COMBINING MODEL-BASED AND LEARNING-BASED APPROACHES FOR SPEECH ENHANCEMENT

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

von
Herrn Marvin Tammen
geboren am 15. November 1993
in Oldenburg (Deutschland)

Marvin Tammen: *Combining Model-Based and Learning-Based Approaches for Speech Enhancement*

ERSTGUTACHTER:
Prof. Dr. ir. Simon Doclo, *Carl von Ossietzky Universität Oldenburg, Deutschland*

WEITERE GUTACHTER:
Prof. Dr. Bernd T. Meyer, *Carl von Ossietzky Universität Oldenburg, Deutschland*
Prof. Dr.-Ing. Reinhold Häb-Umbach, *Universität Paderborn, Deutschland*

TAG DER DISPUTATION:
30. Juni 2025

ACKNOWLEDGMENTS

This thesis was written in the Signal Processing group at the Department of Medical Physics and Acoustics, Carl von Ossietzky Universität Oldenburg, Germany. I am grateful to the many people who walked this road with me, in whole or in part.

First and foremost, thank you, Simon, for giving me the chance to write my master's thesis—and then this dissertation—in your group. Your insightful guidance, paired with the freedom to follow my own ideas and many interesting and often fun discussions, made all the difference. I learned a lot and I'm deeply grateful.

I also thank Reinhold Häb-Umbach and Bernd T. Meyer for reviewing this thesis and for their interest in my work. I greatly appreciate the time and care you invested.

A special thanks to the current and former members of the Signal Processing group: my office mates Daniel, for lending me his ears for complaints about reviewer 2, and Henning, for successfully steering spaceships to their destinations. Wiebke, undisputed banana connoisseur; Tong, my Asian-food dealer; and Henri, Klaus, and Anselm for many interesting and fun conversations. I won't forget our shared adventures: conference trips featuring wild animals hiding in parking-lot bushes, endless coffee (and the occasional beer) chats, and fiercely contested SingStar nights.

Thanks to Nils Westhausen and Hendrik Kayser for intense yet insanely exciting collaborations. To my internship colleagues at Meta and NTT for broadening my perspective—especially Ochiai-san for introducing me to the world's best ramen spot.

Nick, thanks for being you, and for the board-game nights that let me escape stressful days into cardboard adventures.

To my family: my parents, Susanne and Gerhard, for making sure I always have somewhere to go and never have to worry about falling; my siblings, grandparents, and the rest of my family and friends for steady encouragement.

Finally, to my partner Saskia—thank you for enduring me when times were tough, for your constant support, and for helping me recharge during the weekends.

Oldenburg, September 2025

Marvin Tammen

ABSTRACT

In many speech communication devices, such as smartphones, smartspeakers, and hearing devices, the microphones capture not only the target speaker but also undesired ambient noise, degrading speech quality and speech intelligibility. Speech enhancement algorithms aim at extracting the target speech from the recorded microphone signals by suppressing noise while not distorting the target speech. Over the past decade, there has been a shift from model-based statistical signal processing approaches to learning-based data-driven approaches. Although model-based approaches offer interpretability and theoretical guarantees, they often struggle in complex, real-world acoustic scenarios where their assumptions are violated. In contrast, learning-based approaches generally achieve higher performance in such scenarios due to their strong representation capacity but may lack interpretability, theoretical guarantees, and robustness when the data observed during inference does not match the training data.

Motivated by the potential to combine the interpretability of model-based approaches with the strong representation capacity of learning-based approaches, the primary objective of this thesis is to develop and evaluate hybrid speech enhancement algorithms that employ a learning-based stage to estimate quantities required by a model-based enhancement stage. The main focus is on investigating whether imposing structure on the estimated quantities—such as correlation matrix structure, correlation vector structure, or spatial structure—improves speech enhancement performance, interpretability, and computational complexity. Another focus is on developing geometry-robust hybrid speech enhancement algorithms that can operate with arbitrary microphone array configurations. While the developed algorithms can be used for various speech enhancement applications, our focus is on hearing devices, where low latency is crucial. To this end, we mainly consider causal multi-frame filters in the short-time Fourier transform domain as the model-based enhancement stage, leveraging their inherent low-latency capabilities.

As a first contribution, we propose a hybrid single-microphone speech enhancement approach by embedding the multi-frame minimum variance distortionless response (MFMVDR) filter within a deep learning framework, imposing structure on the required temporal covariance matrices. Simulation results using the deep noise suppression (DNS) 1 challenge dataset demonstrate that the resulting deep MFMVDR filter improves speech enhancement performance compared to a purely learning-based algorithm that does not impose the MFMVDR structure on the filter coefficients. Additionally, imposing structure on the temporal covariance matrices reduces computational complexity while maintaining speech enhancement performance.

Second, we extend the hybrid single-microphone approach to multi-microphone speech enhancement for binaural hearing devices by embedding the binaural spatio-temporal Wiener filter within a deep learning framework, imposing structure on the required spatio-temporal correlation vectors. Simulation results using the DNS 1, DNS 2, CEC 1, and CEC 3 datasets demonstrate that the Kronecker factorization of the speech spatio-temporal correlation vectors into a spatial factor (the relative transfer function (RTF) vector) and a temporal factor reduces computational complexity while maintaining speech enhancement performance and preserving binaural cues, outperforming two causal state-of-the-art binaural speech enhancement algorithms.

Third, we investigate the acoustic interpretability of the estimated RTF vector in the Kronecker factorization of the speech spatio-temporal correlation vector. Since the estimated RTF vector does not reflect the spatial characteristics of the acoustic scenario, we propose a spatial regularization procedure to improve interpretability by imposing spatial structure. Simulation results using the CHiME-3 microphone array demonstrate that the proposed spatial regularization procedure yields accurate estimates of the RTF vector even in reverberant environments without sacrificing speech enhancement performance or increasing computational complexity.

Finally, we propose three procedures to improve the robustness of the mask-based beamformer with attention-based spatial covariance matrix aggregator (ASA) against varying microphone array configurations. These procedures include incorporating random microphone array configurations during training, employing the transform-average-concatenate (TAC) method, and using geometry-robust input features. Simulation results for a moving source using the CHiME-3 and DEMAND microphone arrays demonstrate that the combination of these procedures enables the application to unseen microphone array configurations, consistently outperforming both a baseline mask-based beamformer with recursive smoothing and the original mask-based beamformer with ASA.

ZUSAMMENFASSUNG

In vielen Sprachkommunikationsgeräten wie Smartphones, Smartspeakern und Hörgeräten erfassen die Mikrofone nicht nur den Zielsprecher, sondern auch unerwünschte Umgebungsgeräusche, was die Sprachqualität und Sprachverständlichkeit beeinträchtigt. Sprachverbesserungsalgorithmen zielen darauf ab, die Zielsprache aus den aufgenommenen Mikrofonsignalen zu extrahieren, indem sie Störgeräusche unterdrücken, ohne die Zielsprache zu verzerren.

Innerhalb des letzten Jahrzehnts hat sich der Fokus von modellbasierten, statistischen Signalverarbeitungsansätzen hin zu lernbasierten, datengetriebenen Ansätzen verlagert. Während modellbasierte Ansätze Interpretierbarkeit und theoretische Garantien bieten, stoßen sie oft in komplexen, realen akustischen Szenarien an ihre Grenzen, in denen ihre zugrundeliegenden Annahmen nicht erfüllt sind. Im Gegensatz dazu erreichen lernbasierte Ansätze in solchen Szenarien aufgrund ihrer starken Repräsentationsfähigkeit üblicherweise eine höhere Leistungsfähigkeit, weisen jedoch potenzielle Nachteile hinsichtlich Interpretierbarkeit, theoretischer Fundierung und Robustheit auf, insbesondere wenn die Daten während der Inferenz nicht mit den Trainingsdaten übereinstimmen.

Motiviert durch das Potenzial, die Interpretierbarkeit modellbasierter Ansätze mit der starken Repräsentationsfähigkeit lernbasierter Ansätze zu kombinieren, besteht das Hauptziel dieser Dissertation darin, hybride Sprachverbesserungsalgorithmen zu entwickeln und zu evaluieren, die eine lernbasierte Stufe zur Schätzung von Größen nutzen, die für eine modellbasierte Signalverarbeitungsstufe erforderlich sind. Der Schwerpunkt liegt darauf, zu untersuchen, ob das Erzwingen von Struktur für die geschätzten Größen – beispielsweise Korrelationsmatrixstruktur, Korrelationsvektorstruktur oder räumliche Struktur – die Sprachverbesserungsleistung, Interpretierbarkeit und rechnerische Effizienz verbessert. Ein weiterer Fokus liegt auf der Entwicklung geometrie-robuster hybrider Sprachverbesserungsalgorithmen, die mit beliebigen Mikrofonanordnungen arbeiten können. Während die entwickelten Algorithmen für verschiedene Sprachverbesserungsanwendungen einsetzbar sind, liegt der Fokus auf Hörgeräten, bei denen geringe Latenz essenziell ist. Daher betrachten wir vorrangig kausale Multi-Frame-Filter im Short-Time-Fourier-Transform-Bereich als die modellbasierte Signalverarbeitungsstufe, da diese inhärent niedrige algorithmische Latenzen ermöglichen.

Als erster Beitrag dieser Arbeit wird ein hybrider Einzelmikrofon-Sprachverbesserungsansatz vorgeschlagen, bei dem das MFMVDR-Filter in ein Deep-Learning-Framework eingebettet wird und eine Struktur für die benötigten zeitlichen Kovarianzmatrizen erzwungen wird. Simulationsergebnisse basierend auf

dem DNS 1-Datensatz zeigen, dass das resultierende tiefe MFMVDR-Filter die Sprachverbesserungsleistung im Vergleich zu einem rein lernbasierten Algorithmus, der keine MFMVDR-Struktur für die Filterkoeffizienten erzwingt, verbessert. Zusätzlich reduziert das Erzwingen von Struktur für die zeitlichen Kovarianzmatrizen die rechnerische Komplexität, ohne die Sprachverbesserungsleistung zu beeinträchtigen.

Als zweiter Beitrag wird der hybride Einzelmikrofon-Ansatz auf mehrkanalige Sprachverbesserung für binaurale Hörgeräte erweitert, indem das binaurale räumlich-zeitliche Wiener Filter in ein Deep-Learning-Framework eingebettet und eine Struktur für die benötigten räumlich-zeitlichen Sprachkovarianzvektoren erzwungen wird. Simulationsergebnisse basierend auf den DNS 1-, DNS 2-, CEC 1- und CEC 3-Datensätzen zeigen, dass die Kronecker-Faktorisierung der räumlich-zeitlichen Sprachkovarianzvektoren in einen räumlichen Faktor – den Vektor der relativen Übertragungsfunktionen (RTF-Vektor) – und einen zeitlichen Faktor die rechnerische Komplexität reduziert, während die Sprachverbesserungsleistung erhalten bleibt und binaurale Cues bewahrt werden. Dabei wird eine bessere Leistung erzielt als mit zwei kausalen, dem aktuellen Stand der Technik entsprechenden binauralen Sprachverbesserungsalgorithmen.

Als dritter Beitrag wird die akustische Interpretierbarkeit des geschätzten RTF-Vektors in der Kronecker-Faktorisierung der räumlich-zeitlichen Sprachkovarianzvektoren untersucht. Da der geschätzte RTF-Vektor nicht die räumlichen Eigenschaften des akustischen Szenarios widerspiegelt, wird ein räumliches Regularisierungsverfahren vorgeschlagen, um die Interpretierbarkeit durch das Auferlegen räumlicher Struktur zu verbessern. Simulationsergebnisse mit der CHiME-3-Mikrofonanordnung zeigen, dass das vorgeschlagene räumliche Regularisierungsverfahren auch in halligen Umgebungen eine präzise Schätzung des RTF-Vektors ermöglicht, ohne die Sprachverbesserungsleistung zu beeinträchtigen oder die rechnerische Komplexität zu erhöhen.

Abschließend werden drei Verfahren vorgeschlagen, um die Robustheit des maskenbasierten Beamformers mit aufmerksamkeitsbasierter Kovarianzmatrixaggregation gegenüber variierenden Mikrofonarray-Konfigurationen zu verbessern. Diese Verfahren umfassen die Einbeziehung zufälliger Mikrofonarray-Konfigurationen während des Trainings, die Anwendung der transform-average-concatenate (TAC)-Methode und die Nutzung geometrie-robuster Eingangsfeatures. Simulationsergebnisse für einen beweglichen Zielsprecher mit den CHiME-3- und DEMAND-Mikrofonanordnungen zeigen, dass die Kombination dieser Verfahren eine Anwendung auf zuvor ungesehene Mikrofonanordnungen ermöglicht und dabei sowohl einen maskenbasierten Beamformer mit rekursiver Glättung als auch den ursprünglichen maskenbasierten Beamformer mit aufmerksamkeitsbasierter Kovarianzmatrixaggregation durchgängig übertrifft.

GLOSSARY

Acronyms and Abbreviations

ASA attention-based spatial covariance matrix aggregator
ATF acoustic transfer function

BCCTN binaural complex convolutional transformer network
BN bottleneck
BRIR binaural room impulse response

CD Cholesky decomposition
CEC Clarity Enhancement Challenge
CNN convolutional neural network
COSPA complex-valued spatial autoencoder
CRNN convolutional recurrent neural network
CW covariance whitening

DCCRN-MC deep complex convolutional recurrent network
DCUNET-MC deep complex U-Net
DDA decision-directed approach
DF deep filter
DNN deep neural network
DNS deep noise suppression
DNSMOS Deep Noise Suppression Mean Opinion Score
DOA direction of arrival
DPRNN dual-path recurrent neural network

FLOPS floating point operations per second
FSB-LSTM full- and subband long short-term memory
FT-JNF frequency-time joint nonlinear filter

GRU gated recurrent unit

HASPI hearing aid speech perception index
HASQI hearing aid speech quality and speech intelligibility index

ILD interaural level difference
IPD interaural phase difference
ISCM instantaneous SCM
ITD interaural time difference

LCMV linearly constrained minimum variance
LSTM long short-term memory

MACS multiply-accumulate operations per second
MF matched filter
MFMVDR multi-frame minimum variance distortionless response
MHA multi-head attention
MMSE minimum mean square error
MSE mean square error
MVDR minimum variance distortionless response

OLA overlap-add

PDT positive-definite Toeplitz
PESQ perceptual evaluation of speech quality
POLQA perceptual objective listening quality assessment
PReLU parametric rectified linear unit
PSD power spectral density
PYIN probabilistic YIN (from Chinese yin and yang philosophy)

R1 rank-1
RF real-time factor
RIR room impulse response
RNN recurrent neural network
RS recursive smoothing
RTF relative transfer function

SCM spatial covariance matrix
SDI speech distortion index
SDR signal-to-distortion ratio
SHA single-head attention
SI-SDR scale-invariant signal-to-distortion ratio
SIR signal-to-interference ratio
SNR signal-to-noise ratio
SPP speech presence probability
STCM spatio-temporal covariance matrix
STCV spatio-temporal correlation vector
STFT short-time Fourier transform
STOI short-time objective intelligibility
STWF spatio-temporal Wiener filter

TAC transform-average-concatenate
TCM temporal covariance matrix
TCN temporal convolutional network
TCV temporal correlation vector

WSJ0 Wall Street Journal 0

Mathematical Notation and Operators

x	scalar x
$\mathbf{x}$	vector $\mathbf{x}$
$\mathbf{X}$	matrix $\mathbf{X}$
$\cdot^{\mathsf{T}}$	transpose operator
$\cdot^{*}$	complex conjugate operator
$\cdot^{\mathsf{H}}$	complex conjugate transpose operator
$\lfloor \cdot \rfloor$	flooring operator
$*$	convolution operator
$\otimes$	Kronecker product
$\widehat{\cdot}$	estimate of $\cdot$
$\mathrm{trace}(\cdot)$	trace operator
$\mathbb{E}(\cdot)$	expectation operator
$\lvert \cdot \rvert$	magnitude of $\cdot$
$\lVert \mathbf{x} \rVert_p$	p-norm of vector $\mathbf{x}$
$\lVert \mathbf{X} \rVert_{\mathrm{F}}$	Frobenius norm of matrix $\mathbf{X}$
$\Re(\cdot)$	real part of $\cdot \in \mathbb{C}$
$\Im(\cdot)$	imaginary part of $\cdot \in \mathbb{C}$
$\angle(\cdot)$	wrapped phase of $\cdot$
$\mathbf{X}^{-1}$	inverse of matrix $\mathbf{X}$
$\mathrm{vec}(\mathbf{X})$	vectorized matrix $\mathbf{X}$
$\mathrm{ReLU}(\cdot)$	rectified linear unit activation function
$\mathrm{PReLU}(\cdot)$	parametric rectified linear unit activation function
$\mathrm{sigmoid}(\cdot)$	sigmoid activation function
$\mathrm{softmax}(\cdot)$	softmax activation function

Fixed Symbols

$\mathbb{R}$	set of real numbers
$\mathbb{C}$	set of complex numbers
j	imaginary unit ($j^2 = -1$)
m	microphone index

M_L	number of microphones in left hearing device
M_R	number of microphones in right hearing device
M	total number of microphones
L	reference microphone at the left hearing device
R	reference microphone at the right hearing device
f	frequency bin index
F	total number of frequency bins
t	time frame index
T	total number of time frames
t_d	sample index in the time domain
T_d	total number of samples in the time domain
N_d	frame length
$y_{f,t,m}$	noisy microphone signal
$x_{f,t,m}$	target speech component
$n_{f,t,m}$	noise component
$x'_{f,t,m}$	uncorrelated speech component
$i_{f,t,m}$	interference component
$\phi^m_{x,f,t}$	speech PSD
$\phi^m_{n,f,t}$	noise PSD
$\phi^m_{i,f,t}$	interference PSD
$m^{\mathbb{R}}_{f,t,m}$	real-valued time-frequency mask
$m^{\mathbb{C}}_{f,t,m}$	complex-valued time-frequency mask
$a_{x,t,\tau}$	speech attention weight relating time frame t to τ
$a_{n,t,\tau}$	noise attention weight relating time frame t to τ
$h^r_{f,t,m}$	RTF relating the speech component at the m-th microphone to the reference microphone r

$\mathbf{y}_{f,t}$	noisy multi-microphone multi-frame vector
$\mathbf{x}_{f,t}$	speech multi-microphone multi-frame vector
$\mathbf{n}_{f,t}$	noise multi-microphone multi-frame vector
$\mathbf{x}'_{f,t}$	uncorrelated speech multi-microphone multi-frame vector
$\mathbf{i}_{f,t}$	interference multi-microphone multi-frame vector
$\check{\mathbf{y}}_{f,t}$	noisy multi-microphone vector
$\check{\mathbf{x}}_{f,t}$	speech multi-microphone vector
$\check{\mathbf{n}}_{f,t}$	noise multi-microphone vector
$\bar{\mathbf{y}}_{f,t,m}$	noisy multi-frame vector
$\bar{\mathbf{x}}_{f,t,m}$	speech multi-frame vector

$\bar{\mathbf{n}}_{f,t,m}$	noise multi-frame vector
$\bar{\mathbf{x}}'_{f,t,m}$	uncorrelated speech multi-frame vector
$\bar{\mathbf{i}}_{f,t,m}$	interference multi-frame vector
$\check{\mathbf{h}}_{f,t,\text{glob}}$	global RTF vector
$\check{\mathbf{h}}^{L}_{f,t,\text{ipsi}}$	ipsilateral RTF vector of the left hearing device
$\check{\mathbf{h}}^{R}_{f,t,\text{ipsi}}$	ipsilateral RTF vector of the right hearing device
$\check{\mathbf{b}}_{t}$	ATF vector between the speech source and each microphone
$\boldsymbol{\gamma}^{m}_{f,t}$	speech spatio-temporal correlation vector between the N most recent speech STFT coefficients at each microphone and the current speech STFT coefficient at the m-th microphone.
$\boldsymbol{\gamma}^{L}_{f,t,m}$	speech spatio-temporal correlation vector between the N most recent speech STFT coefficients at the m-th microphone and the current speech STFT coefficient at the left hearing device's reference microphone.
$\boldsymbol{\gamma}^{R}_{f,t,m}$	speech spatio-temporal correlation vector between the N most recent speech STFT coefficients at the m-th microphone and the current speech STFT coefficient at the right hearing device's reference microphone.
$\mathbf{w}^{r}_{f,t}$	spatio-temporal filter vector for reference microphone r
$\check{\mathbf{w}}^{r}_{f,t}$	spatial filter vector for reference microphone r
$\bar{\mathbf{w}}^{r}_{f,t}$	temporal filter vector for reference microphone r
$\mathbf{e}_{m}$	spatio-temporal selection vector with a 1 at the position corresponding to the current frame of the m-th microphone and 0 elsewhere
$\check{\mathbf{e}}_{m}$	spatial selection vector with a 1 at the position corresponding to the m-th microphone and 0 elsewhere
$\bar{\mathbf{e}}$	temporal selection vector with a 1 at the first position (corresponding to the current frame) and 0 elsewhere
$\boldsymbol{\chi}_{f,t,m}$	feature vector at the m-th microphone
$\mathbf{a}_{x,t}$	speech attention weights relating time frame t to all time frames
$\mathbf{a}_{n,t}$	noise attention weights relating time frame t to all time frames
$\boldsymbol{\Phi}_{y,f,t}$	noisy spatio-temporal covariance matrix
$\boldsymbol{\Phi}_{x,f,t}$	speech spatio-temporal covariance matrix
$\boldsymbol{\Phi}_{n,f,t}$	noise spatio-temporal covariance matrix
$\boldsymbol{\Phi}^{L}_{x',f,t}$	uncorrelated speech spatio-temporal covariance matrix for the left hearing device
$\boldsymbol{\Phi}^{R}_{x',f,t}$	uncorrelated speech spatio-temporal covariance matrix for the right hearing device

$\mathbf{\Phi}_{i,f,t}^{L}$	interference spatio-temporal covariance matrix for the left hearing device
$\mathbf{\Phi}_{i,f,t}^{R}$	interference spatio-temporal covariance matrix for the right hearing device
$\tilde{\mathbf{\Phi}}_{y,f,t}^{LL}$	submatrix of $\mathbf{\Phi}_{y,f,t}$ containing only correlations between the microphones of the left hearing device
$\tilde{\mathbf{\Phi}}_{y,f,t}^{RR}$	submatrix of $\mathbf{\Phi}_{y,f,t}$ containing only correlations between the microphones of the right hearing device
$\tilde{\mathbf{\Phi}}_{y,f,t}^{LR}$	submatrix of $\mathbf{\Phi}_{y,f,t}$ containing only correlations between contralateral microphones
$\check{\mathbf{\Phi}}_{y,f,t}$	noisy spatial covariance matrix
$\check{\mathbf{\Phi}}_{x,f,t}$	speech spatial covariance matrix
$\check{\mathbf{\Phi}}_{n,f,t}$	noise spatial covariance matrix
$\widehat{\check{\mathbf{\Psi}}}_{x,f,t}$	instantaneous speech spatial covariance matrix estimate
$\widehat{\check{\mathbf{\Psi}}}_{n,f,t}$	instantaneous noise spatial covariance matrix estimate
$\bar{\mathbf{\Phi}}_{y,f,t}$	noisy temporal covariance matrix
$\bar{\mathbf{\Phi}}_{x,f,t}$	speech temporal covariance matrix
$\bar{\mathbf{\Phi}}_{n,f,t}$	noise temporal covariance matrix
$\bar{\mathbf{\Phi}}_{x',f,t}^{r}$	uncorrelated speech temporal covariance matrix w.r.t. reference microphone r
$\bar{\mathbf{\Phi}}_{i,f,t}^{r}$	interference temporal covariance matrix w.r.t. reference microphone r

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

Speech is the primary means of human communication, enabling both information exchange and emotional expression. Technical advances in recent decades have enabled the widespread integration of speech communication technology into various devices, including smartphones, smartwatches, smartspeakers, headphones, and hearing devices. These devices rely on one or more microphones to capture the target speech signal but inevitably also pick up undesired ambient noise and reverberation, degrading speech quality and speech intelligibility. While all users are affected, this poses a particular challenge for listeners with hearing impairment, who struggle to understand speech even at signal-to-noise ratios where normal-hearing listeners maintain good intelligibility [1]. Moreover, humans rely on binaural cues—differences in time of arrival, level, and spectral content between the ears—to localize sound sources and improve the separation of target speech from noise in adverse acoustic scenarios. However, listeners with hearing impairment frequently have reduced sensitivity to these binaural cues, further impairing their ability to understand the target speaker. As a result, hearing impairment can negatively impact speech communication, social engagement, and cognitive health, which increases the risk of depression and cognitive decline, particularly in older adults [2]. Hence, among the various applications of speech enhancement algorithms, hearing devices represent one of the most impactful applications, since effective noise suppression can greatly improve the auditory experience for users with hearing impairment. However, also normal-hearing listeners can benefit from speech enhancement in adverse acoustic scenarios, highlighting the need for robust speech enhancement algorithms that enhance the target speech signal across a wide range of devices and scenarios.

