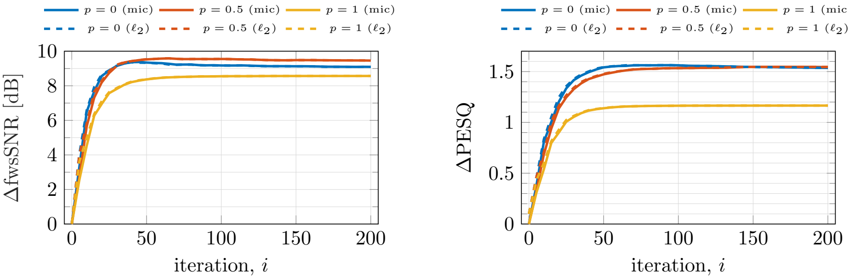
Improvements in terms of the considered performance measures obtained using the rIRLS-p method are depicted in Figs. 3.6–3.8. Firstly, we consider the results for a relatively short filter length $L_g = 5$, depicted in Fig. 3.6. It can be observed that all methods converge after about 100 iterations for all values of the shape parameter p, and result in best-case improvements in terms of ΔfwsSNR and ΔPESQ comparable to the one of the unregularized IRLS-p (cf. Fig. 3.3). Contrary to the unregularized case, initialization has virtually no influence on the performance of rIRLS-p. This can be attributed to the regularization strategy, which helps to avoid the local minima [223], but also requires a much larger number of iterations. The method with $p = 1$ shows a slightly better performance in terms of ΔfwsSNR, while $p = 0$ and $p = 0.5$ perform better in terms of ΔPESQ.



(a) ΔfwsSNR

(b) ΔPESQ

Fig. 3.6: Performance of the rIRLS-p method with $L_g = 5$ for different initializations and number of iterations in terms of ΔfwsSNR (left) and ΔPESQ (right). Initialization is denoted with the label in parentheses.



(a) ΔfwsSNR

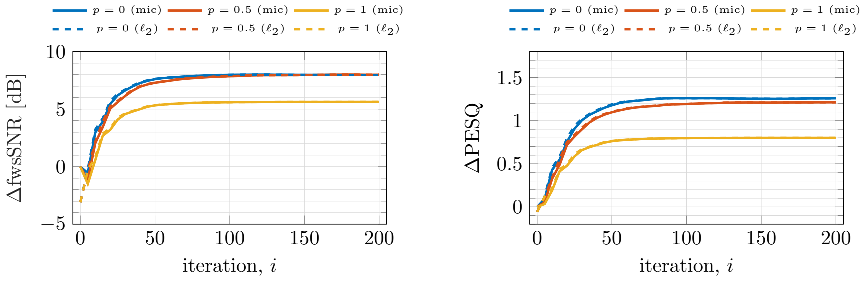
(b) ΔPESQ

Fig. 3.7: Performance of the rIRLS-p method with $L_g = 20$ for different initializations and number of iterations in terms of ΔfwsSNR (left) and ΔPESQ (right). Initialization is denoted with the label in parentheses.

Secondly, we consider the results for $L_g = 20$, depicted in Fig. 3.7. Again, it can be observed that all methods converge after about 100 iterations for all values of the shape parameter p, and result in considerably larger improvements in terms of ΔfwsSNR and ΔPESQ than for $L_g = 5$. Also, rIRLS achieves a similar best-case performance as the unregularized IRLS (cf. Fig. 3.4), independently of the used initialization. In terms of both ΔfwsSNR and ΔPESQ, $p = 0$ and $p = 0.5$ perform better than $p = 1$.

Thirdly, we consider the results for $L_g = 40$, depicted in Fig. 3.8. As before, it can be observed that all methods converge after about 100 iterations for all values of the shape parameter p, and result in lower improvements in terms of ΔfwsSNR and ΔPESQ than $L_g = 20$ and compared to unregularized IRLS (cf. Fig. 3.4).

In conclusion, the results in Figs. 3.6–3.8 indicate that the (regularized) rIRLS-p leads to a similar performance independent of initialization, as opposed to the (unregularized) IRLS-p method. As expected, rIRLS is less influenced by the shape



(a) ΔfwsSNR (b) ΔPESQ

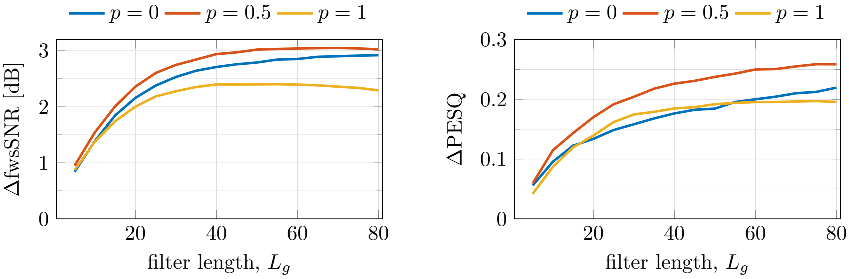
Fig. 3.8: Performance of the rIRLS-p method with $L_g = 40$ for different initializations and number of iterations in terms of ΔfwsSNR (left) and ΔPESQ (right). Initialization is denoted with the label in parentheses.

parameter $p < 1$ due to the regularization strategy, with $p < 1$ performing better than $p = 1$ due to a stronger enforcement of sparsity of the desired signal coefficients (with the exception of relatively small differences in case of short filters, cf. Fig. 3.6a). However, the consistency of rIRLS comes at the price of a much larger number of iterations required to achieve convergence. Furthermore, in the case of relatively long filters it achieves a somewhat lower performance than the IRLS-p method (cf. Figs. 3.5b and 3.8b). Therefore, in time-constrained practical applications the IRLS-p method with a suitably selected shape parameter is preferred over the rIRLS-p method. Hence, in the remainder of this chapter we will consider only the IRLS-p method.

3.5.3 *Influence of filter length L_g and number of microphones M*

In this section, we investigate the influence of the filter length L_g and the number of microphones M on the dereverberation performance of the considered IRLS-p method for the scenario with $T_{60} \approx 700$ ms. The number of microphones is set $M \in \{1, 2, 4\}$, the shape parameter is set to $p \in \{0, 0.5, 1\}$, and the filter length L_g is varied to cover the total number of coefficients (ML_g) between 5 and 80. In all experiments the optimization is initialized using the coefficients of the reference microphone signal (cf. Section 3.5.2). The maximum number of iterations is set to $I = 20$, and the stopping criterion for Alg. 1 is set as described in Section 3.5.1.

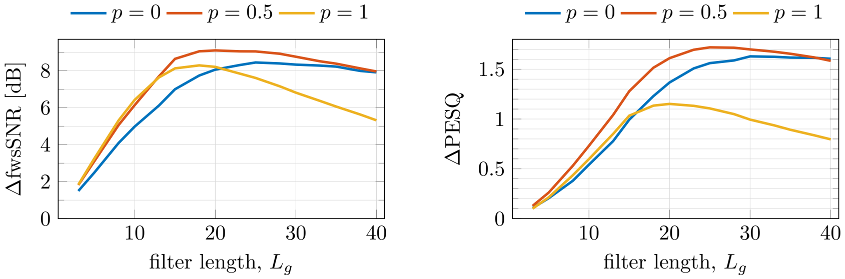
Firstly, we consider the results for the setup with $M = 1$ microphone, depicted in Fig. 3.9. It can be observed that the IRLS-p method results in improvements in terms of ΔfwsSNR and ΔPESQ for all considered shape parameter values. This shows that sparse linear prediction can also be used for single-channel dereverberation, although the MCLP-based signal model does not hold for $M = 1$. In general, increasing the filter length L_g improves the dereverberation performance, although the improvements for filter lengths larger than $L_g = 50$ seem to be marginal. More-



(a) ΔfwsSNR

(b) ΔPESQ

Fig. 3.9: Performance of the IRLS-p method with $M = 1$ for different values of the filter length L_g in terms of ΔfwsSNR (left) and ΔPESQ (right).



(a) ΔfwsSNR

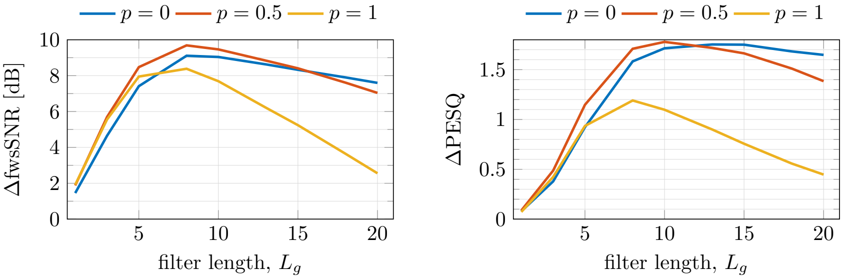
(b) ΔPESQ

Fig. 3.10: Performance of the IRLS-p method with $M = 2$ for different values of the filter length L_g in terms of ΔfwsSNR (left) and ΔPESQ (right).

over, $p = 0.5$ performs best in terms of both measures across all values of the filter length.

Secondly, we consider the results for the setup with $M = 2$ microphones, depicted in Fig. 3.10. The dereverberation performance in terms of both ΔfwsSNR and ΔPESQ depends significantly on the filter length, with the best results obtained with L_g between 20 and 30 for $p = 0$ and $p = 0.5$, and L_g between 15 and 20 for $p = 1$. Interestingly, the performance decreases as the filter length is further increased. The performance also depends on the shape parameter p, with $p = 0.5$ consistently performing better than the other two in terms of both measures and across all filter lengths. It can also be observed that the best-case performance is significantly higher than for the single-channel case (i.e., $M = 1$), with $M = 2$ resulting in an improvement of approximately 5 dB in ΔfwsSNR and more than 1 point in ΔPESQ over $M = 1$.

Thirdly, we consider the results for the setup with $M = 4$ microphones, depicted in Fig. 3.11. Again, the dereverberation performance in terms of both ΔfwsSNR and ΔPESQ depends significantly on the filter length, with the best results obtained with

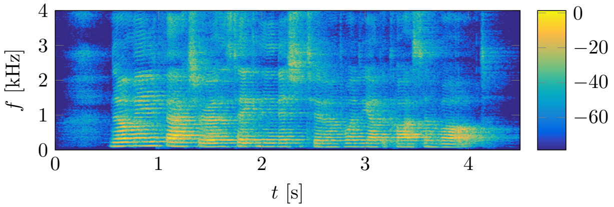


(a) ΔfwsSNR (b) ΔPESQ

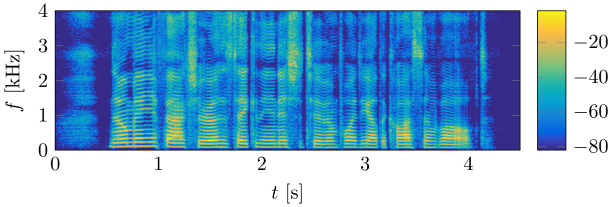
Fig. 3.11: Performance of the IRLS-p method with $M = 4$ for different values of the filter length L_g in terms of ΔfwsSNR (left) and ΔPESQ (right).

L_g between 8 and 12 for $p = 0$ and $p = 0.5$, and L_g between 5 and 10 for $p = 1$. As before, the performance decreases for relatively large filter lengths. The performance also depends on the shape parameter p, with $p = 0.5$ performing better than $p = 0$ and $p = 1$ in terms of both measures (except for large filter lengths). It can also be observed that the best-case performance is somewhat higher than for $M = 2$, with $M = 4$ resulting in an improvement of approximately 1 dB in ΔfwsSNR and 0.1 point in ΔPESQ over $M = 2$. This indicates that the performance improvement is not proportional to the number of microphones, and increasing the number of microphones further might only result in marginal improvements.

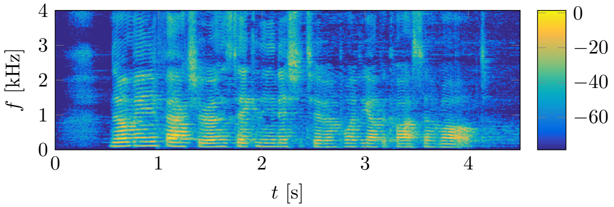
In summary, these results show that the dereverberation performance increases as the number of microphones is increased, with a very large performance difference between $M = 1$ and $M = 2$ and a much smaller difference between $M = 2$ and $M = 4$, indicating marginal improvements for further increasing the number of microphones. Overall, the shape parameter $p = 0.5$ performs better than $p = 0$ and $p = 1$ in terms of both performance measures. The differences between the estimated desired speech signal when using $M \in \{1, 2, 4\}$ microphones can also be observed from the spectrograms of the corresponding signals depicted in Fig. 3.12. By comparing the estimated desired speech signal with the microphone signal, it can be observed that $M = 1$ results in some dereverberation. However, $M = 2$ and $M = 4$ achieve result in much better dereverberation than $M = 1$, with a very small difference between $M = 2$ and $M = 4$. Although the reverberation time is fixed in this scenario, the optimal filter length highly depends on the number of microphones as suggested with the expression in (2.8). For the used reverberation time $T_{60} \approx 700$ ms, this would correspond to theoretical length of the time-domain prediction filter of approximately 1.4 s for $M = 2$ and approximately 900 ms for $M = 4$. Since a subband model is used here, this would correspond to the length of approximately 85 time frames for $M = 2$ and 55 frames for $M = 4$. However, the experimentally observed optimal values are much lower than this. This can be attributed to two effects: (i) the very late reverberation, corresponding to the reverberant tail, can be difficult to predict, since it is not strongly structured and resembles random noise, and (ii) for $M > 1$ and a fixed number of time frames N,



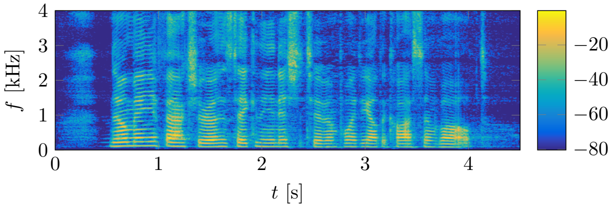
(a) Microphone signal



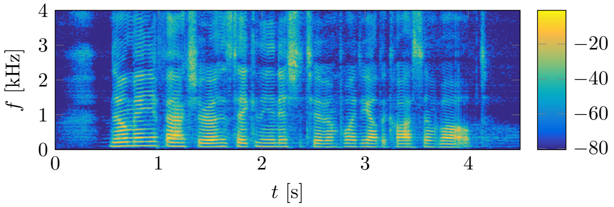
(b) Direct speech signal



(c) $M = 1$



(d) $M = 2$



(e) $M = 4$

Fig. 3.12: Spectrograms of the microphone signal, direct speech signal and the desired speech signal estimated using IRLS-p with $p = 0.5$ and using $M \in \{1, 2, 4\}$ microphones.

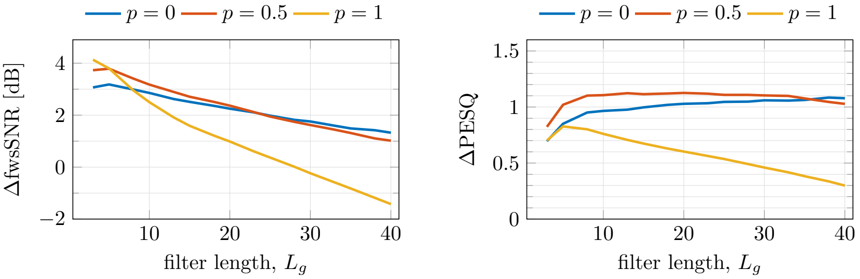
very long filter lengths L_g would result in the desired signal obtained by solving the optimization problem in (3.27) being equal to zero, independently of the considered ℓ_p-norm. The first effect explains the marginal improvements for the single-channel case (i.e., $M = 1$) for large filter lengths, and a similar behavior for $M > 1$ for filter lengths slightly larger than the optimal filter length. The second effect explains why large filter lengths result in a performance deterioration for $M > 1$, which can be attributed to over-suppression of the speech signal. As a rule of thumb, it seems that a simple heuristic can be used to determine a near-optimal filter length L_g for a given number of microphones M by keeping the total number of coefficients ML_g constant. For the considered scenario and STFT setup the optimal performance is obtained with approximately $ML_g \approx 50$, resulting in the filter length of 50, 25, and 12 coefficients for 1, 2, and 4 microphones, respectively. A similar observation has been reported in the context of speech recognition in [164], where the word error rate has been analyzed as a function of the filter length. Of course, the optimal filter length depends on the application, scenario and algorithm setup.

3.5.4 *Influence of filter length L_g for different acoustic scenarios*

In this section, we investigate the influence of the filter length L_g on the dereverberation performance of the considered IRLS-p method for sparse MCLP-based dereverberation for three different reverberation times, namely $T_{60} \approx \{250, 600, 700\}$ ms. The number of microphones is fixed to $M = 2$, the considered shape parameter values are $p \in \{0, 0.5, 1\}$, and the filter length L_g is varied between 3 and 40. In all experiments the optimization is initialized using the coefficients of the reference microphone signal. The maximum number of iterations is set to $I = 20$, and the stopping criterion in Alg. 1 is set as described in Section 3.5.1.

Firstly, we consider the results for the scenario with $T_{60} \approx 250$ ms, depicted in Fig. 3.13. It can be observed that increasing the filter length initially results in an improved performance in terms of ΔPESQ and in minor changes in terms of ΔfwsSNR. Large filter lengths result in a decreasing performance in terms of ΔfwsSNR and a stagnating or decreasing performance in terms of ΔPESQ. Furthermore, $p = 0$ and $p = 0.5$ improve the performance over the microphone signal for all considered filter lengths, while $p = 1$ even may result in a decreased performance for long filters, especially visible in terms of ΔfwsSNR. Overall, the best performance is obtained for $p = 0.5$ with L_g between 5 and 10 providing a reasonable tradeoff between the improvements in ΔfwsSNR and ΔPESQ.

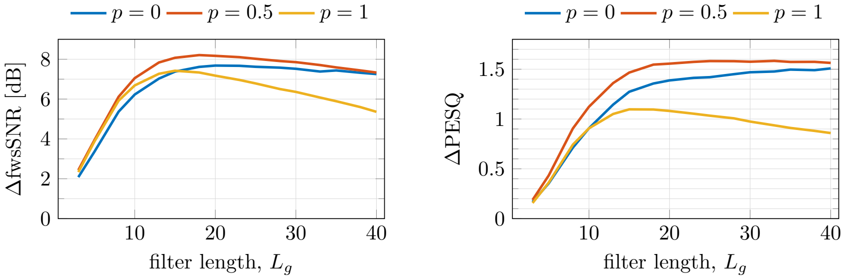
Secondly, we consider the results for the scenario with $T_{60} \approx 600$ ms, depicted in Fig. 3.14. Similar trends as for $T_{60} \approx 250$ ms can be observed in the behavior of the performance measures with respect to the filter length. More specifically, increasing the filter length initially results in an improved performance for both measures, until optimal performance is reached. Further increasing the filter length results in a decreasing or stagnating performance. Overall, the best performance is obtained for $p = 0.5$ with L_g between 15 and 25, providing significant improvements in ΔfwsSNR and ΔPESQ.



(a) ΔfwsSNR

(b) ΔPESQ

Fig. 3.13: Performance of the IRLS-p method with $T_{60} \approx 250$ ms for different values of the filter length L_g in terms of ΔfwsSNR (left) and ΔPESQ (right).



(a) ΔfwsSNR

(b) ΔPESQ

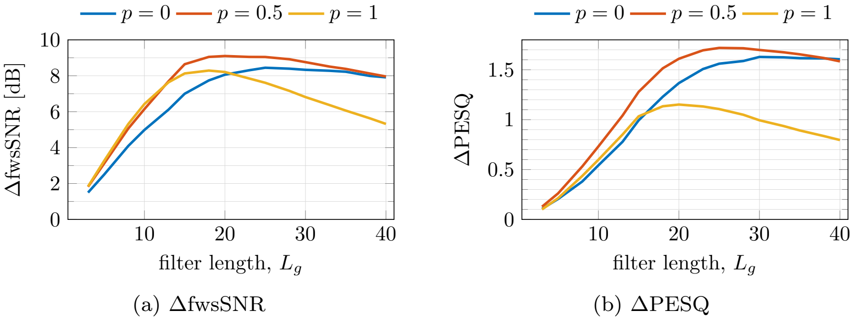
Fig. 3.14: Performance of the IRLS-p method with $T_{60} \approx 600$ ms for different values of the filter length L_g in terms of ΔfwsSNR (left) and ΔPESQ (right).

Thirdly, we consider the results for the scenario with $T_{60} \approx 700$ ms, depicted in Fig. 3.15. Again, a similar trend as earlier can be observed, with the best performance obtained for $p = 0.5$ with L_g between 20 and 30.

In summary, the filter length corresponding to the optimal performance highly depends on the reverberation time, as expected. A simple heuristic for determining a useful filter length for a given T_{60} with $M = 2$ would be to use L_g slightly larger than half of the length corresponding to the reverberation time, which would correspond to 7, 18, and 22 for the considered reverberation times. Again, the actual optimal filter length depends on the application, scenario and algorithm setup.

3.6 Summary

In this chapter, we have presented a novel MCLP-based blind speech dereverberation method, based on a sparse prior for modeling the desired speech signal, with a special emphasis on circular priors from the complex generalized Gaussian family. We have estimated the prediction filter by iteratively maximizing the corresponding



(a) ΔfwsSNR (b) ΔPESQ

Fig. 3.15: Performance of the IRLS-p method with $T_{60} \approx 700$ ms for different values of the filter length L_g in terms of ΔfwsSNR (left) and ΔPESQ (right).

likelihood function, using a variational representation of the sparse prior. The proposed signal model can be interpreted as a generalization of the TVG model, with an additional hyperprior on the unknown variances. It has also been shown that the underlying prior in the conventional WPE method strongly promotes sparsity of the desired speech signal and can be obtained as a special case of the proposed method with $p = 0$. The proposed method has also been reformulated as a constrained optimization problem minimizing the ℓ_p-norm of the desired speech signal. Furthermore, solving this optimization problem by an iteratively reweighted LS algorithm results in a set of updates corresponding to the probabilistic formulation with a complex generalized Gaussian prior.

The experimental results for various setups and acoustic scenarios show that the speech enhancement performance can be consistently improved by the proposed general method by selecting an appropriate value of the shape parameter p in the sparsity-promoting cost function. While the improvements are mild compared to the conventional WPE method, it is important to keep in mind that these come at virtually no cost with just a minor modification of the weight/variance update. As we have analytically shown using the ℓ_p-norm-based formulation, speech dereverberation is achieved by exploiting the fact that the desired speech signal is more sparse than the reverberant observations in the STFT domain. Furthermore, the highlighted role of sparsity-promoting cost functions also suggests that different cost functions and optimization methods could be applied to achieve speech dereverberation. These insights could be useful not only for the considered MCLP-based dereverberation method but also for integration of MCLP-based dereverberation with other speech enhancement methods, such as denoising and source separation.

4

GROUP SPARSE MULTI-CHANNEL LINEAR PREDICTION FOR MULTIPLE-INPUT MULTIPLE-OUTPUT DEREVERBERATION

In Chapter 3, a blind single-output speech dereverberation method based on sparse MCLP has been proposed, estimating the desired speech signal at one of the microphones. However, in many applications it is beneficial to have a multi-channel output signal which can be used for further processing, e.g., for source localization, denoising or source separation.

In [161], a generalized WPE (GWPE) method for multiple-input multiple-output (MIMO) dereverberation based on a time-varying multivariate complex Gaussian model has been proposed. The estimation of the prediction filter in GWPE is formulated as minimization of inter-frame dependence, quantified using a correlation measure called Hadamard-Fischer mutual correlation.

In this chapter, we propose a principled way to obtain a multi-channel dereverberated output signal based on the concept of group sparsity. More specifically, as a multi-output extension of the sparse MCLP method from Chapter 3, we propose to achieve MIMO speech dereverberation using group sparse MCLP, by promoting sparsity across time and taking into account grouping of the coefficients across the microphones. In Section 4.1, we formulate the general problem of blind MIMO speech dereverberation using the subband signal model. In Section 4.2, we introduce the concept of group sparsity and mixed norms, used to quantify structured sparsity. As a multi-channel extension of the ℓ_p-norm based optimization problem for sparse MCLP from Chapter 3, in Section 4.1 we formulate a non-convex optimization problem based on a mixed $\ell_{p,2}$-norm, which is solved using an IRLS algorithm. We show that the proposed formulation generalizes several existing methods. The performance of the proposed method is evaluated in Section 4.4, where the obtained

---

This chapter is partly based on:

[179] A. Jukić, T. van Waterschoot, T. Gerkmann, S. Doclo, “Group sparsity for MIMO speech dereverberation,” in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, Oct. 2015.

results show the advantage of the non-convex cost functions compared with the convex cost function.

4.1 Problem formulation

Similarly as in Chapter 3, we consider an acoustic scenario with a single static speech source captured by M microphones in a reverberant enclosure without the presence of additive noise. While in Chapter 3 the goal was to estimate the desired speech signal at one of the microphones, i.e., the reference microphone, here we aim to jointly compute an M-channel output signal corresponding to the M-channel desired speech signal at all microphones. This can be advantageous in many applications, since a multi-channel output signal can be used for further multi-channel processing, e.g., for direction-of-arrival estimation [161], denoising [82, 110, 168], or source separation [160, 189]. Using (2.29), the MCLP-based signal model for all M reverberant microphone signals in the k-th subband can be written as

$$\mathbf{X}(k) = \mathbf{D}(k) + \tilde{\mathbf{X}}_\tau(k)\mathbf{G}(k), \tag{4.1}$$

where $\mathbf{X}(k) \in \mathbb{C}^{N \times M}$ is the multi-channel reverberant speech matrix, $\mathbf{D}(k) \in \mathbb{C}^{N \times M}$ is the multi-channel desired speech matrix, $\tilde{\mathbf{X}}_\tau(k) \in \mathbb{C}^{N \times ML_g}$ is the multi-channel convolution matrix with delay τ, and $\mathbf{G}(k) \in \mathbb{C}^{ML_g \times M}$ is the MIMO prediction filter. The problem of blind speech dereverberation can now be formulated as blind estimation of the desired M-channel speech matrix $\mathbf{D}(k)$ using only the reverberant observations $\mathbf{X}(k)$, i.e., without using the ATFs between the speech source and the microphones. Using the signal model in (4.1) and given an estimate $\hat{\mathbf{G}}(k)$ of the MIMO prediction filter, the desired speech signal can be estimated as

$$\hat{\mathbf{D}}(k) = \mathbf{X}(k) - \tilde{\mathbf{X}}_\tau(k)\hat{\mathbf{G}}(k). \tag{4.2}$$

Similarly as in Chapter 3, the M-channel desired speech signal can be interpreted as the multi-channel prediction error of the delayed linear prediction model [154]. MIMO dereverberation can be performed by estimating the MIMO prediction filter $\mathbf{G}(k)$ for each subband k and applying (4.2). A block scheme of this MCLP-based speech dereverberation system is depicted in Fig. 4.1. In the remainder of this chapter each subband will be processed independently and the index k will be omitted where possible for notational convenience.

In the following section we formulate speech dereverberation as an optimization problem with a cost function promoting group-sparsity, and propose to solve it using an IRLS algorithm. We start with defining mixed norms and briefly review their relationship to group sparsity.

4.2 Group-sparse modeling

Sparse modeling has been extensively employed in many inverse problem, including speech dereverberation. In many applications it is also possible to exploit

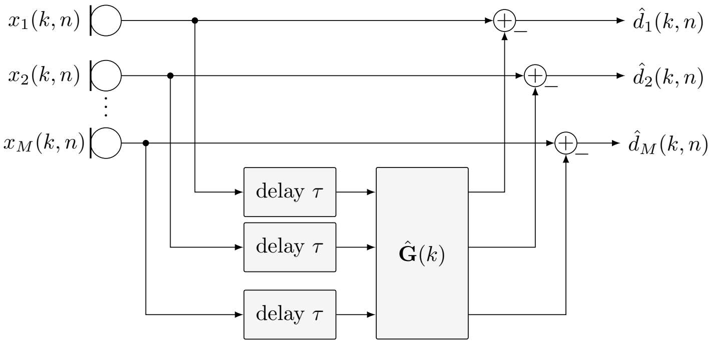


Fig. 4.1: A block scheme of an MCLP-based MIMO dereverberation system.

additional structure of the desired signal. For example, in certain cases the desired signal exhibits sparsity with a certain pattern, e.g., when the significant nonzero coefficients naturally appear together in groups. This concept, typically referred to as group, joint, or block sparsity, has been used in signal processing and machine learning [236–239]. Group structure is usually enforced by using mixed norms [236, 237, 240] or probabilistic models [233, 241].

Mixed norms generalize the usual matrix and vector norms [236, 237, 240, 242]. Let $d_m(n)$ be the element in the n-th row and m-th column of the matrix $\mathbf{D} \in \mathbb{C}^{N \times M}$. Let the elements of the n-th row of $\mathbf{D}$ be contained in a (column) vector $\mathbf{d}(n)$ as

$$\mathbf{d}(n) = [d_1(n), \dots, d_M(n)]^{\mathsf{T}}, \tag{4.3}$$

i.e., the vector $\mathbf{d}(n) \in \mathbb{C}^M$ contains the coefficients of the multi-channel desired speech signal at the n-th time frame. Let $p \in (0, 2]$ be a shape parameter and $\boldsymbol{\Phi} \in \mathbb{C}^{M \times M}$ be a positive definite matrix. The mixed $\ell_{p,2;\boldsymbol{\Phi}}$-norm of the matrix $\mathbf{D}$ is then defined as

$$\|\mathbf{D}\|_{p,2;\boldsymbol{\Phi}} = \left(\sum_{n=1}^{N} \|\mathbf{d}(n)\|_{2;\boldsymbol{\Phi}}^{p} \right)^{\frac{1}{p}}, \tag{4.4}$$

where

$$\|\mathbf{d}(n)\|_{2;\boldsymbol{\Phi}} = \left(\mathbf{d}^{\mathsf{H}}(n) \boldsymbol{\Phi}^{-1} \mathbf{d}(n) \right)^{\frac{1}{2}} \tag{4.5}$$

is the $\ell_{2;\boldsymbol{\Phi}}$-norm of the vector $\mathbf{d}(n)$. In this context, each row of $\mathbf{D}$, i.e., $\mathbf{d}(n)$, represents a group. The role of the matrix $\boldsymbol{\Phi}$ is to model the correlation structure within the group, i.e., within the rows of $\mathbf{D}$. When $\boldsymbol{\Phi}$ is the identity matrix, i.e., $\boldsymbol{\Phi} = \mathbf{I}$, we denote the corresponding norm as the mixed $\ell_{p,2}$-norm.

Fig. 4.2 provides an illustration of the computation of a mixed norm for a given matrix. In words, the mixed $\ell_{p,2;\boldsymbol{\Phi}}$-norm of $\mathbf{D}$ is composed of the inner $\ell_{2;\boldsymbol{\Phi}}$-norm applied on each row of $\mathbf{D}$ in the first step, and the outer ℓ_p-norm applied on the vector composed of the values obtained in the first step. Intuitively, the inner $\ell_{2;\boldsymbol{\Phi}}$-norm measures the weighted energy of the coefficients in each row, while the outer ℓ_p-norm measures the number of rows with significant energies, i.e., the $\ell_{p,2;\boldsymbol{\Phi}}$-norm

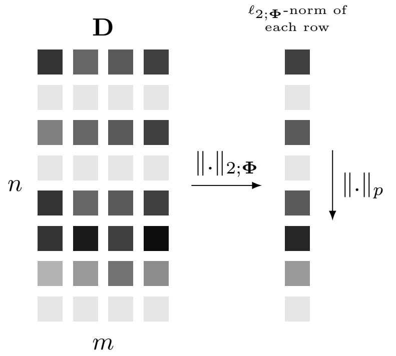

D
$\ell_{2;\boldsymbol{\Phi}}$-norm of each row
$\|\cdot\|_{2;\boldsymbol{\Phi}}$
$\|\cdot\|_p$
n
m

Fig. 4.2: Illustration of computation of the mixed norm $\|\mathbf{D}\|_{p,2;\boldsymbol{\Phi}}$.

of $\mathbf{D}$ provides a measure of group sparsity of the matrix $\mathbf{D}$. Therefore, minimization of (4.4) aims at estimating a matrix $\mathbf{D}$ that has a relatively small number of rows with significant energy, in terms of the $\ell_{2;\boldsymbol{\Phi}}$ norm, and a relatively high number of rows with small energy. Similarly as for the vector norms, for $p < 1$ in (4.4) the obtained functional is not a norm since it is not convex. Nevertheless, we will still refer to the $\ell_{p,2;\boldsymbol{\Phi}}$-norm for $p < 1$ as a norm.

The defined mixed norm includes many matrix norms as a special case, e.g., the $\ell_{2,2}$-norm is the Frobenius norm of a matrix. A commonly used mixed norm is the $\ell_{1,2}$-norm, which is typically referred to as the group Lasso [236] or joint sparsity [243] penalty, and it is often used in sparse regression with the goal of keeping or discarding entire groups (here rows) of elements in a matrix [242].

4.3 MIMO dereverberation using a group-sparse penalty

As a multi-channel extension of the ℓ_p-norm-based optimization problem in (3.27) for estimating the prediction filter $\hat{\mathbf{g}}_{\text{ref}}$, in this section we propose to estimate the prediction filter $\hat{\mathbf{G}}$ for MIMO speech dereverberation by solving the optimization problem based on a mixed norm, i.e.,

$$\begin{aligned} \min_{\mathbf{G}} \quad & \|\mathbf{D}\|_{p,2;\boldsymbol{\Phi}}^p \\ \text{subject to} \quad & \mathbf{D} + \tilde{\mathbf{X}}_\tau \mathbf{G} = \mathbf{X} \end{aligned} \tag{4.6}$$

for $p \leq 1$. The motivation behind the proposed mixed norm cost function is to estimate a a prediction filter $\hat{\mathbf{G}}$ that results in some rows with significant energy in $\mathbf{D}$, and suppresses the coefficients in the remaining rows. For $p = 1$ and $\boldsymbol{\Phi} = \mathbf{I}$ the cost function in (4.6) is the $\ell_{1,2}$-norm as in group Lasso, with the groups being defined across the M microphones. While for $p = 1$ the cost function in (4.6) is convex, it is known that non-convex penalty functions, i.e., $p < 1$, can be more useful in enforcing sparsity [244], similarly as for the ℓ_p-norm-based cost function used in Chapter 3.

The proposed cost function for MIMO speech dereverberation is motivated by the following common assumptions. Firstly, as discussed in Section 2.2, reverberation makes the TF coefficients of the microphone signals less sparse than the TF coefficients of the corresponding clean speech signal. Therefore, as already done in Chapter 3, dereverberation can be achieved by estimating a prediction filter that makes the estimate of the desired speech signal more sparse across time than the microphone signal. Secondly, for a relatively small microphone array it is plausible to assume that at a given time frame the speech signal is present or absent simultaneously at all microphones. This assumption is satisfied, e.g., when the relative delay between the microphone signals is smaller than the frame shift of the TF transform, which is virtually always true for relatively compact, non-distributed arrays. Hence, MIMO dereverberation can be formulated as estimation of the MIMO prediction filter using a cost function promoting group sparsity as in (4.6), with the groups defined across the microphones and the matrix $\mathbf{\Phi}$ capturing the spatial correlation of the speech source between the microphones. The MIMO prediction filter obtained by solving (4.6) hence aims to estimate the desired speech signal matrix $\mathbf{D}$ that is more sparse than the reverberant speech matrix $\mathbf{X}$, by simultaneously keeping or discarding the coefficients across the microphones. Therefore, the undesired reverberation will be suppressed by enforcing sparsity over time, with the spatial correlation, i.e., the group structure, being taken into account.

4.3.1 *Non-convex minimization using IRLS*

In Section 3.4.1, the ℓ_p-norm-based optimization problem in (3.27)/(3.35) has been solved using the IRLS algorithm, and a similar approach can be used here to solve (4.6). More specifically, the mixed $\ell_{p,2;\mathbf{\Phi}}$-norm in (4.6) is approximated with a convex weighted $\ell_{2,2;\mathbf{\Phi}}$-norm. Therefore, in the i-th iteration of the IRLS algorithm the non-convex ℓ_p-norm of the energies of the rows of $\mathbf{D}$ is replaced by a convex weighted ℓ_2-norm, resulting in the following approximation of the cost function

$$\|\mathbf{D}\|_{p,2;\mathbf{\Phi}}^{p} = \sum_{n=1}^{N} \|\mathbf{d}(n)\|_{2;\mathbf{\Phi}}^{p} \approx \sum_{n=1}^{N} \hat{w}^{i}(n) \|\mathbf{d}(n)\|_{2;\mathbf{\Phi}}^{2} \tag{4.7}$$

where $\hat{w}^i(n)$ is the estimated weight for the n-th time frame n in the i-th iteration. The quadratic approximation of the original cost function can then be written in matrix form as

$$\|\mathbf{D}\|_{p,2;\mathbf{\Phi}}^{p} \approx \operatorname{tr}\left[\hat{\mathbf{W}}^{i}\mathbf{D}\mathbf{\Phi}^{-\mathsf{T}}\mathbf{D}^{\mathsf{H}}\right], \tag{4.8}$$

where $\hat{\mathbf{W}}^i$ is a diagonal matrix with the weights $\hat{w}^i(n)$ on its diagonal, and $\operatorname{tr}[.]$ denotes the matrix trace operator. Similarly as in Section 3.4.1, the weights $\hat{w}^i(n)$ are selected such that the approximation in (4.7) is a first-order approximation of the corresponding $\ell_{p,2;\mathbf{\Phi}}$ cost function (cf. Appendix B.1). Therefore, similarly to (3.40), the weights $\hat{w}^i(n)$ in the i-th iteration are computed from the previous estimate of the desired speech matrix $\hat{\mathbf{D}}^{i-1}$ as

$$\hat{w}^i(n) = \left(\frac{1}{M}\left\|\hat{\mathbf{d}}^{i-1}(n)\right\|_{2;\boldsymbol{\Phi}}^2 + \varepsilon_{\min}\right)^{\frac{p}{2}-1}, \tag{4.9}$$

where a small positive constant $\varepsilon_{\min}$ is included in the weight update to prevent division by zero. Given the weights $\hat{w}^i(n)$ and using the convex approximation in (4.8), the optimization problem in (4.6) in the i-th iteration can be rewritten as

$$\begin{aligned} \hat{\mathbf{G}}^i &= \arg\min_{\mathbf{G}} \operatorname{tr}\left[\mathbf{D}^{\mathsf{H}}\hat{\mathbf{W}}^i\mathbf{D}\boldsymbol{\Phi}^{-\mathsf{T}}\right] \\ &= \arg\min_{\mathbf{G}} \operatorname{tr}\left[\left(\mathbf{X}-\tilde{\mathbf{X}}_\tau\mathbf{G}\right)^{\mathsf{H}}\hat{\mathbf{W}}^i\left(\mathbf{X}-\tilde{\mathbf{X}}_\tau\mathbf{G}\right)\boldsymbol{\Phi}^{-\mathsf{T}}\right], \end{aligned} \tag{4.10}$$

with the closed-form solution for the MIMO prediction filter in the i-th iteration given as

$$\hat{\mathbf{G}}^i = \left(\tilde{\mathbf{X}}_\tau^{\mathsf{H}}\hat{\mathbf{W}}^i\tilde{\mathbf{X}}_\tau\right)^{-1}\tilde{\mathbf{X}}_\tau^{\mathsf{H}}\hat{\mathbf{W}}^i\mathbf{X}. \tag{4.11}$$

Note that the obtained solution for the prediction filter $\hat{\mathbf{G}}^i$ does not depend on the matrix $\boldsymbol{\Phi}$. Nevertheless, the choice of $\boldsymbol{\Phi}$ affects the calculation of the weights $\hat{w}^i(n)$ in (4.9), and can therefore influence the final estimate of the prediction filter and hence the dereverberation performance. To take the spatial (within-group) correlation into account, the matrix $\boldsymbol{\Phi}$ in the i-th iteration can be updated using the estimate of the desired speech matrix $\hat{\mathbf{D}}^i$ as

$$\hat{\boldsymbol{\Phi}}^i = \frac{1}{N}\sum_{n=1}^{N}\hat{w}^i(n)\hat{\mathbf{d}}^i(n)\left(\hat{\mathbf{d}}^i(n)\right)^{\mathsf{H}} = \frac{1}{N}\left(\hat{\mathbf{D}}^i\right)^{\mathsf{T}}\hat{\mathbf{W}}^i\left(\hat{\mathbf{D}}^i\right)^*, \tag{4.12}$$

with $(.)^*$ denoting the complex conjugate. This update can be obtained by minimizing the cost function in (4.10) with an additional additive term ($N\log\det\boldsymbol{\Phi}$), which corresponds to a ML estimator of $\boldsymbol{\Phi}$ when $\mathbf{d}(n)$ is modeled using a zero-mean complex Gaussian distribution with covariance $\hat{w}^{-1}(n)\boldsymbol{\Phi}$, as commonly used in group sparse learning [233, 245]. In a practical implementation, a small diagonal loading $\varepsilon_{\boldsymbol{\Phi}}$ can be used to regularize the matrix $\hat{\boldsymbol{\Phi}}^i$. The complete algorithm for solving (4.6) using IRLS is outlined in Alg. 3.

