SINGLE-MICROPHONE MULTI-FRAME SPEECH ENHANCEMENT EXPLOITING SPEECH INTERFRAME CORRELATION

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

von
Frau Dörte Fischer
geboren am 31. Juli 1990
in Ribnitz-Damgarten (Deutschland)

Dörte Fischer: *Single-Microphone Multi-Frame Speech Enhancement Exploiting Speech Interframe Correlation*

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

WEITERE GUTACHTER:
Prof. Dr.-Ing. Timo Gerkmann, *University of Hamburg, Germany*
Prof. Dr. Mads Græsbøll Christensen, *Aalborg University, Denmark*

TAG DER DISPUTATION:
30. November 2020

ACKNOWLEDGMENTS

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

First of all, I would like to express my gratitude to my supervisor Simon Doclo for giving me the opportunity to write this thesis and the freedom to pursue my interests. I am deeply grateful for his guidance, support and advice through the years. Especially during the writing process, his scientific advices cannot be appreciated enough. I would also like to thank Timo Gerkmann who supervised me in my first years, for the interesting discussions and his influential advice, as well as for reviewing my thesis and participating in my thesis committee. I am thankful to Mads Christensen for reviewing my thesis, showing much interest in my work and participating in the thesis committee. Furthermore, I would like to thank Steven van de Par as a member of the examination committee.

I am grateful to Jörg Bitzer, Uwe Simmer, Danilo Hollosi and Jan Wellmann whose early guidance in my bachelor's and master's studies led to my increased interest in research. I would also like to thank Tino Just for his support in the recent years.

I want to thank all the members of the Department of Medical Physics and Acoustics and the Cluster of Excellence Hearing4all. A special thanks goes to all current and former members of the Signal Processing Group for providing an open and friendly work environment, for the many scientific and non-scientific discussions and for many enjoyable conferences. Particularly, I would like to thank my office mate Nico Gößling, as well as Henning Schepker, Marvin Tammen and Klaus Brümann for their help and even more for all the fun and the interesting conversations. I would also like to thank Christopher Hauth, with whom I enjoyed the daily lunch breaks, for all the laughs and easy chats, especially when things were getting tough.

Most importantly, I would like to thank my family and friends for their continuous support and encouragement throughout these years.

Güstrow, January 2021
Dörte Fischer

ABSTRACT

Speech communication devices such as hearing aids or mobile phones are often used in acoustically challenging situations, where the desired speech signal is affected by undesired background noise. Since, in these situations speech quality and speech intelligibility may be degraded, speech enhancement algorithms are required to suppress the undesired background noise, while preserving the desired speech signal. In this thesis, we focus on single-microphone speech enhancement algorithms in the short-time Fourier transform domain, more in particular on multi-frame algorithms that aim at exploiting speech correlation across time-frames. In principle, exploiting the speech interframe correlation enables to suppress the undesired background noise, while keeping speech distortion low. Existing single-microphone multi-frame speech enhancement algorithms, such as the multi-frame minimum variance distortionless response (MFMVDR) filter and the multi-frame minimum power distortionless response (MFMPDR) filter, depend on the normalized speech correlation vector, which is highly time-varying and hence difficult to be accurately estimated.

The main objective of this thesis is to develop and evaluate novel robust methods to estimate the normalized speech correlation vector from the noisy microphone signal, either based on robust beamforming approaches or exploiting a low-rank speech model.

First, in order to better understand the performance of the MFMVDR and MFMPDR filters, we investigate the sensitivity of both filters to estimation errors in the normalized speech correlation vector. We compare the practically feasible MFMPDR filter with two oracle versions of the MFMVDR filter for different oracle and blind estimates of the normalized speech correlation vector. Simulation results show that accurately estimating the normalized speech correlation vector is crucial, since even small estimation errors degrade the performance of the MFMVDR and MFMPDR filters, resulting in speech distortion and unpleasant artifacts in the background noise.

Second, in order to improve the robustness of the practical feasible MFMPDR filter against estimation errors in the normalized speech correlation vector, we investigate the potential of using concepts from robust MPDR beamforming in the context of single-microphone multi-frame speech enhancement. We propose two constrained MFMPDR filters that estimate the normalized speech correlation vector as the vector maximizing the total signal output power spectral density within a spherical uncertainty set. This corresponds to imposing a quadratic inequality constraint on the mismatch vector with respect to the presumed normalized speech correlation vector, e.g., the state-of-the-art maximum-likelihood (ML) estimate. Whereas the singly-constrained (SC) MFMPDR filter only considers the quadratic inequality constraint to estimate the (non-normalized) speech correlation vector, the doubly-

constrained (DC) MFMPDR filter integrates a linear normalization constraint into the optimization problem to directly estimate the normalized speech correlation vector. The main novelty is to set the upper bound of the spherical uncertainty set using a trained non-linear mapping function that depends on the time-varying a-priori SNR estimate for each time-frequency point. Simulation results show that the proposed constrained approaches yield a more accurate estimate of the normalized speech correlation vector than the ML estimate. An instrumental and a perceptual evaluation show that both constrained MFMPDR filters lead to a more natural speech quality and less noise distortion, but a more conservative noise reduction performance than the state-of-the-art ML-MFMPDR filter, where the DC-MFMPDR filter is preferred in terms of overall quality compared to the SC-MFMPDR filter and the ML-MFMPDR filter.

Third, assuming that speech signals can be modeled using a low-rank model, we propose two matrix-based methods to estimate the normalized speech correlation vector, namely the matrix-subtraction (MS) method and the subspace-decomposition (SD) method. Both methods are based on the eigenvalue decomposition of a matrix, which is either constructed by subtracting the estimated normalized noise correlation matrix from the estimated normalized noisy speech correlation matrix or by prewhitening the estimated normalized noisy speech correlation matrix with the estimated normalized noise correlation matrix. We propose to estimate the speech model order for each time-frequency point by incorporating the a-priori SNR into the minimum description length selection criterion. Simulation results show that the proposed matrix-based SD method yields a more accurate estimate of the normalized speech correlation vector than the vector-based ML estimate. Instrumental performance measures indicate that the MFMPDR filter using the proposed SD estimator leads to a better speech quality and more noise reduction than the ML-MFMPDR filter, while keeping speech distortion low.

Finally, the results of a subjective listening test confirm that the overall quality for the MFMPDR filters using the proposed SD estimator and the proposed DC estimator are significantly better than for the state-of-the-art ML-MFMPDR filter.

ZUSAMMENFASSUNG

Sprachkommunikationsgeräte wie Hörgeräte oder Mobiltelefone werden häufig in akustisch schwierigen Situationen verwendet, in denen das gewünschte Sprachsignal durch ein unerwünschtes Hintergrundgeräusch beeinflusst wird. Da in diesen Situationen die Sprachqualität und das Sprachverstehen beeinträchtigt werden können, sind Sprachverbesserungsalgorithmen erforderlich, die das unerwünschte Hintergrundgeräusch unterdrücken und gleichzeitig das erwünschte Sprachsignal nicht verändern. In dieser Arbeit werden Sprachverbesserungsalgorithmen mit einem Mikrofon im Bereich der Kurzzeit-Fourier-Transformation betrachtet. Es werden insbesondere mehrrahmige Algorithmen untersucht und entwickelt, die Sprachkorrelationen über Zeitrahmen hinweg auszunutzen. Im Prinzip ermöglicht die Ausnutzung der sprachlichen Zwischenrahmenkorrelation die Unterdrückung unerwünschter Hintergrundgeräusche bei gleichzeitig geringen Sprachverzerrungen. Bestehende mehrrahmige Sprachverbesserungsalgorithmen mit einem Mikrofon, wie z.B. das MFMVDR (engl. multi-frame minimum variance distortionless response) Filter und das MFMPDR (engl. multi-frame minimum power distortionless response) Filter, hängen von dem normierten Sprachkorrelationsvektor ab, welcher stark zeitvariant ist und daher nur schwer exakt geschätzt werden kann.

Das Hauptziel dieser Arbeit ist die Entwicklung und Bewertung neuer robuster Methoden zur Schätzung des normierten Sprachkorrelationsvektors aus dem verrauschten Mikrofonsignal entweder auf der Grundlage robuster Strahlenformer-Ansätze oder unter Verwendung eines niedrig-rangigen Sprachmodells.

Erstens, um die Leistung der MFMVDR und MFMPDR Filter besser zu verstehen, untersuchen wir die Empfindlichkeit beider Filter gegenüber Schätzfehlern im normierten Sprachkorrelationsvektor. Wir vergleichen das praktisch realisierbare MFMPDR Filter mit zwei orakel Versionen des MFMVDR Filters für verschiedene orakel und blinde Schätzungen des normierten Sprachkorrelationsvektors. Die Simulationsergebnisse zeigen, dass die genaue Schätzung des normierten Sprachkorrelationsvektors von entscheidender Bedeutung ist, da bereits kleine Schätzfehler die Leistung der MFMVDR und MFMPDR Filter verschlechtern, was zu Sprachverzerrungen und unangenehmen Artefakten im Hintergrundrauschen führt.

Zweites, um die Robustheit des praktisch realisierbaren MFMPDR Filters gegenüber Schätzfehlern im normierten Sprachkorrelationsvektor zu verbessern, untersuchen wir das Potenzial von Konzepten aus robusten MPDR Strahlenformern im Kontext der mehrrahmigen Sprachverbesserung mit einem Mikrofon. Wir schlagen zwei MFMPDR Filter mit Nebenbedingungen vor, die den normierten Sprachkorrelationsvektor als den Vektor schätzen, der die spektrale Dichte der gesamten Signalausgangsleistung innerhalb eines sphärischen Unsicherheitsbereichs maximiert. Dies entspricht dem Auferlegen einer quadratischen Ungleichheitsbedingung auf den

Fehlervektor in Bezug auf den angenommenen normierten Sprachkorrelationsvektor, z.B. die State-of-the-Art maximum-likelihood (ML) Schätzung. Während das MFMPDR Filter mit einer Nebenbedingung (engl. singly-constrained, SC) nur die quadratische Ungleichheitsbedingung berücksichtigt, um den (nicht-normierten) Sprachkorrelationsvektor zu schätzen, integriert das MFMPDR Filter mit zwei Nebenbedingungen (engl. doubly-constrained, DC) zusätzlich eine lineare Normalisierungsbedingung in das Optimierungsproblem, um den normierten Sprachkorrelationsvektor direkt zu schätzen. Die Hauptneuheit besteht darin, die obere Grenze des sphärischen Unsicherheitsbereichs mit Hilfe einer trainierten nichtlinearen Abbildungsfunktion zu bestimmen, die von der zeitvarianten a-priori SNR Schätzung für jeden Zeit-Frequenz-Punkt abhängt. Die Simulationsergebnisse zeigen, dass die vorgeschlagenen Ansätze mit Nebenbedingung(en) eine genauere Schätzung des normierten Sprachkorrelationsvektors liefern als die ML-Schätzung. Eine instrumentelle und eine wahrnehmungsbezogene Evaluierung zeigen, dass beide MFMPDR Filter mit Nebenbedingung(en) zu einer natürlicheren Sprachqualität und weniger Verzerrungen im Hintergrundgeräusch führen, aber zu einer konservativeren Störgeräuschunterdrückungsleistung als das State-of-the-Art ML-MFMPDR Filter. Das DC-MFMPDR Filter wird in Bezug auf die Gesamtqualität gegenüber des SC-MFMPDR Filters und des ML-MFMPDR Filters bevorzugt.

Drittens, auf der Grundlage das Sprachsignale mittels eines niedrig-rangigen Sprachmodells beschrieben werden können, schlagen wir zwei matrixbasierte Methoden zur Schätzung des normierten Sprachkorrelationsvektors vor. Zum einen die Matrix-Subtraktionsmethode (MS) und zum anderen die Subraum-Zerlegungsmethode (engl. subspace-decomposition, SD). Beide Methoden basieren auf der Eigenwertzerlegung einer Matrix, welche entweder durch Subtraktion der geschätzten normierten Rauschkorrelationsmatrix von der geschätzten normierten rauschbehafteten Sprachkorrelationsmatrix oder durch Weißung der geschätzten normierten rauschbehafteten Sprachkorrelationsmatrix mit der geschätzten normierten Rauschkorrelationsmatrix konstruiert wird. Für die Schätzung der Ordnung des Sprachmodells schlagen wir vor, dass der a-priori SNR in das Kriterium der minimalen Beschreibungslänge integriert wird. Die Simulationsergebnisse zeigen, dass die vorgeschlagene matrixbasierte SD-Methode eine genauere Schätzung des normierten Sprachkorrelationsvektors liefert als die vektorbasierte ML-Schätzung. Instrumentelle Leistungsmessungen zeigen, dass das MFMPDR Filter unter Verwendung des vorgeschlagenen SD-Schätzers zu einer besseren Sprachqualität und einer stärkeren Rauschunterdrückung führt als das ML-MFMPDR Filter, wobei die Sprachverzerrung konstant niedrig bleibt.

Abschließend bestätigen die Ergebnisse eines subjektiven Hörtests, dass die Gesamtqualität für das MFMPDR Filter unter Verwendung des vorgeschlagenen SD-Schätzers und des vorgeschlagenen DC-Schätzers signifikant besser ist als für das State-of-the-Art ML-MFMPDR Filter.

GLOSSARY

Acronyms and Abbreviations

AIC	Akaike information criterion
ANOVA	analysis of variance
DC	doubly-constrained
DDA	decision-directed approach
DFT	discrete Fourier transform
DNN	deep neural network
EVD	eigenvalue decomposition
FIR	finite-impulse response
iDFT	inverse discrete Fourier transform
iSTFT	inverse short-time Fourier transform
MAP	maximum a-posteriori
MDL	minimum description length
MFMPDR	multi-frame minimum power distortionless response
MFMVDR	multi-frame minimum variance distortionless response
MFWF	multi-frame Wiener filter
ML	maximum-likelihood
MMSE	minimum mean-square error
MOS	mean opinion score
MPDR	minimum power distortionless response
MS	matrix-subtraction
MSE	mean-square error
MUSHRA	multi stimulus test with hidden reference and anchor
MVDR	minimum variance distortionless response
MWF	multi-channel Wiener filter
PDF	probability density function
PESQ	perceptual evaluation of speech quality
PSD	power spectral density
RTF	relative transfer function
SC	singly-constrained

SD	subspace-decomposition
SDW-MFWF	speech-distortion-weighted multi-frame Wiener filter
SDW-MWF	speech-distortion-weighted multi-channel Wiener filter
segNR	segmental noise reduction
segSNR	segmental signal-to-noise ratio
segSSNR	segmental speech signal-to-noise ratio
SNR	signal-to-noise ratio
SPP	speech presence probability
STFT	short-time Fourier transform
STSA	short-time spectral amplitude
VAD	voice activity detector
WG	Wiener gain

Mathematical Notation

a	scalar a
$\boldsymbol{a}$	vector $\boldsymbol{a}$
$\boldsymbol{A}$	matrix $\boldsymbol{A}$
a^*	complex conjugate of a
$\boldsymbol{a}^T, \boldsymbol{A}^T$	transpose of vector $\boldsymbol{a}$, matrix $\boldsymbol{A}$
$\boldsymbol{a}^H, \boldsymbol{A}^H$	Hermitian transpose of vector $\boldsymbol{a}$, matrix $\boldsymbol{A}$
$\boldsymbol{a}^w, \boldsymbol{A}^w$	prewhitened vector $\boldsymbol{a}$, matrix $\boldsymbol{A}$
$\boldsymbol{A}^{-1}$	inverse of matrix $\boldsymbol{A}$
$\mathrm{tr}\,[\boldsymbol{A}]$	trace of matrix $\boldsymbol{A}$ (sum of diagonal elements)
$\det[\boldsymbol{A}]$	determinant of matrix $\boldsymbol{A}$
$\widehat{a}$, $\widehat{\boldsymbol{a}}$, $\widehat{\boldsymbol{A}}$	estimate of scalar a, vector $\boldsymbol{a}$, matrix $\boldsymbol{A}$
$a(t)$	discrete-time sequence at discrete time index t
$A(k,m)$	short-time Fourier transform of $a(t)$ at frequency-bin index k and time-frame index m
$\phi_A(k,m)$	power spectral density of $A(k,m)$
$\boldsymbol{R}_{\boldsymbol{a}}$	correlation matrix of vector $\boldsymbol{a}$
$\boldsymbol{\Gamma}_{\boldsymbol{a}}$	normalized correlation matrix of vector $\boldsymbol{a}$
$\mathcal{E}[\cdot]$	expectation operator
$\lvert\,\cdot\,\rvert$	absolute value
$\lVert\cdot\rVert_2$	l_2-norm

$f(a)$	probability density function of a
$f(a\|b)$	conditional probability density function of a given b

Fixed Symbols

t	discrete time index
m	time-frame index
k	frequency-bin index
l	time-lag / filter index
M	number of time-frames m
K	number of frequency-bins k / size of the DFT
L	number of consecutive time-frames l / filter length
N	number of independent and identically distributed zero-mean multivariate Gaussian observations
T	time-frame length
S	time-segment length
R	frame shift
Q	speech model order / rank of speech correlation matrix
$\mathbb{P}_Y$	set of time-frequency points containing speech-and-noise
$\mathbb{P}_N$	set of time-frequency points containing noise-only
$\mathbb{T}_Y$	set of time-segments containing speech-and-noise
$\mathbb{U}$	spherical uncertainty set
$\mathbb{Q}^{\mathrm{MS,pos}}$	set of positive prewhitened speech eigenvalues
$\mathbb{Q}^{\mathrm{SD,pos}}$	set of positive speech eigenvalues
$\mathbb{Q}^{\mathrm{Thre}}$	set of eigenvalues larger than a threshold ϑ
$\mathbb{Q}^{\mathrm{O}}$	set of speech eigenvalues covering 98 % of the total energy of speech eigenvalues
$\mathbb{Y}$	set of N observations
$w_a(t)$	analysis window
$w_s(t)$	synthesis window
$y(t)$	noisy speech signal
$x(t)$	speech signal
$n(t)$	noise signal
$\tilde{x}(t)$	processed speech signal
$\tilde{n}(t)$	processed noise signal

$\phi_Y(k,m)$	noisy speech PSD
$\phi_X(k,m)$	speech PSD
$\phi_N(k,m)$	noise PSD
$\phi_{\boldsymbol{y}}^{\text{out}}(k,m)$	signal output PSD
$\phi_{\boldsymbol{n}}^{\text{out}}(k,m)$	noise output PSD
$\phi_{\boldsymbol{u}}^{\text{out}}(k,m)$	undesired output PSD
$\xi(k,m)$	a-priori SNR
$Y(k,m)$	noisy speech STFT coefficient
$X(k,m)$	speech STFT coefficient
$N(k,m)$	noise STFT coefficient
$\tilde{N}(k,m)$	processed noise STFT coefficient
$\boldsymbol{y}(k,m)$	noisy speech vector
$\boldsymbol{x}(k,m)$	speech vector
$\boldsymbol{s}(k,m)$	correlated speech component
$\boldsymbol{x'}(k,m)$	uncorrelated speech component
$\boldsymbol{n}(k,m)$	noise vector
$\boldsymbol{u}(k,m)$	undesired signal vector
$\boldsymbol{R}_{\boldsymbol{y}}(k,m)$	noisy speech correlation matrix
$\boldsymbol{R}_{\boldsymbol{x}}(k,m)$	speech correlation matrix
$\boldsymbol{R}_{\boldsymbol{s}}(k,m)$	correlation matrix of the correlated speech component
$\boldsymbol{R}_{\boldsymbol{x'}}(k,m)$	correlation matrix of the uncorrelated speech component
$\boldsymbol{R}_{\boldsymbol{n}}(k,m)$	noise correlation matrix
$\boldsymbol{R}_{\boldsymbol{u}}(k,m)$	undesired correlation matrix
$\boldsymbol{\gamma}_{\boldsymbol{y}}(k,m)$	normalized noisy speech correlation vector
$\boldsymbol{\gamma}_{\boldsymbol{x}}(k,m)$	normalized speech correlation vector
$\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}(k,m)$	presumed normalized speech correlation vector
$\breve{\boldsymbol{\gamma}}_{\boldsymbol{x}}(k,m)$	(non-normalized) speech correlation vector
$\boldsymbol{\gamma}_{\boldsymbol{n}}(k,m)$	normalized noise correlation vector
$\boldsymbol{\Gamma}_{\boldsymbol{y}}(k,m)$	normalized noisy speech correlation matrix
$\boldsymbol{\Gamma}_{\boldsymbol{x}}(k,m)$	normalized speech correlation matrix
$\boldsymbol{\Gamma}_{\boldsymbol{x}'}(k,m)$	normalized correlation matrix of the uncorrelated speech component
$\boldsymbol{\Gamma}_{\boldsymbol{n}}(k,m)$	normalized noise correlation matrix
$\boldsymbol{\Gamma}_{\boldsymbol{u}}(k,m)$	normalized undesired correlation matrix
$\boldsymbol{R}_{\boldsymbol{\gamma}_{\boldsymbol{n}}}(k,m)$	correlation matrix of the normalized noise correlation vector

$\boldsymbol{\mu}_{\boldsymbol{\gamma_n}}(k)$	mean of the normalized noise correlation vector
$\boldsymbol{U}(k,m)$	matrix containing the orthogonal eigenvectors of $\boldsymbol{R}_{\boldsymbol{y}}(k,m)$
$\boldsymbol{\Psi}(k,m)$	matrix containing the orthogonal eigenvectors of $\boldsymbol{S}_{\boldsymbol{y}}(k,m)$
$\boldsymbol{V}(k,m)$	matrix containing the orthonormal eigenvectors of $\boldsymbol{\Gamma}^{w}_{\boldsymbol{y}}(k,m)$
$\boldsymbol{W}(k,m)$	matrix containing the orthonormal eigenvectors of $\boldsymbol{\Gamma}_{\boldsymbol{x}}(k,m)$
$\boldsymbol{\Upsilon}(k,m)$	diagonal matrix containing the eigenvalues of $\boldsymbol{R}_{\boldsymbol{y}}(k,m)$
$\boldsymbol{\Theta}(k,m)$	diagonal matrix containing the eigenvalues of $\boldsymbol{S}_{\boldsymbol{y}}(k,m)$
$\boldsymbol{\Lambda}^{w}_{\boldsymbol{y}}(k,m)$	diagonal matrix containing the eigenvalues of $\boldsymbol{\Gamma}^{w}_{\boldsymbol{y}}(k,m)$
$\boldsymbol{\Lambda}_{\boldsymbol{x}}(k,m)$	diagonal matrix containing the eigenvalues of $\boldsymbol{\Gamma}_{\boldsymbol{x}}(k,m)$
$\boldsymbol{\Lambda}^{w}_{\boldsymbol{x}}(k,m)$	diagonal matrix containing the eigenvalues of $\boldsymbol{\Gamma}^{w}_{\boldsymbol{x}}(k,m)$
$\boldsymbol{v}_q(k,m)$	q-th eigenvector of $\boldsymbol{V}(k,m)$
$\boldsymbol{w}_q(k,m)$	q-th eigenvector of $\boldsymbol{W}(k,m)$
$\upsilon_l(k,m)$	l-th eigenvalue of $\boldsymbol{\Upsilon}(k,m)$
$\varphi_l(k,m)$	l-th eigenvalue of $\boldsymbol{\Theta}(k,m)$
$\lambda^{w}_{\boldsymbol{y},q}(k,m)$	q-th eigenvalue of $\boldsymbol{\Lambda}^{w}_{\boldsymbol{y}}(k,m)$
$\lambda_{\boldsymbol{x},q}(k,m)$	q-th eigenvalue of $\boldsymbol{\Lambda}_{\boldsymbol{x}}(k,m)$
$\lambda^{w}_{\boldsymbol{x},q}(k,m)$	q-th eigenvalue of $\boldsymbol{\Lambda}^{w}_{\boldsymbol{x}}(k,m)$
$G(k,m)$	(real-valued) gain
$G_{\mathrm{WG}}(k,m)$	Wiener gain
$\boldsymbol{h}(k,m)$	FIR filter
$H_l(k,m)$	l-th FIR filter coefficient
$\boldsymbol{h}_{\mathrm{MFMVDR}}(k,m)$	MFMVDR filter
$\boldsymbol{h}_{\mathrm{MFMPDR}}(k,m)$	MFMPDR filter
$\boldsymbol{h}_{\mathrm{MFWF}}(k,m)$	MFWF filter
$\boldsymbol{h}_{\mathrm{MFMVDR\text{-}WG}}(k,m)$	MFMVDR filter and a WG as postfilter
$\boldsymbol{h}_{\mathrm{MFMPDR\text{-}WG}}(k,m)$	MFMPDR filter and a WG as postfilter
$J_{\mathrm{WG}}[\cdot]$	cost function of the WG
$\mathcal{L}_{\mathrm{MFMVDR}}[\cdot]$	Lagrangian function of the MFMVDR filter
$J_{\mathrm{MFWF}}[\cdot]$	cost function of the MFWF filter
$\mathcal{L}_{\mathrm{SC}}[\cdot]$	Lagrangian function of the SC optimization problem
$\mathcal{L}_{\mathrm{DC}}[\cdot]$	Lagrangian function of the DC optimization problem
$\epsilon(k,m)$	upper bound of the spherical uncertainty set
$\boldsymbol{\theta}(k,m)$	parameter vector in MDL selection criterion
$\boldsymbol{\theta}^{\mathrm{MDL}}(k,m)$	parameter vector in classical MDL selection criterion

$\boldsymbol{\theta}^{\mathrm{MDL}_\xi}(k,m)$	parameter vector in a-priori SNR-based MDL selection criterion
$\boldsymbol{S_y}(k,m)$	Schur complement
$\boldsymbol{C}(k,m)$	lower triangular Cholesky factor of $\boldsymbol{\Gamma_n}(k,m)$
$\beta_{\mathrm{SPP}}(k,m)$	SPP-based smoothing parameter for $\widehat{\phi}_N^{\mathrm{SPP}}(k,m)$
α_{SPP}	smoothing parameter for $\beta_{\mathrm{SPP}}(k,m)$
α_{DDA}	weighting parameter for $\widehat{\xi}^{\mathrm{DDA}}(k,m)$
α_y	smoothing parameter for $\widehat{\phi}_Y(k,m)$, $\widehat{\boldsymbol{R}}_{\boldsymbol{y}}(k,m)$
α_x	smoothing parameter for $\widehat{\boldsymbol{R}}_{\boldsymbol{x}}(k,m)$
α_n	smoothing parameter for $\widehat{\boldsymbol{R}}_{\boldsymbol{n}}(k,m)$
ζ	adaptation speed for $\widehat{\phi}_N^{\mathrm{Min}}(k,m)$ in dB/s
$\xi_{\mathcal{H}_1}$	fixed a-priori SNR
$\xi_{\min}$	lower limit of the a-priori SNR
$G_{\min}$	lower limit of the WG
μ	Lagrange multiplier
P_f	false alarm rate
κ	scaling parameter for diagonal loading
r	degrees of freedom in $\boldsymbol{\theta}(k,m)$
$\mathcal{H}_1$, $\mathcal{H}_0$	hypothesis 1, hypothesis 0
$\boldsymbol{e}$	selection vector with first element equal to 1 and all other elements equal to 0
$\boldsymbol{E}$	$L\times(L-1)$-dimensional identity matrix with first column equal to 0
$\boldsymbol{I}_L$	$L \times L$-dimensional identity matrix
$\mathbf{0}_Q$	$Q \times Q$-dimensional zero matrix

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

Speech is one of the most natural forms of human communication, providing a way to convey thoughts, feelings and information. Technical developments in the last decades have led to a wide, diverse and flexible use of speech communication technology in many devices such as telephones, watches, headphones and hearing aids. In these speech communication devices, one or more microphones are used to capture the desired speech signal, while also unfortunately capturing undesired signals such as background noise (e.g., multi-talker babble noise) and reverberation, i.e., sound reflections from walls and other surfaces. In general, background noise and reverberation negatively affect speech communication, leading to a reduced speech quality and speech intelligibility, especially at low signal-to-noise ratios (SNRs). Hence, for many decades speech enhancement algorithms have been used in these devices, aiming at improving both speech quality as well as speech intelligibility [1–9].

While single-microphone speech enhancement algorithms exploit temporal and spectral features [3, 5, 6], multi-microphone speech enhancement algorithms are able to additionally exploit spatial features [1, 7, 9], such that they usually result in a higher speech enhancement performance. Nevertheless, single-microphone algorithms are required, e.g., when due to economical reasons or physical size limitations no additional microphone is available, or when multi-microphone algorithms are extended by a spectro-temporal filter to improve the overall speech enhancement performance [10, 11].

The main objective of single-microphone speech enhancement algorithms is to suppress the undesired background noise while, preserving the desired speech signal, i.e., not introducing speech distortion or other artifacts. Several single-microphone speech enhancement algorithms were proposed in the literature, either in the time-domain or in the time-frequency-domain, e.g., based on statistical models, subspace decomposition, Kalman filtering or machine-learning. Although most single-microphone speech enhancement algorithms are able to improve speech quality to some extent, only recently some algorithms were proposed that are also able to improve speech intelligibility, e.g., [12, 13].

In this thesis, we will focus on single-microphone speech enhancement algorithms in the short-time Fourier transform (STFT)-domain. It is frequently assumed that consecutive STFT coefficients are uncorrelated, such that an estimate of the speech

STFT coefficients can be obtained by applying a gain to the noisy speech STFT coefficients at each time-frequency point independently [5,6]. Since it is more realistic to assume that consecutive STFT coefficients are correlated, single-microphone multi-frame algorithms were proposed that aim at exploiting speech correlation across time-frames [4, 14–16]. Exploiting the speech interframe correlation in principle enables to keep speech distortion low while reducing the undesired background noise. However, the main challenge of multi-frame algorithms is to accurately estimate the highly time-varying speech interframe correlation from the noisy microphone signal. Motivated by the potential of single-microphone multi-frame speech enhancement algorithms, to keep speech distortion low while suppressing the undesired background noise, the main objective of this thesis is to **develop and evaluate novel robust methods to estimate the speech interframe correlation** from the noisy microphone signal, either based on robust beamforming approaches or exploiting a low-rank speech model.

This chapter is structured as follows. In Section 1.1, we discuss the typical temporal and spectral characteristics of speech and noise signals. In Section 1.2, we provide a general overview of single-microphone and multi-microphone speech enhancement algorithms. In Section 1.3, we give a separate overview of single-microphone multi-frame speech enhancement algorithms exploiting speech interframe correlation. In Section 1.4, we present the main contributions and outline of this thesis.

1.1 Spectro-temporal Characteristics of Signals

In practice only the noisy microphone signal containing a mixture of the desired speech signal and the undesired background noise is available. In order to develop appropriate speech enhancement algorithms, it is hence crucial to understand and exploit the typical temporal and spectral characteristics of speech and noise signals, which are discussed in the following sections.

1.1.1 *Speech Signals*

Although speech signals are highly non-stationary, i.e., the power spectral density (PSD) varies rapidly over time, short-term stationarity can be assumed for periods of about 10-30 ms [5]. In order to analyze both temporal and spectral characteristics of a signal, a time-frequency transform such as the STFT can be used [17]. Fig. 1.1 depicts the spectrogram of an exemplary speech signal at a sampling frequency of 16 kHz, using a STFT with a relatively short frame length of 8 ms, an overlap of 75 % and a square-root Hann analysis window. In general, speech signals can be roughly segmented into voiced sounds, unvoiced sounds and silent gaps [5]. While voiced sounds can be characterized as (quasi-)periodic signals having a harmonic structure and energy mainly at frequencies below 5 kHz, unvoiced sounds can be characterized as aperiodic noise-like signals with energy mainly at frequencies above 4 kHz.

Although it is frequently assumed that consecutive speech STFT coefficients are uncorrelated over time and frequency, it is more realistic to assume that they are correlated, especially when using short STFT analysis frames and/or large overlap between consecutive time-frames. In order to visualize the relationship of speech STFT coefficients between consecutive time-frames and consecutive frequency-bins, Fig. 1.2a and 1.2b show scatter plots of consecutive speech spectral amplitudes for time-frames at a center frequency of 500 Hz and 6 kHz, while Fig. 1.2c and 1.2d show scatter plots of consecutive speech spectral amplitudes for frequency-bins in voiced and unvoiced sounds, (based on the speech spectrogram from Fig. 1.1). From Fig. 1.2a and 1.2b it can be observed that the speech spectral amplitudes are highly correlated across consecutive time-frames, with decreasing correlation for higher frequencies. From Fig. 1.2c and 1.2d it can be observed that the speech spectral amplitudes are also correlated across consecutive frequency-bins, where the correlation is clearly larger in voiced sounds than in unvoiced sounds. Since the correlation across time-frames is larger than the correlation across frequency-bins, in this thesis we will only **exploit the speech correlation across time-frames** and neglect the speech correlation across frequency-bins.

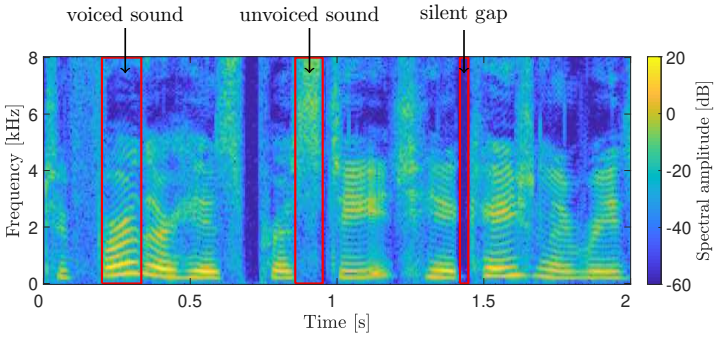


Fig. 1.1: Spectrogram of a speech signal with red boxes pointing to a voiced sound, an unvoiced sound and a silent gap. The STFT is performed using a frame length of 8 ms, an overlap of 75 % and a square-root Hann analysis window at a sampling frequency of 16 kHz.

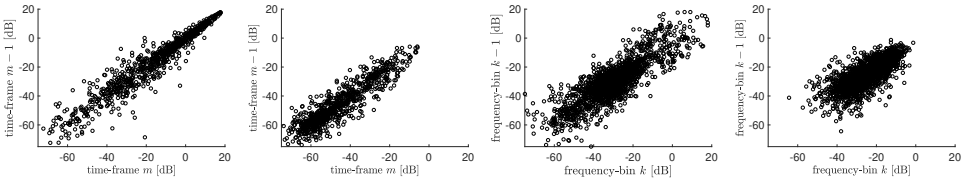


Fig. 1.2: Scatter plots of consecutive speech spectral amplitudes for (a) time-frames at a center frequency of 500 Hz, (b) time-frames at a center frequency of 6 kHz, (c) frequency-bins in voiced sound and (d) frequency-bins in unvoiced sound.

In [18, 19], it was proposed to mathematically describe a speech signal using a low-rank model, e.g., as linear combination of a finite number of complex exponentials. Due to the periodicity in voiced sounds the number of exponentials is smaller than in unvoiced sounds. In this thesis, we will **consider this low-rank speech model** to estimate the speech correlation across time-frames.

1.1.2 *Noise Signals*

Background noise occurs in our everyday life in a variety of situations, e.g., traffic on the street, fans in an office or people talking and dish clattering in a restaurant. Background noise can be roughly divided into stationary noise such as fan noise, and non-stationary noise such as babble noise, i.e., multiple people talking simultaneously in the background. In general, we assume that noise is more stationary than speech, i.e., its PSD varies slower over time, and is less correlated across time-frames. This implies that an interfering speaker is not considered as noise.

To investigate the temporal and spectral characteristic of different noise types, Fig. 1.3 depicts the exemplary spectrograms of PC-fan noise, traffic noise, i.e., a car passing a street, and restaurant noise, i.e., babble noise mixed with dish clattering. It can be observed that PC-fan noise is clearly more stationary than traffic noise

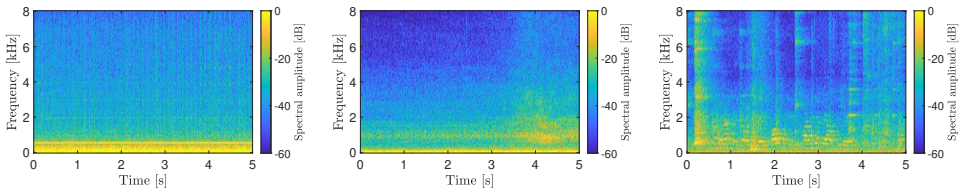


Fig. 1.3: Spectrogram of (a) PC-fan noise, (b) traffic noise and (c) restaurant noise. The STFT is performed using a frame length of 8 ms, an overlap of 75 % and a square-root Hann analysis window at a sampling frequency of 16 kHz.

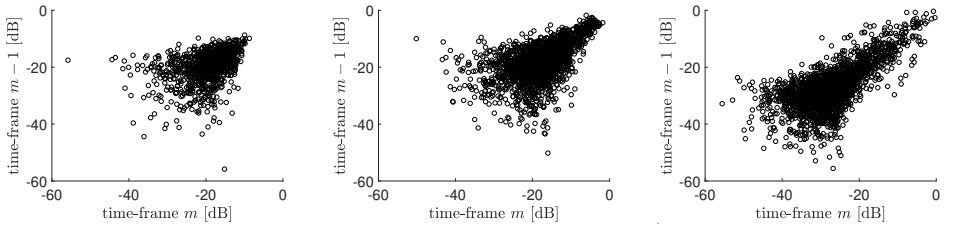


Fig. 1.4: Scatter plots of consecutive noise spectral amplitudes for time-frames at a center frequency of 500 Hz for (a) PC-fan noise, (b) traffic noise and (c) restaurant noise.

and restaurant noise. While most of the energy in the traffic noise is below 2 kHz, the non-stationary restaurant noise has a speech-like frequency range. Comparing the spectrograms of the noise signals in Fig. 1.3 with the spectrogram of the speech signal in Fig. 1.1, it can be observed that the considered noise signals are more stationary and have a different time-frequency structure than the speech signal.

Since in this thesis we aim at exploiting interframe correlation, Fig. 1.4 shows scatter plots of the noise spectral amplitudes for consecutive time-frames at a center frequency of 500 Hz for the three considered noise types. It can be observed that the spectral amplitudes of the more stationary PC-fan noise are less correlated over consecutive time-frames than those of the traffic noise and restaurant noise. Comparing the scatter plots of the noise spectral amplitudes in Fig. 1.4 with the scatter plots of the speech spectral amplitudes in Fig. 1.2a and 1.2b, it can be observed that speech is much more correlated over consecutive time-frames than noise, even for the speech-like restaurant noise.

1.2 Overview of Speech Enhancement Algorithms

The main objective of speech enhancement algorithms is to improve the speech quality and speech intelligibility of the noisy microphone signal(s) by suppressing the undesired background noise, while preserving the desired speech signal, i.e., not introducing speech distortion or other artifacts. In this section, we present a short overview of existing speech enhancement algorithms. A more detailed overview can be found, e.g., in [2–9]. To estimate the desired speech signal from the noisy microphone signal(s), a variety of speech enhancement algorithms were proposed, either in the time-domain or the time-frequency-domain, e.g., based on statistical models [6,20–28], subspace decomposition [29–45], Kalman filtering approaches [46–49] or machine-learning-based approaches [12, 50–62]. In this thesis, we focus on speech enhancement algorithms in the time-frequency-domain using the STFT.

In Section 1.2.1, we discuss single-microphone single-frame speech enhancement algorithms, where a (real-valued) gain is applied to the noisy speech STFT coefficients at each time-frequency point independently. In Section 1.2.2, we discuss multi-microphone speech enhancement algorithms, where a (complex-valued) filter is applied to the noisy speech STFT coefficients of the microphone signals. The focus of this thesis is on single-microphone multi-frame speech enhancement algorithms exploiting speech interframe correlation, where a (complex-valued) filter is applied to the noisy speech STFT coefficients. We will provide a separate overview of these algorithms in Section 1.3.

1.2.1 *Single-Microphone Single-Frame Speech Enhancement*

In single-microphone single-frame algorithms, it is generally assumed that consecutive time-frames and frequency-bins are uncorrelated, which is a valid assumption when using a sufficiently long frame length in the order of 20-30 ms and a small overlap of, e.g., 50 % [3, 5, 6]. Hence, each time-frequency point can be processed independently. Fig. 1.5 depicts the block scheme of a typical single-microphone

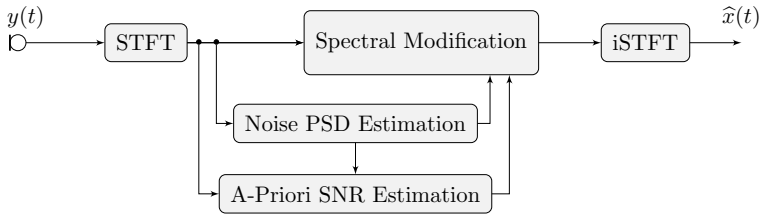


Fig. 1.5: A block scheme of a typical single-microphone speech enhancement algorithms in the STFT-domain, where $y(t)$ denotes the noisy microphone signal and $\widehat{x}(t)$ denotes the estimated speech signal in the time-domain.

speech enhancement algorithm in the STFT-domain. After transforming the noisy speech signal into the STFT-domain, an estimate of the speech STFT coefficients is obtained by modifying the noisy speech STFT coefficients, i.e., by applying a real-valued or a complex-valued gain. This gain is typically computed based on estimates of the noise PSD and the a-priori SNR. The estimated speech signal in the time-domain is then obtained by using the inverse STFT (iSTFT). In the following we present an overview of several (i) spectral modification approaches, (ii) noise PSD estimators and (iii) a-priori SNR estimators.

(i) *Spectral Modification*

One of the first proposed approaches for single-microphone speech enhancement was spectral subtraction [63, 64], where the speech spectral amplitude is estimated by simply subtracting an estimate of the noise spectral amplitude from the noisy speech spectral amplitude. To reconstruct the complex-valued speech STFT coefficient, the phase of the noisy speech STFT coefficient is used in combination with the estimated speech spectral amplitude. Since the spectral subtraction approach may lead to speech distortion and artifacts in the background noise, such as musical noise, many improvements were proposed, e.g., [65–68].

Although, the spectral subtraction approach is rather intuitive, it should be realized that it is not optimal. In contrast, the well-known Wiener gain (WG) [3, Sec. 11.3.1] [69] is derived by minimizing the mean-square error (MSE) between the speech STFT coefficient and the estimated speech STFT coefficient. The WG is also the optimal linear estimator assuming that the speech and noise STFT coefficients follow complex-valued Gaussian distributions.

Using the same statistical model for the speech and noise STFT coefficients and assuming that the phase of the speech STFT coefficients follow a uniform distribution in $[0; 2\pi]$, the minimum mean-square error (MMSE) short-time spectral amplitude (STSA) estimator and the MMSE logarithmic STSA estimator were derived in [20, 21]. Similarly, the maximum-likelihood (ML) and maximum a-posteriori (MAP) spectral amplitude estimator were derived in [65, 70]. Since for frame lengths larger than 5 ms it was observed that the speech STFT coefficients are more likely to

follow a super-Gaussian distribution [71], e.g., a Laplace or Gamma distribution, several spectral amplitude estimators were derived under this statistical model assumption [22–25, 27, 28, 53].

All aforementioned spectral amplitude estimators use the phase of the noisy speech STFT coefficients to estimate the complex-valued speech STFT coefficients, since it is assumed that the phase is perceptually less relevant than the amplitude. However, in the last years it was shown that the speech enhancement performance can be improved by utilizing phase information, especially for challenging non-stationary noise types [72–75].

Most single-microphone speech enhancement algorithms require estimates of the noise PSD and the a-priori SNR, which need to be estimated blindly from the noisy speech STFT coefficients. It should be noted that the estimation accuracy of both quantities has a strong influence on the speech enhancement performance. In general, overestimating the noise PSD may lead to speech distortion, while underestimating the noise PSD may lead to a large amount of residual noise.

(ii) *Noise PSD Estimation*

Exploiting the fact that speech signals contain silent gaps during which only the noise signal is active, the noise PSD can simply be estimated during these speech pauses, which can be determined using a voice activity detector (VAD), e.g., [76–78]. However, since the noise PSD is only updated during detected speech pauses, the performance of VAD-based noise PSD estimators is rather limited, especially for non-stationary noise, because the time-varying noise PSD can not be tracked when speech is active.

In order to also update the noise PSD when speech is active, a minimum statistics estimator was proposed in [79]. This noise PSD estimator is based on tracking the minimum of temporally smoothed noisy speech periodograms within a large time-frame window of, e.g., 1.5 s. An additional bias compensation is required between the tracked minimum and the mean of a normally distributed random variable [80]. Although this approach is an improvement compared to VAD-based noise PSD estimators, its performance for non-stationary noise is still unsatisfactory, especially when the noise PSD increases rapidly, leading to an underestimation of the noise PSD.

In order to improve the noise PSD estimation accuracy for non-stationary noise, optimal noise periodogram estimators were proposed, which are subsequently smoothed over time to obtain an estimate of the noise PSD [26, 81, 82]. For instance, in [26] the MMSE estimator was derived under a speech presence and absence model, leading to a noise PSD estimator where the noisy speech periodogram is recursively smoothed using a speech presence probability (SPP)-based time- and frequency-dependent smoothing parameter. When speech is likely to be absent, the noise tracking is fast, whereas when speech is likely to be present, the noise tracking is slowed down in order to avoid an overestimation of the noise PSD. Other noise PSD estimators are based on, e.g., subspace decomposition [38] or

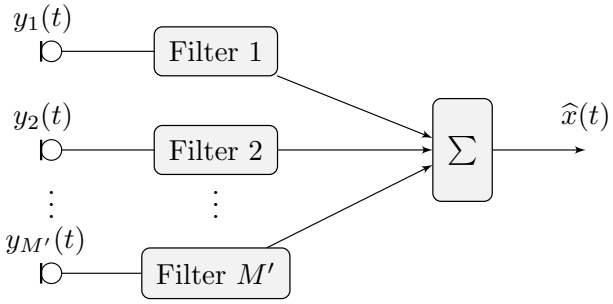


Fig. 1.6: Block scheme of a typical multi-microphone speech enhancement algorithm using a filter-and-sum structure, where $y_{m'}(t)$ denotes the m'-th microphone signal with $m' \in \{1, 2, \ldots, M'\}$ and $\widehat{x}(t)$ denotes the estimated speech signal.

deep neural networks (DNNs) [55, 62] were proposed.

(iii) *A-Priori SNR Estimation*

Using an estimate of the noise PSD, the a-priori SNR can be estimated. A ML estimator was proposed in [20], which may lead to unpleasant artifacts in the background noise, e.g., musical noise caused by random outliers. To reduce such artifacts, the decision-directed approach (DDA) was proposed in [20], where the ML estimate is weighted with an estimate of the previous speech STFT coefficient. This simple but effective approach is one of the most commonly used a-priori SNR estimators, e.g., in [83–86]. Alternatively, the more sophisticated cepstro-temporal smoothing approach was proposed in [87, 88], where the ML estimate is transformed in the cepstral-domain and only cepstral coefficients that are irrelevant for speech are smoothed, while important cepstral coefficients, e.g. related to the fundamental frequency, are preserved.

A simple but effective way to mask artifacts in the background noise is to apply a spectral floor, i.e., a lower limit, to the a-priori SNR estimate or the real-valued gain before processing the noisy speech STFT coefficients [5, 6]. This however comes with a trade-off between noise reduction and artifact reduction.

1.2.2 *Multi-Microphone Speech Enhancement*

While single-microphone speech enhancement algorithms exploit temporal and spectral information of the speech and noise signals, multi-microphone speech enhancement algorithms are able to additionally exploit spatial information of the speech and noise sources. To obtain an estimate of the desired speech signal, multi-microphone speech enhancement algorithms typically use a filter-and-sum structure (see Fig. 1.6), where complex-valued filter coefficients are applied to the noisy speech STFT coefficients of the microphone signals before being summed up. Multi-microphone speech enhancement algorithms, also commonly referred to as beam-

formers, can be roughly classified in data-independent (fixed) beamformers and data-dependent (adaptive) beamformers [1, 89].

Data-independent beamformers are designed to focus on signals arriving from a certain direction. The most commonly used examples are the delay-and-sum beamformer [90], simply summing up delay-compensated microphone signals, and the superdirective beamformer [91–93], maximizing the array gain for a spatially isotropic noise field. Steering these beamformers to the desired speech source requires either a-priori knowledge or an estimate of the direction of the desired speech source and the steering vector for this direction, e.g., based on an analytical model or a database with measured steering vectors.