Many speech enhancement approaches have been proposed, which can be broadly categorized into model-based and learning-based approaches [3]–[9]. Model-based approaches rely on explicit assumptions about speech, noise, and the acoustic scenario, often incorporating models of spectro-temporal signal characteristics and of sound propagation. Although their theoretical foundation often provides interpretability and theoretical guarantees, model-based approaches are inherently limited by their assumptions; when these assumptions are violated, speech enhancement performance degrades. Moreover, model-based approaches often require quantities that cannot be reliably estimated from the noisy microphone signals, particularly

in adverse acoustic scenarios. In contrast, learning-based approaches avoid explicit assumptions and instead learn complex relationships between the noisy microphone signals and the target speech signal directly from data, thereby relying on implicit assumptions instead. Learning-based approaches have achieved impressive performance across nearly all engineering and scientific applications [10], including speech enhancement even in adverse acoustic scenarios [8], [9]. However, the "black-box" nature of many learning-based approaches makes their behavior difficult to interpret. Additionally, learning-based approaches typically exhibit a strong dependence on their training data, including the speech and noise material as well as the acoustic environment, which may limit their generalizability and make their performance difficult to predict in conditions not represented in the training data. For instance, in multi-microphone speech enhancement, many learning-based algorithms overfit to a specific array geometry, rendering them unsuitable for applications involving ad-hoc microphone arrays or the possibility of microphone failures.

Motivated by the potential to combine the interpretability of model-based approaches with the strong representation capacity of learning-based approaches, the primary objective of this thesis is to **develop and evaluate hybrid single- and multi-microphone speech enhancement algorithms that employ deep neural networks to estimate the quantities required by a model-based enhancement stage**. The main focus is on investigating whether imposing structure on estimated quantities—such as correlation matrix structure, correlation vector structure, or spatial structure—improves speech enhancement performance, interpretability, and computational complexity. Another focus is on developing geometry-robust hybrid speech enhancement algorithms that can operate with arbitrary microphone array configurations. While the proposed algorithms can be used for various speech enhancement applications, the main focus is on hearing devices, where low latency is crucial. To this end, we mainly consider causal multi-frame filters in the short-time Fourier transform (STFT) domain as the model-based enhancement stage, leveraging their inherent low-latency capabilities and applicability to dynamic acoustic scenarios.

The remainder of this chapter is organized as follows. In Section 1.1, we describe the general acoustic scenario considered in this thesis. In Section 1.2, we provide an overview of model-based and learning-based speech enhancement approaches and discuss their respective advantages and limitations. In Section 1.2.3, we provide an overview of hybrid speech enhancement approaches, which combine model-based and learning-based approaches. Finally, in Section 1.3, we present an outline of the remaining chapters of this thesis and summarize the main contributions.

1.1 Acoustic Scenario

Real-world acoustic scenarios involve complex mixtures of sounds reaching a listener's ears. The target speaker often competes with multiple noise sources, degrading speech quality and speech intelligibility. In this thesis, we consider multiple speech enhancement configurations that allow for different levels of spatial process-

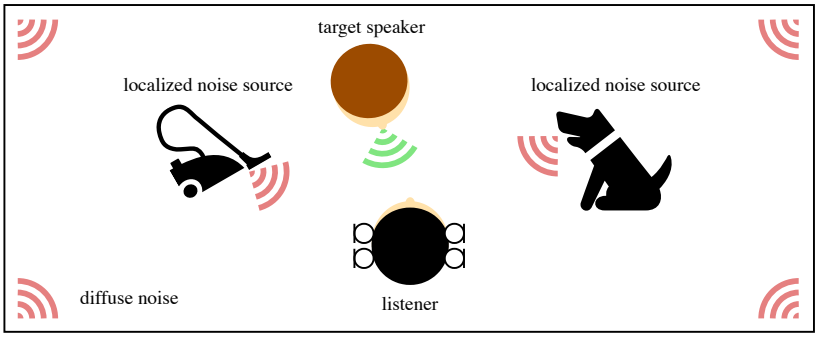


Figure 1.1: Acoustic scenario with a listener wearing binaural hearing devices, a target speaker, two localized (non-speech) noise sources, and diffuse noise, enclosed within a reverberant room.

ing: single-microphone monaural speech enhancement, where no spatial information is available; multi-microphone monaural speech enhancement, which benefits from spatial information within a single hearing device; and multi-microphone binaural speech enhancement, which can additionally exploit binaural cues for improved source separation and spatial awareness. Figure 1.1 illustrates the most general configuration considered in this thesis, where a listener wearing binaural hearing devices is exposed to a target speaker, two localized (non-speech) noise sources, and diffuse noise, all enclosed within a reverberant room. Each device is equipped with multiple microphones that record the resulting mixture. The speech enhancement problem investigated in this thesis consists of extracting the target speech component from the recorded microphone signals while suppressing the noise component, ensuring that the target speech component remains undistorted.

In this section, we describe the acoustic scenario investigated in this thesis, including the target speech source, noise sources, acoustic environment, and binaural cues, before summarizing the acoustic cues that can be exploited for speech enhancement.

1.1.1 *Target Speech Source*

To motivate the use of multi-frame filters, which leverage the temporal correlation of speech, we first describe the target speech source and introduce properties that are particularly relevant to this thesis.

In terms of spatial properties, the target speech source is assumed to be a directional source, producing coherent wavefronts that result in characteristic spatial coherence patterns between signals recorded at the microphones. These patterns depend on frequency as well as the relative positioning of the microphones and the target speech source.

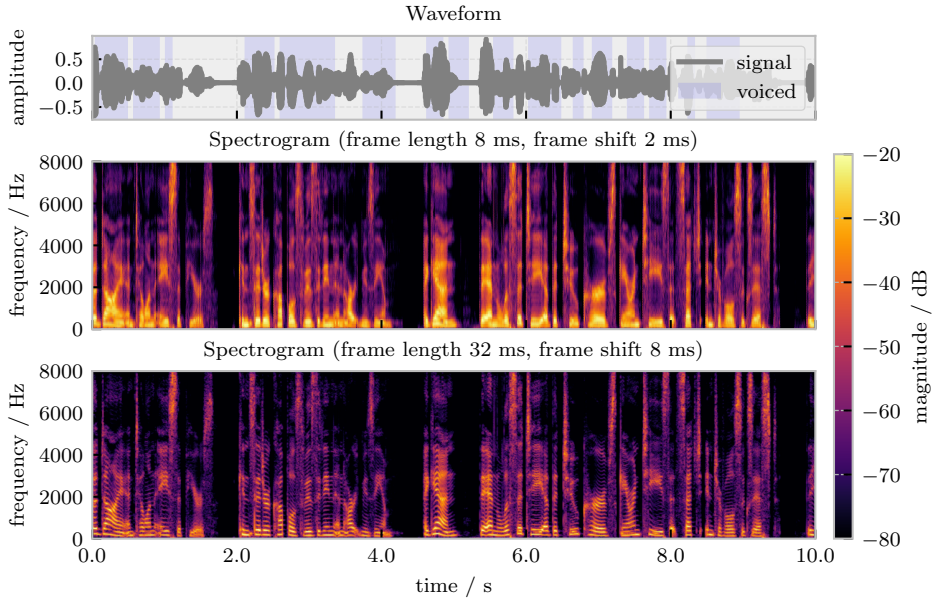


Figure 1.2: Female speech signal, visualized as a waveform with voiced speech segments shaded in blue (determined using the probabilistic YIN (PYIN) algorithm [11]), a spectrogram with a frame length of 8 ms and a frame shift of 2 ms, as well as a spectrogram with a frame length of 32 ms and a frame shift of 8 ms, both using a $\sqrt{\text{Hann}}$ window.

To introduce spectro-temporal properties, we start by briefly summarizing the human speech production mechanism. This mechanism involves airflow from the lungs that is modulated by the vocal cords and filtered by the vocal tract resonances, generating two primary types of speech sounds. Voiced speech occurs when the vocal cords vibrate periodically, whereas unvoiced speech results from turbulent airflow through constrictions in the vocal tract. Time-frequency transforms such as the STFT [12] can reveal distinct spectro-temporal patterns for these different speech sound types.

Voiced speech is characterized by a harmonic structure that includes a fundamental frequency (typically 85 Hz to 180 Hz for male speakers and 165 Hz to 255 Hz for female speakers) and overtones, where most acoustic energy is distributed below 5 kHz. The resonances of the vocal tract determine the spectral envelope of speech and encode information required for speech intelligibility. To reveal this structure, the STFT settings should be chosen to result in sufficient spectral resolution. As illustrated in Fig. 1.2, the bottom spectrogram, computed with a frame length of 32 ms, provides sufficient spectral resolution, allowing the periodic structure of voiced speech to be clearly resolved. In contrast, the top spectrogram, computed with a shorter frame length of 8 ms, lacks the spectral resolution necessary to resolve the periodic structure.

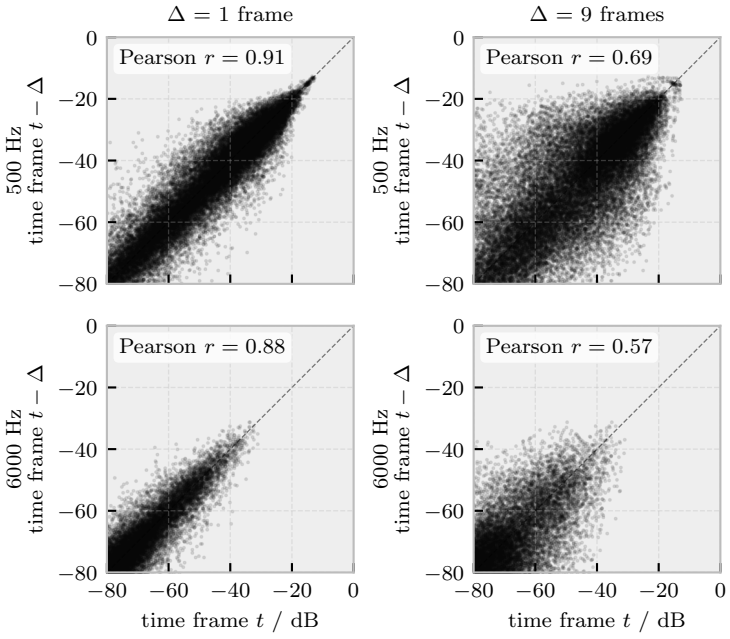


Figure 1.3: Scatter plots of STFT magnitudes (frame length 8 ms, frame shift 2 ms, $\sqrt{\text{Hann}}$ window) with a lag of 1 or 9 frames at 500 Hz and 6000 Hz, including the corresponding Pearson correlation coefficient r, computed from four female and four male speech signals.

Compared to voiced speech, unvoiced speech lacks a harmonic structure and instead appears as noise-like signals with energy distributed across a broad frequency range, typically extending from 4 kHz to 12 kHz (see, e.g., the unvoiced speech segment around 1.6 s in Fig. 1.2).

Overall, speech exhibits considerable spectro-temporal sparsity, with energy concentrated in specific time-frequency regions corresponding to harmonics and formants [13]. Sparsity can be observed both within individual time frames, where energy is localized at certain frequencies, and within individual frequency bins, where speech alternates between voiced segments, unvoiced segments, and speech pauses, resulting in characteristic temporal patterns. These patterns vary across different timescales. Over short durations of 10 ms to 30 ms, speech can be considered quasi-stationary, particularly at low frequencies, meaning its statistical properties remain relatively constant [14]. Over longer durations exceeding 100 ms, speech is highly non-stationary, including switches between different phonemes. Figure 1.3 illustrates the strong temporal correlation of speech in the STFT domain by showing scatter plots of consecutive STFT magnitudes with a frame lag of $\Delta = 1$ or $\Delta = 9$ at 500 Hz and 6000 Hz (using an STFT with a frame length of 8 ms, a frame shift of 2 ms, and a $\sqrt{\text{Hann}}$ window). As expected, the highest temporal correlation occurs

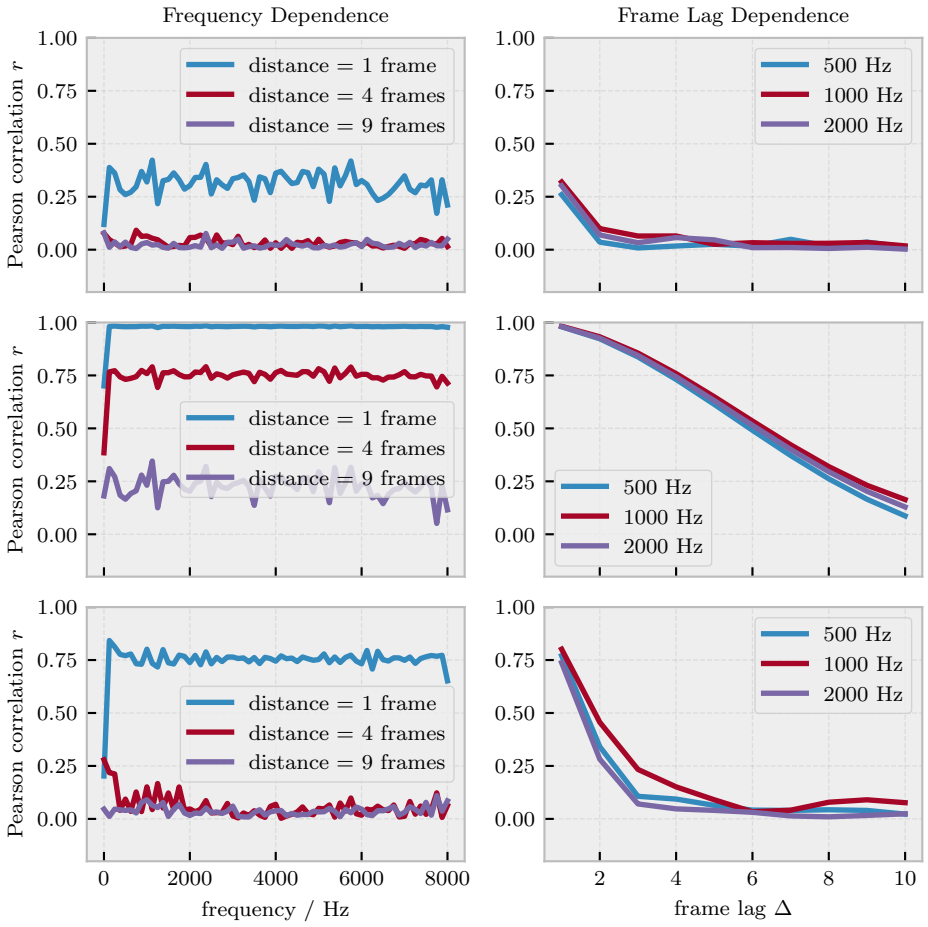


Figure 1.4: Absolute value of Pearson correlation coefficients between consecutive STFT coefficients of speech signals for an STFT with 32 ms frame length and 16 ms frame shift (top), an STFT with 32 ms frame length and 2 ms frame shift (center), and an STFT with 8 ms frame length and 2 ms frame shift (bottom), all using a $\sqrt{\text{Hann}}$ window, computed from four female and four male speech signals. Left: as a function of frequency for different frame lags Δ (1, 4, and 9 frames corresponding to 2 ms, 8 ms, and 18 ms). Right: as a function of frame lag Δ for different frequencies (500 Hz, 1000 Hz, and 2000 Hz).

at a frame lag of $\Delta = 1$ and 500 Hz (reflected by a Pearson correlation coefficient of $r = 0.91$), while the lowest temporal correlation occurs at a frame lag of $\Delta = 9$ and 6000 Hz (reflected by a Pearson correlation coefficient of $r = 0.57$). Although these magnitude-based scatter plots provide an intuitive impression of how consecutive speech frames are correlated in terms of energy, they do not capture any phase relationships. Since many multi-frame speech enhancement algorithms rely on complex-valued STFT correlations (which incorporate both magnitude and phase), it is instructive to analyze the correlation of consecutive complex-valued speech STFT coefficients rather than just their magnitudes. Accordingly, Fig. 1.4 depicts the absolute value of the Pearson correlation coefficient between consecutive complex-valued speech STFT coefficients as a function of frequency for different frame lags Δ (left) and as a function of frame lag Δ for different frequencies (right). The results are shown for three different STFT configurations, ordered to highlight the trade-offs in spectral resolution, temporal correlation, and computational complexity:

- common STFT configuration for speech enhancement (32 ms frame length, 16 ms shift, top): This configuration provides good spectral resolution due to the long frame length but exhibits relatively low temporal correlation due to small frame overlap. Additionally, this configuration maintains a relatively low computational complexity because, despite the large number of frequency bins, the number of frames per second is small.
- common frame length with a small frame shift (32 ms frame length, 2 ms shift, center): This configuration preserves the good spectral resolution of the previous configuration while significantly increasing temporal correlation due to the greater frame overlap. However, the increased number of frames per second substantially increases computational complexity, as a larger number of STFT frames must be processed per second.
- short frame length with a small frame shift (8 ms frame length, 2 ms shift, bottom): This configuration achieves high temporal correlation due to the small frame shift while maintaining a relatively low computational complexity, as the number of frames per second is high but the number of frequency bins per frame is low. However, the shorter frame length results in reduced spectral resolution, which could be a disadvantage for speech enhancement algorithms.

As can be observed, temporal correlation decreases with increasing frame lag Δ for all configurations but is largely independent of frequency. The configurations with a small frame shift (center and bottom) clearly exhibit a stronger temporal correlation than the configuration with a large frame shift (top), as smaller shifts increase the proportion of shared signal content between consecutive frames. Comparing the configurations with a small frame shift, the long frame length (32 ms) results in higher spectral resolution and stronger temporal correlation than the short frame length (8 ms), but at the cost of significantly higher computational complexity. This analysis highlights fundamental trade-offs in selecting an appropriate STFT configuration: First, a longer frame length improves spectral resolution and—when combined with a small frame shift—enhances temporal correlation. However, it also

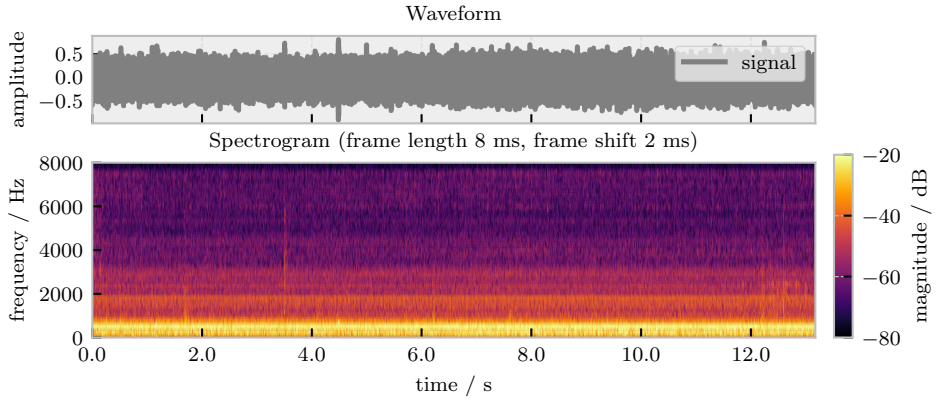


Figure 1.5: Continuous PC fan noise, visualized as a waveform and a spectrogram.

increases computational complexity due to the larger number of frequency bins and introduces greater algorithmic latency.[1] Conversely, a shorter frame length reduces computational complexity and latency but at the expense of spectral resolution. Second, a smaller frame shift increases temporal correlation but also computational complexity due to the increased number of frames per second. In this thesis, we seek a balance between high temporal correlation, low computational complexity, and small algorithmic latency. Fortunately, multi-frame speech enhancement algorithms can compensate for lower spectral resolution by processing multiple overlapping frames simultaneously, effectively increasing the analysis window without introducing additional latency. Hence, we primarily employ an STFT with a short frame length of 8 ms and a small frame shift of 2 ms unless stated otherwise.

1.1.2 *Noise Sources*

The noise sources present in real-world acoustic scenarios exhibit diverse spatial and spectro-temporal characteristics that often differ from target speech. Spatially, noise signals typically originate from either localized sources or spatially (quasi-)diffuse sound fields. Similarly to the target speech source, localized noise sources, such as moving vehicles or machinery, produce coherent wavefronts from specific directions, resulting in characteristic spatial coherence patterns between signals recorded at different positions. In contrast, spatially diffuse sound fields can originate from multiple distributed sources around the microphones, such as restaurant babble, leading to low coherence between signals captured at different microphones, particularly at

1 Approaches such as asymmetric analysis and synthesis windows or predicted future STFT frames can help mitigate the increased algorithmic latency [15], [16].

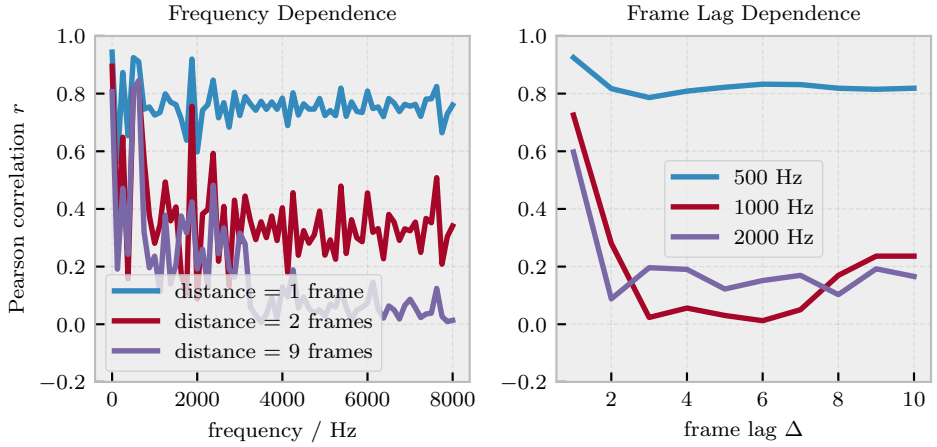


Figure 1.6: Absolute value of Pearson correlation coefficients between consecutive STFT coefficients of continuous PC fan noise (frame length 8 ms, frame shift 2 ms, $\sqrt{\text{Hann}}$ window). Left: as a function of frequency for different frame lags Δ (1, 4, and 9 frames corresponding to 2 ms, 8 ms, and 18 ms). Right: as a function of frame lag Δ for different frequencies (500 Hz, 1000 Hz, and 2000 Hz).

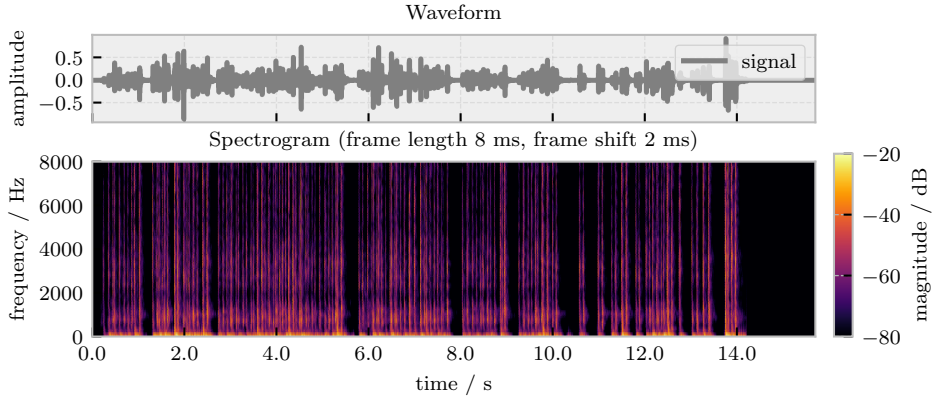


Figure 1.7: Impulsive keyboard typing noise, visualized as a waveform and a spectrogram.