4.3.2 *Relation to existing methods*

The GWPE method in [161] was derived based on a locally Gaussian model for the multi-channel desired signal, i.e., $\mathbf{d}(n)$ was modeled using a multivariate complex Gaussian distribution with an unknown and time-varying covariance matrix. The optimization problem for the MIMO prediction filter $\hat{\mathbf{G}}$ was formulated using a cost function based on the Hadamard-Fischer mutual correlation, which favors temporally uncorrelated random vectors, i.e., the prediction filter was estimated by decorrelating the vectors $\mathbf{d}(n)$ across time. In order to derive a practical algorithm,

Alg. 3 Group sparse MCLP with a mixed $\ell_{p,2;\boldsymbol{\Phi}}$-norm cost function using the IRLS algorithm (gIRLS-p).

```
parameters: filter length $L_g$ and prediction delay $\tau$ in (4.1), $p$ in (4.6), regularization parameters $\varepsilon_{\min}$, $\varepsilon_{\boldsymbol{\Phi}}$, maximum number of iterations $I$, tolerance $\eta$
input: $M$-channel reverberant microphone signal coefficients $\mathbf{X}(k)$, $\forall k$
for all $k$ do
    $i \leftarrow 0$
    $\hat{\mathbf{D}}^0 \leftarrow \mathbf{X}$, $\hat{\boldsymbol{\Phi}}^0 \leftarrow \mathbf{I}$                                   ▷ initialization
    repeat
        $i \leftarrow i + 1$
        $\hat{w}^i(n) \leftarrow \left( \frac{1}{M} \left\| \hat{\mathbf{d}}^{i-1}(n) \right\|^2_{2;\hat{\boldsymbol{\Phi}}^{i-1}} + \varepsilon_{\min} \right)^{\frac{p}{2}-1}$, $\forall n$          ▷ equation (4.9)
        $\hat{\mathbf{G}}^i \leftarrow \left( \tilde{\mathbf{X}}_\tau^{\mathsf{H}} \hat{\mathbf{W}}^i \tilde{\mathbf{X}}_\tau \right)^{-1} \tilde{\mathbf{X}}_\tau^{\mathsf{H}} \hat{\mathbf{W}}^i \mathbf{X}$          ▷ equation (4.11)
        $\hat{\mathbf{D}}^i \leftarrow \mathbf{X} - \tilde{\mathbf{X}}_\tau \hat{\mathbf{G}}^i$                                  ▷ equation (4.2)
        if $\boldsymbol{\Phi}$ is estimated then
            $\hat{\boldsymbol{\Phi}}^i \leftarrow \frac{1}{N} \hat{\mathbf{D}}^{(i)\mathsf{T}} \hat{\mathbf{W}}^i \hat{\mathbf{D}}^{(i)*} + \varepsilon_{\boldsymbol{\Phi}} \mathbf{I}$          ▷ equation (4.12)
        end if
    until $i = I$ or $\frac{\left\| \hat{\mathbf{D}}^i - \hat{\mathbf{D}}^{i-1} \right\|_F}{\left\| \hat{\mathbf{D}}^{i-1} \right\|_F} < \eta$
end for
output: estimated desired $M$-channel signal coefficients $\hat{\mathbf{D}}(k) = \hat{\mathbf{D}}^i(k)$, $\forall k$
```

a suitable auxiliary majorizing function was derived, which is minimized using alternating optimization. By comparing Alg. 3 with the updates in [161], it can be seen that the GWPE method corresponds a special case of the proposed method with $p = 0$, i.e., to the minimization of the $\ell_{0,2;\boldsymbol{\Phi}}$-norm in (4.6). Therefore, similarly as in Chapter 3, the success of the GWPE-based dereverberation can be attributed to the sparsifying behavior of the underlying cost function used to estimate the prediction filter $\hat{\mathbf{G}}$.

Furthermore, if an $\ell_{p,p}$-norm would be used as the cost function in (4.6), the proposed method would be decomposed into a multiple-input single-output method from Chapter 3 applied M times to generate M outputs, with each microphone being selected as the reference microphone exactly once. This is a direct consequence of ignoring the group structure when using the $\ell_{p,p}$-norm, i.e.,

$$\|\mathbf{D}\|_p^p = \sum_{n=1}^{N} \sum_{m=1}^{M} |d_m(n)|^p, \tag{4.13}$$

as the cost function. Since the signal model in (4.1) can then be decoupled into M independent models, as in (2.28), this leads to independent estimation for the M prediction filters.

4.4 Simulations

In this section, the dereverberation performance of the proposed group sparse MCLP method is investigated. More specifically, the performance for different values of the shape parameter p and the number of microphones M is compared, with and without the presence of additive noise.

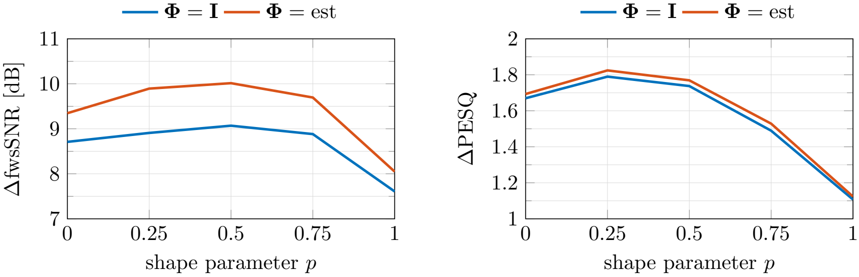
The considered acoustic scenario and the implementation details are outlined in Section 4.4.1. The influence of the shape parameter p for an acoustic scenario without the presence of additive noise is investigated in Section 4.4.2. The influence of additive noise on the speech dereverberation performance is investigated in Section 4.4.3.

4.4.1 *Acoustic scenario and algorithmic setup*

We consider an acoustic scenario with a single static speech source and $M \in \{2, 4\}$ omni-directional microphones placed at a distance of about 2 m from the source, with the same array configuration as in Section 3.5. To generate the noisy reverberant signals, recorded noise signal has been added to the reverberant speech signal to achieve a desired signal-to-noise ratio (SNR) with respect to the direct speech signal at the first microphone. The noise has been recorded in the same conditions as the RIRs, and consists mainly of a stationary background noise caused by the air conditioning system (cf. REVERB challenge [22, 23]).

Similarly to the parameter setup for the single-output methods in Chapter 3, the analysis and synthesis STFT is computed using a tight window based on a 64 ms Hamming window with a 16 ms window shift. The prediction delay in (4.1) is set to $\tau = 2$ for all experiments, and the filter length is set to $L_g = 25$ for $M = 2$ microphones and $L_g = 10$ for $M = 4$ microphones. The filter length L_g is selected as suggested by the results in Section 3.5.3. The proposed group sparse MCLP dereverberation method based on IRLS (gIRLS-p) is implemented as in Alg. 3, with the conventional GWPE method corresponding to $p = 0$. The weights are regularized with $\varepsilon_{\min} = 10^{-8}$. The iterative algorithm is initialized with the microphone signal coefficients. The tolerance for the relative change of the ℓ_2-norm of the estimated desired signal is set to $\eta = 10^{-6}$ and the maximum number of iterations is set to $I = 20$.

The dereverberation performance is evaluated in terms of the instrumental measures described in Section 2.3, i.e., improvement in fwsSNR (ΔfwsSNR) and PESQ (ΔPESQ). The reference signal used for the instrumental measures is the direct speech signal at the microphone, obtained by convolving the anechoic speech signal with the direct component of the corresponding RIR. The reported improvements of the instrumental measures are obtained by averaging over all microphones and speech samples.



(a) ΔfwsSNR (b) ΔPESQ

Fig. 4.3: Performance of the gIRLS-p method with $M = 2$ and $L_g = 25$ in terms of ΔfwsSNR and ΔPESQ. The correlation matrix $\mathbf{\Phi}$ was fixed ($\mathbf{\Phi} = \mathbf{I}$) or estimated using (4.12) ($\mathbf{\Phi} = \text{est}$).

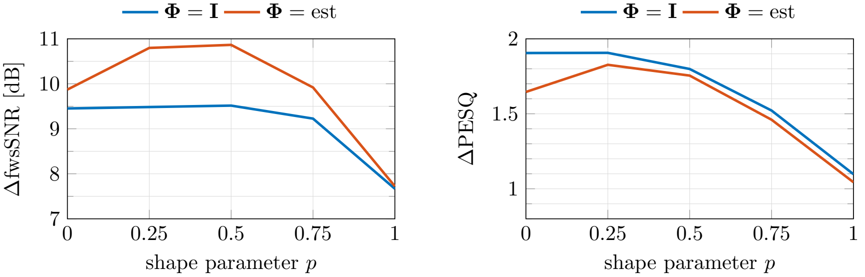
4.4.2 *Influence of the shape parameter p in a noiseless scenario*

In this section, we investigate the influence of the shape parameter p on the performance of the proposed gIRLS-p speech dereverberation method. The shape parameter is varied between $p = 0$ and $p = 1$, with $p = 0$ corresponding to the conventional GWPE method and $p = 1$ corresponding to the convex group Lasso cost function.

Firstly, we consider the case with $M = 2$ microphones. Fig. 4.3 depicts the improvement of the considered performance measures, either using a fixed within-group correlation matrix ($\mathbf{\Phi} = \mathbf{I}$) or an estimated within-group correlation matrix ($\mathbf{\Phi} = \text{est}$) using (4.12). It can be observed that both performance measures exhibit a similar trend for both correlation matrices. In general, the obtained improvements highly depend on the shape parameter p. More specifically, the performance significantly deteriorates as the cost function becomes closer to the convex case, i.e., as the shape parameter p approaches 1. In general, smaller values of the shape parameter p, corresponding to non-convex cost functions which promote sparsity more aggressively, result in a better performance. In particular, values of the shape parameter between $p = 0.25$ and $p = 0.5$ achieve somewhat better performance than when using $p = 0$, similarly as in the single-channel output case in Section 3.5.

In terms of reverberation suppression, as indicated by ΔfwsSNR, the estimated correlation matrix $\mathbf{\Phi}$ results in a better performance (about 1 dB) than the fixed correlation matrix, whereas both correlation matrices achieve a similar perceptual speech quality improvement, as indicated by ΔPESQ.

Secondly, we consider the case with $M = 4$ microphones, with Fig. 4.4 depicting the improvements of the considered performance measures. Similarly as for $M = 2$, both performance measures exhibit a similar trend for both correlation matrices, with the obtained improvements highly depending on the shape parameter p. Again, the performance significantly deteriorates as the cost function becomes closer to the convex case, and the non-convex cost functions (for $p < 1$) result in a better performance. The estimated correlation matrix $\mathbf{\Phi}$ again results in better performance in terms



(a) ΔfwsSNR (b) ΔPESQ

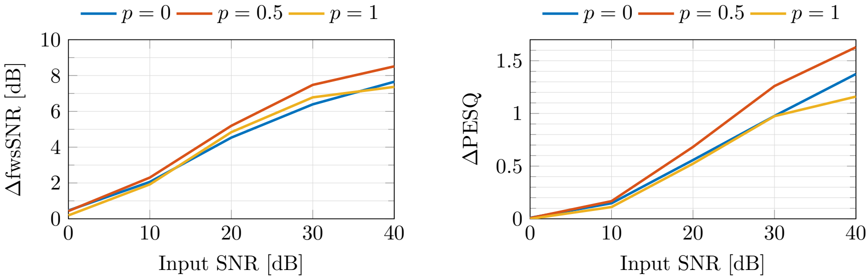
Fig. 4.4: Performance of the gIRLS-p method with $M = 4$ and $L_g = 10$ in terms of ΔfwsSNR and ΔPESQ. The correlation matrix $\mathbf{\Phi}$ was fixed ($\mathbf{\Phi} = \mathbf{I}$) or estimated using (4.12) ($\mathbf{\Phi} = \mathrm{est}$).

of reverberation suppression, as indicated by ΔfwsSNR, with both correlation matrices achieving a similar perceptual speech quality improvement, as indicated by ΔPESQ, except for $p = 0$.

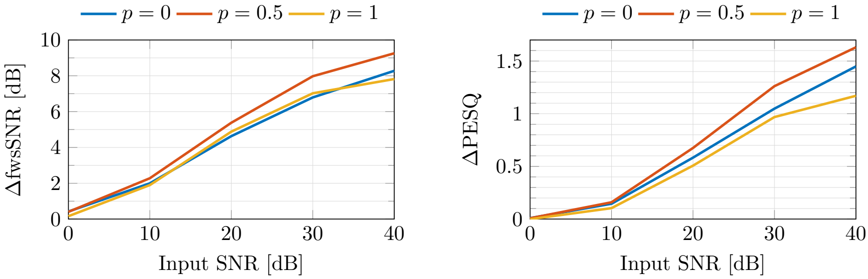
4.4.3 *Influence of noise*

In this section, we investigate the influence of noise on the performance of the proposed gIRLS-p dereverberation method. The input SNR is varied between 0 dB and 40 dB, the shape parameter is set to $p \in \{0, 0.5, 1\}$, and both fixed and estimated correlation matrices are considered.

Fig. 4.5 depicts the improvement of the considered performance measures for $M = 2$. It can be observed that in all cases the method provides an improvement over the microphone signal. However, the obtained improvements are considerably larger for high SNRs, when the microphone signals are mainly corrupted by reverberation, and are significantly reduced for lower SNRs, when the microphone signals are mainly corrupted by noise. Since the MCLP signal model does not explicitly include the noise component, the improvements stem from dereverberation while the noise component is typically not strongly affected. This is due to the fact that the noise signal is typically less predictable than reverberation, such that the estimated prediction filters capture almost exclusively reverberation. Similarly as for the noiseless case in Section 4.4.2, $p = 0.5$ achieves the best performance for all SNR values. The estimated correlation matrix $\mathbf{\Phi}$ results in better performance in terms of reverberation suppression, as indicated by ΔfwsSNR, than the fixed correlation matrix, whereas both correlation matrices achieve similar perceptual speech quality improvements, as indicated by ΔPESQ. Fig. 4.6 depicts the improvement of the considered performance measures for $M = 4$. Again, it can be observed that in all cases the method provides improvements over the microphone signal. Similarly as for $M = 2$, the obtained improvements are considerably larger for high SNRs, and



(a) $\mathbf{\Phi} = \mathbf{I}$ (b) $\mathbf{\Phi} = \mathbf{I}$



(c) $\mathbf{\Phi} = \text{est}$ (d) $\mathbf{\Phi} = \text{est}$

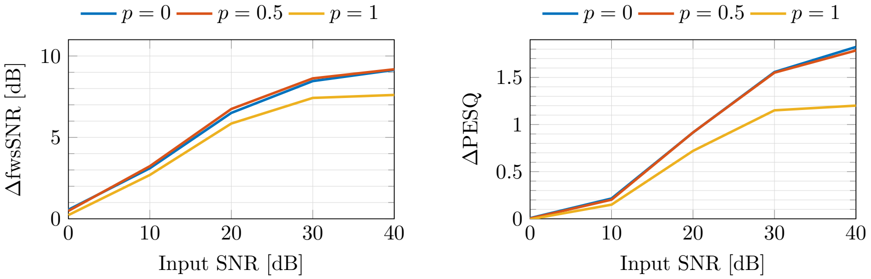
Fig. 4.5: Performance of the gIRLS-p method with $M = 2$ and $L_g = 25$ in terms of ΔfwsSNR (left) and ΔPESQ (right) for different SNRs. The correlation matrix $\mathbf{\Phi}$ was either fixed (top) or estimated using (4.12) (bottom).

are significantly reduced for lower SNRs, with $p = 0.5$ typically performing slightly better or comparable to $p = 0$.

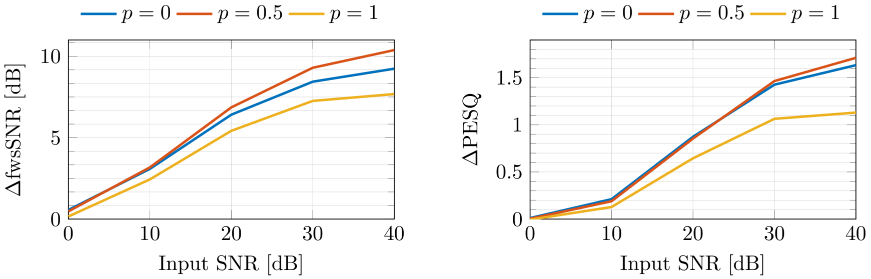
4.5 Summary

In this chapter, we have presented a novel formulation for MCLP-based MIMO speech dereverberation using the concept of group sparsity. The dereverberation is formulated as a non-convex optimization problem based on a mixed norm, which takes into account group sparsity of the TF coefficients of the desired multi-channel speech signal. Intuitively, the cost function promotes sparsity of the TF coefficients across time frames while taking into account grouping across the microphones. The optimization problem solved using the IRLS algorithm generalizes several previously proposed MCLP-based methods, including the single-output method from Chapter 3.

The experimental results obtained for noiseless as well as noisy scenarios show that speech enhancement performance can be improved by selecting an appropriate shape of the group sparsity-promoting cost function. The results for the noiseless acoustic scenario show that the speech enhancement performance, compared to the conven-



(a) $\mathbf{\Phi} = \mathbf{I}$ (b) $\mathbf{\Phi} = \mathbf{I}$



(c) $\mathbf{\Phi} = \text{est}$ (d) $\mathbf{\Phi} = \text{est}$

Fig. 4.6: Performance of the gIRLS-p method with $M = 4$ and $L_g = 10$ in terms of ΔfwsSNR (left) and ΔPESQ (right) for different SNRs. The correlation matrix $\mathbf{\Phi}$ was either fixed (top) or estimated using (4.12) (bottom).

tional method, can be consistently improved in the proposed framework by selecting an appropriate value of the shape parameter p and including the within-group correlation. The experimental results for the noisy acoustic scenario show that the performance is significantly reduced in low-SNR scenarios, since the used signal model does not explicitly take noise into account. Nevertheless, MCLP-based dereverberation can lead to considerable improvements in moderate and high-SNR scenarios, such that it can be used as a preprocessor for further multi-channel signal processing.

5

CONSTRAINED SPARSE MULTI-CHANNEL LINEAR PREDICTION FOR ADAPTIVE SPEECH DEREVERBERATION

In Chapter 4, a blind MIMO speech dereverberation method based on group sparse MCLP has been proposed, aiming to estimate the multi-channel desired speech signal. The proposed method operates in batch mode, i.e., the MIMO prediction filter is estimated by using the complete signal captured at the microphone and does not change over time. While the simulation results show that the proposed method performs well, it is based on the assumption of a static acoustic scenario. However, in many applications the source-microphone geometry is not fixed, e.g., if the speaker or the microphone array (e.g., hearing aids) is moving inside the enclosure or if multiple speakers are taking turns. Furthermore, for real-time applications the microphone signals should be processed online, with latency requirements depending on the application. Adaptive versions of MCLP-based dereverberation that are suitable for online processing have been proposed in [156, 163] where the filter updates are based on the recursive least squares (RLS) algorithm. However, since these methods typically do not include additional knowledge about the undesired speech signal, they may lead to a significant overestimation of the undesired signal and severe distortions of the output signal.

In this chapter, we present an adaptive speech dereverberation method suitable for online processing, based on the batch group sparse MCLP method considered in Chapter 4. To prevent overestimation of the undesired signal, we propose to integrate additional knowledge about the reverberant speech signal. More precisely, we propose to constrain the power of the MCLP-based estimate of the undesired rever-

---

This chapter is partly based on:

[180] A. Jukić, Z. Wang, T. van Waterschoot, T. Gerkmann, S. Doclo, "Constrained multi-channel linear prediction for adaptive speech dereverberation," in *Proceedings of the International Workshop on Acoustic Signal Enhancement (IWAENC)*, Xi'an, China, Sept. 2016.

[181] A. Jukić, T. van Waterschoot, S. Doclo, "Adaptive speech dereverberation using constrained sparse multi-channel linear prediction," *IEEE Signal Processing Letters*, vol. 24, no. 1, pp. 101–105, Jan. 2017.

berant signal using an estimate of the late reverberant power spectral density (PSD). The resulting constrained optimization problem is solved using the alternating direction method of multipliers (ADMM) algorithm [246], which can be implemented efficiently as a variant of the RLS algorithm. Experimental results demonstrate the advantages of the proposed constrained method when the prediction filter needs to adapt quickly, e.g., for a moving source.

In Section 5.1, we formulate the problem of blind MIMO adaptive speech dereverberation. In Section 5.2, we present an adaptive dereverberation method based on group sparse MCLP proposed in Chapter 4. In Section 5.3, we propose a constrained version of the adaptive algorithm by including a bound on the power of the estimated undesired signal. In Section 5.4, we present a diagonal approximation which can be used to significantly reduce computational complexity. Experimental results for a moving source and alternating sources are presented in Section 5.5.

5.1 Problem formulation

We consider an acoustic scenario with a single speech source captured by M microphones. As in (2.22), the m-th microphone signal $x_m(k,n)$ can be decomposed as $x_m(k,n) = d_m(k,n) + u_m(k,n)$, where $d_m(k,n)$ is the desired speech signal, consisting of the direct signal and early reflections, and $u_m(k,n)$ is the undesired speech signal, consisting of late reflections. The multi-channel model at the n-th time frame can be written as

$$\mathbf{x}(n) = \mathbf{d}(n) + \mathbf{u}(n), \tag{5.1}$$

where $\mathbf{x}(n) = [x_1(n), \ldots, x_M(n)]^{\mathsf{T}}$ is the multi-channel reverberant signal and $\mathbf{d}(n)$ and $\mathbf{u}(n)$ are defined similarly. As shown in Section 2.1.2 and used in Section 4.2, $\mathbf{u}(n)$ can be modeled using MCLP as the sum of filtered (delayed) microphone signals, i.e.,

$$\mathbf{u}(n) = \mathbf{G}^{\mathsf{H}}(n)\tilde{\mathbf{x}}_\tau(n), \tag{5.2}$$

where $\mathbf{G}(n) = [\mathbf{g}_1(n), \ldots, \mathbf{g}_M(n)] \in \mathbb{C}^{ML_g \times M}$ denotes the MIMO prediction filter, with $\mathbf{g}_m(n) \in \mathbb{C}^{ML_g}$ containing L_g taps per microphone, and $\tilde{\mathbf{x}}_\tau(n) \in \mathbb{C}^{ML_g}$ is a signal buffer defined as

$$\begin{aligned}\tilde{\mathbf{x}}_\tau(n) = [&x_1(n-\tau), \ldots, x_1(n-\tau-L_g+1), \ldots \\ &\ldots, x_M(n-\tau), \ldots, x_M(n-\tau-L_g+1)]^{\mathsf{T}}. \end{aligned} \tag{5.3}$$

As opposed to Chapters 3 and 4, the prediction filter $\mathbf{G}(n)$ in the signal model in (5.2) is not fixed for all frames, i.e., it can change over time. As discussed in Section 3.1, the prediction delay τ ensures preservation of the short-time speech correlation and early reflections in the desired speech signal.

The goal of blind speech dereverberation is to recover the multi-channel desired speech signal $\mathbf{d}(n)$, which can be achieved by estimating the undesired speech signal $\mathbf{u}(n)$ in (5.2) and subtracting it from the reverberant microphone signals $\mathbf{x}(n)$, i.e., the estimated desired speech signal is equal to the prediction error.

5.2 Adaptive group sparse MCLP

Assuming that the MIMO prediction filter $\mathbf{G}(n)$ does not change over time, i.e., $\mathbf{G}(n) = \mathbf{G}$, a batch dereverberation method based on the signal model in (5.1) has been derived in Chapter 4. More specifically, given a batch of N time frames, the prediction filter $\mathbf{G}$ has been estimated by maximizing sparsity across time, which has been formulated as minimizing the mixed $\ell_{p,2}$-norm of the desired speech signal, cf. (4.6). As shown in Section 4.3, the obtained non-convex optimization problem can be solved using the IRLS algorithm by approximating the ℓ_p-norm using a weighted ℓ_2-norm, cf. (4.7), i.e., the batch prediction filter can be estimated as

$$\hat{\mathbf{G}} = \arg\min_{\mathbf{G}} \sum_{n=1}^{N} \hat{w}(n) \|\mathbf{d}(n)\|_2^2, \tag{5.4}$$

where the weights $\hat{w}(n)$ are set to obtain a first-order approximation of the ℓ_p-norm. For a known $\mathbf{d}(n)$, the weights $\hat{w}(n)$ can be set to the ideal weights

$$\hat{w}(n) = \left(\frac{1}{M} \|\mathbf{d}(n)\|_2^2 + \varepsilon_{\min}\right)^{\frac{p}{2}-1}, \tag{5.5}$$

where $\varepsilon_{\min}$ is a small positive regularization constant. As noted in Section 3.4.1, the weights $\hat{w}(n)$ put more emphasis on frames where the desired signal $\mathbf{d}(n)$ should have a relatively small energy, and therefore mimic sparsity-promoting behavior of the ℓ_p-norm. Since in practice the true $\mathbf{d}(n)$ is obviously not known, the weights $\hat{w}(n)$ are usually computed using the estimated $\hat{\mathbf{d}}(n)$ from the previous iteration, cf. (4.9). Alternatively, the weights can also be computed using the average PSD of the desired speech signal, i.e.,

$$\hat{w}(n) = \left(\frac{1}{M} \|\hat{\boldsymbol{\sigma}}_d(n)\|_2^2 + \varepsilon_{\min}\right)^{\frac{p}{2}-1}, \tag{5.6}$$

with $\hat{\boldsymbol{\sigma}}_d^2(n) = [\hat{\sigma}_{d,1}^2(n), \ldots, \hat{\sigma}_{d,M}^2(n)]^{\mathsf{T}}$ containing the PSDs of the desired speech signal in all microphones. Alg. 5 describes recursive PSD estimators, where the PSD of the desired speech signal is estimated using an exponential decay model for the late reverberation [74, 75, 77], which requires an estimate of the reverberation time T_{60}. Alternatively, other late reverberant PSD estimators could be employed [115, 116].

Similarly as in [163], an adaptive version of group sparse MCLP, estimating the prediction filter $\mathbf{G}(n)$ for each time frame n, can be derived by incorporating an exponential window in (5.4). This leads to the following optimization problem for estimating the adaptive prediction filter $\mathbf{G}(n)$ at the n-th time frame

$$\hat{\mathbf{G}}(n) = \arg\min_{\mathbf{G}(n)} \sum_{l=1}^{n} \gamma^{n-l} \hat{w}(l) \|\mathbf{d}(l)\|_2^2, \tag{5.7}$$

where $\gamma \in (0,1]$ is the forgetting factor. The prediction filter $\hat{\mathbf{G}}(n)$ in (5.7) can be computed by solving the unconstrained optimization problem

$$\hat{\mathbf{G}}(n) = \arg\min_{\mathbf{G}(n)} F\left(\mathbf{G}(n)\right), \tag{5.8}$$

with $F : \mathbb{C}^{ML_g \times M} \rightarrow \mathbb{R}$ a quadratic cost function equal to

$$F\left(\mathbf{G}(n)\right) = \operatorname{tr}\left[\mathbf{G}^{\mathsf{H}}(n)\hat{\mathbf{Q}}(n)\mathbf{G}(n)\right] - 2\Re\left\{\operatorname{tr}\left[\mathbf{G}^{\mathsf{H}}(n)\hat{\mathbf{R}}(n)\right]\right\}, \tag{5.9}$$

with the matrices $\hat{\mathbf{Q}}(n)$ and $\hat{\mathbf{R}}(n)$ defined as

$$\hat{\mathbf{Q}}(n) = \sum_{l=1}^{n} \gamma^{n-l}\hat{w}(l)\tilde{\mathbf{x}}_{\tau}(l)\tilde{\mathbf{x}}_{\tau}^{\mathsf{H}}(l), \tag{5.10a}$$

$$\hat{\mathbf{R}}(n) = \sum_{l=1}^{n} \gamma^{n-l}\hat{w}(l)\tilde{\mathbf{x}}_{\tau}(l)\mathbf{x}^{\mathsf{H}}(l). \tag{5.10b}$$

The closed-form solution for the prediction filter in (5.8) can hence be written as

$$\hat{\mathbf{G}}(n) = \hat{\mathbf{Q}}^{-1}(n)\hat{\mathbf{R}}(n). \tag{5.11}$$

Since the matrices $\hat{\mathbf{Q}}(n)$ and $\hat{\mathbf{R}}(n)$ in (5.10) are rank-1 perturbations of $\gamma\hat{\mathbf{Q}}(n-1)$ and $\gamma\hat{\mathbf{R}}(n-1)$, the matrix inversion lemma can be used to obtain a variant of the recursive least squares (RLS) algorithm [247], as given in Alg. 4. The computational complexity of Alg. 4 is quadratic in the number of prediction filter coefficients per channel, with $\mathcal{O}\left(M^2L_g^2\right)$ operations.

Alg. 4 Adaptive group sparse MCLP-based speech dereverberation.

```
parameters: forgetting factor $\gamma$, shape parameter $p$, regularization parameter $\varepsilon_{\min}$
input: $\mathbf{x}(n)$, $\hat{\mathbf{G}}(n-1)$, $\hat{\mathbf{Q}}^{-1}(n-1)$
  compute $\hat{w}(n)$                                   ▷ equation (5.6) and Alg. 5
  $\hat{\mathbf{k}}(n) = \frac{\hat{\mathbf{Q}}^{-1}(n-1)\tilde{\mathbf{x}}_{\tau}(n)}{\frac{\gamma}{\hat{w}(n)} + \tilde{\mathbf{x}}_{\tau}^{\mathsf{H}}(n)\hat{\mathbf{Q}}^{-1}(n-1)\tilde{\mathbf{x}}_{\tau}(n)}$
  $\hat{\mathbf{G}}(n) = \hat{\mathbf{G}}(n-1) + \hat{\mathbf{k}}(n)\left[\mathbf{x}(n) - \hat{\mathbf{G}}^{\mathsf{H}}(n-1)\tilde{\mathbf{x}}_{\tau}(n)\right]^{\mathsf{H}}$
  $\hat{\mathbf{Q}}^{-1}(n) = \frac{1}{\gamma}\left[\mathbf{I} - \hat{\mathbf{k}}(n)\tilde{\mathbf{x}}_{\tau}^{\mathsf{H}}(n)\right]\hat{\mathbf{Q}}^{-1}(n-1)$
  $\hat{\mathbf{u}}(n) = \hat{\mathbf{G}}^{\mathsf{H}}(n)\tilde{\mathbf{x}}_{\tau}(n)$
output: $\hat{\mathbf{u}}(n)$, $\hat{\mathbf{G}}(n)$, $\hat{\mathbf{Q}}^{-1}(n)$
  $\hat{\mathbf{d}}(n) = \mathbf{x}(n) - \hat{\mathbf{u}}(n)$
```

Alg. 5 Recursive PSD estimation. All operations applied element-wise.

```
parameters: smoothing constant $\alpha$, duration of the early part $T_d$ (seconds) and
    $n_d$ (frames), decay constant $\Delta = \frac{3 \ln 10}{T_{60}/T_d}$
input: $\mathbf{x}(n)$
    $\hat{\boldsymbol{\sigma}}_x^2(n) = \alpha \, \hat{\boldsymbol{\sigma}}_x^2(n-1) + (1-\alpha) \, |\mathbf{x}(n)|^2$
    $\hat{\boldsymbol{\sigma}}_r^2(n) = e^{-2\Delta} \, \hat{\boldsymbol{\sigma}}_x^2(n - n_d)$
    $\hat{\boldsymbol{\sigma}}_d^2(n) = \alpha \, \hat{\boldsymbol{\sigma}}_d^2(n-1) + (1-\alpha) \max\left(|\mathbf{x}(n)|^2 - \hat{\boldsymbol{\sigma}}_r^2(n), 0\right)$
output: $\hat{\boldsymbol{\sigma}}_r(n), \hat{\boldsymbol{\sigma}}_d(n)$
```

5.3 Constrained adaptive group sparse MCLP

For dynamic scenarios, e.g., with a moving speaker or multiple speakers taking turns, the ATFs between the active speaker and the microphones inevitably vary over time. In such a scenario, the variations in the ATFs should be tracked, and small values of the forgetting factor γ are generally preferred for quick tracking. However, small values of the forgetting factor result in a prediction error that approaches zero [247]. Since the output signal $\hat{\mathbf{d}}(n)$ in (5.1) is equal to the prediction error, this may result in overestimation of the undesired signal $\hat{\mathbf{u}}(n)$ and excessive cancellation of the speech signal [163]. Similarly, small values of the forgetting factor may lead to ill-conditioning of the matrix $\hat{\mathbf{Q}}(n)$ in (5.10), resulting in an unstable output [163].

To prevent overestimation of the undesired signal, we propose to incorporate prior knowledge about the undesired reverberation. More specifically, we propose to constrain the power of the MCLP-based estimate of the undesired signal power by an estimate of the PSD of the late reverberation based on the temporal exponential decay model, leading to the following optimization problem for estimating the constrained prediction filter $\check{\mathbf{G}}(n)$

$$\begin{aligned} \check{\mathbf{G}}(n) = \arg \min_{\mathbf{G}(n)} \quad & F\left(\mathbf{G}(n)\right) \\ \text{subject to} \quad & \left|\mathbf{G}^{\mathsf{H}}(n)\tilde{\mathbf{x}}_{\tau}(n)\right|^2 \leq \hat{\boldsymbol{\sigma}}_u^2(n). \end{aligned} \tag{5.12}$$

The vector $\hat{\boldsymbol{\sigma}}_u(n) = [\hat{\sigma}_{u,1}(n), \ldots, \hat{\sigma}_{u,M}(n)]^{\mathsf{T}}$ contains the bounds for the undesired speech signal power, and is defined as

$$\hat{\boldsymbol{\sigma}}_u(n) = \min\left(\hat{\boldsymbol{\sigma}}_r(n), |\mathbf{x}(n)|\right) \tag{5.13}$$

with $\hat{\boldsymbol{\sigma}}_r(n)$ the late reverberant PSD estimate, e.g., estimated using Alg. 5. By using the constrained optimization problem in (5.12) instead of the unconstrained optimization problem in (5.8), we aim to prevent the excessive speech cancellation for small values of the forgetting factor γ, while not significantly deteriorating the performance for large values of the forgetting factor γ.

The constrained optimization problem in (5.12) does not have a closed-form solution such that we need to resort to an iterative algorithm. By introducing a splitting variable $\mathbf{z}(n) \in \mathbb{C}^M$ [246], (5.12) can be rewritten as

$$\begin{aligned} \min_{\mathbf{G}(n),\mathbf{z}(n)} \quad & F\left(\mathbf{G}(n)\right) + C\left(\mathbf{z}(n)\right) \\ \text{subject to} \quad & \mathbf{G}^{\mathsf{H}}(n)\tilde{\mathbf{x}}_{\tau}(n) = \mathbf{z}(n), \end{aligned} \tag{5.14}$$

where the inequality constraint in (5.12) is replaced with a convex barrier function $C : \mathbb{C}^M \rightarrow \bar{\mathbb{R}}$, which is defined as $C\left(\mathbf{z}(n)\right) = 0$ when the constraint is satisfied, i.e., $|z_m(n)| \leq \hat{\sigma}_{u,m}(n)$ for all m, and $C\left(\mathbf{z}(n)\right) = \infty$ otherwise. Since $F(.)$ and $C(.)$ are convex functions, the optimization problem in (5.14) can be solved efficiently by applying the ADMM algorithm [246] (cf. Appendix B.3). The augmented Lagrangian for the optimization problem in (5.14) can be written as

$$\begin{aligned} L_{\rho}\left(\mathbf{G}(n), \mathbf{z}(n), \boldsymbol{\mu}\right) = F\left(\mathbf{G}(n)\right) + C\left(\mathbf{z}(n)\right) + & \\ + \frac{\rho}{2}\left\|\mathbf{G}^{\mathsf{H}}(n)\tilde{\mathbf{x}}_{\tau}(n) - \mathbf{z}(n) - \boldsymbol{\mu}\right\|_2^2 - \frac{\rho}{2}\left\|\boldsymbol{\mu}\right\|_2^2, & \end{aligned} \tag{5.15}$$

where ρ is a penalty parameter and $\boldsymbol{\mu}$ is the so-called dual variable [246]. The ADMM algorithm proceeds by minimizing $L_{\rho}(.)$ alternately with respect to $\mathbf{G}(n)$ and $\mathbf{z}(n)$ followed by an ascent over $\boldsymbol{\mu}$ [248], i.e., in the j-th iteration we have

$$\check{\mathbf{G}}^j(n) \leftarrow \arg\min_{\mathbf{G}} F\left(\mathbf{G}\right) + \frac{\rho}{2}\left\|\mathbf{G}^{\mathsf{H}}\tilde{\mathbf{x}}_{\tau}(n) - \left(\check{\mathbf{z}}^{j-1}(n) + \boldsymbol{\mu}^{j-1}\right)\right\|_2^2, \tag{5.16a}$$

$$\check{\mathbf{z}}^j(n) \leftarrow \arg\min_{\mathbf{z}} C\left(\mathbf{z}\right) + \frac{\rho}{2}\left\|\mathbf{z} - \left(\left(\check{\mathbf{G}}^j(n)\right)^{\mathsf{H}}\tilde{\mathbf{x}}_{\tau}(n) - \boldsymbol{\mu}^{j-1}\right)\right\|_2^2, \tag{5.16b}$$

$$\boldsymbol{\mu}^j \leftarrow \boldsymbol{\mu}^{j-1} + \check{\mathbf{z}}^j(n) - \left(\check{\mathbf{G}}^j(n)\right)^{\mathsf{H}}\tilde{\mathbf{x}}_{\tau}(n). \tag{5.16c}$$

Since the function $F(.)$ is quadratic, the constrained prediction filter $\check{\mathbf{G}}^j(n)$ in (5.16a) can be computed using a closed-form expression, similarly to (5.11), as

$$\check{\mathbf{G}}^j(n) \leftarrow \check{\mathbf{Q}}^{-1}(n)\check{\mathbf{R}}^j(n), \tag{5.17}$$

where the matrices $\check{\mathbf{Q}}(n)$ and $\check{\mathbf{R}}^j(n)$ are defined as

$$\check{\mathbf{Q}}(n) = \hat{\mathbf{Q}}(n) + \frac{\rho}{2}\tilde{\mathbf{x}}_{\tau}(n)\tilde{\mathbf{x}}_{\tau}^{\mathsf{H}}(n), \tag{5.18a}$$

$$\check{\mathbf{R}}^j(n) = \hat{\mathbf{R}}(n) + \frac{\rho}{2}\tilde{\mathbf{x}}_{\tau}(n)\left(\check{\mathbf{z}}^{j-1}(n) + \boldsymbol{\mu}^{j-1}\right)^{\mathsf{H}}. \tag{5.18b}$$

Since the matrices $\check{\mathbf{Q}}(n)$ and $\check{\mathbf{R}}^j(n)$ in (5.18) are rank-1 perturbations of $\hat{\mathbf{Q}}(n)$ and $\hat{\mathbf{R}}(n)$ in (5.10), the matrix inversion lemma can be used to obtain an RLS-like algorithm for updating the constrained prediction filter $\check{\mathbf{G}}^j(n)$ [247]. Computing

$\check{\mathbf{z}}^j(n)$ in (5.16b) corresponds to a projection on the constraint set, i.e., clipping of the magnitudes, and can be written element-wise as

$$\check{z}_m^j(n) \leftarrow \min\left(\frac{\hat{\sigma}_{u,m}(n)}{\left|\check{u}_m^j(n) - \mu_m^{j-1}\right|}, 1\right)\left(\check{u}_m^j(n) - \mu_m^{j-1}\right), \tag{5.19}$$

where $\check{\mathbf{u}}^j(n) = \left(\check{\mathbf{G}}^j(n)\right)^{\mathsf{H}} \tilde{\mathbf{x}}_\tau(n)$ is the undesired signal estimated using linear filtering with the constrained MIMO prediction filter $\check{\mathbf{G}}^j(n)$ at j-th iteration.

The complete iterative procedure of the ADMM algorithm for solving the constrained optimization problem in (5.12) is given in Alg. 6. The iterative updates in the ADMM algorithm in Alg. 6 can be interpreted as an iterative correction of the unconstrained filter $\hat{\mathbf{G}}(n)$ to obtain a constrained filter $\check{\mathbf{G}}(n)$ which satisfies the inequality constraint in (5.12). Note that the equality constraint in (5.14) will be satisfied as the number of iterations becomes arbitrarily large, i.e., $j \to \infty$ [246]. However, for a relatively small number of iterations, $\check{\mathbf{u}}^j(n)$ and $\check{\mathbf{z}}^j(n)$ will not necessarily be equal, and only the latter will definitely satisfy the inequality constraint in (5.12). Nevertheless, it is possible to use either $\check{\mathbf{u}}^J(n)$, or the splitting variable $\check{\mathbf{z}}^J(n)$ as an estimate of the undesired signal for dereverberation, leading to two variants of the dereverberation algorithm.

The complexity of Alg. 6 is quadratic and is dominated by the computation of the gain vector $\check{\mathbf{k}}(n)$ with $\mathcal{O}\left(M^2 L_g^2\right)$ operations, equivalent to the computation of the gain vector in Alg. 4, with the additional complexity of the iterative updates being $\mathcal{O}\left(JM^2 L_g\right)$.

Alg. 6 ADMM algorithm for the constrained problem in (5.12). Operations in step 7 are applied element-wise.

```
parameters: penalty parameter $\rho$, number of iterations $J$
input: $\mathbf{x}(n)$, $\hat{\mathbf{G}}(n)$, $\hat{\mathbf{u}}(n)$, $\hat{\mathbf{Q}}^{-1}(n)$ estimated using Alg. 4, $\hat{\boldsymbol{\sigma}}_r(n)$ estimated using Alg. 5
 1: initialize: $\check{\mathbf{z}}^0(n) = \mathbf{0}, \boldsymbol{\mu}^0(n) = \mathbf{0}$
 2: $\check{\mathbf{k}}(n) = \frac{\hat{\mathbf{Q}}^{-1}(n)\tilde{\mathbf{x}}_\tau(n)}{\frac{2}{\rho} + \tilde{\mathbf{x}}_\tau^{\mathsf{H}}(n)\hat{\mathbf{Q}}^{-1}(n)\tilde{\mathbf{x}}_\tau(n)}$
 3: $\hat{\boldsymbol{\sigma}}_u(n) = \min\left(\hat{\boldsymbol{\sigma}}_r(n), |\mathbf{x}(n)|\right)$
 4: for $j = 1, \ldots, J$ do
 5:     $\check{\mathbf{G}}^j(n) \leftarrow \hat{\mathbf{G}}(n) + \check{\mathbf{k}}(n)\left[\check{\mathbf{z}}^{j-1}(n) + \boldsymbol{\mu}^{j-1} - \hat{\mathbf{u}}(n)\right]^{\mathsf{H}}$
 6:     $\check{\mathbf{u}}^j(n) \leftarrow \left(\check{\mathbf{G}}^j(n)\right)^{\mathsf{H}} \tilde{\mathbf{x}}_\tau(n)$
 7:     $\check{\mathbf{z}}^j(n) \leftarrow \min\left(\frac{\hat{\boldsymbol{\sigma}}_u(n)}{|\check{\mathbf{u}}^j(n) - \boldsymbol{\mu}^{j-1}|}, 1\right)\left(\check{\mathbf{u}}^j(n) - \boldsymbol{\mu}^{j-1}\right)$
 8:     $\boldsymbol{\mu}^j \leftarrow \boldsymbol{\mu}^{j-1} + \check{\mathbf{z}}^j(n) - \check{\mathbf{u}}^j(n)$
 9: end for
output: $\check{\mathbf{u}}^J(n), \mathbf{z}^J(n)$
10: $\hat{\mathbf{d}}(n) = \mathbf{x}(n) - \check{\mathbf{u}}^J(n)$                    ▷ u-variant
11: $\hat{\mathbf{d}}(n) = \mathbf{x}(n) - \check{\mathbf{z}}^J(n)$                    ▷ z-variant
```

5.4 Complexity reduction using diagonal approximation

For some applications, the computational complexity of the obtained RLS-based algorithms Algs. 4 and 6 may be prohibitively high, and a more efficient algorithm should be used. A relatively straightforward way to reduce the computational complexity would be to replace the RLS algorithm with an algorithm with linear complexity, such as the least mean squares (LMS) algorithm or its normalized variant (NLMS) [247]. However, in our initial experiments these algorithms did not perform well, presumably since they use an instantaneous approximation of the correlation matrix [247] and essentially do not exploit the temporal information for estimating the correlation matrix.