In order to adapt the beamformer to time-varying noise fields, data-dependent beamformers were proposed, such as the minimum variance distortionless response (MVDR) beamformer [94] and the multi-channel Wiener filter (MWF) [34, 95, 96]. In general, data-dependent beamformers lead to a higher speech enhancement performance than data-independent beamformers, but require an estimate of the signal statistics, e.g., the noise correlation matrix, and the relative transfer function (RTF) vector of the desired speech source [9].

The MVDR beamformer aims at minimizing the output PSD of the undesired background noise, while not distorting the desired speech signal in a reference microphone [9,94,97,98]. In theory, the MVDR beamformer is equivalent to the minimum power distortionless response (MPDR) beamformer, which aims at minimizing the total signal output PSD, while not distorting the desired speech signal in a reference microphone [89]. Both beamformers require an estimate of a correlation matrix (i.e., the noise correlation matrix for the MVDR beamformer and the noisy speech correlation matrix for the MPDR beamformer) and the RTF vector of the desired speech source. When the correlation matrix and the RTF vector are perfectly estimated, no speech distortion occurs and the output SNR is maximized for both beamformers. However, it was shown that the MPDR beamformer is more sensitive to estimation errors (especially of the RTF vector/steering vector) than the MVDR beamformer [99,100]. Hence, several extensions of the MPDR beamformer were proposed to increase robustness against estimation errors in the steering vector. One of the most popular approaches is diagonal loading, which corresponds to imposing a quadratic inequality constraint on the filter vector [101]. However, since diagonal loading does not explicitly address uncertainty of the steering vector, several other approaches were proposed, e.g., by imposing (equality and/or inequality) constraints on the so-called mismatch vector, i.e., the difference between the steering vector and the presumed steering vector [102–110]. The robust MPDR beamformers in [104, 106] estimate the steering vector as the vector maximizing the total signal output PSD of the MPDR beamformer within a spherical uncertainty set.

Similarly to the WG (cf. Section 1.2.1), the MWF aims at minimizing the MSE between the output speech signal and the desired speech signal in a reference microphone [34, 95, 96]. Compared to the MVDR beamformer, the MWF leads to more noise reduction but also more speech distortion. In order to allow for a trade-off between noise reduction and speech distortion, the speech-distortion-weighted MWF (SDW-MWF) was derived by introducing a trade-off parameter in the MWF

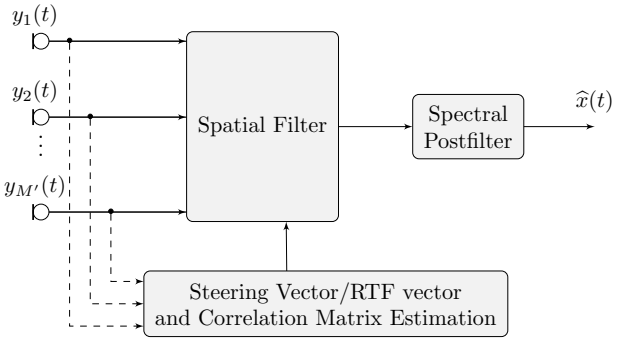


Fig. 1.7: Block scheme of a typical multi-microphone speech enhancement algorithm with spectral postfilter, where $y_{m'}(t)$ denotes the m'-th microphone signal with $m' \in \{1, 2, \ldots, M'\}$ and $\widehat{x}(t)$ denotes the estimated speech signal.

optimization problem [95, 96]. The larger the trade-off parameter, the more noise reduction is achieved but the more speech distortion occurs. For one desired speech source, it was shown in [11,111,112] that the MWF can be decomposed into a MVDR beamformer and a single-channel WG as spectral postfilter, which is statistically optimal in the MMSE sense under a Gaussian noise assumption. Fig. 1.7, depicts the block scheme of a beamformer with a spectral postfilter. The spectral postfilter typically allows to adapt much faster to spectral noise changes than only using a beamformer, leading to an improved speech enhancement performance [10, 11].

Similarly to the estimation of the noise PSD for single-microphone speech enhancement (cf. Section 1.2.1), the noise correlation matrix can be estimated based on a multi-microphone SPP, e.g., [113], where the noisy speech correlation matrix is recursively smoothed using an SPP-based smoothing parameter for each microphone. To estimate the RTF vector of the desired speech source, several methods were proposed, e.g., the covariance subtraction method [42, 114–117] and the covariance whitening method [39,42,115,116], which both rely on estimates of the noisy speech and noise correlation matrices.

1.3 Single-Microphone Multi-Frame Speech Enhancement Algorithms

For the single-microphone speech enhancement algorithms discussed in Section 1.2.1 it was assumed that consecutive STFT coefficients are uncorrelated over time, which is a valid assumption when using a sufficiently long frame length in the order of 16-32 ms and a small overlap of, e.g., 50 % [3, 5, 6] (see Fig. 1.8a). Hence, to obtain an estimate of the speech STFT coefficients, a (real-valued) gain can be applied to the noisy speech STFT coefficients at each time-frequency point independently. As mentioned before, although many single-frame speech enhancement algorithms are able to improve the speech quality, noise reduction is often accompanied by speech distortion, possibly affecting the speech intelligibility of the processed speech signal [118]. Aiming at exploiting that speech and/or noise STFT coefficients are

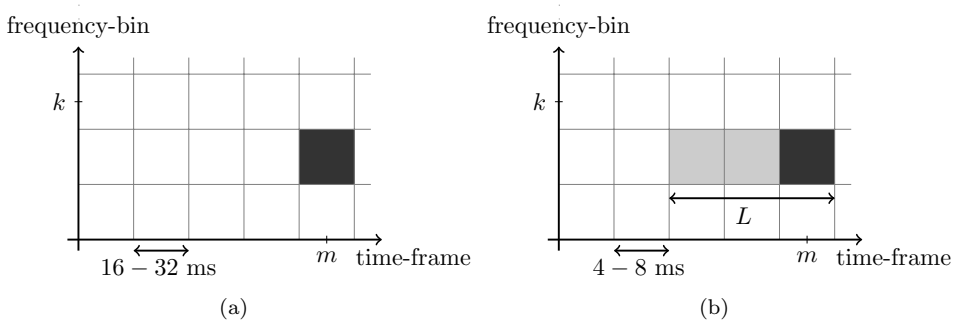


Fig. 1.8: Illustration of the considered time-frames using (a) single-frame signal model and (b) multi-frame signal model, where L denotes the number of consecutive time-frames.

correlated across time-frames (cf. Section 1.1), which is a valid assumption when using a short frame length in the order of 4-8 ms and a large overlap of, e.g., 50-85 % (see Fig. 1.8b), in the last decades, several single-microphone multi-frame speech enhancement algorithms were proposed [4,14–16,48,49,119–127]. In principle, these speech enhancement algorithms have the potential to reduce noise without introducing speech distortion.

Exploiting the speech interframe correlation using complex-valued linear prediction models in [48, 120, 127], modified Kalman filters were applied to the noisy speech STFT coefficients to obtain an estimate of the speech STFT coefficients. An extension was proposed in [49] by exploiting both speech and noise interframe correlation using a real-valued magnitude predictor. However, the speech enhancement performance of Kalman filters strongly depends on the highly time-varying speech transition matrix which is difficult to estimate blindly, i.e., only given the noisy speech signal.

Exploiting the speech interframe correlation using a multi-frame signal model [4, 14, 119], complex-valued finite-impulse response (FIR) filters were applied to the noisy speech vector, containing the current and previous noisy speech STFT coefficients (see Fig. 1.8b), to obtain an estimate of the speech STFT coefficients. In the multi-frame signal model the normalized speech correlation vector represents the speech interframe correlation with respect to the current speech STFT coefficient. Conceptually, this multi-frame signal model is similar to a multi-microphone signal model when interpreting time-frames as microphone inputs and the normalized speech correlation vector as the RTF vector of the desired speech source. Based on this multi-frame signal model, several beamformer-like algorithms were proposed for single-microphone speech enhancement, e.g., the multi-frame MVDR (MFMVDR) filter [14,119], the multi-frame Wiener filter (MFWF) [14] and the speech-distortion-weighted MFWF (SDW-MFWF) [14, 16]. The MFMVDR filter aims at minimizing the output PSD of the undesired signal, i.e., all signal components that are uncorrelated to the current speech STFT coefficient, while not distorting correlated speech component [14, 119]. This filter depends on the normalized speech correlation vec-

tor and the correlation matrix of the undesired signal. Typically, both quantities are highly time-varying and hence quite difficult to accurately estimate in practice. Therefore, similarly to the MPDR beamformer (cf. Section 1.2.2), the multi-frame MPDR (MFMPDR) filter was derived [14, 119] using the noisy speech correlation matrix, which can be directly estimated from the noisy speech vector. Using an accurate estimate of the normalized speech correlation vector, it was shown in [14,119] that the MFMPDR filter achieves a good noise reduction performance and impressive results in terms of speech distortion, especially for a single-microphone speech enhancement algorithm. It should however be realized that the MFMVDR and MFMPDR filters are only equivalent when the normalized speech correlation vector can be perfectly estimated.

Several methods were proposed to estimate the highly time-varying normalized speech correlation vector from the noisy speech STFT coefficients. Based on the assumption that the speech and noise signals are uncorrelated and that speech is generally much more correlated over time-frames than noise (cf. Section 1.1.2), in [14,15,119] it was proposed to use a VAD to estimate and update the normalized noise correlation vector in speech pauses and to subtract this vector from the normalized noisy speech correlation vector to obtain an estimate of the normalized speech correlation vector. However, since this estimator strongly depends on the VAD and the stationarity assumption of the noise signal, more sophisticated estimators were proposed in [15]. Based on a statistical analysis of the normalized speech and noise correlation vectors, the real and imaginary parts of the normalized speech and noise correlation vectors can be assumed to follow multivariate Gaussian distributions, such that a MAP estimator and a ML estimator were proposed in [15]. The MAP estimator relies on pre-trained data of the correlation matrices of the normalized speech and noise correlation vectors. The accuracy of the MAP estimator strongly depends on the validity of the pre-trained correlation matrices and the a-priori SNR estimate. The ML estimator is purely data-driven. The accuracy of the ML estimator only depends on the a-priori SNR estimate. Especially for low a-priori SNRs, the ML estimate may become very large, such that the estimation error between the normalized speech correlation vector and the ML estimate may become very large. In addition, outliers in the a-priori SNR estimate may negatively affect both the MAP and ML estimates, resulting in unpleasant artifacts in the background noise or even introducing speech distortion in the processed speech signal [15, 128]. In [16], it was proposed to estimate the normalized speech correlation vector by first estimating the speech periodogram in a high frequency-resolution STFT filterbank and then applying the Wiener-Khinchin theorem, which states that the correlation of a wide-sense stationary process is given by the inverse discrete Fourier transform (iDFT) of the PSD. This estimator strongly depends on the accuracy of the estimated speech periodogram in the high frequency-resolution STFT filterbank.

1.4 Outline of the Thesis and Main Contributions

Motivated by the potential to keep speech distortion low while reducing the undesired background noise, this thesis deals with single-microphone multi-frame speech enhancement algorithms exploiting speech interframe correlation. The main focus is to analyze and improve the speech enhancement performance of the practically feasible MFMPDR filter by developing novel robust methods to estimate the normalized speech correlation vector.

The main contributions of this thesis are threefold. First, in order to better understand the performance of multi-frame speech enhancement algorithms, we **investigated the sensitivity of the MFMVDR and MFMPDR filters to estimation errors** in the normalized speech correlation vector. We showed that accurately estimating the normalized speech correlation vector is crucial, since even small estimation errors lead to a degraded performance, resulting in speech distortion and unpleasant artifacts. Second, **inspired by robust beamforming algorithms, we developed robust constrained MFMPDR filters** that estimate the normalized speech correlation vector by maximizing the total signal output PSD within a spherical uncertainty set. The main novelty lies in setting the upper bound of this spherical uncertainty set based on the time-varying a-priori SNR for each time-frequency point. Simulation results show that the proposed constrained approaches yield a more accurate estimate of the normalized speech correlation vector than the state-of-the-art ML estimate, leading to a more natural speech quality and less noise distortion but a more consecutive noise reduction. Third, **based on a low-rank speech model we derived matrix-based methods to estimate the normalized speech correlation vector**, where we estimate the speech model order based on the time-varying a-priori SNR for each time-frequency point. Simulation results show that the proposed matrix-based estimation methods yield a more accurate estimate of the normalized speech correlation vector than the vector-based ML estimator, leading to a better speech quality and more noise reduction while keeping speech distortion low. We extensively compared the performance of all proposed algorithms using instrumental performance measures and evaluated the most promising algorithms using a subjective listening test.

In the remainder of this section, we provide a chapter-by-chapter overview of this thesis summarizing the main contributions. In addition, we provide a list of publications that were produced in the context of this thesis. A structured overview of the thesis is given in Fig. 1.9.

In **Chapter 2**, we introduce the single-frame and multi-frame signal models in the STFT-domain and their usage for single-microphone speech enhancement. Furthermore, we define the instrumental performance measures used in this thesis to evaluate the speech enhancement performance.

In **Chapter 3**, we briefly review single-frame speech enhancement algorithms and multi-frame speech enhancement algorithms exploiting speech interframe correlation. More in particular, we discuss the single-frame WG, which is used as reference speech enhancement algorithm in this thesis, and discuss a-priori SNR and noise PSD estimators that are used in this thesis. Moreover, we introduce several multi-frame filters, namely the MFMVDR filter, the MFMPDR filter and the MFWF, which serve as the basis for the proposed algorithms in this thesis. We show that the MFWF can be decomposed into the MFMPDR filter and a single-frame WG as postfilter, which will be discussed in Appendix A and is related to the work published in [129, 130]. Furthermore, since the considered multi-frame filters are related to beamformers, we discuss the conceptual similarities and differences between single-microphone multi-frame and multi-microphone algorithms.

In **Chapter 4**, we investigate the sensitivity of the MFMVDR and MFMPDR filters to estimation errors in the normalized speech correlation vector. We compare the practically feasible MFMPDR filter with two oracle versions of the MFMVDR filter for different oracle and blind estimates of the normalized speech correlation vector. Simulation results show that accurately estimating the normalized speech correlation vector is crucial, since even small estimation errors may lead to a degraded performance of the MFMVDR and MFMPDR filters, resulting in speech distortion and unpleasant artifacts in the background noise. On the one hand, when using oracle normalized speech correlation vector estimates, the performance of the MFMPDR filter is very close to the performance of an oracle MFMVDR filter. On the other hand, when using the state-of-the-art blind ML estimate, the performance of the MFMPDR filter is strongly reduced due to large estimation errors. The content of this chapter is related to the work published in [131]. Motivated by these results, in Chapter 5 and Chapter 6 we derive novel normalized speech correlation vector estimators, aiming at improving the speech enhancement performance of the MFMPDR filter.

In **Chapter 5**, we investigate the potential of using concepts proposed for robust MPDR beamforming in the context of single-microphone multi-frame speech enhancement. We propose two constrained MFMPDR filters that estimate the normalized speech correlation vector as the vector maximizing the total signal output PSD within a spherical uncertainty set. This corresponds to imposing a quadratic inequality constraint on the mismatch vector with respect to the presumed normalized speech correlation vector, e.g., the ML estimate. Whereas the proposed singly-constrained (SC) MFMPDR filter only considers the quadratic inequality constraint to estimate the (non-normalized) speech correlation vector, the proposed doubly-constrained (DC) MFMPDR filter integrates a linear normalization constraint into the optimization problem to directly estimate the normalized speech correlation vector. To set the upper bound of the spherical uncertainty set, we propose to use a trained non-linear mapping function that depends on the time-varying a-priori SNR estimate for each time-frequency point. Simulation results for different speech signals, noise types and SNRs show that the proposed constrained approaches yield a more accurate estimate of the normalized

speech correlation vector than the state-of-the-art ML estimate. An instrumental and a perceptual evaluation indicate that both constrained MFMPDR filters lead to a more natural speech quality and less noise distortion, but a more conservative noise reduction performance than the state-of-the-art ML-MFMPDR filter, where the DC-MFMPDR filter is preferred in terms of overall quality compared to the SC-MFMPDR filter and the ML-MFMPDR filter. The content of this chapter is based on the following publications: [132–134].

In **Chapter 6**, we assume that speech signals can be modeled using a low-rank model, e.g., as a linear combination of a limited number of complex exponentials, such that the speech correlation matrix can be assumed to be rank-deficient. Based on this speech model, we propose two matrix-based methods to estimate the normalized speech correlation vector. Both methods are based on the eigenvalue decomposition of a matrix, which is either constructed by subtracting the estimated normalized noise correlation matrix from the estimated normalized noisy speech correlation matrix (i.e., the matrix-subtraction (MS) method) or by prewhitening the estimated normalized noisy speech correlation matrix with the estimated normalized noise correlation matrix (i.e., the subspace-decomposition (SD) method). The speech model order is either assumed to be fixed, i.e., time- and frequency-independent, or needs to be estimated for each time-frequency point. When using a limited amount of data, which is typically the case for the considered multi-frame filters, most classical model order selection criteria, such as the minimum description length (MDL) selection criterion, have a poor estimation accuracy. Hence, we propose to estimate the speech model order by incorporating the a-priori SNR into a classical model selection criterion, i.e., the MDL selection criterion. Simulation results for different speech signals, noise types and SNRs show that the proposed matrix-based methods yield a more accurate estimate of the normalized speech correlation vector than the vector-based ML estimate. An instrumental evaluation indicates that the SD-MFMPDR filter using the proposed a-priori SNR-based speech model order estimator leads to a better speech quality and more noise reduction than the state-of-the-art ML-MFMPDR filter, while keeping speech distortion low. The content of this chapter is based on the following publications: [135, 136].

In **Chapter 7**, we compare the speech enhancement performance of the most promising MFMPDR filters from Chapters 4, 5 and 6 with an oracle MFMVDR filter and with oracle and blind single-frame WGs. For several speech signals and noise types, we evaluate the algorithms using both instrumental performance measures as well as a subjective listening test. Using oracle estimators, the instrumental performance measures and the results from the subjective listening test show that the overall quality and the speech distortion for the oracle MFMVDR filter are better than for the oracle WG, while the noise reduction is similar. Using blind estimators, the instrumental performance measures indicate that the proposed SD-MFMPDR filter from Chapter 6 leads to a clearly better noise reduction performance than the proposed DC-MFMPDR filter from Chapter 5 and the state-of-the-art ML-MFMPDR filter and to a similar noise reduction performance

as the WG, while keeping speech distortion as low as the ML-MFMPDR filter. The results from the subjective listening test show that the perceived overall quality for the proposed DC-MFMPDR and SD-MFMPDR filters is significantly better than for the state-of-the-art ML-MFMPDR filter but shows no statistically significant difference to the WG.

In **Chapter 8**, we summarize the main contributions of the thesis and discuss possible directions for further research.

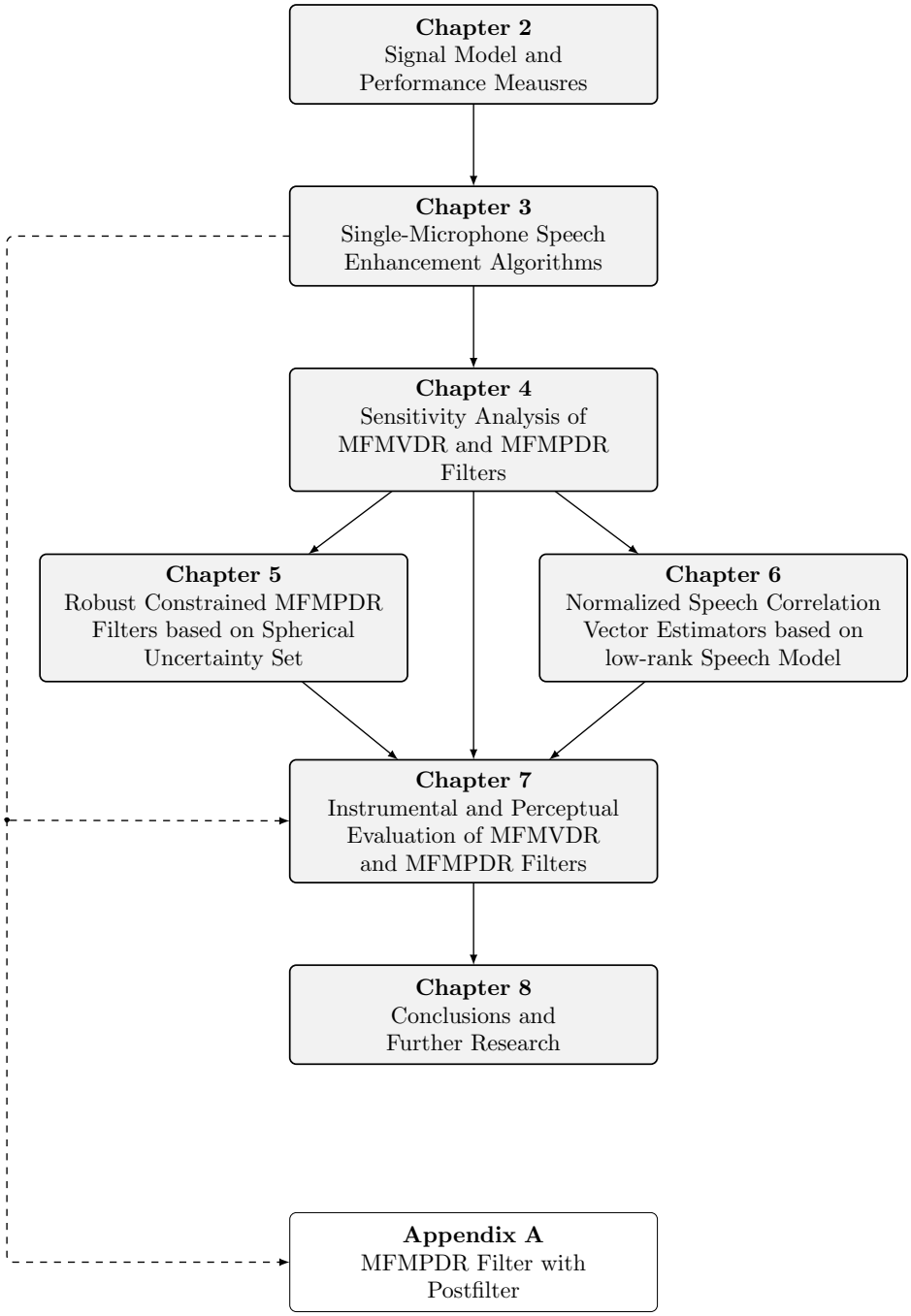


Fig. 1.9: Structure of the thesis.

2

PROBLEM FORMULATION AND INSTRUMENTAL PERFORMANCE MEASURES

In this chapter, we introduce the general notation, the single-microphone single-frame and multi-frame signal models in the short-time Fourier transform (STFT) domain, as well as the instrumental performance measures used in this thesis. In Section 2.1, we describe the single-frame signal model and present its extension to the multi-frame signal model, exploiting speech correlation across consecutive time-frames. Furthermore, the objective of speech enhancement is mathematically formulated using both the single-frame and the multi-frame signal model. In Section 2.2, we introduce several instrumental performance measures that will be used to evaluate the speech enhancement performance in the remainder of this thesis.

2.1 Problem Formulation

Consider the acoustic scenario depicted in Fig. 2.1, where a desired speech signal $x(t)$, with t denoting the discrete time index, is degraded by an undesired background noise $n(t)$ and is captured by a single microphone. Hence, the noisy speech signal $y(t)$ is given by

$$y(t) = x(t) + n(t)\,. \tag{2.1}$$

To exploit the temporal and spectral characteristics of speech and noise signals, single-microphone speech enhancement algorithms are often derived in the STFT-domain. Fig. 2.2 depicts the typical STFT processing scheme. The noisy speech signal $y(t)$ is split into time-frames of length T, which typically overlap to reduce boundary artifacts, and is weighted by a sliding analysis window $w_a(t)$ of length T. Since using a rectangular analysis window may lead to undesirable properties, e.g., low sidelobe attenuation, tapered analysis windows such as a Hann window, a Kaiser window or a square-root Hann window are often applied. Using the discrete Fourier transform (DFT) [17] of size $K \geq T$, each windowed time-frame is

transformed to the frequency-domain, resulting in the complex-valued noisy speech STFT coefficient $Y(k,m)$, i.e.,

$$Y(k,m) = \sum_{t=0}^{K-1} w_a(t) y(t+mR) e^{-j2\pi kt/K}\,, \tag{2.2}$$

where $k \in \{0,1,\ldots,K-1\}$ denotes the frequency-bin index, $m \in \{0,1,\ldots,M-1\}$ denotes the time-frame index, M denotes the total number of time-frames, R denotes the frame shift and $j^2 = -1$. Due to the linearity property of the DFT, the superposition in (2.1) also holds in the STFT-domain, i.e.,

$$Y(k,m) = X(k,m) + N(k,m)\,, \tag{2.3}$$

with $X(k,m)$ and $N(k,m)$ denoting the speech and noise STFT coefficients of the speech signal $x(t)$ and noise signal $n(t)$, respectively.

The estimated speech STFT coefficient $\widehat{X}(k,m)$ is obtained by applying a speech enhancement algorithm to the noisy speech STFT coefficient $Y(k,m)$. The estimated speech signal $\widehat{x}(t)$ is obtained by applying the weighted overlap-add method [17], i.e., by applying the inverse DFT (iDFT) to $\widehat{X}(k,m)$, weighting the resulting time-frame with a synthesis window $w_s(t)$ and adding the overlapping weighted time-frames, i.e.,

$$\widehat{x}(t) = \frac{1}{K} \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} w_s(t-mR) \widehat{X}(k,m) e^{j2\pi k(t-mR)/K}\,. \tag{2.4}$$

The synthesis window $w_s(t)$ is designed to provide perfect reconstruction, when no processing is applied, e.g., using a square-root Hann window as analysis and synthesis window.

In the following sections, we define the single-frame and multi-frame signal models, where the speech STFT coefficient $X(k,m)$ is estimated either by applying a (real-valued) gain or a (complex-valued) filter to the noisy speech STFT coefficient(s).

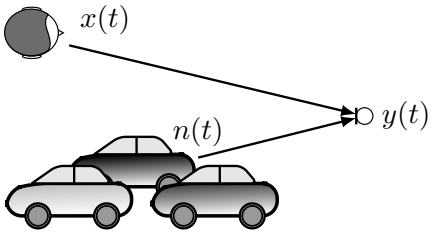


Fig. 2.1: Considered acoustic scenario with the speech signal $x(t)$, the noise signal $n(t)$ and the noisy speech signal $y(t)$.

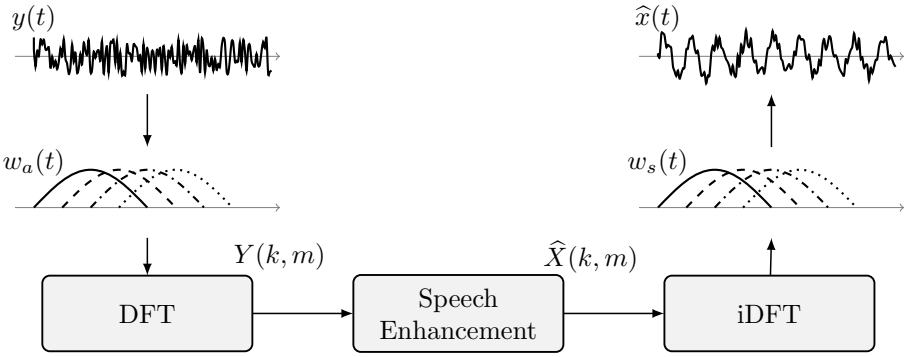


Fig. 2.2: Typical STFT processing scheme for single-microphone speech enhancement.

2.1.1 *Single-Frame Signal Model*

In single-frame speech enhancement algorithms, it is generally assumed that consecutive time-frames and frequency-bins are uncorrelated, which is a valid assumption when using a sufficiently long frame length T in the order of 16-32 ms and a small overlap of, e.g., 50 % [3,5,6]. Hence, each time-frequency point is processed independently, as depicted in Fig. 1.8a. To estimate the speech STFT coefficient, a (real-valued) gain $G(k,m)$ is applied to the noisy speech STFT coefficient $Y(k,m)$ [3,5,6], i.e.,

$$\widehat{X}(k,m) = G(k,m)Y(k,m)\,, \tag{2.5}$$

e.g., using the Wiener gain [3, Sec. 11.3.1] [69] or the minimum mean-square error (MMSE) short-time spectral amplitude estimator [20]. Assuming that the speech and noise signals are uncorrelated, i.e., $\mathcal{E}\left[X(k,m)N^*(k,m)\right] = 0$, with $\mathcal{E}\left[\cdot\right]$ denoting the expectation operator and $[\cdot]^*$ denoting the complex-conjugate operator, the noisy speech power spectral density (PSD) $\phi_Y(k,m) = \mathcal{E}\left[|Y(k,m)|^2\right]$ is given by

$$\phi_Y(k,m) = \phi_X(k,m) + \phi_N(k,m)\,, \tag{2.6}$$

where $\phi_X(k,m) = \mathcal{E}\left[|X(k,m)|^2\right]$ denotes the speech PSD and $\phi_N(k,m) = \mathcal{E}\left[|N(k,m)|^2\right]$ denotes the noise PSD. The a-priori SNR is defined as

$$\xi(k,m) = \frac{\phi_X(k,m)}{\phi_N(k,m)}\,. \tag{2.7}$$

In this thesis, the Wiener gain will be used as a reference single-frame speech enhancement algorithm (see Section 3.1.1).

2.1.2 *Multi-Frame Signal Model*

In multi-frame speech enhancement algorithms as in [4, 14–16], it is assumed that consecutive times-frames are correlated, which is a valid assumption when using a short frame length T in the order of 4-8 ms and a large overlap of, e.g., 50-85 %. To consider multiple time-frames, the m-th and the $L-1$ previous noisy speech STFT coefficients are stacked into an L-dimensional noisy speech vector as depicted in Fig. 1.8b, i.e.,

$$\boldsymbol{y}(k,m) = \Big[Y(k,m),\ Y(k,m-1),\ \ldots\ ,Y(k,m-L+1)\Big]^T, \tag{2.8}$$

where $[\cdot]^T$ denotes the transpose operator. Using (2.3), the noisy speech vector $\boldsymbol{y}(k,m)$ can be written as

$$\boldsymbol{y}(k,m) = \boldsymbol{x}(k,m) + \boldsymbol{n}(k,m), \tag{2.9}$$

where the speech vector $\boldsymbol{x}(k,m)$ and the noise vector $\boldsymbol{n}(k,m)$ are defined similarly as in (2.8).

Assuming that the speech and noise signals are uncorrelated, i.e., $\mathcal{E}\left[\boldsymbol{x}(k,m)\boldsymbol{n}^H(k,m)\right] = 0$, with $[\cdot]^H$ denoting the Hermitian operator, the $L\times L$-dimensional noisy speech correlation matrix $\boldsymbol{R}_{\boldsymbol{y}}(k,m) = \mathcal{E}\left[\boldsymbol{y}(k,m)\boldsymbol{y}^H(k,m)\right]$ is given by

$$\boldsymbol{R}_{\boldsymbol{y}}(k,m) = \boldsymbol{R}_{\boldsymbol{x}}(k,m) + \boldsymbol{R}_{\boldsymbol{n}}(k,m), \tag{2.10}$$

where $\boldsymbol{R}_{\boldsymbol{x}}(k,m) = \mathcal{E}\left[\boldsymbol{x}(k,m)\boldsymbol{x}^H(k,m)\right]$ and $\boldsymbol{R}_{\boldsymbol{n}}(k,m) = \mathcal{E}\left[\boldsymbol{n}(k,m)\boldsymbol{n}^H(k,m)\right]$ are the speech and noise correlation matrices, respectively. In [18, 19] it was proposed to mathematically model a speech signal using a low-rank model, e.g., as a linear combination of a finite number of complex exponentials, in which case the speech correlation matrix $\boldsymbol{R}_{\boldsymbol{x}}(k,m)$ can be assumed to be of rank Q with $Q \leq L$. We assume that the noise correlation matrix $\boldsymbol{R}_{\boldsymbol{n}}(k,m)$ is of full-rank (i.e., rank L), such that it is invertible and the noisy speech correlation matrix $\boldsymbol{R}_{\boldsymbol{y}}(k,m)$ is also of full-rank.

To exploit the speech correlation across consecutive time-frames, it was proposed in [4,14] to decompose the speech vector $\boldsymbol{x}(k,m)$ into the temporally correlated speech component $\boldsymbol{s}(k,m)$ and the temporally uncorrelated speech component $\boldsymbol{x'}(k,m)$ with respect to the speech STFT coefficient $X(k,m)$, i.e.,

$$\boldsymbol{x}(k,m) = \boldsymbol{s}(k,m) + \boldsymbol{x'}(k,m) = \boldsymbol{\gamma}_{\boldsymbol{x}}(k,m)X(k,m) + \boldsymbol{x'}(k,m)\,. \tag{2.11}$$

The (highly time-varying) *normalized speech correlation vector* $\boldsymbol{\gamma}_{\boldsymbol{x}}(k,m)$ is defined as

$$\boxed{\boldsymbol{\gamma}_{\boldsymbol{x}}(k,m) = \frac{\mathcal{E}\left[\boldsymbol{x}(k,m)X^*(k,m)\right]}{\mathcal{E}\left[|X(k,m)|^2\right]} = \frac{\boldsymbol{R}_{\boldsymbol{x}}(k,m)\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{R}_{\boldsymbol{x}}(k,m)\boldsymbol{e}}} \tag{2.12}$$

where $\boldsymbol{e} = \begin{bmatrix} 1, & 0, & \dots, & 0 \end{bmatrix}^T$ is an L-dimensional selection vector. Due to the normalization term $\boldsymbol{e}^T \boldsymbol{R}_{\boldsymbol{x}}(k,m)\boldsymbol{e}$, which corresponds to the speech PSD $\phi_X(k,m)$, the first element of the normalized speech correlation vector is equal to 1, i.e.,

$$\boldsymbol{e}^T \boldsymbol{\gamma}_{\boldsymbol{x}}(k,m) = 1\,, \tag{2.13}$$

which will be referred to as the normalization constraint in Chapter 5.
Substituting (2.11) into (2.9) we obtain

$$\boldsymbol{y}(k,m) = \boldsymbol{\gamma}_{\boldsymbol{x}}(k,m)X(k,m) + \boldsymbol{x'}(k,m) + \boldsymbol{n}(k,m)\,. \tag{2.14}$$

Considering the uncorrelated speech component $\boldsymbol{x'}(k,m)$ as an interference, we define the undesired signal vector as

$$\boldsymbol{u}(k,m) = \boldsymbol{x'}(k,m) + \boldsymbol{n}(k,m)\,, \tag{2.15}$$

such that the *multi-frame signal model* is given by

$$\boxed{\boldsymbol{y}(k,m) = \boldsymbol{\gamma}_{\boldsymbol{x}}(k,m)X(k,m) + \boldsymbol{u}(k,m)} \tag{2.16}$$

In [4, 14] it was proposed to estimate the speech STFT coefficient $X(k,m)$ by applying a (complex-valued) finite-impulse response (FIR) filter $\boldsymbol{h}(k,m)$ to the noisy speech vector $\boldsymbol{y}(k,m)$, i.e.,

$$\widehat{X}(k,m) = \boldsymbol{h}^H(k,m)\boldsymbol{y}(k,m), \tag{2.17}$$

where the filter $\boldsymbol{h}(k,m)$ contains L time-varying filter coefficients $H_l(k,m)$, i.e.,

$$\boldsymbol{h}(k,m) = \Big[H_0(k,m), H_1(k,m), \dots, H_{L-1}(k,m)\Big]^T. \tag{2.18}$$

Examples are the multi-frame Wiener filter (MFWF) [4, 14, 16], the multi-frame minimum variance distortionless response (MFMVDR) filter [4, 14] and the multi-frame minimum power distortionless response (MFMPDR) filter [4, 14, 119] (see Section 3.2).

Similarly to (2.12), the normalized noisy speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{y}}(k,m)$ and the normalized noise correlation vector $\boldsymbol{\gamma}_{\boldsymbol{n}}(k,m)$ are defined as

$$\boldsymbol{\gamma}_{\boldsymbol{y}}(k,m) = \frac{\boldsymbol{R}_{\boldsymbol{y}}(k,m)\boldsymbol{e}}{\boldsymbol{e}^T \boldsymbol{R}_{\boldsymbol{y}}(k,m)\boldsymbol{e}}, \tag{2.19}$$

$$\boldsymbol{\gamma}_{\boldsymbol{n}}(k,m) = \frac{\boldsymbol{R}_{\boldsymbol{n}}(k,m)\boldsymbol{e}}{\boldsymbol{e}^T \boldsymbol{R}_{\boldsymbol{n}}(k,m)\boldsymbol{e}}, \tag{2.20}$$

where $\boldsymbol{e}^T \boldsymbol{R}_{\boldsymbol{y}}(k,m)\boldsymbol{e}$ and $\boldsymbol{e}^T \boldsymbol{R}_{\boldsymbol{n}}(k,m)\boldsymbol{e}$ correspond to the noisy speech PSD $\phi_Y(k,m)$ and the noise PSD $\phi_N(k,m)$, respectively.

Using (2.6), (2.12), (2.19), and (2.20), it can be easily shown that

$$\phi_Y(k,m)\boldsymbol{\gamma}_{\boldsymbol{y}}(k,m) = \phi_X(k,m)\boldsymbol{\gamma}_{\boldsymbol{x}}(k,m) + \phi_N(k,m)\boldsymbol{\gamma}_{\boldsymbol{n}}(k,m)\,, \tag{2.21}$$

such that the normalized speech correlation vector can be written as

$$\boldsymbol{\gamma}_{\boldsymbol{x}}(k,m) = \frac{\xi(k,m)+1}{\xi(k,m)}\boldsymbol{\gamma}_{\boldsymbol{y}}(k,m) - \frac{1}{\xi(k,m)}\boldsymbol{\gamma}_{\boldsymbol{n}}(k,m)\,, \tag{2.22}$$

with $\xi(k,m)$ the a-priori SNR defined in (2.7).

Since the correlated speech component $\boldsymbol{s}(k,m)$ and the uncorrelated speech component $\boldsymbol{x'}(k,m)$ in (2.11) are uncorrelated by construction, the speech correlation matrix $\boldsymbol{R}_{\boldsymbol{x}}(k,m)$ can be decomposed into a rank-1 correlation matrix $\boldsymbol{R}_{\boldsymbol{s}}(k,m) = \mathcal{E}\left[\boldsymbol{s}(k,m)\boldsymbol{s}^H(k,m)\right]$ and a rank-$(Q-1)$ correlation matrix $\boldsymbol{R}_{\boldsymbol{x'}}(k,m) = \mathcal{E}\left[\boldsymbol{x'}(k,m)\boldsymbol{x'}^H(k,m)\right]$, where all entries of the first row and first column are equal to 0, i.e.,

$$\begin{aligned}\boldsymbol{R}_{\boldsymbol{x}}(k,m) &= \boldsymbol{R}_{\boldsymbol{s}}(k,m) + \boldsymbol{R}_{\boldsymbol{x'}}(k,m)\,,\\ &= \phi_X(k,m)\boldsymbol{\gamma}_{\boldsymbol{x}}(k,m)\boldsymbol{\gamma}_{\boldsymbol{x}}^H(k,m) + \boldsymbol{R}_{\boldsymbol{x'}}(k,m)\,.\end{aligned} \tag{2.23}$$

Substituting (2.23) into (2.10) with the undesired correlation matrix $\boldsymbol{R}_{\boldsymbol{u}}(k,m) = \mathcal{E}\left[\boldsymbol{u}(k,m)\boldsymbol{u}^H(k,m)\right]$ given by

$$\boldsymbol{R}_{\boldsymbol{u}}(k,m) = \boldsymbol{R}_{\boldsymbol{x'}}(k,m) + \boldsymbol{R}_{\boldsymbol{n}}(k,m)\,, \tag{2.24}$$

the noisy speech correlation matrix can be written as

$$\boldsymbol{R}_{\boldsymbol{y}}(k,m) = \phi_X(k,m)\boldsymbol{\gamma}_{\boldsymbol{x}}(k,m)\boldsymbol{\gamma}_{\boldsymbol{x}}^H(k,m) + \boldsymbol{R}_{\boldsymbol{u}}(k,m)\,. \tag{2.25}$$

We now define the normalized speech correlation matrix $\boldsymbol{\Gamma}_{\boldsymbol{x}}(k,m)$, the normalized noisy speech correlation matrix $\boldsymbol{\Gamma}_{\boldsymbol{y}}(k,m)$ and the normalized noise correlation matrix $\boldsymbol{\Gamma}_{\boldsymbol{n}}(k,m)$ as

$$\boldsymbol{\Gamma}_{\boldsymbol{x}}(k,m) = \frac{\boldsymbol{R}_{\boldsymbol{x}}(k,m)}{\boldsymbol{e}^T\boldsymbol{R}_{\boldsymbol{x}}(k,m)\boldsymbol{e}}, \tag{2.26}$$

$$\boldsymbol{\Gamma}_{\boldsymbol{y}}(k,m) = \frac{\boldsymbol{R}_{\boldsymbol{y}}(k,m)}{\boldsymbol{e}^T\boldsymbol{R}_{\boldsymbol{y}}(k,m)\boldsymbol{e}}, \tag{2.27}$$

$$\boldsymbol{\Gamma}_{\boldsymbol{n}}(k,m) = \frac{\boldsymbol{R}_{\boldsymbol{n}}(k,m)}{\boldsymbol{e}^T\boldsymbol{R}_{\boldsymbol{n}}(k,m)\boldsymbol{e}}. \tag{2.28}$$

Using (2.23), the normalized speech correlation matrix in (2.26) can be written as

$$\begin{aligned}\boldsymbol{\Gamma}_{\boldsymbol{x}}(k,m) &= \frac{\boldsymbol{R}_{\boldsymbol{s}}(k,m)}{\boldsymbol{e}^T\boldsymbol{R}_{\boldsymbol{x}}(k,m)\boldsymbol{e}} + \frac{\boldsymbol{R}_{\boldsymbol{x}}^{'}(k,m)}{\boldsymbol{e}^T\boldsymbol{R}_{\boldsymbol{x}}(k,m)\boldsymbol{e}}, \\ &= \boldsymbol{\gamma}_{\boldsymbol{x}}(k,m)\boldsymbol{\gamma}_{\boldsymbol{x}}^H(k,m) + \boldsymbol{\Gamma}_{\boldsymbol{x}'}(k,m)\,.\end{aligned} \tag{2.29}$$

2.2 Instrumental Performance Measures

Although listening tests are the best way to evaluate the performance of a speech enhancement algorithm, they are often quite time-consuming. Hence, listening tests are typically only performed at the end of the development process, while during development the speech enhancement algorithm is evaluated using instrumental performance measures. In this thesis, we will evaluate the speech enhancement performance in terms of speech quality and the amount of noise reduction, speech distortion, and noise distortion.

To evaluate speech quality, we will use the perceptual evaluation of speech quality (PESQ) measure [137] and the segmental SNR [138]. In [139,140], it was shown that the PESQ measure correlates well with the subjectively evaluated speech quality for speech enhancement algorithms. The PESQ measure compares a test signal with a reference signal, for which we use the clean speech signal $x(t)$. The reference signal and the test signal are time-aligned and compared using a model that accounts for perceptual aspects for the human auditory system. The output is a mean opinion score (MOS) in the range -0.5 to 4.5, corresponding to a very bad and a very good perceptual quality, respectively. In the remainder of this thesis, we will consider the PESQ improvement ΔPESQ, i.e., the difference between the PESQ MOS of the noisy speech signal $y(t)$ and the estimated speech signal $\widehat{x}(t)$. Consequently, a positive value indicates an improvement and a negative value a deterioration relative to the noisy speech signal.

The segmental SNR (segSNR) measure [138] considers both noise reduction as well as speech distortion. The segSNR compares a test signal with a reference signal, for which we also use the clean speech signal $x(t)$. The segSNR of the estimated speech signal $\widehat{x}(t)$ is defined as

$$\text{segSNR}(\widehat{x}(t)) = \frac{10}{|\mathbb{T}^Y|}\sum_{t\in\mathbb{T}^Y}\log_{10}\frac{\sum_{s=1}^{S}x^2(tS+s)}{\sum_{s=1}^{S}\left[x(tS+s)-\widehat{x}(tS+s)\right]^2}, \tag{2.30}$$

where S denotes the segment length and $\mathbb{T}^Y$ denotes the set of segments that contain speech-and-noise, defined as segments whose energy is larger than -45 dB with respect to the maximum segment energy. Similarly as in [141], we set the segment length S to 10 ms and use no overlapping time-frames. In the remainder of the the-

sis, we consider the segSNR improvement ΔsegSNR, i.e., the difference between the segSNR of the noisy speech signal $y(t)$ and the estimated speech signal $\widehat{x}(t)$, i.e.,

$$\Delta\text{segSNR} = \text{segSNR}(y(t)) - \text{segSNR}(\widehat{x}(t))\,. \tag{2.31}$$

Consequently, a positive value indicates an improvement and a negative value a deterioration relative to the noisy speech signal.

To evaluate the amount of noise reduction and speech distortion, we will use the segmental noise reduction (segNR) and the segmental speech distortion (segSSNR) [24], defined as

$$\text{segNR} = \frac{10}{|\mathbb{T}^Y|} \sum_{t\in\mathbb{T}^Y} \log_{10} \frac{\sum_{s=1}^{S} n^2(tS+s)}{\sum_{s=1}^{S} \tilde{n}^2(tS+s)}, \tag{2.32}$$

$$\text{segSSNR} = \frac{10}{|\mathbb{T}^Y|} \sum_{t\in\mathbb{T}^Y} \log_{10} \frac{\sum_{s=1}^{S} x^2(tS+s)}{\sum_{s=1}^{S} \left[x(tS+s) - \tilde{x}(sS+s)\right]^2}, \tag{2.33}$$

where $\tilde{x}(t)$ and $\tilde{n}(t)$ denotes the processed speech and noise signal, respectively. Note that higher segNR values indicate more noise reduction and higher segSSNR values indicate less speech distortion, which is both desired for speech enhancement algorithms.

To evaluate the noise distortion, more in particular the presence of musical noise artifacts, we will use the weighted log kurtosis ratio $\Delta\Psi_{\log}$ [142], which was shown to correlate well with perceptual listening results. This measure is defined as the natural logarithm of the ratio of the weighted kurtosis of the processed noise STFT coefficients $\tilde{N}(k,m)$ and the noise STFT coefficient $N(k,m)$, i.e.,

$$\Delta\Psi_{\log} = \log\left(\frac{\frac{1}{M}\sum_{m=1}^{M} \Psi_{\tilde{N}}(m)}{\frac{1}{M}\sum_{m=1}^{M} \Psi_{N}(m)}\right). \tag{2.34}$$

The weighted kurtosis $\Psi_N(m)$ is defined as

$$\Psi_N(m) = \frac{\frac{1}{K}\sum_{k=0}^{K-1} \left[\tau_N(k)|N(k,m)|^2 - \varpi_N(m)\right]^4}{\left(\frac{1}{K}\sum_{k=0}^{K-1} \left[\tau_N(k)|N(k,m)|^2 - \varpi_N(m)\right]^2\right)^2}, \tag{2.35}$$

with

$$\tau_N(k) = \left(\frac{1}{M} \sum_{m=1}^{M} |N(k,m)|^2 \right)^{-1}, \tag{2.36}$$

$$\varpi_N(m) = \frac{1}{K} \sum_{k=0}^{K-1} \tau_N(k) |N(k,m)|^2 . \tag{2.37}$$

The weighted kurtosis $\Psi_{\tilde{N}}(m)$ is defined similarly as $\Psi_N(m)$ in (2.35). Note that the perceived amount of noise distortion, especially musical noise, is lowest when $\Delta\Psi_{\log} = 0$ and higher $\Delta\Psi_{\log}$ values, i.e., $\Delta\Psi_{\log} > 0$ indicate more noise distortion.