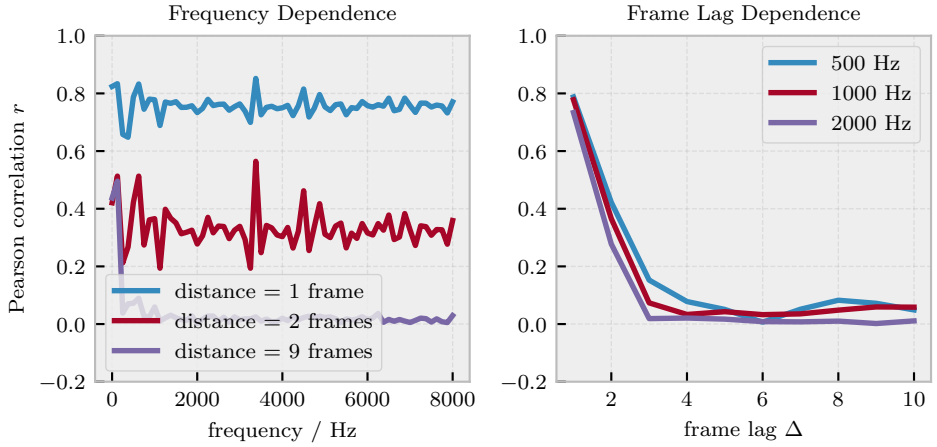


Figure 1.8: Absolute value of Pearson correlation coefficients between consecutive STFT coefficients of impulsive keyboard typing noise (frame length 8 ms, frame shift 2 ms, $\sqrt{\text{Hann}}$ window). Left: as a function of frequency for different frame lags Δ (1, 4, and 9 frames corresponding to 2 ms, 8 ms, and 18 ms). Right: as a function of frame lag Δ for different frequencies (500 Hz, 1000 Hz, and 2000 Hz).

high frequencies. This diffuseness results from the superposition of many sound waves arriving from different directions with random phase relationships.

The spectro-temporal characteristics of noise sources depend on their physical generation mechanisms, which can be broadly categorized into three main classes. First, continuous noise sources, such as PC fans and factory machinery, generate periodic signals with harmonic structures. Compared to speech signals, continuous noise sources typically exhibit more pronounced temporal stationarity (Fig. 1.5), which is also reflected in the temporal correlation (Fig. 1.6), where continuous noise sources exhibit high temporal correlation over a larger time lag Δ at specific frequencies. Second, intermittent noise sources, such as keyboard typing or footsteps, produce brief broadband acoustic events with rapid onset and decay times (Fig. 1.7). Compared to speech signals, intermittent noise sources exhibit a faster decrease in temporal correlation with increasing frame lag Δ (Fig. 1.8). Third, time-varying noise sources, such as music, exhibit more complex temporal patterns. The differences in temporal structure between the target speech and noise signals provide valuable cues that speech enhancement algorithms—particularly multi-frame algorithms—can exploit.

In scenarios with multiple speakers, determining which speaker should be considered as the target speaker and which speakers should be considered as undesired speakers presents an additional problem. This problem can be addressed using heuristics, such as selecting the loudest speaker, or by tracking the auditory attention of the listener [17]–[19]. Since this thesis does not address speaker selection, **we will only consider acoustic scenarios with a single clearly defined target speaker, (non-speech) localized noise sources, and spatially (quasi-)diffuse noise**.

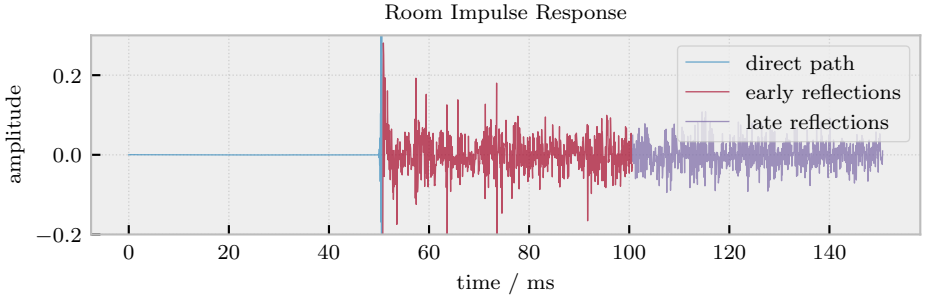


Figure 1.9: Example of a RIR from the REVERB challenge [20] with $T_{60} \approx 0.73\,\text{s}$.

The diffuse noise may include both non-speech noise as well as babble noise, which originates from multiple overlapping speakers at different positions. While diffuse babble noise shares some spectro-temporal characteristics with target speech, it lacks the clear harmonic structure and temporal correlation patterns of an individual speaker. Additionally, it does not exhibit a distinct direction of arrival, further differentiating it from target speech.

1.1.3 *Acoustic Environment*

The acoustic environment significantly influences signal propagation from the sources to the microphones through acoustic reflections (i.e., reverberation) and acoustic shadowing effects caused by walls and objects. When modeled as a linear and time-invariant system, these effects can be interpreted as multiple copies of the source signal arriving at the microphones with different delays and attenuations, characterized by room impulse responses (RIRs). Reverberation presents a particular challenge for speech enhancement because it may decrease the validity of key assumptions made in model-based approaches, such as the independence of consecutive STFT coefficients of speech and noise signals, often leading to performance degradation in reverberant environments.

As shown in Fig. 1.9, a RIR comprises three main components: the direct path, which represents the shortest route from the source to the microphone and includes its propagation delay; early reflections, which arrive within approximately 50 ms to 80 ms after the direct sound and can improve speech intelligibility [21], [22]; and late reverberation, which consists of exponentially decaying, densely overlapping reflections [23]. Reverberation characteristics strongly depend on properties of the acoustic scenario such as room size and surface materials as well as source-microphone geometry. A key metric for quantifying reverberation is the reverberation time T_{60}, defined as the time required for sound energy to decay by 60 dB after the source stops. Although the reverberation time varies with frequency, it is usually expressed as a single (broadband) value. The reverberation time typically ranges from 0.3 s for residential rooms to 1.2 s for class rooms up to over 2 s for larger spaces. The effects

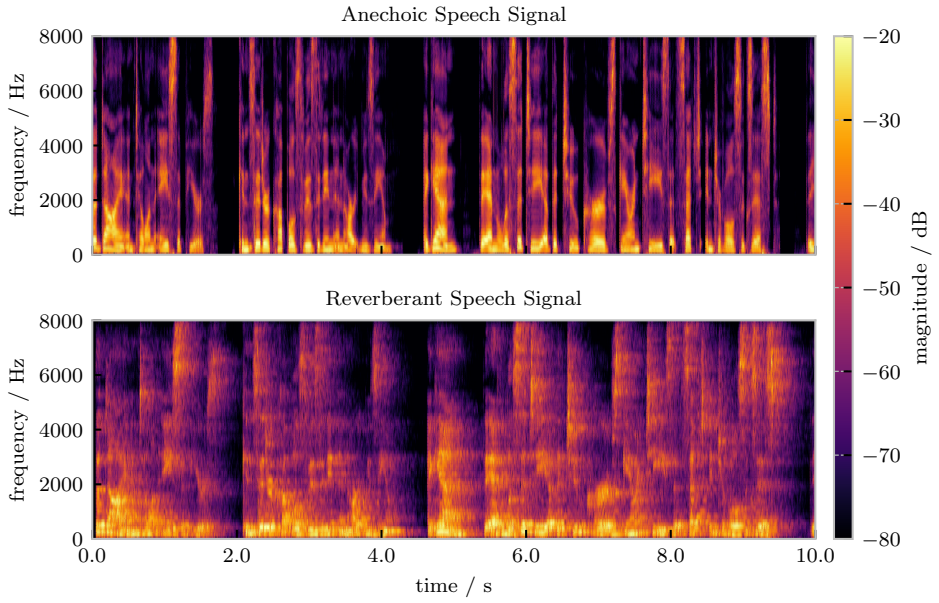


Figure 1.10: Spectrograms of a speech signal, captured in an anechoic environment (top) and in a reverberant environment (bottom) with $T_{60} \approx 0.73\,\mathrm{s}$, corresponding to the RIR in Fig. 1.9.

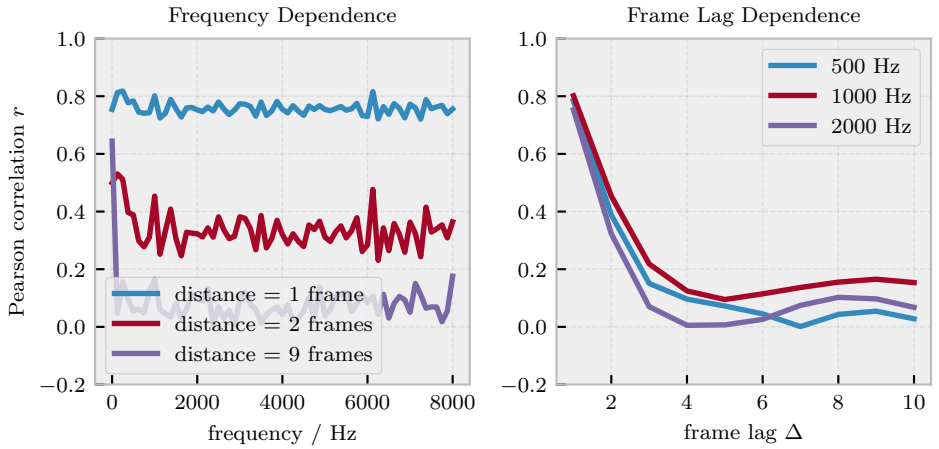


Figure 1.11: Absolute value of Pearson correlation coefficients between consecutive STFT coefficients of reverberant speech signals (frame length 8 ms, frame shift 2 ms, $\sqrt{\text{Hann}}$ window). Left: as a function of frequency for different frame lags Δ (1, 4, and 9 frames corresponding to 2 ms, 8 ms, and 18 ms). Right: as a function of frame lag Δ for different frequencies (500 Hz, 1000 Hz, and 2000 Hz).

of reverberation are demonstrated in Fig. 1.10, which compares the spectrogram of an anechoic speech signal to a simulated reverberant version with $T_{60} \approx 0.73\,\mathrm{s}$. It can be clearly observed that reverberation reduces spectro-temporal sparsity and prolongs energy decay, resulting in temporal smearing. The impact of temporal smearing on the temporal correlation of speech is depicted in Fig. 1.11, which shows Pearson correlation coefficients between consecutive complex-valued STFT coefficients of reverberant speech. Compared to anechoic speech (Fig. 1.4, bottom), reverberation increases temporal correlation, although correlation still decreases with frame lag. Additionally, reverberation decreases spatial coherence, particularly at higher frequencies.

While the RIR characterizes the acoustic environment in the time domain, the acoustic transfer function (ATF) characterizes the acoustic environment in the frequency domain. By applying the multiplicative transfer function approximation in the STFT domain for a specific coherent source, the ATF enables the computation of the STFT coefficients of this source at a given microphone as the product of the source STFT coefficient and the corresponding ATF. Alternatively, relative transfer functions (RTFs) can be used, which relate the ATFs of different microphones to a chosen reference microphone [24].

In summary, reverberation makes it more challenging to separate target speech from noise sources. In this thesis, **we consider a range of reverberation times from 0.2 s to 1.2 s** to investigate the impact of commonly encountered reverberation conditions on speech enhancement performance.

1.1.4 *Binaural Cues*

Although hearing devices often incorporate multiple microphones, it should be realized that the inter-microphone distances on each device are restricted to just a few centimeters. In particular, at frequencies below approximately 2 kHz, wavelengths of sounds may exceed the array dimensions, limiting the spatial resolution and hence the performance of many (purely) spatial filtering algorithms. However, binaural processing can be leveraged by linking left and right hearing devices, with the head between the devices acting as a natural spatial filter. The head shadow effect creates substantial interaural level differences (ILDs) (reaching up to 20 dB at high frequencies) and interaural time differences (ITDs) (reaching up to 0.7 ms) [25], commonly referred to as binaural cues. These binaural cues depend on the source direction and can hence be exploited for localization. For broadband sources, ITDs have been shown to be dominant in human source localization [26]. The ability to localize sound sources enables spatial release from masking, improving the effective signal-to-noise ratio (SNR) for human listeners if the sources are sufficiently spatially separated [27].

1.1.5 *Summary*

In summary, the distinct properties of target speech and noise sources as well as the acoustic environment provide multiple complementary cues that can be exploited by speech enhancement algorithms to distinguish the target speech source from the noise sources. The speech production mechanism introduces characteristic spectro-temporal patterns, including spectro-temporal sparsity, which simplifies the separation of multiple simultaneously active sources in the STFT domain [28]. Additionally, speech exhibits harmonic structures, formant transitions, and amplitude modulations, resulting in characteristic temporal correlation patterns. These spectro-temporal cues can be exploited even in single-microphone configurations. If multiple microphones are available, additional spatial cues can be exploited. In particular, binaural cues such as ILDs and ITDs can aid in spatially separating sources.

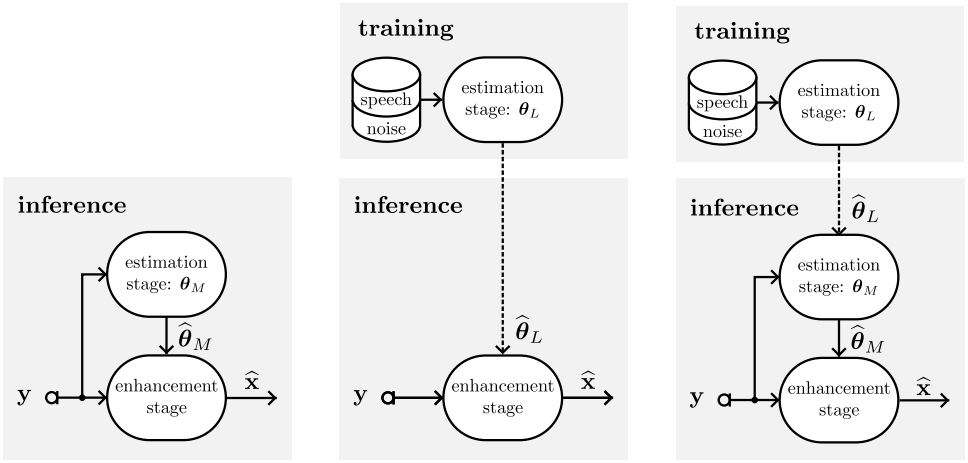
Since the reliability of different cues varies across different acoustic scenarios, speech enhancement algorithms typically exploit multiple cues simultaneously. While some properties can be modeled explicitly (such as the temporal stationarity of certain noise sources [29]) and some cues have well-defined mathematical relationships (such as the link between binaural cues and specific directions of arrival), other separation cues are challenging to represent analytically. This limitation makes learning-based approaches attractive, since they can learn separation cues that are difficult to model explicitly.

1.2 Overview of Speech Enhancement Approaches

The primary objective of single- and multi-microphone speech enhancement algorithms is to extract the target speech from the recorded noisy and reverberant microphone signals, i.e., suppressing noise while not distorting target speech. In binaural hearing device applications, a further objective is to preserve the spatial impression of the acoustic scenario, e.g., by preserving the binaural cues.

This section provides an overview of model-based, learning-based, and hybrid speech enhancement approaches that are particularly relevant to this thesis. For more comprehensive reviews, we refer to [3]–[6], [14], [24], [30], [31] for model-based approaches, to [7]–[9] for learning-based approaches, and to [32] for hybrid approaches.

Speech enhancement algorithms typically consist of two stages (Fig. 1.12), namely an *estimation stage*, where quantities required for enhancement are estimated, and an *enhancement stage*, where the estimated quantities are used to process the noisy microphone signals [32]. The key distinction between model-based and learning-based approaches lies in the estimation stage. In model-based approaches (Fig. 1.12a), the required quantities $\boldsymbol{\theta}_M$ (such as covariance matrices, power spectral densities (PSDs), or a-priori SNRs) are estimated from the single- or multi-microphone noisy reverberant microphone signals $\mathbf{y}$ *during inference*, relying on explicit assumptions about signal characteristics and the acoustic environment. In



(a) model-based (b) learning-based (c) hybrid

Figure 1.12: Definition of model-based, learning-based, and hybrid speech enhancement approaches. $\mathbf{y}$ denotes the single- or multi-microphone noisy microphone signals, $\mathbf{x}$ and $\widehat{\mathbf{x}}$ denote the target and estimated speech component, $\boldsymbol{\theta}_M$ and $\widehat{\boldsymbol{\theta}}_M$ denote the ground truth and estimated model quantities, $\boldsymbol{\theta}_L$ and $\widehat{\boldsymbol{\theta}}_L$ denote the optimal and estimated trainable weights. Adapted from [32].

learning-based approaches (Fig. 1.12b), the required quantities $\boldsymbol{\theta}_L$, i.e., the trainable weights, are estimated by optimizing over (large) speech and noise datasets *during training*. During the enhancement stage, in both model-based and learning-based approaches, the estimated quantities $\widehat{\boldsymbol{\theta}}_M$ or $\widehat{\boldsymbol{\theta}}_L$ are used to process the noisy microphone signals, resulting in an estimate $\widehat{\mathbf{x}}$ of the speech signal. In this thesis, **we define hybrid approaches (Fig. 1.12c) as approaches where a model-based component is combined with a learning-based component, i.e., a deep neural network (DNN), and the required quantities are estimated both *during training and inference***. The quantities $\boldsymbol{\theta}_L$ required by the learning-based component are estimated during training, while the quantities $\boldsymbol{\theta}_M$ required by the model-based component are estimated during inference, aided by the DNN. Crucially, in our definition of hybrid approaches, *the enhancement stage always follows a model-based framework*, ensuring that the enhancement operation is derived based on an optimization problem, which differs from the definition in [32] that also includes approaches with a final enhancement stage that follows a learning-based framework.

Speech enhancement approaches typically operate in either the time domain or a transform domain, including learned transforms and the STFT (with some exceptions that incorporate both time- and transform-domain processing, e.g., [33]). This thesis primarily focuses on speech enhancement approaches in the *STFT domain*, a time-frequency transform that is widely used in both model-based and learning-based approaches. Regarding learning-based approaches, this thesis focuses on *supervised learning*, where DNNs are trained using paired noisy and clean speech data [7].

While alternative approaches based on semi-supervised [34], self-supervised [35], [36], and unsupervised learning [37] have shown promise, their inclusion would extend beyond the scope of this thesis.

In Section 1.2.1, we discuss model-based approaches, which rely on assumptions about the signal characteristics and the acoustic environment. In Section 1.2.2, we discuss learning-based approaches, which utilize DNNs to learn complex relationships between the noisy microphone signals and the target speech signal from large datasets. Finally, in Section 1.2.3, we discuss hybrid approaches, which combine aspects of both model-based and learning-based approaches.

1.2.1 *Model-Based Approaches*

Model-based speech enhancement approaches rely on simplifying assumptions about the speech and noise characteristics and the acoustic environment. For instance, these assumptions often include statistical independence of speech and noise, distributions of the speech and noise STFT coefficients, temporal stationarity of speech and noise components, and spatial coherence of speech and noise components. Additionally, model-based approaches frequently assume that speech and noise STFT coefficients are uncorrelated across both frequency bins and time frames. By leveraging these assumptions, mathematically tractable algorithms can be derived.

In single-microphone configurations, model-based approaches often aim at minimizing the mean square error (MSE) between the target and estimated speech STFT coefficients [38] or their (logarithmic) spectral amplitudes [39], [40]. These approaches commonly assume that speech STFT coefficients are uncorrelated and follow a complex-valued Gaussian distribution. Alternatively, other statistical priors, such as the generalized Gamma distribution, have been proposed for modeling speech STFT coefficients. This distribution provides a more accurate fit to the empirical distribution of speech STFT coefficients than the Gaussian assumption and has been shown to yield a better speech enhancement performance [41], [42].

The assumption that consecutive speech STFT coefficients are uncorrelated holds roughly when using sufficiently long time frames and large frame shift [4], [30]. Under this assumption, independent (real-valued) masks can be applied to each STFT coefficient to suppress the noise component, however, thereby also introducing speech distortion [43]. To address this issue, multi-frame algorithms have been proposed, which leverage the fact that speech and noise STFT coefficients indeed do exhibit temporal correlation, especially if using a small frame shift (Fig. 1.4). By jointly processing multiple consecutive STFT frames, these algorithms can achieve noise reduction while preventing speech distortion [44]–[46]. This makes them particularly useful in single-microphone configurations, where spatial filtering and the potential use of distortionless constraints (such as those used in the minimum variance distortionless response (MVDR) beamformer) are otherwise not available, but they can also be applied in multi-microphone configurations [3], [47]. Similar to the MVDR beamformer for multi-microphone configurations, a frequently used structure for

single-microphone configurations is the multi-frame minimum variance distortionless response (MFMVDR) filter [3], [44], [45], which estimates the target speech STFT coefficients by minimizing the output interference power while preserving the speech temporal correlations.

Model-based algorithms typically rely on estimates of quantities, such as the noise PSD [29], [48], [49], the a-priori SNR [39], [50], or the speech presence probability [51], [52]. For multi-frame algorithms such as the MFMVDR filter, estimates of the highly time-varying speech temporal correlation vector (TCV) and the interference temporal covariance matrix (TCM) are required, for which several model-based estimators have been proposed. For example, the speech TCV has been estimated using a maximum likelihood procedure that assumes a Gaussian interference component [53], or using a low-rank model of the speech TCM. Although it has been shown that the MFMVDR filter yields a good noise reduction with little speech distortion provided that accurate estimates of these quantities are available, its performance is rather sensitive to estimation errors, especially in the speech TCV [54]. However, none of the currently available model-based approaches is able to yield sufficiently accurate estimates mainly due to the highly time-varying nature of these quantities, and hence the potential of multi-frame algorithms has not been fully exploited. This challenge motivates one goal of this thesis, which is to **leverage learning-based approaches to drastically improve the estimation of quantities for multi-frame algorithms**.

In multi-microphone configurations, model-based approaches can further exploit spatial cues (Section 1.1.4), typically outperforming single-microphone approaches [5], [24], [55]–[60]. Popular algorithms include the aforementioned MVDR beamformer, which minimizes the output noise PSD while preserving target speech, and the multi-microphone Wiener filter, which minimizes the MSE between the target speech STFT coefficients and the output STFT coefficients. These algorithms require estimates of the noisy, speech, and/or noise spatial covariance matrices, as well as estimates of the ATFs or RTFs.

Provided that accurate estimates are available, model-based algorithms exhibit highly desirable properties. Their performance can be theoretically shown to be optimal and analyzed for predefined model mismatches using metrics such as input and output SNRs, offering a degree of interpretability. Additionally, since these algorithms do not rely on training data, they are inherently robust to train–inference mismatch. However, performance can degrade rapidly if their underlying assumptions are violated, such as if in the presence of non-stationary or non-Gaussian noise, if STFT coefficients are not uncorrelated, or if spatial coherence assumptions do not hold.

1.2.2 *Learning-Based Approaches*

Learning-based approaches, particularly those leveraging deep learning, have emerged as powerful alternatives to model-based approaches, generally yielding a

much better speech enhancement performance [8]. Instead of relying on simplifying assumptions about speech and noise characteristics and the acoustic environment, these approaches learn complex nonlinear relationships between the noisy microphone signals and the target speech signal directly from data, typically using DNNs, which are often described as "universal function approximators" [61]. Although DNNs can also be viewed as models, this thesis primarily uses the term "model" to refer to model-based approaches.

The design of an effective learning-based speech enhancement algorithm involves selecting and configuring several key components, including training data, input features, DNN architectures, DNN outputs, and loss functions. Unlike model-based approaches, where theoretical optimality often guides algorithm design, learning-based approaches heavily rely on empirical evaluation. As a result, much of the progress in this field originates from combining and optimizing different components, making a clear understanding of available design choices essential. In this section, we provide an overview of these key components and their relevance to speech enhancement.

1.2.2.1 *Training Data*

In model-based algorithms, speech enhancement performance is affected by the mismatch between the assumptions and the signals observed during inference. In contrast, in learning-based algorithms, performance is largely affected by the mismatch between the training data and the signals observed during inference [62], [63]. Therefore, to obtain robust learning-based speech enhancement algorithms, it is crucial to use training datasets that cover a wide range of acoustic scenarios, including diverse languages, speakers, noise types, acoustic environments, and recording setups. The choice of speech material is particularly important, as it has been shown to have a greater impact on robustness than noise material or recording setups, especially in single-microphone configurations [64], [65]. Extensive efforts in data collection, along with initiatives such as the deep noise suppression (DNS) challenge series [66]–[68], the Clarity Enhancement Challenge (CEC) series [69], and the CHiME challenge series [70] have provided the research community with large-scale speech, noise, and (binaural) RIR datasets for training and evaluation. Additional training data can be simulated by assuming that microphones operate linearly, such that speech and noise signals can be additively mixed. The effects of the acoustic environment (Section 1.1.3) and binaural cues (Section 1.1.4) can be simulated using datasets of measured RIR such as the ones in [71]–[73] or using open-source tools such as the `RIR-Generator`[2], `pyroomacoustics` [74], and `gpuRIR` [75] (which are mainly based on the image source model), as well as more sophisticated tools that incorporate human perception [76] or DNNs [77]. Although simulated signals do not capture all real-world variabilities, they allow to include a large amount of variation during training [78].