By inspecting the unconstrained RLS algorithm in Alg. 4 and the constrained ADMM algorithm in Alg. 6, it can be observed that the computation of the gain vectors $\hat{\mathbf{k}}(n)$ and $\check{\mathbf{k}}(n)$, respectively, dominates the computational complexity. In both cases, the number of operations required to compute the gain vector is quadratic in the number of the prediction coefficients per channel, i.e., $\mathcal{O}\left(M^2L_g^2\right)$, due to the computation of the matrix-vector product $\hat{\mathbf{Q}}^{-1}(n)\tilde{\mathbf{x}}_\tau(n)$ and the update of the matrix $\hat{\mathbf{Q}}^{-1}(n)$.

In order to reduce the computational complexity, we propose to use a diagonal approximation of the matrix $\hat{\mathbf{Q}}^{-1}(n)$, i.e., to track only the vector $\hat{\mathbf{q}}^{-1}(n) \in \mathbb{C}^{ML_g}$, corresponding to the diagonal of the matrix $\hat{\mathbf{Q}}^{-1}(n)$, similarly as in frequency-domain adaptive filtering [249, 250]. Based on the update for the matrix $\hat{\mathbf{Q}}^{-1}(n)$ in Alg. 4, the vector $\hat{\mathbf{q}}^{-1}(n)$ can be updated as

$$\hat{\mathbf{q}}^{-1}(n) = \frac{1}{\gamma}\left[\mathbf{1} - \hat{\mathbf{k}}(n) \odot \tilde{\mathbf{x}}_\tau^*(n)\right] \odot \hat{\mathbf{q}}^{-1}(n-1), \tag{5.20}$$

where $\mathbf{1}$ is an ML_g-dimensional vector of ones, and $\odot$ denotes element-wise multiplication. Using this diagonal approximation, the gain vector $\hat{\mathbf{k}}(n)$ in Alg. 4 can be computed as

$$\hat{\mathbf{k}}(n) = \frac{\hat{\mathbf{q}}^{-1}(n-1) \odot \tilde{\mathbf{x}}_\tau(n)}{\frac{\gamma}{w(n)} + \tilde{\mathbf{x}}_\tau^{\mathsf{H}}(n)\left[\hat{\mathbf{q}}^{-1}(n-1) \odot \tilde{\mathbf{x}}_\tau(n)\right]}. \tag{5.21}$$

The gain vector $\check{\mathbf{k}}(n)$ in Alg. 6 can be computed similarly to (5.21).

In effect, this diagonal approximation reduces the computational complexity of the operations involving the matrix $\hat{\mathbf{Q}}^{-1}(n)$ from quadratic to linear, i.e., $\mathcal{O}\left(ML_g\right)$. Therefore, the overall complexity for the unconstrained Alg. 4 is reduced from $\mathcal{O}\left(M^2L_g^2\right)$ to $\mathcal{O}\left(M^2L_g\right)$, while the overall complexity for the constrained Alg. 6 is reduced from $\mathcal{O}\left(M^2L_g^2\right)$ to $\mathcal{O}\left(JM^2L_g\right)$.

5.5 Simulations

In this section, the dereverberation performance of the proposed adaptive sparse MCLP-based methods is investigated. More specifically, the performance of the adaptive unconstrained method (denoted as ADA) given in Alg. 4, the two proposed variants of the constrained adaptive method (denoted as cADA-u and cADA-z) given in Alg. 6, and their variants with the diagonal approximation is evaluated.

The considered acoustic scenarios and the implementation details are outlined in Section 5.5.1. The influence of the forgetting factor on the performance of the considered methods is investigated in Section 5.5.2. The influence of the filter length on the performance is investigated in Section 5.5.3. The performance using a real recording of a moving speaker is investigated in Section 5.5.4.

5.5.1 *Acoustic scenario and algorithmic setup*

For the experimental results we consider two acoustic scenarios: a speech source alternating between two positions (Sections 5.5.2 and 5.5.3) and a moving human speaker (Section 5.5.4).

For the alternating speech source, we use the same setup from the REVERB challenge [22, 23] as used for simulations in Chapters 3 and 4. We use $M \in \{2, 4\}$ microphones with a source-microphone distance of approximately 2 m, where the speech source is alternating between two positions located 45° to the left and 45° to the right of the center of the array. The room has a reverberation time $T_{60} \approx 700$ ms and the sampling frequency is $f_s = 16$ kHz. The clean speech consisted of 6 utterances, and the microphone signal with a total length of approximately 27 s has been generated by alternating the source position for each utterance.

For the moving speaker, we use a recording with $M = 2$ microphones, with the distance between the microphones approximately 11 cm [139]. The signals have been recorded in a room with $T_{60} \approx 750$ ms and contain some background noise, at a reverberant signal-to-noise ratio (RSNR)[1] of approximately 20 dB (cf. [139]).

Similarly to the parameter setup for the batch method in Chapter 4, the analysis and synthesis STFT is computed using a tight window based on a 64 ms Hamming window with a 16 ms window shift. The prediction delay in (5.2) is set to $\tau = 2$ for all experiments, and the filter length is set to $L_g = 25$ for $M = 2$ microphones and $L_g = 10$ for $M = 4$ microphones. The PSDs of the desired speech signal and the late reverberation are estimated using Alg. 5 with $\alpha = 0.65$, $T_d = 50$ ms, and T_{60} is assumed to be known. For Alg. 6, the ADMM penalty parameter is set to $\rho = 10^3$ and the number of iterations is $J = 25$. The weights are regularized with $\varepsilon_{\min} = 10^{-8}$ for subbands. The forgetting factor γ is varied between 0.75 and 0.999. The estimate of the prediction filter $\hat{\mathbf{G}}(n)$ is initialized with zeros and the matrix $\hat{\mathbf{Q}}^{-1}(n)$ is initialized as the identity matrix. Before processing the signal under investigation, an additional 5 s signal has been processed to reduce the influence of initialization.

The dereverberation performance is evaluated in terms of the instrumental measures described in Section 2.3, i.e., the improvement in fwsSNR (ΔfwsSNR) and PESQ (ΔPESQ). For the simulated signals, the reference signal used for the instrumental

---

[1] Reverberant signal-to-noise ratio is defined as the ratio of the power of the noiseless reverberant signal and the power of the noise signal.

measures is the direct speech signal at the first microphone, obtained by convolving the anechoic speech signal with the direct component of the corresponding RIR. For the recorded signals, the reference signal used for the instrumental measures is the close-talking speech signal.

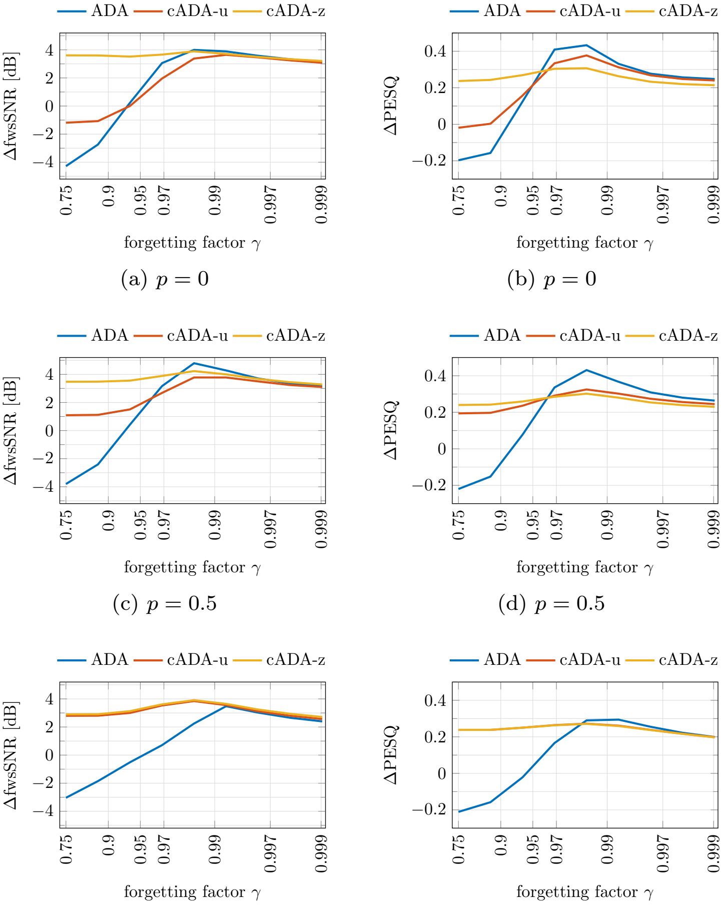
5.5.2 *Influence of the forgetting factor γ and the shape parameter p*

In this section, we investigate the influence of the forgetting factor γ and the shape parameter p on the performance of the adaptive MCLP methods for the simulated alternating speaker scenario.

Fig. 5.1 depicts the improvements the considered performance measures for the adaptive MCLP methods (without the diagonal approximation) with $M = 2$ microphones. On the one hand, the performance of the unconstrained ADA method depends strongly on the forgetting factor γ, even resulting in a significant performance degradation with respect to the microphone signal. This effect is very noticeable for small values of the forgetting factor γ, due to overestimation of the undesired reverberant signal and excessive cancellation of the desired speech signal. It can also be observed that ADA performs similarly for $p = 0$ and $p = 0.5$, with $p = 1$ resulting in a decreased performance. On the other hand, although the constrained cADA-u and cADA-z methods achieve a somewhat lower best-case performance than the optimally tuned ADA, both constrained methods are much more robust with respect to the value of the forgetting factor γ and the shape parameter p. By comparing cADA-u and cADA-z, it can be observed that the latter is more robust to the value of the forgetting factor than the former, since the constraint in (5.12) is always satisfied for the z-variant.

Fig. 5.2 depicts improvements of the considered performance measures for the adaptive MCLP methods using the diagonal approximation with $M = 2$ microphones. In general, the diagonal approximation results in a decreased performance, in the best case achieving approximately 1 dB less in ΔfwsSNR and 0.1 in ΔPESQ than the adaptive methods without the approximation. Furthermore, the performance of both unconstrained and the constrained methods depends on the forgetting factor. As opposed to the methods without the diagonal approximation (cf. Fig. 5.1), it can be observed that for small values of γ the constrained d-cADA-u and d-cADA-z methods result in no improvement with respect to the microphone signal, whereas the unconstrained d-ADA method results in deterioration.

Fig. 5.3 depicts the improvements of the considered performance measures for the adaptive MCLP methods (without diagonal approximation) with $M = 4$ microphones. Overall, the obtained performance with $M = 4$ is better than with $M = 2$. Similarly as for $M = 2$, the performance of the unconstrained ADA method depends strongly on the forgetting factor γ, with a significant performance degradation with respect to the microphone signal for small values of the forgetting factor γ. Although the constrained cADA-u and cADA-z methods achieve a somewhat lower best-case performance than the optimally tuned ADA, both constrained methods are much more robust with respect to the value of the forgetting factor γ.



(a) $p = 0$

(b) $p = 0$

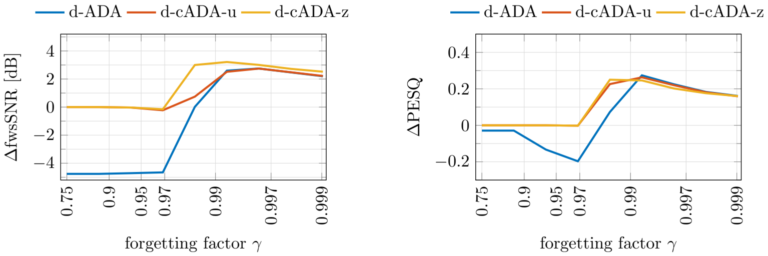
(c) $p = 0.5$

(d) $p = 0.5$

(e) $p = 1$

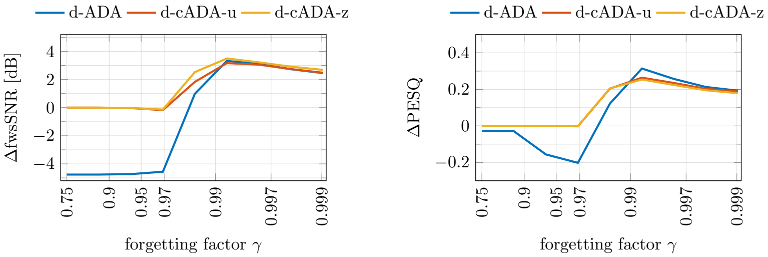
(f) $p = 1$

Fig. 5.1: Performance of the adaptive methods without diagonal approximation (ADA, cADA-u, cADA-z) with $M = 2$ and $L_g = 25$ in terms of ΔfwsSNR (left) and ΔPESQ (right) for different values of the forgetting factor γ and shape parameter p.



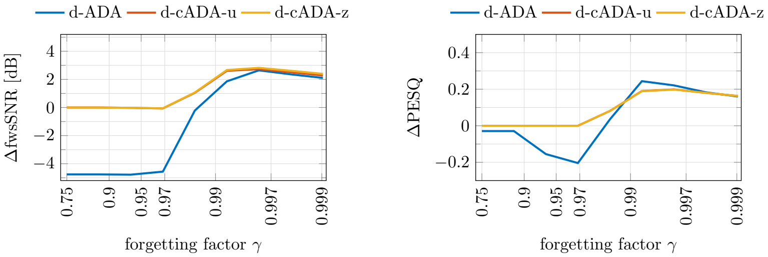
(a) $p = 0$

(b) $p = 0$



(c) $p = 0.5$

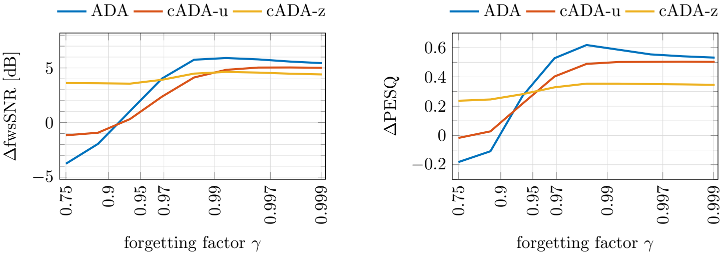
(d) $p = 0.5$



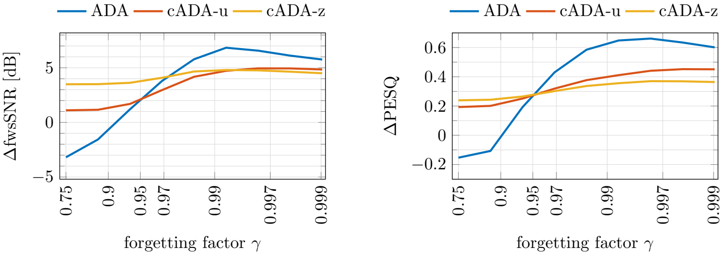
(e) $p = 1$

(f) $p = 1$

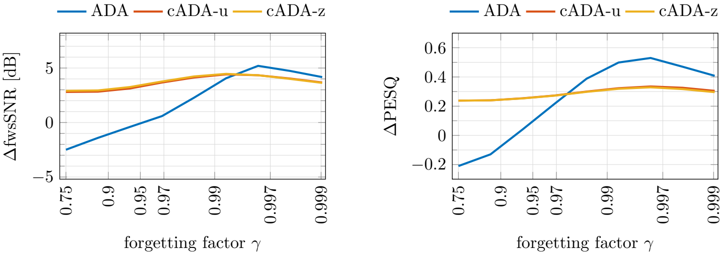
Fig. 5.2: Performance of the adaptive methods with diagonal approximation (d-ADA, d-cADA-u, d-cADA-z) with $M = 2$ and $L_g = 25$ in terms of ΔfwsSNR (left) and ΔPESQ (right) for different values of the forgetting factor γ and shape parameter p.



(a) $p = 0$ (b) $p = 0$



(c) $p = 0.5$ (d) $p = 0.5$



(e) $p = 1$ (f) $p = 1$

Fig. 5.3: Performance of the adaptive methods without the diagonal approximation (ADA, cADA-u, cADA-z) with $M = 4$ and $L_g = 10$ in terms of ΔfwsSNR (left) and ΔPESQ (right) for different values of the forgetting factor γ and shape parameter p.

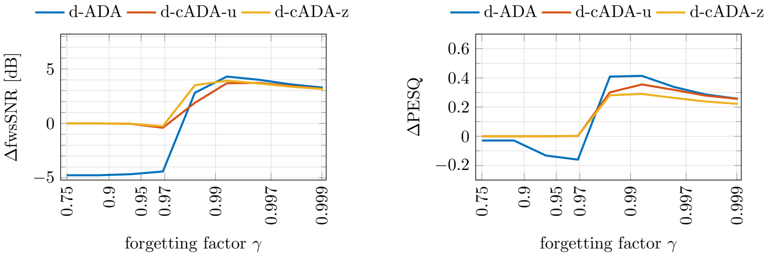
Table 5.1: Real-time factors of the adaptive algorithms.

	ADA	cADA	d-ADA	d-cADA
$M = 2$, $L_g = 25$	1.86	4.04	0.11	1.96
$M = 4$, $L_g = 10$	1.23	4.34	0.15	2.97

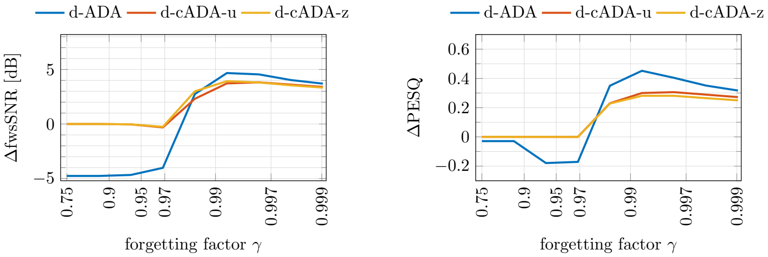
Fig. 5.4 depicts improvements of the considered performance measures for the adaptive MCLP methods using diagonal approximation with $M = 4$ microphones. Similarly as for $M = 2$, the diagonal approximation results in a decreased performance, in the best case achieving approximately 1 dB less in ΔfwsSNR and 0.1 in ΔPESQ than the adaptive methods without the approximation. Also, it can be observed that for small values of γ the constrained d-cADA-u and d-cADA-z methods result in no improvement, whereas the unconstrained d-ADA method results in deterioration, as for $M = 2$.

To investigate the relative computational complexity between the different adaptive methods (with and without diagonal approximation), Table 5.1 contains the real-time factors (RTFs) for the considered adaptive methods, measured on a Windows 7 machine with a 3.4 GHz CPU and algorithms running in MATLAB 2015b. It can be observed that for the considered setup (i.e., number of microphones M, filter length L_g, and number of ADMM iterations J), the constrained methods have about two times larger RTFs than the unconstrained ADA method. This is in line with the expected computational complexity, which states that the RLS iteration in ADA (cf. Alg. 4) and the ADMM algorithm in cADA (cf. Alg. 6) have the same complexity for the considered setup. On the one hand, the diagonal approximation results in an approximately ten times lower RTF than the the unconstrained ADA, which might make a large difference in practice. On the other hand, the diagonal approximation results in an approximately 1.5-2 times lower RTF than the constrained cADA method. In this case, the overall RTF is mainly determined by the number of ADMM iterations J in Alg. 6, and hence the RTF could be further linearly decreased by decreasing the number of ADMM iterations. In our simulations (not depicted here), we noticed that using a smaller number of iterations, e.g., around 10, typically did not result in a significantly decreased performance.

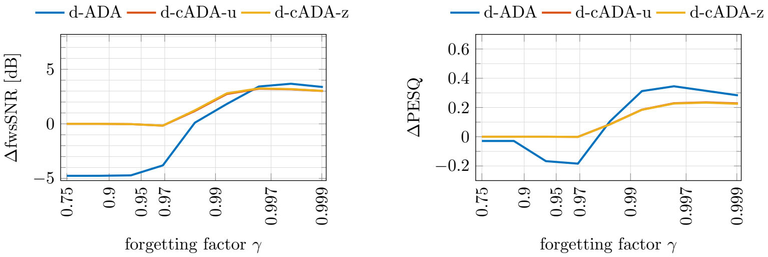
In summary, the performance of the unconstrained ADA method strongly depends on the value of the forgetting factor γ, even resulting in a large performance degradation for small values of the forgetting factor. The constrained methods resolve this issue at the expense of some performance degradation compared to the optimally tuned ADA and a higher computational complexity. It has been observed that the constrained c-ADA-z method is more robust with respect to the forgetting factor than the c-ADA-u, since the inequality constraint in (5.12) is always satisfied for the former. Furthermore, the diagonal approximation results in some performance degradation for all considered methods, with the advantage of a significantly reduced computational complexity.



(a) $p = 0$ (b) $p = 0$



(c) $p = 0.5$ (d) $p = 0.5$



(e) $p = 1$ (f) $p = 1$

Fig. 5.4: Performance of the adaptive methods with diagonal approximation (d-ADA, d-cADA-u, d-cADA-z) with $M = 4$ and $L_g = 10$ in terms of ΔfwsSNR (left) and ΔPESQ (right) for different values of the forgetting factor γ and shape parameter p.

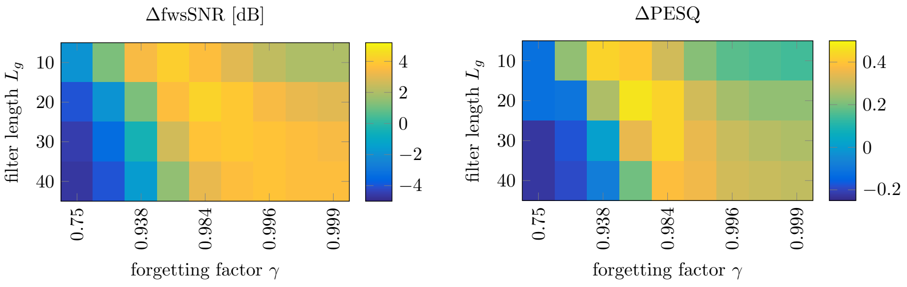
5.5.3 *Influence of the filter length L_g*

In this section, we investigate the influence of the filter length L_g and the forgetting factor γ on the performance of the adaptive MCLP methods for simulated alternating speaker scenario. We have used $M = 2$ microphones and the shape parameter is set to $p = 0$.

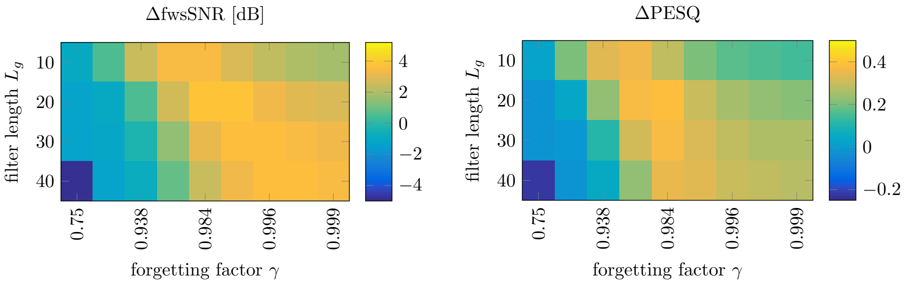
Fig. 5.5 depicts the improvements of the considered performance measures for the adaptive MCLP methods without diagonal approximation. It can be observed that the unconstrained ADA method becomes very sensitive to relatively small forgetting factors as the filter length L_g increases. More specifically, combining a large L_g with a small γ results in significant distortions and possible instability of the output. While cADA-u is less influenced by γ, the performance still drops for small γ, especially for large L_g when the matrix $\hat{\mathbf{Q}}(n)$ becomes ill-conditioned. Finally, it can be observed that cADA-z is quite robust with respect to the filter length and the forgetting factor, since the inequality constraint in (5.12) prevents overestimation and possible instability of the output. Overall, the best-case performance of all methods is very similar in terms of ΔfwsSNR, while the constrained methods yield a somewhat lower ΔPESQ, with approximately 0.1 points lower best-case performance.

5.5.4 *Evaluation using a real recording*

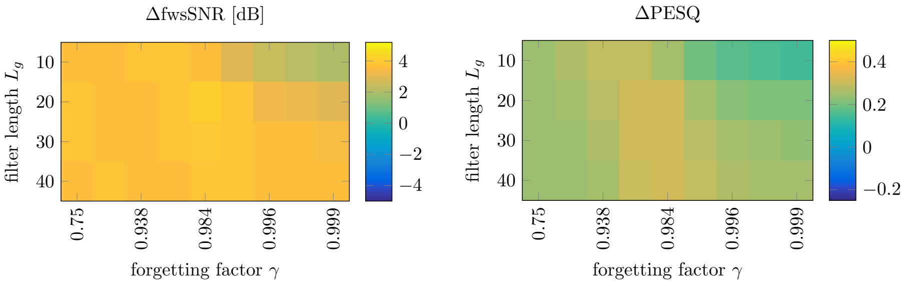
In this section, we investigate the performance of the adaptive methods using a recording of a moving speaker with $M = 2$ microphones, containing some background noise (RSNR ≈ 20 dB). The speaker is naturally moving between different locations in the room. The total length of the used recordings is approximately 42 s, where the speaker is first static and then starts moving around 8 s. The filter length is set to $L_g = 25$, the shape parameter to $p = 0$, and the forgetting factor to $\gamma \in \{0.99, 0.85\}$. To illustrate the temporal dependency of the performance, an excerpt of the frame-wise values of the fwsSNR (smoothed across 15 frames) is shown in Fig. 5.6, while the overall PESQ values are shown in Table 5.2. On the one hand, it can be observed both from fwsSNR and PESQ values that for a relatively large forgetting factor $\gamma = 0.99$ all methods perform similarly. From Fig. 5.6 it can also be observed that all methods result in improvements compared to the microphone signal for the static part and relatively small improvements for the dynamic part. On the other hand, for a relatively small forgetting factor $\gamma = 0.85$ the unconstrained ADA method results in excessive speech cancellation due to overestimation of the undesired signal, resulting in an fwsSNR value of about 0 dB for the complete signal and a significantly reduced PESQ value when compared to the microphone signal. This is also illustrated in the spectrograms of the corresponding signals with $\gamma = 0.85$ in Fig. 5.7. On the contrary, using the smaller forgetting factor $\gamma = 0.85$ with the constrained methods results in a performance improvement for the dynamic part, with cADA-z performing generally better than cADA-u in terms of fwsSNR, and both achieving the same overall performance in terms of PESQ.



(a) ADA (b) ADA



(c) cADA-u (d) cADA-u

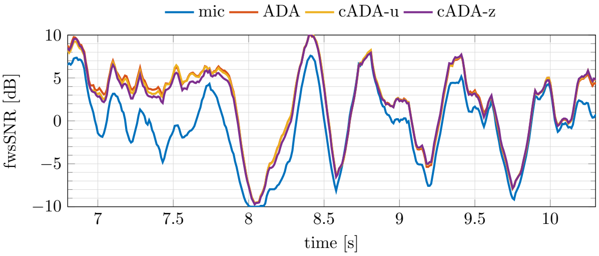


(e) cADA-z (f) cADA-z

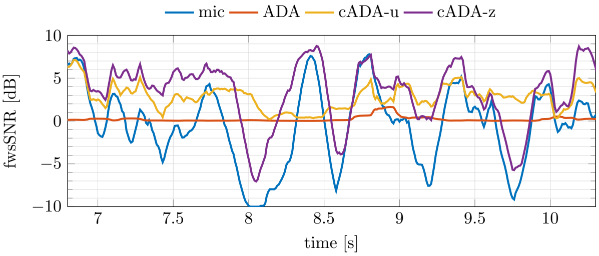
Fig. 5.5: Performance of the adaptive methods without diagonal approximation (ADA, cADA-u, cADA-z) with $p = 0$ for different values of the forgetting factor γ and filter length L_g.

Table 5.2: Overall PESQ values for the microphone signal and the output signal obtained using ADA, cADA-u and cADA-z for a moving speaker with different values of the forgetting factor.

γ	mic	ADA	cADA-u	cADA-z
0.99	1.32	1.49	1.47	1.46
0.85	1.32	1.18	1.43	1.43



(a) $\gamma = 0.99$



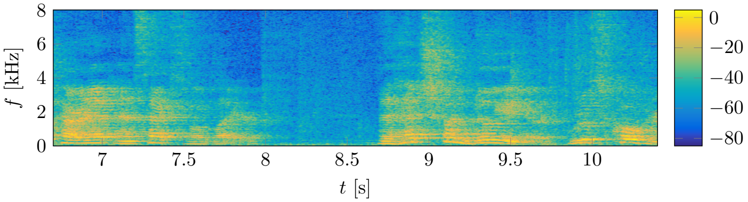
(b) $\gamma = 0.85$

Fig. 5.6: Smoothed fwsSNR for a moving speaker vs. time for the microphone signal and the output signal obtained using ADA, cADA-u and cADA-z with $\gamma = 0.99$ (top) and $\gamma = 0.85$ (bottom). The speaker starts walking around 8 s.

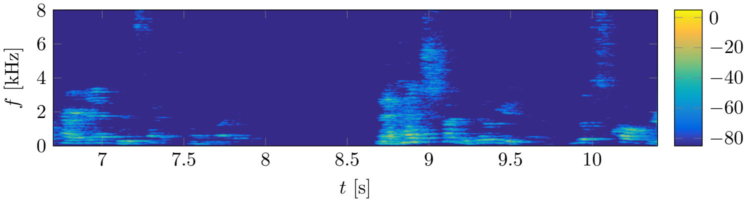
5.6 Summary

In this chapter, we have presented a novel adaptive formulation for group sparse MCLP-based speech dereverberation. To prevent overestimation of the undesired speech signal, leading to speech distortion, we have constrained the power of the MCLP-based estimate of the undesired signal with an estimate of the late reverberant PSD. We have used a late reverberant PSD estimator based on a statistical model requiring a T_{60} estimate, although other estimators could be readily employed. The obtained constrained optimization problem is solved using the ADMM algorithm, resulting in an efficient implementation similar to the RLS algorithm. Moreover, we have proposed a diagonal approximation to reduce the computational complexity of the derived constrained and unconstrained adaptive methods.

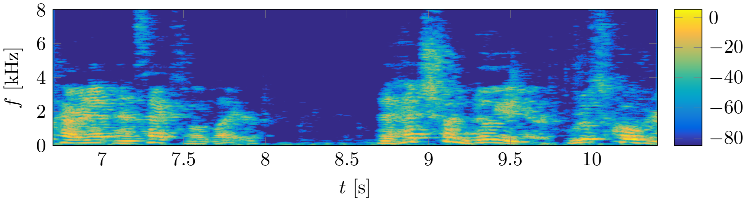
The experimental results show that in comparison to the unconstrained ADA method, both versions of the proposed constrained adaptive methods increase the robustness with respect to the forgetting factor and the filter length, with the cADA-z variant outperforming the cADA-u variant. Hence, the proposed methods can be used to improve the robustness and the performance of MCLP-based adaptive dereverberation in dynamic scenarios, e.g., when the prediction filters need to adapt quickly and the optimal forgetting factor is not known. Furthermore, the evaluation shows that the computational complexity can be significantly decreased by using the diagonal approximation, however at the expense of some loss in the performance.



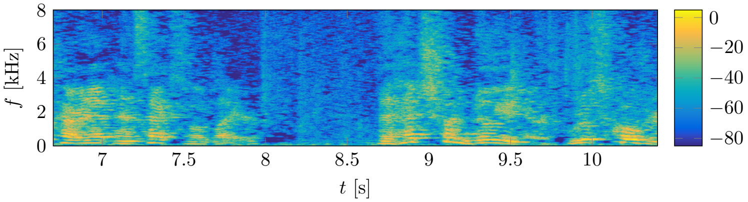
(a) Microphone signal



(b) ADA



(c) cADA-u



(d) cADA-z

Fig. 5.7: Spectrograms of the microphone signal (top) and the output signal obtained using ADA, cADA-u and cADA-z with $\gamma = 0.85$. The speaker starts walking around 8 s.

6

GENERAL FRAMEWORK FOR SPARSITY-BASED SPEECH DEREVERBERATION

In this chapter, we present a general framework for blind speech dereverberation based on the MCLP-based signal model and exploiting sparsity of the desired speech signal in the TF domain. Whereas in Chapters 3, 4 and 5 we have considered a subband MCLP-based signal model in the STFT domain, in this chapter we propose a more general framework for blind speech dereverberation, either using a wideband MCLP-based signal model in the time domain or a subband MCLP-based signal model in the TF domain. We formulate several optimization problems, combining either the wideband or the subband signal model with a sparse analysis or synthesis prior to exploit sparsity of the speech signal coefficients. The obtained optimization problems can again be solved using the ADMM algorithm.

The proposed framework supports general TF transforms by using corresponding analysis/synthesis operators, e.g., the STFT, the ERBlet transform [251], or adaptive non-stationary Gabor transforms [192]. To promote sparsity, we will consider the commonly used weighted ℓ_1- and ℓ_2-norm, although other sparsity-promoting functions can be used in the proposed framework. In addition to locally computed weights for the weighted norms, we also consider structured weights by using a neighborhood in the TF domain or a low-rank approximation of the speech power spectrogram.

The wideband and subband MCLP-based signal models are briefly reviewed in Section 6.1, and analysis and synthesis sparsity are introduced in Section 6.2. Several

---

This chapter is partly based on:

[182] A. Jukić, T. van Waterschoot, T. Gerkmann, S. Doclo, “A general framework for multi-channel speech dereverberation by exploiting sparsity,” in *Proceedings of the AES 60th Conference on DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)*, Leuven, Belgium, Feb. 2016.

[183] A. Jukić, T. van Waterschoot, T. Gerkmann, S. Doclo, “A general framework for incorporating time-frequency domain sparsity in multi-channel speech dereverberation,” *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, pp. 17–30, Jan./Feb. 2017.

optimization problems combining both signal models with analysis or synthesis sparsity are formulated in Sections 6.2–6.5, followed by a discussion on the selection of the sparsity-promoting cost function in Section 6.6. In Sections 6.7 and 6.8 we discuss the relation to existing methods and possible extensions. The performance of the considered methods is evaluated in Section 6.9.

6.1 Problem formulation

We consider an acoustic scenario with a single static speech source captured by M microphones in a reverberant enclosure without the presence of additive noise. As discussed in Section 2.1.2, the subband MCLP-based signal model is given by

$$\mathbf{x}_{\text{ref}}(k) = \mathbf{d}_{\text{ref}}(k) + \tilde{\mathbf{X}}_{\tau}(k)\mathbf{g}_{\text{ref}}(k), \tag{6.1}$$

with $\mathbf{x}_{\text{ref}}(k), \mathbf{d}_{\text{ref}}(k) \in \mathbb{C}^{N}$, $\tilde{\mathbf{X}}_{\tau}(k) \in \mathbb{C}^{N\times ML_g}$, $\mathbf{g}_{\text{ref}}(k) \in \mathbb{C}^{ML_g}$, and ref denoting the reference microphone, ref $\in \{1, \ldots, M\}$. As discussed in Section 2.1.1, the wideband MCLP-based signal model can be written similarly as

$$\underline{\mathbf{x}}_{\text{ref}} = \underline{\mathbf{d}}_{\text{ref}} + \underline{\tilde{\mathbf{X}}}_{\underline{\tau}}\underline{\mathbf{g}}_{\text{ref}}, \tag{6.2}$$

with $\underline{\mathbf{x}}_{\text{ref}}, \underline{\mathbf{d}}_{\text{ref}} \in \mathbb{R}^{T}$, $\underline{\tilde{\mathbf{X}}}_{\underline{\tau}} \in \mathbb{R}^{T\times M\underline{L}_g}$, and $\underline{\mathbf{g}}_{\text{ref}} \in \mathbb{R}^{M\underline{L}_g}$.

As discussed in Section 2.1, the wideband model in (6.2) holds perfectly when the MINT conditions are fulfilled. However, the time-domain prediction filter $\mathbf{g}_{\text{ref}}$ is typically very long, such that dereverberation based on the wideband model in (6.2) can be computationally demanding [154, 183]. In order to reduce the length of the filters, a common approximation is to use the subband model in (6.1), as we have done in Chapters 3–5.

In the following, we formulate blind speech dereverberation as estimation of the desired speech signal at the reference microphone by using a sparsity promoting cost function and either the wideband signal model in (6.2) or the subband signal model in (6.1), assuming batch processing.

6.2 Analysis and synthesis sparsity

Given a batch of T time-domain samples, let $\boldsymbol{\Psi} \in \mathbb{C}^{T\times F}$, with F the number of TF coefficients such that $F > T$, denote the overcomplete linear operator corresponding to a TF transform. In general, many TF transforms of interest can be represented with such an operator $\boldsymbol{\Psi}$, e.g., the STFT, the Gabor transform, the ERBlet transform and general adaptive linear transforms [192, 251, 252]. In the following, we will use $\boldsymbol{\Psi}$ corresponding to the STFT, with N time frames and K subbands and $F = KN$. Furthermore, for simplicity we assume that $\boldsymbol{\Psi}$ is a Parseval tight frame [253], i.e., $\boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathsf{H}} = \mathbf{I}$.

Using the linear operator $\mathbf{\Psi}$, the TF coefficients $\mathbf{d}$ of the time-domain signal $\underline{\mathbf{d}}$ can then be obtained by applying the analysis transform $\mathbf{\Psi}^{\mathsf{H}}$, while the time-domain signal can be recovered by applying the synthesis transform $\mathbf{\Psi}$, i.e,

$$\mathbf{d} = \mathbf{\Psi}^{\mathsf{H}}\underline{\mathbf{d}} \in \mathbb{C}^{KN}, \tag{6.3a}$$

$$\underline{\mathbf{d}} = \mathbf{\Psi}\mathbf{d} \in \mathbb{R}^{T}. \tag{6.3b}$$

The vector $\mathbf{d}$ contains all TF coefficients of the time-domain signal $\underline{\mathbf{d}}$, i.e., its elements are the TF coefficients $d(k,n)$. The subvectors of $\mathbf{d}$ are the vectors of coefficients for individual subbands $\mathbf{d}(k) \in \mathbb{C}^{N}$, $k \in \{1,\dots,K\}$.

Sparsity has been used in various inverse problems in signal processing and machine learning, and has typically been used in the following two paradigms: synthesis sparsity and analysis sparsity [254]. On the one hand, synthesis sparsity is based on the assumption that a signal can be expressed as a linear combination of a relatively small number of elements from a dictionary. In the considered scenario, this would imply that the time-domain desired speech signal $\underline{\mathbf{d}}$ can be represented as a sum of scaled prototype functions contained in the columns of $\mathbf{\Psi}$, corresponding to a relatively small number of estimated TF coefficients, i.e., $\underline{\mathbf{d}} \approx \mathbf{\Psi}\mathbf{d}$ with a sparse $\mathbf{d}$. On the other hand, analysis sparsity is based on the assumption that a signal has a sparse representation when a suitable analysis operator is applied. In the considered scenario, this would imply that the estimated time-domain speech signal $\mathbf{d}$ has a sparse STFT representation, i.e., that $\mathbf{d} = \mathbf{\Psi}^{\mathsf{H}}\underline{\mathbf{d}}$ is sparse. While both paradigms assume sparsity of the TF coefficients, synthesis sparsity leads to estimation of the TF coefficients, while analysis sparsity leads to estimation of the time-domain signal. The paradigms are equivalent only if the analysis operator is equal to the inverse of the synthesis operator [254]. In the considered case this is not fulfilled since the STFT synthesis operator $\mathbf{\Psi}$ is overcomplete (i.e., redundant, since $KN > T$) and thus not invertible, and hence the two paradigms differ. In this context, in Chapters 3–5 we have used synthesis sparsity for MCLP-based dereverberation (cf. Section 6.5).

In the remainder of this chapter, we present different formulations of MCLP-based speech dereverberation exploiting sparsity in the TF domain. In Sections 6.3 and 6.4, we first consider the wideband signal model in (6.2) with the analysis and synthesis sparsity prior, respectively. In Section 6.5, we then consider the subband signal model in (6.1) with the synthesis sparsity prior, which can be considered a generalization of the sparse MCLP method from Chapter 3.

6.3 Wideband model and analysis sparsity

In this section, we consider the wideband model in (6.2) in combination with the analysis sparsity prior. This corresponds to estimating the desired speech signal $\underline{\mathbf{d}}_{\text{ref}}$ in the time domain and enforcing its TF coefficients to be sparse in terms of the sparsity-promoting cost function $P(.)$, leading to the following optimization problem

$$\begin{aligned} \min_{\underline{\mathbf{d}}_{\text{ref}},\underline{\mathbf{g}}_{\text{ref}}} \quad & P\left(\boldsymbol{\Psi}^{\mathsf{H}}\underline{\mathbf{d}}_{\text{ref}}\right) \\ \text{subject to} \quad & \underline{\mathbf{d}}_{\text{ref}} + \underline{\tilde{\mathbf{X}}}_{\mathcal{T}}\underline{\mathbf{g}}_{\text{ref}} = \underline{\mathbf{x}}_{\text{ref}}, \end{aligned} \tag{6.4}$$

By applying the ADMM algorithm (cf. Appendix B.3), the obtained problem can be solved using the following iterative updates

$$\hat{\underline{\mathbf{d}}}^{j}_{\text{ref}} \leftarrow \arg\min_{\underline{\mathbf{d}}} P\left(\boldsymbol{\Psi}^{\mathsf{H}}\underline{\mathbf{d}}\right) + \frac{\rho}{2}\left\|\underline{\mathbf{d}} + \underline{\tilde{\mathbf{X}}}_{\mathcal{T}}\hat{\underline{\mathbf{g}}}^{j-1}_{\text{ref}} - \underline{\mathbf{x}}_{\text{ref}} + \underline{\boldsymbol{\mu}}^{j-1}\right\|_2^2, \tag{6.5a}$$

$$\hat{\underline{\mathbf{g}}}^{j}_{\text{ref}} \leftarrow \arg\min_{\underline{\mathbf{g}}} \left\|\hat{\underline{\mathbf{d}}}^{j}_{\text{ref}} + \underline{\tilde{\mathbf{X}}}_{\mathcal{T}}\underline{\mathbf{g}} - \underline{\mathbf{x}}_{\text{ref}} + \underline{\boldsymbol{\mu}}^{j-1}\right\|_2^2, \tag{6.5b}$$

$$\underline{\boldsymbol{\mu}}^{j} \leftarrow \underline{\boldsymbol{\mu}}^{j-1} + \eta\left(\hat{\underline{\mathbf{d}}}^{j}_{\text{ref}} + \underline{\tilde{\mathbf{X}}}_{\mathcal{T}}\hat{\underline{\mathbf{g}}}^{j}_{\text{ref}} - \underline{\mathbf{x}}_{\text{ref}}\right), \tag{6.5c}$$

where ρ is the penalty parameter, $\underline{\boldsymbol{\mu}}$ is the dual variable and η is a parameter used for faster convergence.