3

SINGLE-MICROPHONE SPEECH ENHANCEMENT ALGORITHMS

In this chapter, we briefly review single-microphone single-frame speech enhancement algorithms and single-microphone multi-frame speech enhancement algorithms exploiting speech interframe correlation. In Section 3.1, we discuss the single-frame Wiener gain (WG), as well as typical estimators for the required quantities, i.e., the a-priori SNR estimators and the noise PSD estimators. In Section 3.2, we discuss the multi-frame minimum variance distortionless response (MFMVDR) filter, the multi-frame minimum power distortionless response (MFMPDR) filter and the multi-frame Wiener filter (MFWF) and show the relation between these filters. Since these multi-frame filters are related to multi-microphone speech enhancement algorithms, we discuss the conceptual similarities and differences between single-microphone multi-frame and multi-microphone algorithms.

3.1 Single-Frame Speech Enhancement Algorithms

In this section, we consider the single-frame estimation problem in (2.5), where the speech STFT coefficient $X(k,m)$ is estimated by applying a (real-valued) gain $G(k,m)$ to the noisy speech STFT coefficient $Y(k,m)$ for each time-frequency point independently. Typically, this gain requires estimates of the a-priori SNR and the noise PSD, which need to be estimated blindly from the noisy speech STFT coefficients. In the following, we introduce the single-frame WG [3, Sec. 11.3.1] [69], which is used as the reference speech enhancement algorithm in this thesis, as well as typical a-priori SNR and noise PSD estimators that are used in this thesis. For conciseness, the frequency-bin index k will be omitted in this chapter if not required. However, it should be realized that all calculations are performed for each time-frequency point.

3.1.1 *Wiener Gain*

The WG aims at minimizing the MSE between the speech STFT coefficient $X(m)$ and the estimated speech STFT coefficient $\widehat{X}(m)$ in (2.5) [3, Sec. 11.3.1] [69]. Using (2.3), the cost function of the WG is given by

$$J_{\mathrm{WG}}\left[G(m)\right] = \mathcal{E}\Big[\big|X(m) - G(m)\Big(X(m) + N(m)\Big)\big|^2\Big]. \tag{3.1}$$

Assuming that the speech and noise STFT coefficients are uncorrelated, i.e., $\mathcal{E}\left[X(m)N^*(m)\right] = 0$, and that $G(m)$ is real-valued, the cost function in (3.1) can be written as

$$J_{\mathrm{WG}}\left[G(m)\right] = \phi_X(m) - 2G(m)\phi_X(m) + G^2(m)\Big(\phi_X(m) + \phi_N(m)\Big). \tag{3.2}$$

Setting the derivative of $J_{\mathrm{WG}}\left[G(m)\right]$ with respect to $G(m)$ equal to zero results in the WG [3, Sec. 11.3.1] [69], i.e.,

$$G_{\mathrm{WG}}(m) = \frac{\phi_X(m)}{\phi_X(m) + \phi_N(m)}. \tag{3.3}$$

Using the definition of the a-priori SNR in (2.7), the WG can be expressed as

$$\boxed{G_{\mathrm{WG}}(m) = \frac{\xi(m)}{\xi(m) + 1}} \tag{3.4}$$

with $0 \leq G_{\mathrm{WG}}(m) \leq 1$. Fig. 3.1 shows the attenuation curve of the WG $G_{\mathrm{WG}}(m)$ in (3.4) as a function of the a-priori SNR $\xi(m)$. When the a-priori SNR $\xi(m)$ is large, i.e., $\xi(m) > 15\,\mathrm{dB}$, $G_{\mathrm{WG}}(m)$ is close to one, resulting in almost no attenuation of $Y(m)$, whereas when the a-priori SNR is low, i.e., $\xi(m) < -15\,\mathrm{dB}$, $G_{\mathrm{WG}}(m)$ is close to zero, resulting in a strong suppression of $Y(m)$.

In order to avoid unpleasant artifacts in the estimated speech signal or in the residual noise, e.g., musical noise caused by random outliers, a simple but effective way to mask such artifacts is to apply a lower limit $G_{\min}$ to the WG [3, 5, 6], i.e.,

$$\widehat{G}_{\mathrm{WG}}(m) = \max\left[G_{\mathrm{WG}}(m), G_{\min}\right]. \tag{3.5}$$

This however comes with a trade-off between noise reduction and artifact reduction. Setting $G_{\min}$ to high values results in less artifacts, but obviously reduces the amount of noise reduction.

In practice, the a-priori SNR needs to be estimated blindly from the noisy speech STFT coefficients, while the estimation accuracy has a strong influence on the speech enhancement performance. In general, underestimating the a-priori SNR may lead to speech distortion, while overestimating the a-priori SNR may lead to a large amount of residual noise. Since most a-priori SNR estimators depend on an estimate of the noise PSD, in the following sections we present different estimators for the a-priori SNR and the noise PSD, which will be used in the remainder of the thesis.

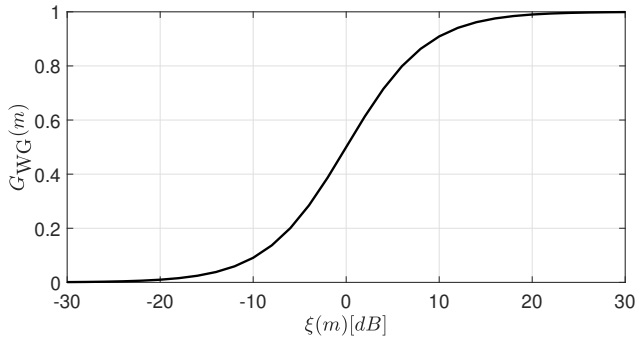


Fig. 3.1: Attenuation curve of the WG $G_{\mathrm{WG}}(m)$ as a function of the a-priori SNR $\xi(m)$.

3.1.2 *A-priori SNR Estimators*

To estimate the a-priori SNR, we will either use the maximum-likelihood (ML) estimator [20] or the decision-directed approach (DDA) [20].

Assuming that the speech and noise STFT coefficients are uncorrelated and follow complex Gaussian distributions, the noisy speech STFT coefficient $Y(m)$ also follows a complex Gaussian distribution given by

$$f\left(Y(m)|\phi_X(m),\phi_N(m)\right) = \frac{1}{\pi(\phi_X(m)+\phi_N(m))}\exp\left(-\frac{|Y(m)|^2}{\phi_X(m)+\phi_N(m)}\right) \quad (3.6)$$

Setting the derivative of (3.6) with respect to $\phi_X(m)$ equal to zero results in the ML estimate for the speech PSD $\phi_X(m)$ [20, 65], i.e.,

$$\widehat{\phi}_X^{\mathrm{ML}}(m) = |Y(m)|^2 - \widehat{\phi}_N(m), \quad (3.7)$$

with $\widehat{\phi}_N(m)$ denoting a noise PSD estimate. Dividing (3.7) by the noise PSD estimate $\widehat{\phi}_N(m)$, the ML estimate for the a-priori SNR is obtained as

$$\widehat{\xi}^{\mathrm{ML}}(m) = \frac{|Y(m)|^2}{\widehat{\phi}_N(m)} - 1, \quad (3.8)$$

where $\frac{|Y(m)|^2}{\widehat{\phi}_N(m)}$ represents an estimate of the a-posteriori SNR. Since $\widehat{\xi}^{\mathrm{ML}}(m)$ directly depends on the noise PSD estimate, estimation errors may lead to unpleasant artifacts in the background noise, e.g., musical noise caused by random outliers. Similarly to (3.5), to mask such artifacts, a lower limit $\xi_{\min}$ is applied to $\widehat{\xi}^{\mathrm{ML}}(m)$ [3,5,6]. This however comes with a trade-off between noise reduction and artifact reduction.

In [20], the DDA was proposed, where the ML estimate in (3.8) is weighted with an estimate of the a-priori SNR in the previous frame, i.e.,

$$\widehat{\xi}^{\mathrm{DDA}}(m) = \alpha_{\mathrm{DDA}} \frac{|\hat{X}(m-1)|^2}{\widehat{\phi}_N(m-1)} + (1-\alpha_{\mathrm{DDA}}) \max\left[\frac{|Y(m)|^2}{\widehat{\phi}_N(m)} - 1, \xi_{\min}\right], \quad (3.9)$$

with α_{DDA} denoting a weighting parameter. Setting α_{DDA} to higher values, i.e., close to 1, yields smoother estimates of the a-priori SNR than the ML estimate $\widehat{\xi}^{\mathrm{ML}}(m)$, typically leading to less musical tones but introducing more speech distortion than $\widehat{\xi}^{\mathrm{ML}}(m)$ [83].

3.1.3 *Noise PSD Estimators*

To estimate the noise PSD, we will either use the MMSE estimator proposed in [26] or the minimum tracking approach proposed in [143], as a simplified minimum statistics estimator.

In [26], it was proposed to estimate the noise PSD under a speech presence and absence model, i.e., given the hypotheses $\mathcal{H}_1$ that speech is present and $\mathcal{H}_0$ that speech is absent, these models are given by

$$\mathcal{H}_1 : Y(m) = X(m) + N(m), \quad (3.10)$$

$$\mathcal{H}_0 : Y(m) = N(m). \quad (3.11)$$

Assuming that the speech and noise STFT coefficients are uncorrelated and follow complex Gaussian distributions, the likelihoods under the hypotheses $\mathcal{H}_1$ and $\mathcal{H}_0$, i.e., $f(Y(m)|\mathcal{H}_1)$ and $f(Y(m)|\mathcal{H}_0)$, are also modeled by complex Gaussian distributions. Using the Bayes theorem and assuming that the prior probability $f(\mathcal{H}_1) = f(\mathcal{H}_0) = 0.5$, the posterior probability that speech is present, i.e., the SPP $f(\mathcal{H}_1|Y(m))$ is given by [26]

$$f(\mathcal{H}_1|Y(m)) = \left(1 + (1+\xi_{\mathcal{H}_1}) \exp\left(-\frac{Y(m)}{\widehat{\phi}_N(m-1)} \frac{\xi_{\mathcal{H}_1}}{\xi_{\mathcal{H}_1}+1}\right)\right)^{-1}, \quad (3.12)$$

with $\xi_{\mathcal{H}_1}$ the fixed a-priori SNR that is expected when $\mathcal{H}_1$ holds. In [26] an optimal value of $\xi_{\mathcal{H}_1} = -15\,\mathrm{dB}$ was derived by minimizing the total probability of error that depends on the false alarm rate, i.e., the probability that the SPP is lower than $f(\mathcal{H}_0)$ even though speech is present, and the missed-hit rate, i.e., the probability that the SPP is larger than $f(\mathcal{H}_1)$ even though speech is absent. Using the SPP in (3.12), the noise PSD $\phi_N(m)$ is estimated by recursive smoothing of the noisy speech periodogram using a SPP-based time- and frequency-dependent smoothing parameter $\beta_{\mathrm{SPP}}(m)$, i.e.,

$$\widehat{\phi}_N^{\mathrm{SPP}}(m) = \beta_{\mathrm{SPP}}(m)\widehat{\phi}_N^{\mathrm{SPP}}(m-1) + (1-\beta_{\mathrm{SPP}}(m))|Y(m)|^2, \quad (3.13)$$

with

$$\beta_{\mathrm{SPP}}(m) = \alpha_{\mathrm{SPP}} + (1 - \alpha_{\mathrm{SPP}}) f\left(\mathcal{H}_1 | Y(m)\right) , \tag{3.14}$$

where α_{SPP} denotes a smoothing parameter. When speech is likely to be absent, $\beta_{\mathrm{SPP}}(m) \approx \alpha_{\mathrm{SPP}}$, whereas when speech is likely to be present, $\beta_{\mathrm{SPP}}(m)$ is close to 1, such that the noise PSD in (3.13) is not updated. If $f\left(\mathcal{H}_1 | Y(m)\right)$ stuck at 1 for a longer time period, the noise PSD may strongly underestimated. To avoid such stagnation, the SPP is then set to a lower value.

In [79], a minimum statistics estimator was proposed, which is based on tracking the minimum value of temporally smoothed noisy speech periodograms within a large window of, e.g., 1.5 s. An additional bias compensation is required between the tracked minimum and the mean of a normally distributed random variable [80]. Since in this thesis we will typically use rather short time-frames of, e.g., 4 ms, the performance of the minimum statistics estimator however results in an underestimation of the noise PSD, especially when the noise PSD increases rapidly. Therefore, we will use a simplified minimum tracking approach proposed in [143]. This noise PSD estimator is based on tracking the minimum of the noisy speech PSD $\phi_Y(m)$ and assuming an adaptation speed for the noise PSD. First, the noisy speech PSD is estimated by recursive smoothing of the noisy speech periodogram as

$$\widehat{\phi}_Y(m) = \alpha_y \widehat{\phi}_Y(m-1) + (1 - \alpha_y)|Y(m)|^2 , \tag{3.15}$$

with α_y denoting a smoothing parameter. The noise PSD $\phi_N(m)$ is then estimated as

$$\widehat{\phi}_N^{\mathrm{Min}}(m) = \min\left[\widehat{\phi}_Y(m), \widehat{\phi}_N^{\mathrm{Min}}(m-1)\right](1 + \zeta) , \tag{3.16}$$

where the parameter ζ defines the adaptation speed in dB/s.

3.2 Multi-Frame Speech Enhancement Algorithms

In this section, we consider the multi-frame estimation problem in (2.17), where the speech STFT coefficient $X(m)$ is estimated by applying a complex-valued FIR filter $\boldsymbol{h}(m)$ to the noisy speech vector $\boldsymbol{y}(m)$, aiming at exploiting the speech interframe correlation. In the following, we introduce the MFMVDR and MFMPDR filters [4, 14, 119] and the MFWF [4, 14] and show the relation between these filters, i.e., the decomposition of the MFWF into the MFMPDR filter and a single-frame WG as postfilter. Since most multi-frame filters are inspired by multi-microphone algorithms, i.e., beamformers, we discuss the relationship between single-microphone multi-frame and multi-microphone speech enhancement algorithms, showing conceptual similarities and differences.

3.2.1 *Multi-Frame MVDR and MPDR Filters*

The MFMVDR filter aims at minimizing the output PSD of the undesired signal vector $\boldsymbol{u}(m)$ in (2.15), while not distorting the correlated speech component $\boldsymbol{s}(m)$ in (2.12) [4, 14, 119]. The cost function of the MFMVDR filter is given by

$$\min_{\boldsymbol{h}(m)} \ \mathcal{E}\left[|\boldsymbol{h}^H(m)\boldsymbol{u}(m)|^2\right], \quad \text{s.t.} \quad \boldsymbol{h}^H(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m) = 1. \tag{3.17}$$

Using the method of Lagrange multipliers [144], the Lagrangian function of the cost function in (3.17) is given by

$$\begin{aligned}\mathcal{L}_{\text{MFMVDR}}\left[\boldsymbol{h}(m), \mu\right] = \boldsymbol{h}^H(m)\boldsymbol{R}_{\boldsymbol{u}}(m)\boldsymbol{h}(m) + \mu\Big(\boldsymbol{h}^H(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m) - 1\Big) \\ +\mu^*\Big(\boldsymbol{\gamma}_{\boldsymbol{x}}^H(m)\boldsymbol{h}(m) - 1\Big)\end{aligned} \tag{3.18}$$

with μ the Lagrange multiplier. Setting the gradient of $\mathcal{L}_{\text{MFMVDR}}\left[\boldsymbol{h}(m), \mu\right]$ with respect to $\boldsymbol{h}(m)$ equal to zero leads to

$$\boldsymbol{h}(m) = -\frac{\mu}{2}\boldsymbol{R}_{\boldsymbol{u}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)\,. \tag{3.19}$$

Substituting (3.19) into (3.17), solving for μ and substituting this result into (3.19), results in the *MFMVDR filter* [4, 14, 119]

$$\boxed{\boldsymbol{h}_{\text{MFMVDR}}(m) = \frac{\boldsymbol{R}_{\boldsymbol{u}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)}{\boldsymbol{\gamma}_{\boldsymbol{x}}^H(m)\boldsymbol{R}_{\boldsymbol{u}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)}} \tag{3.20}$$

where the term $(\boldsymbol{\gamma}_{\boldsymbol{x}}^H(m)\boldsymbol{R}_{\boldsymbol{u}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m))^{-1}$ denotes the undesired output PSD

$$\phi_{\boldsymbol{u}}^{\text{out}}(m) = \mathcal{E}\left[\left|\boldsymbol{h}_{\text{MFMVDR}}^H(m)\boldsymbol{u}(m)\right|^2\right] = \frac{1}{\boldsymbol{\gamma}_{\boldsymbol{x}}^H(m)\boldsymbol{R}_{\boldsymbol{u}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)}. \tag{3.21}$$

The formula of the MFMVDR filter in (3.20) is very similar to the well-known MVDR beamformer for multi-microphone speech enhancement [89]. It should however be realized that the MFMVDR filter depends on the normalized speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$ in (2.12) and the undesired correlation matrix $\boldsymbol{R}_{\boldsymbol{u}}(m)$ in (2.24). Typically, both quantities are highly time-varying, making it quite difficult to accurately estimate them from the noisy speech STFT coefficients. In addition, it should be realized that the undesired correlation matrix $\boldsymbol{R}_{\boldsymbol{u}}(m)$ does not only contain the noise correlation matrix $\boldsymbol{R}_{\boldsymbol{n}}(m)$ but also the correlation matrix of the uncorrelated speech component $\boldsymbol{R}_{\boldsymbol{x}'}(m)$. Therefore, similarly to the MPDR beamformer (see Section 1.2.2), the MFMPDR filter was derived [4, 14, 119] using the noisy speech correlation matrix, which can be directly estimated from the noisy speech STFT coefficients. The MFMPDR filter aims at minimizing the output PSD of the noisy

speech vector, while not distorting the correlated speech component [4,14,119]. The cost function of the MFMPDR filter is given by

$$\min_{\boldsymbol{h}(m)} \ \mathcal{E}\left[|\boldsymbol{h}^H(m)\boldsymbol{y}(m)|^2\right], \quad \text{s.t.} \quad \boldsymbol{h}^H(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m) = 1. \tag{3.22}$$

Solving this optimization problem, using similar steps as in (3.18) and (3.19), yields the *MFMPDR filter* [4, 14, 119]

$$\boxed{\boldsymbol{h}_{\text{MFMPDR}}(m) = \frac{\boldsymbol{R}_{\boldsymbol{y}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)}{\boldsymbol{\gamma}_{\boldsymbol{x}}^H(m)\boldsymbol{R}_{\boldsymbol{y}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)}} \tag{3.23}$$

where the term $(\boldsymbol{\gamma}_{\boldsymbol{x}}^H(m)\boldsymbol{R}_{\boldsymbol{y}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m))^{-1}$ denotes the signal output PSD

$$\phi_{\boldsymbol{y}}^{\text{out}}(m) = \mathcal{E}\left[\left|\boldsymbol{h}_{\text{MFMPDR}}^H(m)\boldsymbol{y}(m)\right|^2\right] = \frac{1}{\boldsymbol{\gamma}_{\boldsymbol{x}}^H(m)\boldsymbol{R}_{\boldsymbol{y}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)}. \tag{3.24}$$

By applying the matrix inversion lemma [144]

$$(\boldsymbol{A} + \boldsymbol{a}\boldsymbol{a}^H)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{a}\boldsymbol{a}^H\boldsymbol{A}^{-1}}{1 + \boldsymbol{a}^H\boldsymbol{A}^{-1}\boldsymbol{a}}, \tag{3.25}$$

with $\boldsymbol{A} = \boldsymbol{R}_{\boldsymbol{u}}(m)$ and $\boldsymbol{a} = \sqrt{\phi_X(m)}\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$, to the noisy speech correlation matrix $\boldsymbol{R}_{\boldsymbol{y}}(m)$ in (2.25), it can be easily shown that

$$\boldsymbol{R}_{\boldsymbol{y}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m) = \boldsymbol{R}_{\boldsymbol{u}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)\,. \tag{3.26}$$

Using (3.26), it is shown that the MFMPDR filter in (3.23) is equivalent to the MFMVDR filter in (3.20). Although in practice it is obviously much easier to estimate the noisy speech correlation matrix $\boldsymbol{R}_{\boldsymbol{y}}(m)$ required in (3.23) instead of the undesired correlation matrix $\boldsymbol{R}_{\boldsymbol{u}}(m)$ required in (3.20), it should be realized that the MFMVDR filter and the MFMPDR filters are only equivalent when the normalized speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$ can be perfectly estimated. Similarly to the corresponding MVDR and MPDR beamformers for multi-microphone speech enhancement [9,89,100], it is however to be expected that the MFMPDR filter is more sensitive to estimation errors in the normalized speech correlation vector than the MFMVDR filter, which will be investigated in detail in Chapter 4. In Chapter 4, we will also discuss the ML estimator for the normalized speech correlation vector [15], which we will be used as the reference estimator in this thesis, while in Chapter 5 and 6, we will present novel estimators, either based on robust beamformer approaches or a low-rank speech model.

3.2.2 *Multi-Frame Wiener Filter*

Similarly to the single-frame WG in Section 3.1.1, the MFWF aims at minimizing the MSE between the speech STFT coefficient $X(m)$ and the estimated speech

STFT coefficient $\widehat{X}(m)$ in (2.17) [4, 14, 16]. Using the multi-frame signal model in (2.16), the cost function of the MFWF is given by

$$J_{\text{MFWF}}\left[\boldsymbol{h}(m)\right]=\mathcal{E}\left[\left|X(m)-\boldsymbol{h}^{H}(m)\Big(\boldsymbol{\gamma}_{\boldsymbol{x}}(m)X(m)+\boldsymbol{u}(m)\Big)\right|^{2}\right]. \tag{3.27}$$

Assuming that the speech and noise STFT coefficients are uncorrelated, the cost function in (3.27) can be written as

$$\begin{aligned} J_{\text{MFWF}}\left[\boldsymbol{h}(m)\right]=&\phi_X(m)-\phi_X(m)\boldsymbol{h}^{H}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)-\phi_X(m)\boldsymbol{\gamma}_{\boldsymbol{x}}^{H}(m)\boldsymbol{h}(m) \\ &+\boldsymbol{h}^{H}(m)\Big(\phi_X(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}^{H}(m)+\boldsymbol{R}_{\boldsymbol{u}}(m)\Big)\boldsymbol{h}(m). \end{aligned} \tag{3.28}$$

Setting the gradient of $J_{\text{MFWF}}\left[\boldsymbol{h}(m)\right]$ with respect to $\boldsymbol{h}(m)$ equal to zero results in the MFWF [4, 14, 16], i.e.,

$$\boldsymbol{h}_{\text{MFWF}}(m)=\phi_X(m)\Big(\phi_X(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}^{H}(m)+\boldsymbol{R}_{\boldsymbol{u}}(m)\Big)^{-1}\boldsymbol{\gamma}_{\boldsymbol{x}}(m)\,. \tag{3.29}$$

Using the definition of the noisy speech correlation matrix in (2.25), the MFWF can be written as

$$\boxed{\boldsymbol{h}_{\text{MFWF}}(m)=\phi_X(m)\boldsymbol{R}_{\boldsymbol{y}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)} \tag{3.30}$$

The MFWF in (3.30) depends on the speech PSD $\phi_X(m)$, the normalized speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$ and the noisy speech correlation matrix $\boldsymbol{R}_{\boldsymbol{y}}(m)$ in (3.30). When comparing the MFWF in (3.30) with the MFMPDR filter in (3.23), the filters can be summarized as

$$\boldsymbol{h}_{\text{MFWF}}(m)=c_{\text{MFWF}}(m)\boldsymbol{R}_{\boldsymbol{y}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m), \tag{3.31}$$

$$\boldsymbol{h}_{\text{MFMPDR}}(m)=c_{\text{MFMPDR}}(m)\boldsymbol{R}_{\boldsymbol{y}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m), \tag{3.32}$$

with

$$c_{\text{MFWF}}(m)=\phi_X(m), \tag{3.33}$$

$$c_{\text{MFMPDR}}(m)=\frac{1}{\boldsymbol{\gamma}_{\boldsymbol{x}}^{H}(m)\boldsymbol{R}_{\boldsymbol{y}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)}. \tag{3.34}$$

While in the case of the MFWF the term $\boldsymbol{R}_{\boldsymbol{y}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$ is multiplied by the speech PSD $\phi_X(m)$, resulting to $\boldsymbol{h}_{\text{MFWF}}(m)=0$ during speech pauses, in the case of the MFMPDR filter the term $\boldsymbol{R}_{\boldsymbol{y}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$ is normalized by $\boldsymbol{\gamma}_{\boldsymbol{x}}^{H}(m)\boldsymbol{R}_{\boldsymbol{y}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$. In practice, due to this normalization the MFMPDR filter may be more robust against numerical errors in the normalized speech correlation vector and the inverse of the noisy speech correlation matrix than the MFWF in (3.30). More robust results may be obtained by decomposing the MFWF into the MFMPDR filter and a single-frame WG as postfilter. Applying the matrix inversion lemma in (3.25) to (3.29),

the MFWF can be decomposed into the MFMVDR filter in (3.20) and a single-frame WG as postfilter, i.e.,

$$\boldsymbol{h}_{\text{MFMVDR-WG}}(m) = \boldsymbol{h}_{\text{MFMVDR}}(m)\frac{\phi_X(m)}{\phi_X(m) + \left(\boldsymbol{\gamma}_{\boldsymbol{x}}^H(m)\boldsymbol{R}_{\boldsymbol{u}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)\right)^{-1}}. \quad (3.35)$$

This postfilter operates on the output of the MFMVDR filter, where $\left(\boldsymbol{\gamma}_{\boldsymbol{x}}^H(m)\boldsymbol{R}_{\boldsymbol{u}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)\right)^{-1}$ denotes the undesired output PSD $\phi_{\boldsymbol{u}}^{\text{out}}(m)$ in (3.21). As shown in Section 3.2.1, using (3.26) the MFMVDR filter in (3.20) is equivalent to the MFMPDR filter in (3.23). Hence, (3.35) can also be formulated as

$$\boldsymbol{h}_{\text{MFMPDR-WG}}(m) = \boldsymbol{h}_{\text{MFMPDR}}(m)\frac{\phi_X(m)}{\phi_X(m) + \phi_{\boldsymbol{u}}^{\text{out}}(m)}. \quad (3.36)$$

While the MFMPDR filter is designed to avoid speech distortion, the postfilter aims at minimizing the MSE between the output of the MFMPDR filter and the speech STFT coefficient. Hence, the filter $\boldsymbol{h}_{\text{MFMPDR-WG}}(m)$ is capable to reduce the undesired signal more strongly than the MFMPDR filter, but speech distortion may be introduced. In Appendix A, we will evaluate the MFWF in (3.36) using different estimators for the undesired output PSD. In the remainder of the thesis, we will however focus on the MFMVDR and MFMPDR filters.

3.2.3 *Relationship between Single-Microphone Multi-Frame and Multi-Microphone Speech Enhancement Algorithms*

While single-microphone speech enhancement algorithms exploit temporal and spectral information of the speech and noise signals, multi-microphone speech enhancement algorithms, also referred to as beamformers, are able to additionally exploit spatial information of the speech and noise sources. In a frequently used multi-microphone signal model [7–9], the vector of the speech STFT coefficients of the microphone signals $\boldsymbol{x}(m)$ is modeled as the speech STFT coefficient $X(m)$ of a reference microphone signal multiplied by a relative transfer function (RTF) vector plus a residual speech component $\boldsymbol{x'}(m)$ (e.g., modeling late reverberation). The RTF vector is defined as the acoustic transfer function vector between the desired speech source and all microphones with respect to a reference microphone and hence depends on, e.g., the position of the desired source and the reverberation of the room. To estimate the desired speech STFT coefficients, complex-valued filter coefficients are applied to the noisy speech vector, containing the noisy speech STFT coefficients of the microphone signals, before being summed up. As mentioned in Section 1.2.2, commonly used beamformers are the delay-and-sum beamformer [89], the MVDR beamformer, [94] and its robust extensions [109], and the MWF [34, 95, 96]. In the single-microphone multi-frame model in (2.16), the speech vector $\boldsymbol{x}(m)$ is modeled as the speech STFT coefficient $X(m)$ in the current (reference) time-frame multiplied by the normalized speech correlation vector plus the temporally uncorrelated speech component $\boldsymbol{x'}(m)$. The normalized speech correlation vector contains statistical information about the speech correlation across consecutive time-frames with

respect to the current time-frame. To estimate the desired speech STFT coefficients, complex-valued filter coefficients are applied to the noisy speech vector, containing the noisy speech STFT coefficients in multiple time-frames, before being summed up.

Conceptually, this multi-frame signal model is hence very similar to the aforementioned multi-microphone signal model when interpreting time-frames as microphone inputs and the normalized speech correlation vector as the RTF vector of the desired speech source, such that concepts from multi-microphone speech enhancement can be applied to single-microphone speech enhancement when using a multi-frame signal model. However, when comparing both models in more detail, there are three major differences:

1. since the normalized speech correlation vector in the multi-frame signal model contains statistical information about speech correlation across consecutive time-frames, it is highly time-varying and needs to be estimated for each time-frequency point, whereas the RTF vector of the desired source in the multi-microphone signal model depends on spatial information and can hence be assumed to be more stationary than the normalized speech correlation vector.
2. while it can be assumed for the multi-microphone signal model that the correlation between the speech STFT coefficients in the different microphone signals is relatively high for each frequency-bin, for the multi-frame signal model the correlation between the speech STFT coefficients in consecutive time-frames can be rather low for certain time-frequency points, e.g., for noise-like speech sounds, (see Section 1.1)
3. while it can be assumed for the multi-microphone signal model that the term $\boldsymbol{x}'(m)$ is either not dominant (often this term is even completely neglected) or spatially stationary, for the multi-frame signal model the term $\boldsymbol{x}'(m)$ is temporally highly non-stationary and can even be dominant compared to the temporally correlated speech component $\boldsymbol{s}(m)$. This means that the influence of estimation errors in the normalized speech correlation vector for the MFMVDR and MFMPDR filters may be larger than the influence of estimation errors in the RTF vector of the desired source for the MVDR and MPDR beamformers.

3.3 Summary

In this chapter, we introduced several single-microphone single-frame and multi-frame speech enhancement algorithms. We discussed the single-frame WG, which will be used as the reference speech enhancement algorithm in the remainder of this thesis. In addition, we reviewed the ML estimator and the DDA to estimate the a-priori SNR and the MMSE estimator and the minimum tracking approach to estimate the noise PSD. Using the multi-frame signal model, we presented the MFMVDR filter, the related MFMPDR filter and the MFWF, which will serve as the basis for multi-frame speech enhancement algorithms proposed in this thesis. We showed the relation between these filters, i.e., the decomposition of the MFWF into the MFMPDR filter and a single-frame WG as postfilter. Furthermore, we showed

that the MFMVDR filter requires estimates of the undesired correlation matrix and the normalized speech correlation vector, where both quantities are highly time-varying. In Chapter 4, we will provide a sensitivity analysis of the MFMVDR and MFMPDR filters to estimation errors in both quantities, while in Chapters 5 and 6 we will propose novel methods to estimate the normalized speech correlation vector from the noisy speech STFT coefficients. Finally, since the presented multi-frame speech enhancement algorithms are related to multi-microphone algorithms, we discussed the conceptual similarities and differences between single-microphone multi-frame and multi-microphone algorithms.

4

SENSITIVITY ANALYSIS OF THE MFMVDR AND MFMPDR FILTERS TO ESTIMATION ERRORS

In Chapter 3, we introduced several single-microphone multi-frame filters, i.e., the multi-frame minimum variance distortionless response (MFMVDR) filter using the undesired correlation matrix and the multi-frame minimum power distortionless response (MFMPDR) filter using the noisy speech correlation matrix. In this chapter, we investigate the sensitivity of the MFMVDR and MFMPDR filters to estimation errors in the normalized speech correlation vector, which is highly time-varying and therefore difficult to accurately estimate. In practice, using an oracle estimate of the normalized speech correlation vector, it was shown in [4,14] that the MFMPDR filter achieves a good noise reduction performance and impressive results in terms of speech distortion, especially for a single-microphone speech enhancement algorithm. However, this oracle estimate requires the noise signal to be available. To blindly estimate the normalized speech correlation vector from the noisy speech STFT coefficients, several approaches were proposed. Based on the assumption that the real and imaginary parts of the normalized speech and noise correlation vectors follow multivariate Gaussian distributions in [15], a ML and a MAP estimator were proposed. Alternatively, in [16] it was proposed to estimate the normalized speech correlation vector by applying the Wiener-Khinchin theorem based on an estimated speech periodogram in a high frequency-resolution filterbank. In order to better understand the performance of the MFMVDR and MFMPDR filters, in this chapter we investigate the sensitivity of two (oracle) versions of the MFMVDR filter and the (blind) MFMPDR filter for different oracle and blind estimates of the normalized speech correlation vector. In Sections 4.1 and 4.2, we present several oracle and blind estimators for the normalized speech correlation vector and the undesired

This chapter is partly based on:

[131] D. Fischer and S. Doclo, "Sensitivity analysis of the multi-frame MVDR filter for single-microphone speech enhancement," in *Proc. of Europ. Signal Process. Conf. (EUSIPCO)*, Kos, Greece, Aug. 2017, pp. 633–637.

correlation matrix, respectively. In Section 4.3, we evaluate the speech enhancement performance of the MFMVDR filters and MFMPDR filter for different speech material, noise types and SNRs using instrumental performance measures. Simulation results show that even small estimation errors in the normalized speech correlation vector may strongly decrease the speech quality. For conciseness, the frequency-bin index k will be omitted in this chapter if not required. However, it should be realized that all calculations are performed for each time-frequency point.

4.1 Normalized Speech Correlation Vector Estimators

In this section, we present several oracle and blind estimators for the normalized speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$. For the oracle estimators, we make the (unrealistic) assumption that either the speech STFT coefficients $X(m)$ or the noise STFT coefficients $N(m)$ are available. In this case, the oracle speech and noise correlation matrix estimates can be computed using recursive smoothing as

$$\widehat{\boldsymbol{R}}_{\boldsymbol{x}}^{\mathrm{O}}(m) = \alpha_x \widehat{\boldsymbol{R}}_{\boldsymbol{x}}^{\mathrm{O}}(m-1) + (1-\alpha_x)\boldsymbol{x}(m)\boldsymbol{x}^H(m)\,, \tag{4.1}$$

$$\widehat{\boldsymbol{R}}_{\boldsymbol{n}}^{\mathrm{O}}(m) = \alpha_n \widehat{\boldsymbol{R}}_{\boldsymbol{n}}^{\mathrm{O}}(m-1) + (1-\alpha_n)\boldsymbol{n}(m)\boldsymbol{n}^H(m)\,, \tag{4.2}$$

with α_x and α_n denoting smoothing parameters. In practice, only the noisy speech STFT coefficients $Y(m)$ are obviously available such that a (blind) estimate of the noisy speech correlation matrix $\boldsymbol{R}_{\boldsymbol{y}}(m)$ can be computed as

$$\widehat{\boldsymbol{R}}_{\boldsymbol{y}}(m) = \alpha_y \widehat{\boldsymbol{R}}_{\boldsymbol{y}}(m-1) + (1-\alpha_y)\boldsymbol{y}(m)\boldsymbol{y}^H(m)\,. \tag{4.3}$$

4.1.1 *Oracle Estimators*

For the first oracle estimator, we assume that the oracle estimate of the speech correlation matrix $\widehat{\boldsymbol{R}}_{\boldsymbol{x}}^{\mathrm{O}}(m)$ in (4.1) is available. Hence, similarly to (2.12), the normalized speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$ can be estimated as

$$\boxed{\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{I}}(m) = \frac{\widehat{\boldsymbol{R}}_{\boldsymbol{x}}^{\mathrm{O}}(m)\boldsymbol{e}}{\boldsymbol{e}^T\widehat{\boldsymbol{R}}_{\boldsymbol{x}}^{\mathrm{O}}(m)\boldsymbol{e}}} \tag{4.4}$$

with $\boldsymbol{e}^T\widehat{\boldsymbol{R}}_{\boldsymbol{x}}^{\mathrm{O}}(m)\boldsymbol{e}$ an oracle estimate of the speech PSD $\phi_X(m)$.

For the second oracle estimator, we assume that the oracle estimate of the noise correlation matrix $\widehat{\boldsymbol{R}}_{\boldsymbol{n}}^{\mathrm{O}}(m)$ in (4.2) is available. Similarly to (2.19) and (2.20), the

normalized noisy speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{y}}(m)$ and the normalized noise correlation vector $\boldsymbol{\gamma}_{\boldsymbol{n}}(m)$ can then be estimated as

$$\widehat{\boldsymbol{\gamma}}_{\boldsymbol{y}}(m) = \frac{\widehat{\boldsymbol{R}}_{\boldsymbol{y}}(m)\boldsymbol{e}}{\boldsymbol{e}^T\widehat{\boldsymbol{R}}_{\boldsymbol{y}}(m)\boldsymbol{e}}, \tag{4.5}$$

$$\widehat{\boldsymbol{\gamma}}^{\mathrm{O}}_{\boldsymbol{n}}(m) = \frac{\widehat{\boldsymbol{R}}^{\mathrm{O}}_{\boldsymbol{n}}(m)\boldsymbol{e}}{\boldsymbol{e}^T\widehat{\boldsymbol{R}}^{\mathrm{O}}_{\boldsymbol{n}}(m)\boldsymbol{e}}, \tag{4.6}$$

with $\boldsymbol{e}^T\widehat{\boldsymbol{R}}_{\boldsymbol{y}}(m)\boldsymbol{e}$ an estimate of the noisy speech PSD $\phi_Y(m)$ and $\boldsymbol{e}^T\widehat{\boldsymbol{R}}^{\mathrm{O}}_{\boldsymbol{n}}(m)\boldsymbol{e}$ an oracle estimate of the noise PSD $\phi_N(m)$. Based on (2.22), an oracle estimate of the normalized speech correlation vector can then be obtained as

$$\boxed{\widehat{\boldsymbol{\gamma}}^{\mathrm{II}}_{\boldsymbol{x}}(m) = \frac{\widehat{\xi}^{\mathrm{O}}(m)+1}{\widehat{\xi}^{\mathrm{O}}(m)}\widehat{\boldsymbol{\gamma}}_{\boldsymbol{y}}(m) - \frac{1}{\widehat{\xi}^{\mathrm{O}}(m)}\widehat{\boldsymbol{\gamma}}^{\mathrm{O}}_{\boldsymbol{n}}(m)} \tag{4.7}$$

with $\widehat{\xi}^{\mathrm{O}}(m)$ an oracle estimate of the a-priori SNR $\xi(m)$, given by

$$\widehat{\xi}^{\mathrm{O}}(m) = \frac{\boldsymbol{e}^T\left(\widehat{\boldsymbol{R}}_{\boldsymbol{y}}(m) - \widehat{\boldsymbol{R}}^{\mathrm{O}}_{\boldsymbol{n}}(m)\right)\boldsymbol{e}}{\boldsymbol{e}^T\widehat{\boldsymbol{R}}^{\mathrm{O}}_{\boldsymbol{n}}(m)\boldsymbol{e}}. \tag{4.8}$$

Since $\widehat{\boldsymbol{R}}_{\boldsymbol{y}}(m)$ is not exactly equal to $\widehat{\boldsymbol{R}}^{\mathrm{O}}_{\boldsymbol{x}}(m) + \widehat{\boldsymbol{R}}^{\mathrm{O}}_{\boldsymbol{n}}(m)$, typically there will be a small difference between the oracle estimate $\widehat{\boldsymbol{\gamma}}^{\mathrm{I}}_{\boldsymbol{x}}(m)$ in (4.4) and the oracle estimate $\widehat{\boldsymbol{\gamma}}^{\mathrm{II}}_{\boldsymbol{x}}(m)$ in (4.7).

4.1.2 *Blind Estimators*

To blindly estimate the normalized speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$ from the noisy speech STFT coefficients $Y(m)$, different estimators were proposed, e.g., in [15, 16]. Based on a statistical analysis of the normalized speech and noise correlation vectors, an ML and a MAP estimator for the normalized speech correlation vector were proposed in [15]. While the MAP estimator relies on pre-trained data of the correlation matrices of the normalized speech and noise correlation vectors, the ML estimator is purely data-driven.

Using (2.21) and assuming that the real and imaginary parts of the normalized speech and noise correlation vectors are identically distributed, multivariate Gaus-

sian random variables, the likelihood of the normalized noisy speech correlation vector $\boldsymbol{\gamma_y}(m)$ given $\boldsymbol{\gamma_x}(m)$ is equal to [15]

$$f\left(\boldsymbol{\gamma_y}(m)|\boldsymbol{\gamma_x}(m)\right) = \frac{1}{\pi^L \det\left[\boldsymbol{R}_{\boldsymbol{\gamma_n}}(m)\right]} \exp\Bigg(-\Big((\xi(m)+1)\boldsymbol{\gamma_y}(m) - \xi(m)\boldsymbol{\gamma_x}(m) - \boldsymbol{\mu}_{\boldsymbol{\gamma_n}}\Big)^H \boldsymbol{R}_{\boldsymbol{\gamma_n}}^{-1}(m)\Big((\xi(m)+1)\boldsymbol{\gamma_y}(m) - \xi(m)\boldsymbol{\gamma_x}(m) - \boldsymbol{\mu}_{\boldsymbol{\gamma_n}}\Big)\Bigg), \tag{4.9}$$

with

$$\boldsymbol{\mu}_{\boldsymbol{\gamma_n}} = \mathcal{E}\left[\boldsymbol{\gamma_n}(m)\right], \tag{4.10}$$

$$\boldsymbol{R}_{\boldsymbol{\gamma_n}}(m) = \mathcal{E}\left[(\boldsymbol{\gamma_n}(m) - \boldsymbol{\mu}_{\boldsymbol{\gamma_n}})(\boldsymbol{\gamma_n}(m) - \boldsymbol{\mu}_{\boldsymbol{\gamma_n}})^H\right], \tag{4.11}$$

denoting the mean normalized noise correlation vector and the correlation matrix of the normalized noise correlation vector, respectively, and $\det[\cdot]$ the determinant of a matrix. Maximizing the logarithm of the likelihood in (4.9) results in the *ML estimator* for the normalized speech correlation vector, given by [15]

$$\boxed{\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{ML}}(m) = \frac{\widehat{\xi}(m)+1}{\widehat{\xi}(m)}\widehat{\boldsymbol{\gamma}}_{\boldsymbol{y}}(m) - \frac{1}{\widehat{\xi}(m)}\widehat{\boldsymbol{\mu}}_{\boldsymbol{\gamma_n}}} \tag{4.12}$$

with $\widehat{\xi}(m)$ an estimate of the a-priori SNR (cf. Section 3.1.2) and $\widehat{\boldsymbol{\mu}}_{\boldsymbol{\gamma_n}}$ an estimate of the mean normalized noise correlation vector. It should be noted that the ML estimator in (4.12) is very similar to the oracle estimator in (4.7). They mainly differ in the estimate of the normalized noise correlation vector $\boldsymbol{\gamma_n}(m)$, which is assumed to be constant for all time-frames in (4.12), such that it can be replaced by an estimate of its mean value $\widehat{\boldsymbol{\mu}}_{\boldsymbol{\gamma_n}}$.

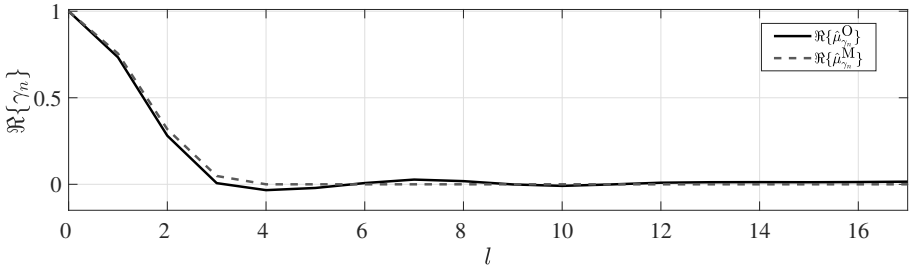


Fig. 4.1: Comparison of the (real part of the) oracle data-based estimate $\widehat{\boldsymbol{\mu}}_{\boldsymbol{\gamma_n}}^{\mathrm{O}}$ and the model-based estimate $\widehat{\boldsymbol{\mu}}_{\boldsymbol{\gamma_n}}^{\mathrm{M}}$ at frequency-bin $k = 5$. The STFT is performed using a frame length of 4 ms, an overlap of 75 % and a square-root Hann analysis window $w_a(t)$ at a sampling frequency of 16 kHz. The number of consecutive time-frames is set to $L = 18$. The smoothing parameter in (4.2) is set to $\alpha_n = 0.90$.

Fig. 4.1 depicts two estimates for the mean normalized noise correlation vector $\boldsymbol{\mu}_{\boldsymbol{\gamma}_{\boldsymbol{n}}}$ over the time-lag l, i.e., an oracle data-based estimate $\widehat{\boldsymbol{\mu}}^{\mathrm{O}}_{\boldsymbol{\gamma}_{\boldsymbol{n}}}$ and a model-based estimate $\widehat{\boldsymbol{\mu}}^{\mathrm{M}}_{\boldsymbol{\gamma}_{\boldsymbol{n}}}$. For the oracle data-based estimate, we pre-trained a mean normalized noise correlation vector based on the noise signals from the NOISEX-92 database [145]. During training, the noise STFT coefficients $N(m)$ are assumed to be available, such that the oracle noise correlation matrix $\widehat{\boldsymbol{R}}^{\mathrm{O}}_{\boldsymbol{n}}(m)$ can be computed as in (4.2). The oracle data-based estimate of the mean normalized noise correlation vector is subsequently obtained by averaging $\widehat{\boldsymbol{R}}^{\mathrm{O}}_{\boldsymbol{n}}(m)$ over all training data and applying (4.6). For the model-based estimate, white Gaussian noise is assumed as the input signal, such that for each frequency-bin k the mean normalized noise correlation vector is completely defined by the frame shift R and the STFT analysis window $w_a(t)$ [15], i.e.,

$$\widehat{\boldsymbol{\mu}}^{\mathrm{M}}_{\boldsymbol{\gamma}_{\boldsymbol{n}},l}(k) = \left(\frac{\sum_{t=0}^{K-1} w_a(t) w_a(t+lR)}{\sum_{t=0}^{K-1} w_a^2(t)} \right) e^{j2\pi Rkl/K} \quad \text{with } l = 0, 1, \ldots, L-1\,. \tag{4.13}$$

It can be observed that the oracle data-based estimate and the model-based estimate in Fig. 4.1 are very similar. This was to be expected, since averaging over a large dataset of noise signals is similar to white Gaussian noise. Hence, in the remainder of the thesis we will use the model-based estimate $\widehat{\boldsymbol{\mu}}^{\mathrm{M}}_{\boldsymbol{\gamma}_{\boldsymbol{n}}}$ in (4.13) to compute the ML estimate $\widehat{\boldsymbol{\gamma}}^{\mathrm{ML}}_{\boldsymbol{x}}$ in (4.12).

The ML estimate in (4.12) strongly depends on the a-priori SNR estimate $\widehat{\xi}(m)$. Especially for low a-priori SNRs, the ML estimate may become very large, such that the estimation error between $\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$ and $\widehat{\boldsymbol{\gamma}}^{\mathrm{ML}}_{\boldsymbol{x}}(m)$ may become very large. Furthermore, outliers in the a-priori SNR estimate may negatively affect the normalized speech correlation vector estimate $\widehat{\boldsymbol{\gamma}}^{\mathrm{ML}}_{\boldsymbol{x}}(m)$, resulting in unpleasant artifacts in the background noise or even introducing speech distortion in the processed speech signal as reported in [15, 128]. In Section 4.3.1, we evaluate the influence of the ML estimate for the different a-priori SNR estimators discussed in Section 3.1.2, i.e., the ML estimate $\widehat{\xi}^{\mathrm{ML}}(m)$ in (3.8) and the DDA estimate $\widehat{\xi}^{\mathrm{DDA}}(m)$ in (3.9), together with the noise PSD estimators discussed in Section 3.1.3, i.e., the SPP-based noise PSD estimate $\widehat{\phi}^{\mathrm{SPP}}_N(m)$ in (3.13) and the minimum tracking estimate $\widehat{\phi}^{\mathrm{Min}}_N(m)$ in (3.16).

4.2 Undesired Correlation Matrix Estimators

In this section, we present several oracle estimators for the undesired correlation matrix $\boldsymbol{R}_{\boldsymbol{u}}(m)$ of the MFMVDR filter in (3.20). For the oracle estimators, we make the (unrealistic) assumption that oracle estimates of the speech correlation matrix $\boldsymbol{R}_{\boldsymbol{x}}(m)$ and/or the noise correlation matrix $\boldsymbol{R}_{\boldsymbol{n}}(m)$ are available. Since both correlation matrices are typically highly time-varying, it is hardly feasible to blindly estimate them from the noisy speech STFT coefficients in practice. Therefore, in

practice we will only consider the MFMPDR filter in (3.23) using an estimate of the noisy speech correlation matrix $\widehat{\boldsymbol{R}}_{\boldsymbol{y}}(m)$ in (4.3).