2 Available at `https://github.com/ehabets/RIR-Generator`.

The use of diverse training datasets such as those included in recent challenges [66]–[69], [79] allows DNNs to learn far more complex relationships between the noisy microphone signals and the target speech signal than would be feasible to design manually in model-based approaches, often leading to substantial performance improvements over model-based algorithms [8]. Moreover, the use of standardized datasets facilitates reproducibility and comparability of results.

1.2.2.2 *Input Features*

The selection of input features determines a trade-off between information content and computational complexity. While the time-domain waveforms obviously contain all information about the microphone signals, they represent this information in a highly redundant and inefficient way. Good features for speech enhancement should retain relevant speech and noise characteristics, be compact, and result in a robust speech enhancement performance.

Commonly used features include STFT-domain magnitude spectra [80], log-magnitude and log-power spectra [81], and compressed magnitude spectra [82], which provide a more structured representation compared to the time-domain waveform. Since phase information is important for speech enhancement [31], magnitude-based features can be complemented with phase-based features [83], though phase wrapping should be taken into account. Alternatively, the real and imaginary parts of the STFT coefficients can be used directly as feature [84].

Motivated by the human auditory system, also auditory-motivated features have been proposed. For example, features such as amplitude modulation spectrograms, relative spectral transform perceptual linear prediction coefficients, Gammatone and Mel-frequency cepstral coefficients, as well as fundamental frequency-based features have been proposed, which aim to capture speech characteristics relevant to human perception [85]. However, since these features often involve a high computational complexity, computationally efficient features have been proposed to mimic auditory processing. For example, the reduced sensitivity of human hearing at higher frequencies motivates the use of logarithmic frequency spacing, which decreases input feature dimensionality compared to linear spacing [86], [87].

Particularly over the last years, an increasing fraction of speech enhancement algorithms rely on learned features, often using a trainable one-dimensional convolutional layer to replace the STFT analysis operation, typically without coupling the analysis and synthesis operations [9], [88], [89]. Interestingly, [89] found that the learned analysis transformation, which can be interpreted as a set of finite impulse response filters, is primarily sensitive to low frequencies, similar to human auditory perception. While such learned features can achieve high speech enhancement performance even for short frame lengths such as 1 ms, they tend to lack robustness in reverberant environments [90], [91]. To improve robustness, various approaches have been proposed, such as imposing an analytic constraint based on the Hilbert transform to couple the analysis and synthesis transformations.

In multi-microphone configurations, single-microphone features from different microphones are either simply concatenated or spatial features such as inter-microphone phase or level differences are considered. These features enable DNNs to exploit spatial cues (Section 1.1.5), although the extent to which spatial cues—rather than spectro-temporal cues—are actually used depends heavily on the DNN architecture [92], [93]. Constructing spatial features by concatenating single-microphone features from multiple microphones or using inter-microphone phase/level differences inherently creates a dependence on specific microphone array geometries. Consequently, any change in the microphone array configuration necessitates modifying the algorithm, e.g., by retraining the DNN. To overcome this issue, the DNN architecture needs to be adapted (see also the discussion in Section 1.2.2.3).

Summarizing, a wide range of features has been investigated for single- and multi-microphone speech enhancement. State-of-the-art algorithms typically use simple features, such as the real and imaginary parts of STFT coefficients, their magnitude and phase, or learned features from a trainable convolutional layer [9]. This thesis favors compressed magnitude and phase representations of STFT coefficients.

1.2.2.3 *DNN Architectures*

Commonly used DNN architectures for speech enhancement employ network components consisting of fully connected networks, recurrent neural networks (RNNs), convolutional neural networks (CNNs), attention-based networks, and combinations thereof. Early architectures often relied on fully connected network components, requiring the concatenation of multiple time frames to provide temporal context [81], which is crucial given the strong temporal correlations observed in speech and noise signals (Section 1.1). Alternatively, RNN-based architectures, such as long short-term memory (LSTM) networks [15] and gated recurrent units (GRUs) [94], have been employed for modeling temporal context. RNN-based architectures have been particularly popular in real-time applications, due to their ability to capture temporal context while maintaining relatively low computational complexity [15], [16], [86], [94]–[97]. CNNs have been particularly attractive due to their ability to efficiently extract multi-microphone features. To incorporate sufficient temporal context at a reasonable computational complexity, so-called temporal convolutional networks (TCNs) employ dilated convolutions [98] (discussed in detail in Section 3.2.2.3). However, fully convolutional architectures are typically associated with higher memory consumption and larger computational complexity than RNNs. Convolutional recurrent neural network (CRNN) can leverage both the feature extraction capability of CNNs and the temporal representation capacity of RNNs and were found to result in a good and scalable trade-off between speech enhancement performance and computational complexity [33], [80], [84], [99]–[101]. More recently, attention-based architectures have demonstrated remarkable performance, usually at the cost of a high number of parameters and significant computational complexity, which grows quadratically with the sequence length [102], [103]. Although efforts have been made to reduce this computational complexity through appropriate hyperparameter choices [104] or by utilizing attention mechanisms that are less computationally

complex than full self-attention [105], the associated computational complexity still makes the implementation of attention-based architectures difficult, especially in applications with low-latency constraints.

Beyond these specific network components, many DNN architectures employ an encoder-decoder structure. In this structure, the encoder transforms the input features into a latent representation, which corresponds to a compressed representation of the input often termed "bottleneck". The decoder then reconstructs the enhanced speech signal from this latent representation. A widely used example is the U-Net structure, which incorporates skip connections between the encoder and decoder to preserve fine-grained details that might otherwise be lost during the encoding process [106]–[109]. The U-Net structure can utilize different network components, such as CNNs [107], RNNs [106], or attention mechanisms [108].

In addition to general advancements in DNN architectures (which often originate from the computer vision field), there have also been speech enhancement-specific architectural advancements. Note that these advancements—although inspired by model-based methods—are still considered purely learning-based and do not count as hybrid approaches in the context of this thesis, as they do not involve the estimation of quantities during both training and inference (Fig. 1.12c). For instance, architectures motivated by beamforming algorithms have gained popularity. These architectures perform independent subband processing of multiple microphones in the STFT domain, which is motivated by the fact that spatial cues, such as inter-microphone phase differences, are relatively temporally stable for stationary sources while varying over frequency (due to the linear phase property of the STFT) [103], [110], [111]. To still allow for fullband spectral context, these architectures typically separate processing into subband processing blocks and fullband processing blocks. Moreover, algorithms that explicitly process separate physical axes—frequency, time, and space—have recently been particularly successful [112].

Additionally, complex-valued DNN architectures have been investigated for speech enhancement [99], [109], [113], [114]. In real-valued architectures, complex-valued STFT coefficients are typically separated into their real and imaginary (or magnitude and phase) components, which are then processed independently. In contrast, complex-valued DNN architectures perform complex-valued operations, such as complex-valued additions and multiplications, hence maintaining the coupling of magnitude and phase. These architectures require the choice of appropriate complex-valued activation functions, which do not directly follow from their real-valued counterparts, thus often requiring additional design choices. However, despite their increased computational complexity, no consistent significant performance improvements over real-valued architectures have been found [113], [114].

A key task in designing learning-based multi-microphone speech enhancement approaches is how to use spatial features. As mentioned before, if features from multiple microphones are simply concatenated at the input, the resulting DNN usually becomes dependent on a specific microphone array geometry, requiring retraining if the geometry changes. This lack of generalization is particularly problematic in applications involving ad-hoc microphone arrays or possible microphone failures.

To address this issue, various approaches have been proposed to achieve microphone permutation and number invariance [112], [115], [116]. For instance, the transform-average-concatenate (TAC) method [115] takes as input a set of feature streams (one stream per microphone), shares information across the streams in a non-linearly transformed space, and outputs a set of modified feature streams. The modified output feature streams contain stream-specific information as well as shared information influenced by all streams in a permutation-invariant fashion. This is achieved through a combination of weight sharing across the transformations for each stream and the application of the averaging operation, which is inherently permutation-invariant. To replace the fixed averaging operation used in the TAC method, channel attention-based methods have been proposed (which are also permutation-invariant) [112], [116]. Using the TAC method or channel attention-based methods enables the training with different microphone arrays as well as the use of arrays during inference that were not seen during training, while maintaining performance in conditions with matched microphone array geometry [116].

Summarizing, many DNN architectures have been investigated for speech enhancement, each offering distinct trade-offs between speech enhancement performance and computational complexity. For real-time applications that require low computational complexity, CNNs, RNNs, and combinations thereof currently tend to result in the best trade-off [100] and will be favored in this thesis, whereas attention-based architectures result in the best speech enhancement performance if computational complexity is unconstrained. In addition, separating the processing along distinct physical axes seems to be a good choice when designing DNN architectures. If the specific application involves multiple microphones, the use of microphone geometry-agnostic DNN architectures may be considered. However, drawing definitive conclusions about DNN architecture choices is challenging, if not impossible, due to the complex interaction between DNN architecture, hyperparameters, dataset, enhancement approach, and other factors.

1.2.2.4 *DNN Outputs*

Learning-based approaches can be categorized based on the DNN output into two main classes, namely those that output an estimate of the target speech component and those that output filter coefficients. In the first class, the DNN directly outputs the enhanced waveform or spectrum, without explicitly processing the noisy microphone signal or its representation [15], [84], [117]–[123]. Consequently, these approaches lack a direct signal path between the input and output signals. For example, the denoising WaveNet algorithm [117] performs denoising in the time domain, estimating multiple samples simultaneously, while complex spectral mapping algorithms predict target speech STFT coefficients [15], [84], [118]–[121], [123].

In the second class, the DNN outputs mask or filter coefficients, which are used to modify the noisy microphone signal or its representation, thereby creating a direct signal path between the input and output signals that bypasses the DNN for the actual signal modification [92], [93], [101], [115], [124]–[132]. In single-microphone con-

figurations, this includes algorithms that apply binary masks [124], ratio masks [125], phase-sensitive masks [126], complex ratio masks [92], [127], [132], and extends to spectro-temporal filters [129]. The popular Conv-TasNet algorithm [89] also falls into this category, as it employs an encoder to transform the input signal into a latent representation, applies a real-valued mask estimated by a separator network to this representation, and constructs the estimated speech signal using a decoder. In multi-microphone configurations, algorithms have been proposed that apply filters to the noisy signals in the time-domain [128], [133], in a latent domain [130], or in the STFT domain [92], [101], [131], [132], [134]–[136].

We want to emphasize the clear distinction between the internal representation of spectral, temporal, and/or spatial context within the DNN (such as temporal context in RNN, Section 1.2.2.3) and the use of such context in the filtering operation. For instance, spatial information can be represented at the feature level without necessarily being exploited in the filtering operation. Whereas spatial filtering approaches can exploit spatial information both within the DNN and in the filtering operation, masking approaches are limited to internally representing spatial information [92]. Comparing the frequency-time joint nonlinear filter (FT-JNF) (single-microphone masking) algorithm with the complex-valued spatial autoencoder (COSPA) (spatial filtering) algorithm (both using multi-microphone input featurs), [93] showed that while both algorithms *can* represent spatial information internally, whether the FT-JNF algorithm does so depends on the specific training target signal and the acoustic scenario. For example, when trained with the reverberant speech component at a reference microphone as the target signal, spatial filtering is not explicitly required—in simple (noise-free) scenarios, the task of the DNN reduces to merely passing the signal of the reference microphone. In contrast, when trained with the output of an oracle time-varying beamformer that is adaptively steered towards the currently active source as the target signal, the DNN is implicitly trained to perform spatial filtering by emulating this beamformer. The COSPA algorithm consistently represented spatial information internally, irrespective of the specific training target signal and the acoustic scenario. This distinction highlights that not all learning-based algorithms inherently utilize the separation cues they are provided with, depending on the specific DNN architecture, filtering operation, training target signal, and acoustic scenario.

Comparing the two classes of learning-based approaches defined above, in the first class, the DNN can be seen as itself acting as the filter, giving it full control over the output signal, thereby circumventing any upper bounds on speech enhancement performance that may limit the second class of approaches. In particular, complex spectral mapping approaches often achieve higher performance than masking approaches [120]. However, this typically comes at the cost of requiring more powerful DNN architectures and increasing the risk of robustness issues. These challenges have made the combination of (learning-based) complex spectral mapping into model-based approaches particularly attractive, resulting in hybrid approaches (Section 1.2.3.1).

1.2.2.5 *Loss Functions*

Loss functions quantify the discrepancy between the outputs of the DNN and the target outputs during training, and they provide the gradients necessary for updating the trainable weights of the DNN through backpropagation. An effective loss function should be differentiable to ensure compatibility with backpropagation, aligned with the objective of robust speech enhancement, stable to mitigate the risk of exploding or vanishing gradients, and computationally efficient to prevent excessive resource consumption. In purely learning-based speech enhancement, loss functions can generally be categorized into filter approximation loss functions and signal approximation loss functions.[3]

Filter approximation loss functions are defined directly on the estimated filter, requiring the definition of target filters. These target filters depend on design choices that can significantly impact speech enhancement performance [137]. Common examples include the binary cross-entropy loss for ideal binary masks and the MSE loss for ideal ratio masks [137] or delay-and-sum beamformer coefficients [134].

Signal approximation loss functions are defined on the estimated speech signal, circumventing the definition of target filters. Variants of MSE are widely used, including spectral magnitude MSE, phase-sensitive spectral magnitude MSE [138], and complex spectral MSE [139] in the STFT domain, as well as time-domain MSE [138]. Variants of the signal-to-distortion ratio (SDR) are also commonly employed, both in the STFT domain [140] and the time domain [138], including the popular scale-invariant signal-to-distortion ratio (SI-SDR) [141].[4] Unlike the MSE, SDR-based loss functions operate on an energy ratio, weighting errors relative to the energy of the target signal. In particular, distortions in lower-energy regions are weighted more strongly than those in higher-energy regions, aligning more closely with human perception. The SI-SDR, in particular, has been widely adopted, especially for single-microphone speech enhancement configurations in anechoic environments and where sensitivity to signal scaling is not a critical concern (see also the discussion on SI-SDR in Section 2.2.1). The SI-SDR is closely related to the time-domain MSE but incorporates a scaling operation on the estimated speech component along with logarithmic compression, in addition to operating on an energy ratio, which aligns it with human loudness perception to some extent [90]. However, as a waveform-matching loss function, SI-SDR is highly sensitive to time shifts as small as a single sample, whereas human auditory perception is robust to such minor shifts [138], [140]. For algorithms capable of modifying signal phase, this strong emphasis on precise phase alignment during training may artificially limit the achievable speech enhancement performance.

Variations of the ℓ_1 loss have also been investigated in both the STFT domain [120] and combined time and STFT domains [92], [108]. One motivation for favoring the

3 Masks are viewed as a special case of filters with a single coefficient.

4 The SDR is often denoted as SNR in the literature, where "noise" refers to distortion rather than ambient noise.

ℓ_1 loss over the more common MSE in regression tasks comes from a maximum-likelihood perspective: if we model the real and imaginary parts of speech STFT coefficients with a certain probability distribution, the negative log-likelihood of that distribution emerges as a natural choice for the loss function to be minimized [142]. In many model-based approaches, the real and imaginary parts of speech STFT coefficients are assumed to follow heavy-tailed distributions, such as the Laplacian distribution [143], whose negative log-likelihood corresponds to the absolute difference used in the ℓ_1 loss. In contrast, a Gaussian assumption corresponds to the squared difference used in the MSE loss. Beyond this probabilistic argument, the ℓ_1 loss also promotes spectro-temporal sparsity in the output signal, aligning with the well-known spectro-temporal sparsity property of speech signals in the STFT domain (Section 1.1.1). Empirical results support these theoretical benefits, demonstrating that ℓ_1-based loss functions can outperform MSE-based losses in speech enhancement [144].

While the previously discussed loss functions prioritize simplicity and computational complexity, more sophisticated approaches incorporate perceptually motivated metrics, such as short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ). To enable differentiability for gradient-based optimization, these metrics have been approximated using differentiable operations [145], integrated into generative adversarial network-based frameworks [146], or replaced with surrogate DNNs trained to mimic the behavior of the metric [147]. Alternatively, reinforcement learning has been employed to optimize these perceptual metrics directly, bypassing the need for gradient-based optimization [148]. In addition, it has been proposed to predict mean opinion scores directly using a DNN [149]–[151], such that the resulting DNN—with its trainable weights frozen—can be used as a loss function. Despite their perceptually motivated design, these loss functions often provide only marginal improvements in speech enhancement performance compared to simpler loss functions [145]–[148]. Moreover, if substantial improvements do occur, they are typically limited to the specific metric used during training, limiting generalizability [151], [152]. For instance, [152] demonstrated that training with a PESQ-based loss function resulted in high PESQ scores but low scores on other metrics, introducing distortions on the output signal that PESQ fails to capture. Similarly, [151] found that training with a Deep Noise Suppression Mean Opinion Score (DNSMOS)-based loss function improved DNSMOS scores but degraded performance on other metrics under evaluation conditions not represented in the training data of DNSMOS. These findings highlight robustness issues associated with many sophisticated perceptually inspired loss functions [144]. Taking these findings into account, in this thesis, we focus on simpler, potentially more robust loss functions, such as the SI-SDR (for anechoic environments) or the ℓ_1 loss in the STFT domain (for reverberant environments).

A recent development in loss function design leverages pre-trained speech foundation models, such as wav2vec [153] and WavLM [154], to derive perceptually motivated loss functions. Speech foundation models are trained on large-scale unlabeled datasets using self-supervised, task-agnostic objectives (often referred to as "pretext tasks") designed to extract generalizable speech representations. Distances between

these representations for the reference signal and the degraded signal have been shown to correlate strongly with objective metrics (e.g., PESQ and STOI) as well as human intelligibility, and have lead to their use in both single-microphone [155], [156] and binaural hearing device [157] applications. However, similarly as for the perceptually motivated loss functions, the use of speech foundation models for loss functions has so far not resulted in significantly improved speech enhancement performance compared with simpler alternatives, such as the SI-SDR loss [155]–[157].

To ensure that signal approximation loss functions accurately reflect the quality of the *final* estimated speech signal, they should be computed on the reconstructed waveform rather than directly on the estimated STFT coefficients. For STFT-based algorithms, defining the loss function directly on the estimated STFT coefficients may seem computationally efficient, as it avoids explicit time-domain synthesis and subsequent re-analysis. However, this approach neglects the impact of overlap-add (OLA) processing, which combines overlapping frames and significantly affects the final output signal. Synthesizing the time-domain signal before re-analyzing it for loss computation not only accounts for OLA processing but also allows for the choice of different STFT configurations for enhancement and loss function evaluation. This even enables the combination of STFTs with different settings, e.g., combining high temporal and spectral resolution [158].

Overall, signal approximation loss functions have been shown to outperform filter approximation loss functions [159] and are generally preferred in current state-of-the-art speech enhancement algorithms [144], potentially due to their direct alignment with the primary objective of speech enhancement: obtaining an accurate estimate of the speech signal (and not a filter estimate). Notably, in certain cases, signal and filter approximation loss functions have a direct correspondence. For instance, [160] demonstrated that an MSE loss applied to the estimated speech signal in the STFT domain is equivalent to an MSE loss on the estimated (real-valued) mask, weighted by the squared magnitude of the noisy microphone signal, thereby giving more weight to STFT coefficients with high energy.

It is important to note that the classification of DNN outputs into speech estimates and filter estimates in the previous section does not directly correspond to the classification of loss functions into signal approximation and filter approximation. For instance, a DNN that outputs filter coefficients can still employ a signal approximation loss function by synthesizing the enhanced signal and computing the loss based on the estimated speech [129].

1.2.3 *Hybrid Approaches*

Model-based speech enhancement approaches offer interpretability and—given accurate estimates of the required quantities—theoretical guarantees. However, their reliance on simplifying assumptions can limit performance in real-world acoustic scenarios. Learning-based speech enhancement approaches, particularly those using DNNs, can learn complex relationships from data and achieve high speech en-

hancement performance, but they typically suffer from train-inference mismatch and lack the interpretability and theoretical guarantees of model-based approaches. This thesis focuses on hybrid approaches, which aim to combine the strengths of both model-based and learning-based approaches. As mentioned before, we define hybrid approaches as those that use a model-based enhancement stage, and where the estimation of quantities is performed both during training (for the DNN) and during inference (for the model-based enhancement stage) (Fig. 1.12c). This definition excludes approaches where the DNN architecture is motivated by model-based approaches [103], [110]–[112] (Section 1.2.2.3), since—from our perspective—such combinations do not fully leverage the interpretability offered by incorporating a model-based enhancement stage.

Hybrid speech enhancement approaches, as defined above, can be categorized based on the interaction between the DNN and the model-based enhancement stage during training. A fundamental distinction is made between *decoupled* and *coupled* approaches, as illustrated in Fig. 1.13. In *decoupled approaches*, the DNN operates independently of the model-based enhancement stage during training. There is no gradient flow from the final estimated speech signal back into the DNN (denoted in blue in Fig. 1.13). Consequently, the DNN is trained without considering its impact on the final speech estimate. However, decoupled approaches offer a high degree of modularity, allowing estimated quantities such as the noise PSD or a-priori SNR to be used with a variety of model-based algorithms, including minimum mean square error (MMSE) estimators [39], [40], [42] and MFMVDR filters [161], without requiring retraining. In contrast, *coupled approaches* integrate the model-based enhancement stage as a differentiable operation within the DNN's computational graph. This integration ensures that all operations, including those of the model-based enhancement stage, are accounted for during training. The categorization presented in this section reflects an increasing influence of the DNN on the final estimated speech signal, moving from decoupled approaches with a greater reliance on the model-based enhancement stage to coupled approaches where the DNN plays a more dominant role.

1.2.3.1 *Class I: Decoupled Approaches*

Decoupled approaches are further divided into two subcategories, based on what the DNN is trained to estimate: *component-directed* approaches and *quantity-directed* approaches (see Fig. 1.13).

CLASS I–A: DECOUPLED, COMPONENT-DIRECTED APPROACHES In component-directed approaches, the DNN is trained to estimate preliminary speech or noise components ($\widehat{\mathbf{x}}^{\text{prel}}$ and $\widehat{\mathbf{n}}^{\text{prel}}$, respectively). These components are either estimated indirectly via time-frequency masks [162]–[164] or directly [78], [103], [120], [165]. These preliminary component estimates are then used to compute the quantities required by the model-based enhancement stage ($\widehat{\boldsymbol{\theta}}_M$).

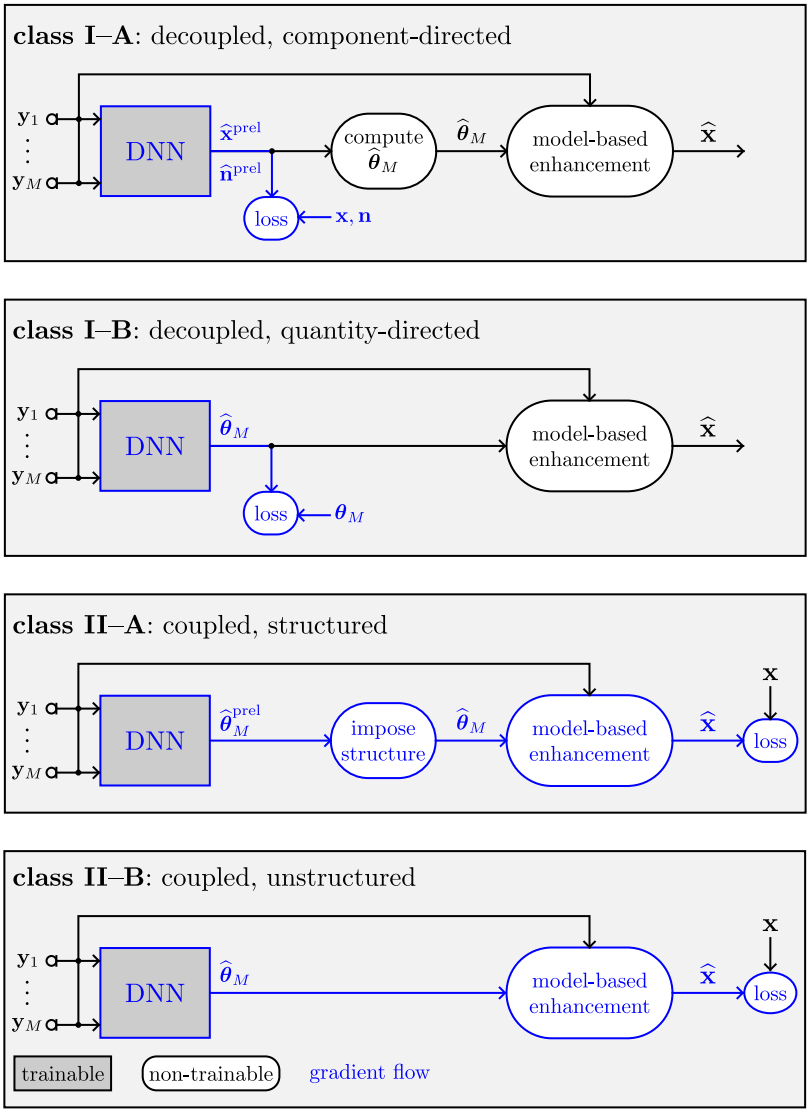


Figure 1.13: Categorization of hybrid speech enhancement approaches. $\mathbf{y}_m$ denotes the m-th noisy microphone signal, $\widehat{\mathbf{x}}^{\text{prel}}$ and $\widehat{\mathbf{n}}^{\text{prel}}$ denote preliminary estimates of the speech and noise components, $\mathbf{x}$ and $\mathbf{n}$ denote the target speech and noise components (obtained directly or by applying a mask/filter), $\boldsymbol{\theta}_M$ and $\widehat{\boldsymbol{\theta}}_M$ denote the ground truth and estimated quantities required by the model-based enhancement stage, $\widehat{\mathbf{x}}$ denotes the final estimated speech signal, and $\widehat{\boldsymbol{\theta}}_M^{\text{prel}}$ denotes a preliminary estimate of the required quantities. From top to bottom, the impact of the DNN on the final estimated speech signal tends to increase.