The update for the time-domain signal $\underline{\mathbf{d}}_{\text{ref}}$ in (6.5a) corresponds to a generalized Lasso problem [246] and can be efficiently solved using the ADMM algorithm, as shown in Appendix B.4. The update for the filter $\underline{\mathbf{g}}_{\text{ref}}$ in (6.5b) is a LS problem with closed-form solution given as

$$\hat{\underline{\mathbf{g}}}^{j}_{\text{ref}} \leftarrow \left(\underline{\tilde{\mathbf{X}}}^{\mathsf{T}}_{\mathcal{T}}\underline{\tilde{\mathbf{X}}}_{\mathcal{T}}\right)^{-1}\underline{\tilde{\mathbf{X}}}^{\mathsf{T}}_{\mathcal{T}}\left(\underline{\mathbf{x}}_{\text{ref}} - \hat{\underline{\mathbf{d}}}^{j}_{\text{ref}} - \underline{\boldsymbol{\mu}}^{j-1}\right) = \hat{\underline{\mathbf{g}}}_{\text{ref},\ell_2} - \hat{\underline{\mathbf{g}}}^{j}_{\text{ref,iter}}, \tag{6.6}$$

where

$$\hat{\underline{\mathbf{g}}}_{\text{ref},\ell_2} = \left(\underline{\tilde{\mathbf{X}}}^{\mathsf{T}}_{\mathcal{T}}\underline{\tilde{\mathbf{X}}}_{\mathcal{T}}\right)^{-1}\underline{\tilde{\mathbf{X}}}^{\mathsf{T}}_{\mathcal{T}}\underline{\mathbf{x}}_{\text{ref}}, \tag{6.7}$$

is an iteration-independent term, and

$$\hat{\underline{\mathbf{g}}}^{j}_{\text{ref,iter}} = \left(\underline{\tilde{\mathbf{X}}}^{\mathsf{T}}_{\mathcal{T}}\underline{\tilde{\mathbf{X}}}_{\mathcal{T}}\right)^{-1}\underline{\tilde{\mathbf{X}}}^{\mathsf{T}}_{\mathcal{T}}\left(\hat{\underline{\mathbf{d}}}^{j}_{\text{ref}} + \underline{\boldsymbol{\mu}}^{j-1}\right), \tag{6.8}$$

is an iteration-dependent correction term. The iteration-independent term $\hat{\underline{\mathbf{g}}}_{\text{ref},\ell_2}$ is equal to the closed-form solution when using the ℓ_2-norm as the cost function in (6.4), i.e., $P(.) = \|.\|_2^2$. As shown in Section 3.5.2, filters obtained minimizing the ℓ_2-norm typically do not perform very well for dereverberation. However, similarly as in [255], the iteration-dependent term $\hat{\underline{\mathbf{g}}}_{\text{ref,iter}}$ can be seen as a correction which sparsifies the TF coefficients of the estimated of desired speech signal, which has been shown in Chapter 3 to be crucial for MCLP-based dereverberation. Note that the matrix $\underline{\tilde{\mathbf{X}}}^{\mathsf{T}}_{\mathcal{T}}\underline{\tilde{\mathbf{X}}}_{\mathcal{T}}$ is the same for all iterations, such that it only needs to be factored once and its factorization can be used for solving the corresponding linear system in the subsequent iterations [246]. Moreover, since $\underline{\tilde{\mathbf{X}}}_{\mathcal{T}}$ is a block-convolution matrix, both $\underline{\tilde{\mathbf{X}}}^{\mathsf{T}}_{\mathcal{T}}\underline{\tilde{\mathbf{X}}}_{\mathcal{T}}$ and $\underline{\tilde{\mathbf{X}}}^{\mathsf{T}}_{\mathcal{T}}\underline{\mathbf{x}}_{\text{ref}}$ can be obtained through multi-channel correlation. Additionally, the block-Toeplitz structure of $\tilde{\underline{\mathbf{X}}}_{\mathcal{T}}^{\mathsf{T}}\tilde{\underline{\mathbf{X}}}_{\mathcal{T}}$ can be further exploited to design a fast linear solver, similarly as in [255], but generalized to the multi-channel case.

6.4 Wideband model and synthesis sparsity

In this section, we consider the wideband model in (6.2) but now in combination with the synthesis sparsity prior. This corresponds to estimating the desired speech signal coefficients $\mathbf{d}_{\text{ref}}$ in the TF domain and enforcing them to be sparse in terms of the cost function $P(.)$, leading to the following optimization problem

$$\begin{aligned} \min_{\mathbf{d}_{\text{ref}},\underline{\mathbf{g}}_{\text{ref}}} \quad & P\left(\mathbf{d}_{\text{ref}}\right) \\ \text{subject to} \quad & \boldsymbol{\Psi}\mathbf{d}_{\text{ref}} + \tilde{\underline{\mathbf{X}}}_{\mathcal{T}}\underline{\mathbf{g}}_{\text{ref}} = \underline{\mathbf{x}}_{\text{ref}}. \end{aligned} \tag{6.9}$$

The desired speech signal in the time domain can then be obtained by performing the inverse STFT of the estimated coefficients, i.e., $\hat{\underline{\mathbf{d}}}_{\text{ref}} = \boldsymbol{\Psi}\hat{\mathbf{d}}_{\text{ref}}$. By applying the ADMM algorithm (cf. Appendix B.3), the obtained problem can be solved using the following iterative updates

$$\hat{\mathbf{d}}_{\text{ref}}^{j} \leftarrow \arg\min_{\mathbf{d}} P\left(\mathbf{d}\right) + \frac{\rho}{2}\left\|\boldsymbol{\Psi}\mathbf{d} + \tilde{\underline{\mathbf{X}}}_{\mathcal{T}}\hat{\underline{\mathbf{g}}}_{\text{ref}}^{j-1} - \underline{\mathbf{x}}_{\text{ref}} + \underline{\boldsymbol{\mu}}^{j-1}\right\|_2^2, \tag{6.10a}$$

$$\hat{\underline{\mathbf{g}}}_{\text{ref}}^{j} \leftarrow \arg\min_{\underline{\mathbf{g}}} \left\|\boldsymbol{\Psi}\hat{\mathbf{d}}_{\text{ref}}^{j} + \tilde{\underline{\mathbf{X}}}_{\mathcal{T}}\underline{\mathbf{g}} - \underline{\mathbf{x}}_{\text{ref}} + \underline{\boldsymbol{\mu}}^{j-1}\right\|_2^2, \tag{6.10b}$$

$$\hat{\underline{\boldsymbol{\mu}}}^{j} \leftarrow \underline{\boldsymbol{\mu}}^{j-1} + \eta\left(\boldsymbol{\Psi}\hat{\mathbf{d}}_{\text{ref}}^{j} + \tilde{\underline{\mathbf{X}}}_{\mathcal{T}}\hat{\underline{\mathbf{g}}}_{\text{ref}}^{j} - \underline{\mathbf{x}}_{\text{ref}}\right), \tag{6.10c}$$

where ρ is the penalty parameter, $\underline{\boldsymbol{\mu}}$ is the dual variable and η is a parameter used for faster convergence.

The update for the TF coefficients $\hat{\mathbf{d}}_{\text{ref}}$ in (6.10a) corresponds to a Lasso problem [256], and can be efficiently solved using the iterative shrinkage/thresholding algorithm (ISTA), or using its fast variant (FISTA) [257], as shown in Appendix B.5. Similarly as in (6.6), the update for the prediction filter $\underline{\mathbf{g}}$ is a LS problem with closed-form solution given as

$$\hat{\underline{\mathbf{g}}}_{\text{ref}}^{j} \leftarrow \left(\tilde{\underline{\mathbf{X}}}_{\mathcal{T}}^{\mathsf{T}}\tilde{\underline{\mathbf{X}}}_{\mathcal{T}}\right)^{-1}\tilde{\underline{\mathbf{X}}}_{\mathcal{T}}^{\mathsf{T}}\left(\underline{\mathbf{x}}_{\text{ref}} - \boldsymbol{\Psi}\hat{\mathbf{d}}_{\text{ref}}^{j} - \underline{\boldsymbol{\mu}}^{j-1}\right) = \hat{\underline{\mathbf{g}}}_{\text{ref},\ell_2} - \hat{\underline{\mathbf{g}}}_{\text{ref,iter}}^{j}, \tag{6.11}$$

where $\hat{\underline{\mathbf{g}}}_{\text{ref},\ell_2}$ is the same iteration-independent term as in (6.7), and

$$\hat{\underline{\mathbf{g}}}_{\text{ref,iter}}^{j} = \left(\tilde{\underline{\mathbf{X}}}_{\mathcal{T}}^{\mathsf{T}}\tilde{\underline{\mathbf{X}}}_{\mathcal{T}}\right)^{-1}\tilde{\underline{\mathbf{X}}}_{\mathcal{T}}^{\mathsf{T}}\left(\boldsymbol{\Psi}\mathbf{d}_{\text{ref}}^{j} + \underline{\boldsymbol{\mu}}^{j-1}\right), \tag{6.12}$$

is the iteration-dependent term.

6.5 Subband model

In this section, we consider the subband model in (6.1) in combination with the synthesis prior. Similarly as in Section 6.4, we estimate the desired speech signal coefficients $\mathbf{d}_{\text{ref}}$ in the TF domain and enforce them to be sparse in terms of the cost function $P(.)$. Since the subband model is independent across subbands and assuming that the cost function $P(.)$ is also separable, the speech signal coefficients $\mathbf{d}_{\text{ref}}(k)$ can then be estimated for each subband k independently, leading to the following optimization problem in the k-th subband

$$\begin{aligned} \min_{\mathbf{d}_{\text{ref}}(k),\mathbf{g}_{\text{ref}}(k)} \quad & P\left(\mathbf{d}_{\text{ref}}(k)\right) \\ \text{subject to} \quad & \mathbf{d}_{\text{ref}}(k) + \tilde{\mathbf{X}}_{\tau}(k)\mathbf{g}_{\text{ref}}(k) = \mathbf{x}_{\text{ref}}(k). \end{aligned} \tag{6.13}$$

The desired speech signal in the time domain can then be obtained by performing the inverse STFT of the estimated coefficients, i.e., $\hat{\underline{\mathbf{d}}}_{\text{ref}} = \mathbf{\Psi}\hat{\mathbf{d}}_{\text{ref}}$. Note that this formulation is a generalization of the sparse MCLP based on ℓ_p-norm considered in (3.27) to a general cost function $P(.)$. By applying the ADMM algorithm (cf. Appendix B.3), the obtained problem can be solved in each subband using the following iterative updates

$$\hat{\mathbf{d}}^{j}_{\text{ref}}(k) \leftarrow \arg\min_{\mathbf{d}(k)} P\left(\mathbf{d}(k)\right) + \frac{\rho}{2}\left\|\mathbf{d}(k) + \tilde{\mathbf{X}}_{\tau}(k)\hat{\mathbf{g}}^{j-1}_{\text{ref}}(k) - \mathbf{x}_{\text{ref}}(k) + \boldsymbol{\mu}^{j-1}(k)\right\|_2^2, \tag{6.14a}$$

$$\hat{\mathbf{g}}^{j}_{\text{ref}}(k) \leftarrow \arg\min_{\mathbf{g}_{\text{ref}}(k)} \left\|\hat{\mathbf{d}}^{j}_{\text{ref}}(k) + \tilde{\mathbf{X}}_{\tau}(k)\mathbf{g}_{\text{ref}}(k) - \mathbf{x}_{\text{ref}}(k) + \boldsymbol{\mu}^{j-1}(k)\right\|_2^2, \tag{6.14b}$$

$$\boldsymbol{\mu}^{j}(k) \leftarrow \boldsymbol{\mu}^{j-1}(k) + \eta\left(\hat{\mathbf{d}}^{j}_{\text{ref}}(k) + \tilde{\mathbf{X}}_{\tau}(k)\hat{\mathbf{g}}^{j}_{\text{ref}}(k) - \mathbf{x}_{\text{ref}}(k)\right), \tag{6.14c}$$

where ρ is the penalty parameter, $\boldsymbol{\mu}(k)$ is the dual variable and η is a parameter used for faster convergence.

The update for the TF coefficients $\hat{\mathbf{d}}_{\text{ref}}(k)$ in the k-th subband in (6.14a) corresponds to the proximal operator of the cost function $P(.)$ (cf. Appendix B.2), and can be written as

$$\hat{\mathbf{d}}^{j}_{\text{ref}}(k) \leftarrow \text{prox}^{\rho}_{P}\left(\mathbf{x}_{\text{ref}}(k) - \tilde{\mathbf{X}}_{\tau}(k)\hat{\mathbf{g}}^{j-1}_{\text{ref}}(k) - \boldsymbol{\mu}^{j-1}(k)\right), \tag{6.15}$$

where $\text{prox}^{\rho}_{P}(.)$ is the proximal operator of $P(.)$ as defined in (B.14). Similarly as in (6.6) and (6.11), the update for the prediction filter $\hat{\mathbf{g}}_{\text{ref}}(k)$ in the k-th subband is a LS problem with closed-form solution given as

$$\begin{aligned} \hat{\mathbf{g}}^{j}_{\text{ref}}(k) \leftarrow & \left(\tilde{\mathbf{X}}^{\mathsf{H}}_{\tau}(k)\tilde{\mathbf{X}}_{\tau}(k)\right)^{-1}\tilde{\mathbf{X}}^{\mathsf{H}}_{\tau}(k)\left(\mathbf{x}_{\text{ref}}(k) - \hat{\mathbf{d}}^{j}_{\text{ref}}(k) - \boldsymbol{\mu}^{j-1}(k)\right) \\ & = \hat{\mathbf{g}}_{\text{ref},\ell_2}(k) - \hat{\mathbf{g}}^{j}_{\text{ref,iter}}(k), \end{aligned} \tag{6.16}$$

where

$$\hat{\mathbf{g}}_{\text{ref},\ell_2}(k) = \left(\tilde{\mathbf{X}}^{\mathsf{H}}_{\tau}(k)\tilde{\mathbf{X}}_{\tau}(k)\right)^{-1}\tilde{\mathbf{X}}^{\mathsf{H}}_{\tau}(k)\mathbf{x}_{\text{ref}}(k), \tag{6.17}$$

is the iteration-independent term, and

$$\hat{\mathbf{g}}^{j}_{\mathrm{ref,iter}}(k) = \left(\tilde{\mathbf{X}}^{\mathsf{H}}_{\tau}(k)\tilde{\mathbf{X}}_{\tau}(k)\right)^{-1}\tilde{\mathbf{X}}^{\mathsf{H}}_{\tau}(k)\left(\hat{\mathbf{d}}^{j}_{\mathrm{ref}}(k) + \boldsymbol{\mu}^{j-1}\right), \tag{6.18}$$

is the iteration-dependent term. Similarly as for the wideband model, the matrix $\tilde{\mathbf{X}}^{\mathsf{H}}_{\tau}(k)\tilde{\mathbf{X}}_{\tau}(k)$ only needs to be factored once and can be used to solve the corresponding linear system in subsequent iterations. Note that this matrix is much smaller than the corresponding matrix in the wideband model (since $L_g \ll L_{\underline{g}}$), and the resulting iterations do not involve analysis and/or synthesis operators since all computations are performed in the TF domain, resulting in a much lower computational complexity.

6.6 Sparsity-promoting cost function

The presented dereverberation methods in Sections 6.3–6.5 enforce sparsity of the TF coefficients in terms of the cost function $P(.)$, i.e., $P(\mathbf{d}_{\mathrm{ref}})$ quantifies the level of sparsity of the TF-domain coefficients $\mathbf{d}_{\mathrm{ref}}$. Hence, an appropriate sparsity-promoting cost function $P(.)$ needs to be selected. Frequently used cost functions for enforcing sparsity include the convex ℓ_1-norm, the non-convex ℓ_p-norms with $p \in (0,1)$ and the ℓ_0-norm, as already used in Chapters 3–5.

Although the proposed framework can be used with any sparsity-promoting function $P(.)$, as long as its proximal operator $\mathrm{prox}^{\rho}_{P}(.)$ can be computed, cf. (B.14), we confine ourselves to the weighted ℓ_1- and ℓ_2-norm, which are one of the most commonly used sparsity-promoting cost functions [196, 198, 223, 229, 231, 242]. In general, the weighted ℓ_1- and ℓ_2-norm have been shown to be more effective for audio applications than their non-weighted counterparts, since they can be used to approximate the non-convex ℓ_p-norms (cf. Appendix B.1).

In Chapter 3, we have considered the (squared) weighted ℓ_2-norm as cost function $P(.)$, i.e.,

$$P\left(\mathbf{d}_{\mathrm{ref}}\right) = \|\mathbf{d}_{\mathrm{ref}}\|^{2}_{\hat{\mathbf{w}},2} = \sum_{k,n}\hat{w}(k,n)\left|d_{\mathrm{ref}}(k,n)\right|^{2}, \tag{6.19}$$

where $\hat{\mathbf{w}}$ is a vector of nonnegative weights. In addition, in this chapter we will consider the weighted ℓ_1-norm as cost function $P(.)$, i.e.,

$$P\left(\mathbf{d}_{\mathrm{ref}}\right) = \|\mathbf{d}_{\mathrm{ref}}\|_{\hat{\mathbf{w}},1} = \sum_{k,n}\hat{w}(k,n)\left|d_{\mathrm{ref}}(k,n)\right|. \tag{6.20}$$

The weights $\hat{w}(k,n)$ are selected in such a way that the weighted ℓ_1 and ℓ_2-norms simulate the behavior of a non-convex ℓ_p-norm [196, 223, 229] (cf. Appendix B.1).

6.6.1 *Proximal operator*

The ADMM algorithms in Sections 6.3–6.5 rely on the proximal operator $\mathrm{prox}_P^\rho(.)$ of the cost function $P(.)$. The proximal operator for the weighted ℓ_1-norm in (6.20) can be computed element-wise using soft thresholding as (cf. Appendix B.2)

$$\mathrm{prox}_P^\rho\left(d_{\mathrm{ref}}(k,n)\right) = \underbrace{\left(1 - \frac{\rho^{-1}\hat{w}(k,n)}{|d_{\mathrm{ref}}(k,n)|}\right)_+}_{\text{real-valued gain}} d_{\mathrm{ref}}(k,n), \tag{6.21}$$

where $(G)_+ = \max(G,0)$ [246]. In the context of speech enhancement, the proximal operator in (6.21) can be interpreted as applying a real-valued gain to the complex-valued coefficients in $\mathbf{d}_{\mathrm{ref}}$. As noted in [19], in speech enhancement a lower bound $G_{\min}$ on the gain is often introduced, i.e., $(G)_+ = \max(G, G_{\min})$, in order to prevent suppression of small coefficients $d(k,n)$ to exactly zero. As shown in Appendix B.2.1, this corresponds to a cost function $P(.)$ in the form of a Huber function [246], which is quadratic for small magnitudes and equal to a scaled absolute value for large magnitudes, where the transition point depends on the penalty parameter ρ, the weight $\hat{w}(k,n)$ and the lower bound $G_{\min}$. Similarly, the proximal operator for the weighted ℓ_2-norm in (6.19) can be computed element-wise using shrinkage as

$$\mathrm{prox}_P^\rho\left(d_{\mathrm{ref}}(k,n)\right) = \underbrace{\left(\frac{1}{1+2\rho^{-1}\hat{w}(k,n)}\right)_+}_{\text{real-valued gain}} d_{\mathrm{ref}}(k,n). \tag{6.22}$$

Again, the real-valued gain can be bounded from below using a lower bound $G_{\min}$. Estimating the sparse TF coefficients $\mathbf{d}_{\mathrm{ref}}$ using the weighted norms in (6.19) or (6.20) is an iterative two-step procedure. In the first step, the weights $\hat{\mathbf{w}}$ are computed based on the previous estimate $\hat{\mathbf{d}}_{\mathrm{ref}}$ of the desired speech signal in the TF domain. In the second step, an optimization problem with the cost function in (6.19) or (6.20) is solved, and consequently a new estimate of the TF coefficients $\hat{\mathbf{d}}_{\mathrm{ref}}$ is obtained. All previously presented ADMM-based methods will be employed in such a reweighted procedure.

6.6.2 *Weights*

The weights $\hat{w}(k,n)$ for the weighted norms in (6.19) and (6.20) are typically computed locally, using a single TF coefficient, i.e., for the weighted ℓ_1-norm as

$$\hat{w}(k,n) = \left(\left|\hat{d}_{\mathrm{ref}}(k,n)\right|^2 + \varepsilon_{\min}\right)^{\frac{p-1}{2}}, \tag{6.23}$$

and for the weighted ℓ_2-norm as, cf. (3.40),

$$\hat{w}(k,n) = \left(\left| \hat{d}_{\mathrm{ref}}(k,n) \right|^2 + \varepsilon_{\min} \right)^{\frac{p}{2}-1}, \tag{6.24}$$

where $\varepsilon_{\min}$ is a small regularization constant to prevent division by zero.

However, computing the weights as in (6.23) or (6.24) does not take into account the TF structure of a typical speech signal. We consider two approaches to take into account this structure: TF neighborhoods and low-rank approximation. To take into account the TF structure of the desired signal, the concept of neighborhoods for shrinkage operators has been introduced in [242]. Here, we adopt this neighborhood concept for computing the weights. Assuming that a neighborhood $\mathcal{N}(k,n)$ of the coefficient $d_{\mathrm{ref}}(k,n)$ is defined, the corresponding weight $\hat{w}(k,n)$ can be computed by averaging across the neighborhood. For the weighted ℓ_1-norm, the weights can be computed as

$$\hat{w}(k,n) = \left(\sum_{(k',n') \in \mathcal{N}(k,n)} \eta(k',n') \left| \hat{d}_{\mathrm{ref}}(k',n') \right|^2 + \varepsilon_{\min} \right)^{\frac{p-1}{2}}, \tag{6.25}$$

where the coefficients of the neighborhood $\eta(k',n')$ should sum to one. Similarly as in [198,242], we will employ rectangular neighborhoods with equal weights. A similar expression can be used for the weighted ℓ_2-norm. Intuitively, computing weights using a neighborhood around each TF coefficient is similar to using smoothing for estimating the PSD at the current TF point.

Alternatively, it is well known that speech spectrograms can be modeled well using a low-rank approximation [258]. Similarly as in [259], the weights can then be obtained by first computing a low-rank approximation $\hat{\mathbf{P}}$ of the power spectrogram, which is a nonnegative matrix containing the squared magnitudes of the TF coefficients. The weights for the weighted ℓ_1-norm can then be computed as

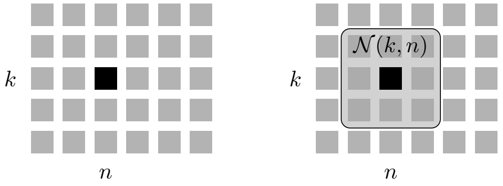
$$\hat{w}(k,n) = \left(\hat{p}(k,n) + \varepsilon_{\min} \right)^{\frac{p-1}{2}}, \tag{6.26}$$

with a similar expression for the weighted ℓ_2-norm. The low-rank approximation $\hat{\mathbf{P}}$ can be computed, e.g., using nonnegative matrix approximation (NMF) [258,260].

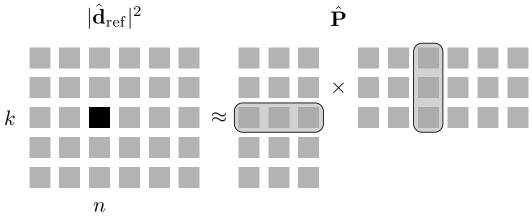
The three different considered ways of computing weights for (6.20) and (6.19) are illustrated in Fig. 6.1. For the illustration we use a 3 × 3 neighborhood for the neighborhood weights and a rank-3 approximation for the low-rank NMF weights.

6.7 Extension to multiple outputs

In this section, we briefly outline how the ADMM-based algorithms in Sections 6.3–6.5 can be extended to MIMO speech dereverberation. More specifically, instead of estimating the desired speech signal at the reference microphone, we reformulate



(a) Local weight (b) Neighborhood weight



(c) NMF weight

Fig. 6.1: Computation of the weight $\hat{w}(k,n)$ for the TF coefficient marked with a black square (■): (a) locally computed weight, (b) weight computed using a neighborhood with dimension 3 across time frames and subbands, and (c) weight computed using an NMF-based low-rank approximation with rank equal to 3.

the problems to estimate the desired speech signal at all microphones, similarly as in Chapter 4 for the subband model.

A multiple-output extension of the wideband signal model with analysis sparsity in (6.4) leads to the following optimization problem

$$\begin{aligned} \min_{\underline{\mathbf{D}},\underline{\mathbf{G}}} \quad & P\left(\mathbf{\Psi}^{\mathsf{H}}\underline{\mathbf{D}}\right) \\ \text{subject to} \quad & \underline{\mathbf{D}} + \underline{\tilde{\mathbf{X}}}_{\mathcal{T}}\underline{\mathbf{G}} = \underline{\mathbf{X}}, \end{aligned} \tag{6.27}$$

where $\underline{\mathbf{D}} \in \mathbb{R}^{T\times M}$ is the multi-channel desired speech component in the time domain, and $\underline{\mathbf{G}} \in \mathbb{R}^{ML_g\times M}$ is the MIMO prediction filter in the time domain. Similarly, a multiple-output extension of the wideband signal model with synthesis sparsity in (6.9) leads to the following optimization problem

$$\begin{aligned} \min_{\mathbf{D},\underline{\mathbf{G}}} \quad & P\left(\mathbf{D}\right) \\ \text{subject to} \quad & \mathbf{\Psi}\mathbf{D} + \underline{\tilde{\mathbf{X}}}_{\mathcal{T}}\underline{\mathbf{G}} = \underline{\mathbf{X}}, \end{aligned} \tag{6.28}$$

where $\mathbf{D} \in \mathbb{C}^{KN\times M}$ are the TF coefficients of the multi-channel desired speech component. As a generalization of the MIMO MCLP-based optimization problem in (4.6), a multiple-output extension of the subband model in (6.13) leads to the following optimization problem

$$\begin{aligned} \min_{\mathbf{D}(k),\mathbf{G}(k)} \quad & P\left(\mathbf{D}(k)\right) \\ \text{subject to} \quad & \mathbf{D}(k) + \tilde{\mathbf{X}}_\tau(k)\mathbf{G}(k) = \mathbf{X}(k), \end{aligned} \tag{6.29}$$

where $\mathbf{D}(k) \in \mathbb{C}^{N \times M}$ are the TF coefficients of the multi-channel desired speech signal in the k-th subband.

The extended MIMO formulations in (6.27)–(6.29) can again be solved using the ADMM algorithm. For example, the subband MIMO dereverberation problem in (6.29) can be solved using the following iterative updates

$$\hat{\mathbf{D}}^j(k) \leftarrow \operatorname{prox}_P^\rho \left(\mathbf{X}(k) - \tilde{\mathbf{X}}_\tau(k)\hat{\mathbf{G}}^{j-1}(k) - \mathbf{M}^{j-1}(k)\right), \tag{6.30a}$$

$$\hat{\mathbf{G}}^j(k) \leftarrow \left(\tilde{\mathbf{X}}_\tau^{\mathsf{H}}(k)\tilde{\mathbf{X}}_\tau(k)\right)^{-1} \tilde{\mathbf{X}}_\tau^{\mathsf{H}}(k) \left(\mathbf{X}(k) - \hat{\mathbf{D}}^j(k) - \mathbf{M}^{j-1}(k)\right), \tag{6.30b}$$

$$\mathbf{M}^j(k) \leftarrow \mathbf{M}^{j-1}(k) + \eta \left(\hat{\mathbf{D}}^j(k) + \tilde{\mathbf{X}}_\tau(k)\hat{\mathbf{G}}^j(k) - \mathbf{X}(k)\right), \tag{6.30c}$$

where $\mathbf{M}(k)$ is the dual variable.

Similarly as in Chapter 4, the cost function $P(.)$ should promote sparsity across the temporal dimension and take into account the group structure across the microphones. This can be achieved by using, e.g., a mixed $\ell_{p,2}$-norm which can be approximated in a reweighting procedure by a weighted $\ell_{2,2}$-norm (as in Chapter 4.3.1) or a weighted $\ell_{1,2}$-norm, with their corresponding proximal operators given in Appendix B.2.2.

6.8 Relation to existing methods

The wideband signal model has been employed for MCLP-based dereverberation in the time-domain in [107, 150, 153, 154], however without explicitly enforcing sparsity of the TF coefficients of the desired speech signal. For example, in [107, 150] the time-domain prediction filters have been estimated by minimizing the output energy, which is equivalent to using the ℓ_2-norm of $\underline{\mathbf{d}}_{\text{ref}}$ as the cost function, i.e.,

$$P\left(\underline{\mathbf{d}}_{\text{ref}}\right) = \|\underline{\mathbf{d}}_{\text{ref}}\|_2^2 = \sum_{t=1}^{T} |\underline{d}_{\text{ref}}(t)|^2 . \tag{6.31}$$

Note that this is a special case of the formulation in (6.4), with the ℓ_2-norm as the cost function and without the analysis operator $\mathbf{\Psi}^{\mathsf{H}}$. In this case, the closed-form solution for the prediction filter is given as $\hat{\underline{\mathbf{g}}}_{\text{ref},\ell_2}$ in (6.7). In [150], the prediction delay $\underline{\tau}$ has not been used, and it has been observed that the obtained prediction filter typically results in excessive whitening of the speech signal, since it removes both the effect of the reverberation but also the short-time correlation of the speech signal. To compensate for the whitening, the output signal is post-processed using an estimated whitening filter. In [107], the microphone signals have been pre-whitened to reduce the effect of the short-time correlation of the speech signal on the estimation of the prediction filter for the MCLP-based signal model. However, the estimated prediction filter has not been used to perform dereverberation using MCLP-based inverse filtering, but to obtain an estimate of the late reverberation for spectral subtraction-based dereverberation.

A different cost function has been used in [154]. More specifically, a time-varying Gaussian model has been used for the desired speech signal in the time-domain, and the time-domain prediction filter has been estimated by iterative maximization of the likelihood function, similarly as in Chapter 3, with pre-whitening applied on the microphone signals. It can be shown that the obtained optimization problem is equivalent to using the weighted ℓ_2-norm as the cost function, i.e.,

$$P\left(\underline{\mathbf{d}}_{\text{ref}}\right) = \|\underline{\mathbf{d}}_{\text{ref}}\|^2_{\hat{\underline{\mathbf{w}}},2} = \sum_{t=1}^{T} \hat{\underline{w}}(t)\, |\underline{d}_{\text{ref}}(t)|^2 \,, \tag{6.32}$$

with the pre-whitened microphone signals. This is a special case of the formulation in (6.4), with the weighted ℓ_2-norm as the cost function and without the analysis operator. For fixed weights, the obtained weighted least-squares optimization problem has a closed-form solution for the prediction filter. In [154], the weights $\hat{\underline{w}}(t)$ have been computed from the previous estimate of the desired speech signal by averaging the energy of the samples across a short frame centered at t [154]. When employed in a reweighting procedure, this can be related to promoting sparsity of the frames of the desired time-domain signal $\underline{\mathbf{d}}$, since the weights take into account the short-term energy of the desired speech signal. Furthermore, a single reweighting iteration has been used in the original contribution, and it has been reported that multiple iterations do not always improve performance [154].

As mentioned throughout this chapter, the MCLP-based speech dereverberation methods proposed in Chapters 3 and 4 can be considered a special case of the subband model in (6.13) and (6.29). More specifically, the cost function for the single-output method in Chapter 3 is equal to the weighted ℓ_2-norm, cf. (3.36),

$$P\left(\mathbf{d}_{\text{ref}}(k)\right) = \|\mathbf{d}_{\text{ref}}(k)\|^2_{\hat{\mathbf{w}}(k),2} = \sum_{n} \hat{w}(k,n)\, |d_{\text{ref}}(k,n)|^2 \,, \tag{6.33}$$

while the cost function for the multi-output method in Chapter 4 is equal to the weighted $\ell_{2,2;\boldsymbol{\Phi}}$ norm, cf. (4.8). Furthermore, in Chapters 3 and 4, only local weights have been considered.

6.9 Simulations

In this section, the dereverberation performance of the ADMM-based methods proposed in Sections 6.2–6.5 is investigated. More specifically, we consider the ADMM methods using the wideband model with analysis sparsity, the wideband model with synthesis sparsity, and the subband signal model with the weighted ℓ_1-norm and

ℓ_2-norm as the cost function. In addition, for the subband model with the weighted ℓ_2-norm as cost function we also consider the IRLS algorithm from Chapter 3.

The considered acoustic scenario and the implementation details are outlined in Section 6.9.1. The influence of the cost function and the penalty parameter on the performance is investigated in Section 6.9.2. The influence of the weights on the performance is investigated in Section 6.9.3. The wideband model is considered in Section 6.9.4.

6.9.1 *Acoustic scenario and algorithmic setup*

We consider the same acoustic scenario from the REVERB challenge [22, 23] used for the simulations in Chapters 3 and 4, i.e., a single speech source and $M = 2$ microphones placed at a distance of about 2 m from the source. The room has a reverberation time $T_{60} \approx 700$ ms and the sampling frequency is $f_s = 16$ kHz. The reverberant signals have been generated by convolving each of the 10 speech samples (5 male and 5 female speakers) [234] with an average length of approximately 5.2 s with the measured RIRs.

Similarly to the parameter setup for the simulations in Chapter 3, the analysis and synthesis STFT is computed using a tight window based on a 64 ms Hamming window with a 16 ms window shift. For the subband model in (6.1) the filter length and the prediction delay are set to $L_g = 25$ and $\tau = 2$. For the wideband model in (6.2) the filter length and the prediction delay are set to $\underline{L}_g = 6400$ and $\underline{\tau} = 512$, i.e., corresponding to the filter length of 400 ms and the prediction delay of 32 ms, as used for the subband model. The weights $\hat{w}(k, n)$ used in the weighted norms are computed either locally as in (6.23) or (6.24), using a rectangular neighborhood as in (6.25), or using an NMF-based low-rank approximation as in (6.26). The weights are regularized with $\varepsilon_{\min} = 10^{-8}$. The low-rank approximation is computed using NMF with Itakura-Saito divergence with multiplicative updates [261]. The number of reweighting iterations I is varied in the experiments, while the maximum number of ADMM iterations was set to $J = 50$ with $\eta = 1.6$. For the proximal operator used in the wideband analysis method (cf. Appendix B.4), we set the penalty parameter δ equal to the penalty parameter ρ of the ADMM algorithm for the wideband analysis problem. For the LASSO problem used in the wideband synthesis method (cf. Section 6.4), we used FISTA with the maximum number of iterations set to 50 with early stopping when the relative change of the estimate is smaller than 10^{-3} (cf. Appendix B.5). In all experiments we used the lower bound $G_{\min} = 0.01$ for the real-valued gain (cf. Section 6.6.1), with smaller values typically resulting in more suppression of unwanted reverberation but also in stronger processing artifacts due to the application of the proximal operator, and larger values resulting in less reverberation suppression.

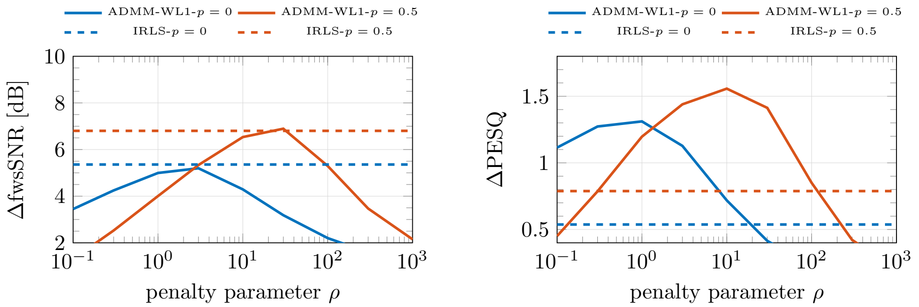
The dereverberation performance is evaluated in terms of the instrumental measures described in Section 2.3, i.e., the improvement in fwsSNR (ΔfwsSNR) and PESQ (ΔPESQ) [210, 211] between the processed output signal and the reverberant input signal. The reference signal used for the instrumental measures is the direct signal on the reference microphone, obtained by convolving the anechoic speech signal with the direct component of the corresponding RIR. The reported improvements of the instrumental measures are obtained by averaging over all speech samples.

6.9.2 *Influence of the cost function and the penalty parameter*

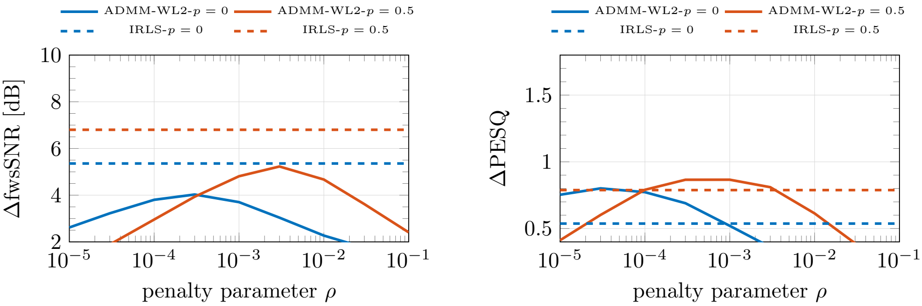
In this section, we investigate the influence of the sparsity-promoting cost function $P(.)$ and the penalty parameter ρ of the ADMM algorithm on the dereverberation performance of the proposed methods. We consider the subband method using the weighted ℓ_1-norm in (6.20), based on the ADMM algorithm in Section 6.5 (ADMM-WL1-p), and the subband method using the weighted ℓ_2-norm in (6.19), either based on the ADMM algorithm in Section 6.5 (ADMM-WL2-p) or the IRLS algorithm from Section 3.3.3 (IRLS-p). For all methods the weights $\hat{w}(k,n)$ are computed locally as in (6.23) or (6.24). We consider two values of the shape parameter $p \in \{0, 0.5\}$ and a suitable range of values for the penalty parameter ρ.

Firstly, the performance of the ADMM and the IRLS-based methods for a single reweighting iteration, i.e., $I = 1$, is depicted in Fig. 6.2. It can be observed that all considered methods result in improvements in terms of the instrumental measures. The IRLS-based method, which does not dependent on the penalty parameter ρ, results in large improvements, with $p = 0.5$ performing better than $p = 0$, as demonstrated also in Section 3.5. The performance of the ADMM-based methods strongly depends on the value of the penalty parameter ρ, for both types of the cost function and values of the shape parameter p. Both ΔfwsSNR and ΔPESQ exhibit a similar behavior, with the performance first increasing and then decreasing with the penalty parameter ρ. This behavior can be explained by referring to the proximal operators in (6.21) and (6.22). On the one hand, small values of the penalty parameter ρ result in a relatively strong suppression of the TF coefficients and over-suppression of the desired speech signal in each ADMM iteration. On the other hand, large values of the penalty parameter result in a relatively weak suppression of the TF coefficients and a relatively low suppression of the desired speech signal in each ADMM iteration. It can also be observed that in general ADMM-WL1 performs better than ADMM-WL2. The ADMM-WL2 method performs worse than the IRLS method in terms of ΔfwsSNR and achieves a similar performance in terms of ΔPESQ for both values of the shape parameter p. The ADMM-WL1 performs equally well as the IRLS in terms of ΔfwsSNR and achieves a significantly better performance in terms of ΔPESQ. Overall, the best performance using $I = 1$ reweighting iteration is obtained using ADMM-WL1 with the shape parameter $p = 0.5$.

Secondly, the performance of the ADMM and the IRLS-based methods for $I = 20$ reweighting iterations is depicted in Fig. 6.3. It can be observed that all considered methods result in improvements in terms of the instrumental measures. The IRLS-based method results in large improvements in both measures and performs significantly better than with $I = 1$ reweighting iteration. As already observed for $I = 1$, the performance of the ADMM-based methods again strongly depends on the value of the penalty parameter ρ. Furthermore, ADMM-WL1 and the ADMM-WL2 now achieve almost the same performance in terms of both instrumental measures, with $p = 0.5$ still performing better than $p = 0$. The similar best-case performance



(a) WL1 (b) WL1



(c) WL2 (d) WL2

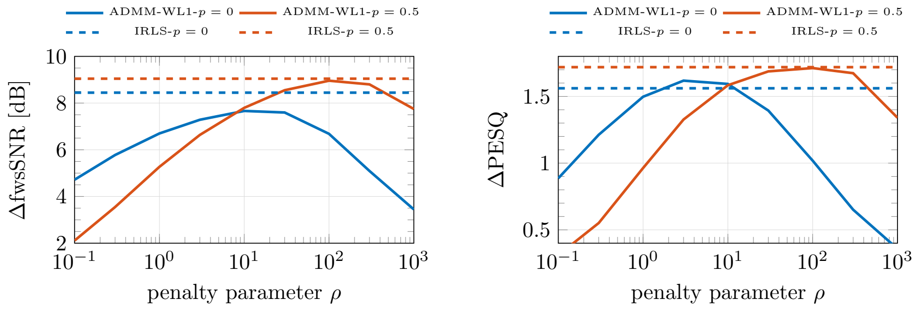
Fig. 6.2: Performance of the ADMM and IRLS-based methods with $M = 2$ and $L_g = 25$ using local weights and $I = 1$ reweighting iteration in terms of ΔfwsSNR (left) and ΔPESQ (right), using the weighted ℓ_1-norm (top) and the weighted ℓ_2-norm (bottom).

obtained using all methods can be attributed to the underlying cost function, since both the weighted ℓ_1-norm and the weighted ℓ_2-norm aim to approximate the non-convex ℓ_p-norm as a measure of sparsity. Overall, the best performance using $I = 20$ reweighting iteration is obtained using both the ADMM and the IRLS-based algorithms with the shape parameter $p = 0.5$.

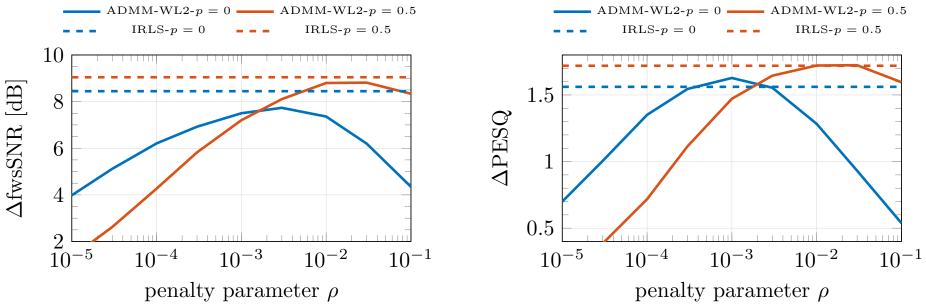
6.9.3 *Influence of the structured weights*

In this section, we investigate the influence of the structured weights used in the reweighting iterations (cf. Section 6.6.2) on the dereverberation performance of the proposed methods.