Using (2.10) and (2.25), the undesired correlation matrix can be written as

$$\boldsymbol{R}_{\boldsymbol{u}}(m) = \boldsymbol{R}_{\boldsymbol{x}}(m) - \phi_X(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}^H(m) + \boldsymbol{R}_{\boldsymbol{n}}(m)\,. \tag{4.14}$$

For the first oracle estimator of the undesired correlation matrix, we use the oracle estimates of the speech correlation matrix $\widehat{\boldsymbol{R}}_{\boldsymbol{x}}^{\mathrm{O}}(m)$ in (4.1) and the noise correlation matrix $\widehat{\boldsymbol{R}}_{\boldsymbol{n}}^{\mathrm{O}}(m)$ in (4.2), and the oracle estimate of the normalized speech correlation vector $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{I}}$ in (4.4), i.e.,

$$\boxed{\widehat{\boldsymbol{R}}_{\boldsymbol{u}}^{\mathrm{I}}(m) = \widehat{\boldsymbol{R}}_{\boldsymbol{x}}^{\mathrm{O}}(m) - \widehat{\phi}_X^{\mathrm{O}}(m)\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{I}}(m)\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{I},H}(m) + \widehat{\boldsymbol{R}}_{\boldsymbol{n}}^{\mathrm{O}}(m)} \tag{4.15}$$

with $\widehat{\phi}_X^{\mathrm{O}}(m) = \boldsymbol{e}^T\widehat{\boldsymbol{R}}_{\boldsymbol{x}}^{\mathrm{O}}(m)\boldsymbol{e}$ an oracle estimate of the speech PSD. Using $\widehat{\boldsymbol{R}}_{\boldsymbol{u}}^{\mathrm{I}}(m)$ in (4.15) will be referred to as the perfect MFMVDR filter (MFMVDR$_{\mathrm{p}}$).

For the second oracle estimator, we assume that the correlation matrix of the uncorrelated speech component $\boldsymbol{R}_{\boldsymbol{x}'}(m)$ in (2.24) can be neglected, i.e., $\boldsymbol{R}_{\boldsymbol{u}}(m) = \boldsymbol{R}_{\boldsymbol{n}}(m)$. Using the oracle estimate of the noise correlation matrix $\widehat{\boldsymbol{R}}_{\boldsymbol{n}}^{\mathrm{O}}(m)$ in (4.2), the undesired correlation matrix can then be approximated as

$$\boxed{\widehat{\boldsymbol{R}}_{\boldsymbol{u}}^{\mathrm{II}}(m) = \widehat{\boldsymbol{R}}_{\boldsymbol{n}}(m)} \tag{4.16}$$

Using $\widehat{\boldsymbol{R}}_{\boldsymbol{u}}^{\mathrm{II}}(m)$ in (3.20) results in an approximated MFMVDR filter (MFMVDR$_{\mathrm{a}}$), which aims at minimizing the noise output PSD while not distorting the correlated speech component.

To avoid small numerical problems when computing the MFMVDR filter and the MFMPDR filter, we apply diagonal loading before computing the inverse of the undesired correlation matrix and the noisy speech correlation matrix, as suggested in [15], i.e.,

$$\widehat{\boldsymbol{R}}_{\boldsymbol{u}}^{-1}(m) = \left(\widehat{\boldsymbol{R}}_{\boldsymbol{u}}(m) + \frac{\kappa\mathrm{tr}\left[\widehat{\boldsymbol{R}}_{\boldsymbol{u}}(m)\right]}{L}\boldsymbol{I}_L\right)^{-1} \tag{4.17}$$

where κ denotes a small scaling parameter, the operator $\mathrm{tr}\,[\cdot]$ denotes the trace of a matrix and $\boldsymbol{I}_L$ denotes the $L \times L$-dimensional identity matrix.

4.3 Simulation Results

In this section, we compare the speech enhancement performance of the practically feasible MFMPDR filter with the oracle MFMVDR filters using instrumental performance measures. In Section 4.3.1, we describe the used audio material and discuss the algorithmic settings. In Section 4.3.2, we compare the speech enhancement performance between the MFMPDR filter and the oracle MFMVDR filters for the different oracle estimators of the normalized speech correlation vector presented in

Section 4.1. In Section 4.3.3, we investigate the performance of the MFMPDR filter using the blind state-of-the-art ML estimator for the normalized speech correlation vector presented in Section 4.1.2.

4.3.1 *Audio Material and Algorithmic Settings*

For the evaluation, we used 260 s of speech material (131 s female, 129 s male) from the TIMIT database [146] as speech signals, sampled at a sampling frequency of 16 kHz. As noise signals, we used two traffic noise signals, babble noise and factory noise taken from the NOISEX-92 database [145]. The considered SNRs ranged from 0 dB to 15 dB in 5 dB steps. The speech enhancement performance was evaluated in terms of the speech quality using the PESQ improvement (ΔPESQ) and the segmental SNR improvement ΔsegSNR (cf. Section 2.2). All performance measures were averaged over all considered speech signals and noise types.

Similarly as in [15], to exploit the speech correlation across consecutive time-frames, we used a highly temporally resolved STFT framework with a frame length of 4 ms ($T = K = 64$) and an overlap of 75 %, resulting in a frame shift of 1 ms ($R = 16$). As the STFT analysis window $w_a(t)$ and the synthesis window $w_s(t)$, we used a square-root Hann window.

The recursive smoothing parameters for the estimation of the correlation matrices in (4.1), (4.2) and (4.3) are experimentally set to $\alpha_x = 0.65$, $\alpha_n = 0.90$ and $\alpha_y = 0.92$, corresponding to a smoothing window of 2.5 ms, 10 ms and 12 ms, respectively. The scaling parameter in (4.17) is set to $\kappa = 0.001$.

For the blind ML estimate of the normalized speech correlation vector in (4.12), estimates of the a-priori SNR and the noise PSD are required. To estimate the a-priori SNR, we either used the ML estimate $\widehat{\xi}^{\mathrm{ML}}(m)$ in (3.8) or the DDA estimate $\widehat{\xi}^{\mathrm{DDA}}(m)$ in (3.9) with a weighting parameter of $\alpha_{\mathrm{DDA}} = 0.97$. To reduce fluctuations in the estimation of the a-priori SNR, we only updated the estimated speech STFT coefficient $\widehat{X}(m-1)$ in (3.9) every 4 ms, i.e., every 4 frames. To reduce the amount of speech distortion and to mask artifacts in the background noise, we applied a lower limit of $\xi_{\min} = -8\,\mathrm{dB}$ to the a-priori SNR estimate. To estimate the noise PSD, we either used the SPP-based noise PSD estimate $\widehat{\phi}_N^{\mathrm{SPP}}(m)$ in (3.13), with a smoothing parameter of $\alpha_{\mathrm{SPP}} = 0.90$ and a fixed a-priori SNR of $\xi_{\mathcal{H}_1} = -15\,\mathrm{dB}$, or the noise PSD estimate $\widehat{\phi}_N^{\mathrm{Min}}(m)$ in (3.16), with an adaptation speed of $\zeta = -5\,\mathrm{dB}$, as suggested in [15].

To set the filter length L, we investigated the influence of the filter length on the performance of the $\mathrm{MFMVDR_p}$, $\mathrm{MFMVDR_a}$ and MFMPDR filters using the (quasi-perfect) oracle estimate $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{I}}(m)$ in (4.4) for filter lengths between 2 and 20, corresponding to a data analysis length between 5 and 23 ms. For an SNR of 0 dB, Fig. 4.2 depicts the performance for different filter lengths L in terms of ΔPESQ and ΔsegSNR, averaged over all speech signals and noise types. It can be seen that both in terms of ΔPESQ and ΔsegSNR the performance increases with increasing filter length. While the ΔPESQ results for the MFMPDR filter saturate at about $L = 12$, the ΔPESQ results for the $\mathrm{MFMVDR_p}$ and $\mathrm{MFMVDR_a}$ filters saturate at about $L = 18$. For a fair comparison (independent of the computational complexity), in

Table 4.1: Overview of the applied correlation matrices.

Label	**Correlation matrix**
MFMVDR_p	Oracle undesired correlation matrix $\widehat{\boldsymbol{R}}_{\boldsymbol{u}}^{\text{I}}(m)$ in (4.15)
MFMVDR_a	Oracle noise correlation matrix $\widehat{\boldsymbol{R}}_{\boldsymbol{n}}(m)$ in (4.2)
MFMPDR	Blind noisy speech correlation matrix $\widehat{\boldsymbol{R}}_{\boldsymbol{y}}(m)$ in (4.3)

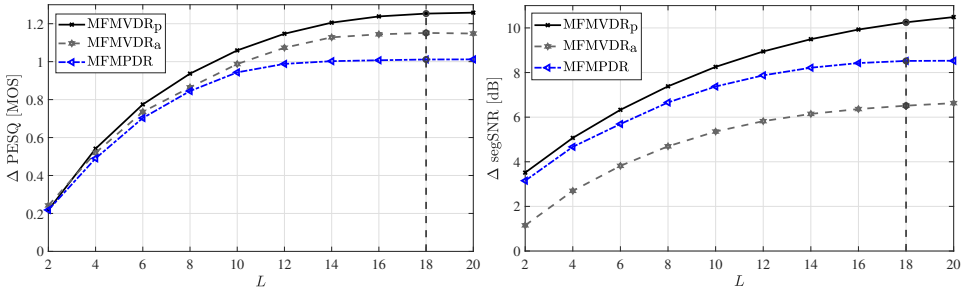


Fig. 4.2: Influence of the filter length L on the average (a) PESQ improvement and (b) segSNR improvement for the MFMVDR_p, MFMVDR_a and MFMPDR filters using $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\text{I}}(m)$ in (4.4) for 0 dB SNR.

this chapter we set the filter length to $L = 18$ for all filters, resulting in 21 ms of analysis data used in each filtering operation.

4.3.2 *Speech Enhancement Performance Using Oracle Normalized Speech Correlation Vector Estimates*

In this section, we compare the performance between the two oracle versions of the MFMVDR filter and the MFMPDR filter (see Table 4.1) for the oracle normalized speech correlation vector estimates $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\text{I}}(m)$ in (4.4) and $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\text{II}}(m)$ in (4.7).

Fig. 4.3(a) depicts the average performance in terms of ΔPESQ and ΔsegSNR of the MFMVDR_p, MFMVDR_a and MFMPDR filters (cf. Table 4.1) for different SNRs, using the oracle estimate $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\text{I}}(m)$ in (4.4). First, it can be seen that for all SNRs the MFMVDR_p filter achieves the highest PESQ and segSNR improvements, which are quite impressive for a single-microphone speech enhancement algorithm. Although the MFMPDR filter should theoretically be equivalent to the MFMVDR filter when using a (quasi-perfect) estimate of $\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$ (cf. Section 3.2.1), it can be observed that the performance of the MFMPDR filter is a bit worse than the MFMVDR_p filter, although still extremely good. This can be explained by the fact that in practice $\widehat{\boldsymbol{R}}_{\boldsymbol{y}}(m)$ is not exactly equal to $\widehat{\phi}_X(m)\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\text{I}}(m)\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\text{I},H}(m) + \widehat{\boldsymbol{R}}_{\boldsymbol{u}}^{\text{I}}(m)$ (cf. (2.25)). The

Table 4.2: Overview of the considered a-priori SNR and noise PSD estimates in the ML estimate of the normalized speech correlation vector.

Label	**Description**
$\text{MFMPDR}^{\text{ML}}_{\text{SPP}}$	$\widehat{\xi}^{\text{ML}}(m)$ in (3.8) with $\widehat{\phi}_N^{\text{SPP}}(m)$ in (3.13)
$\text{MFMPDR}^{\text{ML}}_{\text{Min}}$	$\widehat{\xi}^{\text{ML}}(m)$ in (3.8) with $\widehat{\phi}_N^{\text{Min}}(m)$ in (3.16)
$\text{MFMPDR}^{\text{DDA}}_{\text{SPP}}$	$\widehat{\xi}^{\text{DDA}}(m)$ in (3.9) with $\widehat{\phi}_N^{\text{SPP}}(m)$ in (3.13)
$\text{MFMPDR}^{\text{DDA}}_{\text{Min}}$	$\widehat{\xi}^{\text{DDA}}(m)$ in (3.9) with $\widehat{\phi}_N^{\text{Min}}(m)$ in (3.16)

performance of the MFMVDR_{a} filter, i.e., assuming that the uncorrelated speech correlation matrix can be neglected, is the worst of all considered filters, especially in terms of ΔsegSNR and for high SNRs. This implies that the influence of the uncorrelated speech component is crucial, especially at high SNRs, and neglecting this component increases the amount of speech distortion leading to a reduced speech quality.

Fig. 4.3(b) depicts the average performance in terms of ΔPESQ and ΔsegSNR of the MFMVDR_{p}, MFMVDR_{a} and MFMPDR filters using the oracle estimate $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\text{II}}(m)$ in (4.7), i.e., considering small estimation errors. Compared to the results in Fig. 4.3(a) using the (quasi-perfect) estimate $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\text{I}}(m)$, it can be observed that the performance for all filters decreases. For instance, for an SNR of 5 dB, the ΔPESQ results are reduced by 0.2 MOS for the MFMVDR_{p} filter, by 0.18 MOS for the MFMPDR filter and by 0.3 MOS for the MFMVDR_{a} filter. Nevertheless, the MFMVDR_{p} filter still outperforms both the MFMVDR_{a} filter and the MFMPDR filter, while the MFMPDR filter still yields excellent results. These results show that even small estimation errors in $\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$ decrease the overall performance. In addition, informal listening tests revealed slight artifacts in the background noise, which is due to the fact that $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\text{II}}(m)$ in (4.7) is not accurately estimate during speech pauses.

4.3.3 *Speech Enhancement Performance Using Blind Normalized Speech Correlation Vector Estimates*

In this section, we evaluate the performance of the practically feasible MFMPDR filter using the blind ML estimate for the normalized speech correlation vector $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\text{ML}}(m)$ in (4.12). Using the different a-priori SNR estimates in Section 3.1.2 and noise PSD estimates in Section 3.1.3 in the ML estimate, we compare four different versions of the blind MFMPDR filter (see Table 4.2). Fig. 4.3(c) depicts the average performance in terms of ΔPESQ and ΔsegSNR of these MFMPDR filters.

First, it can be observed that the $\text{MFMPDR}^{\text{ML}}_{\text{SPP}}$ filter and the $\text{MFMPDR}^{\text{DDA}}_{\text{SPP}}$ filter using the SPP-based noise PSD estimator achieve the highest ΔPESQ and ΔsegSNR results, where the performance difference between the ML estimate and the DDA estimate for the a-priori SNR is minor. The performance of the $\text{MFMPDR}^{\text{ML}}_{\text{Min}}$ filter is the worst, especially in terms of ΔPESQ. More importantly, comparing the per-

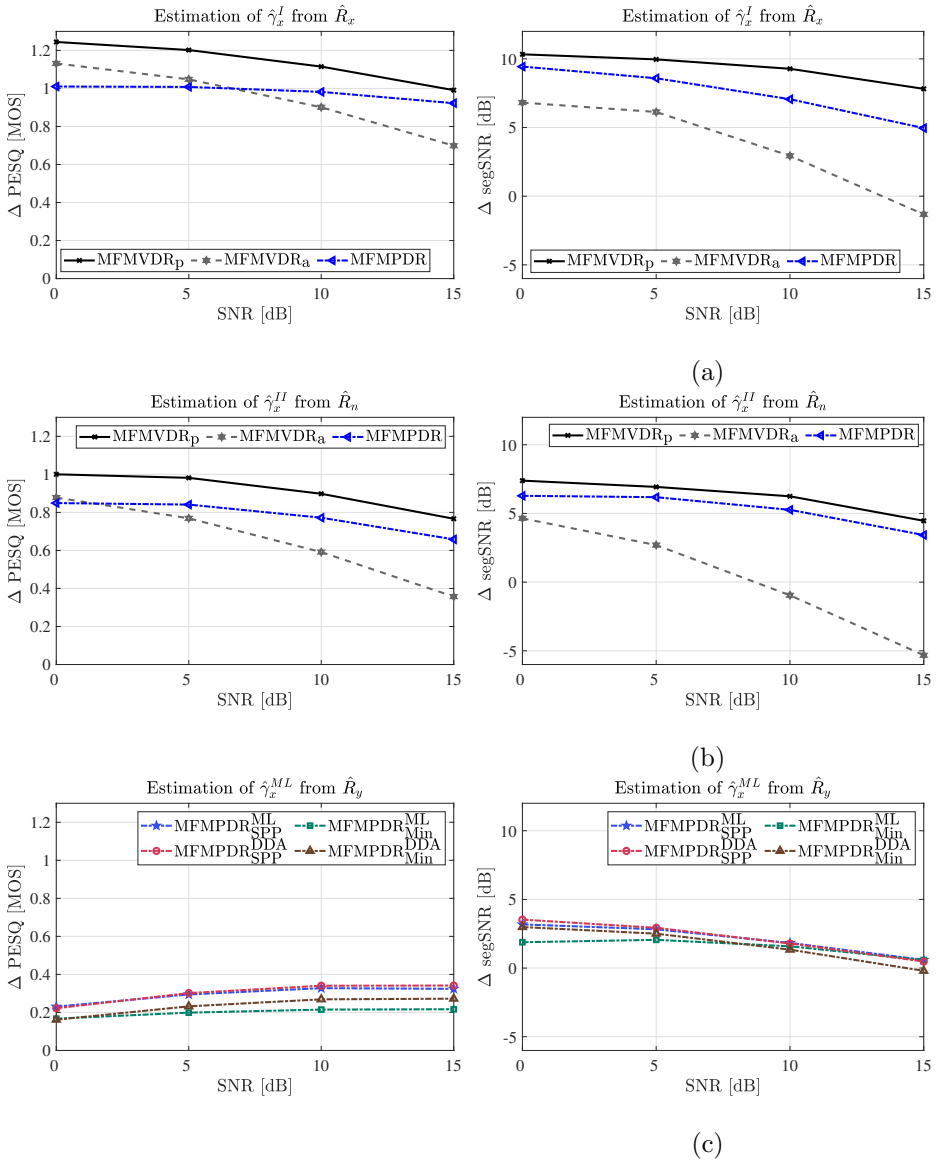


Fig. 4.3: Influence of different estimators for the normalized speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$ on the MFMVDR$_\mathrm{p}$, MFMVDR$_\mathrm{a}$ and MFMPDR filters: (a) oracle estimate $\widehat{\boldsymbol{\gamma}}^{\mathrm{I}}_{\boldsymbol{x}}(m)$, (b) oracle estimate $\widehat{\boldsymbol{\gamma}}^{\mathrm{II}}_{\boldsymbol{x}}(m)$ and (c) blind estimate $\widehat{\boldsymbol{\gamma}}^{\mathrm{ML}}_{\boldsymbol{x}}(m)$. The plots show the average PESQ and segmental SNR improvements.

formance of the MFMPDR filter using the blind ML estimate for the normalized speech correlation vector in Fig. 4.3(c) with the related oracle estimate $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{II}}(m)$ in Fig. 4.3(b), it can be observed that the performance is substantially degraded. For instance, for an input SNR of 5 dB, ΔPESQ is reduced by 0.5 MOS. These results indicate that, as expected, estimation errors in the normalized speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$ lead to a strongly reduced speech enhancement performance for the MFMPDR filter. Hence, further work is required to either improve the accuracy of blind estimators for the highly time-varying normalized speech correlation vector, or to improve the robustness of the MFMPDR filter against estimation errors.

4.4 Summary

In this chapter, we investigated the speech enhancement performance of the practically feasible MFMPDR filter with two oracle versions of the MFMVDR filter for different oracle and blind estimates of the normalized speech correlation vector. The simulation results show that, as expected, the oracle MFMVDR filter using a quasi-perfect estimate of the undesired correlation matrix achieves the best results, but that even small estimation errors in the normalized speech correlation vector decrease the performance. The results also show that the performance of the MFMPDR filter is very close to the oracle (quasi-perfect) MFMVDR filter when using oracle estimates of the normalized speech correlation vector. When using the state-of-the-art blind ML estimator for the normalized speech correlation vector, the performance of the MFMPDR is strongly reduced due to large estimations errors in the normalized speech correlation vector. Hence, in the remainder of this thesis, we will focus on estimating the normalized speech correlation vector to improve the performance of the MFMPDR filter. In Chapter 5, we will investigate the potential of using concepts proposed for robust MPDR beamforming in the context of single-microphone multi-frame speech enhancement, by estimating the normalized speech correlation vector as the vector maximizing the total signal output PSD within a spherical uncertainty set. In Chapter 6, based on a low-rank speech model we will propose different matrix-based methods to estimate the normalized speech correlation vector.

5

ROBUST CONSTRAINED MFMPDR FILTERS BASED ON SPHERICAL UNCERTAINTY SET

As shown in Chapter 4, accurately estimating the highly time-varying normalized speech correlation vector is crucial, since even small estimation errors may degrade the speech enhancement performance of the multi-frame minimum power distortionless response (MFMPDR) filter. To improve the robustness of the MFMPDR filter against estimation errors in the normalized speech correlation vector, in this chapter we propose to estimate the normalized speech correlation vector within a spherical uncertainty set.

In the area of multi-channel processing, i.e., beamforming, several approaches were proposed to increase the robustness of MPDR beamformers (also called robust Capon beamformers) against estimation errors in the steering vector. One of the most popular approaches is diagonal loading, imposing a quadratic inequality constraint on the filter vector [101]. However, since diagonal loading does not explicitly address uncertainty of the steering vector, several other approaches were proposed, e.g., by imposing (equality and/or inequality) constraints on the so-called mismatch vector, i.e., the difference between the steering vector and the presumed steering vector [102–110]. The robust MPDR beamformers in [104, 106] estimate the steering vector as the vector maximizing the total signal output PSD of the MPDR beamformer within a spherical uncertainty set. Inspired by these robust multi-channel approaches, in this chapter we investigate the potential of estimating the normal-

This chapter is partly based on:

[132] D. Fischer and S. Doclo, "Robust constrained MFMVDR Filtering for single-microphone speech enhancement," in *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, Sep. 2018, pp. 41-45.

[133] D. Fischer and S. Doclo, "Evaluation of robust constrained MFMVDR Filtering for single-channel speech enhancement," in *Proc. ITG Conference on Speech Commun.*, Oldenburg, Germany, Oct. 2018, pp. 156-160.

[134] D. Fischer and S. Doclo, "Robust constrained MFMVDR filters for single-channel speech enhancement based on spherical uncertainty set," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 2020, manuscript submitted for publication.

ized speech correlation vector as the vector maximizing the total signal output PSD of the MFMPDR filter within a spherical uncertainty set. This corresponds to imposing a quadratic inequality constraint on the mismatch vector, i.e., the difference between the speech correlation vector and the presumed normalized speech correlation vector, e.g., the maximum-likelihood (ML) estimate in [15]. We propose two constrained MFMPDR filters. The *singly-constrained (SC) MFMPDR filter* only considers the quadratic inequality constraint on the mismatch vector to estimate the (non-normalized) speech correlation vector and applies normalization afterwards. On the other hand, the *doubly-constrained (DC) MFMPDR filter* integrates the (linear) normalization constraint into the optimization problem and directly estimates the normalized speech correlation vector by solving an optimization problem with two constraints. Since oracle simulations using several speech and noise signals at different SNRs show that the norm of the mismatch vector decreases with increasing SNR, we propose to train a non-linear mapping function that depends on the a-priori SNR to set the upper bound of the spherical uncertainty set for each time-frequency point.

In Section 5.1, the proposed constrained MFMPDR filters based on a spherical uncertainty set as well as the mapping function to set the upper bound of the spherical uncertainty set are presented. In Section 5.2, an instrumental and perceptual performance comparison for different speech signals, noise types and SNRs is provided between both constrained MFMPDR filters, the state-of-the-art ML-MFMPDR filter and the single-frame Wiener gain (WG) as reference speech enhancement algorithm. The results show that the proposed constrained MFMPDR filters result in a more conservative noise reduction performance with a more natural speech quality and less noise distortion than the ML-MFMPDR filter, where the DC-MFMPDR filter is preferred in terms of overall quality.

5.1 Constrained MFMPDR Filters

Aiming at improving the robustness against estimation errors in the normalized speech correlation vector, in this section we propose two constrained MFMPDR filters. Inspired by the robust MPDR beamformers in [104, 106], we propose to estimate the normalized speech correlation vector as the vector maximizing the total signal output PSD of the MFMPDR filter within a spherical uncertainty set. Section 5.1.1 presents the SC-MFMPDR filter, which only impose a quadratic inequality constraint on the mismatch vector with respect to the presumed normalized speech correlation vector, whereas Section 5.1.2 presents the DC-MFMPDR filter, which jointly considers the quadratic inequality constraint as well as a (linear) normalization constraint. Section 5.1.3 discusses a trained non-linear mapping function to set the upper bound of the spherical uncertainty set for each time-frequency point. For conciseness, the frequency-bin index k and the time-frame index m will be omitted in this chapter if not required. However, it should be realized that all calculations are performed for each time-frequency point.

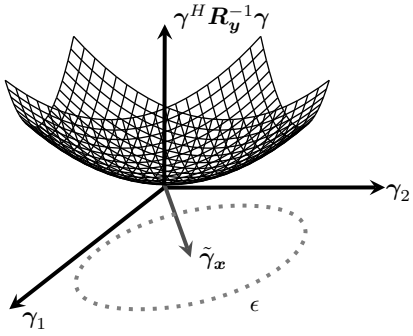


Fig. 5.1: Quadratic cost function in (5.3) with exemplary presumed normalized speech correlation vector $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}$ and bound ϵ.

5.1.1 *Singly-Constrained MFMPDR Filter*

Given a presumed normalized speech correlation vector $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}$, e.g., the ML estimate $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{ML}}$ in (4.12), the mismatch vector with respect to the (unknown) normalized speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{x}}$ is defined as $\boldsymbol{\delta}_{\boldsymbol{x}} = \boldsymbol{\gamma}_{\boldsymbol{x}} - \tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}$, with $\epsilon_{\boldsymbol{x}} = \|\boldsymbol{\gamma}_{\boldsymbol{x}} - \tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}\|_2^2$. We now define the spherical uncertainty set comprising all vectors whose squared distance to the presumed normalized speech correlation vector $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}$ is smaller than or equal to a bound $\epsilon \geq 0$, i.e.,

$$\mathbb{U} = \left\{ \boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}} + \boldsymbol{\delta} \mid \|\boldsymbol{\delta}\|_2^2 \leq \epsilon \right\}. \tag{5.1}$$

Similarly to the robust MPDR beamformer in [104], we propose to compute the (non-normalized) speech correlation vector for the SC-MFMPDR filter as the vector maximizing the total signal output PSD of the MFMPDR filter in (3.24) within the spherical uncertainty set in (5.1), i.e.,

$$\breve{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SC}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmax}} \quad \frac{1}{\boldsymbol{\gamma}^H \boldsymbol{R}_{\boldsymbol{y}}^{-1} \boldsymbol{\gamma}}, \quad \text{s.t. } \|\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}\|_2^2 \leq \epsilon\,, \tag{5.2}$$

which is equivalent to

$$\boxed{\breve{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SC}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \quad \boldsymbol{\gamma}^H \boldsymbol{R}_{\boldsymbol{y}}^{-1} \boldsymbol{\gamma}, \quad \text{s.t. } \|\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}\|_2^2 \leq \epsilon} \tag{5.3}$$

For an exemplary noisy speech correlation matrix $\boldsymbol{R}_{\boldsymbol{y}}$ and $L = 2$, Fig. 5.1 visualizes the quadratic cost function $\boldsymbol{\gamma}^H \boldsymbol{R}_{\boldsymbol{y}}^{-1} \boldsymbol{\gamma}^H$ in (5.3), together with an exemplary presumed normalized speech correlation vector $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}$ and bound ϵ. Obviously, the bound ϵ in (5.3) plays an important role and should be chosen in accordance with the accuracy of the presumed normalized speech correlation vector $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}$, i.e., if $\|\boldsymbol{\gamma}_{\boldsymbol{x}} - \tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}\|_2^2$ is large, then ϵ should be large, whereas if $\|\boldsymbol{\gamma}_{\boldsymbol{x}} - \tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}\|_2^2$ is small, then ϵ should be small.

The minimum of the quadratic cost function $\boldsymbol{\gamma}^H \boldsymbol{R}_{\boldsymbol{y}}^{-1}\boldsymbol{\gamma}$ is given by $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}} = 0$, which is obviously undesired. In order to avoid this solution, the bound ϵ should be chosen such that

$$\epsilon < \|\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}\|_2^2 \,. \tag{5.4}$$

Under this condition and considering the convex nature of the quadratic cost function in (5.3), the inequality constraint in (5.3) can be replaced by an equality constraint, i.e.,

$$\breve{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SC}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \quad \boldsymbol{\gamma}^H \boldsymbol{R}_{\boldsymbol{y}}^{-1}\boldsymbol{\gamma}, \quad \text{s.t. } \|\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}\|_2^2 = \epsilon \,. \tag{5.5}$$

This constrained optimization problem can be solved using the method of Lagrange multipliers [144]. The Lagrangian function is given by

$$\mathcal{L}_{\mathrm{SC}}\left[\boldsymbol{\gamma}, \mu\right] = \boldsymbol{\gamma}^H \boldsymbol{R}_{\boldsymbol{y}}^{-1}\boldsymbol{\gamma} + \mu\left(\|\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}\|_2^2 - \epsilon\right) , \tag{5.6}$$

with μ the Lagrange multiplier. Setting the gradient of $\mathcal{L}_{\mathrm{SC}}\left[\boldsymbol{\gamma}, \mu\right]$ with respect to $\boldsymbol{\gamma}$

$$\nabla_{\boldsymbol{\gamma}} \mathcal{L}_{\mathrm{SC}}\left[\boldsymbol{\gamma}, \mu\right] = 2\boldsymbol{R}_{\boldsymbol{y}}^{-1}\boldsymbol{\gamma} + 2\mu\left(\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}\right) \tag{5.7}$$

equal to zero, yields

$$\boldsymbol{\gamma} = \mu\left(\boldsymbol{R}_{\boldsymbol{y}}^{-1} + \mu\boldsymbol{I}_{\mathrm{L}}\right)^{-1}\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}} \,. \tag{5.8}$$

Applying the matrix inversion lemma [144], we obtain the SC speech correlation vector $\breve{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SC}}(\mu)$ as

$$\boxed{\breve{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SC}}(\mu) = \tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}} - \left(\mu\boldsymbol{R}_{\boldsymbol{y}} + \boldsymbol{I}_{\mathrm{L}}\right)^{-1}\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}} \tag{5.9}$$

Setting the partial derivative of $\mathcal{L}_{\mathrm{SC}}\left[\boldsymbol{\gamma}, \mu\right]$ in (5.6) with respect to μ equal to zero and substituting (5.9) results in

$$g_{\mathrm{SC}}(\mu) = \frac{\partial \mathcal{L}_{\mathrm{SC}}\left[\boldsymbol{\gamma}, \mu\right]}{\partial \mu} = \left\| \left(\mu\boldsymbol{R}_{\boldsymbol{y}} + \boldsymbol{I}_{\mathrm{L}}\right)^{-1}\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}} \right\|_2^2 - \epsilon = 0, \tag{5.10}$$

which should be solved for the Lagrange multiplier μ.

Let the eigenvalue decomposition (EVD) of the noisy speech correlation matrix be given by

$$\boldsymbol{R}_{\boldsymbol{y}} = \boldsymbol{U}\boldsymbol{\Upsilon}\boldsymbol{U}^H \,, \tag{5.11}$$

where the columns of $\boldsymbol{U}$ contain the orthogonal eigenvectors and the diagonal elements of the diagonal matrix $\boldsymbol{\Upsilon}$ are the corresponding eigenvalues, with $\upsilon_0 \geq \upsilon_1 \geq \ldots \geq \upsilon_{L-1}$. By defining

$$\boldsymbol{z}_{\tilde{\boldsymbol{\gamma}}} = \boldsymbol{U}^H\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}} \,, \tag{5.12}$$

and using (5.11) and (5.12) in (5.10), we obtain

$$\boxed{g_{\mathrm{SC}}(\mu) = \sum_{l=0}^{L-1} \frac{|\boldsymbol{z}_{\tilde{\gamma},l}|^2}{(1+\mu\upsilon_l)^2} = \epsilon} \tag{5.13}$$

with $\boldsymbol{z}_{\tilde{\gamma},l}$ denoting the l-th element of $\boldsymbol{z}_{\tilde{\gamma}}$. This non-linear equation in the Lagrange multiplier μ can be solved, e.g., using Newton's method [144]. The solution is then used in (5.9), yielding the SC speech correlation vector $\breve{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SC}}$. Since the normalization constraint in (2.13) (i.e., the first element of $\boldsymbol{\gamma}_{\boldsymbol{x}}$ is equal to 1) is typically not satisfied, resulting in a scaling inaccuracy, normalization is performed, i.e.,

$$\boxed{\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SC}} = \frac{\breve{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SC}}}{\boldsymbol{e}^T\breve{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SC}}}} \tag{5.14}$$

However, there is no guarantee that the *normalized SC speech correlation vector* $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SC}}$ in (5.14) satisfies the quadratic inequality constraint in (5.1), i.e., lies within the spherical uncertainty set. Using the normalized SC speech correlation vector in (3.23) results in the SC-MFMPDR filter.

5.1.2 *Doubly-Constrained MFMPDR Filter*

Since it is not guaranteed that the normalized SC speech correlation vector satisfies both the quadratic inequality constraint in (5.1), as well as the (linear) normalization constraint in (2.13), in this section we propose to estimate the normalized speech correlation vector as the vector maximizing the total signal output PSD of the MFMPDR filter while satisfying both constraints, i.e.,

$$\boxed{\begin{aligned}\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{DC}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \quad & \boldsymbol{\gamma}^H\boldsymbol{R}_{\boldsymbol{y}}^{-1}\boldsymbol{\gamma}, \quad \text{s.t. } \|\boldsymbol{\gamma}-\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}\|_2^2 \leq \epsilon, \\ & \boldsymbol{e}^T\boldsymbol{\gamma} = 1\end{aligned}} \tag{5.15}$$

This doubly-constrained optimization problem can be transformed into a singly-constrained optimization problem by decomposing the L-dimensional vector $\boldsymbol{\gamma}$ as

$$\boldsymbol{\gamma} = \begin{bmatrix} 1 \\ -\boldsymbol{d} \end{bmatrix} = \boldsymbol{e} - \boldsymbol{E}\boldsymbol{d}\,, \tag{5.16}$$

with $\boldsymbol{d}$ an $(L-1)$-dimensional vector and the $L \times (L-1)$-dimensional matrix $\boldsymbol{E}$ defined as

$$\boldsymbol{E} = \begin{bmatrix} \mathbf{0}_{1\times \mathrm{L\text{-}1}} \\ \boldsymbol{I}_{\mathrm{L\text{-}1}} \end{bmatrix}. \tag{5.17}$$

Similarly, the L-dimensional vectors $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{DC}}$ and $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}$ can be decomposed as

$$\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{DC}} = \begin{bmatrix} 1 \\ -\widehat{\boldsymbol{d}}_{\boldsymbol{x}}^{\mathrm{DC}} \end{bmatrix} = \boldsymbol{e} - \boldsymbol{E}\widehat{\boldsymbol{d}}_{\boldsymbol{x}}^{\mathrm{DC}}, \tag{5.18}$$

$$\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}} \quad = \begin{bmatrix} 1 \\ -\tilde{\boldsymbol{d}}_{\boldsymbol{x}} \end{bmatrix} = \boldsymbol{e} - \boldsymbol{E}\tilde{\boldsymbol{d}}_{\boldsymbol{x}}\,. \tag{5.19}$$

Instead of estimating $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{DC}}$, it is hence sufficient to estimate $\widehat{\boldsymbol{d}}_{\boldsymbol{x}}^{\mathrm{DC}}$, which can be done by substituting (5.16), (5.18) and (5.19) into (5.15), i.e.,

$$\widehat{\boldsymbol{d}}_{\boldsymbol{x}}^{\mathrm{DC}} = \underset{\boldsymbol{d}}{\operatorname{argmin}}\ (\boldsymbol{e} - \boldsymbol{E}\boldsymbol{d})^H \boldsymbol{R}_{\boldsymbol{y}}^{-1} (\boldsymbol{e} - \boldsymbol{E}\boldsymbol{d}), \quad \text{s.t. } \left\| \boldsymbol{d}_{\boldsymbol{x}} - \tilde{\boldsymbol{d}}_{\boldsymbol{x}} \right\|_2^2 \leq \epsilon, \tag{5.20}$$

transforming the doubly-constrained optimization problem in (5.15) into a singly-constrained optimization problem.

Based on the definition of the normalized noisy speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{y}}$ in (2.19) and using the decomposition

$$\boldsymbol{\gamma}_{\boldsymbol{y}} = \begin{bmatrix} 1 \\ -\boldsymbol{d}_{\boldsymbol{y}} \end{bmatrix}, \tag{5.21}$$

the $L \times L$-dimensional noisy speech correlation matrix $\boldsymbol{R}_{\boldsymbol{y}}$ can be decomposed as

$$\boldsymbol{R}_{\boldsymbol{y}} = \left[\begin{array}{c|c} \phi_Y & -\phi_Y \boldsymbol{d}_{\boldsymbol{y}}^H \\ \hline -\phi_Y \boldsymbol{d}_{\boldsymbol{y}} & \boldsymbol{D}_{\boldsymbol{y}} \end{array}\right], \tag{5.22}$$

with $\boldsymbol{D}_{\boldsymbol{y}}$ an $(L-1)\times(L-1)$-dimensional matrix. Using blockwise inversion, the matrix $\boldsymbol{R}_{\boldsymbol{y}}^{-1}$ is equal to

$$\boldsymbol{R}_{\boldsymbol{y}}^{-1} = \left[\begin{array}{c|c} a_Y & \boldsymbol{b}_{\boldsymbol{y}}^H \\ \hline \boldsymbol{b}_{\boldsymbol{y}} & \boldsymbol{S}_{\boldsymbol{y}}^{-1} \end{array}\right], \tag{5.23}$$

with $\boldsymbol{S}_{\boldsymbol{y}}$ the $(L-1) \times (L-1)$-dimensional Schur complement [144], i.e.,

$$\boldsymbol{S}_{\boldsymbol{y}} = \boldsymbol{D}_{\boldsymbol{y}} - \phi_Y \boldsymbol{d}_{\boldsymbol{y}} \boldsymbol{d}_{\boldsymbol{y}}^H, \tag{5.24}$$

and

$$a_Y = \phi_Y^{-1} + \boldsymbol{d}_{\boldsymbol{y}}^H \boldsymbol{S}_{\boldsymbol{y}}^{-1} \boldsymbol{d}_{\boldsymbol{y}}, \tag{5.25}$$

$$\boldsymbol{b}_{\boldsymbol{y}} = \boldsymbol{S}_{\boldsymbol{y}}^{-1} \boldsymbol{d}_{\boldsymbol{y}}. \tag{5.26}$$

Using (5.23), the optimization problem in (5.20) can be reformulated as

$$\boxed{\widehat{\boldsymbol{d}}_{\boldsymbol{x}}^{\mathrm{DC}} = \underset{\boldsymbol{d}}{\operatorname{argmin}}\ \ a_Y - \boldsymbol{b}_{\boldsymbol{y}}^H \boldsymbol{d} - \boldsymbol{d}^H \boldsymbol{b}_{\boldsymbol{y}} + \boldsymbol{d}^H \boldsymbol{S}_{\boldsymbol{y}}^{-1} \boldsymbol{d}, \quad \text{s.t. } \left\| \boldsymbol{d} - \tilde{\boldsymbol{d}}_{\boldsymbol{x}} \right\|_2^2 \leq \epsilon} \tag{5.27}$$

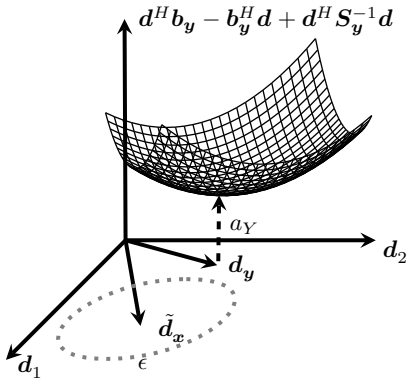


Fig. 5.2: Quadratic cost function in (5.27) with exemplary vector $\tilde{\boldsymbol{d}}_{\boldsymbol{x}}$ (part of $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}$), vector $\boldsymbol{d}_{\boldsymbol{y}}$ (part of $\boldsymbol{\gamma}_{\boldsymbol{y}}$) and bound ϵ.

which is similar but obviously not the same as the optimization problem in (5.3). For an exemplary noisy speech correlation matrix $\boldsymbol{R}_{\boldsymbol{y}}$ and $L = 3$, Fig. 5.2 visualizes the quadratic cost function $a_Y - \boldsymbol{b}_{\boldsymbol{y}}^H\boldsymbol{d} - \boldsymbol{d}^H\boldsymbol{b}_{\boldsymbol{y}} + \boldsymbol{d}^H\boldsymbol{S}_{\boldsymbol{y}}^{-1}\boldsymbol{d}$, together with an exemplary presumed vector $\tilde{\boldsymbol{d}}_{\boldsymbol{x}}$ (part of $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}$) and bound ϵ. In comparison to Fig. 5.1, the quadratic cost function is shifted upwards by the scalar a_Y and shaped by the term $-\boldsymbol{b}_{\boldsymbol{y}}^H\boldsymbol{d} - \boldsymbol{d}^H\boldsymbol{b}_{\boldsymbol{y}}$. Similarly as in Section 5.1.1, for the optimization problem in (5.27) the bound ϵ plays an important role and should be chosen in accordance with the accuracy of the presumed vector $\tilde{\boldsymbol{d}}_{\boldsymbol{x}}$.

The minimum of the quadratic cost function $a_Y - \boldsymbol{b}_{\boldsymbol{y}}^H\boldsymbol{d} - \boldsymbol{d}^H\boldsymbol{b}_{\boldsymbol{y}} + \boldsymbol{d}^H\boldsymbol{S}_{\boldsymbol{y}}^{-1}\boldsymbol{d}$ is given by $\widehat{\boldsymbol{d}}_{\boldsymbol{x}} = \boldsymbol{S}_{\boldsymbol{y}}\boldsymbol{b}_{\boldsymbol{y}}$, which using (5.26) is equal to $\boldsymbol{d}_{\boldsymbol{y}}$. Using this solution, or consequently $\boldsymbol{\gamma}_{\boldsymbol{y}}$, in (3.23) results in the MFMPDR filter being equal to the selection vector $\boldsymbol{e}$, which is obviously undesired. In order to avoid this solution, the bound ϵ should be chosen such that

$$\epsilon < \left\| \tilde{\boldsymbol{d}}_{\boldsymbol{x}} - \boldsymbol{d}_{\boldsymbol{y}} \right\|_2^2 . \tag{5.28}$$

Under this condition and considering the convex nature of the quadratic cost function in (5.27), the inequality constraint in (5.27) can be replaced by an equality constraint, i.e.,

$$\widehat{\boldsymbol{d}}_{\boldsymbol{x}}^{\mathrm{DC}} = \underset{\boldsymbol{d}}{\operatorname{argmin}} \quad a_Y - \boldsymbol{b}_{\boldsymbol{y}}^H\boldsymbol{d} - \boldsymbol{d}^H\boldsymbol{b}_{\boldsymbol{y}} + \boldsymbol{d}^H\boldsymbol{S}_{\boldsymbol{y}}^{-1}\boldsymbol{d}, \quad \text{s.t.} \ \left\| \boldsymbol{d} - \tilde{\boldsymbol{d}}_{\boldsymbol{x}} \right\|_2^2 = \epsilon. \tag{5.29}$$

Similarly to (5.5), this constrained optimization problem can be solved using the method of Lagrange multipliers [144], where the Lagrangian function is now given by

$$\mathcal{L}_{\mathrm{DC}}\left[\boldsymbol{d}, \mu\right] = a_Y - \boldsymbol{b}_{\boldsymbol{y}}^H\boldsymbol{d} - \boldsymbol{d}^H\boldsymbol{b}_{\boldsymbol{y}} + \boldsymbol{d}^H\boldsymbol{S}_{\boldsymbol{y}}^{-1}\boldsymbol{d} + \mu\left(\left\| \boldsymbol{d} - \tilde{\boldsymbol{d}}_{\boldsymbol{x}} \right\|_2^2 - \epsilon\right), \tag{5.30}$$

with μ the Lagrange multiplier. Setting the gradient of $\mathcal{L}_{\mathrm{DC}}\left[\boldsymbol{d},\mu\right]$ with respect to $\boldsymbol{d}$

$$\nabla_{\boldsymbol{d}}\mathcal{L}_{\mathrm{DC}}\left[\boldsymbol{d},\mu\right] = 2\boldsymbol{S}_{\boldsymbol{y}}^{-1}\boldsymbol{d} - 2\boldsymbol{b}_{\boldsymbol{y}} + 2\mu(\boldsymbol{d} - \tilde{\boldsymbol{d}}_{\boldsymbol{x}}), \tag{5.31}$$

equal to zero, yields

$$\boldsymbol{d} = \left(\boldsymbol{S}_{\boldsymbol{y}}^{-1} + \mu\boldsymbol{I}_{\text{L-1}}\right)^{-1}\left(\boldsymbol{b}_{\boldsymbol{y}} + \mu\tilde{\boldsymbol{d}}_{\boldsymbol{x}}\right). \tag{5.32}$$

Applying the matrix inversion lemma [144], we obtain the vector $\widehat{\boldsymbol{d}}_{\boldsymbol{x}}^{\mathrm{DC}}(\mu)$ as

$$\boxed{\widehat{\boldsymbol{d}}_{\boldsymbol{x}}^{\mathrm{DC}}(\mu) = \left(\boldsymbol{I}_{\text{L-1}} - \left(\mu\boldsymbol{S}_{\boldsymbol{y}} + \boldsymbol{I}_{\text{L-1}}\right)^{-1}\right)\left(\frac{1}{\mu}\boldsymbol{b}_{\boldsymbol{y}} + \tilde{\boldsymbol{d}}_{\boldsymbol{x}}\right)} \tag{5.33}$$

Setting the partial derivative of $\mathcal{L}_{\mathrm{DC}}\left[\boldsymbol{d},\mu\right]$ in (5.30) with respect to μ equal to zero and substituting (5.33) results in

$$g_{\mathrm{DC}}(\mu) = \frac{\partial\mathcal{L}_{\mathrm{DC}}\left[\boldsymbol{d},\mu\right]}{\partial\mu} = \left\|\left(\mu\boldsymbol{S}_{\boldsymbol{y}} + \boldsymbol{I}_{\text{L-1}}\right)^{-1}\left(\frac{1}{\mu}\boldsymbol{b}_{\boldsymbol{y}} + \tilde{\boldsymbol{d}}_{\boldsymbol{x}}\right) - \frac{1}{\mu}\boldsymbol{b}_{\boldsymbol{y}}\right\|_2^2 - \epsilon = 0\,, \tag{5.34}$$

which should be solved for the Lagrange multiplier μ.
Let the EVD of the Schur complement $\boldsymbol{S}_{\boldsymbol{y}}$ in (5.24) be given by

$$\boldsymbol{S}_{\boldsymbol{y}} = \boldsymbol{\Psi}\boldsymbol{\Theta}\boldsymbol{\Psi}^{H}, \tag{5.35}$$

where the columns of $\boldsymbol{\Psi}$ contain the orthogonal eigenvectors and the diagonal elements of the diagonal matrix $\boldsymbol{\Theta}$ are the corresponding eigenvalues, with $\varphi_0 \geq \varphi_1 \geq \ldots \geq \varphi_{L-2}$. By defining

$$\boldsymbol{z}_{\tilde{\boldsymbol{d}}} = \boldsymbol{\Psi}^{H}\tilde{\boldsymbol{d}}_{\boldsymbol{x}}, \tag{5.36}$$

$$\boldsymbol{z}_{\boldsymbol{b}} = \boldsymbol{\Psi}^{H}\boldsymbol{b}_{\boldsymbol{y}}, \tag{5.37}$$

and using (5.35), (5.36) and (5.37) in (5.34), we obtain

$$\boxed{g_{\mathrm{DC}}(\mu) = \sum_{l=0}^{L-2}\frac{\left|\boldsymbol{z}_{\boldsymbol{b},l}\varphi_l - \boldsymbol{z}_{\tilde{\boldsymbol{d}},l}\right|^2}{(1+\mu\varphi_l)^2} = \epsilon} \tag{5.38}$$

with $\boldsymbol{z}_{\tilde{\boldsymbol{d}},l}$ and $\boldsymbol{z}_{\boldsymbol{b},l}$ denoting the l-th element of $\boldsymbol{z}_{\tilde{\boldsymbol{d}}}$ and $\boldsymbol{z}_{\boldsymbol{b}}$, respectively. This non-linear equation in the Lagrange multiplier μ can be solved similarly to (5.13), e.g., using Newton's method [144]. The solution is then used in (5.33), to obtain $\widehat{\boldsymbol{d}}_{\boldsymbol{x}}^{\mathrm{DC}}$, and subsequently in (5.18), yielding the *normalized DC speech correlation vector* $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{DC}}$. Using $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{DC}}$ in (3.23) results in the DC-MFMPDR filter. It should be noted that due to the EVD in (5.11) and (5.35) and solving the non-linear equations in (5.13) and (5.38), the computational complexity is obviously larger for the constrained MFMPDR filters than for the ML-MFMPDR filter, where the computational complexity for the SC-MFMPDR filter and the DC-MFMPDR filter is similar.