A prominent group of algorithms that estimate preliminary components indirectly via time-frequency masks is given by mask-based beamformers introduced in [162]. Specifically, [162] trained a bidirectional LSTM network to estimate ideal binary masks, which were then used to compute the covariance matrices required by a generalized eigenvalue beamformer. Mask-based beamformers have since then seen many improvements. For example, [163] investigated different strategies for using masks to derive the quantities required by an MVDR beamformer, including using an independent mask per channel, a single mask for all channels, and additionally applying a mask as a post-filter. In [164], a CNN-based mask estimator was combined with a convolutional beamformer, aiming at simultaneous denoising, dereverberation, and source separation.

Algorithms that estimate preliminary components directly have has also been explored. The Beam-TasNet algorithm [165] applies variations of Conv-TasNet [89] to estimate the target speech component at each microphone. These estimates are then used to compute an MVDR beamformer. Another successful group of algorithms, known as "sequential beamformers," comprises several model-based and learning-based enhancement stages that each estimate the speech component at a reference microphone [78], [103], [120]. Typically, the first stage is a DNN that estimates the speech component. Based on this estimate, the second stage estimates the quantities required by a model-based filter that is applied to the noisy microphone signals, with the goal of resulting in an estimate of the speech signal with lower speech distortion compared to the estimate of the first stage. In the third stage, the noisy microphone signals and the estimated speech signals by the first stage and the second stage are then used by another DNN to output a further improved estimated speech signal. This procedure can be iterated, potentially further improving speech enhancement performance. For example, [120] used a multi-microphone complex spectral mapping algorithm in the first stage, where a temporal convolutional network (TCN) with a Dense U-Net structure estimates the speech component at a reference microphone. The second stage was a time-varying MVDR beamformer, and the third stage was another TCN with a Dense U-Net structure. A similar approach was proposed in [103], but the (time-varying) MVDR beamformer was replaced with a time-invariant spatio-temporal Wiener filter, and the TCN architecture was replaced with the TF-GridNet architecture. Despite being trained separately, these algorithms [78], [103], [120], [165] achieved impressive speech enhancement performance. However, the contribution of the model-based stage is not always clear. In [103], the model-based stage provided only modest improvements, and in [120], a direct comparison with a purely learning-based multi-microphone baseline algorithm was missing.

CLASS I–B: DECOUPLED, QUANTITY-DIRECTED APPROACHES In contrast to component-directed approaches, quantity-directed approaches bypass the intermediate step of estimating separated components or masks. Instead, the DNN is trained to directly estimate the quantities required by the subsequent model-based enhancement stage ($\widehat{\boldsymbol{\theta}}_M$). The key motivation is to avoid potential estimation

errors introduced by the heuristic computation of these quantities from preliminary component estimates.

For example, DNNs were employed to estimate the noise PSD, outperforming model-based estimators due to the ability of DNNs to adapt to quick noise level changes [166]. Similarly, DNNs were used to estimate quantities such as the a-priori SNR [167]–[169] and speech presence probability (SPP) [161], [170], [171]. For instance, [167] used a residual LSTM network to estimate the a-priori SNR, which was then used by the (model-based) MMSE short-time spectral amplitude estimator derived under a Gaussian assumption [39] on the real and imaginary parts of the STFT coefficients. The DeepMMSE algorithm [166] uses a TCN to estimate the noise PSD, which is then used by the (model-based) MMSE estimator derived under a Gamma assumption [42]. In another example [161], a bidirectional LSTM was used to estimate the SPP, which was then used by the (model-based) MFMVDR filter. These algorithms outperformed not only model-based estimators [161], [166], [167], but also purely learning-based speech enhancement algorithms [166], [167].

1.2.3.2 *Class II: Coupled Approaches*

In contrast to decoupled approaches, *coupled approaches* integrate the model-based enhancement stage as a differentiable operation within the computational graph. This enables end-to-end training, meaning the DNN is trained with direct awareness of how its estimates influence the final enhanced speech signal. Coupled approaches are further categorized into *structured estimation* and *unstructured estimation* approaches.

CLASS II–A: COUPLED, STRUCTURED ESTIMATION APPROACHES Structured estimation approaches explicitly impose structure derived from the underlying model on the preliminary estimated quantities ($\widehat{\boldsymbol{\theta}}_M^{\text{prel}}$), such as enforcing that estimated covariance matrices are Hermitian positive-definite through specific linear algebra operations. This ensures that the final estimated quantities ($\widehat{\boldsymbol{\theta}}_M$) are mathematically compliant within the model-based enhancement stage. As an illustrative example of how structure can be imposed on estimated quantities is given by an extension of mask-based beamformers. In conventional mask-based beamformers—categorized as decoupled approaches (class I–A)—time-frequency masks are estimated, which are then used to derive spatial covariance matrices (SCMs) by temporally aggregating instantaneous SCM estimates, often using heuristic procedures like recursive smoothing. These procedures implicitly impose structure on the resulting estimated quantities. Instead of relying on heuristic procedures, [172] proposed an attention mechanism to learn the temporal aggregation of instantaneous SCMs, outperforming conventional mask-based beamformers that rely on heuristic temporal aggregation especially in acoustic scenarios involving moving sources. This attention mechanism linearly aggregates instantaneous SCM estimates (which exhibit a rank-1 structure) to compute time-varying speech and noise SCMs that are generally Hermitian positive-definite and full-rank (assuming a sufficient number of

time frames are considered during aggregation). However, integrating the attention mechanism proposed in [172] created a dependence on the specific microphone array configuration used during training, losing the flexibility to operate with arbitrary microphone array configurations, which is one of the key benefits of conventional mask-based beamformers. Hence, another goal of this thesis is to **develop an algorithm that combines the benefits of learned temporal aggregation for tracking moving sources with the robustness against varying microphone array configurations of conventional mask-based beamformers**.

CLASS II–B: COUPLED, UNSTRUCTURED ESTIMATION APPROACHES Unstructured estimation approaches do not explicitly impose structure derived from the underlying model on the estimated quantities ($\widehat{\boldsymbol{\theta}}_M$). Instead, the DNN is free to estimate these quantities without restrictions like positive-definiteness. For example, [173] extended the algorithm in [172] by exploring unstructured, non-linear instantaneous SCM aggregation. Results showed similar performance between linear (structured) and non-linear (unstructured) SCM aggregation for stationary sources, but a slight advantage for the structured approach in dynamic scenarios. Another example is the ADL-MVDR algorithm [174], which replaces explicit matrix inversion and eigenvector computation with a learned nonlinear transformation. The estimated SCMs are not constrained to be positive-definite, which is inconsistent with the theoretical derivations underlying the MVDR beamformer. Follow-up studies to the ADL-MVDR algorithm, such as [135] in the STFT domain and [133] in the time domain, replaced the MVDR beamformer itself with a learnable layer. This represents a shift towards purely learning-based approaches (Section 1.2.2), where DNNs directly estimate filter coefficients rather than quantities for a model-based enhancement stage.

1.2.3.3 *Loss Functions*

The choice of loss function in hybrid speech enhancement algorithms depends on whether the algorithm follows a decoupled or a coupled approach. In decoupled approaches, the loss function is defined on intermediate quantities that precede the model-based enhancement stage, rather than on the final estimated speech signal. In contrast, coupled approaches allow for the use of signal approximation loss functions defined on the final estimated speech signal, enabling end-to-end training [135], [172]–[175]. This allows the DNN to be trained without explicitly defining target values for the intermediate quantities. While signal approximation loss functions can improve overall speech enhancement performance compared to loss functions defined on intermediate quantities, they do not necessarily ensure that the DNN outputs estimates that accurately reflect the true underlying quantities, even if mathematical structure is imposed. For instance, in the ADL-MVDR algorithm, one might expect the DNN to output accurate estimates of the inverse noise SCMs and the RTF vector due to the incorporation of the MVDR structure. An argument against this expectation is the fact that the optimization objectives of the MVDR beamformer and the DNN—as defined by the loss function—are typically not aligned. Addressing

this research question, another goal of this thesis is to **investigate and improve the acoustic interpretability of estimated quantities in a specific coupled, structured estimation algorithm**.

1.2.3.4 *Summary*

In summary, decoupled approaches (class I) leverage the strengths of model-based and learning-based approaches while maintaining a clear separation between their respective roles. A key advantage of decoupled approaches is their modularity, allowing estimated quantities to be used with a variety of model-based algorithms without retraining. However, this modularity comes at the cost of potentially reduced speech enhancement performance due to the lack of coupling between the DNN and the model-based enhancement stage during training. In contrast, coupled approaches integrate the model-based enhancement stage as a differentiable operation within the computational graph, enabling end-to-end optimization and hence potentially resulting in a higher speech enhancement performance, but lacking the modularity of decoupled approaches. When not imposing structure on estimated quantities, the DNN in a coupled approach has more degrees of freedom, potentially resulting in a higher performance ceiling (particularly on the training dataset). However, this freedom comes at the cost of interpretability and potential robustness issues. In this thesis, we argue that imposing structure on estimated quantities, motivated by model-based approaches, is crucial for designing interpretable and robust hybrid speech enhancement algorithms. Incorporating mathematical and physical knowledge into the algorithm may improve robustness by regularizing the end-to-end training [32]. Therefore, one goal of this thesis is to **investigate different procedures to impose structure on estimated quantities** in hybrid speech enhancement approaches.

1.3 Thesis Outline and Main Contributions

Motivated by the potential to combine the interpretability of model-based approaches with the strong representation capacity of learning-based approaches, the primary objective of this thesis is to develop and evaluate hybrid single- and multi-microphone speech enhancement algorithms that employ deep neural networks to estimate the quantities required by a model-based enhancement stage. The main focus is on investigating whether imposing structure on estimated quantities—such as correlation matrix structure, correlation vector structure, or spatial structure—improves speech enhancement performance, interpretability, and computational complexity. Another focus is on developing geometry-robust hybrid speech enhancement algorithms that can operate with arbitrary microphone array configurations. While the proposed algorithms can be used for various speech enhancement applications, the main focus is on hearing devices, where low latency is crucial. To this end, we mainly consider causal multi-frame filters in the STFT domain as the

model-based enhancement stage, leveraging their inherent low-latency capabilities and applicability to dynamic acoustic scenarios.

The main contributions of this thesis are as follows. As a first contribution, **we propose a hybrid speech enhancement approach by embedding the single-microphone MFMVDR filter within a deep learning framework, explicitly imposing structure on the estimated temporal covariance matrices**. By coupling the MFMVDR-based enhancement stage with a DNN-based quantity estimation stage, substantial performance improvements are achieved over both a decoupled MFMVDR filter and a purely learning-based algorithm that does not impose structure on the filter coefficients. Furthermore, we demonstrate that imposing structure on the estimated temporal covariance matrices reduces computational complexity while preserving speech enhancement performance. As a second contribution, **we extend the hybrid single-microphone approach to multi-microphone speech enhancement for binaural hearing devices by embedding the binaural spatio-temporal Wiener filter within a deep learning framework, explicitly imposing structure on the estimated speech spatio-temporal correlation vectors**. Unlike the single-microphone MFMVDR filter, which exploits only temporal correlations, the binaural spatio-temporal Wiener filter explicitly exploits both temporal and spatial correlations of the speech and noise components. By decomposing the speech spatio-temporal correlation vectors into a spatial factor (corresponding to the RTF vector) and a temporal factor, we show that imposing correlation vector structure reduces computational complexity while maintaining speech enhancement performance and preserving binaural cues, outperforming two purely learning-based causal state-of-the-art binaural speech enhancement algorithms. Since the estimated RTF vectors however do not reflect the spatial characteristics of the acoustic scenario, as a third contribution **we propose a spatial regularization procedure to improve the interpretability of the estimated RTF vector, implicitly imposing spatial structure**. The proposed spatial regularization procedure yields accurate estimates of the RTF vector even in reverberant environments without sacrificing speech enhancement performance or increasing computational complexity. The aforementioned algorithms are designed for a fixed microphone array configuration and cannot generalize to unseen configurations. Addressing this limitation for the mask-based beamformer with attention-based spatial covariance matrix aggregator (ASA), as a fourth contribution **we propose three procedures to improve the robustness against varying microphone array configurations**: incorporating random microphone array configurations during training, employing the TAC method to enable permutation-invariant processing, and using robust input features. The combination of these procedures enables the application to unseen microphone array configurations and consistently outperforms both a baseline mask-based beamformer with recursive smoothing and the original mask-based beamformer with ASA.

In the remainder of this section, we provide a chapter-by-chapter overview of this thesis, highlighting the content and main contributions of each chapter. An overview of the thesis structure is illustrated in Fig. 1.14.

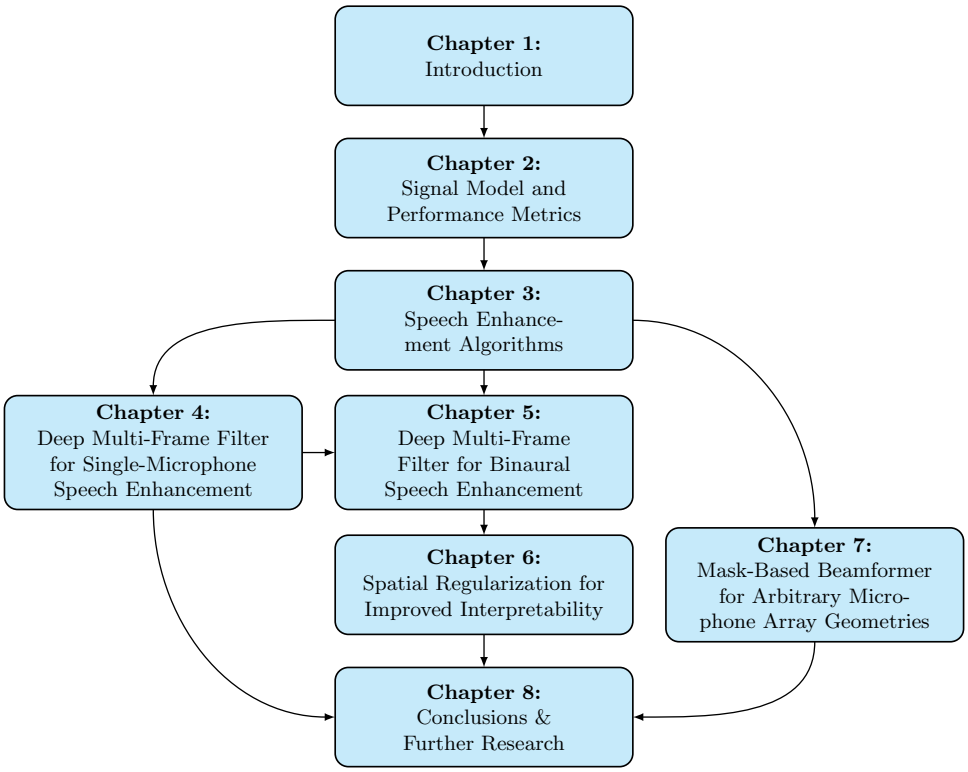


Figure 1.14: Overview of thesis structure.

In Chapter 2, we introduce the notation and STFT-domain signal models used throughout this thesis. Specifically, we present single-frame and multi-frame signal models in both single-microphone and multi-microphone configurations, as well as an extension for binaural hearing devices (Section 2.1). Additionally, we describe the objective performance measures used to evaluate the speech enhancement algorithms (Section 2.2).

In Chapter 3, we review the model-based, learning-based, and hybrid speech enhancement algorithms used throughout this thesis. First, we discuss two model-based optimal filters, namely the spatio-temporal minimum variance distortionless response filter and the spatio-temporal Wiener filter, which serve as the foundation for the proposed algorithms in this thesis. As special cases of the spatio-temporal MVDR filter with either a single time frame or a single microphone, we introduce the MFMVDR filter and the MVDR beamformer; and as a special case of the spatio-temporal Wiener filter with a single time frame, we introduce the binaural spatial Wiener filter. Furthermore, we review two learning-based speech enhancement algorithms, namely the deep filter algorithm and the Conv-TasNet algorithm, which serve as baseline speech enhancement algorithms in this thesis. Finally, we review a hybrid speech enhancement algorithm, namely the mask-based MVDR beamformer with attention-based spatial covariance matrix aggregator (ASA).

In Chapter 4, we propose a coupled, structured estimation hybrid speech enhancement approach by embedding the single-microphone MFMVDR filter within an end-to-end deep learning framework. Specifically, we train TCNs to estimate the noisy and interference TCMs as well as the a-priori SNR from the noisy speech STFT coefficients, minimizing the signal approximation SI-SDR loss function at the output of the MFMVDR filter. Since this estimation procedure yields strong speech enhancement performance, it is also adopted in subsequent chapters. For the interference TCM, we investigate imposing different matrix structures: Hermitian positive-definite, Hermitian positive-definite Toeplitz, and rank-1. For the Hermitian positive-definite structure, we consider estimation procedures based on both recursive smoothing and the Cholesky decomposition. The main differences between the investigated procedures lie in the number of parameters that need to be estimated by the TCNs and the required linear algebra operations, resulting in varying computational complexity. We show that with a rank-1 structure, the MFMVDR filter can be reformulated as a linear combination of the TCN outputs, avoiding computationally complex matrix inversions and thereby significantly reducing computational complexity. Using the DNS 1 challenge dataset, simulation results demonstrate that the TCM estimation procedure using the Hermitian positive-definite structure based on the Cholesky decomposition yields the best performance, while the rank-1 structure achieves almost the same performance at a lower computational complexity. Importantly, using the coupled hybrid approach leads to a substantial improvement in speech enhancement performance compared to a decoupled SPP-driven hybrid approach, as well as a slight but consistent improvement compared to a purely learning-based approach that does not impose structure on the multi-frame filter coefficients. The content in this chapter is related to the publications [176], [177].

In Chapter 5, we extend the coupled, structured estimation hybrid approach from Chapter 4 to binaural speech enhancement by embedding the binaural spatio-temporal Wiener filter (STWF) within an end-to-end deep learning framework. Aiming at reducing computational complexity while preserving speech enhancement performance and binaural cues, we impose various procedures to impose spatio-temporal correlation structures on the required interference spatio-temporal covariance matrices (STCMs) and speech spatio-temporal correlation vectors (STCVs) at both hearing devices, which mainly differ in the assumed relationship between microphones (particularly between the left and right device) and the number of parameters that need to be estimated. First, assuming that the spatial correlation of the speech component is stationary over the length of the multi-frame filter, we decompose the speech STCVs as the Kronecker product of an RTF vector and a TCV, separating the estimation process into a spatial factor and a temporal factor. We either consider a single "global" reference microphone, requiring the speech TCV to be estimated only for this microphone, or a reference microphone for each hearing device, requiring speech TCVs to be estimated for both (left and right) reference microphones. Second, we propose to replace the left and right interference STCMs with a common interference STCM, as the difference between both STCMs can be assumed to be negligible. Additionally, we consider a bilateral STWF by assuming no spatio-temporal correlation between both hearing devices, both for the speech STCVs and for the interference STCM. Using the DNS 1, DNS 2, CEC 1, and CEC 3 datasets, simulation results demonstrate that the binaural STWF using a combination of the speech STCV structure using two reference microphones and a common interference STCM significantly reduces computational complexity while achieving similar speech enhancement and binaural cue preservation performance compared to not imposing any spatio-temporal correlation structure. Furthermore, this (causal) deep binaural STWF outperforms the deep bilateral STWF, the binaural deep filter (DF) algorithm, and the binaural Conv-TasNet algorithm, approaching the performance of the non-causal binaural complex convolutional transformer network (BC-CTN) algorithm. The content in this chapter is related to the publications [178], [179].

In Chapter 6, we investigate the acoustic interpretability of the estimated RTF vector in the deep spatio-temporal MVDR algorithm. Similarly as for the binaural STWF in Chapter 5, we decompose the speech STCV into the Kronecker product of an RTF vector and a TCV. While end-to-end training with a signal approximation loss function is effective for speech enhancement, it does not ensure that the DNN outputs RTF vector estimates that reflect the spatial characteristics of the acoustic scenario, compromising interpretability. To address this issue, we propose a spatial regularization procedure that incorporates an additional loss term penalizing discrepancies between the estimated and ground truth RTF vectors. This loss term incentivizes the DNN to output estimates that are not only effective for speech enhancement but also reflect the spatial characteristics of the acoustic scenario. To automatically balance the individual loss terms, we employ an adaptive weighting method based on homoscedastic uncertainty. Using the DNS 1 and DNS 2 challenge datasets and simulated RIRs, simulation results demonstrate that the proposed spatial regularization procedure yields accurate estimates of the RTF vector, even in

reverberant environments, without sacrificing speech enhancement performance or increasing computational complexity. We hypothesize that this regularization procedure can be extended to other coupled hybrid speech enhancement approaches to improve the acoustic interpretability of estimated quantities.

In Chapter 7, we propose three procedures to improve the robustness of the mask-based beamformer with attention-based spatial covariance matrix aggregator (ASA) against varying microphone array configurations. First, we incorporate random channel configurations during training to prevent the DNN from overfitting to specific channel permutations and channel numbers. Second, we employ the TAC method to process multi-microphone features, allowing the algorithm to adapt to different channel numbers and enabling permutation invariance. Third, we utilize input features that are relatively insensitive to variations in channel configuration. Using the CHiME-3 and DEMAND datasets with simulated moving speakers, simulation results demonstrate that combining all three procedures improves generalization to unseen microphone arrays while maintaining speech enhancement performance under matched conditions. Furthermore, the mask-based beamformer combining all three procedures consistently outperforms both a baseline mask-based beamformer with recursive smoothing and the mask-based beamformer with the original ASA. The content in this chapter is related to the publication [180].

In Chapter 8, we summarize the main findings of this thesis and point out possible directions for further research.

2

SIGNAL MODEL AND PERFORMANCE METRICS

In this chapter, we introduce the signal models and notation that form the foundation for the speech enhancement algorithms presented in this thesis (Section 2.1). Additionally, we define the objective performance measures used to evaluate these algorithms (Section 2.2).

2.1 Signal Model

We consider an acoustic scenario where a single speech source and ambient noise are recorded in a reverberant environment. The noise may originate from both localized sources, such as keyboard typing and car engines, as well as diffuse sound fields, common in crowded spaces like restaurants. The acoustic scenario is captured using either a single device equipped with M microphones (single-device processing) or a binaural configuration consisting of two hearing devices with M_L and M_R microphones on the left and right devices, respectively, totaling $M = M_L + M_R$ microphones.

2.1.1 *Time Domain*

In the time domain, a single microphone captures a superposition of the reverberant speech and noise components. The noisy microphone signal $\acute{y}_{t_d,m} \in \mathbb{R}$ at discrete time index t_d and microphone index m is expressed as

$$\acute{y}_{t_d,m} = \acute{x}_{t_d,m} + \acute{n}_{t_d,m}, \tag{2.1}$$

where $\acute{x}_{t_d,m} \in \mathbb{R}$ denotes the reverberant speech component and $\acute{n}_{t_d,m} \in \mathbb{R}$ denotes the noise component. The vector containing the noisy microphone signal of the complete utterance with T_d samples at the m-th microphone can be written as

$$\acute{\mathbf{y}}_m = \begin{bmatrix} \acute{y}_{1,m} & \acute{y}_{2,m} & \dots & \acute{y}_{T_d,m} \end{bmatrix}^\mathsf{T} \in \mathbb{R}^{T_d} = \acute{\mathbf{x}}_m + \acute{\mathbf{n}}_m, \tag{2.2}$$

with $\acute{\mathbf{x}}_m$ and $\acute{\mathbf{n}}_m$ denoting the speech and noise vectors, respectively. For binaural configurations, the microphone indices are partitioned between devices, with indices $m \in \{1, \ldots, M_L\}$ corresponding to the left device and $m \in \{M_L+1, \ldots, M\}$ corresponding to the right device.