We consider the subband method using the weighted ℓ_1-norm based on the ADMM algorithm (ADMM-WL1-p), since it in general perform comparable to or better than the weighted ℓ_2-norm, and the subband method using the weighted ℓ_2-norm based on the IRLS algorithm (IRLS-p). The neighborhood weights in (6.25) are computed using a square-shaped neighborhood, with neighborhood sizes ranging between 3



(a) WL1 (b) WL1

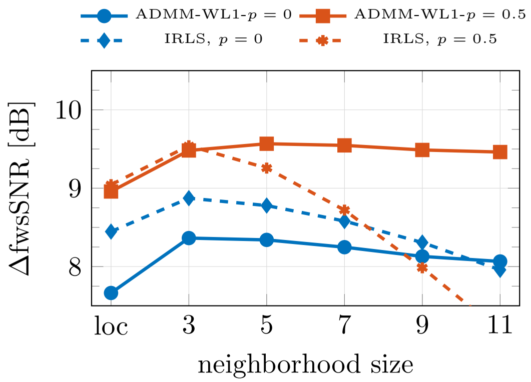


(c) WL2 (d) WL2

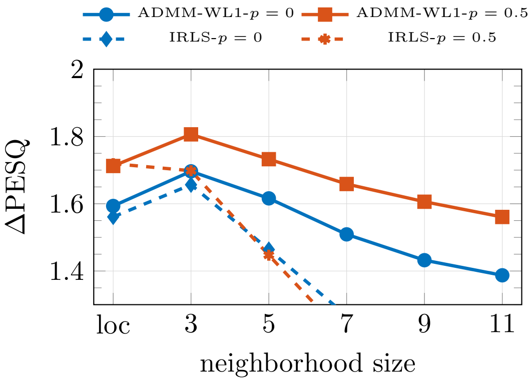
Fig. 6.3: Performance of the ADMM and IRLS-based methods with $M = 2$ and $L_g = 25$ using local weights and $I = 20$ reweighting iterations in terms of ΔfwsSNR (left) and ΔPESQ (right), using the weighted ℓ_1-norm (top) and the weighted ℓ_2-norm (bottom).

and 11. Note that the locally computed weights correspond to a neighborhood with size 1. The low-rank approximation-based weights in (6.26) are computed using NMF with a rank between 20 and 80. We consider two values of the shape parameter $p \in \{0, 0.5\}$, and the penalty parameter ρ is set for the corresponding p to $\{10, 100\}$ for local weights and $\{30, 100\}$ for structured weights. The number of reweighting iterations has been set to $I = 20$.

Firstly, we consider the neighborhood-based weights, with the results depicted in Fig. 6.4. It can be observed that the performance of the ADMM-WL1-p and the IRLS-p method depends on the neighborhood size. For the IRLS-p method, the neighborhood size of 3 leads to a better performance than the local weights (neighborhood size of 1) in terms of both ΔfwsSNR and ΔPESQ, while larger neighborhoods result in a large performance degradation. For the ADMM-WL-p method, the neighborhood size has a similar influence, however with less performance degradation as the neighborhood size increases. Relatively small neighborhoods result in an improved performance, since including a neighborhood around the current TF coefficient can be seen as a form of smoothing for estimating the PSD of the desired



(a) ΔfwsSNR

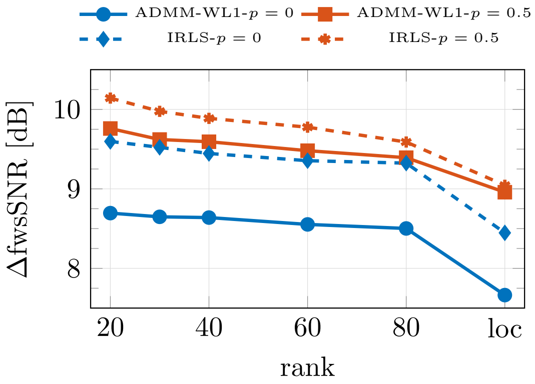


(b) ΔPESQ

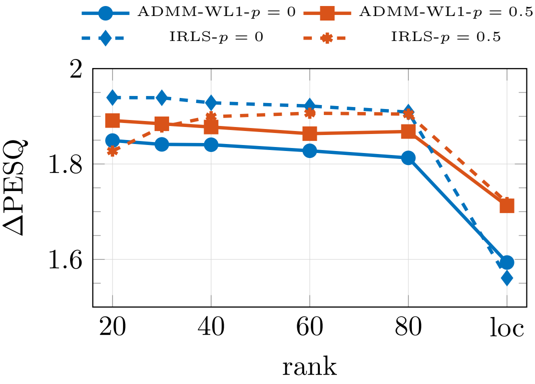
Fig. 6.4: Performance of the ADMM-WL1 and IRLS-based methods for $p = 0$ and $p = 0.5$ with $M = 2$ and $L_g = 25$ using neighborhood weights and local weights (loc) with $I = 20$ reweighting iterations in terms of ΔfwsSNR (top) and ΔPESQ (bottom).

speech signal at the current TF point. However, increasing the neighborhood size results in more smoothing, which is contradicting the actual goal of making the output signal more sparse in the TF domain, and therefore results in a decreased performance. Overall, the best performance using neighborhood weights is obtained using the ADMM-WL1-p and the IRLS-p methods with the neighborhood size equal to 3 and the shape parameter $p = 0.5$, with ADMM-WL1-p performing somewhat better in terms of ΔPESQ (approximately 0.1 points).

Secondly, we consider the NMF-based weights, with the results depicted in Fig. 6.5. It can be observed that the performance of both the ADMM and the IRLS-based methods is relatively insensitive to the rank of the low-rank approximation for ranks larger than 40, although a small decline in performance can be observed, as expected from the low-rank model. Both the ADMM and the IRLS-based methods benefit from the NMF weights, with performance improvements compared to the



(a) ΔfwsSNR

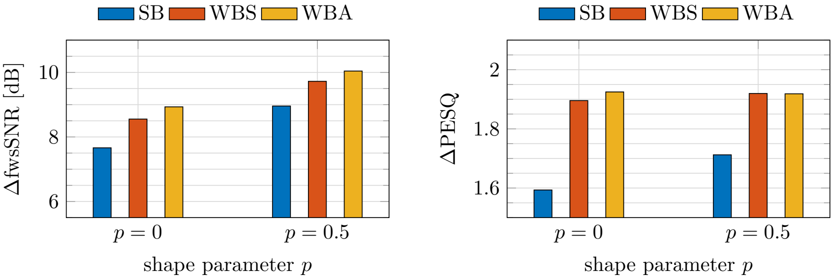


(b) ΔPESQ

Fig. 6.5: Performance of the ADMM-WL1 and IRLS-based methods for $p = 0$ and $p = 0.5$ with $M = 2$ and $L_g = 25$ using NMF weights and the local weights (loc) with $I = 20$ reweighting iterations in terms of ΔfwsSNR (top) and ΔPESQ (bottom).

local weights for all considered ranks. Overall, the best performance is achieved using IRLS-p with $p = 0.5$ and rank 40, with the ADMM-WL1-p with $p = 0.5$ and the IRLS-p with $p = 0$ with rank 20 performing slightly worse. Furthermore, the best-case performance is considerably higher than when using the neighborhood weights and the local weights (cf. Fig. 6.4).

In summary, structured weights can lead to considerable improvements for speech dereverberation methods exploiting sparsity. The relatively simple neighborhood weights, which include the local structure around each TF point, result in improvements compared to the locally computed weights. The NMF weights, which include modeling of the whole spectral and temporal profile, result in further improvements, with the additional cost of computing the NMF-based low-rank approximation in each reweighting iteration. In general, including additional structure, beyond sparsity, hence shows to be beneficial for the dereverberation performance.



(a) ΔfwsSNR (b) ΔPESQ

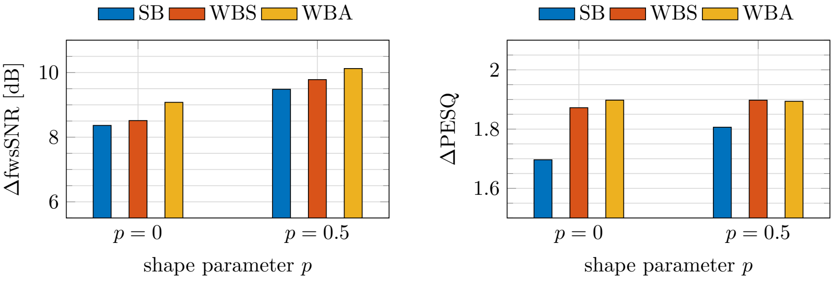
Fig. 6.6: Performance of the wideband analysis (WBA), wideband synthesis (WBS) and subband (SB) ADMM-WL1 methods using local weights with $I = 20$ reweighting iterations in terms of ΔfwsSNR (left) and ΔPESQ (right).

6.9.4 *Comparison between subband and wideband methods*

In this section, we investigate the performance differences between the wideband methods and the subband method. We consider the formulation with the wideband signal model and analysis sparsity (WBA) from Section 6.3, the wideband signal model with synthesis sparsity (WBS) from Section 6.4, and the subband (SB) signal model from Section 6.5. In all cases we use the weighted ℓ_1-norm in (6.20) as the cost function, i.e., the SB corresponds to the ADMM-WL1-p method investigated in Section 6.9.3. The shape parameter has been set to $p \in \{0, 0.5\}$, and the penalty parameter ρ for SB set as in the previous experiment (cf. Section 6.9.3), for WBA set to $\{30, 300\}$, and for WBS set to $\{10, 100\}$. For all methods we used $I = 20$ reweighting iterations, with the other implementation details outlined in Section 6.9.1.

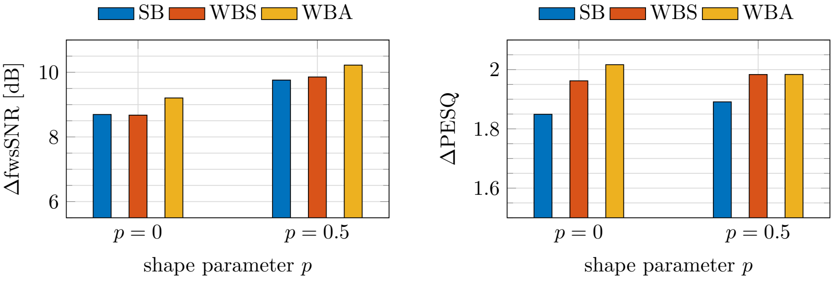
Firstly, we consider the WBA, WBS and SB methods using local weights in (6.23). The obtained results are depicted in Fig. 6.6. It can be observed that both wideband methods perform significantly better than the subband method in both ΔPESQ and ΔfwsSNR, with the WBA method performing better than the WBS method. Furthermore, the obtained performance of both wideband methods is better for $p = 0.5$ than for $p = 0$, similarly as for the subband method.

Secondly, we consider the WBA, WBS and SB methods using neighborhood weights in (6.25), with the neighborhood size equal to 3 (as suggested by the results in Section 6.9.3). The obtained results are depicted in Fig. 6.7. Again, it can be observed that both wideband methods perform better than the subband method in terms of both ΔPESQ and ΔfwsSNR, with the WBA method performing better than the WBS method, and the performance of both wideband methods being better for $p = 0.5$ than for $p = 0$. However, the performance of the WBA and WBS methods using neighborhood weights is in some cases lower than when using local weights, e.g., for the WBS method in terms of both ΔfwsSNR and ΔPESQ.



(a) ΔfwsSNR (b) ΔPESQ

Fig. 6.7: Performance of the wideband analysis (WBA), wideband synthesis (WBS) and subband (SB) ADMM-WL1 methods using neighborhood weights with $I = 20$ reweighting iterations in terms of ΔfwsSNR (left) and ΔPESQ (right).



(a) ΔfwsSNR (b) ΔPESQ

Fig. 6.8: Performance of the wideband analysis (WBA), wideband synthesis (WBS) and subband (SB) ADMM-WL1 methods using NMF weights with $I = 20$ reweighting iterations in terms of ΔfwsSNR (left) and ΔPESQ (right).

Thirdly, we consider the WBA, WBS and SB methods with NMF weights in (6.26), with the rank equal to 20 (as suggested by the results in Section 6.9.3). The obtained results are depicted in Fig. 6.8. Similarly as for the local and the neighborhood weights, it can be observed that both wideband methods perform significantly better than the subband method in terms of both ΔPESQ and ΔfwsSNR, with the WBA method performing better than WBS method, and the performance of both wideband methods being typically better for $p = 0.5$ than for $p = 0$, except for WBA in terms of ΔPESQ. Overall, the WBA and WBS methods with NMF weights achieve a better performance than when using local weights. The best performance for $p = 0$ and $p = 0.5$ is obtained using the WBA method, with $p = 0.5$ performing better in terms of ΔfwsSNR and $p = 0$ performing better in terms of ΔPESQ.

6.9.5 *Summary of the results*

In this section, we summarize the results obtained using selected variants of sparse MCLP, using the subband or the wideband signal model and using the IRLS or the ADMM algorithm to minimize a weighted ℓ_1- or ℓ_2-norm with local or structured weights. In all cases we used $I = 20$ reweighting iterations, and the ADMM penalty parameter ρ has been set as described in the previous sections.

The obtained results are summarized in Table 6.1, including the dereverberation performance in terms of the considered performance measures and the average RTFs. It can be observed that the shape parameter p of the cost function improves the performance of the subband methods (cf. $\mathcal{M}_1$ vs. $\mathcal{M}_2$), improving ΔfwsSNR and ΔPESQ by approximately 0.5 dB and 0.1 points, without increasing the computational complexity. Using the structured weights further improves the performance of the subband methods, with the NMF weights (cf. $\mathcal{M}_2$ vs. $\mathcal{M}_3$) improving ΔfwsSNR and ΔPESQ by approximately 0.5 dB and 0.2 points. However, the computational complexity is also increased, since an NMF of the power spectrogram needs to be computed at each iteration. Furthermore, similar performance can be obtained with the IRLS and ADMM algorithms for the subband model (cf. $\mathcal{M}_3$ vs. $\mathcal{M}_4$), with the latter having a larger computational complexity. It should also be pointed out that a suitable value for the penalty parameter ρ for the ADMM algorithm needs to be selected, as described in the previous sections. The wideband signal model with analysis sparsity offers some advantage over the subband signal model (cf. $\mathcal{M}_2$ vs. $\mathcal{M}_5$), improving ΔfwsSNR and ΔPESQ by approximately 1 dB and 0.2 points. However, the wideband method has a much higher computational complexity, due to the long prediction filter ($M\underline{L}_g$) and since the analysis and synthesis operators (i.e., $\mathbf{\Psi}^{\mathsf{H}}$ and $\mathbf{\Psi}$) need to be applied in each iteration. The subband method has a shorter filter (ML_g) and the analysis and synthesis needs to be performed only once. Using the structured weights improves the performance for the wideband method, with the NMF weights (cf. $\mathcal{M}_5$ vs. $\mathcal{M}_6$) resulting in minor improvements in ΔfwsSNR and ΔPESQ, with approximately the same computational complexity. The observed improvements are however smaller than the ones for the subband method (cf. $\mathcal{M}_2$ vs. $\mathcal{M}_3$). This can be attributed to the fact that using NMF weights with the subband model includes information about the global structure of the speech signal (e.g., across subbands) in the estimation procedure.

Overall, the best performance is obtained using the wideband signal model with analysis sparsity and NMF weights ($\mathcal{M}_6$), improving ΔfwsSNR and ΔPESQ by approximately 1.7 dB and 0.4 points compared to the subband signal model with local weights ($\mathcal{M}_1$). The differences can also be observed from the spectrograms of the corresponding signals depicted in Fig. 6.9 (the differences were largest in the shown frequencies up to 4 kHz). By comparing the estimated signals with the reverberant microphone signal, it can be observed that both the subband method $\mathcal{M}_1$ and the wideband method $\mathcal{M}_6$ achieve a high level of dereverberation. However, it can also be observed that $\mathcal{M}_6$ achieves better dereverberation, e.g., by removing more reverberant energy in speech pauses or even between harmonics. Although resulting in a better performance, the wideband methods are in general much more computationally expensive than the subband methods [183], which makes the sub-

Table 6.1: Summary of the results obtained with the selected variants of sparse MCLP with different signal models, cost functions and iterative optimization algorithms in terms of ΔfwsSNR and ΔPESQ and real-time factors.

	model	alg.	cost $P(.)$	weigh.	ΔfwsSNR	ΔPESQ	RTF
$\mathcal{M}_1$	SB	IRLS	WL2-$p = 0$	loc	8.45	1.56	3
$\mathcal{M}_2$	SB	IRLS	WL2-$p = 0.5$	loc	9.04	1.72	3
$\mathcal{M}_3$	SB	IRLS	WL2-$p = 0.5$	nmf	9.78	1.91	7
$\mathcal{M}_4$	SB	ADMM	WL1-$p = 0.5$	nmf	9.76	1.89	15
$\mathcal{M}_5$	WBA	ADMM	WL1-$p = 0.5$	loc	10.04	1.92	76
$\mathcal{M}_6$	WBA	ADMM	WL1-$p = 0.5$	nmf	10.22	1.98	77

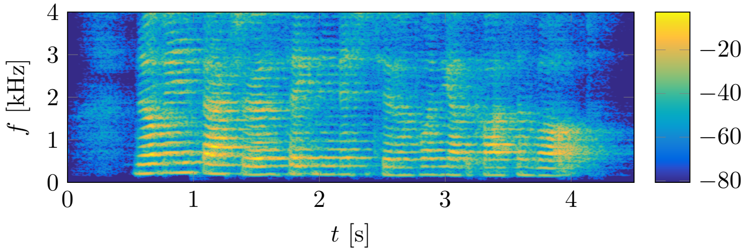
band processing more appealing for practical application. Nevertheless, using the wideband methods offers more flexibility in the selection of the TF transform, and could be used even when the subband model does not hold, e.g., if there is a strong influence between adjacent bands in the TF domain.

6.10 Summary

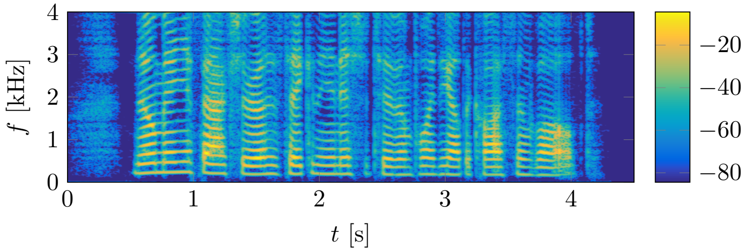
In this chapter we have presented a general framework for multi-channel speech dereverberation exploiting sparsity of the speech signal in the time-frequency domain. We have formulated MCLP-based speech dereverberation as an optimization problem with a general cost function aiming to promote sparsity of the desired speech signal in the time-frequency domain. The presented framework enables to employ either a wideband or a subband MCLP-based signal model, as well as an analysis or a synthesis prior for the desired speech signal. While the discussion in this chapter has been limited to sparsity in the STFT domain, other time-frequency transforms could be easily adopted in this framework by using a suitable pair of analysis-synthesis operators. We have shown that all resulting optimization problems can be efficiently solved using the ADMM algorithm, and that different sparsity-promoting cost functions can be employed by selecting an appropriate proximal operator.

Simulation results show that the proposed ADMM-based methods using the weighted ℓ_1-norm as the sparsity-promoting cost function perform better than the conventional IRLS-based method for a single reweighting iteration, and achieve a similar performance for multiple reweighting iterations. In addition, we have shown that using structured weights in the reweighting iterations can improve the dereverberation performance of the sparsity-based methods.

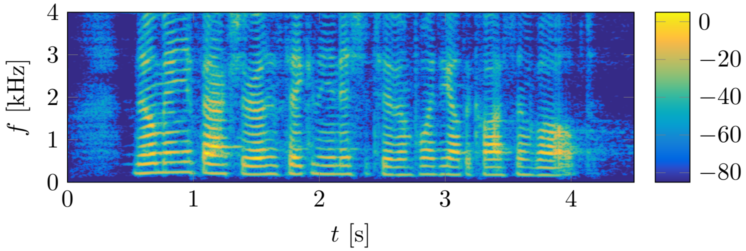
In conclusion, even though the performance of the wideband methods is better than the subband methods, the subband methods appear to be more relevant in practice, since they achieve a very good dereverberation performance with a significantly lower computational complexity than the wideband methods.



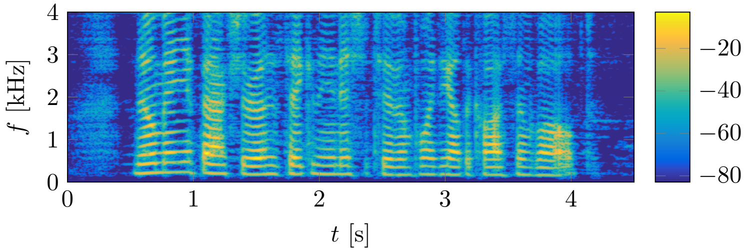
(a) Microphone signal



(b) Direct speech signal



(c) $\mathcal{M}_1$



(d) $\mathcal{M}_6$

Fig. 6.9: Spectrograms of the microphone signal, direct speech signal and the desired speech signal estimated using $\mathcal{M}_1$ and $\mathcal{M}_6$ (showing frequencies up to 4 kHz).

7

SPARSITY-BASED MULTI-CHANNEL DEREVERBERATION AND DENOISING

In Chapters 3–6, we have considered different formulations of blind speech dereverberation based on the MCLP-based signal model and sparsity of the speech signal in the TF domain. However, the additive noise has not been explicitly taken into account. Although the simulation results in Chapter 4 show that sparse MCLP-based dereverberation is to some extent robust to additive noise, its performance is substantially degraded when the noise is the dominant disturbance.

Sparsity in the TF domain has been often exploited for denoising of audio signals [197–199]. Typically, it is assumed that the desired signal has a sparse representation in the TF domain, as opposed to the undesired noise signal, and denoising is formulated as an optimization problem with a sparsity-promoting cost function with a wideband or a subband signal model [195, 198, 199]. Joint dereverberation and denoising based on MCLP has been considered in [158, 159, 165, 167, 168]. In [158, 159], a probabilistic formulation based on a locally Gaussian for the speech signal has been used, leading to an iterative algorithm for ML parameter estimation. Similarly, a locally Gaussian model and iterative ML estimation have been used in [165], with a Kalman smoother used to solve a structured LS problem [262], and extended to online processing in [167]. In [168], a similar probabilistic MCLP-based formulation has been combined with a probabilistic diffuse noise model, aiming to reduce a non-stationary noise while assuming that the spatial properties are known.

In this chapter, we extend sparse MCLP-based dereverberation methods from Chapters 3 and 4 by taking into account the additive noise signal. We propose batch subband methods for denoising and for joint dereverberation and denoising by exploiting sparsity of the speech signal. More specifically, the optimization problem for denoising is formulated using a sparsity-promoting cost function with a subband signal model and a constraint for the noise energy. The optimization problem for joint dereverberation and denoising is formulated using a sparsity-promoting cost function with a subband MCLP-based signal model for the microphone signal and a constraint for the noise energy. Similarly as in the previous chapters, the obtained optimization problems can be solved using the ADMM algorithm.

In Sections 7.1 and 7.2, we formulate the problem of joint dereverberation and denoising and define the signal model for the noise. In Section 7.3, we formulate MIMO speech denoising using the subband signal model and a sparsity-promoting cost function. In Section 7.4, we propose a joint method for dereverberation and denoising, by including the noise term in the MCLP-based model and a bound for the noise energy. The cost function is discussed in Section 7.5, and the performance of the proposed methods is evaluated in Section 7.6.

7.1 Problem formulation

We consider an acoustic scenario with a single static speech source captured by M microphones in a reverberant enclosure in the presence of additive noise. Given a batch of N time frames, the signal model for the reverberant and noisy microphone signal $\mathbf{Y}(k) \in \mathbb{C}^{N \times M}$ in the k-th subband is given by, cf. (2.20),

$$\mathbf{Y}(k) = \mathbf{X}(k) + \mathbf{V}(k), \tag{7.1}$$

where $\mathbf{X}(k) \in \mathbb{C}^{N \times M}$ is the reverberant speech matrix and $\mathbf{V}(k) \in \mathbb{C}^{N \times M}$ is the noise matrix. Since the signal model in (7.1) is used independently in each subband, the subband index k will be omitted in the remainder of the chapter for notational convenience.

Using the subband MCLP-based signal model for the reverberant signal in (2.31), the signal model in (7.1) can be written as

$$\mathbf{Y} = \mathbf{D} + \tilde{\mathbf{X}}_\tau \mathbf{G} + \mathbf{V}, \tag{7.2}$$

where $\mathbf{D} \in \mathbb{C}^{N \times M}$ is the multi-channel desired speech signal matrix and $\tilde{\mathbf{X}}_\tau \mathbf{G}$ is the MCLP-based multi-channel undesired reverberant signal, with the convolution matrix $\tilde{\mathbf{X}}_\tau \in \mathbb{C}^{N \times ML_g}$ and the prediction filter $\mathbf{G} \in \mathbb{C}^{ML_g \times M}$.

In the following, we formulate joint dereverberation and denoising as estimation of the desired speech signal signal $\mathbf{D}$ without using any information about the ATFs, but assuming the noise correlation matrix to be available. Similarly as in Chapter 6, the optimization problem is formulated in terms of a general sparsity-promoting cost function $P(.)$, assuming batch processing.

7.2 Noise model

In this section, we briefly discuss the signal model for the additive noise. We assume that the multi-channel noise signal is Gaussian, zero-mean, stationary, and independent over time frames. More specifically, consider the multi-channel noise matrix $\mathbf{V} \in \mathbb{C}^{N \times M}$ and its corresponding correlation matrix $\mathbf{\Phi}_V \in \mathbb{C}^{M \times M}$, i.e.,

$$\mathbf{\Phi}_V = \mathcal{E}\left\{\mathbf{v}(n)\mathbf{v}^{\mathsf{H}}(n)\right\}, \tag{7.3}$$

where $\mathcal{E}\{.\}$ is the mathematical expectation operator. Assuming the noise correlation matrix $\mathbf{\Phi}_V$ is positive definite (PD), it can be decomposed as

$$\mathbf{\Phi}_V = \mathbf{\Phi}_V^{1/2} \mathbf{\Phi}_V^{\mathsf{H}/2}, \tag{7.4}$$

e.g., using the Cholesky decomposition [263]. The matrix $\mathbf{\Phi}_V^{1/2}$ can then be used to perform spatial decorrelation and normalization of the multi-channel noise signal $\mathbf{V}$. It can be easily shown that the matrix $\mathbf{V}\mathbf{\Phi}_V^{-\mathsf{T}/2}$ contains independent identically distributed Gaussian random variables, where each element of $\mathbf{V}\mathbf{\Phi}_V^{-\mathsf{T}/2}$ is a zero-mean complex Gaussian random variable with variance equal to one. Furthermore, it can be shown that $2\left\|\mathbf{V}\mathbf{\Phi}_V^{-\mathsf{T}/2}\right\|_F^2$ has a χ^2-distribution with $2MN$ degrees of freedom. Therefore, using the properties of the χ^2-distribution, it follows that the expected value and the variance of the random variable $\left\|\mathbf{V}\mathbf{\Phi}_V^{-\mathsf{T}/2}\right\|_F^2$ are both equal to MN.

In order to perform denoising, an upper bound for the noise energy will be required. Using the assumed noise model, a reasonable upper bound for the energy of the (whitened) noise can be expressed using the expected value of $\left\|\mathbf{V}\mathbf{\Phi}_V^{-\mathsf{T}/2}\right\|_F^2$, i.e., as

$$\left\|\mathbf{V}\mathbf{\Phi}_V^{-\mathsf{T}/2}\right\|_F^2 \leq MN. \tag{7.5}$$

Alternatively, in order to increase the noise suppression, a larger value for the bound has been proposed in [196] as a sum of the expected value and double the standard deviation of the corresponding random variable, i.e., as

$$\left\|\mathbf{V}\mathbf{\Phi}_V^{-\mathsf{T}/2}\right\|_F^2 \leq MN + 2\sqrt{MN}. \tag{7.6}$$

In this case, the bound on the right hand side of (7.6) implies that $\left\|\mathbf{V}\mathbf{\Phi}_V^{-\mathsf{T}/2}\right\|_F^2$ will not exceed its mean by more than two standard deviations, which holds with very high probability [196].

In the following we assume that the noise correlation matrix $\mathbf{\Phi}_V$ is known. In general, the noise correlation matrix can be estimated from a noise-only segment of the input signal $\mathbf{Y}$ or using a noise correlation matrix estimation method, e.g., [264].

7.3 Sparsity-based denoising

In this section, we formulate the problem of speech denoising using the subband signal model in (7.1), a sparsity-promoting cost function $P(.)$, and the noise model considered in Section 7.2. More specifically, the reverberant speech signal $\mathbf{X}$ can be estimated by solving the following optimization problem

$$\begin{aligned} \min_{\mathbf{X}} \quad & P(\mathbf{X}) \\ \text{subject to} \quad & \left\|(\mathbf{Y}-\mathbf{X})\,\mathbf{\Phi}_V^{-\mathsf{T}/2}\right\|_F^2 \leq \beta, \end{aligned} \tag{7.7}$$

where β is an appropriate upper bound for the noise energy, e.g., as in (7.5). Similarly as in Section 5.3, the optimization problem in (7.7) can be rewritten by introducing a splitting variable $\mathbf{Z} \in \mathbb{C}^{N \times M}$ as

$$\begin{aligned} \min_{\mathbf{D}} \quad & P(\mathbf{X}) + C_V(\mathbf{Z}) \\ \text{subject to} \quad & \mathbf{X} = \mathbf{Z}, \end{aligned} \tag{7.8}$$

where the inequality constraint in (7.7) is replaced with a convex barrier function $C_V : \mathbb{C}^{N \times M} \to \bar{\mathbb{R}}$, which is defined as

$$C_V(\mathbf{Z}) = \begin{cases} 0, & \text{if} \quad \left\| (\mathbf{Y} - \mathbf{Z}) \boldsymbol{\Phi}_V^{-\mathsf{T}/2} \right\|_F^2 \leq \beta \\ +\infty, & \qquad \text{otherwise} \end{cases} . \tag{7.9}$$

The function $C_V(.)$ is an indicator function for the feasible set of the optimization problem in (7.7). Since $P(.)$ and $C_V(.)$ are convex functions, the optimization problem in (7.8) can be efficiently solved using the ADMM algorithm. The augmented Lagrangian for the optimization problem in (7.8) can be written as

$$L_\rho(\mathbf{X}, \mathbf{Z}, \mathbf{M}) = P(\mathbf{X}) + C_V(\mathbf{Z}) + \frac{\rho}{2} \|\mathbf{X} - \mathbf{Z} + \mathbf{M}\|_F^2 - \frac{\rho}{2} \|\mathbf{M}\|_F^2, \tag{7.10}$$

where ρ is a penalty parameter and $\mathbf{M}$ is the dual variable [246]. The ADMM algorithm proceeds by minimizing $L_\rho(.)$ alternately with respect to $\mathbf{X}$ and $\mathbf{Z}$ followed by an ascent over $\mathbf{M}$ [246], i.e., in the j-th iteration we have the following update equations

$$\hat{\mathbf{X}}^j \leftarrow \operatorname{prox}_P^\rho \left(\hat{\mathbf{Z}}^{j-1} - \mathbf{M}^{j-1} \right), \tag{7.11a}$$

$$\hat{\mathbf{Z}}^j \leftarrow \operatorname{prox}_{C_V}^\rho \left(\hat{\mathbf{X}}^j + \mathbf{M}^{j-1} \right), \tag{7.11b}$$

$$\mathbf{M}^j \leftarrow \mathbf{M}^{j-1} + \eta \left(\hat{\mathbf{X}}^j - \hat{\mathbf{Z}}^j \right), \tag{7.11c}$$

where η is a parameter for faster convergence. The update for the denoised signal $\hat{\mathbf{X}}$ is obtained by computing the proximal operator of the cost function $P(.)$, cf. Section 7.5. The update for the splitting variable $\hat{\mathbf{Z}}$ is obtained by computing the proximal operator of the barrier function $C_V(.)$ in (7.9). Since the function $C_V(.)$ is an indicator function of the feasible set in (7.7), the corresponding proximal operator $\operatorname{prox}_{C_V}^\rho(.)$ is a projection on the feasible set, and is in fact independent of the penalty parameter ρ. An iterative algorithm for computing $\operatorname{prox}_{C_V}^\rho(.)$ can be found in Appendix B.2.3.

7.4 Joint dereverberation and denoising

In this section, we extend the sparsity-based denoising method presented in Section 7.3 to a joint dereverberation and denoising method, by integrating the additive noise in the MCLP-based signal model.

The subband signal model in (7.2) naturally leads to a formulation of joint denoising and dereverberation as the following optimization problem

$$\begin{aligned} \min_{\mathbf{D},\mathbf{G},\mathbf{X}} \quad & P(\mathbf{D}) \\ \text{subject to} \quad & \left\| (\mathbf{Y} - \mathbf{X}) \, \mathbf{\Phi}_V^{-\mathsf{T}/2} \right\|_F^2 \leq \beta \\ & \mathbf{D} + \tilde{\mathbf{X}}_\tau \mathbf{G} = \mathbf{X}. \end{aligned} \tag{7.12}$$

In this case, the reverberant speech $\mathbf{X}$ is obtained by denoising the microphone signal $\mathbf{Y}$, and the desired (dereverberated) speech signal is obtained by using MCLP-based dereverberation from the estimated $\mathbf{X}$, with denoising and dereverberation performed jointly. However, the constraint in (7.12) is not linear since it includes a product of the unknown convolution matrix $\tilde{\mathbf{X}}$ and the prediction filter $\mathbf{G}$, and joint estimation of the unknowns in this case is somewhat involved and computationally complex (cf. Appendix C).

Here we use an alternative MCLP-based signal model, which has also been used in the literature [159]. Since $\tilde{\mathbf{X}}_\tau = \tilde{\mathbf{Y}}_\tau - \tilde{\mathbf{V}}_\tau$, the signal model in (7.2) can be rewritten as

$$\mathbf{Y} = \mathbf{D} + \tilde{\mathbf{Y}}_\tau \mathbf{G} + \mathbf{V}_f, \tag{7.13}$$

where the filtered noise signal $\mathbf{V}_f$ is given by

$$\mathbf{V}_f = \mathbf{V} - \tilde{\mathbf{V}}_\tau \mathbf{G}. \tag{7.14}$$

The main difference between the signal model in (7.2) and the signal model in (7.13) is that the prediction filter $\mathbf{G}$ in (7.13) is applied on the delayed microphone signal $\mathbf{Y}$, and not on the (unknown) delayed reverberant signal $\mathbf{X}$ as in (7.2).

By combining the sparsity-based optimization problem in (7.7) with the signal model in (7.13) which includes both noise and reverberation, we obtain an optimization problem for joint dereverberation and denoising as

$$\begin{aligned} \min_{\mathbf{D},\mathbf{G},\mathbf{V}_f} \quad & P(\mathbf{D}) \\ \text{subject to} \quad & \mathbf{D} + \tilde{\mathbf{Y}}_\tau \mathbf{G} + \mathbf{V}_f = \mathbf{Y} \\ & \left\| \mathbf{V}_f \mathbf{\Phi}_{V_f}^{-\mathsf{T}/2} \right\| \leq \beta \end{aligned} \tag{7.15}$$

where it is assumed that the noise correlation matrix $\mathbf{\Phi}_{V_f}$ is known, and β is an appropriate bound for the noise energy (cf. Section 7.2). Note that if $\beta = 0$, the noise $\mathbf{V}_f$ is constrained to be zero, i.e., the noise is not considered and the optimization problem becomes equal to the subband MCLP formulated in (6.29).

In the optimization problem in (7.15), the prediction filter for dereverberation is applied on the delayed (noisy and reverberant) microphone signal $\mathbf{Y}$. In this case, the dereverberated but noisy speech signal is obtained using MCLP-based dereverberation from the microphone signal $\mathbf{Y}$, and the desired speech signal is then obtained by further denoising, with dereverberation and denoising performed iteratively in a joint optimization procedure. This corresponds to a processing structure composed of MCLP-based dereverberation followed by sparsity-based denoising. Similarly, a structure consisting of MCLP-based dereverberation followed by denoising has been used, e.g., in [159,163,164,265]. However, in [163,164,265], dereverberation and denoising have been performed independently, whereas in (7.15), dereverberation and denoising are performed simultaneously, similarly to the probabilistic formulation in [159].

The optimization problem for joint dereverberation and denoising in (7.15) can be rewritten as

$$\begin{aligned} \min_{\mathbf{D},\mathbf{G},\mathbf{V}_f} \quad & P(\mathbf{D}) + C_{V_f}(\mathbf{V}_f) \\ \text{subject to} \quad & \mathbf{D} + \tilde{\mathbf{Y}}_\tau \mathbf{G} + \mathbf{V}_f = \mathbf{Y}, \end{aligned} \tag{7.16}$$

where the inequality constraint in (7.15) is replaced with a barrier function $C_{V_f} : \mathbb{C}^{N \times M} \to \bar{\mathbb{R}}$, which is defined as

$$C_{V_f}(\mathbf{V}_f) = \begin{cases} 0, & \text{if} \quad \|\mathbf{V}_f \mathbf{\Phi}_{V_f}^{-\mathsf{T}/2}\|_F^2 \leq \beta \\ +\infty, & \text{otherwise} \end{cases}. \tag{7.17}$$

Since $P(.)$ and $C_{V_f}(.)$ are convex functions, the optimization problem in (7.16) can be efficiently solved using the ADMM algorithm. The augmented Lagrangian for the optimization problem in (7.16) can be written as

$$\begin{aligned} L_\rho(\mathbf{D}, \mathbf{G}, \mathbf{V}_f, \mathbf{M}) = P(\mathbf{D}) + C_{V_f}(\mathbf{V}_f) \\ + \frac{\rho}{2}\|\mathbf{D} + \tilde{\mathbf{Y}}_\tau \mathbf{G} + \mathbf{V}_f - \mathbf{Y} + \mathbf{M}\|_F^2 - \frac{\rho}{2}\|\mathbf{M}\|_F^2 \end{aligned} \tag{7.18}$$

where ρ is a penalty parameter, and $\mathbf{M}$ is the dual variable. Applying the ADMM algorithm leads to the following iterative updates for the unknown variables

$$\hat{\mathbf{D}}^j \leftarrow \arg\min_{\mathbf{D}} P(\mathbf{D}) + \frac{\rho}{2}\left\|\mathbf{D} - \left(\mathbf{Y} - \tilde{\mathbf{Y}}_\tau \hat{\mathbf{G}}^{j-1} - \hat{\mathbf{V}}_f^{j-1} - \mathbf{M}^{j-1}\right)\right\|_F^2, \tag{7.19a}$$

$$\hat{\mathbf{G}}^j \leftarrow \arg\min_{\mathbf{G}} \left\|\tilde{\mathbf{Y}}_\tau \mathbf{G} - \left(\mathbf{Y} - \hat{\mathbf{D}}^j - \hat{\mathbf{V}}_f^{j-1} - \mathbf{M}^{j-1}\right)\right\|_F^2, \tag{7.19b}$$

$$\hat{\mathbf{V}}_f^j \leftarrow \arg\min_{\mathbf{V}_f} C_{V_f}(\mathbf{V}_f) + \frac{\rho}{2}\left\|\mathbf{V}_f - \left(\mathbf{Y} - \hat{\mathbf{D}}^j - \tilde{\mathbf{Y}}_\tau \hat{\mathbf{G}}^j - \mathbf{M}^{j-1}\right)\right\|_F^2, \tag{7.19c}$$

which are followed by an update for the dual variable $\mathbf{M}$. The update for the desired speech $\hat{\mathbf{D}}$ is obtained by computing the proximal operator of the cost function $P(.)$, cf. Section 7.5. The update for the prediction filter $\hat{\mathbf{G}}$ is a LS problem with a closed-form solution, similarly as in the ADMM-based dereverberation algorithm in (6.30b). The update for the noise signal $\hat{\mathbf{V}}_f$ is obtained by computing the proximal operator of the barrier function $C_{V_f}(.)$, cf. Appendix B.2.3. Finally, the iterative updates for the ADMM algorithm are given by

$$\hat{\mathbf{D}}^j \leftarrow \operatorname{prox}_P^{\rho}\left(\mathbf{Y} - \tilde{\mathbf{Y}}_\tau \hat{\mathbf{G}}^{j-1} - \hat{\mathbf{V}}_f^{j-1} - \mathbf{M}^{j-1}\right), \tag{7.20a}$$

$$\hat{\mathbf{G}}^j \leftarrow \left(\tilde{\mathbf{Y}}_\tau^{\mathsf{H}} \tilde{\mathbf{Y}}_\tau\right)^{-1} \tilde{\mathbf{Y}}_\tau^{\mathsf{H}} \left(\mathbf{Y} - \hat{\mathbf{D}}^j - \hat{\mathbf{V}}_f^{j-1} - \mathbf{M}^{j-1}\right), \tag{7.20b}$$

$$\hat{\mathbf{V}}_f^j \leftarrow \operatorname{prox}_{C_{V_f}}^{\rho}\left(\mathbf{Y} - \hat{\mathbf{D}}^j - \tilde{\mathbf{Y}}_\tau \hat{\mathbf{G}}^j - \mathbf{M}^{j-1}\right), \tag{7.20c}$$

$$\mathbf{M}^j \leftarrow \mathbf{M}^{j-1} + \eta\left(\hat{\mathbf{D}}^j + \tilde{\mathbf{Y}}_\tau \hat{\mathbf{G}}^j + \hat{\mathbf{V}}_f^j - \mathbf{Y}\right). \tag{7.20d}$$

In the obtained algorithm, the matrix $\tilde{\mathbf{Y}}_\tau \tilde{\mathbf{Y}}_\tau^{\mathsf{H}}$ is the same for all iterations, such that it only needs to be factored once and its factorization can be used for solving the corresponding linear system in (7.20b) in the subsequent iterations. In contrast, when employing the signal model in (7.2), a new linear system needs to be solved in each iteration, resulting in a much higher complexity per iteration (cf. Appendix C). Note that the estimated filtered noise signal $\hat{\mathbf{V}}_f$ depends on the estimated prediction filter $\hat{\mathbf{G}}$, cf. (7.14). Assuming a noise segment is available, the noise correlation matrix $\mathbf{\Phi}_{V_f}$ can then be updated after computing the current estimate of the prediction filter in (7.20b) by filtering the noise segment, e.g., as in (7.14), and estimating the noise correlation matrix $\mathbf{\Phi}_{V_f}$. Alternatively, the noise correlation matrix $\mathbf{\Phi}_{V_f}$ could be related to the noise correlation matrix $\mathbf{\Phi}_V$ using (7.14). Also, since the noise $\mathbf{V}$ is assumed to be stationary, the noise signal $\mathbf{V}_f$ will as well be stationary although temporally correlated. However, the expected value of $\|\mathbf{V}_f \mathbf{\Phi}_{V_f}^{-\mathsf{T}/2}\|_F^2$ remains the same as discussed in Section 7.2.