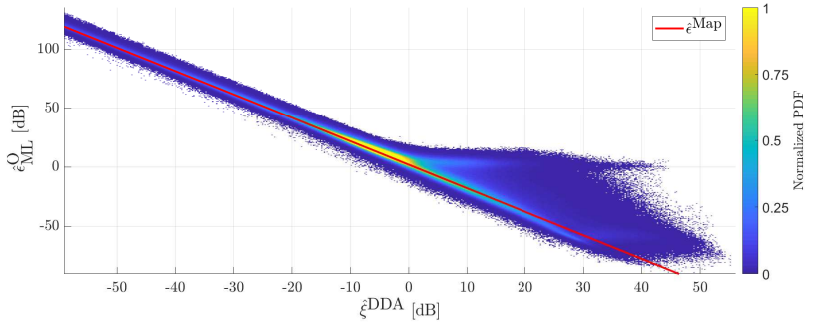


Fig. 5.3: Normalized joint PDF of the oracle bound $\widehat{\epsilon}^{\text{O}}_{\text{ML}}$ and the a-priori SNR estimate $\widehat{\xi}^{\text{DDA}}$ with the mapping function $\widehat{\epsilon}^{\text{Map}}$ in red.

5.1.3 *Bound of the Spherical Uncertainty Set*

As already mentioned, the bound ϵ of the spherical uncertainty set in (5.1) plays a crucial role for both constrained optimization problems in that it should be chosen in accordance with the accuracy of the presumed normalized speech correlation vector $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}$. In this chapter, we define the oracle normalized speech correlation vector $\widehat{\boldsymbol{\gamma}}^{\text{II}}_{\boldsymbol{x}}$ in (4.7) as the (optimal) solution aiming to be found for the constrained optimization problems. In order to ensure that $\widehat{\boldsymbol{\gamma}}^{\text{II}}_{\boldsymbol{x}}$ lies within the spherical uncertainty set (and hence can be found as a solution of the constrained optimization problems), the bound ϵ should be larger than or equal to the oracle bound $\widehat{\epsilon}^{\text{O}} = \left\|\widehat{\boldsymbol{\gamma}}^{\text{II}}_{\boldsymbol{x}} - \tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}\right\|_2^2$. In this chapter, we will use the ML estimate $\widehat{\boldsymbol{\gamma}}^{\text{ML}}_{\boldsymbol{x}}$ in (4.12) as the presumed normalized speech correlation vector $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}$. Since we showed in Chapter 4 that the accuracy of the ML estimate strongly depends on the a-priori SNR estimate $\widehat{\xi}$ (cf. Section 4.3.3), we propose to train a (non-linear) mapping function between the oracle bound, i.e.,

$$\widehat{\epsilon}^{\text{O}}_{\text{ML}} = \left\|\widehat{\boldsymbol{\gamma}}^{\text{II}}_{\boldsymbol{x}} - \widehat{\boldsymbol{\gamma}}^{\text{ML}}_{\boldsymbol{x}}\right\|_2^2 , \tag{5.39}$$

and an estimate of the a-priori SNR. To estimate the a-priori SNR, we use the DDA estimate $\widehat{\xi}^{\text{DDA}}$ in (3.9) with the SPP-based noise PSD estimate $\widehat{\phi}^{\text{SPP}}_N$ in (3.13). For a wide range of speech and noise signals (30 TIMIT sentences [146], speech-shaped noise, two traffic and babble noise signals [145]) and an SNR range of 0 dB to 15 dB in 5 dB steps, Fig. 5.3 shows the normalized joint probability density function (PDF) of the oracle bound $\widehat{\epsilon}^{\text{O}}_{\text{ML}}$ and the a-priori SNR estimate $\widehat{\xi}^{\text{DDA}}$ (for all time-frequency points). It can be clearly observed that the oracle bound decreases with increasing a-priori SNR. Fitting a linear function (in log-log scale) to the maximum value of the normalized PDF for each a-priori SNR estimate $\widehat{\xi}^{\text{DDA}}$ yields the mapping function

$$\boxed{\widehat{\epsilon}^{\text{Map}}(\widehat{\xi}^{\text{DDA}}_{\text{dB}}) = 10^{\left(-1.983\widehat{\xi}^{\text{DDA}}_{\text{dB}}+2\right)/10}} \tag{5.40}$$

with $\widehat{\xi}^{\text{DDA}}_{\text{dB}} = 10\log_{10}(\widehat{\xi}^{\text{DDA}})$. This mapping function is shown in red in Fig. 5.3.

5.2 Simulation Results

In this section, we analyze the performance of the proposed constrained MFMPDR filters based on a spherical uncertainty set. In Section 5.2.1 we discuss the audio material and the algorithmic settings. In Section 5.2.2 we compare the accuracy of the proposed normalized SC and DC speech correlation vector estimates with the ML estimate. For different speech signals, noise types and SNRs, in Section 5.2.3 and 5.2.4 we compare the instrumental and perceptual speech quality of the proposed SC-MFMPDR and DC-MFMPDR filters with the state-of-the-art ML-MFMPDR filter and the single-frame Wiener gain (WG) as reference speech enhancement algorithm.

5.2.1 *Audio Material and Algorithmic Settings*

For the evaluation, we used 60 sentences from the TIMIT database [146], spoken by different speakers (10 male, 10 female) as speech signals, sampled at a sampling frequency of 16 kHz. As noise signals, we used speech-shaped noise, traffic noise, babble noise and factory noise signals taken from the NOISEX-92 database [145]. The considered SNRs ranged from -5 dB to 20 dB in 5 dB steps. We made sure that the evaluation data differs from the data used for training the mapping function in Section 5.1.3.

For the MFMPDR filter similarly as in Section 4.3, in order to achieve a high speech interframe correlation, we used a highly temporally resolved STFT framework with a frame length of 4 ms ($T = K = 64$) and an overlap of 75 %, resulting in a frame shift of 1 ms ($R = 16$). As the STFT analysis window $w_a(t)$ and the synthesis window $w_s(t)$, we used a square-root Hann window. Similarly as in Section 4.3 and [15], the number of consecutive time-frames is set to $L = 18$, resulting in 21 ms of data used in each filtering operation.

The noisy speech correlation matrix $\boldsymbol{R}_{\boldsymbol{y}}$ and the oracle noise correlation matrix estimate (required for the oracle estimate $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{II}}$ in (4.7)) are estimated using recursive smoothing as in (4.3) and (4.2), with smoothing parameters experimentally set to $\alpha_y = \alpha_n = 0.90$, corresponding to a smoothing window of 10 ms. To avoid numerical problems for the MFMPDR filter, we performed diagonal loading as in (4.17) before computing the inverse of the noisy speech correlation matrix, with a scaling parameter of $\kappa = 0.001$, similarly as in Section 4.3. The a-priori SNR estimate $\widehat{\xi}$ required for the ML estimate $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{ML}}$ in (4.12) and the bound $\widehat{\epsilon}^{\mathrm{Map}}$ in (5.40), is computed using the DDA estimate $\widehat{\xi}^{\mathrm{DDA}}$ in (3.9) with a weighting parameter of $\alpha_{\mathrm{DDA}} = 0.70$ and the SPP-based noise PSD estimate $\widehat{\phi}_N^{\mathrm{SPP}}$ in (3.13).

Although the main objective is to compare the performance of the proposed constrained MFMPDR filters with the ML-MFMPDR filter, we will also consider the single-frame WG as a reference single-microphone speech enhancement algorithm. For a fair comparison, the WG is implemented using an equivalent frame length of 21 ms and an overlap of 50 %. The a-priori SNR for the WG is also computed using the DDA estimate $\widehat{\xi}^{\mathrm{DDA}}$ in (3.9) with a weighting parameter of $\alpha_{\mathrm{DDA}} = 0.98$ and the SPP-based noise PSD estimate $\widehat{\phi}_N^{\mathrm{SPP}}$ in (3.13). To reduce the amount of speech

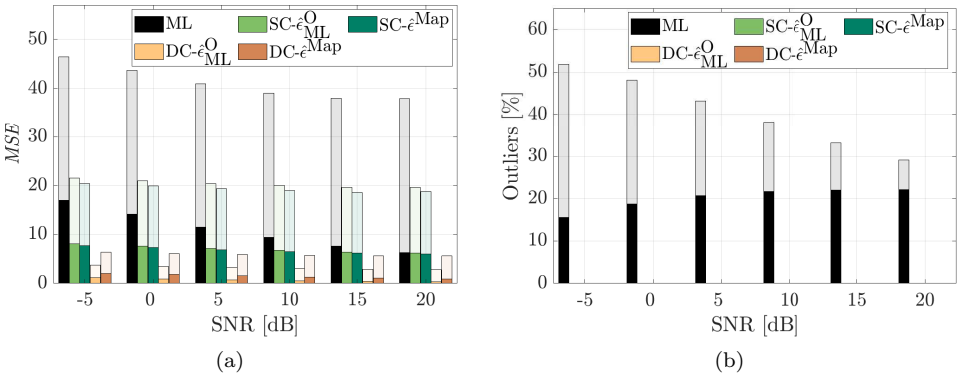


Fig. 5.4: Average (a) *MSE* and (b) percentage of outliers for the normalized ML, SC and DC speech correlation vector estimates using the oracle bound $\widehat{\epsilon}^{\mathrm{O}}_{\mathrm{ML}}$ and the mapping function $\widehat{\epsilon}^{\mathrm{Map}}$ for different SNRs. The lower and upper parts of the bars represent the performance in speech-and-noise and noise-only time-frequency points, respectively.

distortion and to mask artifacts in the background noise, the lower limit for the WG in (3.5) is set to $G_{\min} = -17\,\mathrm{dB}$.

5.2.2 *Accuracy of the Normalized Speech Correlation Vector Estimates*

In this section, we compare the accuracy of the proposed normalized SC and DC speech correlation vector estimates $\widehat{\boldsymbol{\gamma}}^{\mathrm{SC}}_{\boldsymbol{x}}$ (5.14) and $\widehat{\boldsymbol{\gamma}}^{\mathrm{DC}}_{\boldsymbol{x}}$ (5.18) with the ML estimate $\widehat{\boldsymbol{\gamma}}^{\mathrm{ML}}_{\boldsymbol{x}}$ in (4.12). To evaluate the proposed mapping function $\widehat{\epsilon}^{\mathrm{Map}}$ for the bound of the spherical uncertainty set in (5.40), we compare the performance of the SC and DC estimates using the oracle bound $\widehat{\epsilon}^{\mathrm{O}}_{\mathrm{ML}}$ in (5.39) and using the mapping function $\widehat{\epsilon}^{\mathrm{Map}}$.

To evaluate the accuracy of the normalized speech correlation vector estimates, we use the normalized mean-square error (MSE) between the estimated oracle normalized speech correlation vector $\widehat{\boldsymbol{\gamma}}^{\mathrm{II}}_{\boldsymbol{x}}$ in (4.7) and the estimated normalized speech correlation vector $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}$, i.e.,

$$MSE = \frac{1}{|\mathbb{P}^*|} \sum_{k,m \in \mathbb{P}^*} \frac{\left\| \widehat{\boldsymbol{\gamma}}^{\mathrm{II}}_{\boldsymbol{x}}(k,m) - \widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}(k,m) \right\|_2^2}{\left\| \widehat{\boldsymbol{\gamma}}^{\mathrm{II}}_{\boldsymbol{x}}(k,m) \right\|_2^2}, \tag{5.41}$$

where $\mathbb{P}^*$ either denotes the set of time-frequency points that contain noise-only $\mathbb{P}^N$ or speech-and-noise $\mathbb{P}^Y$, defined as time-frequency points with an oracle a-priori SNR estimate $\widehat{\xi}^{\mathrm{O}}$ in (4.8) smaller or larger than -5 dB, respectively. Furthermore, we classify time-frequency points whose normalized squared error is larger than 200 as outliers and exclude them from the *MSE* calculation.

For different SNRs, Fig. 5.4 depicts the performance, averaged over all combinations of speech and noise signals, in terms of the *MSE* and the percentage of outliers in speech-and-noise time-frequency points (lower bar) and noise-only time-frequency points (upper bar). For all SNRs, it can be clearly observed that the SC and DC estimates achieve a considerably lower *MSE* than the ML estimate, where the DC estimate achieves the lowest *MSE* of all considered estimates (both in speech-and-noise and noise-only time-frequency points). This shows that the accuracy of the normalized speech correlation vector estimate can be substantially improved by jointly considering the quadratic inequality constraint and the normalization constraint (see Section 5.1.2).

Furthermore, it can be observed for both the SC and DC estimates that the *MSE* obtained using the proposed mapping function $\widehat{\epsilon}^{\text{Map}}$ is similar to the *MSE* obtained using the oracle bound $\widehat{\epsilon}^{\text{O}}_{\text{ML}}$, showing that the proposed mapping function is a good approximation. In addition, whereas the ML estimate causes a large amount of outliers, resulting in speech distortion and unpleasant artifacts in the background noise, it can be observed that no outliers occur for the SC and DC estimates.

In conclusion, these results show that the SC and DC estimates are more accurate than the state-of-the-art ML estimate, with the DC estimate achieving the highest estimation accuracy.

5.2.3 *Instrumental Speech Enhancement Performance*

In this section, the speech enhancement performance of the proposed SC-MFMPDR and DC-MFMPDR filters using the mapping function $\widehat{\epsilon}^{\text{Map}}$ is evaluated and compared with the ML-MFMPDR filter. As already mentioned, although the main objective is to compare the proposed constrained MFMPDR filters with the ML-MFMPDR filter, we also consider the single-frame WG as reference algorithm. For different SNRs, Fig 5.5 depicts the results, averaged over all combinations of speech and noise signals, in terms of the considered instrumental performance measures, i.e., segSSNR (speech distortion), segNR (noise reduction), $\Delta\Psi_{\log}$ (noise distortion) and ΔPESQ (speech quality) (cf. Section 2.2).

First, it can be observed that the constrained MFMPDR filters yield larger segSSNR values (i.e., less speech distortion) but smaller segNR values (i.e., less noise reduction) than the ML-MFMPDR filter and the WG. Among the MFMPDR filters, the DC-MFMPDR filter yields the largest segSSNR values, which can be explained by the high estimation accuracy of the normalized DC speech correlation vector (see Fig. 5.4). The more conservative noise reduction performance (especially at low SNRs) of the SC-MFMPDR and DC-MFMPDR filters compared to the ML-MFMPDR filter can be explained by the additional robustness constraints. Second, it can be observed that the constrained MFMPDR filters yield lower $\Delta\Psi_{\log}$ values (i.e., less noise distortion) than the ML-MFMPDR filter and the WG, and that among the MFMPDR filters, the DC-MFMPDR filter yields the lowest $\Delta\Psi_{\log}$ values. Third, the ΔPESQ results indicate that at low SNRs a better overall quality is obtained by the constrained MFMPDR filters than the ML-MFMPDR filter, whereas at high SNRs a better overall quality is obtained by the ML-MFMPDR filter than the

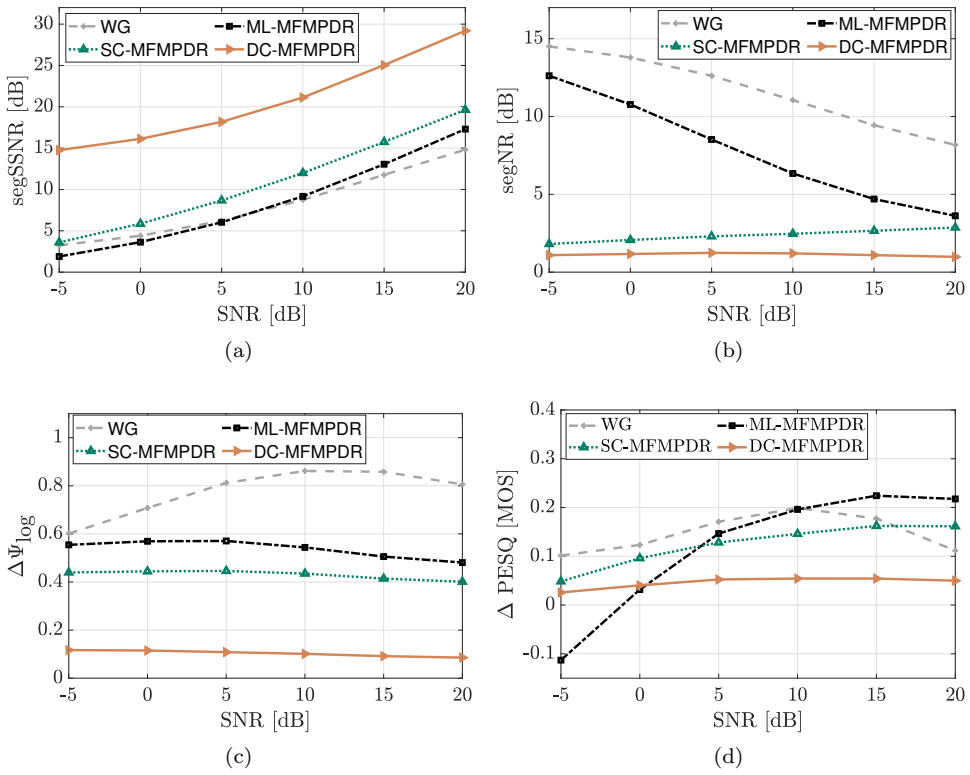


Fig. 5.5: Average (a) segmental speech SNR (segSSNR), (b) segmental noise reduction (segNR), (c) weighted log kurtosis ratio ($\Delta\Psi_{\log}$) and (d) PESQ improvement (ΔPESQ) obtained using the Wiener gain (WG), the ML-MFMPDR filter and the proposed SC-MFMPDR and DC-MFMPDR filters for different SNRs.

constrained MFMPDR filters. However, since for all SNRs informal listening experiments suggest that for the constrained MFMPDR filters the speech sounds more natural and less musical noise is present than for the ML-MFMPDR filter for all SNRs, we decided to conduct a formal listening test.

5.2.4 *Perceptual Speech Enhancement Performance*

The perceptual evaluation was performed using a pairwise preference test, where the SC-MFMPDR filter, the DC-MFMPDR filter, the ML-MFMPDR filter, the WG and the unprocessed noisy speech signal were compared to each other resulting in a comparison of ten signal pairs. All signal pairs were evaluated under four acoustic scenarios in terms of three evaluation criteria: speech quality, noise reduction and overall quality. For the four acoustic scenarios, we mixed traffic and babble noise [145] as noise signals with a randomly selected TIMIT sentence [146] as the speech

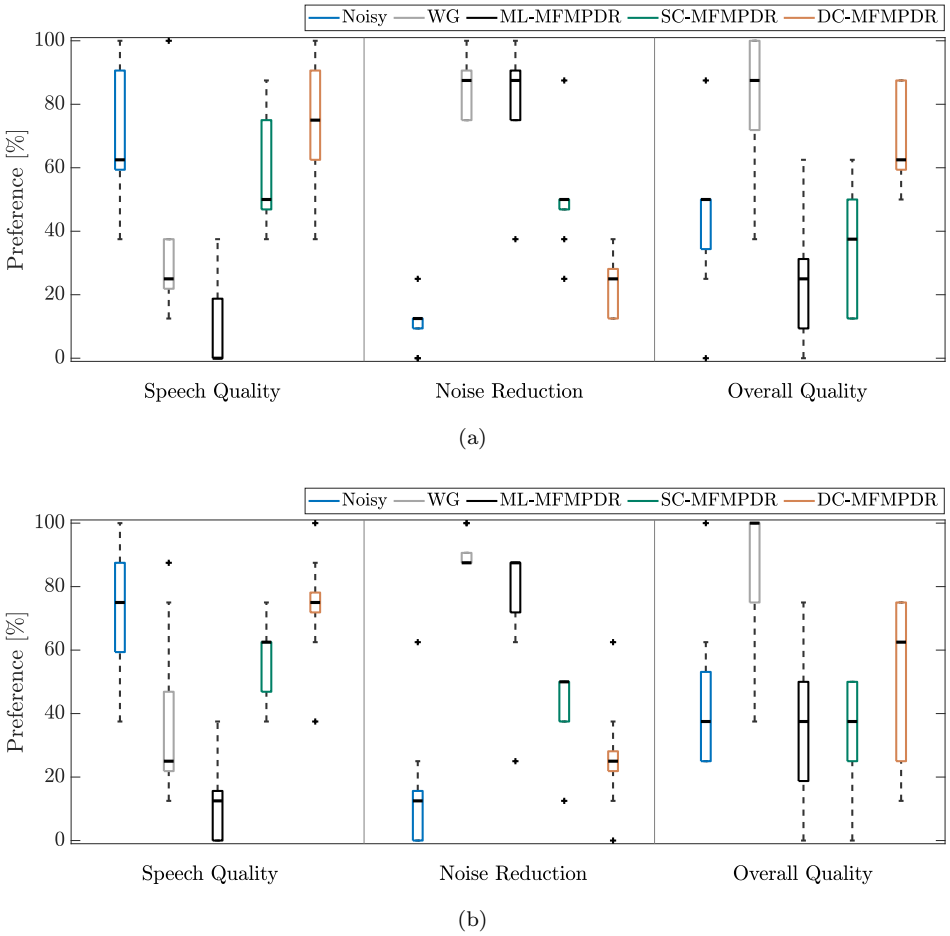


Fig. 5.6: Boxplots of the preference ratings for the unprocessed noisy speech signal, the WG, the ML-MFMPDR filter and the proposed SC-MFMPDR and DC-MFMPDR filters for the criteria speech quality, noise reduction and overall quality for SNRs of (a) 0 dB and (b) 10 dB (averaged over both noise types). On each box, the central horizontal line is the median, the edges of the box are the 25-th and 75-th percentiles and the whiskers extend to 1.5 times the interquartile range from the median. Outliers are indicated by + markers.

signal at an SNR of 0 and 10 dB. We asked 15 self-reported normal-hearing subjects aged between 23 and 39 to judge which of the two presented signals is preferred in terms of the three evaluation criteria per acoustic scenario. The order of the acoustic scenarios as well as the order of the signal pairs was randomized. As a reference, the speech signal was also available. After all signal pairs per acoustic scenario were evaluated for one criterion, the next criterion was rated.

The signals were presented diotically over Sennheiser HDA 200 headphones using an RME Babyface sound card in a quiet office. Similarly as in [147], the ratings are

expressed in the percentage of times each algorithm was preferred, i.e., a value of 100 % indicates that the algorithm won every pairwise comparison it was involved in. To determine statistical significance, we used the non-parametric Friedman test [148] followed by a Wilcoxon signed-rank test [149] with Bonferroni correction [150]. Statistical significance is assumed for $p < 0.05$.

Since the ratings between the two noise types showed no statistically significant difference for both SNRs, we averaged the ratings over the two noise types for each SNR. Fig. 5.6 shows the boxplots of the average preference ratings for the three considered evaluation criteria for both SNRs. In terms of speech quality, the results show that for both SNRs the DC-MFMPDR filter (75 % at 0 dB, 75 % at 10 dB) is preferred over the SC-MFMPDR filter (50 % at 0 dB, 63 % at 10 dB), the WG (25 % at 0 dB, 25 % at 10 dB) and especially the ML-MFMPDR filter (0 % at 0 dB, 13 % at 10 dB), where the differences between the ML-MFMPDR filter and the constrained MFMPDR filters are statistically significant. Moreover, the results show that the preference between the DC-MFMPDR filter and the unprocessed signal (63 % at 0 dB, 75 % at 10 dB) is very similar in terms of speech quality. In terms of noise reduction, the results show that for both SNRs the ML-MFMPDR filter (88 % at 0 dB, 88 % at 10 dB) and the WG (88 % at 0 dB, 88 % at 10 dB) are preferred over the SC-MFMPDR filter (50 % at 0 dB, 50 % at 10 dB), the DC-MFMPDR filter (25 % at 0 dB, 25 % at 10 dB) and the unprocessed signal (13 % at 0 dB, 13 % at 10 dB), with statistically significant differences at both SNRs. In terms of overall quality, the results show that for both SNRs the DC-MFMPDR filter (63 % at 0 dB, 63 % at 10 dB) is preferred over the unprocessed signal (50 % at 0 dB, 38 % at 10 dB), the SC-MFMPDR filter (38 % at 0 dB, 38 % at 10 dB) and the ML-MFMPDR filter (25 % at 0 dB, 38 % at 10 dB). Despite the large speech distortion of the WG, these results also show that the WG (88 % at 0 dB, 100 % at 10 dB) is preferred compared to all MFMPDR filters, presumably due to its larger noise reduction.

In conclusion, the results from the perceptual listening test show that the proposed DC-MFMPDR filter leads to less speech distortion and a more natural speech quality than the SC-MFMPDR and ML-MFMPDR filters and the WG, but that the proposed constrained MFMPDR filters are more conservative in suppressing the background noise than the ML-MFMPDR filter and the WG.

5.3 Summary

In this chapter, we investigated the potential of using concepts proposed for robust MPDR beamforming in the context of single-microphone multi-frame speech enhancement. We proposed two constrained MFMPDR filters that estimate the normalized speech correlation vector as the vector maximizing the total signal output PSD within a spherical uncertainty set. This corresponds to imposing a quadratic inequality constraint on the mismatch vector with respect to the presumed normalized speech correlation vector. While the SC-MFMPDR filter only considers the quadratic inequality constraint and applies the required normalization afterwards, the DC-MFMPDR jointly considers the quadratic inequality constraint and the linear normalization constraint in the optimization problem. To set the upper bound of

the spherical uncertainty set, we proposed to use a trained non-linear mapping function that depends on the a-priori SNR estimate. Simulation results show that the proposed approaches to estimate the normalized speech correlation vector clearly lead to a more accurate estimate than the ML estimate, with the DC estimate achieving the highest estimation accuracy. An instrumental and a perceptual evaluation for different speech signals, noise types and SNRs indicated that although the proposed constrained MFMPDR filters lead to a more conservative noise reduction than the ML-MFMPDR filter and the WG, especially the DC-MFMPDR filter produces less speech and noise distortions than the state-of-the-art ML-MFMPDR filter, such that the speech sounds more natural and less musical noise is present.

6

NORMALIZED SPEECH CORRELATION VECTOR ESTIMATION BASED ON A LOW-RANK SPEECH MODEL

Whereas in Chapter 5 we used concepts from robust beamforming to improve the robustness of the multi-frame minimum power distortionless response (MFMPDR) filter against estimation errors in the normalized speech correlation vector, in this chapter based on a low-rank speech model we propose matrix-based methods to estimate the normalized speech correlation vector.

Most existing estimators of the normalized speech correlation vector, such as the maximum-likelihood (ML) estimator discussed in Section 4.1.2 or the estimator based on a high frequency-resolution STFT filterbank proposed in [16], only make use of the first column of the estimated noisy speech, noise and/or speech correlation matrices and hence do not exploit the information of the complete estimated correlation matrices. Assuming that speech signals can be modeled using a low-rank model [18,19], e.g., as a linear combination of a limited number of complex exponentials, the speech correlation matrix can be assumed to be rank-deficient, e.g., of rank Q. Based on this assumption, we propose two methods to estimate the normalized speech correlation matrix from which the normalized speech correlation vector can be easily computed. The *matrix-subtraction method* first subtracts the estimated normalized noise correlation matrix from the estimated noisy speech correlation matrix and then estimates the normalized speech correlation matrix using the Q largest eigenvalues and the corresponding eigenvectors of this matrix. The *subspace-decomposition method* estimates the normalized speech correlation matrix using the Q largest eigenvalues and the corresponding eigenvectors of the prewhitened normalized noisy speech correlation matrix.

This chapter is partly based on:

[135] D. Fischer and S. Doclo, "Subspace-based speech correlation vector estimation for single microphone multi-frame MVDR Filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 856 - 860.

For both methods, an estimate of the normalized noise correlation matrix and the speech model order is required. Similarly to the vector-based ML estimator in Section 4.1.2, where an estimate of the mean normalized noise correlation vector is used instead of the normalized noise correlation vector, we will use an estimate of the mean normalized noise correlation matrix for the matrix-based estimators. For the speech model order, we either assume that the uncorrelated speech component in the multi-frame signal model can be neglected and use a fixed value equal to 1 for all time-frequency points or we estimate the speech model order for each time-frequency point. Estimating the dimension of the signal subspace from a noisy correlation matrix is a well-known problem, for which several model order selection criteria such as the Akaike information criterion (AIC) [151, 152] and the minimum description length (MDL) selection criterion [152,153] were proposed. However, since for the considered MFMPDR filter the noisy speech correlation matrix is typically estimated using a small amount of data (i.e., 10-20 ms), these estimators lead to an inaccurate estimate of the speech model order. To increase the accuracy of the speech model order estimate, we propose to incorporate the a-priori SNR estimate into the MDL selection criterion, similarly to [38], where the noise PSD estimate was incorporated into the MDL selection criterion.

In Section 6.1, we propose two matrix-based methods to estimate the normalized speech correlation vector based on a low-rank speech model, namely the matrix-subtraction (MS) method and the subspace-decomposition (SD) method. In Section 6.2, we propose several methods to estimate the speech model order for each time-frequency point. In Section 6.3, we provide an instrumental performance comparison for different speech signals, noise types and SNRs between both matrix-based MFMPDR filters, the state-of-the-art vector-based ML-MFMPDR filter and the single-frame WG as reference speech enhancement algorithm. Simulation results show that the proposed SD-MFMPDR filter leads to a better speech quality and more noise reduction than the ML-MFMPDR filter, while keeping speech distortion low.

6.1 Matrix-based Normalized Speech Correlation Vector Estimators

In this section, we propose two matrix-based methods to estimate the normalized speech correlation vector, namely the MS method in Section 6.1.1 and the SD method in Section 6.1.2. Both methods are based on the eigenvalue decomposition (EVD) of a matrix, which is either constructed by subtracting the estimated normalized noise correlation matrix from the estimated normalized noisy speech correlation matrix or by prewhitening the estimated normalized noisy speech correlation matrix with the estimated normalized noise correlation matrix. In Section 6.1.3, we discuss the estimation of the normalized noise correlation matrix. For conciseness, the frequency-bin index k and the time-frame index m will be omitted in this chapter if not required. However, it should be realized that all calculations are performed for each time-frequency point.

6.1.1 *Matrix-Subtraction Method*

Similarly to the vector formulation in (2.22), it can be easily shown that the normalized speech correlation matrix $\mathbf{\Gamma}_{\boldsymbol{x}}$ in (2.26) is equal to subtracting the scaled normalized noise correlation matrix $\mathbf{\Gamma}_{\boldsymbol{n}}$ in (2.28) from the scaled normalized noisy speech correlation matrix $\mathbf{\Gamma}_{\boldsymbol{y}}$ in (2.27), i.e.,

$$\mathbf{\Gamma}_{\boldsymbol{x}} = \frac{\xi+1}{\xi}\mathbf{\Gamma}_{\boldsymbol{y}} - \frac{1}{\xi}\mathbf{\Gamma}_{\boldsymbol{n}}\,, \tag{6.1}$$

with ξ the a-priori SNR defined in (2.7).
Assuming that the speech correlation matrix $\boldsymbol{R}_{\boldsymbol{x}}$ is of rank Q, with $Q \leq L$, it was shown in Section 2.1.2 that

$$\mathbf{\Gamma}_{\boldsymbol{x}} = \boldsymbol{\gamma}_{\boldsymbol{x}}\boldsymbol{\gamma}_{\boldsymbol{x}}^{H} + \mathbf{\Gamma}_{\boldsymbol{x}'}, \tag{6.2}$$

with $\mathbf{\Gamma}_{\boldsymbol{x}'}$ a rank-(Q-1) matrix. The EVD of $\mathbf{\Gamma}_{\boldsymbol{x}}$ is given by

$$\mathbf{\Gamma}_{\boldsymbol{x}} = \boldsymbol{W}\mathbf{\Lambda}_{\boldsymbol{x}}\boldsymbol{W}^{H} = \sum_{q=1}^{Q}\lambda_{\boldsymbol{x},q}\boldsymbol{w}_q\boldsymbol{w}_q^{H}, \tag{6.3}$$

where the $L\times Q$-dimensional matrix $\boldsymbol{W}$ contains the orthonormal eigenvectors $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_Q$, and the diagonal elements of the $Q\times Q$-dimensional matrix $\mathbf{\Lambda}_{\boldsymbol{x}}$ are the corresponding speech eigenvalues, with $\lambda_{\boldsymbol{x},1} \geq \lambda_{\boldsymbol{x},2} \geq \ldots \geq \lambda_{\boldsymbol{x},Q}$.
In practice, let $\widehat{\mathbf{\Gamma}}_{\boldsymbol{x}}^{\mathrm{MS}}$ be an estimate of $\mathbf{\Gamma}_{\boldsymbol{x}}$ based on (6.1), i.e.,

$$\boxed{\widehat{\mathbf{\Gamma}}_{\boldsymbol{x}}^{\mathrm{MS}} = \frac{\widehat{\xi}+1}{\widehat{\xi}}\,\widehat{\mathbf{\Gamma}}_{\boldsymbol{y}} - \frac{1}{\widehat{\xi}}\widehat{\mathbf{\Gamma}}_{\boldsymbol{n}}} \tag{6.4}$$

where $\widehat{\mathbf{\Gamma}}_{\boldsymbol{y}}$ and $\widehat{\mathbf{\Gamma}}_{\boldsymbol{n}}$ denote estimates of the normalized noisy speech and noise correlation matrices, respectively, and $\widehat{\xi}$ denotes an estimate of the a-priori SNR (cf. Section 3.1.2). Since $\widehat{\mathbf{\Gamma}}_{\boldsymbol{x}}^{\mathrm{MS}}$ is typically a full-rank matrix, the EVD of $\widehat{\mathbf{\Gamma}}_{\boldsymbol{x}}^{\mathrm{MS}}$ is given by

$$\widehat{\mathbf{\Gamma}}_{\boldsymbol{x}}^{\mathrm{MS}} = \sum_{q=1}^{L}\widehat{\lambda}_{\boldsymbol{x},q}\widehat{\boldsymbol{w}}_q\widehat{\boldsymbol{w}}_q^{H}, \tag{6.5}$$

where $\widehat{\boldsymbol{w}}_q$ and $\widehat{\lambda}_{\boldsymbol{x},q}$ denote the q-th eigenvector and the corresponding eigenvalue of $\widehat{\mathbf{\Gamma}}_{\boldsymbol{x}}^{\mathrm{MS}}$, respectively. It should be noted that the eigenvalues $\widehat{\lambda}_{\boldsymbol{x},q}$ are not guaranteed

to be larger than or equal to zero. The best rank-$\widehat{Q}$ approximation of $\widehat{\mathbf{\Gamma}}_{\boldsymbol{x}}^{\mathrm{MS}}$, with $\widehat{Q}$ the estimated speech model order (cf. Section 6.2), can be obtained as

$$\boxed{\widehat{\mathbf{\Gamma}}_{\boldsymbol{x}}^{\mathrm{MS},Q} = \sum_{q=1}^{\widehat{Q}} \widehat{\lambda}_{\boldsymbol{x},q} \widehat{\boldsymbol{w}}_q \widehat{\boldsymbol{w}}_q^H} \tag{6.6}$$

The *MS normalized speech correlation vector* estimate $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{MS},Q}$ is then given by

$$\boxed{\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{MS},Q} = \frac{\widehat{\mathbf{\Gamma}}_{\boldsymbol{x}}^{\mathrm{MS},Q} \boldsymbol{e}}{\boldsymbol{e}^T \widehat{\mathbf{\Gamma}}_{\boldsymbol{x}}^{\mathrm{MS},Q} \boldsymbol{e}}} \tag{6.7}$$

where the normalization guarantees that the first element of $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{MS},Q}$ is equal to 1 and $\boldsymbol{e} = \begin{bmatrix} 1, & 0, & \ldots, & 0 \end{bmatrix}^T$ is an L-dimensional selection vector.

6.1.2 *Subspace-Decomposition Method*

For the SD method, we consider the prewhitened normalized noisy speech correlation matrix $\mathbf{\Gamma}_{\boldsymbol{y}}^{w}$, defined as

$$\mathbf{\Gamma}_{\boldsymbol{y}}^{w} = \boldsymbol{C}^{-1} \mathbf{\Gamma}_{\boldsymbol{y}} \boldsymbol{C}^{-H}, \tag{6.8}$$

with $\boldsymbol{C}$ the $L \times L$-dimensional lower triangular Cholesky factor [144] of the normalized noise correlation matrix $\mathbf{\Gamma}_{\boldsymbol{n}}$, i.e.,

$$\mathbf{\Gamma}_{\boldsymbol{n}} = \boldsymbol{C}\boldsymbol{C}^H. \tag{6.9}$$

Using (6.1) and (6.8), the prewhitened normalized speech correlation matrix $\mathbf{\Gamma}_{\boldsymbol{x}}^{w}$ is given by

$$\mathbf{\Gamma}_{\boldsymbol{x}}^{w} = \boldsymbol{C}^{-1} \mathbf{\Gamma}_{\boldsymbol{x}} \boldsymbol{C}^{-H} = \frac{\xi + 1}{\xi} \mathbf{\Gamma}_{\boldsymbol{y}}^{w} - \frac{1}{\xi} \boldsymbol{I}_L, \tag{6.10}$$

with $\boldsymbol{I}_L$ the $L \times L$-dimensional identity matrix. Let the EVD of $\mathbf{\Gamma}_{\boldsymbol{y}}^{w}$ be given by

$$\mathbf{\Gamma}_{\boldsymbol{y}}^{w} = \boldsymbol{V} \mathbf{\Lambda}_{\boldsymbol{y}}^{w} \boldsymbol{V}^H = \sum_{q=1}^{L} \lambda_{\boldsymbol{y},q}^{w} \boldsymbol{v}_q \boldsymbol{v}_q^H, \tag{6.11}$$

where $\boldsymbol{V}$ contains the orthonormal eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_L$, and the diagonal elements of $\mathbf{\Lambda}_{\boldsymbol{y}}^{w}$ are the corresponding prewhitened noisy speech eigenvalues, with $\lambda_{\boldsymbol{y},1}^{w} \geq \lambda_{\boldsymbol{y},2}^{w} \geq \ldots \geq \lambda_{\boldsymbol{y},L}^{w}$. The EVD of the prewhitened normalized speech cor-

relation matrix $\mathbf{\Gamma}_{\boldsymbol{x}}^{w}$ in (6.10) can hence be written using the eigenvectors and the eigenvalues of $\mathbf{\Gamma}_{\boldsymbol{y}}^{w}$ in (6.11) as

$$\mathbf{\Gamma}_{\boldsymbol{x}}^{w} = \boldsymbol{V}\left(\frac{\xi+1}{\xi}\mathbf{\Lambda}_{\boldsymbol{y}}^{w} - \frac{1}{\xi}\boldsymbol{I}_{L}\right)\boldsymbol{V}^{H} = \sum_{q=1}^{L}\left(\frac{\xi+1}{\xi}\lambda_{\boldsymbol{y},q}^{w} - \frac{1}{\xi}\right)\boldsymbol{v}_{q}\boldsymbol{v}_{q}^{H}\,. \tag{6.12}$$

Assuming that $\mathbf{\Gamma}_{\boldsymbol{x}}$ and hence also $\mathbf{\Gamma}_{\boldsymbol{x}}^{w}$ is of rank Q, with $Q \leq L$, it can be easily seen that

$$\mathbf{\Gamma}_{\boldsymbol{x}}^{w} = \sum_{q=1}^{Q}\lambda_{\boldsymbol{x},q}^{w}\boldsymbol{v}_{q}\boldsymbol{v}_{q}^{H}\,, \tag{6.13}$$

with the prewhitened speech eigenvalues $\lambda_{\boldsymbol{x},q}^{w}$ equal to

$$\lambda_{\boldsymbol{x},q}^{w} = \frac{\xi+1}{\xi}\lambda_{\boldsymbol{y},q}^{w} - \frac{1}{\xi}, \qquad q = 1,\ldots,Q\,, \tag{6.14}$$

and

$$\lambda_{\boldsymbol{y},q}^{w} = \frac{1}{\xi+1}, \qquad q = Q+1,\ldots,L\,. \tag{6.15}$$

Hence, $\mathbf{\Gamma}_{\boldsymbol{y}}^{w}$ in (6.11) can be decomposed into a signal subspace of dimension Q and a noise-only subspace of dimension $(L-Q)$, i.e.,

$$\mathbf{\Gamma}_{\boldsymbol{y}}^{w} = \boldsymbol{V}\left[\begin{array}{c|c} \mathbf{\Lambda}_{\boldsymbol{y},Q}^{w} & \mathbf{0}_{Q\times(L-Q)} \\ \hline \mathbf{0}_{(L-Q)\times Q} & \dfrac{1}{\xi+1}\boldsymbol{I}_{L-Q} \end{array}\right]\boldsymbol{V}^{H}, \tag{6.16}$$

with $\mathbf{\Lambda}_{\boldsymbol{y},Q}^{w}$ a $Q\times Q$-dimensional diagonal matrix containing the eigenvalues of the signal subspace and $\mathbf{0}_{Q\times(L-Q)}$ denoting a $Q\times(L-Q)$-dimensional zero matrix. By transforming $\mathbf{\Gamma}_{\boldsymbol{x}}^{w}$ in (6.13) back from the prewhitening-domain, the normalized speech correlation matrix $\mathbf{\Gamma}_{\boldsymbol{x}}$ can hence be written as

$$\mathbf{\Gamma}_{\boldsymbol{x}} = \boldsymbol{C}\left(\sum_{q=1}^{Q}\lambda_{\boldsymbol{x},q}^{w}\boldsymbol{v}_{q}\boldsymbol{v}_{q}^{H}\right)\boldsymbol{C}^{H}\,. \tag{6.17}$$

In practice, let $\widehat{\mathbf{\Gamma}}_{\boldsymbol{y}}^{w}$ be an estimate of $\mathbf{\Gamma}_{\boldsymbol{y}}^{w}$ in (6.8), i.e.,

$$\widehat{\mathbf{\Gamma}}_{\boldsymbol{y}}^{w} = \widehat{\boldsymbol{C}}^{-1}\widehat{\mathbf{\Gamma}}_{\boldsymbol{y}}\widehat{\boldsymbol{C}}^{-H}, \tag{6.18}$$

with $\widehat{\mathbf{\Gamma}}_{\boldsymbol{y}}$ an estimate of the normalized noisy speech correlation matrix and $\widehat{\boldsymbol{C}}$ the Cholesky factor of the estimated normalized noise correlation matrix $\widehat{\mathbf{\Gamma}}_{\boldsymbol{n}}$. Similarly as in (6.11), the EVD of $\widehat{\mathbf{\Gamma}}_{\boldsymbol{y}}^{w}$ is given by

$$\widehat{\mathbf{\Gamma}}_{\boldsymbol{y}}^{w} = \sum_{q=1}^{L} \widehat{\lambda}_{\boldsymbol{y},q}^{w} \widehat{\boldsymbol{v}}_q \widehat{\boldsymbol{v}}_q^H \,, \tag{6.19}$$

where $\widehat{\boldsymbol{v}}_q$ and $\widehat{\lambda}_{\boldsymbol{y},q}^{w}$ denote the q-th eigenvector and the corresponding eigenvalue of $\widehat{\mathbf{\Gamma}}_{\boldsymbol{y}}^{w}$, respectively. Similarly to (6.17), we propose to estimate the normalized speech correlation matrix as

$$\boxed{\widehat{\mathbf{\Gamma}}_{\boldsymbol{x}}^{\mathrm{SD},Q} = \widehat{\boldsymbol{C}} \left(\sum_{q=1}^{\widehat{Q}} \widehat{\lambda}_{\boldsymbol{x},q}^{w} \widehat{\boldsymbol{v}}_q \widehat{\boldsymbol{v}}_q^H \right) \widehat{\boldsymbol{C}}^H} \tag{6.20}$$

with $\widehat{Q}$ the estimated speech model order (cf. Section 6.2) and $\widehat{\lambda}_{\boldsymbol{x},q}^{w}$ an estimate of the q-th prewhitened speech eigenvalue according to (6.14), i.e.,

$$\boxed{\widehat{\lambda}_{\boldsymbol{x},q}^{w} = \frac{\widehat{\xi}+1}{\widehat{\xi}} \widehat{\lambda}_{\boldsymbol{y},q}^{w} - \frac{1}{\widehat{\xi}}, \qquad q = 1, \ldots, \widehat{Q}} \tag{6.21}$$

The *SD normalized speech correlation vector* estimate $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SD},Q}$ is then given by

$$\boxed{\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SD},Q} = \frac{\widehat{\mathbf{\Gamma}}_{\boldsymbol{x}}^{\mathrm{SD},Q} \boldsymbol{e}}{\boldsymbol{e}^T \widehat{\mathbf{\Gamma}}_{\boldsymbol{x}}^{\mathrm{SD},Q} \boldsymbol{e}}} \tag{6.22}$$

where the normalization guarantees that the first element of $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SD},Q}$ is equal to 1.

6.1.3 *Normalized Noise Correlation Matrix*

Both proposed matrix-based methods in Section 6.1.1 and 6.1.2 require an estimate of the normalized noise correlation matrix $\mathbf{\Gamma}_{\boldsymbol{n}}$. In practice, it is rather difficult to accurately estimate this time-varying correlation matrix. Similarly to the vector-based ML estimator in Section 4.1.2, where an estimate of the mean normalized noise correlation vector $\boldsymbol{\mu}_{\boldsymbol{\gamma}_{\boldsymbol{n}}}$ is used instead of the normalized noise correlation vector $\boldsymbol{\gamma}_{\boldsymbol{n}}$, we propose to use an estimate of the mean normalized noise correlation matrix $\mathbf{\Gamma}_{\boldsymbol{n}}^{\mu}$ instead of $\mathbf{\Gamma}_{\boldsymbol{n}}$. Since $\mathbf{\Gamma}_{\boldsymbol{n}}$ is a Hermitian Toeplitz matrix, it is completely defined by its first column, i.e., the normalized noise correlation vector $\boldsymbol{\gamma}_{\boldsymbol{n}}$. Hence, using $\widehat{\boldsymbol{\mu}}_{\boldsymbol{\gamma}_{\boldsymbol{n}}}^{\mathrm{M}}$ in (4.13), an estimate of the mean normalized noise correlation matrix $\mathbf{\Gamma}_{\boldsymbol{n}}^{\mu}$ can be obtained as

$$\widehat{\mathbf{\Gamma}}_{\boldsymbol{n}}^{\mu} = \mathrm{Toeplitz}\left[\widehat{\boldsymbol{\mu}}_{\boldsymbol{\gamma}_{\boldsymbol{n}}}^{\mathrm{M}}\right], \tag{6.23}$$

Using (6.23), it should be noted that for $\widehat{Q} = L$ the MS normalized speech correlation vector estimate in (6.7) and the SD normalized speech correlation vector estimate in (6.22) yield the ML estimate in (4.12).