2.1.2 *STFT-Domain Representation*

To leverage the distinct spectro-temporal characteristics of speech and noise (Section 1.1.5), we transform the time-domain signals to a time-frequency representation using the STFT. The STFT segments the time-domain signal $\acute{y}_{t_d,m}$ into overlapping frames of length N_d samples and applies a tapered analysis window $\acute{w}_{t_d}$, such as the $\sqrt{\text{Hann}}$ window, to reduce spectral leakage. The discrete Fourier transform of each windowed frame yields complex-valued time-frequency coefficients as

$$y_{f,t,m} = \sum_{t_d=0}^{N_d-1} \acute{y}_{tT_s+t_d,m} \acute{w}_{t_d} \exp\left(-j2\pi f t_d / N_d\right) \in \mathbb{C}, \tag{2.3}$$

where f denotes the frequency bin index, t denotes the time frame index, and T_s denotes the frame shift in samples. Due to the conjugate symmetry of the Fourier transform of real-valued signals, we only need to process frequency bins $f \in \{0, \ldots, \lfloor N_d/2 \rfloor\}$, corresponding to positive frequencies up to the Nyquist frequency. The frame shift T_s determines the overlap between adjacent frames, with smaller values providing higher temporal correlation (Fig. 1.4) and temporal resolution at the cost of increased computational requirements. In the STFT domain, (2.1) can be written as

$$y_{f,t,m} = x_{f,t,m} + n_{f,t,m}, \quad m \in \{1, \ldots, M\}, \tag{2.4}$$

where $x_{f,t,m}$ and $n_{f,t,m}$ denote the speech and noise STFT coefficients, respectively. Since each frequency bin is processed independently, we omit the frequency index f in the remainder of this thesis for brevity, except when explicitly required.

2.1.3 *Multi-Microphone Signal Model*

To leverage both spatial information available from multiple microphones, we define the noisy multi-microphone vector $\check{\mathbf{y}}_t$ as

$$\check{\mathbf{y}}_t = \begin{bmatrix} y_{t,1} & \cdots & y_{t,M} \end{bmatrix}^\mathsf{T} = \check{\mathbf{x}}_t + \check{\mathbf{n}}_t \in \mathbb{C}^M, \tag{2.5}$$

where $\cdot^\mathsf{T}$ denotes the transpose operator, and $\check{\mathbf{x}}_t$ and $\check{\mathbf{n}}_t$ denote the multi-microphone speech and noise vectors, respectively.

TARGET SPEECH COMPONENT The relationship between the source signal—as uttered by the target speaker—and the speech components at all microphones

can be modeled using multiplicative ATFs. We can express the multi-microphone speech vector $\check{\mathbf{x}}_t$ as

$$\check{\mathbf{x}}_t = \begin{bmatrix} \check{b}_{t,1} \\ \vdots \\ \check{b}_{t,M} \end{bmatrix} s_t = \check{\mathbf{b}}_t s_t, \tag{2.6}$$

where s_t denotes the source signal and $\check{\mathbf{b}}_t$ denotes the (frequency- and time-dependent) ATF vector between the speech source and each microphone. The model in (2.6) assumes fully correlated speech components across microphones, which is a valid assumption if the STFT frame length is long enough to capture all propagation effects, such as delays, attenuation, reflections, and diffraction. To account for modeling errors that arise if the STFT frame length is not long enough to capture all propagation effects, (2.6) can be extended to include both a spatially correlated speech component and a spatially uncorrelated speech component as

$$\check{\mathbf{x}}_t = \underbrace{\check{\mathbf{b}}_t s_t}_{\text{correlated}} + \underbrace{\check{\mathbf{x}}'_t}_{\text{uncorrelated}}, \tag{2.7}$$

where $\check{\mathbf{x}}'_t$ denotes the spatially uncorrelated speech vector. Since many multi-microphone speech enhancement algorithms employ rather long STFT frames that indeed capture a sufficient fraction of propagation effects, the spatially uncorrelated speech vector is often neglected, leading back to the model in (2.6). However, if the STFT frame length is too short to capture a sufficient fraction of propagation effects, model errors may increase, and a convolutive ATF model may be preferred instead [5], [24].

To avoid the inclusion of the source signal s_t in the model in (2.6), RTFs can be used. RTFs model the relationship between the speech components at a reference microphone and all other microphones. The RTF between the reference microphone with index $r \in \{1, \ldots, M\}$ and the m-th microphone is defined as

$$h^r_{t,m} = \frac{\mathbb{E}\left(x_{t,m} x^*_{t,r}\right)}{\mathbb{E}\left(|x_{t,r}|^2\right)} = \frac{\check{\mathbf{e}}_m^{\mathsf{T}} \check{\mathbf{\Phi}}_{x,t} \check{\mathbf{e}}_r}{\phi^r_{x,t}}, \tag{2.8}$$

where $\mathbb{E}(\cdot)$ denotes the expectation operator, $\check{\mathbf{e}}_m \in \{0,1\}^M$ denotes a spatial selection vector with a 1 at the position corresponding to the m-th microphone and 0 elsewhere, $h^r_{t,r} = 1$ by definition, and $\phi^r_{x,t} = \mathbb{E}(|x_{t,r}|^2)$ denotes the speech PSD at the reference microphone. Using RTFs, we can express the multi-microphone speech vector as

$$\check{\mathbf{x}}_t = \begin{bmatrix} h^r_{t,1} \\ \vdots \\ h^r_{t,M} \end{bmatrix} x_{t,r} = \check{\mathbf{h}}^r_t x_{t,r}, \tag{2.9}$$

where $\check{\mathbf{h}}^r_t \in \mathbb{C}^M$ denotes the RTF vector for reference microphone r. Substituting (2.9) in (2.5), we obtain the multi-microphone signal model

$$\boxed{\check{\mathbf{y}}_t = \check{\mathbf{h}}^r_t x_{t,r} + \check{\mathbf{n}}_t.} \tag{2.10}$$

SPATIAL COVARIANCE MATRICES To capture spatial correlations (across microphones), we define the noisy SCM $\check{\mathbf{\Phi}}_{y,t}$ as

$$\check{\mathbf{\Phi}}_{y,t} = \mathbb{E}\left(\check{\mathbf{y}}_t \check{\mathbf{y}}_t^{\mathsf{H}}\right) \in \mathbb{C}^{M\times M}, \tag{2.11}$$

where $\check{\mathbf{y}}_t$ is the multi-microphone vector defined in (2.5) and $\cdot^{\mathsf{H}}$ denotes the Hermitian transpose operator. Assuming statistical independence between the speech and noise components, $\check{\mathbf{\Phi}}_{y,t}$ can be decomposed as

$$\check{\mathbf{\Phi}}_{y,t} = \check{\mathbf{\Phi}}_{x,t} + \check{\mathbf{\Phi}}_{n,t}, \tag{2.12}$$

where $\check{\mathbf{\Phi}}_{x,t} = \mathbb{E}(\check{\mathbf{x}}_t \check{\mathbf{x}}_t^{\mathsf{H}})$ and $\check{\mathbf{\Phi}}_{n,t} = \mathbb{E}(\check{\mathbf{n}}_t \check{\mathbf{n}}_t^{\mathsf{H}})$ denote the speech and noise SCMs, respectively. Using the RTF model for the multi-microphone speech vector in (2.9), the speech SCM can alternatively be written as a rank-1 matrix, i.e.,

$$\check{\mathbf{\Phi}}_{x,t} = \phi_{x,t}^r \check{\mathbf{h}}_t^r \check{\mathbf{h}}_t^{r,\mathsf{H}}. \tag{2.13}$$

Correspondingly, the RTF can be written using the speech SCM as

$$\check{\mathbf{h}}_t^r = \frac{\check{\mathbf{\Phi}}_{x,t}\check{\mathbf{e}}_r}{\phi_{x,t}^r}. \tag{2.14}$$

Furthermore, the speech PSD $\phi_{x,t}^r$ at the r-th microphone can be written using the speech SCM as

$$\phi_{x,t}^r = \check{\mathbf{e}}_r^{\mathsf{T}} \check{\mathbf{\Phi}}_{x,t} \check{\mathbf{e}}_r. \tag{2.15}$$

Finally, using (2.13), the noisy SCM in (2.12) can be written as

$$\boxed{\check{\mathbf{\Phi}}_{y,t} = \phi_{x,t}^r \check{\mathbf{h}}_t^r \check{\mathbf{h}}_t^{r,\mathsf{H}} + \check{\mathbf{\Phi}}_{n,t}.} \tag{2.16}$$

2.1.4 *Multi-Frame Signal Model*

To leverage the temporal correlations inherent in speech signals, we extend our signal model to incorporate multiple consecutive time frames. For each microphone m, we construct an N-dimensional noisy multi-frame vector containing the current time frame and $N-1$ previous time frames, i.e.,

$$\bar{\mathbf{y}}_{t,m} = \begin{bmatrix} y_{t,m} & \cdots & y_{t-N+1,m} \end{bmatrix}^{\mathsf{T}} = \bar{\mathbf{x}}_{t,m} + \bar{\mathbf{n}}_{t,m} \in \mathbb{C}^N, \tag{2.17}$$

where $\bar{\mathbf{x}}_{t,m}$ and $\bar{\mathbf{n}}_{t,m}$ denote the multi-frame speech and noise vectors, respectively.

TARGET SPEECH COMPONENT It was proposed in [3], [45] to decompose the multi-frame speech vector into a temporally correlated component and a temporally uncorrelated component w.r.t. the current speech STFT coefficient $x_{t,r}$ at the r-th microphone, i.e.,

$$\bar{\mathbf{x}}_{t,m} = \underbrace{\bar{\boldsymbol{\gamma}}_{t,m}^r x_{t,r}}_{\text{correlated}} + \underbrace{\bar{\mathbf{x}}_{t,m}^{\prime r}}_{\text{uncorrelated}}, \tag{2.18}$$

where $\bar{\mathbf{x}}_{t,m}^{\prime r}$ denotes the temporally uncorrelated speech vector, and where the speech temporal correlation vector $\bar{\boldsymbol{\gamma}}_{t,m}^{r} \in \mathbb{C}^N$ describes the correlation between the N most recent speech STFT coefficients at the m-th microphone and the current speech STFT coefficient at the r-th microphone $x_{t,r}$, i.e.,

$$\bar{\boldsymbol{\gamma}}_{t,m}^{r} = \frac{\mathbb{E}\left(\bar{\mathbf{x}}_{t,m} x_{t,r}^{*}\right)}{\phi_{x,t}^{r}} \in \mathbb{C}^N. \tag{2.19}$$

Due to this normalization, the first element of $\bar{\boldsymbol{\gamma}}_{t,m}^{m}$ is equal to 1, i.e.,

$$\bar{\mathbf{e}}^T \bar{\boldsymbol{\gamma}}_{t,m}^{m} = 1, \tag{2.20}$$

where $\bar{\mathbf{e}} = [1\ 0\ \ldots\ 0]^{\mathsf{T}} \in \mathbb{R}^N$ is a temporal selection vector with a 1 at the first position (corresponding to the current frame) and 0 elsewhere. Hence, the first element of the uncorrelated component $\bar{\mathbf{x}}_{t,m}^{\prime m}$ is equal to 0, i.e.,

$$\bar{\mathbf{e}}^T \bar{\mathbf{x}}_{t,m}^{\prime m} = 0. \tag{2.21}$$

It is important to note that the correlated speech component $\bar{\boldsymbol{\gamma}}_{t,m}^{r} x_{t,r}$ in (2.18) contains both the target speech STFT coefficient $x_{t,r}$ at the reference microphone as well as components that are correlated with $x_{t,r}$. The speech components that are uncorrelated with $x_{t,r}$, i.e., the elements of $\bar{\mathbf{x}}_{t,m}^{\prime r}$, are considered undesired.

The decomposition in (2.18) is mathematically similar to the decomposition in (2.7). However, a key distinction is how the uncorrelated component is handled in multi-microphone and multi-frame approaches. In multi-microphone approaches, the spatially uncorrelated speech vector $\check{\mathbf{x}}_{t,m}^{\prime}$ in (2.7) is typically neglected, as the STFT frame length is long enough to capture a sufficient fraction of propagation effects. In contrast, in multi-frame approaches, a short STFT frame length is typically chosen to effectively exploit temporal speech correlation (Section 1.1.1), and the uncorrelated speech vector $\bar{\mathbf{x}}_{t,m}^{\prime}$ in (2.18) is retained.

Substituting (2.18) in (2.17), we obtain the multi-frame signal model

$$\boxed{\bar{\mathbf{y}}_t = \bar{\boldsymbol{\gamma}}_{t,m}^{r} x_{t,r} + \underbrace{\bar{\mathbf{x}}_{t,m}^{\prime r} + \bar{\mathbf{n}}_t}_{=\bar{\mathbf{i}}_{t,m}^{r}},} \tag{2.22}$$

where the multi-frame interference vector $\bar{\mathbf{i}}_{t,m}^{r}$ contains both the uncorrelated speech component and the noise component.

TEMPORAL COVARIANCE MATRICES To capture temporal correlations (across time frames), we define the noisy TCM $\bar{\boldsymbol{\Phi}}_{y,t,m}$ at the m-th microphone as

$$\bar{\boldsymbol{\Phi}}_{y,t,m} = \mathbb{E}\left(\bar{\mathbf{y}}_{t,m} \bar{\mathbf{y}}_{t,m}^{\mathsf{H}}\right) \in \mathbb{C}^{N \times N}, \tag{2.23}$$

where $\bar{\mathbf{y}}_{t,m}$ is the multi-frame vector defined in (2.17). Assuming statistical independence between speech and noise, $\bar{\boldsymbol{\Phi}}_{y,t,m}$ can be decomposed as

$$\bar{\boldsymbol{\Phi}}_{y,t,m} = \bar{\boldsymbol{\Phi}}_{x,t,m} + \bar{\boldsymbol{\Phi}}_{n,t,m}, \tag{2.24}$$

where $\bar{\mathbf{\Phi}}_{x,t,m} = \mathbb{E}(\bar{\mathbf{x}}_{t,m}\bar{\mathbf{x}}_{t,m}^{\mathsf{H}})$ and $\bar{\mathbf{\Phi}}_{n,t,m} = \mathbb{E}(\bar{\mathbf{n}}_{t,m}\bar{\mathbf{n}}_{t,m}^{\mathsf{H}})$ denote the speech and noise TCMs at the m-th microphone, respectively. Using the multi-frame speech vector decomposition in (2.18), the speech TCM can be written as

$$\bar{\mathbf{\Phi}}_{x,t,m} = \phi_{x,t}^{m}\bar{\boldsymbol{\gamma}}_{t,m}^{r}(\bar{\boldsymbol{\gamma}}_{t,m}^{r})^{\mathsf{H}} + \bar{\mathbf{\Phi}}_{x',t,m}^{r}, \tag{2.25}$$

where $\bar{\mathbf{\Phi}}_{x',t,m}^{r} = \mathbb{E}(\bar{\mathbf{x}}_{t,m}^{\prime r}\bar{\mathbf{x}}_{t,m}^{\prime r,\mathsf{H}})$ denotes the TCM of the uncorrelated speech component.

Similarly as in (2.15), the speech PSD at the m-th microphone $\phi_{x,t}^{m}$ can be written using the speech TCM as

$$\phi_{x,t}^{m} = \bar{\mathbf{e}}^{\mathsf{T}}\bar{\mathbf{\Phi}}_{x,t,m}\bar{\mathbf{e}}. \tag{2.26}$$

Because the first element of the uncorrelated speech vector in (2.21) is equal to 0, the first column and the first row of $\bar{\mathbf{\Phi}}_{x',t,m}^{r} = \mathbb{E}(\bar{\mathbf{x}}_{t,m}^{\prime r}\bar{\mathbf{x}}_{t,m}^{\prime r,\mathsf{H}})$ are also equal to 0. Consequently, the speech TCV can be written using the speech TCM as

$$\bar{\boldsymbol{\gamma}}_{t,m}^{m} = \frac{\bar{\mathbf{\Phi}}_{x,t,m}\bar{\mathbf{e}}}{\phi_{x,t}^{m}}. \tag{2.27}$$

Finally, using (2.25), the noisy TCM in (2.24) can be written as

$$\boxed{\bar{\mathbf{\Phi}}_{y,t,m} = \phi_{x,t}^{m}\bar{\boldsymbol{\gamma}}_{t,m}^{m}(\bar{\boldsymbol{\gamma}}_{t,m}^{m})^{\mathsf{H}} + \underbrace{\bar{\mathbf{\Phi}}_{x',t,m}^{m} + \bar{\mathbf{\Phi}}_{n,t,m}}_{=:\bar{\mathbf{\Phi}}_{i,t,m}^{r}},} \tag{2.28}$$

with the interference TCM $\bar{\mathbf{\Phi}}_{i,t,m}^{r} = \mathbb{E}(\bar{\mathbf{i}}_{t,m}^{r}\bar{\mathbf{i}}_{t,m}^{r,\mathsf{H}})$.

2.1.5 *Multi-Microphone, Multi-Frame Signal Model*

To leverage both the spatial information available from multiple microphones and the temporal correlations inherent in speech signals, we define the multi-microphone multi-frame vector $\mathbf{y}_t \in \mathbb{C}^{MN}$ as

$$\mathbf{y}_t = \begin{bmatrix}\bar{\mathbf{y}}_{t,1}^{\mathsf{T}} & \dots & \bar{\mathbf{y}}_{t,M}^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}} = \mathbf{x}_t + \mathbf{n}_t, \tag{2.29}$$

where $\mathbf{x}_t$ and $\mathbf{n}_t$ denote the multi-microphone multi-frame speech and noise vectors, respectively. Hence, the vector $\mathbf{y}_t$ concatenates the multi-frame vectors from all M microphones.

Using the RTF vector from (2.9), the multi-frame speech vector at the m-th microphone can be written using the speech component at the reference microphone $x_{t,r}$ as

$$\bar{\mathbf{x}}_{t,m} = \begin{bmatrix} x_{t,m} \\ x_{t-1,m} \\ \vdots \\ x_{t-N+1,m}\end{bmatrix} = \begin{bmatrix} h_{t,m}^{r}x_{t,r} \\ h_{t-1,m}^{r}x_{t-1,r} \\ \vdots \\ h_{t-N+1,m}^{r}x_{t-N+1,r}\end{bmatrix}. \tag{2.30}$$

Assuming that the RTFs are constant over N frames (i.e., $h^r_{t,m}$, $h^r_{t-1,m}$, $\ldots$, $h^r_{t-N+1,m}$ are equal), the multi-frame speech vector in (2.30) can be written as [3], [47]

$$\bar{\mathbf{x}}_{t,m} = h^r_{t,m}\bar{\mathbf{x}}_{t,r}. \tag{2.31}$$

Without loss of generality, we select the first microphone as the reference microphone ($r = 1$). The multi-microphone multi-frame speech vector can then be modeled as

$$\mathbf{x}_t = \begin{bmatrix} \bar{\mathbf{x}}_{t,1} \\ \bar{\mathbf{x}}_{t,2} \\ \vdots \\ \bar{\mathbf{x}}_{t,M} \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{x}}_{t,1} \\ h^1_{t,2}\bar{\mathbf{x}}_{t,1} \\ \vdots \\ h^1_{t,M}\bar{\mathbf{x}}_{t,1} \end{bmatrix} = \check{\mathbf{h}}^1_t \otimes \bar{\mathbf{x}}_{t,1}, \tag{2.32}$$

where $\otimes$ denotes the Kronecker product.

SPEECH COMPONENT DECOMPOSITION Similarly to the multi-frame decomposition in (2.18), the multi-microphone multi-frame speech vector $\mathbf{x}_t$ can be decomposed into a spatio-temporally correlated component and a spatio-temporally uncorrelated component with respect to the current speech STFT coefficient at the r-th microphone, i.e.,

$$\mathbf{x}_t = \underbrace{\boldsymbol{\gamma}^r_t x_{t,r}}_{\text{correlated}} + \underbrace{\mathbf{x}'^r_t}_{\text{uncorrelated}}, \tag{2.33}$$

where $\mathbf{x}'^r_t$ denotes the spatio-temporally uncorrelated speech vector and the speech STCV $\boldsymbol{\gamma}^r_t$ describes the correlation between the N most recent speech STFT coefficients at each microphone and the current speech STFT coefficient at the r-th microphone. The speech STCV is defined as

$$\boldsymbol{\gamma}^r_t = \frac{\mathbb{E}(\mathbf{x}_t x^*_{t,r})}{\mathbb{E}(|x_{t,r}|^2)} = \frac{\mathbb{E}(\mathbf{x}_t x^*_{t,r})}{\phi^r_{x,t}} \in \mathbb{C}^{MN}, \tag{2.34}$$

and it depends on both the temporal correlation, which is highly time-varying, and the spatial correlation, which is typically less time-varying. To separate the spatial dependence from the temporal dependence, we can substitute the Kronecker product factorization of $\mathbf{x}_t$ in (2.32) into the definition of the STCV in (2.34), i.e.,

$$\boldsymbol{\gamma}^r_t = \frac{\mathbb{E}\left(\left(\check{\mathbf{h}}^1_t \otimes \bar{\mathbf{x}}_{t,1}\right) x^*_{t,r}\right)}{\phi^r_{x,t}} \tag{2.35}$$

$$= \check{\mathbf{h}}^1_t \otimes \frac{\mathbb{E}\left(\bar{\mathbf{x}}_{t,1} x^*_{t,r}\right)}{\phi^r_{x,t}}. \tag{2.36}$$

Using the definition of the speech TCV in (2.19), (2.35) can hence be written as

$$\boldsymbol{\gamma}^r_t = \check{\mathbf{h}}^1_t \otimes \bar{\boldsymbol{\gamma}}^r_{t,1}, \tag{2.37}$$

i.e., the speech STCV can be decomposed into the Kronecker product of the RTF vector and the speech TCV.

Due to the normalization with the speech PSD in (2.34), the element of the speech STCV corresponding to the current frame at the r-th microphone is equal to 1, i.e.,

$$\mathbf{e}_r^{\mathsf{T}}\boldsymbol{\gamma}_t^r = 1, \tag{2.38}$$

where $\mathbf{e}_r \in \{0,1\}^{MN}$ is a spatio-temporal selection vector with a 1 at the position corresponding to the current frame of the r-th microphone and 0 elsewhere. Hence, the corresponding element of $\mathbf{x}_t'^r$ in (2.33) is equal to 0, i.e.,

$$\mathbf{e}_r^{\mathsf{T}}\mathbf{x}_t'^r = 0. \tag{2.39}$$

Similarly as for the temporal signal model, the spatio-temporally correlated speech component $\boldsymbol{\gamma}_t^r x_{t,r}$ in (2.33) contains both the target speech STFT coefficient $x_{t,r}$ at the reference microphone as well as components that are correlated with $x_{t,r}$. The speech components that are uncorrelated with $x_{t,r}$, i.e., the elements of $\mathbf{x}_t'^r$, are again considered undesired. Substituting (2.33) into (2.29), we obtain the multi-microphone multi-frame signal model

$$\boxed{\mathbf{y}_t = \boldsymbol{\gamma}_t^r x_{t,r} + \underbrace{\mathbf{x}_t'^r + \mathbf{n}_t}_{\mathbf{i}_t^r},} \tag{2.40}$$

where the multi-microphone multi-frame interference vector $\mathbf{i}_t^r$ contains both the spatio-temporally uncorrelated speech component and the noise component.