7.5 Sparsity-promoting cost function

Similarly as in Section 6.6, the proposed algorithms in Sections 7.3 and 7.4 are formulated in terms of a general sparsity-promoting cost function $P(.)$. Since the goal is to estimate a multi-channel speech signal ($\hat{\mathbf{X}}$ or $\hat{\mathbf{D}}$), the cost function $P(.)$ should promote sparsity over time and take into account the multi-channel structure, as the group-sparse cost function used in Chapter 4. In this chapter, we confine ourselves to the weighted $\ell_{1,2}$-norm, i.e.,

$$P(\mathbf{D}) = \sum_{n=1}^{N} \hat{w}(n) \|\mathbf{d}(n)\|_2, \tag{7.21}$$

with the weights computed from the previous reweighting iteration as

$$\hat{w}(n) = \left(\frac{1}{M}\|\hat{\mathbf{d}}(n)\|_2^2 + \varepsilon_{\min}\right)^{\frac{p-1}{2}}, \tag{7.22}$$

where p is the shape parameter and $\varepsilon_{\min}$ is the regularization parameter. The proximal operator for the weighted $\ell_{1,2}$-norm in (7.21) can be computed element-wise as (cf. Appendix B.2.2)

$$\operatorname{prox}_{P}^{\rho}\left(d_m(n)\right) = \left(1 - \frac{\rho^{-1}\hat{w}(n)}{\|\mathbf{d}(n)\|_2}\right)_{+} d_m(n), \tag{7.23}$$

which is referred to as block soft thresholding [248].

7.6 Simulations

In this section, the performance of the denoising and joint dereverberation and denoising methods proposed in Sections 7.3 and 7.4 is evaluated. The considered acoustic scenarios and the implementation details are outlined in Section 7.6.1. The influence of the penalty parameter on the performance of the proposed sparsity-based denoising method is investigated in Section 7.6.3. The influence of the penalty parameter on the performance of the proposed sparsity-based joint dereverberation and denoising method is investigated in Section 7.6.4. The performance of the proposed joint dereverberation and denoising method and two-stage methods is compared in Section 7.6.5.

7.6.1 *Acoustic scenario and algorithmic setup*

We consider the same acoustic scenario from the REVERB challenge [22, 23] used for simulations in Chapter 4, i.e., a single speech source and $M = 2$ microphones placed at a distance of about 2 m from the source. The room has a reverberation time $T_{60} \approx 700$ ms and the sampling frequency is $f_s = 16$ kHz. The reverberant signals have been generated by convolving each of the 10 speech samples (5 male and 5 female speakers) [234] with an average length of approximately 5.2 s with the measured RIRs. As in Chapter 4, the noisy reverberant signals have been generated by adding noise to the reverberant signals to achieve a desired SNR with respect to the direct speech signal at the first microphone.

Similarly to the parameter setup for the simulations in Chapter 4, the analysis and synthesis STFT is computed using a tight window based on a 64 ms Hamming window with 16 ms window shift. The prediction delay is set to $\tau = 2$ in all experiments, and the filter length is set to $L_g = 25$. The weights are regularized with $\varepsilon_{\min} = 10^{-8}$. The iterative algorithms are initialized by using the microphone signal coefficients as the initial estimate of the desired speech signal. The shape parameter for the weighted cost function is set to $p = 0.5$, cf. (7.22), the maximum number of reweighting iterations is set to $I = 20$ for all methods, the maximum number of ADMM iterations is set to $J = 50$, and the lower bound for the real-valued gain of the proximal operator is set to $G_{\min} = 0.01$. The noise correlation matrices $\mathbf{\Phi}_V$ and $\mathbf{\Phi}_{V_f}$ are estimated on a noise-only segment of the signal.

The performance is evaluated in terms of the instrumental measures described in Section 2.3, i.e., the improvement in fwsSNR (ΔfwsSNR) and PESQ (ΔPESQ). The

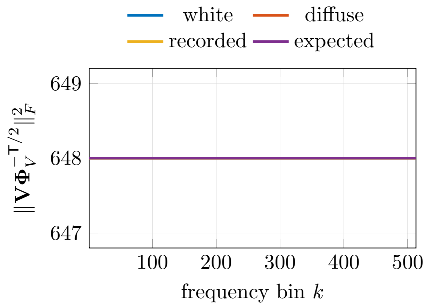


Fig. 7.1: The energy of the spatially whitened noise $\|\mathbf{V}\boldsymbol{\Phi}_V^{-\mathsf{T}/2}\|_F^2$ and its expected value in each subband. The lines for all noise types are virtually identical and equal to the expected value.

reference signal used for the instrumental measures is the direct speech signal at the microphone, obtained by convolving the anechoic speech signal with the direct signal of the corresponding RIR. The reported improvements of the instrumental measures are obtained by averaging over all microphones and speech samples.

7.6.2 *Validation of the noise model*

In this section, we investigate the validity of the noise model from Section 7.2. More specifically, we compute the value of the noise energy $\|\mathbf{V}\boldsymbol{\Phi}_V^{-\mathsf{T}/2}\|_F^2$ in (7.5) and compare it against its expected value MN. We consider three different types of stationary noise: Gaussian white noise (temporally and spatially uncorrelated), Gaussian diffuse noise (temporally uncorrelated and spatially diffuse), and the recorded noise described in Section 7.6.1.

The obtained results for the noise energy for different noise types are depicted in Fig. 7.1. It can be observed that for all considered noise types and in all subbands the value of $\left\|\mathbf{V}\boldsymbol{\Phi}^{-\mathsf{T}/2}\right\|_F^2$ is virtually identical to its expected value. As a practical note, the matrix $\boldsymbol{\Phi}_V^{-\mathsf{T}/2}$ for spatial whitening is never explicitly computed. Instead, we first compute the factor $\boldsymbol{\Phi}_V^{1/2}$ from the Cholesky decomposition in (7.4), and then instead use a linear solver to compute $\mathbf{V}\boldsymbol{\Phi}_V^{-\mathsf{T}/2}$. When $\boldsymbol{\Phi}_V^{-\mathsf{T}/2}$ is explicitly computed using the matrix inverse, the value of $\|\mathbf{V}\boldsymbol{\Phi}_V^{-\mathsf{T}/2}\|_F^2$ can differ significantly from its expected value, especially at low frequencies for the diffuse and the recorded noise. The reason for this is the ill-conditioning of the correlation matrix $\boldsymbol{\Phi}_V$ in low frequencies for the diffuse and the recorded noise, which results in a numerically problematic explicit computation of the inverse.

7.6.3 *Influence of the penalty parameter on the performance of sparsity-based denoising*

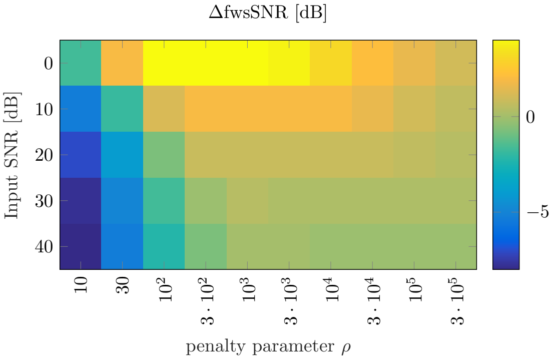
In this section, we investigate the influence of the penalty parameter ρ of the ADMM algorithm on the performance of the sparsity-based denoising method proposed in Section 7.3. We consider the described setup with input SNRs between 0 dB and 40 dB. Based on the result from the previous section, the noise bound β is set to its expected value MN.

The performance of the proposed denoising method in terms of the instrumental measures is depicted in Fig. 7.2. It can be observed that the performance of the ADMM-based denoising method strongly depends on the value of the penalty parameter ρ. Both ΔfwsSNR and ΔPESQ exhibit a similar behavior, with the performance first increasing and then decreasing with the penalty parameter ρ for each input SNR. Similarly as in Section 6.9.2, this can be related to the proximal operator in (7.23), with small values of ρ resulting in a strong suppression, while large values of ρ resulting in a weak suppression of the TF coefficients in each iteration. For the considered algorithm setup, the value of the penalty parameter $\rho = 10^3$ works well across the considered input SNRs. Furthermore, for a fixed penalty parameter ρ, the obtained performance depends on the input SNR. On the one hand, it can be observed that the proposed denoising method results in improvements over the microphone signal for low input SNRs, when additive noise is the dominant disturbance. On the other hand, there are virtually no improvements for high input SNRs, when reverberation is the dominant disturbance and the additive noise is very low.

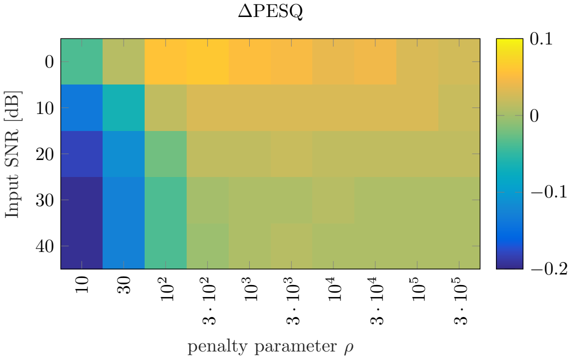
7.6.4 *Influence of the penalty parameter on the performance of sparsity-based joint dereverberation and denoising*

In this section, we investigate the influence of the penalty parameter ρ of the ADMM algorithm on the performance of the sparsity-based dereverberation and denoising method presented in Section 7.4. As in the previous section, we consider the described setup with input SNRs between 0 dB and 40 dB, and the noise bound set to its expected value MN.

The performance on the proposed joint dereverberation and denoising method in terms of the instrumental measures is depicted in Fig. 7.3. Similarly as in the previous section, it can be observed that the performance of the joint denoising and dereverberation method strongly depends on the value of the penalty parameter ρ. Again, both measures exhibit a similar behavior, with the performance first increasing and then decreasing with the penalty parameter ρ for each input SNR. For the considered algorithm setup, the value of the penalty parameter $\rho = 10^3$ works well across the considered input SNRs. Furthermore, for a fixed penalty parameter ρ, the obtained performance in general depends on the input SNR. On the one hand, it can be observed that the joint dereverberation and denoising method results in improvements in terms of ΔfwsSNR that are similar across the input SNRs. On the other hand, the improvements in terms of ΔPESQ are relatively small for low input SNRs and relatively large for high input SNRs. This behavior can be explained by a

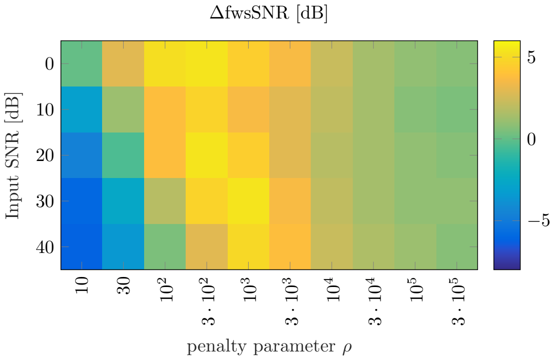


(a) ΔfwsSNR

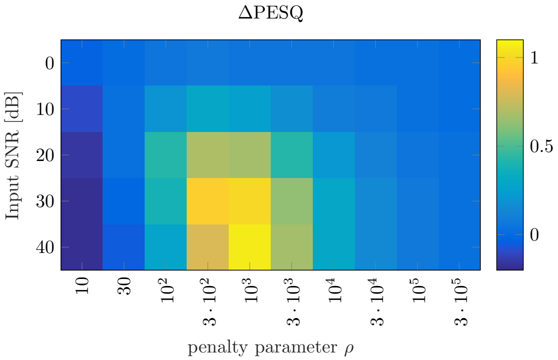


(b) ΔPESQ

Fig. 7.2: Performance of the proposed ADMM-based denoising method in terms of ΔfwsSNR (top) and ΔPESQ (bottom) for different input SNRs and values of the penalty parameter ρ.



(a) ΔfwsSNR



(b) ΔPESQ

Fig. 7.3: Performance of the proposed ADMM-based joint dereverberation and denoising method in terms of ΔfwsSNR (top) and ΔPESQ (bottom) for different input SNRs and values of the penalty parameter ρ.

combined effect of dereverberation and denoising, with the effect of denoising being dominant for low input SNRs, as in the previous section (cf. Fig. 7.2), and the effect of dereverberation being dominant for high input SNRs.

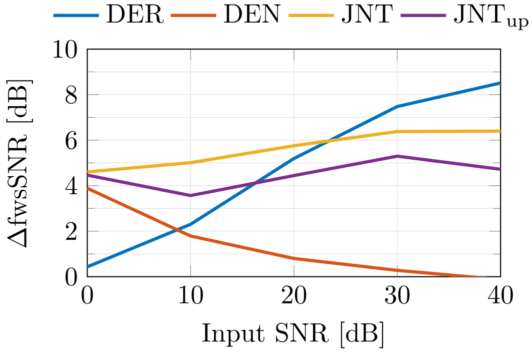
7.6.5 *Comparison of sparsity-based dereverberation and denoising methods*

In this section, we compare the performance of the proposed denoising and dereverberation methods with the MCLP-based dereverberation method from Chapter 4, which does not take into account the additive noise. Furthermore, we compare the performance of the proposed joint dereverberation and denoising method with two different two-stage systems for dereverberation and denoising, consisting of the MCLP-based dereverberation method and the proposed sparsity-based denoising method. The MCLP-based dereverberation is implemented as described in

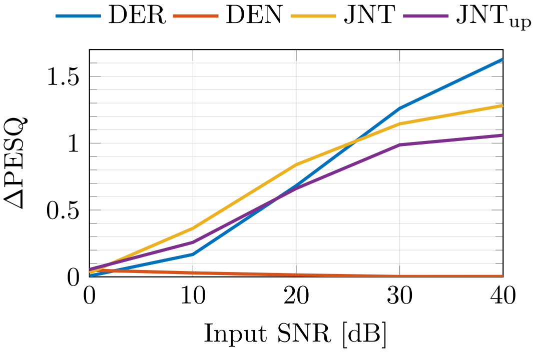
Section 4.4.1, with the identity within-group correlation matrix and $I = 20$ iterations.

Firstly, we compare the performance of the MCLP-based dereverberation method proposed in Chapter 4 (DER), the denoising method proposed in Section 7.3 (DEN), and the joint denoising and dereverberation method proposed in Section 7.4 with the correlation matrix $\mathbf{\Phi}_{V_f}$ either updated in each iteration (JNT_{up}) or fixed to $\mathbf{\Phi}_V$ for all iterations (JNT). Using the results from the previous sections, the penalty parameter ρ is set to 10^3 for DEN and JNT_{up}, and a similar procedure is used to set the penalty parameter ρ to 300 for JNT. The results for the four considered methods are depicted in Fig. 7.4. It can be observed that all considered methods result in improvements when compared to the microphone signal for the considered input SNRs. As noted in Section 4.4.3, it can be observed that the DER method performs very well for high input SNRs, as indicated by both ΔfwsSNR and ΔPESQ values. However, the DER method results in very limited improvements over the microphone signal for low SNRs. On the contrary, the DEN method results in improvements for low input SNRs, especially in terms of ΔfwsSNR, but virtually no improvements are observed for high input SNRs. It can be further observed that the proposed joint JNT and JNT_{up} methods result in a combined effect of dereverberation and denoising, resulting in improvements over the microphone signals for all considered input SNRs. For low SNRs, the joint methods perform better than both DER and DEN. As shown by the ΔPESQ, the performance of the joint methods is slightly better than the performance of the DER method. However, as shown by the ΔfwsSNR, the joint methods achieve a significantly better performance than DER, which can be attributed to the integrated denoising. For high SNRs, the joint methods perform much better than DEN, although worse than DER. Significant improvements of the joint methods when compared to DER can be attributed to the MCLP-based dereverberation integrated in the joint methods. However, performance degradation of the joint methods when compared to DEN can be attributed to the relative inaccuracy of estimation of the prediction filter in the joint methods. This is likely influenced by the selected noise bound β and the setup of the iterative algorithm, e.g., the number of iterations of the ADMM algorithm and the proximal operator $\text{prox}^{\rho}_{C_V}(.)$ (cf. Appendix B.2.3). Reducing the noise bound β would result in a decreased performance for low SNRs, due to less denoising, but in an increased performance in high SNRs, due to a more accurate estimation of the prediction filter. Furthermore, it can be observed that the JNT method in general performs better than the JNT_{up} method, indicating that the update of the noise correlation matrix $\mathbf{\Phi}_{V_f}$ is not required and can even somewhat degrade the performance in the considered scenario.

Secondly, we compare the performance of the DER method for dereverberation and three different methods for combined dereverberation and denoising. More specifically, we compare DER with the joint JNT method, a two-stage method consisting of the MCLP-based dereverberation method followed by the proposed denoising method (DER+DEN), and a two-stage method consisting of the proposed denoising method followed by the MCLP-based dereverberation method (DEN+DER). The penalty parameter for the denoising method in the two-stage methods has been selected similarly as in the previous sections and is set to $\rho = 10^3$. The results for the



(a) Performance in terms of ΔfwsSNR



(b) Performance in terms of ΔPESQ

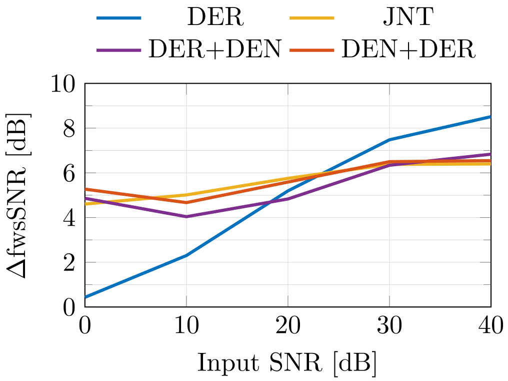
Fig. 7.4: Performance of the dereverberation-only (DER), denoising (DEN), and joint dereverberation and denoising (JNT and JNT$_{up}$) methods in terms of ΔfwsSNR (top) and ΔPESQ (bottom) for different input SNRs.

four considered methods are depicted in Fig. 7.5. It can be observed that all considered methods result in improvements when compared to the microphone signal for the considered input SNRs. Furthermore, it can be observed that all methods for combined dereverberation and denoising exhibit a similar pattern, with better performance than DER for low input SNRs and comparable or worse performance than DER for high input SNRs. By comparing the joint JNT method with its two-stage counterpart DER+DEN, it can be observed that JNT performs somewhat better in terms of ΔfwsSNR for low SNRs. However, DER+DEN performs better than JNT in terms of ΔPESQ for high SNRs. Overall, all three combined methods result in a similar performance, with the two-stage methods performing better than JNT for high input SNRs in terms of ΔPESQ. The joint method offers an improved performance over DER for low input SNRs, and a competitive performance in general. However, two-stage methods might be more appealing for practical applications, due to simpler estimation procedures and easier and more robust control of each processing stage, e.g., for individual control of the amount of denoising or dereverberation.

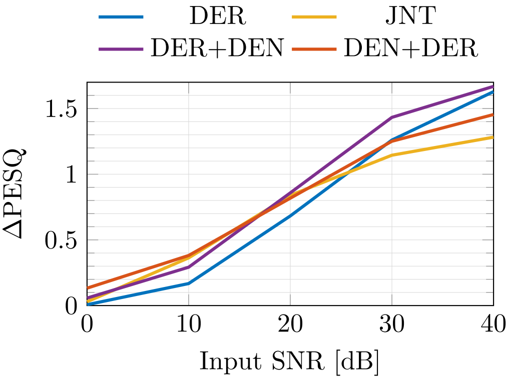
7.7 Summary

In this chapter we have considered combined denoising and dereverberation. We have proposed a batch subband method for denoising based on an optimization problem with a sparsity-promoting cost function and a constraint for the noise energy. Furthermore, we have also proposed a batch subband method for joint dereverberation and denoising by including the noise term in the MCLP-based signal model and imposing a constraint on the noise energy. We have shown that the resulting optimization problems can be efficiently solved using the ADMM algorithm.

Simulation results show that the proposed joint dereverberation and denoising method performs better than the MCLP-based dereverberation method proposed in Chapter 4 for low input SNRs. Furthermore, we have shown that two-stage methods for combined dereverberation and denoising result in a similar performance across the considered input SNRs. In conclusion, even though the joint method offers considerable improvements over MCLP-based dereverberation in low-SNR scenarios, two-stage methods seem to be more relevant in practice due to a very good performance, modularity and relative simplicity.



(a) Performance in terms of ΔfwsSNR



(b) Performance in terms of ΔPESQ

Fig. 7.5: Performance of the the dereverberation-only (DER), joint dereverberation and denoising (JNT), two-stage dereverberation and denoising (DER+DEN), and two-stage denoising and dereverberation (DEN+DER) methods in terms of ΔfwsSNR (top) and ΔPESQ (bottom) for different input SNRs.

8

CONCLUSION AND FURTHER RESEARCH

In this chapter, we provide a summary of the main contributions of the thesis in Chapters 3–7 and discuss possible extensions and directions for further research which could be envisaged as a follow-up to the work presented in this thesis.

8.1 Conclusion

Speech signals are often captured in reverberant and noisy enclosures with microphones placed at a distance from the desired speech source. This is a common scenario encountered in teleconferencing and hands-free communication systems, smart home control assistants with a voice-based interface, and assistive listening devices. Reverberation and noise present in the captured signals typically lead to a reduced effectiveness of speech communication, e.g., resulting in reduced speech quality, intelligibility, and automatic speech recognition performance.

In the context of this thesis, the main goal was to investigate and develop methods for blind speech dereverberation based on sparse multi-channel linear prediction. The proposed methods are relevant for many practical applications requiring speech dereverberation since they do not require measured or estimated RIRs.

In Chapters 3 and 4, we proposed batch dereverberation methods based on sparse MCLP and the subband signal model in the STFT domain. In Chapter 5, we proposed an adaptive dereverberation method based on constrained sparse MCLP. In Chapter 6, we proposed a general framework for dereverberation using wideband and subband signal models and sparse modeling of the TF coefficients. In Chapter 7, we proposed a joint dereverberation and denoising method based on sparse modeling and investigated combined dereverberation and denoising.

In Chapter 3, we considered the noiseless case and proposed a batch speech dereverberation method based on the subband MCLP-based signal model and a general sparse prior for the desired speech signal coefficients. We proposed to estimate the multi-channel prediction filter using an iterative algorithm that maximizes the likelihood function, resulting in an IRLS algorithm. We analytically showed that the variational representation of the sparse prior for the proposed signal model can be interpreted as a generalization of the TVG model and that the underlying prior in the conventional WPE method strongly promotes sparsity of the desired speech signal coefficients and can be obtained as a special case of the proposed framework. We reformulated the estimation of the prediction filter as a minimization of the non-convex ℓ_p-norm of the desired speech signal coefficients, for which a regularized IRLS algorithm was derived. We evaluated the dereverberation performance of the regularized and the unregularized IRLS algorithm and investigated the influence of the initialization and the number of iterations on the performance. The simulation results showed that the regularized IRLS algorithm leads to a more consistent performance than the unregularized IRLS, i.e., depends less on the initialization and the shape of the cost function, however at a price of a significantly larger number of iterations. Furthermore, we investigated the influence of the filter length, the number of microphones, and the acoustic scenario on the dereverberation performance. While the proposed MCLP-based dereverberation method can be used with a single microphone, large improvements were observed when employing two microphones and additional, albeit much smaller, improvements when using four microphones. The simulations indicated that the optimal filter length depends on the number of microphones used and the best performance for two microphones was obtained when the filter length corresponds to approximately half of the reverberation time. Overall, the best performance among the considered methods was obtained using the unregularized IRLS algorithm with a suitably selected shape parameter of the cost function, with $p = 0.5$ in general performing better than the conventional MCLP-based method (e.g., for $T_{60} \approx 700$ ms, $p = 0.5$ consistently improved ΔPESQ by 0.1). In addition to the simulation results, the proposed formulation with a sparsity-promoting cost function gives a transparent entry point for extensions and integration of MCLP-based dereverberation with other speech enhancement methods.

In Chapter 4, we proposed a multiple-output extension of the subband MCLP-based speech dereverberation method from Chapter 3 using a group sparse cost function. The MIMO prediction filter was estimated by solving a non-convex optimization problem using the IRLS algorithm, with the sparsity-promoting cost function taking into account the grouping of the coefficients across the channels. Additionally, we showed that the proposed method generalizes existing MCLP-based dereverberation methods. We evaluated the dereverberation performance of the proposed method and investigated the influence of the shape parameter of the cost function, the number of microphones and the additive noise on the performance. The simulation results for the noiseless scenario showed that the proposed method can be used to improve the dereverberation performance compared to the conventional MCLP-based method (e.g., for $T_{60} \approx 700$ ms, $p \in \{0.25, 0.5\}$ improved ΔPESQ by up to 0.2). The simulation results for the noisy scenario showed that while the method is to some extent robust to noise, the performance is rather limited for the scenarios with a relatively low SNR since the used signal model does not explicitly take into account the additive noise. Nevertheless, significant improvements can still be obtained in the moderate and high SNR scenarios, indicating that multiple-output MCLP-based dereverberation can be used as an effective pre-processor for further multi-channel signal processing, e.g., noise reduction.

In Chapter 5, we extended the batch methods from the previous chapters and proposed a constrained formulation for adaptive speech dereverberation based on group sparse MCLP in the subband domain. In general, unconstrained adaptive methods may lead to an overestimation of the undesired reverberant signal and distortions of the desired speech signal at the output. To alleviate this issue, we proposed to constrain the power of the MCLP-based undesired speech signal estimate using an estimate of the late reverberant PSD. The constrained optimization problem was solved using the ADMM algorithm, where exploiting the rank-1 updates leads to an efficient RLS-like iterative algorithm. To reduce the computational complexity for both the unconstrained and the constrained adaptive MCLP methods, we proposed to use a diagonal approximation of the weighted correlation matrix, resulting in a significantly reduced computational complexity (e.g., reducing the RTF by a factor 10 for the unconstrained method). The simulation results showed that the proposed constrained MCLP method for adaptive dereverberation is more robust to the selection of the parameters, such as the forgetting factor and the filter length, than the unconstrained method (e.g., for $T_{60} \approx 700$ ms with $M = 2$ and $\gamma = 0.9$, the constrained method improved ΔPESQ by up to 0.3). Therefore, the constrained formulation can be used to improve the performance of MCLP-based dereverberation in dynamic scenarios when the optimal forgetting factor or the filter length are not known.

In Chapter 6, we proposed a general framework for speech dereverberation using MCLP-based signal models and exploiting sparsity of the speech signal in the TF domain. More specifically, we proposed to formulate speech dereverberation by combining either a wideband or a subband signal model with an analysis or synthesis sparsity prior for the desired speech signal. All obtained optimization problems have been solved using the ADMM algorithm, and support a general sparsity-promoting function and a TF transform. Furthermore, we proposed to incorporate speech structure in the cost function through structured weights, using neighborhood and low-rank NMF weights, and reviewed the existing methods in the context of the proposed general framework. The simulation results showed that the ADMM-based method with the subband signal model with multiple reweighting iterations performs similarly as the IRLS-based method from Chapter 3, where the ADMM-based method performs better for a single reweighting iteration. It was also shown that the wideband signal model leads to a better performance compared to the subband signal model (e.g., for $T_{60} \approx 700$ ms with $M = 2$ and local weights, ΔPESQ was improved by 0.2 using the wideband model). Furthermore, it was demonstrated that using structured weights can improve the performance both for the subband and the wideband signal model, with the low-rank weights performing better than the TF neighborhood (e.g., for $T_{60} \approx 700$ ms with $M = 2$ and the subband signal model, ΔPESQ was improved by 0.1 using the neighborhood weights and by 0.2 using the low-rank weights). Overall, the best performance was obtained using the wideband signal model with analysis sparsity and low-rank weights, improving the ΔPESQ by 0.4 compared to the conventional subband MCLP-based method. Finally, the framework presented in this chapter constitutes a flexible and general formulation of MCLP-based dereverberation exploiting TF sparsity and offers many directions for further research.

As opposed to the previous chapters, where only the noiseless case was considered, in Chapter 7 we considered sparsity-based dereverberation and denoising. Firstly, we proposed to formulate speech denoising by using a sparsity-promoting cost function and by imposing a bound for the energy of the noise term. Secondly, we included the additive noise in the MCLP-based signal model and formulated joint dereverberation and denoising by combining sparse MCLP and imposing a bound for the noise energy. The proposed joint method can be seen as an extension of the group sparse MCLP-based dereverberation method from Chapter 4, taking into account additional noise term in the signal model. Both for the sparsity-based denoising method and for the joint dereverberation and denoising method, the obtained optimization problems can be solved iteratively using the ADMM algorithm. The simulation results for a reverberant and noisy scenario showed that the proposed joint dereverberation and denoising method results in an improved enhancement performance for low and moderate SNRs compared to MCLP-based dereverberation (e.g., for $T_{60} \approx 700$ ms with $M = 2$, ΔfwsSNR was improved by 4 dB). Furthermore, the simulation results showed that in noisy conditions a good performance can be obtained by combining sparse MCLP-based dereverberation with sparsity-based denoising in a two-stage procedure, with the obtained performance similar to the joint method.

8.2 Further research directions

In the following, we summarize possible research directions for further improvements and possible applications of the proposed dereverberation methods based on sparse MCLP.

In Chapter 6, it was demonstrated that exploiting the additional structure of the speech signal, e.g., using TF neighborhoods or a low-rank approximation, can be beneficial for the dereverberation performance. Therefore, the additional structure of the speech signal could be exploited for computing the weights for the reweighting procedure or for the shrinkage operators. In general, temporal averaging for computing the weights in the reweighting procedure reduces the dereverberation performance, e.g., when using large neighborhoods in the TF domain. As an alternative to explicitly defining the neighborhood, an approach based on modulation filtering could be employed for structured estimation [199]. More specifically, the weight at each TF point could be obtained by computing the strength of the filtered modulations, i.e., the weights are computed by filtering the modulation spectrum. Furthermore, instead of manually selecting the neighborhoods in the TF domain, a suitable modulation filter can be learned on a speech database [199]. Similarly, the weights could be computed using instantaneous cepstral weighting, with the model for the speech cepstral coefficients learned on a speech database [81]. In Chapter 6 we exploited the speech signal structure to compute the weights. However, the TF structure of the speech signal, e.g., neighborhood, modulation, or harmonic structure, could be also used for structured proximal operators [199, 242]. Furthermore, sparsity of the desired speech signal in the TF domain could be further increased by removing perceptually irrelevant components [266].

Although the STFT has been used in this thesis, the wideband methods proposed in Chapter 6 can easily use alternative TF transforms. Alternatively, it might be advantageous to use auditory-motivated transforms, e.g., the constant-Q transform or the ERBlet transform [251], or in general adaptive linear TF transforms [192]. Furthermore, a combination of different transforms could be used to exploit different types of structure of the desired speech signal. For example, sparse and low-rank structure of the desired speech signal could be simultaneously exploited by using a suitable dictionary [267], or a combination of different trained dictionaries for voiced and unvoiced speech could be used.

Furthermore, the proposed wideband methods for dereverberation could be integrated with wideband methods for source separation [195, 196], and computational complexity could be reduced by using faster convex optimization algorithms [231,268] and linear solvers which exploit the block-Toeplitz structure of the involved matrices [255,269]. Similarly, it is expected that joint dereverberation and denoising would benefit from including additional speech structure, either through structured weights and shrinkage operators or a structured cost function. Furthermore, wideband formulations of the joint dereverberation and denoising method may be worth investigating.

Adaptive or online processing is typically required for practical applications of speech dereverberation. At the same time, the available computational resources are often limited and the processing delay should be within a certain range. Therefore, a practically relevant and an interesting topic for further research is complexity reduction for the adaptive algorithms. While large computational savings were achieved with the diagonal approximation proposed in Chapter 5, the reduced computational complexity simultaneously resulted in a decreased performance, and improving the performance would be valuable for practical applications. For example, the performance of the adaptive methods could benefit from exploiting additional speech structure, e.g., by using the instantaneous cepstral weighting [81] for weight computation. Moreover, the performance of the adaptive MCLP-based dereverberation should be further investigated when using a low-delay filter-bank to minimize the processing delay.

Since MIMO speech dereverberation can be performed using the proposed group sparse MCLP-based methods, these could be easily applied for binaural dereverberation. Furthermore, the MCLP-based signal model implies that the desired speech signal in each channel contains a delayed and filtered direct path signal, with the relative delays between the channels being preserved. In previous studies, it was demonstrated that the accuracy of direction-of-arrival estimation can be improved by using a MIMO MCLP-based dereverberation method [161]. Therefore, it would certainly be interesting to evaluate the dereverberation performance and cue preservation of the MCLP-based methods in a binaural scenario. Furthermore, integration of the proposed MCLP-based dereverberation methods with binaural denoising [270] is an interesting and practically relevant topic for further research.

A

VARIATIONAL REPRESENTATION OF A SPARSE PRIOR

A.1 Convex variational representation

In this appendix, we give an overview of the convex variational representation of a sparse prior used in Chapter 3. Under certain conditions derived below, a circular sparse prior $\mathrm{p}(z) = e^{-f(|z|)}$ can be represented in the form (3.12) with a scaling function $\psi(.)$ as

$$\mathrm{p}(z) = \max_{\lambda>0} \mathcal{N}_{\mathbb{C}}\left(z; 0, \lambda\right) \psi\left(\lambda\right), \tag{A.1}$$

i.e., the value of p(.) at a fixed point z is obtained by maximizing the value of a scaled Gaussian $\mathcal{N}_{\mathbb{C}}\left(z; 0, \lambda\right) \psi\left(\lambda\right)$ kernel centered at zero over the variance λ. This variational representation of a sparse prior p(.) using a Gaussian kernel is often referred to as the convex type of variational representation [220, 271]. Alternatively, an integral variational representation could be used, such as mixture of scaled Gaussians [272]. The two representations are related, and in [220, 271] it has been shown that the convex representation can be used to represent a broader class of super-Gaussian priors than the integral representation.

An illustration of the convex variational approximation is given in Fig. A.1, where a CGG with $p = 0.5$ is approximated with a scaled Gaussian at several different points. The variance of the depicted scaled Gaussians at a fixed point t is obtained by maximization as in (A.1), and consequently both wider and narrower scaled Gaussians attain a lower value of $\mathrm{p}(t)$ for a fixed t.

Assuming the circular symmetry of p(.), the analysis of the convex representation in (A.1) can be restricted to the positive real-valued axis, similarly as in [220]. The negative log-probability for $t \in (0, \infty)$ can then be written as

$$-\log \mathrm{p}(t) = \inf_{\lambda>0} \frac{t^2}{\lambda} - \log \frac{\psi(\lambda)}{\pi\lambda}, \tag{A.2}$$

By defining a function $g(.)$ such that $\mathrm{p}(t) = e^{-g(t^2)}$, i.e., $f(t) = g(t^2)$, it follows that $g(.)$ can be written as

$$g(t) = \inf_{\lambda>0} \frac{t}{\lambda} - \log \frac{\psi(\lambda)}{\pi\lambda}. \tag{A.3}$$

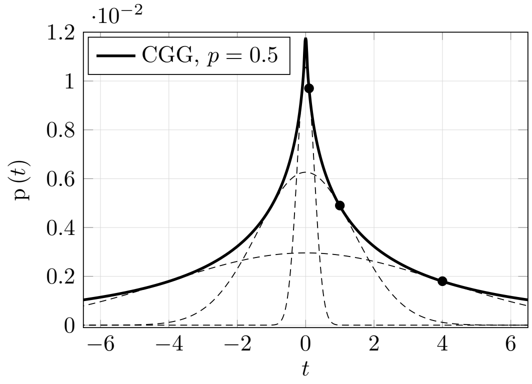


Fig. A.1: An example of a variational approximation of a sparse prior: the true sparse CGG prior with $p = 0.5$ (solid line), and scaled Gaussians at points $t \in \{0.1, 1, 4\}$ (dotted lines).

This implies that $g(.)$ is the concave conjugate of $\log \frac{\psi(\lambda)}{\pi\lambda}$, which is possible if and only if $g(.)$ is closed, increasing and concave on $(0, \infty)$ [220], and in that case the function $g(.)$ and its concave conjugate $g_\star(.)$ are related as [273]

$$g_\star(u) = \inf_t tu - g(t) \tag{A.4a}$$

$$g(t) = \inf_u tu - g_\star(u) \tag{A.4b}$$

implying that the scaling function $\psi(.)$ can be expressed as

$$\psi(\lambda) = \pi\lambda e^{g_\star(\lambda^{-1})}. \tag{A.5}$$

Note that since $g'(t) = \frac{f'(\sqrt{t})}{2\sqrt{t}}$, the concavity requirement for $g(.)$ is equivalent to $f'(t)/t$ being monotonically decreasing on $(0, \infty)$ [221, 222, 273], which is characterization of a strongly super-Gaussian prior [220].

As an example, consider a CGG prior in (3.18) with $p \in (0, 2)$. In this case, (3.19) implies that the function $g(.)$ is equal to

$$g(t) = \frac{t^{p/2}}{\zeta^{p/2}} - \log \frac{p}{2\pi\zeta\Gamma(2/p)}. \tag{A.6}$$

Since the concave conjugate of $\frac{t^{p/2}}{p/2}$ is $\frac{t_\star^{q/2}}{q/2}$ with $\frac{2}{p} + \frac{2}{q} = 1$, and the concave conjugate of $ag(t) + b$, $a > 0$, is $ag_\star\left(\frac{u}{a}\right) - b$, the concave conjugate of $g(.)$ can be easily expressed as

$$g_\star(u) = \left(\frac{p}{2}\right)^{1-q/2} \frac{\zeta^{q/2}}{q/2} u^{q/2} + \log \frac{p}{2\pi\zeta\Gamma(2/p)}, \tag{A.7}$$

resulting in the following scaling function $\psi(.)$ for the CGG prior

$$\psi\left(\lambda\right) = \frac{p}{2\zeta\Gamma(2/p)} \lambda e^{\left(\frac{p}{2}\right)^{1-q/2} \frac{\zeta^{q/2}}{q/2} \lambda^{-q/2}}. \tag{A.8}$$

A.2 Variance estimation

In the variance estimation step in Chapter 3, an optimization problem had to be solved to compute the optimal variance $\hat{\lambda}$ in (3.16) as

$$\hat{\lambda} = \arg\min_{\lambda>0} \frac{|\hat{d}|^2}{\lambda} + \log \pi\lambda - \log \psi\left(\lambda\right). \tag{A.9}$$

Using (A.5), this optimization problem can be rewritten as

$$\hat{\lambda} = \arg\min_{\lambda>0} \frac{t^2}{\lambda} - g_\star\left(\lambda^{-1}\right), \tag{A.10}$$

for some $t \geq 0$, with $g_\star\left(\lambda^{-1}\right) = \log \psi(\lambda) - \log \pi\lambda$. Since $g_\star(.)$ is concave, the cost function in the previous expression is strongly convex and the global minimum can be easily found. Hence, the optimal variance $\hat{\lambda}$ is equal to

$$\hat{\lambda} = \frac{1}{\left(g'_\star\right)^{-1}\left(t^2\right)}, \tag{A.11}$$

where $\left(g'_\star\right)^{-1}(.)$ is the inverse function of $g'_\star(.)$. Based on the results from convex analysis, the following holds for the derivative $g'(.)$ [220, 273]

$$g'(t) = \arg\min_{u} tu - g_\star(u) \tag{A.12}$$

and the inverse of $g'_\star(.)$ can be expressed using $g'(.)$ as

$$\left(g'_\star\right)^{-1}(t) = g'(t). \tag{A.13}$$

Since $f(t) = g(t^2)$, the optimal variance $\hat{\lambda}$ can hence be written as

$$\hat{\lambda} = \frac{1}{g'(t^2)} = \frac{2t}{f'(t)}, \tag{A.14}$$

as has been used in (3.17).

B

OPTIMIZATION

In this appendix, we provide a brief overview of several optimization problems of interest in the context of this thesis. In Section B.1, we discuss iteratively reweighted methods for non-convex minimization employing ℓ_2- and ℓ_1-norm reweighting. In Section B.2, we define the proximal operator and give examples of proximal operators for several functions. In Section B.3, we provide a brief overview of the alternating-direction method of multipliers (ADMM) algorithm. Two special optimization problems are considered in Sections B.4 and B.5, namely the proximal operator of the composition of the analysis operator $\mathbf{\Psi}^{\mathsf{H}}$ and a convex cost function $P(.)$ and the LASSO problem, respectively.

B.1 Iteratively reweighted methods for non-convex minimization

Consider an optimization problem in the following form

$$\begin{aligned} \min_{\mathbf{x}} \quad & \|\mathbf{B}\mathbf{x}\|_p^p \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{c}, \end{aligned} \tag{B.1}$$

where $p \in (0,1)$, $\mathbf{x} \in \mathbb{C}^N$, $\mathbf{y} \in \mathbb{C}^M$, $\mathbf{A} \in \mathbb{C}^{M \times N}$, $\mathbf{B} \in \mathbb{C}^{K \times N}$, and

$$\|\mathbf{x}\|_p = \left(\sum_{n=1}^{N} |x_n|^p \right)^{\frac{1}{p}}, \tag{B.2}$$

is the ℓ_p-norm. Note that $\|.\|_p$ is not actually a norm for $p < 1$, since it is a non-convex functional, but it is still commonly referred to as the ℓ_p-norm. Optimization problems in the form (B.1) are often encountered, e.g., in sparse recovery, where $\mathbf{x}$ is the desired signal, $\mathbf{A}$ is the measurement operator, $\mathbf{c}$ are the measurements, and $\mathbf{B}$ is the analysis operator.

Many algorithms for non-convex optimization in the area of sparse recovery are based on an iterative reweighting procedure [228]. The main idea is to replace the non-convex problem in (B.1) with a series of convex problems which are easily solved. The convex problems are typically obtained by selecting a convenient convex upper bound for the original cost function, and these reweighting algorithms in general fall in the category of majorization-minimization algorithms [274]. A brief review of iteratively reweighted least squares and iteratively reweighted ℓ_1-norm for ℓ_p-norm minimization is given in the following.

B.1.1 *Iteratively reweighted least squares*

The main idea in IRLS is to substitute the non-convex ℓ_p-norm-based cost function in (B.1) with a squared weighted ℓ_2-norm

$$\|\mathbf{x}\|_{\mathbf{w},2}^2 = \sum_{n=1}^{N} w_n \left|x_n\right|^2 . \tag{B.3}$$

The role of the weights $\mathbf{w}$ is to mimic the non-convex behavior of the original cost function. More specifically, the weights are computed in a such a way that the convex cost function $\|.\|_{\mathbf{w},2}^2$ is a first-order approximation of the original cost function $\|.\|_p^p$ in (B.1).