6.2 Speech Model Order

Both proposed matrix-based methods require an estimate of the speech model order Q, cf. (6.6) and (6.20). In this section, we present different estimators for the speech model order. We either assume that the normalized speech correlation matrix has rank 1 and use a fixed value ($\widehat{Q} = 1$) for all time-frequency points, or propose several time- and frequency-dependent estimators.

6.2.1 *Rank-1 Assumption ($\widehat{Q} = 1$)*

Assuming that the uncorrelated speech component can be neglected and $\boldsymbol{\Gamma}_{\boldsymbol{x}'} = 0$ in (6.2), which is of course not the case in practice, $\boldsymbol{\Gamma}_{\boldsymbol{x}} = \boldsymbol{\gamma}_{\boldsymbol{x}}\boldsymbol{\gamma}_{\boldsymbol{x}}^H$ becomes a rank-1 matrix, i.e., $Q = 1$ for each time-frequency point. Hence, the expressions in (6.3) and (6.17) are equal to

$$\boldsymbol{\Gamma}_{\boldsymbol{x}} = \lambda_{\boldsymbol{x},1}\boldsymbol{w}_1\boldsymbol{w}_1^H = \lambda_{\boldsymbol{x},1}^{w}\left(\boldsymbol{C}\boldsymbol{v}_1\right)\left(\boldsymbol{C}\boldsymbol{v}_1\right)^H. \tag{6.24}$$

Using the time- and frequency-independent value $\widehat{Q} = 1$ for the speech model order, the MS and SD estimators for the normalized speech correlation vector in (6.7) and (6.22) then simplify to

$$\boxed{\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{MS},1} = \frac{\widehat{\boldsymbol{w}}_1}{\boldsymbol{e}^T\widehat{\boldsymbol{w}}_1}} \tag{6.25}$$

and

$$\boxed{\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SD},1} = \frac{\widehat{\boldsymbol{C}}\widehat{\boldsymbol{v}}_1}{\boldsymbol{e}^T\widehat{\boldsymbol{C}}\widehat{\boldsymbol{v}}_1}} \tag{6.26}$$

i.e., using only the principal eigenvectors $\widehat{\boldsymbol{w}}_1$ or $\widehat{\boldsymbol{v}}_1$. It should be noted that the estimator in (6.25) is similar to the so-called covariance subtraction method (with rank-1 approximation) [42, 114–117] and the estimator in (6.26) is similar to the so-called covariance whitening method [39, 42, 115, 116], which are frequently used to estimate the RTF vector of the desired speech source in multi-microphone speech enhancement.

6.2.2 *Time- and Frequency-Dependent Estimators*

Instead of using a fixed time- and frequency-independent value for the speech model order Q, in this section we present two methods to estimate a time- and frequency-dependent value, either based on the estimated (prewhitened) speech eigenvalues or on the estimated prewhitened noisy speech eigenvalues.

First, based on the estimated (prewhitened) speech eigenvalues, we can simply estimate the speech model order as the cardinality of the set of estimated positive speech eigenvalues, i.e., for the MS estimate in (6.6) as

$$\boxed{\widehat{Q}^{\mathrm{MS,pos}} = |\mathbb{Q}^{\mathrm{MS,pos}}|, \quad \text{with } \mathbb{Q}^{\mathrm{MS,pos}} = \left\{ \widehat{\lambda}_{\boldsymbol{x},q} > 0 | q = 1, 2, \ldots, L \right\}} \tag{6.27}$$

and for the SD estimate in (6.21) as

$$\boxed{\widehat{Q}^{\mathrm{SD,pos}} = |\mathbb{Q}^{\mathrm{SD,pos}}|, \quad \text{with } \mathbb{Q}^{\mathrm{SD,pos}} = \left\{ \widehat{\lambda}^{w}_{\boldsymbol{x},q} > 0 | q = 1, 2, \ldots, L \right\}} \tag{6.28}$$

Second, based on the estimated prewhitened noisy speech eigenvalues, several approaches were proposed in the literature to estimate the dimension of the signal subspace, which can be roughly classified into hypothesis testing approaches, e.g., [29–31, 38, 45, 135] and approaches based on model order selection criteria, e.g., [32, 40, 151–156]. In hypothesis testing approaches a threshold is used to decide whether a prewhitened noisy eigenvalue belongs to the signal subspace or the noisy-only subspace. The signal model order is then estimated as the cardinality of the set of eigenvalues that are larger than the threshold ϑ, i.e.,

$$\widehat{Q}^{\mathrm{Thre}} = |\mathbb{Q}^{\mathrm{Thre}}|, \quad \text{with } \mathbb{Q}^{\mathrm{Thre}} = \left\{ \widehat{\lambda}^{w}_{\boldsymbol{y},q} > \vartheta | q = 1, 2, \ldots, L \right\}. \tag{6.29}$$

A problem with this approach is the choice of the threshold ϑ, which is either set experimentally, as in [29, 45], or based on a-priori knowledge, as in [30, 38, 135]. Similarly to [38], where the noise PSD estimate of the previous time-frame was used to set the threshold ϑ, we proposed in [135] to use an a-priori SNR-based threshold, i.e.,

$$\widehat{\vartheta} = -\frac{1}{\widehat{\xi} + 1} \log\left[P_f\right], \tag{6.30}$$

with $\widehat{\xi}$ an estimate of the a-priori SNR and P_f the false alarm rate [157].

In approaches based on model order selection criteria, the signal model order is estimated as the value that minimizes an information selection criterion such as the AIC [151, 152] or the MDL criterion [152, 153]. A problem with these approaches is that when the (prewhitened) noisy correlation matrix is estimated using a limited amount of data, which is the case for the considered MFMPDR filter, these estimators tend to lead to inaccurate estimates of the signal model order. To increase the estimation accuracy, we propose to estimate the speech model order Q as the value that minimizes the MDL selection criterion incorporating the a-priori SNR

estimate. In order to show the differences between the classical MDL estimator and the proposed a-priori SNR-based MDL estimator, we will discuss both estimators in the following.

For a set of N observations $\mathbb{Y} = \{\boldsymbol{y}_1^w, \boldsymbol{y}_2^w, \ldots, \boldsymbol{y}_N^w\}$, the MDL selection criterion is defined as [153, 154]

$$\text{MDL} = -\log\left[f\left(\mathbb{Y}|\widehat{\boldsymbol{\theta}}\right)\right] + \frac{1}{2} r \log\left[N\right], \tag{6.31}$$

where $f\left(\mathbb{Y}|\widehat{\boldsymbol{\theta}}\right)$ denotes the joint probability density function of $\mathbb{Y}$ given the ML estimate $\widehat{\boldsymbol{\theta}}$ of the parameter vector $\boldsymbol{\theta}$ and r denotes the number of degrees of freedom in $\boldsymbol{\theta}$. Assuming that the observations are L-dimensional independent and identically distributed zero-mean multivariate Gaussian vectors, the joint probability density function $f\left(\mathbb{Y}|\boldsymbol{\theta}\right)$ is given by

$$f\left(\mathbb{Y}|\boldsymbol{\theta}\right) = \prod_{n=1}^{N} \frac{1}{\pi^L \det\left[\boldsymbol{\Gamma}_{\boldsymbol{y},Q}^w\right]} \exp\left(-\boldsymbol{y}_n^{w,H} \left(\boldsymbol{\Gamma}_{\boldsymbol{y},Q}^w\right)^{-1} \boldsymbol{y}_n^w\right), \tag{6.32}$$

where $\boldsymbol{\Gamma}_{\boldsymbol{y},Q}^w$ is defined as

$$\boldsymbol{\Gamma}_{\boldsymbol{y},Q}^w = \boldsymbol{V} \left[\begin{array}{c|c} \boldsymbol{\Lambda}_{\boldsymbol{y},Q}^w & \boldsymbol{0}_{Q\times(L-Q)} \\ \hline \boldsymbol{0}_{(L-Q)\times Q} & z\boldsymbol{I}_{L-Q} \end{array}\right] \boldsymbol{V}^H, \tag{6.33}$$

with z a scaling term, which is related to the a-priori SNR as $z = (\xi + 1)^{-1}$, cf. (6.16). Taking the logarithm of (6.32), the log-likelihood is given by

$$\log\left[f\left(\mathbb{Y}|\boldsymbol{\theta}\right)\right] = a - N \log\left[\det\left[\boldsymbol{\Gamma}_{\boldsymbol{y},Q}^w\right]\right] - N \text{tr}\left[\left(\boldsymbol{\Gamma}_{\boldsymbol{y},Q}^w\right)^{-1} \widehat{\boldsymbol{\Gamma}}_{\boldsymbol{y}}^w\right], \tag{6.34}$$

where a is a constant term and $\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{y}}^w$ is defined as

$$\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{y}}^w = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{y}_n^w \boldsymbol{y}_n^{w,H}. \tag{6.35}$$

In the classical MDL selection criterion, the parameter vector $\boldsymbol{\theta}$ contains the (real-valued) eigenvalues $\lambda_{\boldsymbol{y},q}^w$, $q = 1, 2, \ldots, Q$, the corresponding orthonormal (complex-valued) eigenvectors $\boldsymbol{v}_q$, $q = 1, 2, \ldots, Q$, as well as the scaling term z, i.e.,

$$\boldsymbol{\theta}^{\text{MDL}} = \left[\lambda_{\boldsymbol{y},1}^w, \lambda_{\boldsymbol{y},2}^w, \ldots \lambda_{\boldsymbol{y},Q}^w, \boldsymbol{v}_1^H, \boldsymbol{v}_2^H, \ldots \boldsymbol{v}_Q^H, z\right]. \tag{6.36}$$

Hence, $\boldsymbol{\theta}^{\text{MDL}}$ consists of $Q + 2LQ + 1$ parameters. However, since we consider orthonormal eigenvectors, not all parameters are independent such that the number of degrees of freedom is reduced by $2Q + Q(Q-1)$ [152], i.e.,

$$r = Q + 1 + 2LQ - 2Q - Q(Q-1) = (2L - Q)Q + 1\,. \tag{6.37}$$

To estimate $\mathbf{\Gamma}^{w}_{\boldsymbol{y},Q}$, we first maximize (6.34) for the parameter vector $\boldsymbol{\theta}^{\mathrm{MDL}}$ in (6.36), leading to the ML estimate $\widehat{\boldsymbol{\theta}}^{\mathrm{MDL}}$ [152], i.e.,

$$\widehat{\lambda}^{w,\mathrm{ML}}_{\boldsymbol{y},q} = \widehat{\lambda}^{w}_{\boldsymbol{y},q}, \qquad q = 1, 2, \ldots, Q, \tag{6.38}$$

$$\widehat{\boldsymbol{v}}^{\mathrm{ML}}_{q} = \widehat{\boldsymbol{v}}_{q}, \qquad q = 1, 2, \ldots, Q, \tag{6.39}$$

$$\widehat{z}^{\mathrm{ML}} = \frac{1}{L-Q} \sum_{q=Q+1}^{L} \widehat{\lambda}^{w}_{\boldsymbol{y},q}, \tag{6.40}$$

with $\widehat{\lambda}^{w}_{\boldsymbol{y},q}$ and $\widehat{\boldsymbol{v}}_{q}$ the eigenvalues and corresponding eigenvector of $\widehat{\mathbf{\Gamma}}^{w}_{\boldsymbol{y}}$, cf. (6.19). Subsequently, the ML estimates in (6.38), (6.39) and (6.40) are substituted into (6.33), resulting in

$$\widehat{\mathbf{\Gamma}}^{w,\mathrm{MDL}}_{\boldsymbol{y},Q} = \widehat{\boldsymbol{V}} \left[\begin{array}{c|c} \widehat{\mathbf{\Lambda}}^{w}_{\boldsymbol{y},Q} & \mathbf{0}_{Q\times(L-Q)} \\ \hline \mathbf{0}_{(L-Q)\times Q} & \widehat{z}^{\mathrm{ML}} \boldsymbol{I}_{L-Q} \end{array} \right] \widehat{\boldsymbol{V}}^{H}, \tag{6.41}$$

with $\widehat{\boldsymbol{V}}$ containing the eigenvectors $\widehat{\boldsymbol{v}}_{q}$, $q = 1, 2, \ldots, L$, and $\widehat{\mathbf{\Lambda}}^{w}_{\boldsymbol{y},Q}$ containing the eigenvalues $\widehat{\lambda}^{w}_{\boldsymbol{y},q}$, $q = 1, 2, \ldots, Q$. The determinant of $\widehat{\mathbf{\Gamma}}^{w,\mathrm{MDL}}_{\boldsymbol{y},Q}$ in (6.41) is equal to

$$\det\left[\widehat{\mathbf{\Gamma}}^{w,\mathrm{MDL}}_{\boldsymbol{y},Q}\right] = \left(\prod_{q=1}^{Q} \widehat{\lambda}^{w}_{\boldsymbol{y},q}\right) \left(\frac{1}{L-Q} \sum_{q=Q+1}^{L} \widehat{\lambda}^{w}_{\boldsymbol{y},q}\right)^{L-Q} \tag{6.42}$$

$$= \frac{\det\left[\widehat{\mathbf{\Gamma}}^{w}_{\boldsymbol{y}}\right]}{\left(\prod_{q=Q+1}^{L} \widehat{\lambda}^{w}_{\boldsymbol{y},q}\right)} \left(\frac{1}{L-Q} \sum_{q=Q+1}^{L} \widehat{\lambda}^{w}_{\boldsymbol{y},q}\right)^{L-Q}. \tag{6.43}$$

By substituting (6.19), (6.41) and (6.43) into (6.34) and neglecting the constant term a, the log-likelihood can be written as

$$\log\left[f\left(\mathbb{Y}|\widehat{\boldsymbol{\theta}}^{\mathrm{MDL}}\right)\right] = -N \log\left[\det\left[\widehat{\mathbf{\Gamma}}^{w,\mathrm{MDL}}_{\boldsymbol{y},Q}\right]\right] - N\mathrm{tr}\left[\left(\widehat{\mathbf{\Gamma}}^{w,\mathrm{MDL}}_{\boldsymbol{y},Q}\right)^{-1} \widehat{\mathbf{\Gamma}}^{w}_{\boldsymbol{y}}\right] \tag{6.44}$$

$$= N \log\left[\frac{\prod_{q=Q+1}^{L} \widehat{\lambda}^{w}_{\boldsymbol{y},q}}{\left(\frac{1}{L-Q} \sum_{q=Q+1}^{L} \widehat{\lambda}^{w}_{\boldsymbol{y},q}\right)^{L-Q}}\right]. \tag{6.45}$$

Substituting (6.37) and (6.45) into (6.31) leads to the classical MDL selection criterion [152], i.e.,

$$\widehat{\mathrm{MDL}}(Q) = -N \log \left[\frac{\prod_{q=Q+1}^{L} \widehat{\lambda}_{\boldsymbol{y},q}^{w}}{\left(\frac{1}{L-Q} \sum_{q=Q+1}^{L} \widehat{\lambda}_{\boldsymbol{y},q}^{w} \right)^{L-Q}} \right] + \left(\frac{1}{2}(2L-Q)Q + \frac{1}{2} \right) \log[N]. \quad (6.46)$$

The classical MDL speech model order estimate $\widehat{Q}^{\mathrm{MDL}}$ is then obtained by minimizing (6.46), i.e.,

$$\widehat{Q}^{\mathrm{MDL}} = \underset{1 \leq Q \leq L}{\operatorname{argmin}} \ \widehat{\mathrm{MDL}}(Q), \quad (6.47)$$

which requires an exhaustive search.

As can be observed from (6.46), the classical MDL selection criterion does not depend on the a-priori SNR ξ, which is due to the fact that the parameter vector in (6.36) contains the scaling term $z = (\xi+1)^{-1}$. Aiming at obtaining an MDL selection criterion that depends on the a-priori SNR, we propose to define the parameter vector without the scaling term z, i.e.,

$$\boldsymbol{\theta}^{\mathrm{MDL}_\xi} = \left[\lambda_{\boldsymbol{y},1}^{w}, \lambda_{\boldsymbol{y},2}^{w}, \ldots \lambda_{\boldsymbol{y},Q}^{w}, \boldsymbol{v}_1^H, \boldsymbol{v}_2^H, \ldots \boldsymbol{v}_Q^H \right]. \quad (6.48)$$

Hence, $\boldsymbol{\theta}^{\mathrm{MDL}_\xi}$ consists of $Q+2LQ$ parameters and the number of degrees of freedom is given by

$$r = (2L-Q)Q\,. \quad (6.49)$$

To estimate $\boldsymbol{\Gamma}_{\boldsymbol{y},Q}^{w}$, we now maximize (6.34) for the parameter vector $\boldsymbol{\theta}^{\mathrm{MDL}_\xi}$ in (6.48), leading to the ML estimate $\widehat{\boldsymbol{\theta}}^{\mathrm{MDL}_\xi}$, i.e., (6.38) and (6.39). Substituting these ML estimates into (6.33) and using the a-priori SNR estimate $\widehat{\xi}$ results in

$$\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{y},Q}^{w,\mathrm{MDL}_\xi} = \widehat{\boldsymbol{V}} \left[\begin{array}{c|c} \widehat{\boldsymbol{\Lambda}}_{\boldsymbol{y},Q}^{w} & \boldsymbol{0}_{Q\times(L-Q)} \\ \hline \boldsymbol{0}_{(L-Q)\times Q} & \frac{1}{\widehat{\xi}+1} \boldsymbol{I}_{L-Q} \end{array} \right] \widehat{\boldsymbol{V}}^H, \quad (6.50)$$

which corresponds to using $(\widehat{\xi}+1)^{-1}$ instead of $\widehat{z}^{ML}$ in (6.40). Similarly to (6.42) and (6.43), the determinant of $\widehat{\mathbf{\Gamma}}_{\boldsymbol{y},Q}^{w,\mathrm{MDL}_\xi}$ in (6.50) is equal to

$$\det\left[\widehat{\mathbf{\Gamma}}_{\boldsymbol{y},Q}^{w,\mathrm{MDL}_\xi}\right] = \left(\prod_{q=1}^{Q}\widehat{\lambda}_{\boldsymbol{y},q}^{w}\right)\left(\frac{1}{\widehat{\xi}+1}\right)^{(L-Q)} \tag{6.51}$$

$$= \frac{\det\left[\widehat{\mathbf{\Gamma}}_{\boldsymbol{y}}^{w}\right]}{\left(\prod_{q=Q+1}^{L}\widehat{\lambda}_{\boldsymbol{y},q}^{w}\right)}\left(\frac{1}{\widehat{\xi}+1}\right)^{(L-Q)}. \tag{6.52}$$

By substituting (6.19), (6.50) and (6.52) into (6.34) and neglecting the constant term a, the log-likelihood can be written as

$$\log\left[f\left(\mathbb{Y}|\widehat{\boldsymbol{\theta}}^{\mathrm{MDL}_\xi}\right)\right] = -N\log\left[\det\left[\widehat{\mathbf{\Gamma}}_{\boldsymbol{y},Q}^{w,\mathrm{MDL}_\xi}\right]\right] - N\mathrm{tr}\left[\left(\widehat{\mathbf{\Gamma}}_{\boldsymbol{y},Q}^{w,\mathrm{MDL}_\xi}\right)^{-1}\widehat{\mathbf{\Gamma}}_{\boldsymbol{y}}^{w}\right] \tag{6.53}$$

$$= N\log\left[(\widehat{\xi}+1)^{(L-Q)}\prod_{q=Q+1}^{L}\widehat{\lambda}_{\boldsymbol{y},q}^{w}\right] - N\log\left[\det\left[\widehat{\mathbf{\Gamma}}_{\boldsymbol{y}}^{w}\right]\right]$$

$$-N\left(Q+(\widehat{\xi}+1)\sum_{q=Q+1}^{L}\widehat{\lambda}_{\boldsymbol{y},q}^{w}\right). \tag{6.54}$$

Since the term $N\log\left[\det\left[\widehat{\mathbf{\Gamma}}_{\boldsymbol{y}}^{w}\right]\right]$ in (6.54) does not depend on Q, we can omit this term. Substituting (6.49) and (6.54) into (6.31) leads to the proposed a-priori SNR-based MDL selection criterion, i.e.,

$$\widehat{\mathrm{MDL}}_\xi(Q) = N\left(Q+(\widehat{\xi}+1)\sum_{q=Q+1}^{L}\widehat{\lambda}_{\boldsymbol{y},q}^{w}\right) - N\log\left[(\widehat{\xi}+1)^{(L-Q)}\prod_{q=Q+1}^{L}\widehat{\lambda}_{\boldsymbol{y},q}^{w}\right] \tag{6.55}$$

$$+\frac{1}{2}(2L-Q)Q\log[N].$$

The *a-priori SNR-based MDL speech model order* estimate $\widehat{Q}^{\mathrm{MDL}_\xi}$ is then obtained by minimizing (6.55), i.e.,

$$\boxed{\widehat{Q}^{\mathrm{MDL}_\xi} = \underset{1\leq Q\leq L}{\mathrm{argmin}}\ \widehat{\mathrm{MDL}}_\xi(Q)} \tag{6.56}$$

which requires an exhaustive search.

6.3 Simulation Results

In this section, we evaluate the performance of the proposed matrix-based methods to estimate the normalized speech correlation vector based on a low-rank speech model. After defining the audio material and discussing the algorithmic settings in Section 6.3.1, in Section 6.3.2 we compare the estimation accuracy of the speech model order estimators presented in Section 6.2 with an oracle speech model order estimator. In Section 6.3.3 we compare the estimation accuracy of the proposed MS and SD methods using different speech model order estimates with the state-of-the-art vector-based ML estimate. For different speech signals, noise types and SNRs, in Section 6.3.4 we compare the speech enhancement performance of the MS- and SD-MFMPDR filters using different speech model order estimates and the ML-MFMPDR filter with the single-frame WG (cf. Section 3.1.1) as a reference speech enhancement algorithm.

6.3.1 *Audio Material and Algorithmic Settings*

For the evaluation, we used 260 s of speech material (131 s female, 129 s male) from the TIMIT database [146] as speech signals, sampled at a sampling frequency of 16 kHz. As noise signals, we used traffic noise, two babble noise signals, factory noise and modulated white Gaussian noise with a modulation frequency of 0.5 Hz taken from the NOISEX-92 database [145]. The considered SNRs ranged from -5 dB to 20 dB in 5 dB steps.

To evaluate the estimation accuracy of the normalized speech correlation vector estimates, we used the normalized MSE in (5.41) between the estimated oracle normalized speech correlation vector $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{II}}$ and the estimated normalized speech correlation vector $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}$ and the percentage of outliers. The speech enhancement performance of the MFMPDR filters and the WG is evaluated in terms of speech quality using the PESQ improvement ΔPESQ (cf. Section 2.2) and in terms of the amount of noise reduction and speech distortion using the segmental noise reduction segNR in (2.32) and the segmental speech distortion segSSNR in (2.33). All performance measures were averaged over all considered speech signals and noise types.

Similarly as in Section 4.3, in order to achieve a high speech interframe correlation we used a highly temporally resolved STFT framework with a frame length of 4 ms ($T = K = 64$) and an overlap of 75 %, resulting in a frame shift of 1 ms ($R = 16$). As the STFT analysis window $w_a(t)$ and synthesis window $w_s(t)$, we used a square-root Hann window.

The noisy speech correlation matrix $\boldsymbol{R}_{\boldsymbol{y}}$ was estimated using recursive smoothing as in (4.3), with a smoothing parameter experimentally set to $\alpha_y = 0.92$, corresponding to a smoothing window of 12 ms. Accordingly, the data length for the a-priori SNR-based MDL selection criterion in (6.55) is set to $N = 9$. To avoid numerical problems for the MFMPDR filter, we performed diagonal loading as in (4.17) before computing the inverse of the noisy speech correlation matrix, with a scaling parameter of $\kappa = 0.001$, similarly as in Section 4.3.

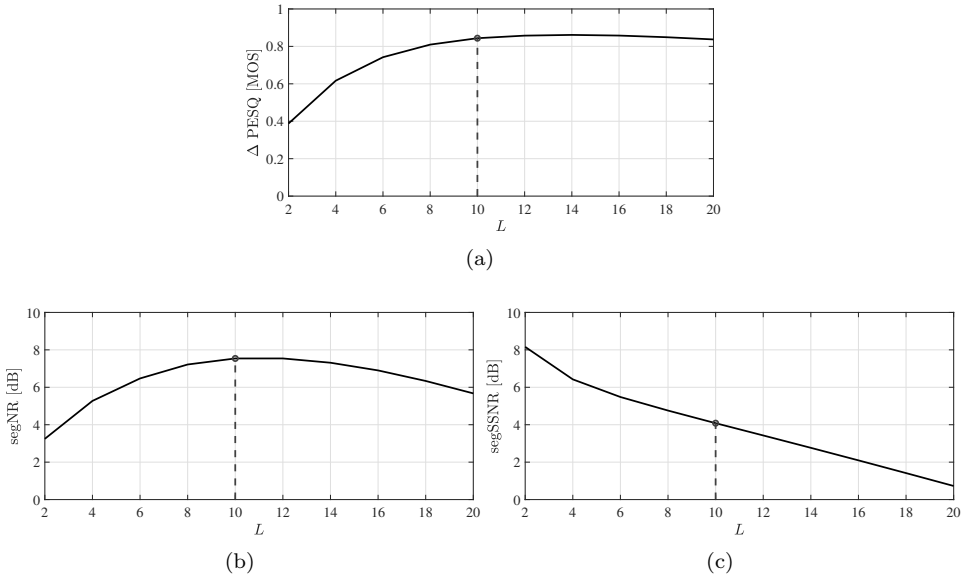


Fig. 6.1: Influence of the filter length L on the average performance (a) PESQ improvement, (b) segNR and (c) segSSNR of the MFMPDR filter using the oracle estimate $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{II}}$ for 0 dB SNR.

To estimate the a-priori SNR ξ, we used the DDA $\widehat{\xi}^{\mathrm{DDA}}$ in (3.9) with a weighting parameter $\alpha_{\mathrm{DDA}} = 0.97$. To reduce unpleasant artifacts in the background noise, e.g., musical noise, we only updated the estimated speech STFT coefficient $\widehat{X}(m-1)$ in (3.9) every 4 ms, i.e., every 4 frames. The noise PSD was estimated using the SPP-based noise PSD estimator $\widehat{\phi}_N^{\mathrm{SPP}}$ in (3.13), with a smoothing parameter $\alpha_{\mathrm{SPP}} = 0.90$ and a fixed a-priori SNR $\xi_{\mathcal{H}_1} = -15$ dB. To reduce the amount of speech distortion and to mask artifacts in the background noise, we applied a lower limit $\xi_{\min} = -8$ dB to the a-priori SNR estimate.

To set the filter length L, we investigated the influence of the filter length on the performance of the MFMPDR filter using the oracle estimate $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{II}}$ in (4.7). We considered filter lengths between 2 and 20, corresponding to a data analysis length between 5 and 23 ms. For an SNR of 0 dB, Fig. 6.1 depicts the performance in terms of ΔPESQ, segNR and segSSNR, averaged over all speech signals and noise types. In terms of ΔPESQ, it can be seen that the performance increases with increasing filter length and saturates at about $L = 10$. In terms of segNR, it can be observed that the maximum performance is obtained at $L = 10$. In terms of segSSNR, it can be seen that the performance decreases with increasing filter length L. To provide a good compromise between speech quality improvement, noise reduction and speech distortion, in this chapter we set the filter length to $L = 10$ for all MFMPDR filters, resulting in 13 ms of analysis data.

The performance of the MFMPDR filters will be compared with the single-frame WG as a reference speech enhancement algorithm. For a fair comparison, the WG is

implemented using an equivalent frame length of 13 ms and an overlap of 50 %. The a-priori SNR and noise PSD were estimated in the same way as for the MFMPDR filters, except for the estimated speech STFT coefficient $\widehat{X}(m-1)$ in (3.9), which was updated in each frame.

6.3.2 *Speech Model Order Estimation Performance*

In this section, we compare the estimation accuracy of the speech model order estimators presented in Section 6.2, namely the estimator $\widehat{Q}^{\mathrm{MS,pos}}$ in (6.27), the estimator $\widehat{Q}^{\mathrm{SD,pos}}$ in (6.28), the hypothesis testing estimator $\widehat{Q}^{\mathrm{Thre}}$ in (6.29) with the a-priori SNR-based threshold in (6.30), the MDL estimator $\widehat{Q}^{\mathrm{MDL}}$ in (6.47), the proposed a-priori SNR-based MDL estimator $\widehat{Q}^{\mathrm{MDL}_\xi}$ in (6.56) and an oracle estimator $\widehat{Q}^{O}$ as reference. Similarly to [38], the oracle estimator $\widehat{Q}^{O}$ is obtained as the cardinality of the set of estimated speech eigenvalues that cover at least 98 % of the energy of an oracle speech correlation matrix estimate $\widehat{\boldsymbol{R}}_{\boldsymbol{x}}^{\mathrm{O}}$, computed as in (4.1), i.e.,

$$\widehat{Q}^{\mathrm{O}} = |\mathbb{Q}^{\mathrm{O}}|, \quad \text{with } \mathbb{Q}^{\mathrm{O}} = \left\{ \frac{\sum_{q=1}^{Q} \widehat{\lambda}_{\boldsymbol{x},q}^{\mathrm{O}}}{\mathrm{tr}\left[\widehat{\boldsymbol{R}}_{\boldsymbol{x}}^{\mathrm{O}}\right]} < 0.98 \right\} \tag{6.57}$$

with $\widehat{\lambda}_{\boldsymbol{x},q}^{\mathrm{O}}$ the eigenvalues of $\widehat{\boldsymbol{R}}_{\boldsymbol{x}}^{\mathrm{O}}$.

For a speech signal from the TIMIT database [146] corrupted by traffic noise [145] at 5 dB SNR, Fig. 6.2 depicts the spectrograms of the speech and noisy speech signals and the estimated speech model order for the considered estimators. First, for the oracle estimator $\widehat{Q}^{O}$ in Fig. 6.2c, it can be observed that for voiced sounds $\widehat{Q}^{O}$ is smaller than for unvoiced sounds and that $\widehat{Q}^{O}$ is equal to 1 in time-frequency points where speech is not active. Comparing the estimators in Fig. 6.2(d)-(h) with the oracle estimate, it can be seen that while $\widehat{Q}^{\mathrm{MS,pos}}$ and $\widehat{Q}^{\mathrm{SD,pos}}$ clearly overestimates Q, especially in noise-only time-frequency points, the classical MDL estimator $\widehat{Q}^{\mathrm{MDL}}$ clearly underestimates Q. Although it is not depicted, using a fixed value $\widehat{Q} = 1$ for each time-frequency point (cf. Section 6.2.1) will clearly underestimates Q, especially in speech-and-noise time-frequency points. Incorporating the a-priori SNR into the MDL selection criterion leads to a more accurate estimate than $\widehat{Q}^{\mathrm{MDL}}$, $\widehat{Q}^{\mathrm{MS,pos}}$ and $\widehat{Q}^{\mathrm{SD,pos}}$. Moreover, while for voiced sounds the proposed estimator $\widehat{Q}^{\mathrm{MDL}_\xi}$ leads to a similar estimation accuracy as the a-priori SNR-based hypothesis testing estimate $\widehat{Q}^{\mathrm{Thre}}$, for unvoiced sounds $\widehat{Q}^{\mathrm{MDL}_\xi}$ is more accurate than $\widehat{Q}^{\mathrm{Thre}}$.

In conclusion, these results show that the a-priori SNR-based speech model order estimators $\widehat{Q}^{\mathrm{Thre}}$ and $\widehat{Q}^{\mathrm{MDL}_\xi}$ lead to more accurate estimates than $\widehat{Q}^{\mathrm{MDL}}$, $\widehat{Q}^{\mathrm{MS,pos}}$, $\widehat{Q}^{\mathrm{SD,pos}}$ or $\widehat{Q} = 1$, with the $\widehat{Q}^{\mathrm{MDL}_\xi}$ estimator achieving the highest estimation accuracy.

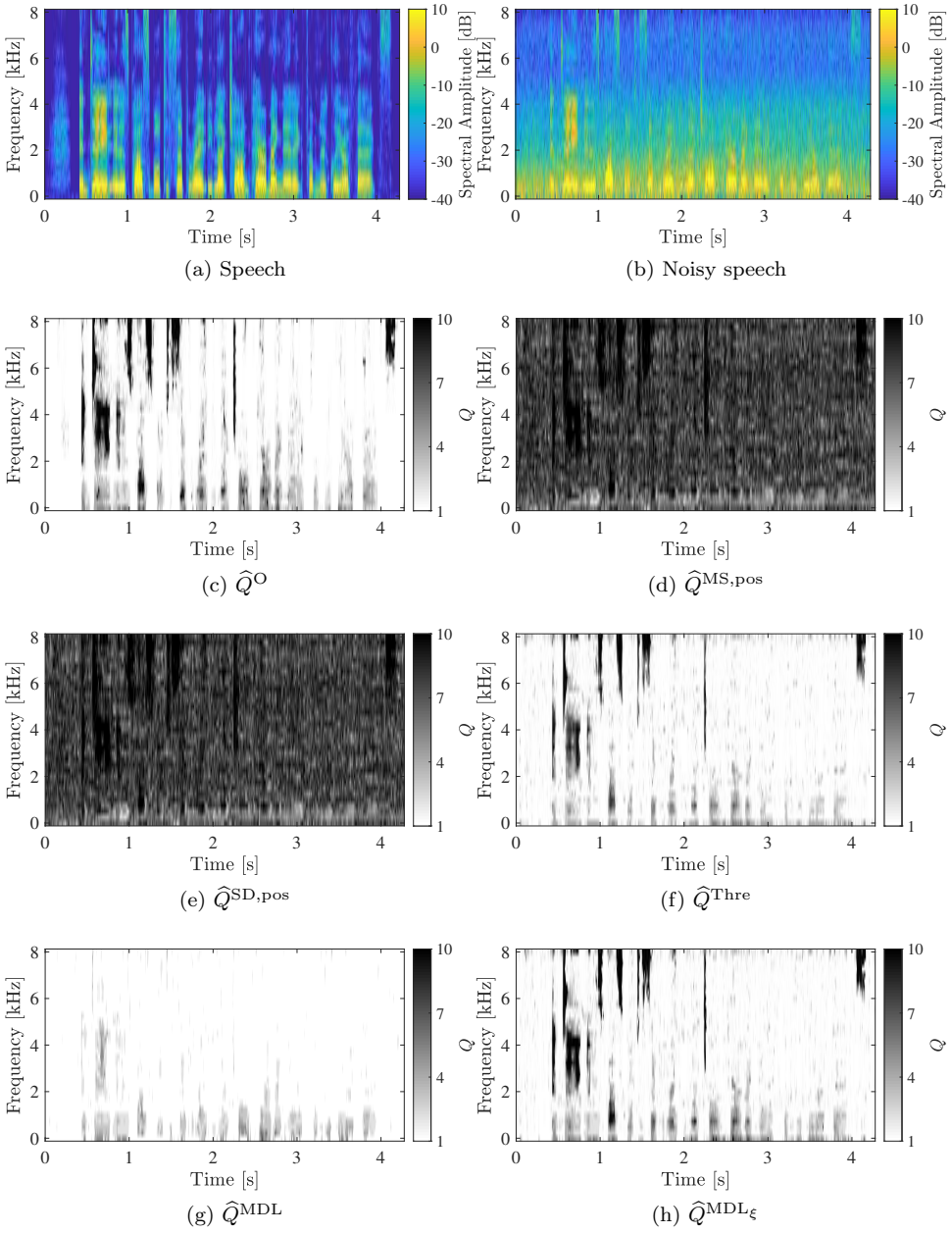


Fig. 6.2: Spectrograms of (a) speech signal and (b) noisy speech signal, corrupted by traffic noise at 5 dB SNR. Estimated speech model order Q using (c) the oracle estimate $\widehat{Q}^{O}$, (d) $\widehat{Q}^{\mathrm{MS,pos}}$ in (6.27), (e) $\widehat{Q}^{\mathrm{SD,pos}}$ in (6.28), (f) the a-priori SNR-based threshold estimator $\widehat{Q}^{\mathrm{Thre}}$ in (6.29), (g) the classical MDL estimator $\widehat{Q}^{\mathrm{MDL}}$ in (6.47), (h) the proposed a-priori SNR-based MDL estimator $\widehat{Q}^{\mathrm{MDL}_{\xi}}$ in (6.56).

Table 6.1: Overview of the considered normalized speech correlation vector estimators.

Label	Description
$\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{ML}}$	State-of-the-art vector-based ML estimator in (4.12)
$\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{MS},1}$	Matrix-based estimator using the MS method in (6.26), i.e., with $\widehat{Q} = 1$
$\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{MS,pos}}$	Matrix-based estimator using the MS method in (6.7) with $\widehat{Q}^{\mathrm{MS,pos}}$ in (6.27)
$\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SD},1}$	Matrix-based estimator using the SD method in (6.25), i.e., with $\widehat{Q} = 1$
$\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SD,pos}}$	Matrix-based estimator using the SD method in (6.22) with $\widehat{Q}^{\mathrm{SD,pos}}$ in (6.28)
$\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SD,MDL}_\xi}$	Matrix-based estimator using the SD method in (6.22) with $\widehat{Q}^{\mathrm{MDL}_\xi}$ in (6.56)

6.3.3 *Normalized Speech Correlation Vector Estimation Performance*

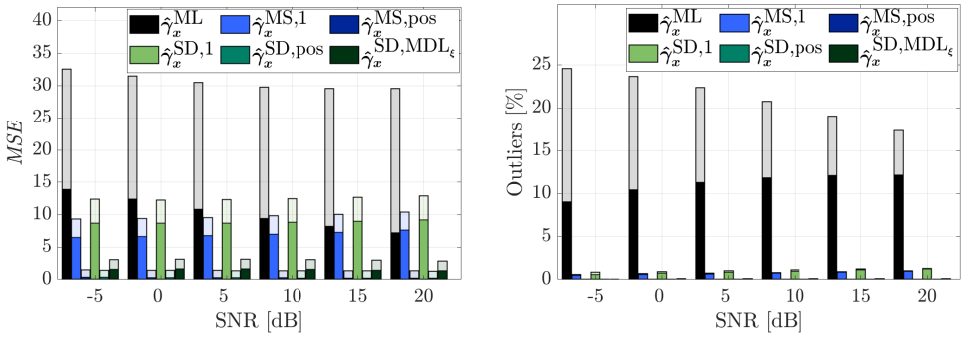
In this section, we compare the estimation accuracy between the proposed matrix-based normalized speech correlation vector estimates and the state-of-the-art vector-based ML estimate. A detailed overview of all considered estimators is given in Table 6.1.

For different SNRs, Fig. 6.3 depicts the performance, averaged over all combinations of speech and noise signals, in terms of the normalized *MSE* (cf. (5.41)) and the percentage of outliers. The lower and upper parts of the bars represent the performance in speech-and-noise and noise-only, respectively. It can be observed that all matrix-based estimates achieve a considerably lower *MSE* with clearly less outliers than the ML estimate. It should however be noted that for SNRs larger than 10 dB, the *MSE* for $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{MS},1}$ and $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SD},1}$ is larger in speech-and-noise time-frequency points than for $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{ML}}$, where $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SD},1}$ shows the largest *MSE*. The matrix-based estimates $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{MS,pos}}$ and $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SD,pos}}$ achieve the smallest *MSE*, both in speech-and-noise and noise-only time-frequency points, although the *MSE* for $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{MDL}_\xi}$ is also very small.

6.3.4 *Speech Enhancement Performance*

In this section, we compare the speech enhancement performance of the proposed MS-MFMPDR and SD-MFMPDR filters using the speech model order estimators presented in Section 6.2 with the ML-MFMPDR filter and the single-frame WG.

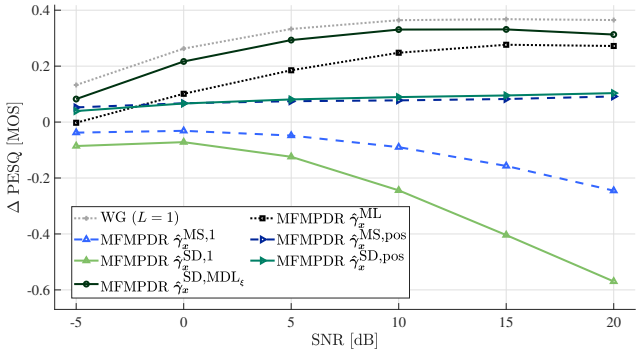
For different SNRs, Fig. 6.4 depicts the performance, averaged over all combinations of speech and noise signals, in terms of speech quality using ΔPESQ, in terms



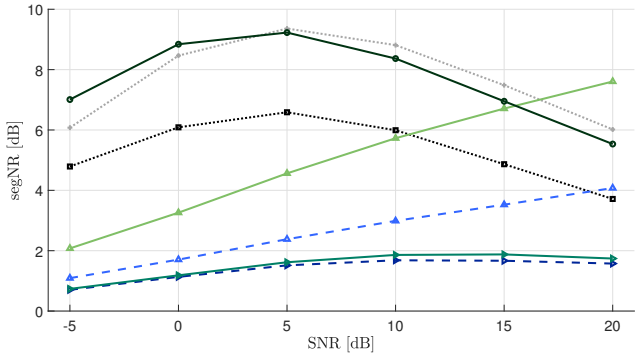
(a) (b)

Fig. 6.3: Average *MSE* in (a) and percentage of outliers in (b) for the ML and the proposed matrix-based normalized speech correlation vector estimates (cf. Table 6.1). The lower and upper parts of the bars represent the performance in speech-and-noise and noise-only, respectively.

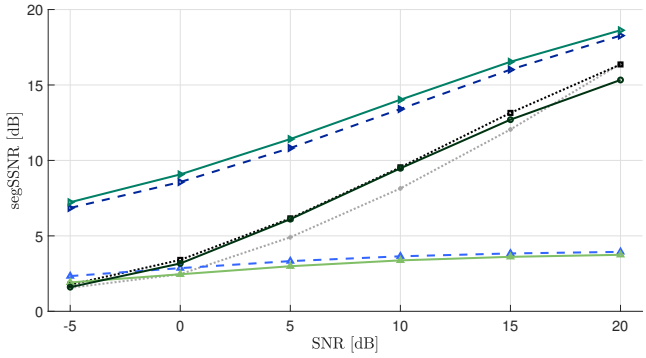
of noise reduction using segNR and in terms of speech distortion using segSSNR. First, it can be observed that neglecting the uncorrelated speech component, i.e., $\widehat{Q} = 1$ in the MS-MFMPDR and SD-MFMPDR filters yields poor results in terms of all considered performance measures for all SNRs. Although the segNR increases with increasing SNR, the ΔPESQ results are negative and strongly decrease with increasing SNR. Assuming a rank-1 speech correlation matrix, i.e., neglecting the uncorrelated speech component and setting $Q = 1$, clearly underestimates Q, especially in speech-and-noise time-frequency points (cf. Section 6.3.2), leading to a strongly degraded speech enhancement performance, especially at high SNRs. This confirms the results in Chapter 4, where it was shown that the influence of the uncorrelated speech component is crucial. Second, it can be observed that the MS-MFMPDR and SD-MFMPDR filters using $\widehat{Q}^{\mathrm{MS,pos}}$ and $\widehat{Q}^{\mathrm{SD,pos}}$ lead to rather conservative results in terms of ΔPESQ and segNR, but outperform all other filters in terms of segSSNR, i.e., less speech distortion. This can be explained by the fact that $\widehat{Q}^{\mathrm{MS,pos}}$ and $\widehat{Q}^{\mathrm{SD,pos}}$ overestimate Q, especially in noise-only time-frequency points (see Section 6.3.2). Third, it can be observed that the proposed SD-MFMPDR filter using the a-priori SNR-based MDL speech model order estimator $\widehat{Q}^{\mathrm{SD,MDL}_\xi}$ outperforms all other filters in terms of segNR for SNRs up to 5 dB, while the WG is better for SNRs larger than 5 dB. Compared to the state-of-the-art ML-MFMPDR filter, the proposed SD-MFMPDR filter using $\widehat{Q}^{\mathrm{SD,MDL}_\xi}$ increases the segNR results by 2 dB for all SNRs. In terms of segSSNR, the proposed SD-MFMPDR filter using $\widehat{Q}^{\mathrm{SD,MDL}_\xi}$ leads to similar results as the ML-MFMPDR filter and to slightly better results as the WG. In terms of ΔPESQ the WG outperforms all MFMPDR filters for all SNRs, but the proposed SD-MFMPDR filter using $\widehat{Q}^{\mathrm{SD,MDL}_\xi}$ clearly outperforms all other MFMPDR filters. For instance, for an SNR of 5 dB the PESQ improvement of the SD-MFMPDR filter using $\widehat{Q}^{\mathrm{SD,MDL}_\xi}$ is 0.03 MOS lower compared to the WG, but 0.11 MOS higher compared to the ML-MFMPDR filter.



(a)



(b)



(c)

Fig. 6.4: Average (a) PESQ improvement, (b) segNR and (c) segSSNR for the WG and for the MFMPDR filters using the state-of-the-art vector-based ML estimate and the proposed matrix-based methods to estimate the normalized speech correlation vector (see Table 6.1) for different SNRs.

These results show that the performance of the MFMPDR filter can be improved by estimating the normalized speech correlation vector using matrix-based methods based on a low-rank speech model compared to using vector-based methods (e.g., the ML estimate). However, as expected, the performance strongly depends on the estimation accuracy of the speech model order Q. While using a time- and frequency-independent value of $\widehat{Q} = 1$ leads to a large amount of speech distortion and a poor speech quality, using a time- and frequency-dependent estimate of Q leads to a conservative noise reduction performance with a low amount of speech distortion. Using the proposed SD estimator with the a-priori SNR-based MDL selection criterion leads to a good noise reduction performance while keeping speech distortion low, resulting in a higher speech quality for all SNRs than the state-of-the-art ML-MFMPDR filter. Although the speech quality of this proposed matrix-based estimator is slightly lower than the single-frame WG, this proposed estimator leads to a similar amount of noise reduction and less speech distortion.

6.4 Summary

Assuming that speech signals can be modeled using a low-rank model, e.g., as a linear combination of a limited number of complex exponentials, the speech correlation matrix can be assumed to be rank-deficient. Based on this low-rank speech model, in this Chapter we proposed two matrix-based methods to estimate the normalized speech correlation vector for the single-microphone MFMPDR filter, namely the matrix-subtraction (MS) method and the subspace-decomposition (SD) method. Both methods are based on the EVD of a matrix, which is either constructed by subtracting the estimated normalized noise correlation matrix from the estimated normalized noisy speech correlation matrix or by prewhitening the estimated normalized noisy speech correlation matrix with the estimated normalized noise correlation matrix. For both methods, an estimate of the normalized noise correlation matrix and the speech model order is required. For the normalized noise correlation matrix, we used an estimate of the mean normalized noise correlation matrix. For the speech model order, we either used a time- and frequency-independent value, or proposed a time- and frequency-dependent estimator that incorporates the a-priori SNR estimate into the MDL selection criterion. Simulation results for different speech signals, noise types and SNRs show that the proposed matrix-based methods yield a more accurate estimate of the normalized speech correlation vector than the vector-based ML estimator. An instrumental evaluation indicates that the SD-MFMPDR filter using the proposed a-priori SNR-based MDL estimator leads to a better speech quality and more noise reduction than the state-of-the-art ML-MFMPDR filter, while keeping speech distortion low. Using this estimator, the SD-MFMPDR filter also leads to less speech distortion and a similar noise reduction performance than the single-frame WG. To verify the instrumental performance results, in the next chapter we will perform a subjective listening test using the most promising MFMPDR filters from Chapters 4, 5 and 6, an oracle MFMVDR filter and oracle and blind single-frame WGs.