SPATIO-TEMPORAL COVARIANCE MATRICES To capture both spatial and temporal correlations, we define the noisy STCM $\boldsymbol{\Phi}_{y,t}$ as

$$\boldsymbol{\Phi}_{y,t} = \mathbb{E}\left(\mathbf{y}_t\mathbf{y}_t^{\mathsf{H}}\right) \in \mathbb{C}^{MN\times MN}, \tag{2.41}$$

where $\mathbf{y}_t$ is the multi-microphone multi-frame vector defined in (2.29). Again, assuming statistical independence between speech and noise, $\boldsymbol{\Phi}_{y,t}$ can be decomposed as

$$\boldsymbol{\Phi}_{y,t} = \boldsymbol{\Phi}_{x,t} + \boldsymbol{\Phi}_{n,t}, \tag{2.42}$$

where $\boldsymbol{\Phi}_{x,t} = \mathbb{E}(\mathbf{x}_t\mathbf{x}_t^{\mathsf{H}})$ and $\boldsymbol{\Phi}_{n,t} = \mathbb{E}(\mathbf{n}_t\mathbf{n}_t^{\mathsf{H}})$ denote the speech and noise STCMs, respectively. Using the decomposition in (2.33), the speech STCM can be written as

$$\boldsymbol{\Phi}_{x,t} = \phi_{x,t}^r\boldsymbol{\gamma}_t^r\boldsymbol{\gamma}_t^{r,\mathsf{H}} + \boldsymbol{\Phi}_{x',t}^r \tag{2.43}$$

where $\boldsymbol{\Phi}_{x',t} = \mathbb{E}(\mathbf{x}_t'^r\mathbf{x}_t'^{r,\mathsf{H}})$ denotes the STCM of the spatio-temporally uncorrelated speech component. Similarly as in (2.15) and (2.26), the speech PSD at the r-th microphone $\phi_{x,t}^r$ can be written using the speech STCM as

$$\phi_{x,t}^r = \mathbf{e}_r^{\mathsf{T}}\boldsymbol{\Phi}_{x,t}\mathbf{e}_r. \tag{2.44}$$

Because the first element of the spatio-temporally uncorrelated speech vector in (2.39) is equal to 0, the column and row of $\mathbf{\Phi}^{r}_{x',t,m} = \mathbb{E}(\mathbf{x}'^{r}_{t,m}\mathbf{x}'^{r,\mathsf{H}}_{t,m})$ corresponding to the current frame at the r-th microphone are also equal to 0. Consequently, the speech STCV can be written using the speech STCM as

$$\boldsymbol{\gamma}^{r}_{t} = \frac{\mathbf{\Phi}_{x,t}\mathbf{e}_{r}}{\phi^{r}_{x,t}}. \tag{2.45}$$

Finally, using (2.43), the noisy STCM can be written as

$$\boxed{\mathbf{\Phi}_{y,t} = \phi^{r}_{x,t}\boldsymbol{\gamma}^{r}_{t}\boldsymbol{\gamma}^{r,\mathsf{H}}_{t} + \underbrace{\mathbf{\Phi}^{r}_{x',t} + \mathbf{\Phi}_{n,t}}_{=:\mathbf{\Phi}^{r}_{i,t}},} \tag{2.46}$$

with the interference STCM $\mathbf{\Phi}^{r}_{i,t} = \mathbb{E}(\mathbf{i}^{r}_{t}\mathbf{i}^{r,\mathsf{H}}_{t})$.

Overall, the multi-microphone signal model and the multi-frame signal model can be seen as special cases of the multi-microphone multi-frame signal model with either $N = 1$ or $M = 1$, respectively.

FILTERING AND SUMMING Many multi-microphone speech enhancement algorithms estimate the target speech component at a reference microphone by filtering and summing all available microphone signals. For the (monaural) multi-microphone case, the estimated speech component at a single reference microphone r is obtained as

$$\widehat{x}_{t,r} = \check{\mathbf{w}}^{r,\mathsf{H}}_{t}\check{\mathbf{y}}_{t}, \tag{2.47}$$

where the spatial filter $\check{\mathbf{w}}^{r}_{t} \in \mathbb{C}^{M}$ is defined similarly to the multi-microphone vector in (2.5). While the theoretical performance of such algorithms is often independent of reference microphone selection, estimation errors can significantly impact speech enhancement performance in practice [181], especially in scenarios where the microphones are spatially separated.

For the multi-frame case, the estimated speech component at the r-th microphone is obtained as

$$\widehat{x}_{t,r} = \bar{\mathbf{w}}^{r,\mathsf{H}}_{t}\bar{\mathbf{y}}_{t,r}, \tag{2.48}$$

where the temporal filter $\bar{\mathbf{w}}^{r}_{t} \in \mathbb{C}^{N}$ is defined similarly to the multi-frame vector in (2.17).

Finally, for the multi-microphone multi-frame case, the estimated speech component at the r-th microphone is obtained as

$$\widehat{x}_{t,r} = \mathbf{w}^{r,\mathsf{H}}_{t}\mathbf{y}_{t}, \tag{2.49}$$

where the spatio-temporal filter $\mathbf{w}^{r}_{t} \in \mathbb{C}^{MN}$ is defined similarly to the multi-microphone multi-frame vector in (2.29).

For consistency, we will use the terms "spatial filter", "temporal filter", and "spatio-temporal filter" in the remainder of this thesis to distinguish between filters based on

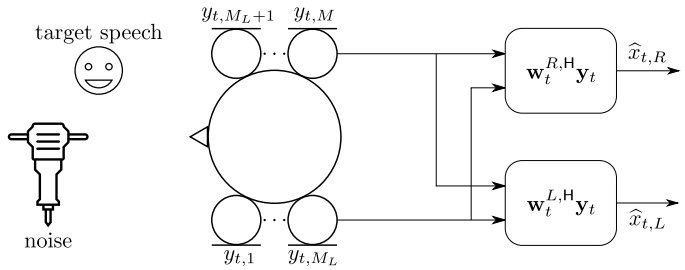


Figure 2.1: Binaural processing scheme, estimating the target speech component at the left and the right hearing device by filtering all available microphone signals.

the multi-microphone signal model in (2.47), the multi-frame signal model in (2.48), and the multi-microphone multi-frame signal model in (2.49), respectively, and we reserve the term "beamformer" for steerable spatial filters such as the MVDR beamformer.

2.1.6 *Binaural Extension*

In contrast to monaural speech enhancement algorithms, binaural speech enhancement algorithms estimate the target speech components at both hearing devices (Fig. 2.1). More in particular, in the case of spatio-temporal filters, the target speech components are estimated using $\mathbf{w}_t^L \in \mathbb{C}^{MN}$ (for the left device) and $\mathbf{w}_t^R \in \mathbb{C}^{MN}$ (for the right device), i.e.,

$$\widehat{x}_{t,L} = \mathbf{w}_t^{L,\mathsf{H}}\mathbf{y}_t, \quad \widehat{x}_{t,R} = \mathbf{w}_t^{R,\mathsf{H}}\mathbf{y}_t, \tag{2.50}$$

where, without loss of generality, the left device reference microphone has been chosen as $L = 1$, the right device reference microphone has been chosen as $R = M_L + 1$, and the filters $\mathbf{w}_{t,m}^L$ and $\mathbf{w}_{t,m}^R$ are defined similarly as in (2.29). In principle, using hearing devices on both ears can generate an important advantage, both from a signal processing perspective, since all microphone signals from both devices can be used, as well as from a perceptual perspective, since interaural cues can be exploited by the auditory system [5], [25]. An important distinction exists between *bilateral* systems, where both devices operate independently, and *binaural* systems, where microphone signals from both devices are processed and combined in each device.

The decomposition in (2.33) can be carried out for both reference microphones ($m = L$ and $m = R$), i.e.,

$$\boxed{\begin{aligned}\mathbf{x}_t &= \boldsymbol{\gamma}_t^L x_{t,L} + \mathbf{x}_t'^L \\ &= \boldsymbol{\gamma}_t^R x_{t,R} + \mathbf{x}_t'^R ,\end{aligned}} \tag{2.51}$$

where $\boldsymbol{\gamma}_t^L$ and $\boldsymbol{\gamma}_t^R$ are the left and right speech STCVs, respectively, and $\mathbf{x}_t'^L$ and $\mathbf{x}_t'^R$ are the corresponding uncorrelated speech components.

Using (2.51), the noisy multi-microphone multi-frame vector can be written with respect to either the left or right reference microphone, i.e.,

$$\boxed{\mathbf{y}_t = \boldsymbol{\gamma}_t^{\nu} x_{t,\nu} + \underbrace{\mathbf{x}_t'^{\nu} + \mathbf{n}_t}_{=:\mathbf{i}_t^{\nu}}, \quad \nu \in \{L, R\},} \tag{2.52}$$

where $\mathbf{i}_t^{\nu}$ denotes the interference vector for the left ($\nu = L$) or right ($\nu = R$) device. The noisy STCM $\boldsymbol{\Phi}_{y,t}$ can then be decomposed w.r.t. either reference microphone, i.e.,

$$\boxed{\boldsymbol{\Phi}_{y,t} = \phi_{x,t}^{\nu} \boldsymbol{\gamma}_t^{\nu} \boldsymbol{\gamma}_t^{\nu,\mathsf{H}} + \underbrace{\boldsymbol{\Phi}_{x',t}^{\nu} + \boldsymbol{\Phi}_{n,t}}_{=\boldsymbol{\Phi}_{i,t}^{\nu}}, \quad \nu \in \{L, R\}.} \tag{2.53}$$

It is important to note that, in general, the speech PSDs ($\phi_{x,t}^{\nu}$), the speech STCVs ($\boldsymbol{\gamma}_t^{\nu}$), the uncorrelated speech STCMs ($\boldsymbol{\Phi}_{x',t}^{\nu}$), and the interference STCMs ($\boldsymbol{\Phi}_{i,t}^{\nu}$) are different for the left and right hearing devices.

Figure 2.2 illustrates the structure of the speech STCM $\boldsymbol{\Phi}_{x,t}$ for three different frequencies (500 Hz, 1000 Hz, and 2000 Hz) in a binaural setup with one microphone on each hearing device ($M_L = M_R = 1$) and three time frames ($N = 3$), computed using the oracle speech component and an STFT configuration with 8 ms frame length, 2 ms frame shift, and a $\sqrt{\text{Hann}}$ window. The matrix $\boldsymbol{\Phi}_{x,t}$ can be partitioned into the submatrices $\tilde{\boldsymbol{\Phi}}_{x,t}^{LL}$ and $\tilde{\boldsymbol{\Phi}}_{x,t}^{RR}$, containing only spatio-temporal correlations of the speech component at the left and right hearing device, respectively, and the submatrix $\tilde{\boldsymbol{\Phi}}_{x,t}^{LR}$ containing spatio-temporal correlations between the contralateral microphones. As the speaker in the depicted scenario is positioned on the right side of the listener, the submatrix $\tilde{\boldsymbol{\Phi}}_{x,t}^{RR}$ is characterized by larger (co-)variance values than the other submatrices. Furthermore, the high temporal correlation of the speech component already discussed in Section 1.1.1 is evident in each of the submatrices.

2.2 Objective Performance Metrics

While subjective listening tests remain the gold standard for evaluating the performance of speech enhancement algorithms, they are time-consuming, costly, and require careful control of experimental conditions. Consequently, objective performance metrics that can reliably predict the outcome of subjective listening tests are highly desirable. These metrics can be broadly categorized into those that assess speech quality, speech intelligibility, or—particularly in the context of binaural speech enhancement—binaural cue preservation. Speech quality refers to the overall pleasantness and naturalness of the speech signal, whereas speech intelligibility refers to the accuracy with which the linguistic content of the speech can be understood. As highlighted in [14], these two aspects, while often correlated, are distinct and may be affected differently by different types of signal degradations or algorithms, especially in the context of highly nonlinear processing often associated

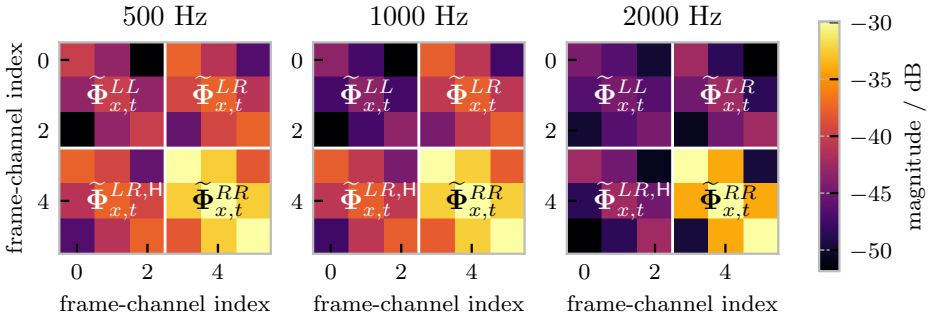


Figure 2.2: Magnitude of speech STCM $\mathbf{\Phi}_{x,t}$ with hearing device-partitioning for three frequencies (500 Hz, 1000 Hz, and 2000 Hz), averaged across an utterance in a binaural setup with one microphone on each hearing device ($M_L = M_R = 1$) and three time frames ($N = 3$), with the target source positioned on the right side of the listener (at an angle of -85°), computed using an STFT configuration with 8 ms frame length, 2 ms frame shift, and a $\sqrt{\text{Hann}}$ window.

with learning-based approaches. In this section, we present the set of objective metrics used in this thesis to evaluate the performance of both monaural and binaural speech enhancement algorithms, encompassing metrics for

- speech quality: scale-invariant signal-to-distortion ratio (SI-SDR), perceptual evaluation of speech quality (PESQ), hearing aid speech quality and speech intelligibility index (HASQI), Deep Noise Suppression Mean Opinion Score (DNSMOS)
- speech intelligibility: short-time objective intelligibility (STOI), hearing aid speech perception index (HASPI)
- binaural cue preservation: interaural level difference (ILD), interaural phase difference (IPD)
- computational complexity: real-time factor (RF), multiply-accumulate operations per second (MACS), number of trainable parameters

2.2.1 *Speech Quality*

SI-SDR The scale-invariant signal-to-distortion ratio is a widely used intrusive metric for evaluating the performance of speech enhancement and source separation algorithms [141]. "Intrusive" metrics require both a clean reference signal and the corresponding degraded (e.g., noisy) or processed signal (termed "estimated signal" in the following). Due to its good trade-off between simplicity and resulting speech enhancement performance, it has also been popular as a loss function for training algorithms. Unlike the signal-to-distortion ratio implemented in the popular `BSS_eval` toolbox [182], SI-SDR addresses issues related to the handling of

scaling and filtering. Specifically, SI-SDR considers the orthogonal projection of the estimated signal onto the reference signal in the time-domain, ensuring that the metric is invariant to the scaling of the estimated signal relative to the reference signal. This is achieved by scaling the target such that the distance to the estimate is minimized, i.e.,

$$\text{SI-SDR} = 10\log_{10}\left(\frac{\|\alpha\acute{\mathbf{x}}\|^2}{\left\|\alpha\acute{\mathbf{x}} - \widehat{\acute{\mathbf{x}}}\right\|^2}\right), \quad \alpha = \frac{\widehat{\acute{\mathbf{x}}}^{\mathsf{T}}\acute{\mathbf{x}}}{\|\acute{\mathbf{x}}\|_2^2}, \tag{2.54}$$

where $\acute{\mathbf{x}}$ and $\widehat{\acute{\mathbf{x}}}$ denote the reference signal and the estimated signal after inverse STFT processing in the time domain, defined similarly as in (2.2). SI-SDR's robustness to scaling discrepancies makes it particularly valuable in scenarios where output levels may vary and these variations do not impact performance. However, in scenarios where scaling discrepancies do matter (such as in hearing device application), SI-SDR may not be appropriate. Furthermore, as SI-SDR is considered a "waveform-matching" metric, even small phase differences between the reference and estimated signals may yield extremely low values, which does not match human perception. Hence, the SI-SDR should be used with caution for algorithms that can modify the phase, especially in reverberant scenarios [138], [140].

PESQ PESQ is an intrusive metric designed to predict the perceived quality of speech as perceived by human listeners [183]. It was developed to address a wide range of network conditions, including those with variable delay, packet loss, different codec types, and background noise. The PESQ algorithm normalizes the reference and estimated signals, models a standard telephone handset using an input filter, time-aligns the signals, and processes them through an auditory transform. The metric accounts for effects such as filtering, additive noise, and time warping. It does so by calculating a disturbance parameter, which represents the audible error between the reference and estimated signals. The disturbance parameter is then mapped to a Mean Opinion Score (MOS) through a cognitive model, providing a prediction of perceived speech quality and speech intelligibility. PESQ was adopted as the ITU-T recommendation P.862 (and its wideband extension P.862.1) and is particularly valuable for evaluating speech quality and speech intelligibility in telecommunication systems, but is less suitable for evaluating distortions resulting from speech enhancement algorithms. Although officially replaced by perceptual objective listening quality assessment (POLQA) [184] as the ITU-T recommendation P.863, PESQ is still widely used, in part due to the ease of accessibility. For example, the `pesq` Python package[1] includes two narrowband variants and the wideband variant, but the specific variant that was used is often not reported in the literature, creating some confusion about the exact scores.

1 Available at `https://pypi.org/project/pesq/`.

HASQI The hearing aid speech quality and speech intelligibility index (HASQI) is an intrusive metric designed to predict the perceived quality of speech processed by hearing devices, applicable to both normal-hearing and hearing-impaired listeners [185]. The index is based on a model of the auditory periphery that simulates the changes caused by hearing impairment, including factors such as reduced audibility, broadened auditory filters, and reduced dynamic range compression. HASQI version 2 combines measures of changes in the signal envelope, temporal fine structure, and long-term spectrum. The model computes a nonlinear term, which is sensitive to noise and nonlinear distortion, and a linear term, which is sensitive to long-term spectral changes. This approach allows HASQI to account for a wide variety of signal degradations, including those commonly encountered in hearing device processing, such as frequency compression, noise reduction, and feedback cancellation.

DNSMOS DNSMOS is a non-intrusive metric designed to evaluate the perceptual quality of speech, particularly for speech processed by speech enhancement algorithms [149]. Unlike most metrics, which are typically intrusive, DNSMOS does not need a reference signal, making it suitable for evaluating real-world recordings where clean reference signals are typically unavailable. DNSMOS was initially trained using subjective quality ratings obtained through the ITU-T Rec. P.808 methodology, which specifies a standardized procedure for conducting online subjective listening tests to assess speech quality and speech intelligibility. The metric employs a DNN, with the log power spectrogram of the evaluated signal as the input feature. The original DNSMOS metric [149] provided an overall quality score. In [150], DNSMOS was extended to predict the quality of speech (SIG), background noise (BAK), and overall quality (OVRL) separately, following the ITU-T Rec. P.835 subjective evaluation framework. The resulting DNSMOS P.835 has shown high correlation with human ratings, achieving a Pearson's correlation coefficient of $r = 0.94$ for SIG and $r = 0.98$ for BAK and OVRL.

2.2.2 *Speech Intelligibility*

STOI STOI is an intrusive metric designed to predict the intelligibility of noisy speech, including speech processed by algorithms that apply time-frequency weighting [186]. The metric evaluates intelligibility in short-time segments (approximately 400 ms) and then averages these segment-based scores to obtain an overall intelligibility score. It utilizes an STFT for time-frequency analysis, groups frequency bins into one-third octave bands, and calculates an intermediate intelligibility metric for each time-frequency bin. The intermediate intelligibility metric is derived from the linear correlation coefficient between the reference and estimated time-frequency bins. The model has shown high correlation ($r = 0.95$) with subjective intelligibility scores for both noisy and time-frequency weighted noisy speech.

HASPI The hearing aid speech perception index (HASPI) is an intrusive metric designed to predict the intelligibility of speech processed by hearing devices, ap-

plicable to both normal-hearing and hearing-impaired listeners [187]. Similarly as HASQI, HASPI is based on a model of the auditory periphery that simulates the changes caused by hearing impairment, including factors such as reduced audibility, broadened auditory filters, and reduced dynamic range compression. HASPI compares the time-frequency envelope and temporal fine structure of an estimated signal to a reference signal. The original HASPI version 1 uses a combination of low-pass filtered envelope and temporal fine structure data, along with a parametric model, to estimate intelligibility. HASPI version 2 introduces two key modifications: First, it replaces the low-pass envelope filter and temporal fine structure analysis with an envelope modulation filterbank to better capture a wider range of modulation rates relevant to speech intelligibility. Second, it uses an ensemble of DNNs instead of a parametric model to combine envelope and temporal fine structure information before mapping it to intelligibility scores. The revised HASPI version 2 has shown improved accuracy compared to version 1 for various degradations including noise, nonlinear distortion, frequency compression, noise reduction, vocoded speech, and reverberation.

2.2.3 *Binaural Cue Preservation*

ILD AND IPD Interaural level differences (ILDs) and interaural phase differences (IPDs) provide cues about sound sources, aiding the human auditory system in source localization, improving speech intelligibility in noisy environments, and providing spatial awareness. Preserving these cues in the output signals is hence crucial for binaural speech enhancement algorithms. To evaluate binaural cue preservation, we consider the ILD and IPD errors between the output signals and the target speech components of the input signals. We use the ILD and IPD errors defined in [188], where both errors are computed only in STFT bins with active speech. The ILD error is defined as

$$\Delta\text{ILD} = \frac{1}{A}\sum_{f=1}^{F}\sum_{t=1}^{T}\mathcal{M}_{f,t}|\text{ILD}_{f,t}^{\text{out}} - \text{ILD}_{f,t}^{\text{in}}|, \tag{2.55}$$

where $\mathcal{M}_{f,t}$ denotes an ideal binary mask indicating STFT bins with active speech computed from the target speech components (see [188] for more details), and $A = \sum_f^F \sum_t^T \mathcal{M}_{f,t}$ denotes the number of STFT bins with active speech. The ILDs between the output signals and the ILDs between the target speech components of the input signals are defined as

$$\text{ILD}_{f,t}^{\text{out}} = \frac{|\widehat{x}_{f,t,L}|^2}{|\widehat{x}_{f,t,R}|^2}, \quad \text{ILD}_{f,t}^{\text{in}} = \frac{|x_{f,t,L}|^2}{|x_{f,t,R}|^2}. \tag{2.56}$$

Similarly, the IPD error is defined as

$$\Delta\text{IPD} = \frac{1}{A}\sum_{f=1}^{F}\sum_{t=1}^{T}\mathcal{M}_{f,t}|\text{IPD}_{f,t}^{\text{out}} - \text{IPD}_{f,t}^{\text{in}}|, \tag{2.57}$$

where the IPDs between the output signals and the IPDs between the target speech components of the input signals are defined as

$$\mathrm{IPD}_{f,t}^{\mathrm{out}} = \angle\left(\frac{\widehat{x}_{f,t,L}}{\widehat{x}_{f,t,R}}\right), \quad \mathrm{IPD}_{f,t}^{\mathrm{in}} = \angle\left(\frac{x_{f,t,L}}{x_{f,t,R}}\right). \tag{2.58}$$

Note that these ILD and IPD errors deviate from traditional definitions, which typically consider the shadow-filtered speech components ($\mathbf{w}_{f,t}^{L,\mathsf{H}}\mathbf{x}_{f,t}$ and $\mathbf{w}_{f,t}^{R,\mathsf{H}}\mathbf{x}_{f,t}$) instead of the estimated speech components [56]. The definitions in (2.55) and (2.57) allow the evaluation even for algorithms that perform non-linear processing of the input signals such as the binaural Conv-TasNet algorithm [189] (which serves as a baseline algorithm in this thesis).

Furthermore, it should be noted that these metrics may provide an indication of the cue preservation performance of an algorithm but conclusions drawn from these metrics should be interpreted cautiously, since these metrics do not include a model of the human auditory system. Hence, the relationship between differences in these metrics and the resulting perceptual impact may not be straightforward [190].

2.2.4 *Computational Complexity*

Computational complexity metrics are essential for assessing the feasibility of implementing speech enhancement algorithms on real-time, resource-constrained devices, such as hearing devices. These devices often have limited processing power, memory, and battery life, making computational complexity a critical consideration. An algorithm that provides excellent enhancement performance but has excessive computational demands may be unsuitable for real-time, on-device processing. Similarly to the binaural cue preservation metrics, the following considered computational complexity metrics may indicate differences between algorithms, but the practical suitability of an algorithm in a specific setting is ultimately determined by the algorithm's exact on-device implementation.

RF The real-time factor (RF) measures the time taken by an algorithm to process a signal relative to the duration of that signal. For real-time applications, the RF *must* be less than or equal to 1 to ensure that the output can be generated continuously without introducing delays. An RF greater than 1 indicates that the implementation cannot keep up with the incoming audio stream in real-time, leading to either output discontinuities or a constantly growing backlog of unprocessed data. However, the observed RF strongly depends on the hardware it is measured on, so it is crucial to measure it on target hardware.

FLOPS In contrast to the RF, floating point operations per second (FLOPS) is a measure of the number of arithmetic operations required to process a signal of length 1 s (including additions, subtractions, multiplications, and divisions), which makes it less dependent on the used hardware. However, a certain dependence persists

due to the availability of different implementations for different devices (such as CPU-optimized vs. GPU-optimized implementations).