Since the ℓ_p-norm is separable and circularly symmetric, the weights can be determined by analyzing the scalar case. Consider the first-order approximation of the function $|t|^p$, $t \in \mathbb{R}$, $p \in (0,1)$, with a quadratic function $\hat{w}^i|t|^2 + \text{const.}$ at a fixed point t^i. In this case, the first-order approximation can be obtained by computing the weight w^i as

$$\hat{w}^i = \frac{p}{2}\left|t^i\right|^{p-2} . \tag{B.4}$$

Since the original cost function is concave, the obtained quadratic approximation is an upper bound for the original cost function, i.e.,

$$|t|^p \leq \hat{w}^i|t|^2 + \left(1 - \frac{p}{2}\right)\left(t^i\right)^p \tag{B.5}$$

An example of quadratic upper bounds of a non-convex cost function for the scalar case is given in Fig. B.1.

Applying IRLS on (B.1) results in the following convex optimization problem in the i-th iteration

$$\begin{aligned} \hat{\mathbf{x}}^i = \arg\min_{\mathbf{x}} \quad & \|\mathbf{B}\mathbf{x}\|_{\hat{\mathbf{w}}^i,2}^2 \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{c}, \end{aligned} \tag{B.6}$$

with the weights computed from the previous iteration as

$$\hat{w}_n^i = \left(\left|\left(\mathbf{B}\hat{\mathbf{x}}^{i-1}\right)_n\right|^2 + \varepsilon^i\right)^{\frac{p}{2}-1} , \tag{B.7}$$

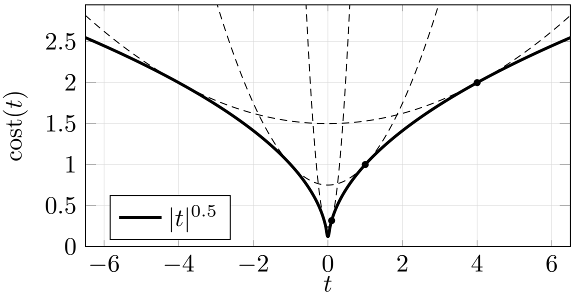


Fig. B.1: Quadratic upper bounds for a non-convex scalar cost function: the true cost function with $p = 0.5$ (solid line), and quadratic approximations at points $t^i \in \{0.1, 1, 4\}$ (dotted lines).

where ε^i is used for regularization. Note that the constant factor in (B.4) is disregarded since it does not affect the solution of (B.6). The convex optimization problem in (B.6) has a closed form solution given as

$$\hat{\mathbf{x}}^i = \left(\mathbf{B}^H \hat{\mathbf{W}}^i \mathbf{B}\right)^{-1} \left(\mathbf{A} \left(\mathbf{B}^H \hat{\mathbf{W}}^i \mathbf{B}\right)^{-1} \mathbf{A}^{\mathsf{H}}\right)^{-1} \mathbf{b}, \tag{B.8}$$

with $\hat{\mathbf{W}}^i = \operatorname{diag}\left(\hat{\mathbf{w}}^i\right)$. The IRLS algorithm iterates between updating the weights (B.7) and computing the solution (B.8). Note that the sparse recovery using the IRLS algorithm can be related to the convex variational representation of a sparse prior with the Gaussian kernel (cf. Appendix A) [233]. More details about the IRLS algorithm and its applications can be found, e.g., in [223, 228].

B.1.2 *Iteratively reweighted ℓ_1-norm minimization*

The main idea in IRL1 is to substitute the non-convex ℓ_p-norm-based cost function in (B.1) with a weighted ℓ_1-norm

$$\|\mathbf{x}\|_{\mathbf{w},1} = \sum_{n=1}^{N} w_n \left|x_n\right|. \tag{B.9}$$

As in IRLS, the role of the weights $\mathbf{w}$ is to mimic the non-convex behavior of the original cost function, and they are computed in such a way that the convex cost function $\|.\|_{\mathbf{w},1}^2$ is a first-order approximation of the original cost function $\|.\|_p^p$ in (B.1).

Consider the scalar real-valued case, i.e., the first-order approximation of the function $|t|^p$, $t \in \mathbb{R}$, $p \in (0, 1)$, with the function $\hat{w}^i|t| + \text{const.}$ at a fixed point t^i. In this case, the first-order approximation can be obtained by computing the weight w^i as

$$\hat{w}^i = p\left|t^i\right|^{p-1}. \tag{B.10}$$

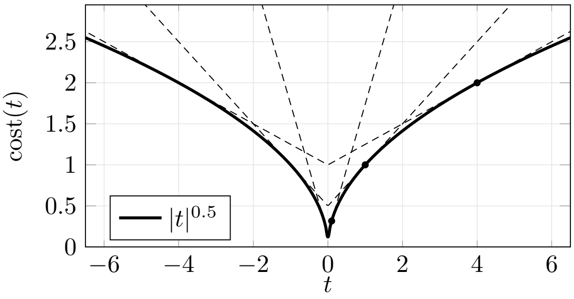


Fig. B.2: Upper bounds for a non-convex scalar cost function: the true cost function with $p = 0.5$ (solid line), and ℓ_1-norm approximations at points $t^i \in \{0.1, 1, 4\}$ (dotted lines).

Since the original cost function is concave, the obtained approximation is an upper bound for the original cost function, i.e.,

$$t^p \leq \hat{w}^i |t| + (1-p)\left(t^i\right)^p \tag{B.11}$$

An example of the upper bounds for a non-convex cost function for the scalar case is given in Fig. B.2.

Applying IRL1 on (B.1) results in the following convex optimization problem in the i-th iteration

$$\begin{aligned} \hat{\mathbf{x}}^i = \arg\min_{\mathbf{x}} \quad & \|\mathbf{B}\mathbf{x}\|_{\hat{\mathbf{w}}^i,1} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{c}, \end{aligned} \tag{B.12}$$

with the weights computed from the previous iteration as

$$\hat{w}_n^i = \left(\left|\left(\mathbf{B}\hat{\mathbf{x}}^{i-1}\right)_n\right|^2 + \varepsilon^i\right)^{\frac{p-1}{2}}, \tag{B.13}$$

where ε^i is used for regularization. Note that the constant factor in (B.10) is disregarded since it does not affect the solution of (B.12). The IRL1 algorithm iterates between updating the weights (B.13) and computing the solution by solving (B.12). The IRL1 algorithm can be seen as a non-smooth alternative to IRLS [274]. Note that the sparse recovery using the IRLS algorithm can be related to the convex variational representation of a sparse prior with the Laplacian kernel (similarly to Appendix A) [176]. More details about the IRL1 algorithm and its applications can be found, e.g., in [228, 229].

B.2 Proximal operator

An important ingredient of many optimization algorithms, e.g. the proximal algorithms and the ADMM, is the proximal operator [246, 248]. The proximal operator

$\text{prox}_P^\rho(.)$ of a closed, proper, and convex function $P(.)$ with the penalty parameter $\rho > 0$ can be defined as

$$\text{prox}_P^\rho(\mathbf{v}) = \arg\min_{\mathbf{x}} P(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{v} - \mathbf{x}\|_2^2, \tag{B.14}$$

For a fixed $\mathbf{v}$, a unique minimizer $\text{prox}_P^\rho(\mathbf{v})$ always exists since the function to be minimized is strongly convex and not everywhere infinite [248]. The result of applying a proximal operator is sometimes referred to as a proximal point, i.e., $\text{prox}_P^\rho(\mathbf{v})$ is a proximal point of of $\mathbf{v}$, since it is obtained as a compromise between being close to $\mathbf{v}$ and minimizing the cost function $P(.)$. Assuming that $P(.)$ is an indicator function of a closed convex set, the proximal operator is equal to the Euclidean projection on that set, and the proximal operator $\text{prox}_P^\rho(.)$ can be seen as a generalization of a projection operator [246, 248]. Among other interpretations, idea of the proximal operator can also be related to Moreau-Yosida regularization and modified gradient methods [248]. For some, relatively simple, functions $P(.)$, the proximal operator in (B.14) can be evaluated analytically and very efficiently [246, 248].

For example, consider a function $P : \mathbb{C} \mapsto \mathbb{R}$ defined as

$$P(x) = w\,|x|^2, \tag{B.15}$$

with $w > 0$, i.e., $P(.)$ is the scalar equivalent of the weighted squared ℓ_2-norm. In this case, a simple closed-form solution for the proximal operator $\text{prox}_P^\rho : \mathbb{C} \mapsto \mathbb{C}$ can be derived as

$$\text{prox}_{w|.|^2}^\rho(v) = \arg\min_x w|x|^2 + \frac{\rho}{2}|x - v|^2 = \frac{1}{1 + 2\rho^{-1}w} \cdot v, \tag{B.16}$$

which is sometimes referred to as the shrinkage operator, since it shrinks the values of v towards zero [242, 248]. Since the weighted ℓ_2-norm of a vector is separable, the proximal point $\text{prox}_{\|\cdot\|_{\mathbf{w},2}^2}^\rho(\mathbf{v})$ can be computed by applying the scalar proximal operator (B.16) on the elements of the vector $\mathbf{v}$.

As another example, consider a function $P : \mathbb{C} \mapsto \mathbb{R}$ defined as

$$P(x) = w\,|x|, \tag{B.17}$$

i.e., $P(.)$ is the scalar equivalent of the weighted ℓ_1-norm. In this case, a simple closed-form solution for the proximal operator $\text{prox}_P^\rho : \mathbb{C} \mapsto \mathbb{C}$ can be derived as

$$\text{prox}_{w|.|}^\rho(v) = \arg\min_x w|x| + \frac{\rho}{2}|x - v|^2 = \max\left(1 - \frac{\rho^{-1}w}{|v|}, 0\right) \cdot v, \tag{B.18}$$

with $w > 0$, i.e., the proximal operator $\text{prox}_P^\rho(.)$ is the soft thresholding operator with the threshold $\rho^{-1}w$. Since the weighted ℓ_1-norm of a vector is separable, the proximal point $\text{prox}_{\|\cdot\|_{\mathbf{w},1}}^\rho(\mathbf{v})$ can be computed by applying the scalar proximal operator (B.18) on the elements of the vector $\mathbf{v}$.

B.2.1 *Soft thresholding with a minimum gain and the corresponding cost function*

In this appendix, analytical expression for the cost function corresponding to the proximal operator in (6.21) with a lower bound on the gain is derived. Consider a proximal operator $\mathrm{prox}_P^{\rho} : \mathbb{C} \mapsto \mathbb{C}$ for a penalty function $P : \mathbb{C} \mapsto \mathbb{R}$ of a complex scalar $v \in \mathbb{C}$, defined as

$$\mathrm{prox}_P^{\rho}(v) = \max\left(1 - \frac{\rho^{-1}w}{|v|}, G_{\min}\right) \cdot v, \tag{B.19}$$

i.e., the proximal mapping is the soft thresholding with a lower bound $G_{\min}$ on the real-valued gain. The lower bound $G_{\min}$ basically prevents the shrinkage of the non-zero values to zero. Since the function $\mathrm{prox}_P^{\rho}(.)$ is circularly symmetric, the analysis can be restricted to the positive part of the real axis, i.e., denoting the independent variable as t, with $t \in \mathbb{R}, t > 0$. As in [275], a function $f(.)$ can be defined as $f(t) = \int_0^t \mathrm{prox}_P^{\rho}(\xi)\, d\xi$, $t > 0$. For the given mapping in (B.19), the function f can be written as

$$f(t) = \begin{cases} \frac{1}{2} G_{\min} t^2, & \text{for } t < \frac{\rho^{-1}w}{1-G_{\min}} \\ \frac{1}{2}\left(t - \rho^{-1}w\right)^2 + \frac{1}{2}\frac{\rho^{-2}w^2 G_{\min}}{1-G_{\min}}, & \text{for } t \geq \frac{\rho^{-1}w}{1-G_{\min}} \end{cases} \tag{B.20}$$

Following [275], the corresponding cost function can be obtained as

$$P(t) = \rho f^{\star}(t) - \rho\frac{t^2}{2}, \tag{B.21}$$

where $f^{\star}(.)$ is the convex conjugate of $f(.)$, i.e., [273]

$$f^{\star}(u) = \sup_t tu - f(t). \tag{B.22}$$

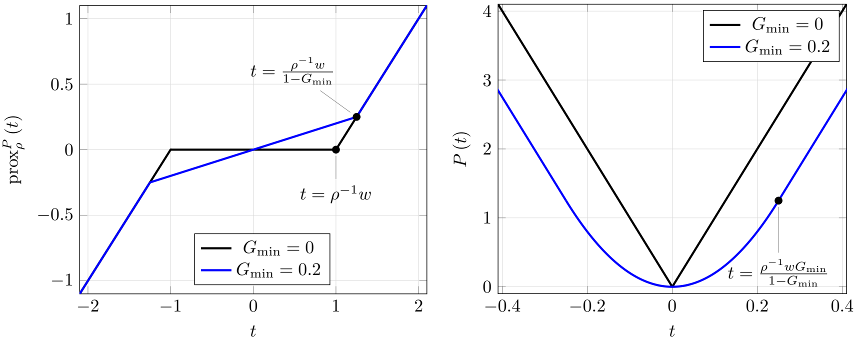
By observing that $f(.)$ is a convex continuous piecewise quadratic function, its convex conjugate can be obtained using [276] as

$$f^{\star}(t) = \begin{cases} \frac{1}{2}\frac{t^2}{G_{\min}}, & \text{for } t < \frac{\rho^{-1}wG_{\min}}{1-G_{\min}} \\ \rho^{-1}wt + \frac{1}{2}t^2 - \frac{1}{2}\frac{\rho^{-2}w^2G_{\min}}{1-G_{\min}}, & \text{for } t \geq \frac{\rho^{-1}wG_{\min}}{1-G_{\min}} \end{cases} \tag{B.23}$$

Finally, by combining with (B.21) the cost function can be written as

$$P(t) = \begin{cases} \rho\frac{1-G_{\min}}{G_{\min}}\frac{t^2}{2}, & \text{for } t < \frac{\rho^{-1}wG_{\min}}{1-G_{\min}} \\ wt - \frac{1}{2}\frac{\rho^{-1}w^2G_{\min}}{1-G_{\min}}, & \text{for } t \geq \frac{\rho^{-1}wG_{\min}}{1-G_{\min}} \end{cases} \tag{B.24}$$

Without the minimum-gain bound, i.e., when $G_{\min} = 0$, the proximal operator in (B.19) reduces to soft thresholding, and the cost function is a linear function of t, corresponding to the weighted ℓ_1-norm. When $G_{\min} > 0$ the cost function is quadratic for small values of t and linear for large values of t, i.e., $P(.)$ has a form of the Huber function [263]. The point of transition between the linear and quadratic



(a) Proximal operator

(b) Cost function

Fig. B.3: Proximal operators (left) and the corresponding cost functions (right) on the real axis with $G_{\min} = \{0, 0.2\}$ and $\rho = w = 10$. For complex-valued inputs the proximal operator and the penalty function are extended using circular symmetry.

behavior depends on the parameters ρ and w, i.e., on the threshold of the soft-thresholding operator. The inclusion of the lower-bound $G_{\min}$ for the real-valued gain can be interpreted as obtaining a smooth approximation of the ℓ_1-norm [248]. An illustration of the effect of the lower bound on the proximal operator and the cost function is given in Fig. B.3.

B.2.2 *Proximal operators for group-sparse penalties*

In this appendix we briefly review proximal operators for two group sparse penalties. More specifically, we give proximal operator for the weighted $\ell_{2,2}$-norm and the weighted $\ell_{1,2}$-norm. For a matrix argument, the proximal mapping of a function $P : \mathbb{C}^{N\times M} \mapsto \mathbb{R}$ can be defined analogously to (B.14) as

$$\mathrm{prox}_P^\rho(\mathbf{V}) = \arg\min_{\mathbf{X}} P(\mathbf{X}) + \frac{\rho}{2}\|\mathbf{V}-\mathbf{X}\|_F^2 . \tag{B.25}$$

Firstly, we consider the weighted $\ell_{2,2}$-norm, i.e.,

$$P(\mathbf{X}) = \sum_{n=1}^{N} w(n)\|\mathbf{x}(n)\|_2^2, \tag{B.26}$$

where $\mathbf{x}(n) \in \mathbb{C}^M$ contains the elements from the n-th row of $\mathbf{X}$. The corresponding proximal operator is given element-wise as

$$\mathrm{prox}_P^\rho(v_m(n)) = \frac{1}{1+2\rho^{-1}w(n)} \cdot v_m(n) \tag{B.27}$$

Secondly, we consider the weighted $\ell_{1,2}$-norm, i.e.,

$$P(\mathbf{X}) = \sum_{n=1}^{N} w(n) \|\mathsf{x}(n)\|_2, \tag{B.28}$$

which is well known as group LASSO or joint sparsity [242]. The corresponding proximal operator is given element-wise as

$$\operatorname{prox}_P^{\rho}(v_m(n)) = \max\left(1 - \frac{\rho^{-1} w(n)}{\|\mathsf{v}(n)\|_2}, 0\right) \cdot v_m(n), \tag{B.29}$$

and it is sometimes referred to as the block soft thresholding [248].

B.2.3 *Proximal operator/projection on a weighted norm ball*

In this appendix we briefly review the proximal operator of the function $C : \mathbb{C}^{N\times M} \mapsto \mathbb{R}$ which has the following form

$$C(\mathbf{X}) = \begin{cases} 0, & \text{if} \quad \|\mathbf{XB}-\mathbf{Y}\|_F^2 \leq \beta, \\ +\infty, & \quad \text{otherwise} \end{cases}, \tag{B.30}$$

where $\mathbf{B}$ is a full-rank matrix. The function $C(.)$ is an indicator function for the set of all matrices $\mathbf{X}$ which satisfy the constraint $\|\mathbf{XB}-\mathbf{Y}\|_F^2 \leq \beta$. The corresponding proximal mapping is

$$\operatorname{prox}_C^{\rho}(\mathbf{V}) = \arg\min_{\mathbf{X}} C(\mathbf{X}) + \frac{\rho}{2}\|\mathbf{V}-\mathbf{X}\|_F^2 = \arg\min_{\mathbf{X}:\|\mathbf{XB}-\mathbf{Y}\|_F^2 \leq \beta} \|\mathbf{V}-\mathbf{X}\|_F^2 \tag{B.31}$$

which is a projection of the matrix $\mathbf{V}$ on the convex feasible set defined by the constraint. Note that the projection does not depend on the penalty parameter ρ. The projection can be computed by applying Lemma 2 from [196], resulting in the following iterative updates

$$\mathbf{T}^j \leftarrow \frac{1}{\nu}\mathbf{U}^{j-1} + \mathbf{P}^{j-1}\mathbf{B} - \mathbf{Y} \tag{B.32a}$$

$$\mathbf{U}^j \leftarrow \nu\left(\mathbf{T}^j - \operatorname{prox}_{i_{\|\cdot\|_F^2 \leq \beta}}\left(\mathbf{T}^j\right)\right) \tag{B.32b}$$

$$\mathbf{P}^j \leftarrow \mathbf{V} - \mathbf{U}^j\mathbf{B}^{\mathsf{H}}, \tag{B.32c}$$

where $0 < \nu < 2/\|\mathbf{BB}^{\mathsf{H}}\|_2$. The mentioned lemma states that $\mathbf{P}^j$ converges to $\operatorname{prox}_C(\mathbf{V})$ linearly [196]. The proximal operator $\operatorname{prox}_{i_{\|\cdot\|_F^2 \leq \beta}}(.)$ of the indicator function of the set $\|\mathbf{X}\|_F^2 \leq \beta$ is defined as

$$\operatorname{prox}_{i_{\|\cdot\|_F^2 \leq \beta}} = \arg\min_{\mathbf{X}:\|\mathbf{X}\|_F^2 \leq \beta} \|\mathbf{V}-\mathbf{X}\|_F^2 = \min\left(1, \frac{\sqrt{\beta}}{\|\mathbf{V}\|_F}\right) \cdot \mathbf{V}. \tag{B.33}$$

B.3 Alternating direction method of multipliers

In this appendix we present a brief overview of the ADMM algorithm. A detailed overview of the ADMM algorithm and its applications can be found in [246]. The ADMM algorithm is suitable for non-smooth convex optimization problems with linear equality constraints in the following form

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}} \quad & P(\mathbf{x}) + Q(\mathbf{y}) \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} = \mathbf{c}. \end{aligned} \tag{B.34}$$

The cost function in the optimization problem in (B.34) conveniently splits into the const functions $P(.)$ and $Q(.)$, which depend on variables $\mathbf{x}$ and $\mathbf{y}$, respectively. The augmented Lagrangian for the constrained optimization problem in (B.34) can be written as

$$L_\rho(\mathbf{x},\mathbf{y},\boldsymbol{\mu}) = P(\mathbf{x}) + Q(\mathbf{y}) + \frac{\rho}{2}\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} - \mathbf{c} + \boldsymbol{\mu}\|_2^2 - \frac{\rho}{2}\|\boldsymbol{\mu}\|_2^2, \tag{B.35}$$

where $\boldsymbol{\mu}$ denotes the dual variable and $\rho > 0$ denotes a penalty parameter. The iterative ADMM algorithm can be obtained by minimizing the augmented Lagrangian L_ρ in (B.35) alternately with respect to $\mathbf{x}$ and $\mathbf{y}$, followed by a dual ascent over $\boldsymbol{\mu}$ [246]. This leads to the following updates

$$\mathbf{x}^j \leftarrow \arg\min_{\mathbf{x}} P(\mathbf{x}) + \frac{\rho}{2}\left\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y}^{j-1} - \mathbf{c} + \boldsymbol{\mu}^{j-1}\right\|_2^2, \tag{B.36a}$$

$$\mathbf{y}^j \leftarrow \arg\min_{\mathbf{y}} Q(\mathbf{y}) + \frac{\rho}{2}\left\|\mathbf{A}\mathbf{x}^j + \mathbf{B}\mathbf{y} - \mathbf{c} + \boldsymbol{\mu}^{j-1}\right\|_2^2, \tag{B.36b}$$

$$\boldsymbol{\mu}^j \leftarrow \boldsymbol{\mu}^{j-1} + \eta\left(\mathbf{A}\mathbf{x}^j + \mathbf{B}\mathbf{y}^j - \mathbf{c}\right), \tag{B.36c}$$

which are iteratively repeated, with j denoting the iteration index. The parameter $\eta \geq 1$ can be used for a faster convergence, and should be smaller than $1 + \sqrt{5}/2$ for a convex cost function [231, 277].

The convergence of the ADMM algorithm has been discussed, e.g., in [246]. For example, assuming that $P(.)$ and $Q(.)$ are closed, proper, and convex real-valued functions and the (non-augmented) Lagrangian has a saddle point, it can be shown that as $j \to \infty$, the iterates $\mathbf{x}^j, \mathbf{y}^j$ approach feasibility, the objective $P\left(\mathbf{x}^j\right) + Q\left(\mathbf{y}^j\right)$ approaches the optimal value, and the dual variable $\boldsymbol{\mu}^j$ approaches the dual optimal point [246]. It is known that the ADMM algorithm can be very slow to converge to high accuracy, requiring many iterations. However, in many cases it converges to modest accuracy already after a few tens of iterations, which is often enough for practical applications [246, 255]. The penalty parameter ρ may have a large effect on the convergence of the algorithm and typically depends on the particular choice of $P(.)$ and $Q(.)$. Although selection of the penalty parameter ρ has been often analyzed [246], its selection is in practice based on heuristics and is considered to be a kind of dark art [32]. Hence, when using a limited number of iterations, an appropriate value for the penalty parameter needs to be selected based on the observed convergence of the performance criteria of interest.

B.4 Proximal operator of the composition of an analysis operator and a convex function

This appendix presents a brief overview of the optimization problem involved in estimation of the time-domain signal in (6.5a), which has a special form of the generalized LASSO [246]. Consider an optimization problem in the form

$$\min_{\underline{\mathbf{x}}} \quad P\left(\mathbf{\Psi}^{\mathsf{H}}\underline{\mathbf{x}}\right) + \frac{\rho}{2}\left\|\underline{\mathbf{x}} - \underline{\mathbf{y}}\right\|_2^2, \tag{B.37}$$

where $P(.)$ is a convex function, and $\underline{\mathbf{x}}, \underline{\mathbf{y}} \in \mathbb{R}^T$.

The minimizer of the cost function in (B.37) is equal to the proximal operator of the composition of the analysis operator $\mathbf{\Psi}^{\mathsf{H}}$ and the function $P(.)$, i.e., the optimal $\underline{\mathbf{x}}$ is equal to $\operatorname{prox}^{\rho}_{P\circ\mathbf{\Psi}^{\mathsf{H}}}\left(\underline{\mathbf{y}}\right)$, with $P \circ \mathbf{\Psi}^{\mathsf{H}}$ denoting the composition of the analysis operator $\mathbf{\Psi}^{\mathsf{H}}$ and the cost function $P(.)$, i.e., $\left(P \circ \mathbf{\Psi}^{\mathsf{H}}\right)\left(\underline{\mathbf{x}}\right) = P\left(\mathbf{\Psi}^{\mathsf{H}}\underline{\mathbf{x}}\right)$ [278]. The optimization problem in (B.37) can be rewritten as

$$\begin{aligned} \min_{\underline{\mathbf{x}},\mathbf{u}} \quad & P\left(\mathbf{u}\right) + \frac{\rho}{2}\left\|\underline{\mathbf{x}} - \underline{\mathbf{y}}\right\|_2^2 \\ \text{subject to} \quad & \mathbf{\Psi}^{\mathsf{H}}\underline{\mathbf{x}} - \mathbf{u} = \mathbf{0}, \end{aligned} \tag{B.38}$$

where $\mathbf{u} \in \mathbb{C}^{KN}$ is the splitting variable, and the corresponding augmented Lagrangian L_ρ is

$$L_\rho\left(\underline{\mathbf{x}}, \mathbf{u}, \boldsymbol{\mu}\right) = P\left(\mathbf{u}\right) + \frac{\rho}{2}\left\|\underline{\mathbf{x}} - \underline{\mathbf{y}}\right\|_2^2 + \frac{\delta}{2}\left\|\mathbf{\Psi}^{\mathsf{H}}\underline{\mathbf{x}} - \mathbf{u} + \boldsymbol{\mu}\right\|_2^2 - \frac{\delta}{2}\|\boldsymbol{\mu}\|_2^2 \tag{B.39}$$

with the penalty parameter $\delta > 0$. Applying the ADMM algorithm results in the following iterative updates

$$\underline{\mathbf{x}}^j \leftarrow \frac{1}{\rho + \delta}\left(\rho\underline{\mathbf{y}} + \delta\mathbf{\Psi}\left(\mathbf{u}^{j-1} - \boldsymbol{\mu}^{j-1}\right)\right), \tag{B.40a}$$

$$\mathbf{u}^j \leftarrow \operatorname{prox}^{\delta}_{P}\left(\mathbf{\Psi}^{\mathsf{H}}\underline{\mathbf{x}}^j + \boldsymbol{\mu}^{j-1}\right), \tag{B.40b}$$

$$\boldsymbol{\mu}^j \leftarrow \boldsymbol{\mu}^{j-1} + \eta\left(\mathbf{\Psi}^{\mathsf{H}}\underline{\mathbf{x}}^j - \mathbf{u}^j\right), \tag{B.40c}$$

where it is assumed that $\mathbf{\Psi}$ is a Parseval frame, i.e., $\mathbf{\Psi}\mathbf{\Psi}^{\mathsf{H}} = \mathbf{I}$.

B.5 Iterative shrinkage/thresholding algorithm

This appendix presents a brief overview of the optimization problem involved in estimation of the TF coefficients in (6.10a), which has a form of the unconstrained least absolute shrinkage and selection operator (LASSO) [256] or an ℓ_1-regularized least squares problem [263]. Consider an optimization problem in the form

$$\min_{\mathbf{x}} P\left(\mathbf{x}\right) + \frac{\rho}{2}\left\|\mathbf{\Psi}\mathbf{x} - \underline{\mathbf{y}}\right\|_2^2, \tag{B.41}$$

where $P(.)$ is a weighted ℓ_1-norm, $\mathbf{x} \in \mathbb{C}^{KN}$, and $\underline{\mathbf{y}} \in \mathbb{R}^{T}$. Since there is no closed-form solution for (B.41), the solution needs to be computed numerically [279].

In general, problems in the form (B.41) can be efficiently solved using proximal algorithms if the proximal operator $\mathrm{prox}_P^{\rho}(.)$ of the cost function $P(.)$ can be easily computed. A simple and effective forward-backward algorithm for solving (B.41) is the iterative shrinkage/thresholding algorithm (ISTA) [257]. The algorithm consists of repeated application of the proximal operator of $P(.)$, i.e., repeated shrinkage/thresholding, on a sequence of vectors, i.e.,

$$\mathbf{x}^{j} \leftarrow \mathrm{prox}_P^{\nu\rho}\left(\mathbf{x}^{j-1} + \frac{1}{\nu}\boldsymbol{\Psi}^{\mathsf{H}}\left(\underline{\mathbf{y}} - \boldsymbol{\Psi}\mathbf{x}^{j-1}\right)\right), \tag{B.42}$$

where $\nu = \|\boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathsf{H}}\|_2$ is the maximum eigenvalue of $\boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathsf{H}}$, e.g., if $\boldsymbol{\Psi}$ is a Parseval frame then $\nu = 1$. While being a very simple first-order method, and therefore typically not very computationally expensive per iteration, ISTA converges relatively slowly as the iterations progress [242, 257]. An accelerated version of the algorithm, namely fast ISTA (FISTA), has been proposed in [257] and consists of the following iterations

$$\mathbf{x}^{j} \leftarrow \mathrm{prox}_P^{\nu\rho}\left(\mathbf{z}^{j-1} + \frac{1}{\nu}\boldsymbol{\Psi}^{\mathsf{H}}\left(\underline{\mathbf{y}} - \boldsymbol{\Psi}\mathbf{z}^{j-1}\right)\right), \tag{B.43a}$$

$$b^{j} \leftarrow \frac{1 + \sqrt{1 + 4\left(b^{j-1}\right)^{2}}}{2} \tag{B.43b}$$

$$\mathbf{z}^{j} \leftarrow \mathbf{x}^{j} + \frac{b^{j-1} - 1}{b^{j}}\left(\mathbf{x}^{j} - \mathbf{x}^{j-1}\right) \tag{B.43c}$$

As shown in [257], FISTA has an improved convergence rate while preserving the simplicity of ISTA. The main difference is that the argument of the proximal operator $\mathrm{prox}_P^{\rho}(.)$ in FISTA is not a function of the previous point $\mathbf{x}^{j-1}$, but of a new point $\mathbf{z}^{j-1}$ which is a specific linear combination of the two previous points $\mathbf{x}^{j-1}$ and $\mathbf{x}^{j-2}$.

C

JOINT DENOISING AND DEREVERBERATION

In this appendix, we formulate the problem of speech denoising using a subband signal model in (7.2), a sparsity-promoting cost function $P(.)$, and the noise model considered in Section 7.2. More specifically, the desired speech component $\mathbf{D}$ can be estimated by solving the following optimization problem

$$\begin{aligned} \min_{\mathbf{D},\mathbf{G},\mathbf{X}} \quad & P\left(\mathbf{D}\right) \\ \text{subject to} \quad & \left\|\left(\mathbf{Y}-\mathbf{X}\right)\mathbf{\Phi}_V^{-\mathsf{T}/2}\right\|_F^2 \leq \beta \\ & \mathbf{D}+\tilde{\mathbf{X}}_\tau\mathbf{G}=\mathbf{X}. \end{aligned} \tag{C.1}$$

where it is assumed that the noise correlation matrix $\mathbf{\Phi}_V$ is known, and β is an appropriate bound for the noise energy (cf. Section 7.2).

In the optimization problem in (C.1), the prediction filter for dereverberation is applied on the uknown delayed reverberant signal $\mathbf{X}$. In this case, denoising is performed to estimate the reverberant but noiseless speech signal, which is further dereverberated to obtain an estimate of the desired speech signal, with both dereverberation and denoising performed iteratively in a joint optimization procedure. This corresponds to a processing structure composed of denoising followed by MCLP-based dereverberation, with the two blocks working jointly. However, since both the unknown filter $\mathbf{G}$ and the unknown denoised signal $\mathbf{X}$ appear in the undesired reverberant component $\tilde{\mathbf{X}}_\tau\mathbf{G}$ in (C.1), the equality constraint is not linear.

The optimization problem for joint denoising and dereverberation in (C.1) can be rewritten as

$$\begin{aligned} \min_{\mathbf{X},\mathbf{G},\mathbf{D}} \quad & P\left(\mathbf{D}\right)+C_V\left(\mathbf{X}\right) \\ \text{subject to} \quad & \mathbf{D}+\tilde{\mathbf{X}}_\tau\mathbf{G}=\mathbf{X}, \end{aligned} \tag{C.2}$$

where the inequality constraint in (C.1) is replaced with a barrier function $C : \mathbb{C}^{N\times M} \rightarrow \mathbb{R}$, which is defined as

$$C_V(\mathbf{X}) = \begin{cases} 0, & \text{if} \quad \|\left(\mathbf{Y}-\mathbf{X}\right)\mathbf{\Phi}_V^{-\mathsf{T}/2}\|_F^2 \leq \beta, \\ +\infty, & \text{otherwise.} \end{cases} \tag{C.3}$$

To decouple the variables $\mathbf{X}$ and $\mathbf{G}$, we introduce a splitting variable $\mathbf{Z}$ as

$$\begin{aligned} \min_{\mathbf{D},\mathbf{G},\mathbf{X},\mathbf{Z}} \quad & P(\mathbf{D}) + C_V(\mathbf{X}) \\ \text{subject to} \quad & \mathbf{D} + \tilde{\mathbf{Z}}_\tau \mathbf{G} = \mathbf{X} \\ & \mathbf{Z} = \mathbf{X} \end{aligned} \tag{C.4}$$

The augmented Lagrangian for the optimization problem in (C.4) can be written as

$$\begin{aligned} L_\rho(\mathbf{X},\mathbf{G},\mathbf{D},\mathbf{Z},\mathbf{M}_1,\mathbf{M}_2) = {} & P(\mathbf{D}) + C_V(\mathbf{X}) \\ & + \frac{\rho}{2}\|\mathbf{D} + \tilde{\mathbf{Z}}_\tau\mathbf{G} - \mathbf{X} + \mathbf{M}_1\|_F^2 + \frac{\rho}{2}\|\mathbf{Z} - \mathbf{X} + \mathbf{M}_2\|_F^2 \\ & - \frac{\rho}{2}\|\mathbf{M}_1\|_F^2 - \frac{\rho}{2}\|\mathbf{M}_2\|_F^2, \end{aligned} \tag{C.5}$$

where ρ is a penalty parameter, and $\mathbf{M}_1$ and $\mathbf{M}_2$ are dual variables. Applying the ADMM algorithm leads to iterative updates for the unknown variables.

The optimization problem for the desired speech signal component $\mathbf{D}$ can be written as

$$\hat{\mathbf{D}}^j \leftarrow \arg\min_{\mathbf{D}} P(\mathbf{D}) + \frac{\rho}{2}\left\|\mathbf{D} - \left(\hat{\mathbf{X}}^{j-1} - \hat{\tilde{\mathbf{Z}}}_\tau^{j-1}\hat{\mathbf{G}}^{j-1} - \mathbf{M}_1^{j-1}\right)\right\|_F^2, \tag{C.6}$$

which has a form of the proximal operator of $P(.)$ (cf. B.2.2). The optimization problem for the prediction filter $\mathbf{G}$ can be written as

$$\hat{\mathbf{G}}^j \leftarrow \arg\min_{\mathbf{G}} \left\|\hat{\tilde{\mathbf{Z}}}_\tau^{j-1}\mathbf{G} - \left(\hat{\mathbf{X}}^{j-1} - \hat{\mathbf{D}}^j - \mathbf{M}_1^{j-1}\right)\right\|_F^2, \tag{C.7}$$

which is a LS problem with a closed-form solution. The optimization problem for the denoised reverberant signal $\mathbf{X}$ can be written as

$$\begin{aligned} \hat{\mathbf{X}}^j \leftarrow \arg\min_{\mathbf{X}} & \, C_V(\mathbf{X}) \\ & + \frac{\rho}{2}\left\|\mathbf{X} - \left(\hat{\mathbf{D}}^j + \hat{\tilde{\mathbf{Z}}}_\tau^{j-1}\hat{\mathbf{G}}^j + \mathbf{M}_1^{j-1}\right)\right\|_F^2 \\ & + \frac{\rho}{2}\left\|\mathbf{X} - \left(\hat{\mathbf{Z}}^{j-1} + \mathbf{M}_2^{j-1}\right)\right\|_F^2, \end{aligned} \tag{C.8}$$

which has a form of the proximal operator $\mathrm{prox}_{C_V}^{\rho}(.)$ of $C_V(.)$ (cf. Appendix B.2.3). The optimization problem for the splitting variable $\mathbf{Z}$ can be written as

$$\begin{aligned} \hat{\mathbf{Z}}^j \leftarrow \arg\min_{\mathbf{Z}} & \, \frac{\rho}{2}\left\|\hat{\tilde{\mathbf{Z}}}_\tau\hat{\mathbf{G}}^j - \left(\hat{\mathbf{X}}^j - \hat{\mathbf{D}}^j - \mathbf{M}_1^{j-1}\right)\right\|_F^2 \\ & + \frac{\rho}{2}\left\|\mathbf{Z} - \left(\hat{\mathbf{X}}^j - \mathbf{M}_2^{j-1}\right)\right\|_F^2, \end{aligned} \tag{C.9}$$

which is a quadratic problem in terms of the elements of $\mathbf{Z}$. Let $\mathbf{z} = \mathrm{vect}(\mathbf{Z}) \in \mathbb{C}^{NM}$ be a vectorized version of the matrix $\mathbf{Z}$, with vectors $\mathbf{x}$ and $\boldsymbol{\mu}_2$ defined similarly, and

$\mathbf{x}_m, \mathbf{d}_m, \boldsymbol{\mu}_{1,m}$ denoting the m-th columns of the corresponding matrices. Then the problem for the splitting variable $\mathbf{Z}$ can be rewritten in terms of its vectorization $\mathbf{z}$ as

$$\min_{\mathbf{z}} \sum_{m=1}^{M} \left\| \hat{\tilde{\mathbf{G}}}^{j}_{m,\tau} \mathbf{z} - \left(\hat{\mathbf{x}}^{j}_{m} - \hat{\mathbf{d}}^{j}_{m} - \boldsymbol{\mu}^{j-1}_{1,m} \right) \right\|_2^2 + \left\| \mathbf{z} - \left(\hat{\mathbf{x}}^{j} - \boldsymbol{\mu}^{j-1}_{2} \right) \right\|_2^2 \tag{C.10}$$

where $\tilde{\mathbf{G}}_{m,\tau}$ is such a matrix that $\tilde{\mathbf{Z}}_{\tau} \mathbf{g}_m = \tilde{\mathbf{G}}_{m,\tau} \mathbf{z}$, i.e., $\tilde{\mathbf{G}}_{m,\tau}$ is a block-convolution matrix constructed using the vector $\mathbf{g}_m$. This results in the following quadratic problem for estimating the vectorized splitting variable $\mathbf{z}$

$$\hat{\mathbf{z}}^{j} \leftarrow \arg\min_{\mathbf{z}} \mathbf{z}^{\mathsf{H}} \mathbf{Q}^{j}_{z} \mathbf{z} - 2\Re \left\{ \mathrm{tr} \left[\mathbf{z}^{\mathsf{H}} \mathbf{r}^{j}_{z} \right] \right\} + \text{const.}, \tag{C.11}$$

where the matrix $\hat{\mathbf{Q}}^{j}_{z}$ and the vector $\hat{\mathbf{r}}^{j}_{z}$ are defined as

$$\hat{\mathbf{Q}}^{j}_{z} = \mathbf{I} + \sum_{m=1}^{M} \left(\hat{\tilde{\mathbf{G}}}^{j}_{m,\tau} \right)^{\mathsf{H}} \hat{\tilde{\mathbf{G}}}^{j}_{m,\tau} \tag{C.12a}$$

$$\hat{\mathbf{r}}^{j}_{z} = \hat{\mathbf{x}}^{j} - \boldsymbol{\mu}^{j-1}_{2} + \sum_{m=1}^{M} \left(\hat{\tilde{\mathbf{G}}}^{j}_{m,\tau} \right)^{\mathsf{H}} \left(\hat{\mathbf{x}}^{j}_{m} - \hat{\mathbf{d}}^{j}_{m} - \boldsymbol{\mu}^{j-1}_{1,m} \right). \tag{C.12b}$$

The estimate of the vectorized splitting variable $\mathbf{z}$ at j-th ADMM iteration is finally given as

$$\hat{\mathbf{z}}^{j} \leftarrow \left(\mathbf{Q}^{j}_{z} \right)^{-1} \mathbf{r}^{j}_{z}, \tag{C.13}$$

and the splitting variable $\hat{\mathbf{Z}}^{j}$ can be obtained by rearranging the elements of the vector $\hat{\mathbf{z}}^{j}$.

Finally, the complete iterative updates for the ADMM algorithm can be written as

$$\hat{\mathbf{D}}^{j} \leftarrow \mathrm{prox}^{\rho}_{P} \left(\hat{\mathbf{X}}^{j-1} - \hat{\tilde{\mathbf{Z}}}^{j-1}_{\tau} \hat{\mathbf{G}}^{j-1} - \mathbf{M}^{j-1}_{1} \right) \tag{C.14a}$$

$$\hat{\mathbf{G}}^{j} \leftarrow \left[\left(\hat{\tilde{\mathbf{Z}}}^{j-1}_{\tau} \right)^{\mathsf{H}} \hat{\tilde{\mathbf{Z}}}^{j-1}_{\tau} \right]^{-1} \left(\hat{\tilde{\mathbf{Z}}}^{j-1}_{\tau} \right)^{\mathsf{H}} \left(\hat{\mathbf{X}}^{j-1} - \hat{\mathbf{D}}^{j} - \mathbf{M}^{j-1}_{1} \right) \tag{C.14b}$$

$$\hat{\mathbf{X}}^{j} \leftarrow \mathrm{prox}^{2\rho}_{C_V} \left(\frac{\hat{\mathbf{D}}^{j} + \hat{\tilde{\mathbf{Z}}}^{j-1}_{\tau} \hat{\mathbf{G}}^{j} + \mathbf{M}^{j-1}_{1} + \hat{\mathbf{Z}}^{j-1} + \mathbf{M}^{j-1}_{2}}{2} \right), \tag{C.14c}$$

$$\hat{\mathbf{Z}}^{j} \leftarrow \mathrm{matr} \left(\hat{\mathbf{z}}^{j} \right), \text{ with } \hat{\mathbf{z}}^{j} \text{ computed using (C.13)} \tag{C.14d}$$

$$\mathbf{M}^{j}_{1} \leftarrow \mathbf{M}^{j-1}_{1} + \eta \left(\hat{\mathbf{D}}^{j} + \hat{\tilde{\mathbf{Z}}}^{j}_{\tau} \hat{\mathbf{G}}^{j} - \hat{\mathbf{X}}^{j} \right), \tag{C.14e}$$

$$\mathbf{M}^{j}_{2} \leftarrow \mathbf{M}^{j-1}_{2} + \eta \left(\hat{\mathbf{Z}}^{j} - \hat{\mathbf{X}}^{j} \right). \tag{C.14f}$$

However, joint estimation of $\mathbf{X}$, $\mathbf{G}$ and $\mathbf{D}$ using this algorithm comes with practical difficulties. Firstly, the constraint in (C.1), obtained using the signal model in (7.2), is not linear since it includes a product of the unknowns $\mathbf{X}$ and $\mathbf{G}$. Secondly, the computational complexity of the algorithm is very high due to estimation of the prediction filter $\mathbf{G}$ after denoising. More specifically, since $\mathbf{G}$ is estimated on the denoised signal, accessed through the splitting variable $\mathbf{Z}$, the matrix of the linear system in (C.7)/(C.14b) has to be recomputed in each iteration j. Therefore, a new linear system needs to be solved in each iteration, as opposed to the joint algorithm in Section 7.4, where dereverberation is performed on the noisy signal and the matrix of the linear system is fixed.