7

INSTRUMENTAL AND PERCEPTUAL EVALUATION OF MFMVDR AND MFMPDR FILTERS

In Chapter 4, we investigated the speech enhancement performance of two oracle versions of the MFMVDR filter and the practically feasible MFMPDR filter using different oracle and blind estimates of the normalized speech correlation vector, e.g., using the state-of-the-art blind ML estimate. Based on instrumental performance measures, simulation results showed that using quasi-perfect estimates both the MFMVDR and MFMPDR filters are able to achieve impressive speech quality improvement. However, it was also shown that even small estimation errors in the normalized speech correlation vector lead to a degraded performance for both multi-frame filters. Aiming at improving the speech enhancement performance of the MFMPDR filter, in Chapter 5 and 6 we proposed several novel normalized speech correlation vector estimators, either based on robust constrained estimation approaches or a low-rank speech model. Simulation results showed that the proposed estimators, more in particular the doubly-constrained (DC) estimator form Chapter 5 and the subspace-decomposition (SD) estimator from Chapter 6, yield more accurate estimates of the normalized speech correlation vector than the ML estimator. Although the proposed DC-MFMPDR filter produces less speech and noise distortions than the ML-MFMPDR filter, leading to a more natural speech quality and to less artifacts in the background noise, it also leads to a more conservative noise reduction. In addition, the SD-MFMPDR filter leads to more noise reduction and a better speech quality than the ML-MFMPDR filter, while keeping speech distortion low. Nevertheless, although the speech distortion for the DC-MFMPDR and SD-MFMPDR filters is lower than for the single-frame Wiener gain (WG), the results also indicated that the overall speech quality is slightly better for the single-frame WG. Aiming at determining the most effective estimator for the normalized speech correlation vector and the most perceptually advantageous speech enhancement algorithm, in this chapter we compare the performance of the most promising MFMPDR filters from Chapters 4, 5 and 6 with an oracle MFMVDR filter and the single-frame WG. Both for the MFMPDR filters as well as for the WG we consider oracle and blind versions. Since instrumental performance measures only provide an indication on how the quality of the processed signals is perceived by humans, we also conduct a subjective listening test.

In Section 7.1, we discuss the algorithmic settings for the (oracle) MFMVDR filter, the (oracle and blind) MFMPDR filters and the (oracle and blind) WGs. For several speech signals, noise types and SNRs, in Section 7.2 we evaluate the speech enhancement performance of the considered algorithms using instrumental performance measures. For the oracle algorithms, the results show that the MFMVDR filter results in a better speech enhancement performance than the WG and the MFMPDR filter. For the blind algorithms, the results indicate that the SD-MFMPDR filter leads to a better speech quality and more noise reduction than the DC-MFMPDR filter and the ML-MFMPDR filter, while keeping speech distortion as low as the ML-MFMPDR filter. For two speech signals and two noise types, in Section 7.3, we perceptually evaluate the considered algorithms in terms of overall quality, speech distortion and noise reduction using a procedure similar to the multi stimulus test with hidden reference and anchor (MUSHRA) [158]. For the oracle algorithms, the perceptual evaluation results show that in terms of overall quality the MFMVDR filter is rated significantly better than the WG and the MFMPDR filter, while in terms of speech distortion the MFMVDR filter is rated significantly better than the WG but shows no significant differences compared to the MFMPDR filter. For the blind algorithms, the perceptual evaluation results show that in terms of overall quality the SD-MFMPDR filter and the DC-MFMPDR filter are rated significantly better than the ML-MFMPDR filter, but show no significant differences compared to the WG. In terms of noise reduction the SD-MFMPDR filter and the WG are rated better than the DC-MFMPDR filter and the ML-MFMPDR filter, while in terms of speech distortion all MFMPDR filters are significantly better rated than the WG.

7.1 Algorithmic Settings

Table 7.1 gives an overview of the considered speech enhancement algorithms: an oracle MFMVDR filter, several MFMPDR filters, either using oracle or blind estimates of the normalized speech correlation vector, and two single-frame WGs, either using oracle or blind estimates of the a-priori SNR and the noise PSD. For the MFMVDR and MFMPDR filters, we used a highly temporally resolved STFT framework at a sampling frequency of 16 kHz with a frame length of 4 ms ($T = K = 64$ samples) and an overlap of 75 %, resulting in a frame shift of 1 ms ($R = 16$ samples). As the STFT analysis window $w_a(t)$ and synthesis window $w_s(t)$, we used a square-root Hann window. Similarly as in Chapter 6, to provide a good compromise between speech quality improvement, noise reduction and speech distortion, the number of consecutive time-frames was set to $L = 10$, resulting in 13 ms of analysis data used in each filtering operation. The noise correlation matrix $\boldsymbol{R}_{\boldsymbol{n}}(k, m)$ and the noisy speech correlation matrix $\boldsymbol{R}_{\boldsymbol{y}}(k, m)$, will estimated using recursive smoothing as in (4.2) and (4.3), respectively. The smoothing parameter were experimentally set to $\alpha_n = 0.88$, corresponding to a smoothing window of 8 ms, and $\alpha_y = 0.90$, corresponding to a smoothing window of 10 ms. Accordingly, for the SD-MFMPDR filter the data length N for the a-priori SNR-based MDL estimator in (6.55) was set to $N = 6$. To avoid numerical problems for the MFMVDR and MFMPDR filters, we performed diagonal loading as in (4.17) before computing the inverse of the

Table 7.1: Overview of the considered MFMVDR filter, MFMPDR filters and WGs.

Label	**Description**
WG_O	Oracle Wiener gain
	- Estimate $\xi(k,m)$ using $\widehat{\xi}^{\text{O}}(k,m)$ in (4.8)
	- Estimate $\phi_N(k,m)$ based on (4.2)
MFMVDR_O	Oracle MFMVDR filter
	- Estimate $\boldsymbol{R}_{\boldsymbol{u}}(k,m)$ using $\widehat{\boldsymbol{R}}^{\text{I}}_{\boldsymbol{u}}(k,m)$ in (4.15)
	- Estimate $\boldsymbol{\gamma}_{\boldsymbol{x}}(k,m)$ using $\widehat{\boldsymbol{\gamma}}^{\text{II}}_{\boldsymbol{x}}(k,m)$ in (4.7)
MFMPDR_O	Oracle MFMPDR filter
	- Estimate $\boldsymbol{R}_{\boldsymbol{y}}(k,m)$ using $\widehat{\boldsymbol{R}}_{\boldsymbol{y}}(k,m)$ in (4.3)
	- Estimate $\boldsymbol{\gamma}_{\boldsymbol{x}}(k,m)$ using $\widehat{\boldsymbol{\gamma}}^{\text{II}}_{\boldsymbol{x}}(k,m)$ in (4.7)
WG	Blind Wiener gain
	- Estimate $\xi(k,m)$ using $\widehat{\xi}^{\text{DDA}}(k,m)$ in (3.9)
	- Estimate $\phi_N(k,m)$ using $\widehat{\phi}^{\text{SPP}}_N(k,m)$ in (3.13)
ML-MFMPDR	Blind (state-of-the-art) MFMPDR filter
	- Estimate $\boldsymbol{R}_{\boldsymbol{y}}(k,m)$ using $\widehat{\boldsymbol{R}}_{\boldsymbol{y}}(k,m)$ in (4.3)
	- Estimate $\boldsymbol{\gamma}_{\boldsymbol{x}}(k,m)$ using $\widehat{\boldsymbol{\gamma}}^{\text{ML}}_{\boldsymbol{x}}(k,m)$ in (4.12)
DC-MFMPDR	Blind MFMPDR filter
	- Estimate $\boldsymbol{R}_{\boldsymbol{y}}(k,m)$ using $\widehat{\boldsymbol{R}}_{\boldsymbol{y}}(k,m)$ in (4.3)
	- Estimate $\boldsymbol{\gamma}_{\boldsymbol{x}}(k,m)$ using $\widehat{\boldsymbol{\gamma}}^{\text{DC}}_{\boldsymbol{x}}(k,m)$ in (5.18)
SD-MFMPDR	Blind MFMPDR filter
	- Estimate $\boldsymbol{R}_{\boldsymbol{y}}(k,m)$ using $\widehat{\boldsymbol{R}}_{\boldsymbol{y}}(k,m)$ in (4.3)
	- Estimate $\boldsymbol{\gamma}_{\boldsymbol{x}}(k,m)$ using $\widehat{\boldsymbol{\gamma}}^{\text{SD}}_{\boldsymbol{x}}(k,m)$ in (6.22) with $\widehat{Q}^{\text{MDL}_\xi}(k,m)$ in (6.56)

undesired or noisy speech correlation matrix with a scaling parameter of $\kappa = 0.001$, similarly as in Chapter 4. For the oracle MFMVDR and MFMPDR filters, the oracle a-priori SNR was estimated using $\widehat{\xi}^{\text{O}}(k,m)$ in (4.8). For all blind MFMPDR filters, the a-priori SNR was estimated using the DDA estimate $\widehat{\xi}^{\text{DDA}}(k,m)$ in (3.9), with a weighting parameter $\alpha_{\text{DDA}} = 0.97$. To reduce unpleasant artifacts in the background noise, e.g., musical noise, we only updated the estimated speech STFT coefficient $\widehat{X}(k,m-1)$ in (3.9) every 4 ms, i.e., every 4 frames. The noise PSD was estimated using the SPP-based noise PSD estimator $\widehat{\phi}^{\text{SPP}}_N(k,m)$ in (3.13), with a smoothing parameter $\alpha_{\text{SPP}} = 0.90$ and a fixed SNR $\xi_{\mathcal{H}_1} = -15\,\text{dB}$. To reduce the amount of speech distortion and to mask artifacts in the background noise, we applied a lower limit of $\xi_{\min} = -6.5\,\text{dB}$ to the a-priori SNR estimate.

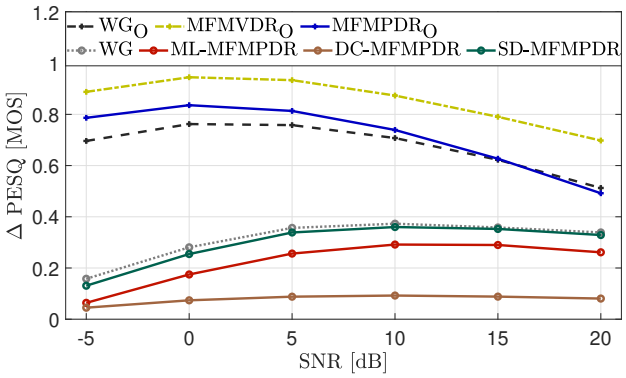
For a fair comparison, the single-frame WGs were implemented using an equivalent frame length of 13 ms and an overlap of 50 %. For the oracle WG, the oracle a-priori SNR was estimated similarly to $\widehat{\xi}^{\mathrm{O}}(k,m)$ in (4.8). The noisy speech PSD and the oracle noise PSD were estimated using recursive smoothing, with the smoothing parameters α_y and α_n, corresponding to a smoothing window of 10 ms and 8 ms, respectively. For the blind WG, the a-priori SNR and the noise PSD were estimated in the same way as for the blind MFMPDR filters, except for the estimated speech STFT coefficient $\widehat{X}(k,m-1)$ in (3.9), which was updated in each frame.

7.2 Instrumental Speech Enhancement Performance

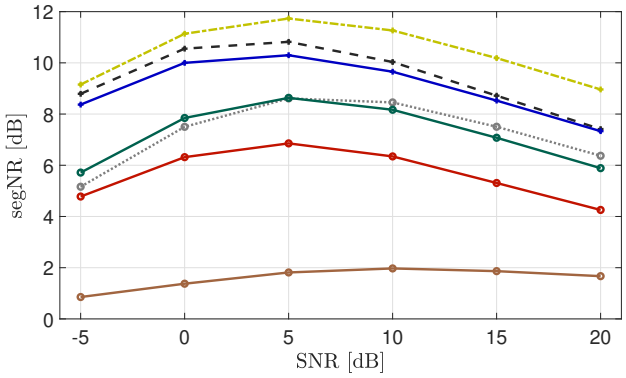
In this section, we compare the speech enhancement performance of the considered algorithms (see Table 7.1) in terms of speech quality using the PESQ improvement (ΔPESQ), in terms of noise reduction using the segmental noise reduction (segNR) measure in (2.32), and in terms of speech distortion, using the segmental speech distortion (segSSNR) measure in (2.33). For the evaluation using instrumental performance measures, we used 260 s of speech material (131 s female, 129 s male) from the TIMIT database [146] as clean speech signals. As noise signals, we used traffic noise, two babble noise signals, factory noise and modulated white Gaussian noise with a modulation frequency of 0.5 Hz taken from the NOISEX-92 database [145]. The considered SNRs ranged from -5 dB to 20 dB in 5 dB steps.

For different SNRs, Fig. 7.1 depicts the performance of the considered algorithms averaged over all considered speech signals and noise types. First, it can be observed that the oracle MFMVDR filter clearly outperforms all other filters in terms of all instrumental performance measures. Second, it can be observed that the oracle MFMPDR filter achieves smaller segNR values (i.e., less noise reduction) but larger segSSNR values (i.e., less speech distortion) and larger ΔPESQ scores than the oracle WG and all blind speech enhancement algorithms (except at SNR = 20 dB). Third, for the blind algorithms it can be observed that in terms of ΔPESQ and segNR the SD-MFMPDR filter consistently outperforms the state-of-the-art ML-MFMPDR filter and the DC-MFMPDR filter and yields similar results as the WG. In terms of segSSNR, it can be observed that the DC-MFMPDR filter consistently outperforms all other blind algorithms, while the SD-MFMPDR filter leads to similar results as the ML-MFMPDR filter and slightly larger segSSNR values than the WG.

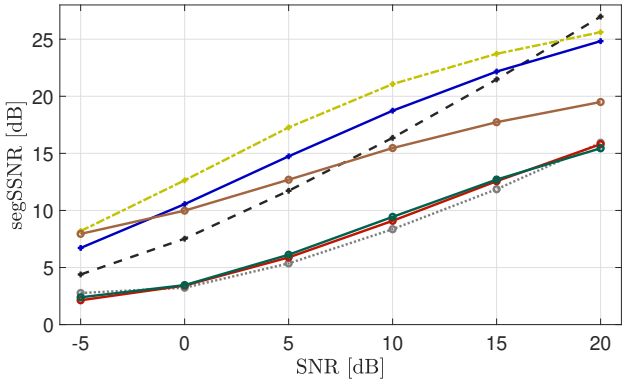
For the oracle algorithms, these results indicate that the MFMVDR and MFMPDR filters yield a better speech quality and lower speech distortion than the WG. For the blind algorithms, these results indicate that the SD-MFMPDR filter yields a clearly better speech quality than the DC-MFMPDR filter and the state-of-the-art ML-MFMPDR filter, while yielding a slightly worse speech quality but slightly lower speech distortion than the WG.



(a)



(b)



(c)

Fig. 7.1: Performance of the considered algorithms (see Table 7.1) in terms of (a) PESQ improvement, (b) segmental noise reduction and (c) segmental speech distortion, averaged over all speech and noise signals for different SNRs.

7.3 Perceptual Speech Enhancement Performance

In this section, we perceptually compare the speech enhancement performance of the considered algorithms using a subjective listening test. For two speech signals and two acoustic scenarios (noisy types), we conducted a procedure similar to the MUSHRA, evaluating three attributes: (a) overall quality, (b) speech distortion and (c) noise reduction. For the attribute (a), the participants were asked to rate the overall signal quality of the test signals with respect to a reference signal. For the attribute (b), the participants were asked to rate how distorted the speech component of the test signals sounds with respect to a reference signal. For the attribute (c), the participants were asked to rate how noticeable the amount of noise reduction of the test signals is with respect to a reference signal. As speech signals, we used two sentences from the TIMIT database [146], spoken by a male and a female speaker. To generate both acoustic scenarios, we mixed the speech signals with traffic noise and babble noise taken from the NOISEX-92 database [145] at 5 dB SNR. For each attribute, acoustic scenario and speech signal, in addition to the signals processed with the considered speech enhancement algorithms in Table 7.1 a noisy speech signal, a hidden reference and an anchor were presented to the participants. For the attributes (a) and (b), the (hidden) reference was the noisy speech signal at 20 dB SNR. The anchor was the speech signal low-pass filtered at 4 kHz, corrupted with the noise at -5 dB and processed with an aggressive WG. For the WG, the a-priori SNR was estimated using the DDA $\widehat{\xi}^{\mathrm{DDA}}(k,m)$ in (3.9) with a weighting parameter $\alpha_{\mathrm{DDA}} = 0.90$, and the noise PSD was estimated using the SPP-based noise PSD estimator $\widehat{\phi}_N^{\mathrm{SPP}}(k,m)$ in (3.13) with $\alpha_{\mathrm{SPP}} = 0.90$ and $\xi_{\mathcal{H}_1} = -15\,\mathrm{dB}$. To achieve a more aggressive performance, we applied a lower limit $G_{\min} = -20\,\mathrm{dB}$ to the WG. For the attribute (c), the (hidden) reference was the unprocessed noisy speech signal at 5 dB SNR, while the anchor was a noisy speech signal at 20 dB SNR. For the sake of completeness, for this attribute we also presented the anchor for the attributes (a) and (b) as a test signal. Hence, for each attribute, acoustic scenario and speech signal, the participants compared ten test signals with a reference signal, i.e., either a noisy speech signal at 20 dB SNR for the attributes (a) and (b) or the unprocessed noisy speech signal at 5 dB SNR for the attribute (c).

A total of 11 self-reported normal-hearing participants in the range of 21 to 36 years participated in the subjective listening test. Due to the COVID-19 pandemic, the experiment took place in quiet rooms at the participants' home. The signals were presented diotically to the participants, using their own sound cards and over-the-ear headphones. Tab. 7.2 lists the used headphones and sound cards for each participant. All signals were normalized in amplitude.

The procedure similar to the MUSHRA consisted of two phases. First, the participants were trained to familiarize themselves with the presented signals and to adjust the volume to a comfortable level. Second, the participants were instructed to rate the test signals according to the three aforementioned attributes on a continuous scale from 0 to 100 using sliders in a graphical user interface. For the attribute (a) overall quality, 0 was labeled with "bad" and 100 with "excellent", while for the attribute (b) speech distortion, 0 was labeled with "extremely distorted" and 100 with "not distorted", and for the attribute (c) noise reduction, 0 was labeled with

Table 7.2: Overview of the sound cards and headphones per participant.

Participant	Sound Card	Headphone
1	RME ADI-2 Pro fs	Sennheiser HDA 200
2	RME ADI-2 Pro fs	Sennheiser HDA 200
3	RME ADI-2 Pro fs	Sennheiser HDA 200
4	Onboad - Realtek HD Audio	Trust GTX26
5	RME Fireface UCX	Audio Technica ATH-M50
6	RME Babyface	Sony MDR-7506
7	Onboad - Realtek HD Audio	Sennheiser HD6xx
8	Onboad - Realtek HD Audio	Sony MDR-7506
9	RME Fireface UCX	Sony MDR-7506
10	Onboad - Realtek HD Audio	Sony MDR-7506
11	MacBook Pro Onboard	Bose QuietComfort 35 ii

"extremely noticeable" and 100 with "not noticeable". The participants were allowed to listen to the reference signal and all test signals as often as they wanted. The participants were instructed to rate at least one test signal with a score of 100, which should correspond to the hidden reference. The order of the presentation of the test signals and acoustic scenarios were randomized between all subjects.

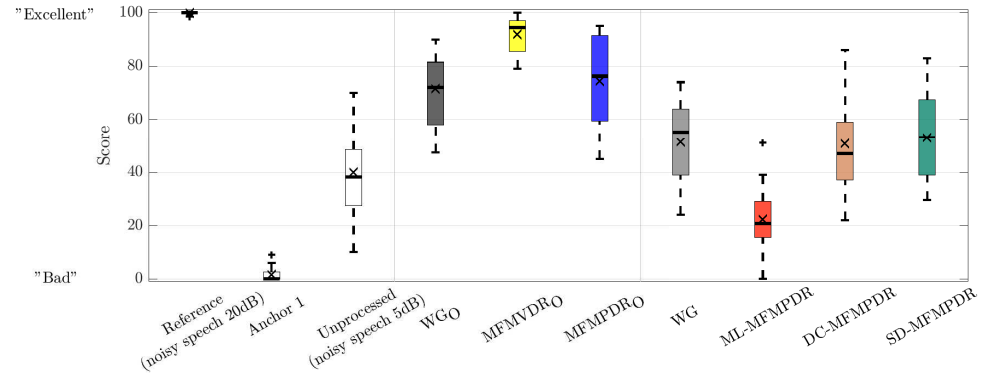
For each attribute, a statistical analysis was conducted using the resulting MUSHRA scores of both speech signals and both acoustic scenarios. Since the data are normally distributed, as shown by the Shapiro-Wilk test [159], a repeated-measures analysis of variance (ANOVA) [160] was performed with factors "acoustic scenario" and "algorithm". First, we tested if the factor "acoustic scenario" has a significant influence. The ANOVA results for the three attributes are (a) (overall quality) $F(1;218) = 0.22, p > 0.05$, (b) (speech distortion) $F(1;218) = 0.91, p > 0.05$, and (c) (noise reduction) $F(1;218) = 1.28, p > 0.05$. Hence, since the statistical analysis showed no significant influence of the factor "acoustic scenario" for all attributes, we averaged the MUSHRA scores over both acoustic scenarios. Second, using these averaged MUSHRA scores we tested if the factor "algorithm" has a significant influence. Since Mauchly's test [161] indicated a violation of sphericity for all three attributes, a Greenhouse-Geisser correction [162] was applied. The ANOVA results for the three attributes are (a) (overall quality) $F(10;100) = 35.92, p < 0.001$, (b) (speech distortion) $F(10;100) = 7.71, p < 0.001$, and (c) (noise reduction) $F(10;100) = 42.31, p < 0.001$. Since the statistical analysis showed a significant influence of the factor "algorithm" for all attributes, we tested for statistically significant differences between the algorithm mean values, by conducting a post-hoc pairwise comparison t-test with Bonferroni correction [150]. Fig. 7.2 depicts the averaged MUSHRA scores for all three attributes using boxplots. The t-test results

are presented in Tables 7.3-7.5, with asterisks denoting statistically significant differences and o denoting not statistically significant differences.

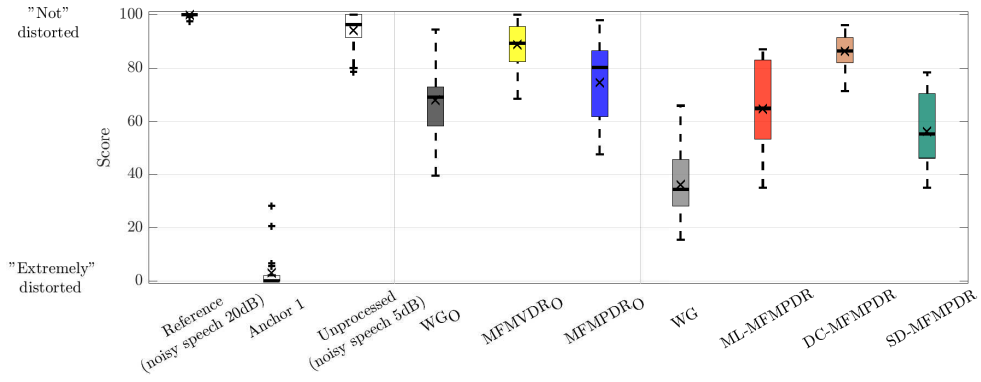
In terms of the attribute (a) overall quality (cf. Fig. 7.2a and Table 7.3), the mean score for the hidden reference was equal to 100, showing that the participants were able to distinguish the hidden reference from the other test signals. As desired, the anchor was rated with the lowest mean score of 1.6. The mean score for the unprocessed noisy speech signal at 5 dB SNR was equal to 39.9, which is significantly lower than for all processed signals, except for the ML-MFMPDR filter. For the oracle WG, the oracle MFMVDR filter and the oracle MFMPDR filter, the mean score was equal to 71.6, 91.82 and 74.5, respectively. For the WG, the DC-MFMPDR filter and the SD-MFMPDR filter, the mean score was equal to 51.5, 50.1 and 54.5, respectively. The statistical analysis showed that all oracle algorithms were rated significantly higher than all blind algorithms, where the oracle MFMVDR filter was rated significantly higher than the oracle MFMPDR filter and the oracle WG. While the differences between the WG, the DC-MFMPDR filter and the SD-MFMPDR filter are not statistically significant, these mean scores are significantly higher than the scores of the state-of-the-art ML-MFMPDR filter and the unprocessed noisy speech signal.

In terms of the attribute (b) speech distortion (cf. Fig. 7.2b and Table 7.4), the mean score of the hidden reference was equal to 100 and the anchor was rated with the lowest mean score of 2.9, as desired. The unprocessed noisy speech signal at 5 dB SNR was rated with the highest mean score of 94.2, followed by the oracle MFMVDR filter with a mean score of 88.9 and the DC-MFMPDR filter with a mean score of 86.5. These differences are not statistically significant, even not with the reference. For the oracle WG and the oracle MFMPDR filter, the mean score was equal to 68.1 and 74.2, respectively. While the oracle MFMVDR filter was rated significantly higher than the oracle WG, the differences between the oracle MFMVDR filter, the oracle MFMPDR and the blind DC-MFMPDR filter were not statistically significant. For the ML-MFMPDR filter and the SD-MFMPDR filter, the mean score was equal to 64.7 and 56.1, respectively, but this difference was not statistically significant. All blind MFMPDR filters were rated significantly higher than the blind WG, with a mean score of 35.9.

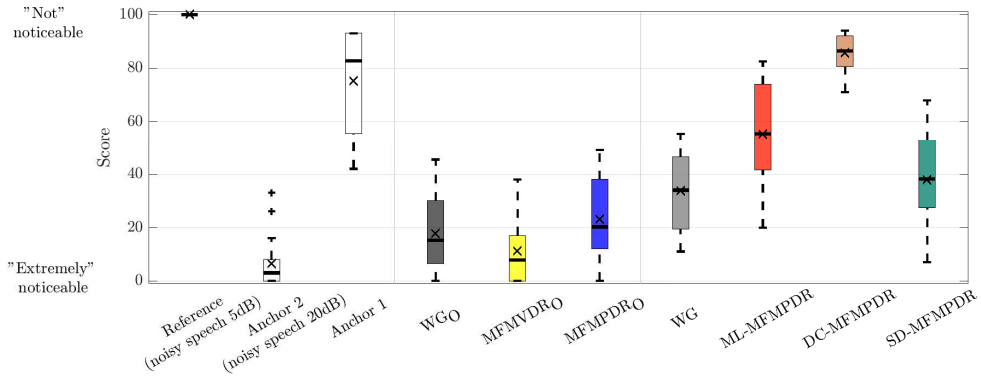
In terms of the attribute (c) noise reduction (cf. Fig. 7.2c and Table 7.5), the mean score of the hidden reference (in this case the unprocessed noisy speech signal at 5 dB) was equal to 100 and the anchor (in this case the noisy speech signal at 20 dB SNR) was rated with the lowest mean score of 6.5, as desired. For the oracle WG, the oracle MFMVDR filter and the oracle MFMPDR filter, the mean score was equal to 17.8, 11.2 and 23.5, respectively, but these differences were not statistically significant. For the WG, the ML-MFMPDR filter, the DC-MFMPDR filter and the SD-MFMPDR filter, the mean score was equal to 33.5, 39.0, 86.1 and 55.3, respectively. All blind algorithms were rated significantly worse than the oracle algorithms, except for the difference between the WG and the oracle MFMPDR filter. Nevertheless, the difference between the WG and the SD-MFMPDR filter is not statistically significant and the WG and the SD-MFMPDR filter are rated significantly better than the ML-MFMPDR filter and the DC-MFMPDR filter.



(a)



(b)



(c)

Fig. 7.2: Averaged MUSHRA scores for the attributes (a) overall quality, (b) speech distortion and (c) noise reduction, for a hidden reference, an anchor, a noisy speech signal and the processed signals using an oracle MFMVDR filter and blind and oracle MFMPDR filters and WGs (see Table 7.1). On each box, the central horizontal line is the median, the edges of the box are the 25-th and 75-th percentiles and the whiskers extend to 1.5 times the interquartile range from the median. The means are indicated by $\times$ markers. Outliers are indicated by $+$ markers.

For the oracle algorithms, the results of the subjective listening test show that the perceived overall quality and speech distortion for the oracle MFMVDR filter are significantly better than for the oracle WG, while the perceived amount of noise reduction is similar. These results confirm the results of the instrumental evaluation in Section 7.2 and confirm the potential of multi-frame speech enhancement algorithms, namely that exploiting speech interframe correlation enables to keep speech distortion low while suppressing undesired background noise. Nevertheless, it should be noted that the differences between the oracle MFMPDR filter and the oracle WG in terms of perceived overall quality, speech distortion and noise reduction are not statistically significant.

For the blind algorithms, the results of the instrumental evaluation in Section 7.2 indicated that the overall quality for the DC-MFMPDR filter is clearly lower than for the ML-MFMPDR filter, the SD-MFMPDR filter and the WG. However, the results of the subjective listening test showed that the perceived overall quality for the DC-MFMPDR filter and the SD-MFMPDR filter is significantly better than for the ML-MFMPDR filter and shows no statistically significant difference to the WG. This can presumably be explained by the fact that all considered blind algorithms produce different artifacts and distortions in the speech and noise signals, which may be perceived and rated differently by the listeners. Although the perceived speech distortion for the proposed SD-MFMPDR filter and the state-of-the-art ML-MFMPDR filter is similar, the perceived overall quality for the SD-MFMPDR filter is significantly better than for the ML-MFMPDR filter, which can probably not only be explained by the fact that the SD-MFMPDR filter reduces more noise than the ML-MFMPDR filter but also by the fact that the ML-MFMPDR filter produces annoying noise artifacts (which were not explicitly assessed in the subjective listening test). Although the perceived amount of noise reduction for the DC-MFMPDR filter is clearly lower than for the ML-MFMPDR filter and the SD-MFMPDR filter, this is apparently compensated by the extremely low perceived speech distortion, such that there is no significant difference between the perceived overall quality of the DC-MFMPDR and SD-MFMPDR filters. Compared to the WG, the perceived speech distortion of the proposed DC-MFMPDR and SD-MFMPDR filters is significantly lower, but the perceived amount of noise reduction is also significantly lower (except for the SD-MFMPDR filter). These effects seen to compensate each other, such that there is no significant difference between the perceived overall quality of the DC-MFMPDR and SD-MFMPDR filters and the WG.

Table 7.3: Overview of the t-test results for the attribute (a) overall quality. The asterisks denote results that are statistically significant (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$) and o denotes results that are not statistically significant ($p > 0.05$).

					Oracle			Blind			
		Reference	**Anchor 1**	**Unprocessed**	**WG_O**	**MFMVDR_O**	**MFMPDR_O**	**WG**	**ML-MFMPDR**	**DC-MFMPDR**	**SD-MFMPDR**
	Reference		***	***	***	o	***	***	***	***	***
	Anchor 1	***		***	***	***	***	***	***	***	***
	Unprocessed	***	***		***	***	***	**	***	**	**
Oracle	**WG_O**	***	***	***		***	o	***	***	***	***
Oracle	**MFMVDR_O**	o	***	***	***		***	***	***	***	***
Oracle	**MFMPDR_O**	***	***	***	o	***		***	***	***	***
Blind	**WG**	***	***	**	***	***	***		***	o	o
Blind	**ML-MFMPDR**	***	***	***	***	***	***	***		***	***
Blind	**DC-MFMPDR**	***	***	**	***	***	***	o	***		o
Blind	**SD-MFMPDR**	***	***	**	***	***	***	o	***	o	

Table 7.5: Overview of the t-test results for the attribute (c) noise reduction. The asterisks denote results that are statistically significant (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$) and o denotes results that are not statistically significant ($p > 0.05$).

					Oracle			Blind			
		Unprocessed	**Anchor 2**	**Anchor 1**	**WG_O**	**MFMVDR_O**	**MFMPDR_O**	**WG**	**ML-MFMPDR**	**DC-MFMPDR**	**SD-MFMPDR**
	Unprocessed		***	***	***	***	***	***	***	*	***
	Anchor 2	***		***	o	o	**	***	***	***	***
	Anchor 1	***	***		***	***	***	***	**	o	***
Oracle	**WG_O**	***	o	***		o	o	**	***	***	**
Oracle	**MFMVDR_O**	***	o	***	o		o	***	***	***	***
Oracle	**MFMPDR_O**	***	**	***	o	o		o	***	***	*
Blind	**WG**	***	***	***	**	***	o		***	***	o
Blind	**ML-MFMPDR**	***	***	**	***	***	***	***		***	**
Blind	**DC-MFMPDR**	*	***	o	***	***	***	***	***		***
Blind	**SD-MFMPDR**	***	***	***	**	***	*	o	**	***	

Table 7.4: Overview of the t-test results for the attribute (b) speech distortion. The asterisks denote results that are statistically significant (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$) and o denotes results that are not statistically significant ($p > 0.05$).

					Oracle			Blind			
		Reference	**Anchor 1**	**Unprocessed**	**$\mathbf{WG_O}$**	**$\mathbf{MFMVDR_O}$**	**$\mathbf{MFMPDR_O}$**	**WG**	**ML-MFMPDR**	**DC-MFMPDR**	**SD-MFMPDR**
	Reference		***	o	***	o	**	***	***	o	***
	Anchor 1	***		***	***	***	***	***	***	***	***
	Unprocessed	o	***		**	o	*	***	***	o	***
Oracle	**$\mathbf{WG_O}$**	***	***	**		**	o	***	o	*	o
Oracle	**$\mathbf{MFMVDR_O}$**	o	***	o	**		o	***	**	o	***
Oracle	**$\mathbf{MFMPDR_O}$**	**	***	*	o	o		***	o	o	*
Blind	**WG**	***	***	***	***	***	***		***	***	**
Blind	**ML-MFMPDR**	***	***	***	o	**	o	***		**	o
Blind	**DC-MFMPDR**	o	***	o	*	o	o	***	**		***
Blind	**SD-MFMPDR**	***	***	***	o	***	*	**	o	***	

7.4 Summary

Aiming at determining the most effective estimator for the normalized speech correlation vector as well as the most perceptually advantageous speech enhancement algorithm, in this chapter we conducted an instrumental and perceptual evaluation of the most promising MFMPDR filters from Chapters 4, 5 and 6 together with an oracle MFMVDR filter and oracle and blind single-frame WGs. For the considered oracle algorithms, the instrumental performance measures and the results from the subjective listening test showed that the overall quality and the speech distortion for the oracle MFMVDR filter are better than for the oracle WG, while the noise reduction is similar. These results confirm the potential of multi-frame speech enhancement algorithms, i.e., exploiting speech interframe correlation enables to keep speech distortion low while suppressing the undesired background noise. For the blind algorithms, the instrumental performance measures indicated that the proposed SD-MFMPDR filter leads to a clearly better noise reduction performance than the DC-MFMPDR filter and the state-of-the-art ML-MFMPDR filter and to a similar noise reduction performance as the WG, while keeping speech distortion as low as the ML-MFMPDR filter. The results of the subjective listening test confirmed these instrumental evaluation results. In addition, the results of the subjective listening test showed that the perceived overall quality for the proposed DC-MFMPDR filter and the proposed SD-MFMPDR filter are significantly better than for the ML-MFMPDR filter and shows no statistically significant difference to the WG. The

perceived amount of noise reduction for the SD-MFMPDR filter and the WG is significantly higher than for the DC-MFMPDR filter and the ML-MFMPDR filter, while the perceived speech distortion for all MFMPDR filters is significantly lower than for the WG.

8

CONCLUSIONS AND FURTHER RESEARCH

In this chapter, we summarize the main contributions of the thesis and discuss possible directions for further research.

8.1 Conclusions

Speech communication devices such as hearing aids or mobile phones are often used in acoustically challenging situations, where the desired speech signal is affected by undesired background noise. Since in these situations speech quality and speech intelligibility may be degraded, especially at low SNRs, speech enhancement algorithms are required to suppress the undesired background noise while preserving the desired speech signal. Depending on the number of available microphones, single-microphone or multi-microphone algorithms can be used, where in this thesis we focused on single-microphone speech enhancement algorithms in the STFT-domain. Since consecutive STFT coefficients can be assumed to be correlated, especially when using a short frame length and a large overlap, we considered single-microphone multi-frame algorithms that aim at exploiting speech correlation across time-frames. In principle, exploiting the speech interframe correlation enables to suppress the undesired background noise, while keeping speech distortion low. The main objective of this thesis was to develop and evaluate novel robust methods to estimate the normalized speech correlation vector from the noisy microphone signal. This estimate can be used in existing single-microphone multi-frame speech enhancement algorithms, such as the multi-frame minimum variance distortionless response (MFMVDR) filter and the multi-frame minimum power distortionless response (MFMPDR) filter. We first investigated the sensitivity of the MFMVDR and MFMPDR filters to estimation errors in the normalized speech correlation vector, showing that even small estimation errors lead to a degraded performance. In order to improve the performance of the practically feasible MFMPDR filter, we proposed two novel methods to estimate the normalized speech correlation vector. On the one hand, we investigated the potential of using concepts from robust MPDR beamforming in the context of single-microphone multi-frame speech enhancement, by estimating the normalized speech correlation vector as the vector maximizing the total signal output PSD within a spherical uncertainty set. On the other hand, based on a low-rank speech model we proposed different matrix-based

normalized speech correlation vector estimators. The proposed algorithms were evaluated using instrumental performance measures and subjective listening tests.

In Chapter 2, we introduced the single-frame signal model and the multi-frame signal model, exploiting speech interframe correlation, in the STFT-domain. The single-frame signal model generally assumes that consecutive time-frames and frequency-bins are uncorrelated, which is a valid assumption when using sufficiently long frame length in the order of 16-32 ms and a small overlap of, e.g., 50 %. Hence, each time-frequency point is processed independently. The single-frame signal model is defined by the superposition of the speech STFT coefficient and the noise STFT coefficient. The speech STFT coefficients are estimated by applying a (real-valued) gain to the noisy speech STFT coefficients. The multi-frame signal model assumes that consecutive time-frames are correlated, which is a valid assumption when using a short frame length in the order of 4-8 ms and a large overlap of, e.g., 50-85 %. Exploiting speech interframe correlation, the speech vector is decomposed into the temporally correlated speech component, containing the (highly time-varying) normalized speech correlation vector, and the temporally uncorrelated speech component with respect to the current speech STFT coefficient. The normalized speech correlation vector contains the statistical information about the speech correlation across consecutive time-frames with respect to the current time-frame. The multi-frame signal model is hence defined by the correlated speech component and the undesired signal vector containing the uncorrelated speech component and the noise vector. The speech STFT coefficients are estimated by applying a (complex-valued) FIR filter to the noisy speech vector.

In Chapter 3, we reviewed single-frame speech enhancement algorithms, more in particular the Wiener gain (WG) and multi-frame speech enhancement algorithms exploiting speech interframe correlation, more in particular the MFMVDR filter and the MFMPDR filter. The MFMVDR filter aims at minimizing the undesired output PSD while not distorting the correlated speech component. The MFMVDR filter depends on the normalized speech correlation vector and the correlation matrix of the undesired signal vector, which are typically both highly time-varying and hence difficult to accurately estimate in practice. As an alternative to the MFMVDR filter, the MFMPDR filter was derived using the noisy speech correlation matrix, instead of the undesired correlation matrix. Since the considered single-microphone multi-frame algorithms are related to multi-microphone beamforming algorithms, we also discussed this conceptual similarities and differences. Conceptually, the multi-frame signal model is very similar to the multi-microphone signal model when interpreting time-frames as microphone inputs, the normalized speech correlation vector as the RTF vector of the desired speech source and the temporally uncorrelated speech component as residual speech component (e.g., modeling the late reverberation). One of the main differences is that the normalized speech correlation vector in the multi-frame signal model contains statistical information, which is highly time-varying and needs to be estimated for each time-frequency point, whereas the RTF vector of the desired speech source in the multi-microphone signal model depends on spatial information and can

be assumed to be more stationary than the normalized speech correlation vector. Furthermore, the temporally uncorrelated speech component in the multi-frame signal model is highly non-stationary and can even be dominant compared to the temporally correlated speech component, whereas the residual speech component in the multi-microphone signal model can be assumed to be either not dominant (often this term is even completely neglected) or spatially stationary. Hence, it is a valid assumption for the multi-microphone signal model to neglect the residual speech component, such that the speech correlation matrix can be assumed as rank 1, whereas the rank 1 assumption for the speech correlation matrix in the multi-frame signal model, i.e., neglecting the uncorrelated speech component, may lead to a degraded speech enhancement performance.

In Chapter 4, we investigated the speech enhancement performance of two oracle versions of the MFMVDR filter and the practically feasible MFMPDR filter using different oracle and blind estimates of the normalized speech correlation vector. For the oracle estimates, we made the unrealistic assumption that either the speech STFT coefficients (referred to as quasi-perfect estimate) and/or the noise STFT coefficients are available. For the blind estimates, only the noisy speech STFT coefficients are obviously available. As blind estimate of the normalized speech correlation vector, we considered the state-of-the-art ML estimator [15], which depends on a mean normalized noise correlation vector estimate and an a-priori SNR estimate. First, we investigated the influence of the filter length on the performance of both oracle MFMVDR filters and the MFMPDR filter using the oracle (quasi-perfect) normalized speech correlation vector. Using instrumental performance measures, simulation results for different speech signals and noise types showed that the performance of the three multi-frame filters increases with increasing filter length but saturates at about 15-21 ms of analysis data. Second, we compared the speech enhancement performance of the considered MFMVDR and MFMPDR filters using the oracle and blind estimates of the normalized speech correlation vector. The simulation results showed that, as expected, the MFMVDR filter using the oracle (quasi-perfect) estimate of the undesired correlation matrix achieves the best results, but that even small estimation errors in the normalized speech correlation vector decrease the performance. Furthermore, the results showed that the influence of the uncorrelated speech component is crucial, especially at high SNRs. The results also showed that the performance of the MFMPDR filter is very close to the performance of the oracle (quasi-perfect) MFMVDR filter when using oracle estimates of the normalized speech correlation vector. When using the blind ML estimate, the performance of the MFMPDR filter is strongly reduced due to large estimation errors in the normalized speech correlation vector, e.g., for an SNR of 5 dB, the PESQ improvement is reduced by 0.5 MOS. Motivated by these results, in Chapter 5 and 6 we proposed novel methods to estimate the normalized speech correlation vector, thereby improving the speech enhancement performance of the MFMPDR filter.

In Chapter 5, we investigated the potential of using concepts proposed for robust MPDR beamforming in the context of single-microphone multi-frame speech

enhancement. We proposed two constrained MFMPDR filters that estimate the normalized speech correlation vector as the vector maximizing the total signal output PSD within a spherical uncertainty set. This corresponds to imposing a quadratic inequality constraint on the mismatch vector with respect to the presumed normalized speech correlation vector. Whereas the singly-constrained (SC) MFMPDR filter only considers the quadratic inequality constraint to estimate the (non-normalized) speech correlation vector, the doubly-constrained (DC) MFMPDR filter integrates a linear normalization constraint into the optimization problem to directly estimate the normalized speech correlation vector. The upper bound of the spherical uncertainty set plays a crucial role for both constrained optimization problems in that it should be chosen in accordance with the accuracy of the presumed normalized speech correlation vector. Since we used the ML estimate as the presumed normalized speech correlation vector, which strongly depends on the a-priori SNR estimate, we proposed to use a trained non-linear mapping function that depends on the a-priori SNR estimate to set the upper bound of the spherical uncertainty set for each time-frequency point. Using instrumental performance measures, simulation results for different speech signals, noise types and SNRs showed that the SC and DC normalized speech correlation vector estimates clearly lead to a more accurate estimate than the ML estimate, with the DC estimate achieving the highest estimation accuracy and no outliers. In addition, we compared the speech enhancement performance of the constrained MFMPDR filters with the ML-MFMPDR filter and the single-frame WG, both using instrumental performance measures and using a pairwise comparison test. The results indicated that although the constrained MFMPDR filters lead to a more conservative noise reduction than the ML-MFMPDR filter and the WG, especially the DC-MFMPDR filter produces less speech and noise distortions than the ML-MFMPDR filter, such that the speech sounds more natural and less musical noise is present. For instance, for an SNR of 0 dB, the overall quality of the DC-MFMPDR filter was preferred to the other filters by 63 %, while the WG was preferred by 88 %, SD-MFMPDR filter by 38 % and the ML-MFMPDR filter by 25 %.

Assuming that speech signals can be modeled using a low-rank model, e.g., as a linear combination of a limited number of complex exponentials, the speech correlation matrix can be assumed to be rank-deficient. Based on this low-rank speech model, in Chapter 6 we proposed two matrix-based methods to estimate the normalized speech correlation vector, namely the matrix-subtraction (MS) method and the subspace-decomposition (SD) method. Both methods are based on the eigenvalue decomposition of a matrix, which is either constructed by subtracting the estimated normalized noise correlation matrix from the estimated normalized noisy speech correlation matrix or by prewhitening the estimated normalized noisy speech correlation matrix with the estimated normalized noise correlation matrix. For the speech model order, we either assumed that the uncorrelated speech component in the multi-frame signal model can be neglected and used a fixed value equal to 1 for all time-frequency points or estimated the speech model order for each time-frequency point. When using a limited amount of data, which is typically

the case for the considered multi-frame filters, most classical model order selection criteria, such as the minimum description length (MDL) selection criterion, have a poor estimation accuracy. Hence, we proposed to estimate the speech model order by incorporating the a-priori SNR into a classical model selection criterion, i.e., the MDL selection criterion. Simulation results for different speech signals, noise types and SNRs showed that the proposed matrix-based methods yield a more accurate estimate of the normalized speech correlation vector than the state-of-the-art vector-based ML estimate. Instrumental performance measures showed that assuming a rank-1 speech correlation matrix, i.e., neglecting the uncorrelated speech component, strongly decreases the speech enhancement performance, especially at high SNRs. The results also indicated that the SD-MFMPDR filter using the proposed a-priori SNR-based speech model order estimate leads to a better speech quality and more noise reduction than the state-of-the-art ML-MFMPDR filter, while keeping speech distortion low. For instance, for an SNR of 5 dB, the PESQ improvement for the SD-MFMPDR filter is 0.03 MOS lower than for the WG, but 0.11 MOS higher than for the ML-MFMPDR filter.

In order to determine the most effective estimator for the normalized speech correlation vector as well as the most perceptually advantageous speech enhancement algorithm, in Chapter 7 we compared the performance of the most promising MFMPDR filters from Chapters 4, 5 and 6 with an oracle MFMVDR filter and with oracle and blind single-frame WGs. For the considered oracle algorithms, the instrumental performance measures and the results from the subjective listening test showed that the overall quality and the speech distortion for the oracle MFMVDR filter are better than for the oracle WG, while the noise reduction is similar. These results confirm the motivation of this thesis, namely that exploiting speech interframe correlation enables to keep speech distortion low while suppressing the undesired background noise. For the blind algorithms, the instrumental performance measures indicated that the proposed SD-MFMPDR filter leads to a clearly better noise reduction performance than the DC-MFMPDR filter and the ML-MFMPDR filter and to a similar noise reduction performance as the WG, while keeping speech distortion as low as the ML-MFMPDR filter. For instance, for an SNR of 5 dB, the segmental noise reduction score of the SD-MFMPDR filter is 6.5 dB better than the DC-MFMPDR filter, 2 dB better than the ML-MFMPDR filter and equal to the WG. The segmental speech distortion score of the SD-MFMPDR filter is equal to the ML-MFMPDR filter and 1 dB better than the WG, but 6 dB worse than the DC-MFMPDR filter. The results of the subjective listening test confirmed these instrumental evaluation results. In addition, the results of the subjective listening test showed that the perceived overall quality for the proposed DC-MFMPDR and SD-MFMPDR filters is significantly better than for the state-of-the-art ML-MFMPDR filter but shows no statistically significant difference to the WG. In conclusion, these results show that the proposed methods to estimate the normalized speech correlation vector substantially improve the overall quality compared to the state-of-the-art ML-MFMPDR filter. This is achieved either for the DC-MFMPDR filter by a significantly lower speech distortion at the expense of a more conservative noise reduction or for the

SD-MFMPDR filter by a similar speech distortion, but more noise reduction and less artifacts in the background noise.

8.2 Suggestion for Further Research

In the following, we discuss possible research directions for further improvements and possible applications of the proposed normalized speech correlation vector estimators.

In Chapter 4, we investigated the speech enhancement performance of the practically feasible MFMPDR filter and two oracle versions of the MFMVDR filter. Using oracle estimates of the normalized speech correlation vector, simulation results showed that the oracle (quasi-perfect) MFMVDR filter achieves a better overall quality than the MFMPDR filter, with a similar noise reduction and speech distortion. In practice, the MFMVDR filter requires an estimate of the highly time-varying undesired correlation matrix. This correlation matrix does not only contain the time-varying noise correlation matrix but also the highly time-varying correlation matrix of the uncorrelated speech component such that it is quite difficult to accurately estimate the matrix from the noisy speech STFT coefficients. While in this thesis we focused on developing robust methods to estimate the normalized speech correlation vector from the noisy speech STFT coefficients, further research could aim at developing methods to estimate the undesired correlation matrix, e.g., similarly to [60, 113, 163–165].