BIBLIOGRAPHY

[1] V. Pulkki and M. Karjalainen, *Communication Acoustics: An Introduction to Speech, Audio and Psychoacoustics.* Wiley, 2015.

[2] G. L. Martin, "The utility of speech input in user-computer interfaces," *International Journal of Man-Machine Studies*, vol. 30, no. 4, pp. 355–375, Mar. 1989.

[3] P. R. Cohen and S. L. Oviat, "The role of voice input for human-machine communication," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, pp. 9921–9927, Oct. 1995.

[4] I. Cohen, J. Benesty, and S. Gannot, Eds., *Speech Processing in Modern Communication.* Springer-Verlag Berlin Heidelberg, 2010.

[5] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation.* Springer, 2010.

[6] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement.* Wiley, 2017 (in preparation).

[7] T. Nakatani, W. Kellermann, P. Naylor, M. Miyoshi, and B. H. Juang, "Introduction to the special issue on processing reverberant speech: Methodologies and applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

[8] 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.

[9] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 331–342, Jul. 2006.

[10] 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.

[11] 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," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. 269–282, Jan. 2013.

[12] D. B. Roe, "Deployment of human-machine dialogue systems," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, pp. 10 017–10 022, Oct. 1995.

[13] M. Brandstein and D. Ward, Eds., *Microphone arrays: Signal processing techniques and applications.* Springer, 2001.

[14] G. M. Davis, Ed., *Noise reduction in speech applications.* CRC Press, 2002.

[15] J. Benesty, S. Makino, and J. Chen, Eds., *Speech enhancement.* Springer, 2005.

[16] P. C. Loizou, *Speech enhancement: Theory and practice.* CRC Press, 2007.

[17] J. Benesty, J. Chen, and Y. A. Huang, Eds., *Microphone array processing.* Springer, 2008.

[18] J. Benesty, J. Chen, Y. A. Huang, and I. Cohen, Eds., *Noise reduction in speech processing.* Springer, 2009.

[19] R. Hendriks, T. Gerkmann, and J. Jensen, "DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art," *Synthesis Lectures on Speech and Audio Processing*, vol. 9, no. 1, pp. 1–80, Jan. 2013.

[20] 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.

[21] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[22] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–19, 2016.

[23] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2013.

[24] H. Kuttruff, *Room acoustics.* Taylor & Francis, 2000.

[25] J. Lochner and J. F. Burger, "The influence of reflections on auditorium acoustics," *Journal of Sound and Vibration*, vol. 1, no. 4, pp. 426–454, Oct. 1964.

[26] H. Haas, "The influence of a single echo on the audibility of speech," *Journal of the Audio Engineering Society*, vol. 20, no. 2, pp. 146–159, 1972.

[27] J. S. Bradley, H. Sasto, and M. Picard, "On the importance of early reflections for speech in rooms," *The Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 3233–3244, Jun. 2003.

[28] I. Arweiler and J. M. Buchholz, "The influence of spectral characteristics of early reflections on speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 130, no. 2, pp. 996–1005, Aug. 2011.

[29] 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.

[30] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 30, pp. 12 186–12 191, 2013.

[31] A. H. Moore, M. Brookes, and P. A. Naylor, "Room geometry estimation from a single channel acoustic impulse response," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Marrakech, Morocco, Sep. 2013.

[32] I. Dokmanić, "Listening to distances and hearing shapes: Inverse problems in room acoustics and beyond," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2015.

[33] Y. Peled and B. Rafaely, "Linearly-constrained minimum-variance method for spherical microphone arrays based on plane-wave decomposition of the sound field," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 12, pp. 2532–2540, Aug. 2013.

[34] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Raking the cocktail party," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 825–836, 2015.

[35] H. A. Javed, A. H. Moore, and P. A. Naylor, "Spherical microphone array acoustic rake receivers," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Shanghai, China, Mar. 2016, pp. 111–115.

[36] R. H. Bolt and A. D. MacDonald, "Theory of speech masking by reverberation," *The Journal of the Acoustical Society of America*, vol. 21, no. 6, pp. 577–580, 1949.

[37] A. K. Nábělek, T. R. Letowski, and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification," *The Journal of the Acoustical Society of America*, vol. 86, no. 4, pp. 1259–1265, Oct. 1989.

[38] A. K. Nábělek and D. Mason, "Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms," *Journal of Speech and Hearing Research*, vol. 24, no. 3, pp. 375–383, Sep. 1981.

[39] A. K. Nábělek and P. K. Robinson, "Monoural and binaural speech perception in reverberation for listeners of various ages," *Journal of Speech and Hearing Research*, vol. 71, no. 5, pp. 1242–1248, May 1982.

[40] Y. Takata and A. K. Nábělek, "English consonant recognition in noise and in reverberation by Japanese and American listeners," *The Journal of the Acoustical Society of America*, vol. 88, no. 2, pp. 663–666, Aug. 1990.

[41] G. W. Elko, E. Diethorn, and T. Gänsler, "Room impulse response variation due to thermal fluctuation and its impact on acoustic echo cancellation," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Kyoto, Japan, Sep. 2003.

[42] J. Mourjopoulos and M. A. Paraskevas, "Pole and zero modeling of room transfer functions," *Journal of Sound and Vibration*, vol. 146, no. 2, pp. 281–302, Apr. 1991.

[43] Y. Haneda, S. Makino, and Y. Kaneda, "Common acoustical pole and zero modeling of room transfer functions," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 320–328, 1994.

[44] G. Vairetti, E. De Sena, M. Catrysse, S. Jensen, M. Moonen, and T. van Waterschoot, "A scalable algorithm for physically motivated and sparse approximation of room impulse responses with orthonormal basis functions," 2017.

[45] M. R. Schroeder, "Integrated-impulse method for measuring sound decay without using impulses," *The Journal of the Acoustical Society of America*, vol. 66, no. 2, pp. 497–500, 1979.

[46] D. D. Rife and J. Vanderkooy, "Transfer-function measurement with maximum-length sequences," *Journal of the Audio Engineering Society*, vol. 37, no. 6, pp. 419–444, 1989.

[47] N. Ream, "Nonlinear identification using inverse-repeat m sequences," *Proceedings of the Institution of Electrical Engineers*, vol. 117, no. 1, pp. 213–218, Jan. 1970.

[48] N. Aoshima, "Computer-generated pulse signal applied for sound measurement," *The Journal of the Acoustical Society of America*, vol. 69, no. 5, pp. 1484–1488, May 1981.

[49] A. J. Berkhout, M. M. Boone, and C. Kesselman, "Acoustic impulse response measurement: A new technique," *Journal of the Audio Engineering Society*, vol. 32, no. 10, pp. 740–746, Oct. 1984.

[50] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proceedings of the AES Convention*, Paris, France, Feb. 2000.

[51] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory.* Prentice Hall, 1993.

[52] I. Kodrasi, "Dereverberation and noise reduction techniques based on acoustic multi-channel equalization," Ph.D. dissertation, University of Oldenburg, Oldenburg, Germany, Dec. 2015.

[53] F. Lim, "Robust multichannel equalization for blind speech dereverberation," Ph.D. dissertation, Imperial College London, London, UK, 2016.

[54] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Transactions on Signal Processing*, vol. 43, no. 12, pp. 2982–2993, Dec. 1995.

[55] W. Xue, M. Brookes, and P. Naylor, "Cross-correlation based under-modelled multichannel blind acoustic system identification with sparsity regularization," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Budapest, Hungary, 2016, pp. 718–722.

[56] J. A. Cadzow, "Blind deconvolution via cumulant extrema," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 24–42, 1996.

[57] W. C. Sabine, *Collected Papers on Acoustics.* Harward University Press, 1922.

[58] E. A. P. Habets, “Fifty years of reverberation reduction: From analog signal processing to machine learning,” in *Proceedings of the AES 60th International Conference on DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)*, Leuven, Belgium, Feb. 2016.

[59] ——, “Single- and multi-microphone speech dereverberation using spectral enhancement,” Ph.D. dissertation, Tehnische Universiteit Eindhoven, Eindhoven, The Netherlands, Jun. 2007.

[60] Y. Ephraim, H. Lev-Ari, and W. J. J. Roberts, “A brief survey of speech enhancement,” in *The electronics handbook*, J. C. Whitaker, Ed. CRC Press, 2005.

[61] I. Cohen and S. Gannot, “Spectral enhancement methods,” in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. A. Huang, Eds. Springer, 2008.

[62] H.-G. Hirsch and C. Ehrlicher, “Noise estimation techniques for robust speech recognition,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Detroit, MI, USA, 1995, pp. 153–156.

[63] J. Sohn and W. Sung, “A voice activity detector employing soft decision based noise spectrum adaptation,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.*

[64] D. Malah, R. V. Cox, and A. J. Accardi, “Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, Mar. 1999, pp. 789–792 vol.2.

[65] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[66] 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.

[67] P. C. Loizou, “Spectral-subtractive algorithms,” in *Speech enhancement: Theory and practice.* CRC Press, 2007, ch. 5, pp. 93–136.

[68] J. Porter and S. Boll, “Optimal estimators for spectral restoration of noisy speech,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 9, San Diego, California, USA, Mar. 1984, pp. 53–56.

[69] 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.

[70] ——, “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.

[71] 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.

[72] 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.

[73] P. C. Loizou, "Statistical-model-based methods," in *Speech enhancement: Theory and practice.* CRC Press, 2007, ch. 7, pp. 203–272.

[74] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica*, vol. 87, no. 3, pp. 359–366, May-Jun 2001.

[75] J. D. Polack, "La transmission de l'energie sonore dans les salles," Ph.D. dissertation, Université du Maine, Le Mans, France, 1988.

[76] E. A. P. Habets, "Multi-channel speech dereverberation based on a statistical model of late reverberation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, USA, Mar. 2005, pp. 173–176.

[77] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–773, Jun. 2009.

[78] E. A. P. Habets, "Speech dereverberation using statistical reverberation models," in *Speech Dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds. Springer, 2010, ch. 3, pp. 57–94.

[79] A. Maezawa, K. Itoyama, K. Yoshii, and H. G. Okuno, "Nonparametric Bayesian dereverberation of power spectrograms based on infinite-order autoregressive processes," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1918–1930, Dec. 2014.

[80] N. Mohammadiha and S. Doclo, "Speech dereverberation using non-negative convolutive transfer function and spectro-temporal modeling," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 2, pp. 276–289, Feb. 2016.

[81] T. Gerkmann, "Cepstral weighting for speech dereverberation without musical noise," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Barcelona, Spain, Sep. 2011.

[82] 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. 61, pp. 1–12, Jul. 2015.

[83] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Advances in Neural Information Processing Systems 13*, Mar. 2001, pp. 758–764.

[84] H. Attias, L. Deng, A. Acero, and J. C. Platt, “A new method for speech denoising and robust speech recognition using probabilistic models for clean speech and for noise,” in *Proceedings of INTERSPEECH*, Aalborg, Denmark, Sep. 2001, pp. 1903–1906.

[85] C. S. Doire, M. Brookes, P. A. Naylor, C. M. Hicks, D. Betts, M. A. Dmour, and S. H. Jensen, “Single-channel online enhancement of speech corrupted by reverberation and noise,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 572–587, Mar. 2017.

[86] N. López, Y. Grenier, G. Richard, and I. Bourmeyster, “Single channel reverberation suppression based on sparse linear prediction,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2014, pp. 5182–5186.

[87] T. Yoshioka, “Speech enhancement in reverberant environments,” Ph.D. dissertation, Graduate School of Informatics, Kyoto University, Kyoto, Japan, 2010.

[88] E. A. Wan and A. T. Nelson, “Networks for speech enhancement,” in *Handbook of Neural Networks for Speech Processing*, S. Katagiri, Ed. Artech House, 1998, ch. 13, pp. 471–504.

[89] L. Deng and D. Yu, “Deep learning: methods and applications,” *Foundations and Trends in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[90] F. Xiong, B. T. Meyer, B. Cauchi, A. Jukić, S. Doclo, and S. Goetze, “Performance comparison of real-time single-channel speech dereverberation algorithms,” in *Proceedings of the Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, San Francisco, CA, USA, Mar. 2017.

[91] Y. Wang, A. Narayanan, and D. Wang, “On training targets for supervised speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[92] A. Maas, Q. V. Le, T. M. O’Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, “Recurrent neural networks for noise reduction in robust ASR,” in *Proceedings of INTERSPEECH*, Portland, OR, USA, Sep. 2012.

[93] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder,” in *Proceedings of INTERSPEECH*, Lyon, France, 2013, pp. 436–440.

[94] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “An experimental study on speech enhancement based on deep neural networks,” *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[95] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, “Integration of speech enhancement and recognition using long-short term memory recurrent neural network,” in *Proceedings of INTERSPEECH*, Dresden, Germany, Sep. 2015.

[96] M. Mimura, S. Sakai, and T. Kawahara, “Speech dereverberation using long short-term memory,” in *Proceedings of INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 2435–2439.

[97] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, “Learning spectral mapping for speech dereverberation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, Jun. 2015.

[98] D. S. Williamson and D. Wang, “Time-frequency masking in the complex domain for speech dereverberation and denoising,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 13, no. 9, May 2010.

[99] 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.

[100] 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.

[101] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and techniques.* Prentice Hall, 1993.

[102] 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.

[103] 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. Liu, Eds. Wiley, 2010.

[104] J. Capon, “High-resolution frequency-wavenumber spectrum analysis,” *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[105] O. L. Frost, “An algorithm for linearly constrained adaptive array processing,” *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[106] M. Brandstein and S. Griebel, *Explicit speech modeling for microphone array applications*, M. Brandstein and D. Ward, Eds. Springer, 2001.

[107] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, “Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 534–545, May 2009.

[108] A. Schwarz, K. Reindl, and W. Kellermann, “A two-channel reverberation suppression scheme based on blind signal separation and Wiener filtering,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Kyoto, Japan, Mar. 2012, pp. 113–116.

[109] 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.

[110] 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 *Proceedings of the REVERB Workshop*, Florence, Italy, May 2014.

[111] S. Wisdom, T. Powers, L. Atlas, and J. Pitton, “Enhancement of reverberant and noisy speech by extending its coherence,” in *Proceedings of the REVERB*

Workshop, Florence, Italy, May 2014.

[112] B. Cauchi, P. A. Naylor, T. Gerkmann, S. Doclo, and S. Goetze, "Late reverberant spectral variance estimation using acoustic channel equalization," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Nice, France, Sep. 2015, pp. 2481–2485.

[113] S. Braun and E. A. P. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Marrakech, Morocco, Sep. 2013.

[114] ——, "A multichannel diffuse power estimator for dereverberation in the presence of multiple sources," *EURASIP Journal on Audio, Speech, and Music Processing*, no. 1, p. 34, 2015.

[115] A. Kuklasiński, S. Doclo, T. Gerkmann, S. H. Jensen, and J. Jensen, "Maximum Likelihood PSD Estimation for Speech Enhancement in Reverberation and Noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1599–1612, Sep. 2016.

[116] I. Kodrasi and S. Doclo, "Late reverberant power spectral density estimation based on an eigenvalue problem," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, New Orleans, LA, USA, Mar. 2017, pp. 611–615.

[117] S. Doclo and M. Moonen, "Combined frequency-domain dereverberation and noise reduction technique for multi-microphone speech enhancement," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Darmstadt, Germany, Sep. 2001, pp. 31–34.

[118] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springer, 2001.

[119] A. Kuklasinski and J. Jensen, "Multichannel Wiener Filters in Binaural and Bilateral Hearing Aids: Speech Intelligibility Improvement and Robustness to DoA Errors," *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, pp. 8–16, Jan./Feb. 2017.

[120] 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.

[121] T. N. Sainatah, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin, A. Misra, and C. Kim, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, May 2017.

[122] J. Heymann, L. Drude, and R. Haeb-Umbach, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Computer Speech & Language*, 2017.

[123] 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.

[124] 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.

[125] 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.

[126] 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.

[127] 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.

[128] ——, “Subband inversion of multichannel acoustic systems,” in *Speech Dereverberation*, P. Naylor and N. D. Gaubitch, Eds. Springer, 2010, ch. 7, pp. 189–218.

[129] M. Kallinger and A. Mertins, “Multi-channel room impulse response shaping - a study,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, May 2006, pp. 101–104.

[130] W. Zhang, E. A. P. Habets, and P. A. Naylor, “On the use of channel shortening in multichannel acoustic system equalization,” in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, Sep. 2010.

[131] 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.

[132] I. Kodrasi and S. Doclo, “Robust partial multichannel equalization techniques for speech dereverberation,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Kyoto, Japan, Mar. 2012, pp. 537–540.

[133] ——, “The effect of inverse filter length on the robustness of acoustic multichannel equalization,” in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, Aug. 2012.

[134] 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.

[135] I. Kodrasi, A. Jukić, and S. Doclo, “Robust sparsity-promoting acoustic multichannel equalization for speech dereverberation,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Shanghai, China, Mar. 2016, pp. 166–170.

[136] I. Kodrasi and S. Doclo, “Signal-dependent penalty functions for robust acoustic multi-channel equalization,” *IEEE/ACM Transactions on Audio, Speech,*

and Language Processing, vol. 25, no. 7, pp. 1512–1525, Jul. 2017.

[137] ——, "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.

[138] B. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation using expectation-maximization and Kalman smoother," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Marrakech, Morocco, Sep. 2013.

[139] ——, "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.

[140] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Speech dereverberation with convolutive transfer function approximation using MAP and variational deconvolution approaches," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Antibes - Juan Les Pins, France, Sep. 2014.

[141] J. Mourjopoulos, P. Clarkson, and J. Hammond, "A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Paris, France, May 1982, pp. 1858–1861.

[142] P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Eindhoven, The Netherlands, Sep. 2005.

[143] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-domain single-channel inverse filtering for speech dereverberation: Theory and practice," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 5177–5181.

[144] D. T. Slock, "Blind fractionally-spaced equalization, perfect-reconstruction filter banks and multichannel linear prediction," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, Adelaide, Australia, 1994, pp. IV–585.

[145] D. Gesbert and P. Duhamel, "Robust blind channel identification and equalization based on multi-step predictors," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, Sep. 1997, pp. 3621–3624.

[146] M. Miyoshi, M. Delcroix, K. Kinoshita, T. Yoshioka, T. Nakatani, and T. Hikichi, "Inverse filtering for speech dereverberation without the use of room acoustics," in *Speech Dereverberation*, P. Naylor and N. D. Gaubitch, Eds. Springer, 2010, ch. 9, pp. 271–310.

[147] M. S. Brandstein, "On the use of explicit speech modeling in microphone array applications," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Seattle, WA, USA, May 1998, pp. 3613–3616.

[148] M. Triki and D. T. Slock, “Blind dereverberation of a single source based on multichannel linear prediction,” in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Eindhoven, The Netherlands, Sep. 2005, pp. 173–176.

[149] M. Triki and D. T. M. Slock, “Delay and predict equalization for blind speech dereverberation,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, May 2006, pp. V–97–V–100.

[150] M. Delcroix, T. Hikichi, and M. Miyoshi, “Precise dereverberation using multichannel linear prediction,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 430–440, Dec. 2007.

[151] ——, “Dereverberation and denoising using multichannel linear prediction,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1791–1801, Aug 2007.

[152] T. Yoshioka, T. Hikichi, and M. Miyoshi, “Dereverberation by using time-variant nature of speech production system,” *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 2, pp. 6–6, Aug. 2007.

[153] T. Nakatani, B.-H. Juang, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, “Speech dereverberation based on maximum-likelihood estimation with time-varying gaussian source model,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1512–1527, Nov. 2008.

[154] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, “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.

[155] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, USA, May 2008, pp. 85–88.

[156] T. Yoshioka, H. Tachibana, T. Nakatani, and M. Miyoshi, “Adaptive dereverberation of speech signals with speaker-position change detection,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, Apr. 2009, pp. 3733–3736.

[157] T. Yoshioka, T. Nakatani, and M. Miyoshi, “Fast algorithm for conditional separation and dereverberation,” in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Glasgow, UK, Sep. 2009, pp. 1432–1436.

[158] ——, “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.

[159] T. Yoshioka, T. Nakatani, K. Kinoshita, and M. Miyoshi, “Speech dereverberation and denoising based on time varying speech model and autoregressive reverberation model,” in *Speech Processing in Modern Communication*, I. Cohen, J. Benesty, and S. Gannot, Eds. Springer, 2010, pp. 151–182.

[160] 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.

[161] T. Yoshioka and T. Nakatani, “Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.

[162] Y. Iwata and T. Nakatani, “Introduction of speech log-spectral priors into dereverberation based on Itakura-Saito distance minimization,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Kyoto, Japan, May 2012, pp. 245–248.

[163] T. Yoshioka and T. Nakatani, “Dereverberation for reverberation-robust microphone arrays,” in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Marrakech, Morocco, Sep. 2013, pp. 1–5.

[164] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, S. Araki, T. Hori, and T. Nakatani, “Strategies for distant speech recognitionin reverberant environments,” *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 60, Jul. 2015.

[165] M. Togami and Y. Kawaguchi, “Noise robust speech dereverberation with Kalman smoother,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 2013, pp. 7447–7451.

[166] 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. 2014.

[167] M. Togami, “Multichannel online speech dereverberation under noisy environments,” in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Nice, France, Sep. 2015, pp. 1083–1087.

[168] N. Ito, S. Araki, and T. Nakatani, “Probabilistic integration of diffuse noise suppression and dereverberation,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 5167–5171.

[169] T. Dietzen, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, “Partitioned block frequency domain Kalman filter for multi-channel linear prediction based blind speech dereverberation,” in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Xi’an, China, Sep. 2016.

[170] S. Braun and E. A. P. Habets, “Online dereverberation for dynamic scenarios using a Kalman filter with an autoregressive model,” *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1741–1745, Dec. 2016.

[171] M. Parchami, W.-P. Zhu, and B. Champagne, “Speech dereverberation using weighted prediction error with correlated inter-frame speech components,”

Speech Communication, vol. 87, pp. 49–57, Mar. 2017.

[172] A. Cohen, G. Stemmer, S. Ingalsuo, and S. Markovich-Golan, "Combined weighted prediction error and minimum variance distortionless response for dereverberation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, New Orleans, LA, USA, Mar. 2017, pp. 611–615.

[173] T. Otsuka, K. Ishiguro, T. Yoshioka, H. Sawada, and H. G. Okuno, "Multichannel sound source dereverberation and separation for arbitrary number of sources based on bayesian nonparametrics," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2218–2232, Dec. 2014.

[174] H. Buchner, R. Aichner, and W. Kellermann, "Blind source separation for convolutive mixtures exploiting nongaussianity, nonwhiteness, and nonstationarity," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Kyoto, Japan, Sep. 2003.

[175] H. Buchner and W. Kellermann, "TRINICON for dereverberation of speech and audio signals," in *Speech Dereverberation*, P. Naylor and N. D. Gaubitch, Eds. Springer, 2010, ch. 10, pp. 311–386.

[176] A. Jukić and S. Doclo, "Speech dereverberation using weighted prediction error with Laplacian model of the desired signal," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 5172–5176.

[177] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Speech dereverberation with multi-channel linear prediction and sparse priors for the desired signal," in *Proceedings of the Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Nancy, France, May 2014, pp. 23–26.

[178] ——, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1509–1520, Sep. 2015.

[179] ——, "Group sparsity for MIMO speech dereverberation," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2015.

[180] A. Jukić, Z. Wang, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Constrained multi-channel linear prediction for adaptive speech dereverberation," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Xi'an, China, Sep. 2016.

[181] A. Jukić, T. van Waterschoot, and S. Doclo, "Adaptive speech dereverberation using constrained sparse multi-channel linear prediction," *IEEE Signal Processing Letters*, vol. 24, no. 1, pp. 101–105, Jan. 2017.

[182] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "A general framework for multi-channel speech dereverberation exploiting sparsity," in *Proceedings of the AES 60th International Conference on DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)*, Leuven, Belgium, Feb. 2016.

[183] A. Jukić, T. van Waterschoot, and S. Doclo, "A general framework for incorporating time-frequency domain sparsity in multi-channel speech dereverberation," *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, pp. 17–30, Jan./Feb. 2017.

[184] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[185] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, May 2009.

[186] ——, "Convolutive transfer function generalized sidelobe canceler," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1420–1434, Sep. 2009.

[187] M. Elad, *Sparse and Redundant Representations.* Springer, 2010.

[188] P. Bofil and M. Zibulevsky, "Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform," in *Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, Helsinki, Finland, Jun. 2000, pp. 87–92.

[189] S. Makino, S. Araki, S. Winter, 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. John Wiley & Sons, 2010.

[190] I. Tashev and A. Acero, "Statistical modeling of the speech signal," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, Sep. 2010.

[191] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: From coding to source separation," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 995–1005, Jun. 2010.

[192] P. Balazs, M. Dörfler, M. Kowalski, and B. Torresani, "Adapted and adaptive linear time-frequency representations: A synthesis point of view," *IEEE Signal Processing Magazine*, vol. 30, no. 6, pp. 20–31, Nov. 2013.

[193] P. Bofil and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, no. 11, pp. 2353–2362, 2001.

[194] P. Georgiev, F. Theis, and A. Cichocki, "Sparse component analysis and blind source separation for underdetermined mixtures," *IEEE Transactions on Neural Networks*, vol. 16, no. 4, pp. 992–996, Jul. 2005.

[195] 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, 2010.

[196] S. Arberet, P. Vandergheynst, R. E. Carrillo, J.-P. 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, July 2013.

[197] C. Févotte, B. Torrésani, L. Daudet, and S. J. Godsill, "Sparse linear regression with structured priors and application to denoising of musical audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 174–185, Jan. 2008.

[198] K. Siedenburg and M. Dörfler, "Audio denoising by generalized time-frequency thresholding," in *Proceedings of the 45th AES Conference on Applications of Time-Frequency Processing in Audio*, Mar. 2012.

[199] K. Siedenburg and P. Depalle, "Modulation filtering for structured time-frequency estimation of audio signals," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2013.

[200] K. Kumatani, J. McDonough, B. Rauch, D. Klakow, P. N. Garner, and W. Li, "Beamforming with a maximum negentropy criterion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 994–1008, Jul. 2009.

[201] N. Epain, T. Noohi, and C. T. Jin, "Sparse recovery method for dereverberation," in *Proceedings of the REVERB Workshop*, Florence, Italy, May 2014.

[202] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, "Audio inpainting," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 922–932, Mar. 2012.

[203] B. Defraene, N. Mansour, S. D. Hertogh, T. van Waterschoot, M. Diehl, and M. Moonen, "Declipping of audio signals using perceptual compressed sensing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2627–2637, Dec. 2013.

[204] K. Siedenburg, M. Dörfler, and M. Kowalski, "Audio declipping with social sparsity," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 1577–1581.

[205] S. Kitić, N. Bertin, and R. Gribonval, "Sparsity and cosparsity for audio declipping: A flexible non-convex approach," in *Proceedings of the Latent Variable Analysis and Signal Separation (LVA/ICA)*, Liberec, Czech Republic, Aug. 2015.

[206] D. Giacobello, "Sparsity in linear predictive coding of speech," Ph.D. dissertation, Aalborg University, Aalborg, Denmark, Aug. 2010.

[207] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1644–1657, 2012.

[208] P. A. Naylor, N. D. Gaubitch, and E. A. P. Habets, "Signal-based performance evaluation of dereverberation," *Journal of Electrical and Computer Engineering*, pp. 1–5, 2013.

[209] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality.* Prentice-Hall, 1988.

[210] 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,” in *International Telecommunication Union (ITU-T) Recommendation P.862*, 2001.

[211] ——, “Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs,” in *International Telecommunication Union (ITU-T) Recommendation P.862.2*, 2001.

[212] T. H. Falk, C. Zheng, and W.-Y. Chan, “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.

[213] J. F. Santos, M. Senoussaoui, and T. H. Falk, “An improved non-intrusive intelligibility metric for noisy and reverberant speech,” in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Antibes - Juan Les Pins, France, Sep. 2014.

[214] B. Cauchi, J. F. Santos, K. Siedenburg, T. H. Falk, P. A. Naylor, S. Doclo, and S. Goetze, “Predicting the quality of processed speech by combining modulation-based features and model trees,” in *Proceedings of the ITG Conference on Speech Communication*, Paderborn, Germany, Oct. 2016.

[215] A. Avila, B. Cauchi, S. Goetze, S. Doclo, and T. H. Falk, “Performance comparison of intrusive and non-intrusive instrumental quality measures for enhanced speech,” in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Xi’an, China, Sep. 2016.

[216] J. Tribolet, P. Noll, B. McDermott, and R. E. Crochiere, “A study of complexity and quality of speech waveform coders,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1978, pp. 586–590.

[217] S. Goetze, E. Albertin, J. Rennies, E. A. P. Habets, and K.-D. Kammeyer, “Speech quality assessment for listening-room compensation,” in *Proceedings of the AES 38th International Conference on Sound Quality Evaluation*, Pitea, Sweden, Jun. 2010, pp. 11–20.

[218] 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 *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Antibes - Juan Les Pins, France, Sep. 2014.

[219] 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, 2008.

[220] J. A. Palmer, K. Kreutz-Delgado, D. P. Wipf, and B. D. Rao, “Variational EM algorithms for non-Gaussian latent variable models,” in *Advances in Neural Information Processing Systems 18.* MIT Press, 2006, pp. 1059–1066.

[221] S. D. Babacan, R. Molina, M. N. Do, and A. K. Katsaggelos, "Bayesian blind deconvolution with general sparse image priors," in *Proceedings of the European Conference on Computer Vision (ECCCV)*, Florence, Italy, Oct. 2012, pp. 341–355.

[222] D. Wipf and H. Zhang, "Analysis of Bayesian blind deconvolution," in *Proceedings of the International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*, Lund, Sweden, Aug. 2013, pp. 40–53.

[223] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, NV, USA, Mar. 2008, pp. 3869–3872.

[224] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, Sep. 2010.

[225] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, Aug. 2005.

[226] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, pp. 1110–1126, 2005.

[227] M. Novey, T. Adali, and A. Roy, "A Complex Generalized Gaussian Distribution - Characterization, Generation, and Estimation," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1427–1433, 2010.

[228] D. Wipf and S. Nagarajan, "Iterative reweighted l1 and l2 methods for finding sparse solutions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 317–329, 2010.

[229] E. Candes, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 877–905, 2008.

[230] S. Nam, M. E. Davies, M. Elad, and R. Gribonval, "The cosparse analysis model and algorithms," *Applied and Computational Harmonic Analysis*, vol. 34, no. 1, pp. 30–56, 2013.

[231] R. Chartrand, E. Y. Sidky, and X. Pan, "Nonconvex compressive sensing for X-ray CT: an algorithm comparison," in *Proceedings of the Asilomar Conference on Signals, Systems, and Computers (ASILOMAR)*, Pacific Grove, California, USA, Nov. 2013, pp. 665–669.

[232] R. Giryes, S. Nam, M. Elad, R. Gribonval, and M. E. Davies, "Greedy-like algorithms for the cosparse analysis model," *Linear Algebra and its Applications*, pp. 22–60, Jan. 2014.

[233] S. D. Babacan, S. Nakajima, and M. N. Do, "Bayesian group-sparse modeling and variational inference," *IEEE Transactions on Signal Processing*, vol. 62, no. 1, pp. 2906–2921, Jun. 2014.

[234] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallet, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus LDC93S1," in *Philadelphia: Linguistic Data Consortium*, 1993.

[235] Z. Průša, P. L. Søndergaard, N. Holighaus, C. Wiesmeyr, and P. Balazs, "The large time-frequency analysis toolbox 2.0," in *Sound, Music, and Motion. CMMR 2013. Lecture Notes in Computer Science*, M. Aramaki, O. Derrien, R. Kronland-Martinet, and S. Ystad, Eds. Springer, 2014, vol. 8905, pp. 419–442.

[236] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B*, vol. 68, no. 1, pp. 49–67, 2006.

[237] M. Kowalski and B. Torrésani, "Structured sparsity: from mixed norms to structured shrinkage," in *Proceedings of the Signal Processing with Adaptive Sparse Structured Representations Workshop (SPARS)*, Saint-Malo, France, Apr. 2009.

[238] J. Huang and T. Zhang, "The benefit of group sparsity," *The Annals of Statistics*, vol. 38, no. 4, pp. 1978–2004, 2010.

[239] A. Asaei, M. Golbabaee, H. Bourlard, and V. Cevher, "Structured sparsity models for reverberant speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 620–633, Mar. 2014.

[240] A. Benedek and R. Panzone, "The space L^p with mixed norm," *Duke Mathematical Journal*, vol. 28, no. 3, pp. 301–324, 1961.

[241] P. J. Garrigues and B. A. Olshausen, "Group sparse coding with a Laplacian scale mixture prior," in *Advances in Neural Information Processing Systems 23*, 2010.

[242] M. Kowalski, K. Siedenburg, and M. Dörfler, "Social Sparsity! Neighborhood Systems Enrich Structured Shrinkage Operators," *IEEE Transactions on Signal Processing*, vol. 61, no. 10, pp. 2498–2511, May 2013.

[243] M. Fornasier and H. Rahut, "Recovery algorithm for vector-valued data with joint sparsity constraints," *SIAM Journal on Numerical Analysis*, vol. 46, no. 2, pp. 577–613, 2008.

[244] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 707–710, Oct. 2007.

[245] Z. Zhao and B. D. Rao, "Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 912–926, Sep. 2011.

[246] S. 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, 2011.

[247] S. Haykin, *Adaptive Filter Theory*, 3rd ed. Prentice Hall, 2013.

[248] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Machine Learning*, vol. 1, no. 3, pp. 127–239, 2014.

[249] H. Buchner, J. Benesty, and W. Kellermann, “Generalized multichannel frequency-domain adaptive filtering: efficient realization and application to hands-free speech communication,” *Signal Processing*, vol. 85, no. 3, pp. 549–570, 2005.

[250] 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, pp. 636–656, 2007.

[251] T. Necciari, P. Balazs, N. Holighaus, and P. L. Søndergaard, “The ERBlet transform: an auditory-based time-frequency representation with perfect reconstruction,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 2013, pp. 7447–7451.

[252] L. Ø. Endelt and A. la Cour-Harbo, “Comparison of methods for sparse representations of musical signals,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, USA, Mar. 2005.

[253] J. Kovacevic and A. Chebira, “Life beyond bases: The advent of frames (part i),” *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 86–104, 2007.

[254] M. Elad, P. Milanfar, and R. Rubinstein, “Analysis versus synthesis in signal priors,” *Inverse problems*, vol. 23, no. 3, p. 947, 2007.

[255] T. L. Jensen, D. Giacobello, T. van Waterschoot, and M. G. Christensen, “Fast algorithms for high-order sparse linear prediction with applications to speech processing,” *Speech Communication*, vol. 76, pp. 143–156, Feb. 2016.

[256] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1996.

[257] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[258] N. Mohammadiha, P. Smaragdis, and A. Leijon, “Supervised and unsupervised speech enhancement using nonnegative matrix factorization,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[259] A. Jukić, N. Mohammadiha, T. van Waterschoot, T. Gerkmann, and S. Doclo, “Multi-channel linear prediction-based speech dereverberation with low-rank power spectrogram approximation,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Brisbane, Australia, Apr. 2015, pp. 96–100.

[260] C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis,” *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[261] A. Cichocki, R. Zdunek, and S. Amari, “Csiszar’s divergences for non-negative matrix factorization: Family of new algorithms,” in *Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, Berlin, Germany, 2006, pp. 32–39.

[262] A. Y. Aravkin, J. V. Burke, and G. Pillonetto, *Optimization Viewpoint on Kalman Smoothing with Applications to Robust and Sparse Estimation.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 237–280.

[263] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge University Press, 2004.

[264] R. C. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 223–233, Jan. 2012.

[265] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, I. Nobutaka, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, and A. Nakamura, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge," in *Proceedings of the REVERB Workshop*, Florence, Italy, May 2014.

[266] P. Balazs, B. Laback, G. Eckel, and W. A. Deutsch, "Time–frequency sparsity by removing perceptually irrelevant components using a simple model of simultaneous masking," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 34–49, Jan. 2010.

[267] C. Févotte and M. Kowalski, "Hybrid sparse and low-rank time-frequency signal decomposition," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Nice, France, Aug. 2015, pp. 464–468.

[268] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, May 2011.

[269] J. Jain, "An efficient algorithm for a large Toeplitz set of linear equations," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 6, pp. 612–615, Dec. 1979.

[270] D. Marquardt, "Development and evaluation of psychoacoustically motivated binaural noise reduction and cue preservation techniques," Ph.D. dissertation, University of Oldenburg, Oldenburg, Germany, Nov. 2015.

[271] J. A. Palmer, "Variational and scale mixture representations of non-gaussian densities for estimation in the bayesian linear model: Sparse coding, independent component analysis, and minimum entropy segmentation," Ph.D. dissertation, University of California San Diego, San Diego, CA, USA, 2006.

[272] Y. Rakvongthai, A. P. Vo, and S. Oraintara, "Complex Gaussian scale mixtures of complex wavelet coefficients," *IEEE Transactions on Signal Processing*, vol. 58, no. 7, pp. 3545–3556, Jul. 2010.

[273] R. T. Rockafellar, *Convex Analysis.* Princeton, 1970.

[274] P. Ochs, A. Dosovitskiy, T. Brox, and T. Pock, "On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision," *SIAM Journal on Imaging Sciences*, vol. 8, no. 1, pp. 331–3372, 2015.

[275] R. Chartrand, "Shrinkage mappings and their induced penalty functions," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 1026–1029.

[276] Y. Lucet, H. H. Bauschke, and M. Trienis, “The piecewise linear-quadratic model for computational convex analysis,” *Computational Optimization and Applications*, vol. 43, no. 1, pp. 95–118, May 2009.

[277] M. Fortin and R. Glowinski, *Augmented Lagrangian Methods.* Elsevier, 1983.

[278] P. L. Combettes and J.-C. Pesquet, *Proximal Splitting Methods in Signal Processing.* New York, NY: Springer New York, 2011, pp. 185–212.

[279] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gonirevsky, “An interior-point method for large-scale ℓ_1-regularized least squares,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, Dec. 2007.

LIST OF PUBLICATIONS

The following publications are related to the work in this thesis.

Peer-reviewed Journal Papers

[J4] A. Jukić, T. van Waterschoot, T. Gerkmann, S. Doclo, "A general framework for incorporating time-frequency domain sparsity in multi-channel speech dereverberation," *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, pp. 17–30, Jan./Feb. 2017.

[J3] A. Jukić, T. van Waterschoot, S. Doclo, "Adaptive speech dereverberation using constrained sparse multi-channel linear prediction," *IEEE Signal Processing Letters*, vol. 24, no. 1, pp. 101–105, Jan. 2017.

[J2] A. Jukić, T. van Waterschoot, T. Gerkmann, S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 9, pp. 1509–1520, Sept. 2015.

[J1] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, 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, pp. 1–12, July. 2015.

Peer-reviewed Conference Papers

[C10] F. Xiong, B. T. Meyer, B. Cauchi, A. Jukić, S. Doclo, S. Goetze, "Performance comparison of real-time single-channel speech dereverberation algorithms," in *Proceedings of the Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, San Francisco, CA, USA, Mar. 2017.

[C9] A. Jukić, Z. Wang, T. van Waterschoot, T. Gerkmann, S. Doclo, "Constrained multi-channel linear prediction for adaptive speech dereverberation," in *Proceedings of the International Workshop on Acoustic Signal Enhancement (IWAENC)*, Xi'an, China, Sept. 2016.

[C8] I. Kodrasi, A. Jukić, and S. Doclo, "Robust sparsity-promoting acoustic multi-channel equalization for speech dereverberation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Shanghai, China, Mar. 2016, pp. 166–170.

[C7] A. Jukić, T. van Waterschoot, T. Gerkmann, S. Doclo, "A general framework for multi-channel speech dereverberation by exploiting sparsity," in *Proceed-*

ings of the AES 60th Conference on DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech), Leuven, Belgium, Feb. 2016.

[C6] A. Jukić, T. van Waterschoot, T. Gerkmann, S. Doclo, "Group sparsity for MIMO speech dereverberation," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, Oct. 2015.

[C5] A. Jukić, N. Mohammadiha, T. van Waterschoot, T. Gerkmann, S. Doclo, "Multi-channel linear prediction-based speech dereverberation with low-rank power spectrogram approximation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, April 2015, pp. 96–100.

[C4] A. Jukić, T. van Waterschoot, T. Gerkmann, S. Doclo, "Speech dereverberation with convolutive transfer function approximation using MAP and variational deconvolution approaches," in *Proceedings of the International Workshop on Acoustic Signal Enhancement (IWAENC)*, Antibes - Juan Les Pins, France, Sept. 2014.

[C3] A. Jukić, T. van Waterschoot, T. Gerkmann, S. Doclo, "Speech dereverberation with multi-channel linear prediction and sparse priors for the desired signal," in *Proceedings of the Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Nancy, France, May 2014, pp. 23–26.

[C2] 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 *Proceedings of the REVERB Challenge Workshop (REVERB'14)*, Florence, Italy, May 2014.

[C1] A. Jukić, S. Doclo, "Speech dereverberation using weighted prediction error with Laplacian model of the desired signal," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 5172–5176.