In Chapter 5, we investigated the potential of using concepts proposed for robust MPDR beamforming in the context of single-microphone multi-frame speech enhancement. We proposed two constrained MFMPDR filters that estimate the normalized speech correlation vector as the vector maximizing the total signal output PSD within a spherical uncertainty set. The main novelty was to set the upper bound of this spherical uncertainty set based on the time-varying a-priori SNR for each time-frequency point. However, these algorithms are computationally quite complex. As further research, the less complex diagonal loading approach could be investigated in the context of multi-frame speech enhancement, which imposes a quadratic inequality constraint on the filter vector instead of the mismatch vector with respect to the presumed normalized speech correlation vector. Since the loading level enables a trade-off between the non-robust MFMPDR filter without additional constraint (i.e., no diagonal loading) and no filtering (i.e., a loading level equal to infinity), it would be interesting to investigate if the loading level could be set based on the a-priori SNR for each time-frequency point.

In Chapter 5, 6 and 7 instrumental and perceptual evaluations showed that the proposed DC-MFMPDR filter from Chapter 5 and the SD-MFMPDR filter from Chapter 6 lead to less speech distortion but also to a lower noise reduction performance than the single-frame WG. In order to benefit both from the good noise reduction performance of the WG, while during speech presence the speech distortion are kept low using the DC-MFMPDR or the SD-MFMPDR filter, further

research could be to integrate the proposed MFMPDR filters with a single-frame WG, similar to the approach in [128]. In this approach we proposed a time- and frequency-dependent SPP-based weighting between the MFMPDR filter and the WG. When speech is likely to be present, the MFMPDR filter is applied to the noisy speech STFT coefficients, whereas when speech is likely to be absent, the WG is applied.

In this thesis, we considered a single-microphone multi-frame signal model exploiting speech interframe correlation, which is conceptually very similar to a multi-microphone signal model (see Section 3.2.3). Since most multi-frame filters are hence inspired by multi-microphone algorithms, i.e., beamformers, further research could be to develop other single-microphone multi-frame speech enhancement algorithms using

1. statistical model-based estimators, e.g., of the multi-channel MAP spectral amplitude estimator in [166].
2. machine-learning-based estimators, e.g., the DNN-based approach in [167] to directly learn the complex-valued filter coefficients.
3. subspace-based estimators, e.g, variable span-filters using joint diagonalization of the speech and noise correlation matrices [44, 168].

Furthermore, in [4, 169, 170] an extension of the multi-microphone MFMVDR filter was proposed, simultaneously exploiting the RTF vector of the desired speech source and the normalized speech correlation vector. This spatial-temporal vector was expressed in terms of the Kronecker product of the RTF vector and the normalized speech correlation vector. Using oracle estimates of the required quantities, simulation results in [4, 169, 170] showed that the speech enhancement performance for the multi-microphone MFMVDR filter is better than for the multi-microphone single-frame MVDR filter, which neglects the speech interframe correlation. In further research, it would be interesting to evaluate this multi-microphone MFMVDR filter using the proposed normalized speech correlation vector estimates in combination with a blind estimate of the RTF vector of the desired speech source, e.g., using the covariance subtraction method [42, 114–117] or the covariance whitening method [39, 42, 115, 116].

Finally, in this thesis we used the multi-frame signal model exploiting speech interframe correlation for noise reduction algorithms. Based on the same signal model in [171] and [172] single-microphone and multi-microphone dereverberation algorithms were proposed, assuming an oracle estimate of the normalized speech correlation vector. Since this resulted in a good dereverberation performance, for further research it would be interesting to investigate the proposed normalized speech correlation vector estimators for dereverberation and/or for joint dereverberation and noise reduction.

A

MFMPDR FILTERING WITH WIENER POSTFILTERING

In Chapter 3, we reviewed single-frame speech enhancement algorithms and multi-frame speech enhancement algorithms exploiting speech interframe correlation. More in particular, we discussed the single-frame Wiener gain (WG), the multi-frame minimum power distortionless response (MFMPDR) filter and the multi-frame Wiener filter (MFWF). We showed that the MFWF can be decomposed into the MFMPDR filter and a single-frame WG as postfilter. In this appendix, we discuss the speech enhancement performance of the MFMPDR filter with postfilter using different undesired output PSD estimates. As shown in Section 3.2.2, the MFWF may be sensitive to numerical errors in the required quantities and more robust results may be obtained by decomposing the MFWF into the MFMPDR filter and a WG as postfilter. While the MFMPDR filter is designed to avoid speech distortion, the WG aims at minimizing the MSE between the output of the MFMPDR filter and the speech STFT coefficient. Hence, the decomposed MFWF is capable to reduce the undesired signal more strongly than the MFMPDR filter, but speech distortion may be introduced. In this appendix, we compare the speech enhancement performance of the MFMPDR filter with postfilter, using different undesired output PSD estimates, with the MFWF and the MFMPDR filter as reference. The undesired output PSD is either estimated directly from the noisy speech STFT coefficients, independently to the output of the MFMPDR filter or based on pre-trained data.

In Section A.1, we recap the decomposition of the MFWF and present the different estimators for the undesired output PSD. In Section A.2, we compare the speech

This appendix is partly based on:

[129] D. Fischer and T. Gerkmann, "Single-microphone speech enhancement using MVDR filtering and Wiener post-filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 201–205.

[130] D. Fischer, K. Brümann, and S. Doclo, "Comparison of parameter estimation methods for single-microphone multi-frame Wiener filtering," in *Proc. of Europ. Signal Process. Conf. (EUSIPCO)*, A Coruña, Spain, Aug. 2019, pp. 1809–1813.

enhancement performance of the decomposed MFWF with the MFWF and the MFMPDR filter for different speech signals, noise types and SNRs using instrumental performance measures. Simulation results show that robust results of the MFWF can be achieved by decomposing the MFWF and that the speech quality can be improved in reference to the MFMPDR filter. However, when estimating the undesired output PSD independently to the output of the MFMPDR filter, no speech quality improvement can be indicated.

A.1 Undesired Output PSD Estimators

In this section, we recap the decomposed MFWF from Section 3.2.2 and present several estimators for the undesired PSD at the output of the MFMPDR filter. For conciseness, the frequency-bin index k and the time-frame index m will be omitted in this appendix if not required. However, it should be realized that all calculations are performed for each time-frequency point.

It was shown in Section 3.2.2 that the MFWF, i.e.,

$$\boldsymbol{h}_{\text{MFWF}} = \phi_X \boldsymbol{R}_{\boldsymbol{y}}^{-1} \boldsymbol{\gamma}_{\boldsymbol{x}} \,, \tag{A.1}$$

depends on the speech PSD ϕ_X, the normalized speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{x}}$ and the noisy speech correlation matrix $\boldsymbol{R}_{\boldsymbol{y}}$. We showed that the MFWF in (A.1) may be more sensitive to numerical errors in the required quantities than the MFMPDR filter (cf. (3.31)-(3.34)) and more robust results may be obtained by decomposing the MFWF into the MFMPDR filter $\boldsymbol{h}_{\text{MFMPDR}}$ in (3.23) and a single-frame WG as postfilter, i.e,

$$\boldsymbol{h}_{\text{MFMPDR-WG}} = \underbrace{\frac{\boldsymbol{R}_{\boldsymbol{y}}^{-1}\boldsymbol{\gamma}_{\boldsymbol{x}}}{\boldsymbol{\gamma}_{\boldsymbol{x}}^{H}\boldsymbol{R}_{\boldsymbol{y}}^{-1}\boldsymbol{\gamma}_{\boldsymbol{x}}}}_{\boldsymbol{h}_{\text{MFMPDR}}} \frac{\phi_X}{\phi_X + \phi_{\boldsymbol{u}}^{\text{out}}}. \tag{A.2}$$

This postfilter operates on the output of the MFMPDR filter, where the undesired output PSD $\phi_{\boldsymbol{u}}^{\text{out}}$ is given by

$$\phi_{\boldsymbol{u}}^{\text{out}} = \boldsymbol{h}_{\text{MFMPDR}}^{H} \boldsymbol{R}_{\boldsymbol{u}} \boldsymbol{h}_{\text{MFMPDR}} \,, \tag{A.3}$$

with the undesired correlation matrix $\boldsymbol{R}_{\boldsymbol{u}}$. The undesired correlation matrix does not only contain the noise correlation matrix $\boldsymbol{R}_{\boldsymbol{n}}$ but also the correlation matrix of the uncorrelated speech component $\boldsymbol{R}_{\boldsymbol{x}'}$ (cf. (2.24)). Typically, both quantities are highly time-varying, making it quite difficult to accurately estimate them from the noisy speech STFT coefficients.

In order to implement the $\boldsymbol{h}_{\text{MFMPDR-WG}}$ in (A.2) estimates of the speech PSD ϕ_X and the undesired output PSD $\phi_{\boldsymbol{u}}^{\text{out}}$ are required. The MFMPDR filter is designed to prevent speech distortion such that in theory, the speech PSD ϕ_X at the output of the MFMPDR filter is equal to the input speech PSD. Using this assumption, we also apply the speech PSD estimate $\widehat{\phi}_X$ at the output of the MFMPDR filter. The undesired output PSD $\phi_{\boldsymbol{u}}^{\text{out}}$ can be estimated in several ways.

First, the most straightforward way is to estimate this PSD as

$$\boxed{\widehat{\phi}_{\boldsymbol{u}}^{\text{out},\boldsymbol{R}_{\boldsymbol{u}}} = \widehat{\boldsymbol{h}}_{\text{MFMPDR}}^{H} \widehat{\boldsymbol{R}}_{\boldsymbol{u}} \widehat{\boldsymbol{h}}_{\text{MFMPDR}}} \tag{A.4}$$

where $\widehat{\boldsymbol{R}}_{\boldsymbol{u}}$ is estimated as

$$\widehat{\boldsymbol{R}}_{\boldsymbol{u}} = \widehat{\boldsymbol{R}}_{\boldsymbol{y}} - \widehat{\phi}_X \widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}} \widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{H}. \tag{A.5}$$

Since $\widehat{\boldsymbol{R}}_{\boldsymbol{u}}$ may not be positive semi-definite due to estimation errors, we set the negative eigenvalues of $\widehat{\boldsymbol{R}}_{\boldsymbol{u}}$ to zero.

Second, instead of estimating the undesired correlation matrix $\boldsymbol{R}_{\boldsymbol{u}}$ we proposed in [129] to use a pre-trained (time-independent but frequency-dependent) normalized undesired correlation matrix $\boldsymbol{\Gamma}_{\boldsymbol{u}}$, which is defined as

$$\boldsymbol{\Gamma}_{\boldsymbol{u}} = \frac{\boldsymbol{R}_{\boldsymbol{u}} \boldsymbol{e}}{\boldsymbol{e}^T \boldsymbol{R}_{\boldsymbol{u}} \boldsymbol{e}}, \tag{A.6}$$

where $\boldsymbol{e}^T \boldsymbol{R}_{\boldsymbol{u}} \boldsymbol{e}$ corresponds to the noise PSD ϕ_N (cf. (4.14)). During training, perfect knowledge of the speech and noise signals are available such that the oracle undesired correlation matrix estimate $\widehat{\boldsymbol{R}}_{\boldsymbol{u}}^{\text{O}}$ in (4.15) can be obtained using oracle estimates of the speech correlation matrix $\boldsymbol{R}_{\boldsymbol{x}}$ and the noise correlation matrix $\boldsymbol{R}_{\boldsymbol{n}}$ as in (4.1) and (4.2), respectively. The pre-trained normalized undesired correlation matrix $\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{u}}^{\text{Tr}}$ is subsequently obtained by averaging $\widehat{\boldsymbol{R}}_{\boldsymbol{u}}^{\text{O}}$ over all training data and normalizing the resulting matrix to its first element, similarly as in (A.6). To estimate $\phi_{\boldsymbol{u}}^{\text{out}}$, the trained $\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{u}}^{\text{Tr}}$ needs to be scaled by the noise PSD estimate $\widehat{\phi}_N$. Hence, the undesired output PSD $\phi_{\boldsymbol{u}}^{\text{out}}$ can be estimated as

$$\boxed{\widehat{\phi}_{\boldsymbol{u}}^{\text{out},\boldsymbol{\Gamma}_{\boldsymbol{u}}} = \widehat{\boldsymbol{h}}_{\text{MFMPDR}}^{H} \left(\widehat{\phi}_N \widehat{\boldsymbol{\Gamma}}_{\boldsymbol{u}}^{\text{Tr}} \right) \widehat{\boldsymbol{h}}_{\text{MFMPDR}}} \tag{A.7}$$

Third, assuming that the correlation matrix of the uncorrelated speech component $\boldsymbol{R}_{\boldsymbol{x}'}$ can be neglected, we can replace $\boldsymbol{R}_{\boldsymbol{u}}$ in (A.3) by the noise correlation matrix $\boldsymbol{R}_{\boldsymbol{n}}$, i.e., $\boldsymbol{R}_{\boldsymbol{u}} = \boldsymbol{R}_{\boldsymbol{n}}$, resulting in the noise output PSD $\phi_{\boldsymbol{n}}^{\text{out}}$, i.e.,

$$\phi_{\boldsymbol{n}}^{\text{out}} = \boldsymbol{h}_{\text{MFMPDR}}^{H} \boldsymbol{R}_{\boldsymbol{n}} \boldsymbol{h}_{\text{MFMPDR}}\,. \tag{A.8}$$

Hence, the undesired output PSD can be estimated independently to the output of the MFMPDR filter, using noise PSD estimators as in Section 3.1.3 at the output signal of the MFMPDR filter, i.e.,

$$\boxed{\widehat{\phi}_{\boldsymbol{u}}^{\text{out},\phi_{\boldsymbol{n}}} = \widehat{\phi}_{\boldsymbol{n}}^{\text{out}}} \tag{A.9}$$

A.2 Simulation Results

In this section, we investigate the speech enhancement performance of the MFMPDR filter with a WG as postfilter using different undesired output PSD estimates. In Section A.2.1, we describe the audio material and discuss the algorithmic settings. In Section A.2.2, we compare the speech enhancement performance of the MFMPDR filter with postfilter, using the different presented estimators for undesired output PSD from Section A.1, with the MFWF in (A.1) and the MFMPDR filter in (3.23) as reference.

A.2.1 *Audio Material and Algorithmic Settings*

For the evaluation, we used 260 s of speech material (131 s female, 129 s male) from the TIMIT database [146] as speech signals, at a sampling frequency of 16 kHz. As noise signals, we used traffic noise, two babble noise signals, factory noise and modulated white Gaussian noise with a modulation frequency of 0.5 Hz taken from the NOISEX-92 database [145]. The considered SNRs ranged from -5 dB to 20 dB in 5 dB steps. For the training of the normalized undesired correlation matrix $\widehat{\mathbf{\Gamma}}_{\boldsymbol{u}}^{\mathrm{Tr}}$ in (A.7), we used 120 s of speech material (58 s female, 62 s male) from the TIMIT database [146] as clean speech signals. As noise signals, we used pink noise, office noise and multi-talker babble noise. The considered SNRs range from -5 dB to 20 dB in 5 dB steps. We make sure that the training data differs from the evaluation data.

Similarly as in Section 4.3, in order to achieve a high speech interframe correlation we used a highly temporally resolved STFT framework with a frame length of 4 ms ($T = K = 64$ samples) and an overlap of 75 %, resulting in a frame shift of 1 ms ($R = 16$ samples). As the STFT analysis window $w_a(t)$ and synthesis window $w_s(t)$, we used a square-root Hann window. Similarly as in Chapter 4 and [15], the number of consecutive time-frames was set to $L = 18$, resulting in 21 ms of analysis data used in each filtering operation.

For the MFWF in (A.1) and the MFMPDR filter in (A.2), the noisy speech correlation matrix $\boldsymbol{R}_{\boldsymbol{y}}$ was estimated using recursive smoothing as in (4.3). The smoothing parameter was experimentally set to $\alpha_y = 0.92$, corresponding to a smoothing window of 12 ms. To avoid numerical problems, we performed diagonal loading as in (4.17) before computing the inverse of the noisy speech correlation matrix with a scaling parameter of $\kappa = 0.001$, similarly as in Chapter 4. The normalized speech correlation vector was estimated using the state-of-the-art ML estimate $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{ML}}$ in (4.12). The required a-priori SNR ξ was estimated using the DDA estimate $\widehat{\xi}^{\mathrm{DDA}}$ in (3.9), with a weighting parameter $\alpha_{\mathrm{DDA}} = 0.97$. To reduce unpleasant artifacts in the background noise, e.g., musical noise, we only updated the estimated speech STFT coefficient $\widehat{X}(m-1)$ in (3.9) every 4 ms, i.e., every 4 frames. The required noise PSD was estimated using the SPP-based noise PSD estimator $\widehat{\phi}_N^{\mathrm{SPP}}$ in (3.13), with a smoothing parameter $\alpha_{\mathrm{SPP}} = 0.90$ and a fixed SNR $\xi_{\mathcal{H}_1} = -15$ dB. To reduce the amount of speech distortion and to mask artifacts in the background noise, we applied a lower limit of $\xi_{\min} = -8$ dB to the a-priori SNR estimate.

Table A.1: Overview of the considered MFMPDR filters with WG as postfilter.

Label	**Description**
MFMPDR-WG$_{\boldsymbol{R}_{\boldsymbol{u}}}$	MFMPDR filter with postfilter in (A.2) using $\widehat{\phi}_{\boldsymbol{u}}^{\text{out},\boldsymbol{R}_{\boldsymbol{u}}}$
MFMPDR-WG$_{\boldsymbol{\Gamma}_{\boldsymbol{u}}}$	MFMPDR filter with postfilter in (A.2) using $\widehat{\phi}_{\boldsymbol{u}}^{\text{out},\boldsymbol{\Gamma}_{\boldsymbol{u}}}$
MFMPDR-WG$_{\phi_{\boldsymbol{n}}}$	MFMPDR filter with postfilter in (A.2) using $\widehat{\phi}_{\boldsymbol{u}}^{\text{out},\phi_{\boldsymbol{n}}}$

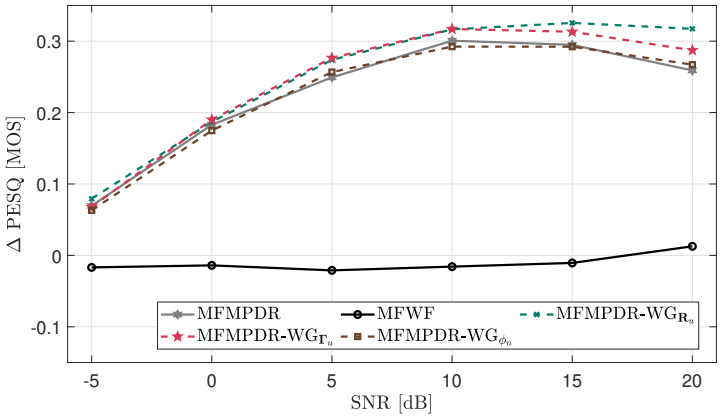


Fig. A.1: Average PESQ improvement for the MFWF and the MFMPDR filter with Wiener gain (WG) as postfilter, using the undesired output PSD estimates from Section A.1, in reference to the MFMPDR filter, for different SNRs.

For the WG as postfilter, the speech PSD ϕ_X was estimated by multiplying the DDA estimate $\widehat{\xi}^{\text{DDA}}$ with the noise PSD estimate, i.e., $\widehat{\phi}_X = \widehat{\phi}_N^{\text{SPP}}\widehat{\xi}^{\text{DDA}}$. To train the normalized undesired correlation matrix, an estimate of the oracle undesired correlation matrix $\widehat{\boldsymbol{R}}_{\boldsymbol{u}}^{\text{O}}$ in (4.15) was required. This matrix based on estimates of the speech and noise correlation matrices, which were estimated using recursive smoothing as in (4.1) and (4.2), respectively. The smoothing parameters were experimentally set to $\alpha_x = 0.65$, corresponding to 2.5 ms and $\alpha_n = 0.90$, corresponding to 10 ms. The noise output PSD in (A.9) was estimated using the SPP-based noise PSD estimator $\widehat{\phi}_N^{\text{SPP}}$ in (3.13), with $\alpha_{\text{SPP}} = 0.90$ and $\xi_{\mathcal{H}_1} = -15\,\text{dB}$, at the output signal of the MFMPDR filter.

A.2.2 *Speech Enhancement Performance*

In this section, we compare the speech enhancement performance of the considered MFMPDR filters with WG as postfilter in (A.2) either using the undesired output PSD estimate $\widehat{\phi}_{\boldsymbol{u}}^{\text{out},\boldsymbol{R}_{\boldsymbol{u}}}$ in (A.4), $\widehat{\phi}_{\boldsymbol{u}}^{\text{out},\boldsymbol{\Gamma}_{\boldsymbol{u}}}$ in (A.7) or $\widehat{\phi}_{\boldsymbol{u}}^{\text{out},\phi_N}$ in (A.9) (cf. Table A.1) with the MFWF in (A.1) and the MFMPDR filter in (3.23) as reference. The

considered algorithms are evaluated in terms of speech quality using the PESQ improvement ΔPESQ (cf. Section 2.2).

For different SNRs, Fig. A.1 depicts the performance of the considered algorithms averaged over all considered speech signals and noise types. First, for all SNRs it can be seen that the MFMPDR filter and all MFMPDR filters with postfilter lead to higher ΔPESQ results than the MFWF in (A.1) Second, it can be seen that the MFMPDR-WG$_{\boldsymbol{R}_{\boldsymbol{u}}}$ filter and the MFMPDR-WG$_{\boldsymbol{\Gamma}_{\boldsymbol{u}}}$ filter achieve higher ΔPESQ results than all other filters, while both results are similar (except for high SNRs). Third, it can be seen that estimating the undesired output PSD independently to the output of the MFMPDR filter leads to similar ΔPESQ results as the MFMPDR filter.

These results show that decomposing the MFWF into the MFMPDR filter and a WG as postfilter leads to more robust results as the MFWF in (A.1). Using the undesired output PSD estimates $\widehat{\phi}_{\boldsymbol{u}}^{\text{out},\boldsymbol{R}_{\boldsymbol{u}}}$ in (A.4) and $\widehat{\phi}_{\boldsymbol{u}}^{\text{out},\boldsymbol{\Gamma}_{\boldsymbol{n}}}$ in (A.7) in the postfilter yield a better speech quality than the MFMPDR filter. However, independently estimating the undesired PSD at the output of the MFMPDR filter, i.e., using $\widehat{\phi}_{\boldsymbol{u}}^{\text{out},\phi_{\boldsymbol{n}}}$ in (A.9), does not yield a PESQ improvement compared to the MFMPDR filter. This results can probably not only be explained by the fact that considering the uncorrelated speech component is crucial (as also shown in Chapters 4 and 6), but also by the fact that the used noise PSD estimator $\widehat{\phi}_N^{\text{SPP}}$, which assumes that the noise is more stationary than the speech, whereas the output of the MFMPDR filter may contain highly fluctuating residual noise leading to an underestimation of the noise output PSD.

A.3 Summary

In this appendix, we compared the speech enhancement performance of the decomposed MFWF filter into the MFMPDR filter and a WG as postfilter, using different estimators for the undesired output PSD, with the MFWF and the MFMPDR filter as reference. The undesired output PSD in the postfilter was either estimated directly from the noisy speech STFT coefficients, independently to the output of the MFMPDR filter or based on pre-trained data. For different speech signals, noise types and SNRs, simulation results show that the MFMPDR filter with postfilter leads to more robust results and a clearly higher speech quality than the MFWF. In reference to the MFMPDR filter, whereas estimating the undesired output PSD independently to the output of the MFMPDR filter does not yield a speech quality improvement, estimating the undesired output PSD directly from the noisy speech STFT coefficients or based on pre-trained data improves the speech quality.

BIBLIOGRAPHY

[1] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, ser. Digital Signal Processing. Springer Science & Business Media, 2001.

[2] J. Benesty, J. Chen, and S. Makino, *Speech Enhancement*, ser. Signals and Communication Technology. Springer Science & Business Media, 2005.

[3] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding And Error Concealment*. John Wiley & Sons, 2006.

[4] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*. Springer Science & Business Media, 2011.

[5] P. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. CRC press, 2013.

[6] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art*. Morgan & Claypool, 2013.

[7] 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.

[8] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.

[9] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[10] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer Handbook of speech processing*. Springer, 2001, pp. 945–978.

[11] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001, ch. 3, pp. 39–60.

[12] M. Kolbæk, Z. Tan, and J. Jensen, "Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 1, pp. 153–167, Jan. 2017.

[13] E. Healy, E. Johnson, M. Delfarah, and D. Wang, "A talker-independent deep learning algorithm to increase intelligibility for hearing-impaired listeners in reverberant competing talker conditions," *J. Acoust. Soc. Amer.*, vol. 147, no. 6, pp. 4106–4118, May 2020.

[14] Y. Huang and J. Benesty, "A multi-frame approach to the frequency-domain single-channel noise reduction problem," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1256–1269, May 2012.

[15] A. Schasse and R. Martin, "Estimation of subband speech correlations for noise reduction via MVDR processing," *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, no. 9, pp. 1355–1365, Sep. 2014.

[16] K. T. Andersen and M. Moonen, "Robust speech-distortion weighted interframe Wiener filters for single-channel noise reduction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 26, no. 1, pp. 97–107, Jan. 2018.

[17] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals.* Prentice Hall, Upper Saddle River, 1978.

[18] J. Flanagan, "Parametric coding of speech spectra," *J. Acoust. Soc. Amer.*, vol. 68, no. 2, pp. 412–419, Aug. 1980.

[19] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[20] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[21] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[22] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Aug. 2005.

[23] S. Gazor and W. Zhang, "Speech enhancement employing Laplacian-Gaussian mixture," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 896–904, Sep. 2005.

[24] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Applied Signal Process.*, vol. 2005, no. 7, pp. 1110–1126, Jan. 2005.

[25] 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 Trans. Audio, Speech, Language Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[26] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[27] N. Mohammadiha, R. Martin, and A. Leijon, "Spectral domain speech enhancement using HMM state-dependent super-Gaussian priors," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 253–256, Mar. 2013.

[28] A. Aroudi, H. Veisi, and H. Sameti, "Hidden Markov model-based speech enhancement using multivariate Laplace and Gaussian distributions," *IET Signal Processing*, vol. 9, no. 2, pp. 177–185, Apr. 2015.

[29] D. N. Lawley, "Tests of Significance for the Latent Roots of Covariance and Correlation Matrices," *Biometrika*, vol. 43, no. 1/2, pp. 128–136, Jun. 1956.

[30] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Commun.*, vol. 10, no. 1, pp. 45–57, Feb. 1991.

[31] S. Bakamidis, M. Dendrinos, and G. Carayannis, "SVD analysis by synthesis of harmonic signals," *IEEE Trans. Signal Process.*, vol. 39, no. 2, pp. 472–477, Feb. 1991.

[32] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[33] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 6, pp. 439–448, Nov. 1995.

[34] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

[35] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise," *IEEE Signal Process. Lett.*, vol. 10, no. 4, pp. 104–106, Apr. 2003.

[36] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 700–708, Nov. 2003.

[37] P. C. Hansen and S. H. Jensen, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis," *EURASIP J. Advances Signal Process.*, vol. 2007, no. 92953, pp. 1–24, Dec. 2007.

[38] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 3, pp. 541–553, Mar. 2008.

[39] S. Markovich, S. Gannot, and I. Cohen, "Multichannel Eigenspace Beamforming in a Reverberant Noisy Environment With Multiple Interfering Speech Signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[40] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Sinusoidal order estimation using angles between subspaces," *EURASIP J. Advances Signal Process.*, vol. 2009, no. 948756, pp. 1–11, Dec. 2009.

[41] S. M. Nørholm, J. Benesty, J. R. Jensen, and M. G. Christensen, "Single-channel noise reduction using unified joint diagonalization and optimal filtering," *EURASIP J. Advances Signal Process.*, vol. 2014, no. 1, p. 37, Mar. 2014.

[42] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank Approximation Based Multichannel Wiener Filter Algorithms for Noise Reduction with Application in Cochlear Implants," *IEEE/ACM Trans. Audio, Speech, Lan-*

guage Process., vol. 22, no. 4, pp. 785–799, Jun. 2014.

[43] J. Benesty, J. R. Jensen, M. G. Christensen, and J. Chen, *Speech enhancement: A signal subspace perspective.* Waltham, MA, USA: Academic, 2014.

[44] J. Benesty, M. G. Christensen, and J. R. Jensen, *Signal enhancement with variable span linear filters.* Springer, 2016, vol. 7.

[45] J. R. Jensen, J. Benesty, and M. G. Christensen, "Noise reduction with optimal variable span linear filters," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 631–644, Apr. 2016.

[46] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 12, Dallas, USA, 1987, pp. 177–180.

[47] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 373–385, Jul. 1998.

[48] E. Zavarehei, S. Vaseghi, and Q. Yan, "Inter-frame modeling of DFT trajectories of speech and noise for speech enhancement using Kalman filters," *Speech Commun.*, vol. 48, no. 11, pp. 1545–1555, Nov. 2006.

[49] T. Esch and P. Vary, "Modified Kalman Filter Exploiting Interframe Correlation of Speech and Noise Magnitudes," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, WA, USA, Sep. 2008, pp. 1–4.

[50] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[51] A. Chinaev, J. Heymann, L. Drude, and R. Haeb-Umbach, "Noise-presence-probability-based noise PSD estimation by using DNNs," in *Proc. ITG Conference on Speech Commun.*, Paderborn, Germany, Oct. 2016, pp. 1–5.

[52] J. Heymann, L. Drude, and R. Haeb-Umbach, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Comput. Speech Lang.*, vol. 46, pp. 374 – 385, Nov. 2017.

[53] R. Rehr and T. Gerkmann, "On the importance of super-Gaussian speech priors for machine-learning based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 2, pp. 357–366, Feb. 2018.

[54] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1702–1726, Oct 2018.

[55] X. Li, S. Leglaive, L. Girin, and R. Horaud, "Audio-noise power spectral density estimation using long short-term memory," *IEEE Signal Process. Lett.*, vol. 26, no. 6, pp. 918–922, Jun 2019.

[56] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[57] N. Saleem and M. I. Khattak, "A review of supervised learning algorithms for single channel speech enhancement," *Int. J. Speech Technol.*, vol. 22, no. 4, pp. 1051–1075, Oct. 2019.

[58] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S. Liu, "FaSNet: Low-Latency Adaptive Beamforming for Multi-Microphone Audio Processing," in *Proc. Int. Workshop on Automat. Speech. Recogn. Understanding (ASRU)*, Singapore, Singapore, Dec. 2019, pp. 260–267.

[59] Y. Liu and D. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 12, pp. 2092–2102, Dec. 2019.

[60] M. Tammen, D. Fischer, B. T. Meyer, and S. Doclo, "DNN-Based Speech Presence Probability Estimation for Multi-Frame Single-Microphone Speech Enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 191–195.

[61] K. Tesch and T. Gerkmann, "Nonlinear spatial filtering for multichannel speech enhancement in inhomogeneous noise fields," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2020, pp. 196–200.

[62] Q. Zhang, A. M. Nicolson, M. Wang, K. Paliwal, and C. Wang, "DeepMMSE: A Deep Learning Approach to MMSE-based Noise Power Spectral Density Estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, no. 1, pp. 1404–1415, Apr. 2020.

[63] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[64] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[65] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

[66] P. Vary, "Noise suppression by spectral magnitude estimation: Mechanism and theoretical limits," *Signal Process.*, vol. 8, no. 4, pp. 387–400, Jul. 1985.

[67] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 328–337, Jul. 1998.

[68] Z. Goh, K.-C. Tan, and T. G. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 3, pp. 287–292, May 1998.

[69] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series*. MIT press Cambridge, MA, 1949, vol. 7.

[70] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *EURASIP J. Advances Signal Process.*, vol. 2003, no. 10, pp. 1043–1051, Sep. 2003.

[71] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Process. Lett.*, vol. 10, no. 7, pp. 204–207, Jul. 2003.

[72] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, no. 4, pp. 465–494, Apr. 2011.

[73] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, Mar. 2015.

[74] M. Krawczyk-Becker, "Phase-aware single-channel speech enhancement," Ph.D. dissertation, Carl von Ossietzky Universität Oldenburg, 2016.

[75] J. K.-W. Stahl, "Contributions to single-channel speech enhancement with a focus on the spectral phase," Ph.D. dissertation, Graz University of Technology, 2019.

[76] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[77] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Commun.*, vol. 42, no. 3-4, pp. 271–287, Apr. 2004.

[78] J.-H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1965–1976, Jun. 2006.

[79] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[80] ——, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Signal Process.*, vol. 86, no. 6, pp. 1215–1229, Jun. 2006.

[81] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[82] J. S. Erkelens and R. Heusdens, "Tracking of nonstationary noise based on data-driven recursive noise power estimation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 6, pp. 1112–1123, Jul. 2008.

[83] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

[84] I. Cohen, "On the decision-directed estimation approach of Ephraim and Malah," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Montreal, Canada, May 2004, pp. I–293–I–296.

[85] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, Aug. 2010.

[86] C. Breithaupt and R. Martin, "Analysis of the Decision-Directed SNR Estimator for Speech Enhancement With Respect to Low-SNR and Transient Conditions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 2, pp. 277–289, Feb. 2011.

[87] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, USA, Mar. 2008, pp. 4897–4900.

[88] T. Gerkmann and R. Martin, "On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling," *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4165–4174, Nov. 2009.

[89] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[90] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Amer.*, vol. 78, no. 5, pp. 1508–1518, Jul. 1985.

[91] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays: Signal Processing Techniques and Applications.* Springer, 2001, ch. 2, pp. 19–638.

[92] H. Cox, R. Zeskind, and T. Kooij, "Practical supergain," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 3, pp. 393–398, Jun. 1986.

[93] S. Doclo and M. Moonen, "Superdirective beamforming robust against microphone mismatch," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 2, pp. 617–631, Feb. 2007.

[94] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[95] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Process.*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004.

[96] 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 Commun. issue on Speech Enhancement*, vol. 49, no. 7-8, pp. 636–656, Jul.-Aug. 2007.

[97] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[98] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 5, pp. 1529–1539, Jul. 2007.

[99] H. Cox, "Resolving power and sensitivity to mismatch of optimum array processors," *J. Acoust. Soc. Amer.*, vol. 54, no. 3, pp. 771–785, Sep. 1973.

[100] L. Ehrenberg, S. Gannot, A. Leshem, and E. Zehavi, "Sensitivity analysis of MVDR and MPDR beamformers," in *26th Conv. Electr. Electron. Eng. in Israel IEEEI*, Eliat, Israel, Nov. 2010, pp. 416–420.

[101] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 10, pp. 1365–1376, Oct. 1987.

[102] S. A. Vorobyov, A. B. Gershman, and Z.-Q. Luo, "Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem," *IEEE Trans. Signal Process.*, vol. 51, no. 2, Feb. 2003.

[103] P. Stoica, Z. Wang, and J. Li, "Robust Capon beamforming," *IEEE Signal Process. Lett.*, vol. 10, no. 6, Jun. 2003.

[104] J. Li, P. Stoica, and Z. Wang, "On robust Capon beamforming and diagonal loading," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 1702–1715, Jul. 2003.

[105] S. Vorobyov, A. Gershman, and Z.-Q. Luo, "Adaptive beamforming with joint robustness against mismatched signal steering vector and interference nonstationarity," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 108–111, Feb. 2004.

[106] J. Li, P. Stoica, and Z. Wang, "Doubly constrained robust Capon beamformer," *IEEE Trans. Signal Process.*, vol. 52, no. 9, pp. 2407–2423, Sep. 2004.

[107] S. Vorobyov, H. Chen, and A. Gershman, "On the relationship between robust minimum variance beamformers with probabilistic and worst-case distortionless response constraints," *IEEE Trans. Signal Process.*, vol. 56, no. 11, pp. 5719–5724, Nov. 2008.

[108] A. Khabbazibasmenj, S. A. Vorobyov, and A. Hassanien, "Robust adaptive beamforming based on steering vector estimation with as little as possible prior information," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2974–2987, Feb. 2012.

[109] S. Vorobyov, "Principles of minimum variance robust adaptive beamforming design," *IEEE Trans. Signal Process.*, vol. 93, no. 12, pp. 3264–3277, Dec. 2013.

[110] Y. Zhao, J. R. Jensen, M. G. Christensen, S. Doclo, and J. Chen, "Experimental study of robust beamforming techniques for acoustic applications," in *Proc. IEEE Workshop Appl. Signal Process. Audio, Acoust. (WASPAA)*, New Paltz, NY, USA, 2017, pp. 86–90.

[111] L. W. Brooks and I. S. Reed, "Equivalence of the likelihood ratio processor, the maximum signal-to-noise ratio filter, and the Wiener filter," *IEEE Trans. Aerosp. Electron. Syst.,*, vol. AES-8, no. 5, pp. 690–692, Sep. 1972.

[112] R. Balan and J. Rosca, "Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase," in *Proc. IEEE Sensor Array Multichannel Signal Process. Workshop (SAM)*, Rosslyn, USA, Aug. 2002, pp. 209–213.

[113] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 5, pp. 1072–1077, Jul. 2010.

[114] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, Sep. 2004.

[115] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 544–548.

[116] S. Markovich-Golan, S. Gannot, and W. Kellermann, "Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function," in *Proc. of Europ. Signal Process. Conf.*

(EUSIPCO), Rome, Italy, Sep. 2018, pp. 2499–2503.

[117] R. Varzandeh, M. Taseska, and E. A. P. Habets, "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation," in *Proc. Hands-free Speech Commun. Microphone Arrays (HSCMA)*, San Francisco, USA, Mar. 2017, pp. 11–15.

[118] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 1, pp. 47–56, Mar. 2011.

[119] J. Benesty and Y. Huang, "A single-channel noise reduction MVDR filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Prague, Czech Republic, May 2011, pp. 273–276.

[120] T. Esch and P. Vary, "Speech enhancement using a modified Kalman filter based on complex linear prediction and supergaussian priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, USA, Apr. 2008, pp. 4877–4880.

[121] K. Helwani, H. Buchner, J. Benesty, and J. Chen, "A Single-Channel MVDR Filter for Acoustic Echo Suppression," *IEEE Signal Process. Lett.*, vol. 20, no. 4, pp. 351–354, Apr. 2013.

[122] D. H. Tran Vu and R. Haeb-Umbach, "Using the turbo principle for exploiting temporal and spectral correlations in speech presence probability estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Vancouver, Canada, May 2013, pp. 863–867.

[123] H. Momeni, E. A. P. Habets, and H. R. Abutalebi, "Single-channel speech presence probability estimation using inter-frame and inter-band correlations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. Florence, Italy: IEEE, May 2014, pp. 2927–2931.

[124] M. Krawczyk-Becker, D. Fischer, and T. Gerkmann, "Utilizing spectro-temporal correlations for an improved speech presence probability based noise power estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 365–369.

[125] H. Huang, C. Hofmann, W. Kellermann, J. Chen, and J. Benesty, "A multiframe parametric Wiener filter for acoustic echo suppression," in *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, Xi'an, China, Sep. 2016, pp. 1–5.

[126] H. Momeni, H. R. Abutalebi, and E. A. P. Habets, "Conditional MMSE-based single-channel speech enhancement using inter-frame and inter-band correlations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5215–5219.

[127] J. Stahl and P. Mowlaee, "Exploiting temporal correlation in pitch-adaptive speech enhancement," *Speech Commun.*, vol. 111, pp. 1–13, Aug. 2019.

[128] D. Fischer, S. Doclo, E. A. P. Habets, and T. Gerkmann, "Combined single-microphone Wiener and MVDR filtering based on speech interframe correlations and speech presence probability," in *Proc. ITG Conference on Speech Commun.*, Paderborn, Germany, Oct. 2016, pp. 292–296.

[129] D. Fischer and T. Gerkmann, "Single-microphone speech enhancement using MVDR filtering and Wiener post-filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 201–205.

[130] D. Fischer, K. Brümann, and S. Doclo, "Comparison of parameter estimation methods for single-microphone multi-frame Wiener filtering," in *Proc. of Europ. Signal Process. Conf. (EUSIPCO)*, A Coruña, Spain, Aug. 2019, pp. 1809–1813.

[131] D. Fischer and S. Doclo, "Sensitivity analysis of the multi-frame MVDR filter for single-microphone speech enhancement," in *Proc. of Europ. Signal Process. Conf. (EUSIPCO)*, Kos, Greece, Aug. 2017, pp. 633–637.

[132] ——, "Robust constrained MFMVDR filtering for single-microphone speech enhancement," in *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, Sep. 2018, pp. 41–45.

[133] ——, "Evaluation of robust constrained MFMVDR filtering for single-channel speech enhancement," in *Proc. ITG Conference on Speech Commun.*, Oldenburg, Germany, Oct. 2018, pp. 156–160.

[134] ——, "Robust constrained MFMVDR filters for single-channel speech enhancement based on spherical uncertainty set," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 618–631, Dec. 2020.

[135] ——, "Subspace-based speech correlation vector estimation for single-microphone multi-frame MVDR filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 856 – 860.

[136] ——, "Normalized speech correlation vector estimation using a low-rank speech model for multi-frame MVDR filtering," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 2020, manuscript in preparation.

[137] "ITU-T recommendation P.862. Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Feb. 2001.

[138] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*. Prentice Hall, 1988.

[139] T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Objective perceptual quality measures for the evaluation of noise reduction schemes," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Eindhoven, The Netherlands, Sep. 2005, pp. 169–172.

[140] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[141] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 5, pp. 910–919, Jul. 2008.

[142] H. Yu and T. Fingscheidt, "A weighted log kurtosis ratio measure for instrumental musical tones assessment in wideband speech," in *Proc. ITG Conference on Speech Commun.*, Braunschweig, Germany, Sep. 2012, pp. 1–4.

[143] E. Hänsler and G. Schmidt, *Acoustic echo and noise control: a practical approach*. John Wiley & Sons, 2005, vol. 40.

[144] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[145] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, Jul. 1993.

[146] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic speech database," in *National Institute of Standards and Technology (NIST)*, 1988.

[147] M. Krawczyk-Becker and T. Gerkmann, "An evaluation of the perceptual quality of phase-aware single-channel speech enhancement," *J. Acoust. Soc. Amer.*, vol. 140, no. 4, pp. EL364–EL369, Oct. 2016.

[148] M. Friedman, "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," *J. Amer. Statist. Assn*, vol. 32, no. 200, pp. 675–701, Dec. 1937.

[149] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, Dec. 1945.

[150] M. Rupert Jr. G., *Simultaneous statistical inference*, ser. Springer Series in Statistics. Springer-Verlag New York, 1981.

[151] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974.

[152] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 387–392, Apr. 1985.

[153] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, Sep. 1978.

[154] G. Schwarz, "Estimating the Dimension of a Model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

[155] N. Merhav, "The estimation of the model order in exponential families," *IEEE Trans. Inform. Theory*, vol. 35, no. 5, pp. 1109–1114, Sep. 1989.

[156] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[157] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*. Prentice-Hall, 1998, vol. 2.

[158] "Recommendation ITU-R BS.1534-3.: Method for the subjective assessment of intermediate quality level of audio systems," Oct. 2015.

[159] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, Dec. 1965.

[160] A. Field, *Discovering statistics using IBM SPSS statistics*, 3rd ed. SAGE, 2009.

[161] J. W. Mauchly, "Significance test for sphericity of a normal n-variate distribution," *Ann. Math. Statist.*, vol. 11, no. 2, pp. 204–209, Jun. 1940.

[162] S. W. Greenhouse and S. Geisser, "On methods in the analysis of profile data," *Psychometrika*, vol. 24, no. 2, pp. 95–112, Jun. 1959.

[163] R. C. Hendriks and T. Gerkmann, "Noise Correlation Matrix Estimation for Multi-Microphone Speech Enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 223–233, Jan. 2012.

[164] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 196–200.

[165] W. Jiang, F. Wen, and P. Liu, "Robust beamforming for speech recognition using DNN-based time-frequency masks estimation," *IEEE Access*, vol. 6, pp. 52 385–52 392, Sep. 2018.

[166] T. Lotter, C. Benien, and P. Vary, "Multichannel direction-independent speech enhancement using spectral amplitude estimation," *EURASIP J. Advances Signal Process.*, vol. 2003, no. 11, p. 705085, Oct. 2003.

[167] Z. Wang, P. Wang, and D. Wang, "Complex Spectral Mapping for Single- and Multi-Channel Speech Enhancement and Robust ASR," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1778–1787, May 2020.

[168] J. R. Jensen, J. Benesty, and M. G. Christensen, "Variable span filters for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. May, Shanghai, China, 2016, pp. 6505–6509.

[169] E. A. P. Habets, J. Benesty, and J. Chen, "Multi-microphone noise reduction using interchannel and interframe correlations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 305–308.

[170] J. Benesty, I. Cohen, and J. Chen, *Array Processing - Kronecker Product Beamforming*. Springer, 2019.

[171] M.-S. Song and H.-G. Kang, "Single-channel dereverberation using a non-causal minimum variance distortionless response filter," *J. Acoust. Soc. Amer.*, vol. 132, no. 1, pp. EL29–EL35, Jun. 2012.

[172] M. Parchami, W. P. Zhu, and B. Champagne, "Speech dereverberation using weighted prediction error with correlated inter-frame speech components," *Speech Commun.*, vol. 87, pp. 49–57, Mar. 2017.

LIST OF PUBLICATIONS

Peer-reviewed Journal Papers

[J1] **D. Fischer** and S. Doclo, "Normalized speech correlation vector estimation using a low-rank speech model for multi-frame MVDR Filtering," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 2020, manuscript in preparation.

[J2] **D. Fischer** and S. Doclo, "Robust constrained MFMVDR filters for single-channel speech enhancement based on spherical uncertainty set," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 618-631, Dec, 2020.

Peer-reviewed Conference Papers

[C1] **D. Fischer** and S. Doclo, "Subspace-based speech correlation vector estimation for single microphone multi-frame MVDR Filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 856 - 860.

[C2] M. Tammen, **D. Fischer**, B. T. Meyer, and S. Doclo, "DNN-Based Speech Presence Probability Estimation for Multi-Frame Single-Microphone Speech Enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 191 - 195.

[C3] **D. Fischer**, K. Brümann and S. Doclo, "Comparison of parameter estimation methods for single-microphone multi-frame Wiener Filtering," in *Proc. of Europ. Signal Process. Conf. (EUSIPCO)*, A Coruña, Spain, Aug. 2019, pp. 1809 - 1813.

[C4] **D. Fischer** and S. Doclo, "Robust constrained MFMVDR Filtering for single-microphone speech enhancement," in *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, Sep. 2018, pp. 41-45.

[C5] **D. Fischer** and S. Doclo, "Evaluation of robust constrained MFMVDR Filtering for single-channel speech enhancement," in *Proc. ITG Conference on Speech Commun.*, Oldenburg, Germany, Oct. 2018, pp. 156-160.

[C6] **D. Fischer** and S. Doclo, "Sensitivity analysis of the multi-frame MVDR Filter for single-microphone speech enhancement," in *Proc. of Europ. Signal Process. Conf. (EUSIPCO)*, Kos, Greece, Aug. 2017, pp. 633-637.

[C7] **D. Fischer**, S. Doclo, E. A. P. Habets, and T. Gerkmann, "Combined single-microphone Wiener and MVDR Filtering based on speech interframe correlations and speech presence probability," in *Proc. ITG Conference on Speech Commun.*, Paderborn, Germany, Oct. 2016, pp. 292-296.

[C8] **D. Fischer** and T. Gerkmann, "Single-microphone speech enhancement using MVDR Filtering and Wiener post-Filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 201-205.

[C9] M. Krawczyk-Becker, **D. Fischer**, and T. Gerkmann, "Utilizing spectro-temporal correlations for an improved speech presence probability based noise power estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 365-369.

Abstracts and Others

[A1] K. Brümann, **D. Fischer** and S. Doclo, "Performance comparison of single-microphone speech enhancement using speech-distortion weighted inter-frame Wiener Filters," in *Fortschritter der Akustik - DAGA 2019*, Rostock, Germany, Mar. 2019, pp. 966-969.

[A2] **D. Fischer** and S. Doclo, "Multi-frame MVDR Filtering for single-microphone speech enhancement," in 44th Erlanger Kolloquium, Erlangen, Germany, Feb. 2017.