NUMBER OF TRAINABLE WEIGHTS An important reason for the impressive representation capacity of DNNs originates from the large number of trainable weights[2] they can have, especially compared to purely model-based algorithms. However, a model with more trainable weights will require more memory for storage and inference, which can be a limiting factor for small devices like hearing devices. Furthermore, the number of weights often correlates with the number of FLOPS. Therefore, a model with fewer trainable weights is generally preferred for resource-constrained applications, as it will likely have lower memory and processing requirements. It should be noted that the relation between the number of trainable weights and the number of computations strongly depends on the employed DNN architecture. For instance, CNNs typically re-use their weights often, leading to more computations per trainable weight than fully connected networks, which use each weight only once.

2 We include both trainable weights and biases under the term "trainable weights" to avoid confusion with the term "parameter" used in the context of model-based approaches.

3

SPEECH ENHANCEMENT ALGORITHMS

In this chapter, we review the speech enhancement algorithms used throughout this thesis, categorized into model-based, learning-based, and hybrid speech enhancement algorithms. Section 3.1.1 introduces two model-based optimal filters, namely the spatio-temporal MVDR filter and the spatio-temporal Wiener filter, which are designed to exploit both spatial and temporal correlations of the speech and noise components. As special cases of the spatio-temporal MVDR filter with either a single time frame or a single microphone, Section 3.1.2 introduces the temporal MVDR filter and the MVDR beamformer; and as a special case of the spatio-temporal Wiener filter with a single time frame, Section 3.1.3 introduces the binaural spatial Wiener filter. Section 3.2 reviews two learning-based speech enhancement algorithms, namely the deep filter algorithm and the Conv-TasNet algorithm, which serve as baseline speech enhancement algorithms in this thesis. Section 3.3 reviews a hybrid speech enhancement algorithm, namely the mask-based MVDR beamformer with attention-based spatial covariance matrix aggregator (ASA), which can adapt to dynamic acoustic scenarios.

3.1 Model-Based Speech Enhancement Algorithms

3.1.1 *Spatio-Temporal Filters*

By combining the benefits of spatial filtering (which exploits the spatial characteristics of the acoustic scenario, Section 1.1.3) and temporal filtering (which exploits the temporal correlation of the speech and noise components, Sections 1.1.1 and 1.1.2), spatio-temporal filters can achieve better speech enhancement performance compared to purely spatial or purely temporal filters. In this section, we introduce the spatio-temporal MVDR filter and the STWF, both of which operate on the multi-microphone multi-frame signal vector $\mathbf{y}_t$ defined in (2.29) to estimate the speech component at the reference microphone.

3.1.1.1 *Spatio-Temporal MVDR Filter*

The spatio-temporal MVDR filter minimizes the output interference PSD while preserving the spatio-temporally correlated speech component at the reference microphone with index r [3], [47]. The constrained optimization problem for the spatio-temporal filter vector $\mathbf{w}_t^r$ in (2.49) is given by

$$\mathbf{w}_t^{\mathrm{MVDR},r} = \underset{\mathbf{w}_t^r}{\operatorname{argmin}}\, \mathbb{E}\left(|\mathbf{w}_t^{r,\mathsf{H}}\mathbf{i}_{t,r}|^2\right) \text{ subject to } \mathbf{w}_t^{r,\mathsf{H}}\mathbf{x}_t = x_{t,r}. \tag{3.1}$$

With the noisy STCM in (2.46), the spatio-temporal MVDR filter [5], [24], [191] is given by

$$\boxed{\mathbf{w}_t^{\mathrm{MVDR},r} = \frac{\left(\mathbf{\Phi}_{i,t}^r\right)^{-1}\boldsymbol{\gamma}_t^r}{\boldsymbol{\gamma}_t^{r,\mathsf{H}}\left(\mathbf{\Phi}_{i,t}^r\right)^{-1}\boldsymbol{\gamma}_t^r}.} \tag{3.2}$$

Hence, to implement the spatio-temporal MVDR filter in (3.2), estimates of the inverse interference STCM $\left(\mathbf{\Phi}_{i,t}^r\right)^{-1}$ and the speech STCV $\boldsymbol{\gamma}_t^r$ are required.

3.1.1.2 *Spatio-Temporal Wiener Filter*

The spatio-temporal Wiener filter (STWF) minimizes the MSE between the output signal and the target speech component at the reference microphone with index r [3]. The optimization problem for the spatio-temporal filter vector $\mathbf{w}_t^{\mathrm{WF},r}$ in (2.49) is given by

$$\mathbf{w}_t^{\mathrm{WF},r} = \underset{\mathbf{w}_t^r}{\operatorname{argmin}}\, \mathbb{E}\left(\left\|x_{t,r} - \mathbf{w}_t^{r,\mathsf{H}}\mathbf{y}_t\right\|_2^2\right), \tag{3.3}$$

yielding

$$\boxed{\mathbf{w}_t^{\mathrm{WF},r} = \mathbf{\Phi}_{y,t}^{-1}\mathbf{\Phi}_{x,t}\mathbf{e}_r.} \tag{3.4}$$

Hence, to implement the STWF, estimates of the inverse noisy STCM $\mathbf{\Phi}_{y,t}^{-1}$ and the speech STCM $\mathbf{\Phi}_{x,t}$ are required.

Using (2.46) and the matrix inversion lemma, it can be easily shown that the STWF in (3.4) can be decomposed as a spatio-temporal MVDR filter and a real-valued scalar postfilter [5], [192], i.e.,

$$\boxed{\mathbf{w}_t^{\mathrm{WF},r} = \underbrace{\frac{\left(\mathbf{\Phi}_{i,t}^r\right)^{-1}\boldsymbol{\gamma}_t^r}{\boldsymbol{\gamma}_x^{r,\mathsf{H}}\left(\mathbf{\Phi}_{i,t}^r\right)^{-1}\boldsymbol{\gamma}_t^r}}_{\text{spatio-temporal MVDR}} \underbrace{\frac{\phi_{x,t}^r}{\phi_{x,t}^r + \boldsymbol{\gamma}_t^{r,\mathsf{H}}\left(\mathbf{\Phi}_{i,t}^r\right)^{-1}\boldsymbol{\gamma}_t^r}}_{\text{postfilter}},} \tag{3.5}$$

where the spatio-temporal MVDR filter minimizes the output interference PSD while preserving the spatio-temporal correlation of the speech component, while the postfilter provides additional noise reduction at the cost of allowing for some speech distortion (see Section 3.1.1.3). This alternative formulation relies on estimates of the inverse interference STCM $\left(\mathbf{\Phi}_{i,t}^r\right)^{-1}$, the speech STCV $\boldsymbol{\gamma}_t^r$, as well as the speech PSD $\phi_{x,t}^r$.

3.1.1.3 *Comparison of Spatio-Temporal MVDR and Wiener Filters*

One major advantage of model-based speech enhancement approaches over learning-based approaches is that their performance can be analytically compared. In this section, we compare the spatio-temporal MVDR filter in Section 3.1.1.1 and the STWF in Section 3.1.1.2, in terms of their signal-to-interference ratio (SIR) improvement and speech distortion index [3].

SIR IMPROVEMENT The subband input SIR at the reference microphone with index r is defined as

$$\mathrm{iSIR}_{f,t}^{r} = \frac{\phi_{x,f,t}^{r}}{\phi_{i,f,t}^{r}}, \tag{3.6}$$

where $\phi_{i,f,t}^{r} = \mathbb{E}(|i_{t,f,r}|^2)$ denotes the interference PSD at the r-th microphone. Similarly, the fullband input SIR is defined as

$$\mathrm{iSIR}_{t}^{r} = \frac{\sum_{f=1}^{F} \phi_{x,f,t}^{r}}{\sum_{f=1}^{F} \phi_{i,f,t}^{r}}. \tag{3.7}$$

The subband output SIR at the reference microphone with index r is defined as the ratio of the PSDs of the filtered speech and interference components (with some filter $\mathbf{w}_{f,t}^{r}$), i.e.,

$$\mathrm{oSIR}_{f,t}^{r}(\mathbf{w}_{f,t}^{r}) = \frac{\mathcal{E}\left(|\mathbf{w}_{f,t}^{r,\mathsf{H}}\mathbf{x}_{f,t}|^2\right)}{\mathcal{E}\left(|\mathbf{w}_{f,t}^{r,\mathsf{H}}\mathbf{i}_{f,t}|^2\right)} \tag{3.8}$$

$$= \frac{\phi_{x,f,t}^{r}|\mathbf{w}_{f,t}^{r,\mathsf{H}}\check{\mathbf{h}}_{f,t}^{r}|^2}{\mathbf{w}_{f,t}^{r,\mathsf{H}}\mathbf{\Phi}_{i,f,t}\mathbf{w}_{f,t}^{r}}, \tag{3.9}$$

and, similarly, the fullband output SIR is defined as

$$\mathrm{oSIR}_{t}^{r}\left(\left\{\mathbf{w}_{f,t}^{r}\right\}_{f=1}^{F}\right) = \frac{\sum_{f=1}^{F} \phi_{x,f,t}^{r}|\mathbf{w}_{f,t}^{r,\mathsf{H}}\check{\mathbf{h}}_{f,t}^{r}|^2}{\sum_{f=1}^{F} \mathbf{w}_{f,t}^{r,\mathsf{H}}\mathbf{\Phi}_{i,f,t}\mathbf{w}_{f,t}^{r}}. \tag{3.10}$$

For the spatio-temporal MVDR filter, the subband output SIR is given by

$$\mathrm{oSIR}_{f,t}^{r}(\mathbf{w}_{f,t}^{\mathrm{MVDR},r}) = \phi_{x,f,t}^{r}\boldsymbol{\gamma}_{f,t}^{r,\mathsf{H}}\left(\mathbf{\Phi}_{i,t}^{r}\right)^{-1}\boldsymbol{\gamma}_{f,t}^{r}, \tag{3.11}$$

which increases with the number of microphones and time frames and which can be shown to always be equal to or exceed the subband input SIR in (3.6) [3]. Furthermore, the output SIR of the spatio-temporal MVDR filter can be shown to be equal to the maximum output SIR that is achievable by a spatio-temporal filter $\bar{\mathbf{w}}_{f,t}^{r}$ that does not introduce speech distortion.

Similarly, the fullband output SIR of the spatio-temporal MVDR filter is equal to

$$\mathrm{oSIR}_t^r\left(\mathbf{w}_{f,t}^{\mathrm{MVDR},r}\right) = \frac{\sum_{f=1}^{F}\phi_{x,f,t}}{\sum_{f=1}^{F}\phi_{x,f,t}\left(\bar{\boldsymbol{\gamma}}_{f,t}^{\mathsf{H}}\left(\boldsymbol{\Phi}_{i,f,t}^r\right)^{-1}\bar{\boldsymbol{\gamma}}_{f,t}\right)^{-1}}, \tag{3.12}$$

which can be shown to always be equal to or exceed the fullband input SIR in (3.7).

For the STWF, the subband output SIR is given by

$$\begin{aligned}\mathrm{oSIR}_{f,t}^r\left(\mathbf{w}_{f,t}^{\mathrm{WF},r}\right) &= \phi_{x,f,t}^r\bar{\boldsymbol{\gamma}}_{f,t}^{r,\mathsf{H}}\left(\boldsymbol{\Phi}_{i,f,t}^r\right)^{-1}\boldsymbol{\gamma}_{f,t}^r &(3.13)\\ &= \mathrm{oSIR}_{f,t}^r\left(\mathbf{w}_{f,t}^{\mathrm{MVDR},r}\right), &(3.14)\end{aligned}$$

i.e., the subband output SIRs of the spatio-temporal MVDR filter and the STWF are equal. The fullband output SIR of the STWF is equal to

$$\mathrm{oSIR}_t^r\left(\mathbf{w}_{f,t}^{\mathrm{WF},r}\right) = \frac{\sum_{f=1}^{F}\phi_{x,f,t}^r\frac{\left(\boldsymbol{\gamma}_{f,t}^{r,\mathsf{H}}\left(\boldsymbol{\Phi}_{i,f,t}^r\right)^{-1}\boldsymbol{\gamma}_{f,t}^{r,\mathsf{H}}\right)^2}{\left(1+\boldsymbol{\gamma}_{f,t}^{r,\mathsf{H}}\left(\boldsymbol{\Phi}_{i,f,t}^r\right)^{-1}\boldsymbol{\gamma}_{f,t}^{r,\mathsf{H}}\right)^2}}{\sum_{f=1}^{F}\phi_{x,f,t}^r\frac{\boldsymbol{\gamma}_{f,t}^{r,\mathsf{H}}\left(\boldsymbol{\Phi}_{i,f,t}^r\right)^{-1}\boldsymbol{\gamma}_{f,t}^{r,\mathsf{H}}}{\left(1+\boldsymbol{\gamma}_{f,t}^{r,\mathsf{H}}\left(\boldsymbol{\Phi}_{i,f,t}^r\right)^{-1}\boldsymbol{\gamma}_{f,t}^{r,\mathsf{H}}\right)^2}}, \tag{3.15}$$

which can be shown to always be equal to or exceed the fullband output SIR of the spatio-temporal MVDR filter in (3.12).

SPEECH DISTORTION INDEX The subband speech distortion index (SDI) at the reference microphone with index r is defined as the ratio of the power of the difference between the filtered speech component and the target speech component as well as the speech PSD, i.e.,

$$\begin{aligned}\mathrm{SDI}_{f,t}^r(\check{\mathbf{w}}_{f,t}^r) &= \frac{\mathbb{E}\left(\left|\check{\mathbf{w}}_{f,t}^{r,\mathsf{H}}\check{\mathbf{x}}_{f,t} - x_{f,t,r}\right|^2\right)}{\phi_{x,f,t}^r} &(3.16)\\ &= \left|\check{\mathbf{w}}_{f,t}^{r,\mathsf{H}}\check{\mathbf{h}}_{f,t}^r - 1\right|^2, &(3.17)\end{aligned}$$

with lower values denoting less speech distortion.

Similarly, the fullband SDI is defined as

$$\mathrm{SDI}_t^r(\{\check{\mathbf{w}}_{f,t}^r\}_{f=1}^F) = \frac{\sum_{f=1}^{F}\mathbb{E}\left(\left|\check{\mathbf{w}}_{f,t}^{r,\mathsf{H}}\check{\mathbf{x}}_{f,t} - x_{f,t,r}\right|^2\right)}{\sum_{f=1}^{F}\phi_{x,f,t}^r}. \tag{3.18}$$

Plugging (3.2) in (3.16) yields the subband SDI of the spatio-temporal MVDR filter

$$\mathrm{SDI}_{f,t}^r(\mathbf{w}_{f,t}^{\mathrm{MVDR},r}) = 0, \tag{3.19}$$

meaning that the spatio-temporal MVDR filter does not distort the target speech signal (due to the distortionless response constraint $\mathbf{w}_{f,t}^{r,\mathsf{H}}\mathbf{x}_{f,t} = x_{f,t,r}$ in (3.1)).

Consequently, also the fullband SDI of the spatio-temporal MVDR filter is equal to 0.

Plugging (3.4) in (3.16) yields the subband SDI of the STWF

$$\mathrm{SDI}_{f,t}^{r}(\mathbf{w}_{f,t}^{\mathrm{WF},r}) = \frac{1}{\left(1+\boldsymbol{\gamma}_{f,t}^{r,\mathsf{H}}\left(\boldsymbol{\Phi}_{i,f,t}^{r}\right)^{-1}\boldsymbol{\gamma}_{f,t}^{r}\right)^{2}} > 0, \tag{3.20}$$

i.e., the STWF introduces speech distortion, which however decreases with an increasing number of microphones and time frames. The fullband output SDI of the STWF is equal to

$$\mathrm{SDI}_{f,t}^{r}(\mathbf{w}_{f,t}^{\mathrm{WF},r}) = \frac{\sum_{f=1}^{F}\phi_{x,f,t}^{r}\left(1+\boldsymbol{\gamma}_{f,t}^{r,\mathsf{H}}\left(\boldsymbol{\Phi}_{i,f,t}^{r}\right)^{-1}\boldsymbol{\gamma}_{f,t}^{r}\right)^{-2}}{\sum_{f=1}^{F}\phi_{x,f,t}^{r}}. \tag{3.21}$$

In summary, both the spatio-temporal MVDR filter and the STWF yield the same subband SIR improvement, with more microphones and time frames generally yielding a better performance. Comparing the STWF and the spatio-temporal MVDR filter, the STWF provides additional fullband SIRs improvement at the cost of introducing speech distortion.

3.1.2 *MVDR Beamformer and Temporal MVDR Filter*

The MVDR beamformer and the temporal MVDR filter can be derived as special cases of the spatio-temporal MVDR filter in Section 3.1.1.1, addressing either the spatial estimation problem in (2.47) (i.e., setting $N = 1$) or the temporal estimation problem in (2.48) (i.e., setting $M = 1$).

3.1.2.1 *MVDR Beamformer*

As mentioned before, spatial filters often neglect the spatially uncorrelated speech vector in (2.7), opting instead to use the simplified speech vector in (2.6), such that the interference component consists of only the noise component. The MVDR beamformer minimizes the output noise PSD while preserving the target speech component at the reference microphone with index r [5], [24], [191]. The constrained optimization problem for the spatial filter vector $\check{\mathbf{w}}_t^r$ in (2.47) is given by

$$\check{\mathbf{w}}_t^{\mathrm{MVDR},r} = \underset{\check{\mathbf{w}}_t^r}{\operatorname{argmin}}\, \mathbb{E}\left(|\check{\mathbf{w}}_t^{r,\mathsf{H}}\check{\mathbf{n}}_t|^2\right) \text{ subject to } \check{\mathbf{w}}_t^{r,\mathsf{H}}\check{\mathbf{x}}_t = x_{t,r}. \tag{3.22}$$

With the noisy SCM in (2.16), the MVDR beamformer [5], [24], [191] is given by

$$\boxed{\check{\mathbf{w}}_t^{\mathrm{MVDR},r} = \frac{\check{\boldsymbol{\Phi}}_{n,t}^{-1}\check{\mathbf{h}}_t^r}{\check{\mathbf{h}}_t^{r,\mathsf{H}}\check{\boldsymbol{\Phi}}_{i,t}^{-1}\check{\mathbf{h}}_t^r}.} \tag{3.23}$$

Hence, to implement the MVDR beamformer, estimates of the inverse noise SCM $\check{\mathbf{\Phi}}_{n,t}^{-1}$ and the RTF vector $\check{\mathbf{h}}_t^r$ of the target speech component are required.

Assuming that the speech covariance matrix is of rank 1 as in (2.13), the MVDR beamformer can be written as [3], [193]

$$\boxed{\check{\mathbf{w}}_t^{\text{MVDR},r} = \frac{\check{\mathbf{\Phi}}_{n,t}^{-1}\check{\mathbf{\Phi}}_{x,t}\check{\mathbf{e}}_r}{\text{trace}\left(\check{\mathbf{\Phi}}_{n,t}^{-1}\check{\mathbf{\Phi}}_{x,t}\right)}.} \tag{3.24}$$

This alternative formulation, often used in mask-based MVDR beamformers [163], replaces the need to estimate the RTF $\check{\mathbf{h}}_t^r$ with the need to estimate the speech SCM $\check{\mathbf{\Phi}}_{x,t}$ (which can be seen as an estimation problem of similar difficulty, since $\check{\mathbf{\Phi}}_{x,t}$ depends on both $\check{\mathbf{h}}_t^r$ and $\phi_{x,t}^r$, with the scaling introduced by $\phi_{x,t}^r$ canceling out in (3.24)).

3.1.2.2 *Temporal MVDR Filter*

As mentioned in Section 1.1.1, speech and noise signals exhibits distinct spectro-temporal patterns that can be exploited by multi-frame approaches, resulting in the potential to reduce noise without introducing any speech distortion. The temporal MVDR filter minimizes the output interference PSD while preserving the target speech component at the reference microphone with index r [3], [45]. The constrained optimization problem for the temporal filter vector $\bar{\mathbf{w}}_t$ in (2.48) is given by

$$\bar{\mathbf{w}}_t^{\text{MVDR},r} = \underset{\bar{\mathbf{w}}_t}{\text{argmin}}\, \mathbb{E}\left(|\bar{\mathbf{w}}_t^{\mathsf{H}}\bar{\mathbf{i}}_{t,r}^r|^2\right) \text{ subject to } \bar{\mathbf{w}}_t^{\mathsf{H}}\bar{\mathbf{x}}_{t,r} = x_{t,r}. \tag{3.25}$$

With the noisy TCM in (2.28), the temporal MVDR filter is given by

$$\boxed{\bar{\mathbf{w}}_t^{\text{MVDR},r} = \frac{(\bar{\mathbf{\Phi}}_{i,t,r}^r)^{-1}\bar{\boldsymbol{\gamma}}_{t,r}^r}{\bar{\boldsymbol{\gamma}}_{t,r}^{r,\mathsf{H}}(\bar{\mathbf{\Phi}}_{i,t,r}^r)^{-1}\bar{\boldsymbol{\gamma}}_{t,r}^r}.} \tag{3.26}$$

Hence, to implement the temporal MVDR filter in (3.26), estimates of the inverse interference TCM $(\bar{\mathbf{\Phi}}_{i,t,r}^r)^{-1}$ and the speech TCV $\bar{\boldsymbol{\gamma}}_{t,r}^r$ are required.

Since the interference TCM includes both the noise TCM and the (highly time-varying) uncorrelated speech TCM, its accurate estimation is quite challenging. To address this issue, the interference TCM has often been replaced either by the noisy TCM, leading to the multi-frame minimum power distortionless response filter [45], [53], [194], [195] (which is very sensitive to estimation errors in the speech TCV [54]), or by the noise TCM, thereby however neglecting the uncorrelated speech vector. In this simplified case, the constrained optimization problem for the temporal filter vector $\bar{\mathbf{w}}_t$ in (2.48) given by

$$\bar{\mathbf{w}}_t^{\text{MVDR},r} = \underset{\bar{\mathbf{w}}_t}{\text{argmin}}\, \mathbb{E}\left(|\bar{\mathbf{w}}_t^{\mathsf{H}}\bar{\mathbf{n}}_{t,r}|^2\right) \text{ subject to } \bar{\mathbf{w}}_t^{\mathsf{H}}\bar{\mathbf{x}}_{t,r} = x_{t,r} \tag{3.27}$$

such that the simplified temporal MVDR filter is given by

$$\boxed{\bar{\mathbf{w}}_t^{\mathrm{MVDR},r} = \frac{(\bar{\bar{\mathbf{\Phi}}}_{n,t})^{-1}\bar{\boldsymbol{\gamma}}_{t,r}^{r}}{\bar{\boldsymbol{\gamma}}_{t,r}^{r,\mathsf{H}}(\bar{\bar{\mathbf{\Phi}}}_{n,t})^{-1}\bar{\boldsymbol{\gamma}}_{t,r}^{r}}.} \tag{3.28}$$

3.1.2.3 *Comparison of Spatial and Temporal MVDR Filters*

While the MVDR beamformer in (3.23) and the temporal MVDR filter in (3.26) share a mathematical resemblance, their corresponding quantities differ significantly. The temporal MVDR filter can be understood as treating time frames as virtual microphones. However, this analogy does not recognize the fundamental difference between the quantities that need to be estimated.

First, the MVDR beamformer relies on estimating the RTF vector, $\check{\mathbf{h}}_t^r$, which depends on spatial characteristics of the acoustic scenario, including the source positions, the microphone array geometry, and propagation effects. Importantly, the RTF is often relatively stationary over short time intervals, particularly in spatially stationary acoustic scenarios. In contrast, the temporal MVDR filter requires the estimation of the speech TCV $\bar{\boldsymbol{\gamma}}_{t,r}^{r}$, which depends on the correlation of the speech component across consecutive time frames. Unlike the RTF vector, the speech TCV is highly time-varying, which makes accurate estimation of the speech TCV considerably more challenging.

Second, the correlation between the speech STFT coefficients at different microphones is typically strong for coherent sources such as speech, while the correlation between the speech STFT coefficients in consecutive time frames can vary drastically and even be relatively weak for some speech sounds such as unvoiced speech. This difference affects the impact of the spatially uncorrelated speech component in (2.7) and the temporally uncorrelated speech component in (2.18) on speech enhancement performance. In the MVDR beamformer, the spatially uncorrelated speech component (representing, for instance, late reverberation) is often neglected. While this simplification can introduce some error, it is frequently a reasonable approximation, especially when the direct path and early reflections are dominant compared to late reflections [5], [24], and because late reflections do not substantially contribute to speech intelligibility [22]. In contrast, in the temporal MVDR filter, the temporally uncorrelated speech component is highly non-stationary and can be considerably large compared to the temporally correlated speech component (strongly depending on the STFT frame shift).

These differences result in varying sensitivities to estimation errors between the spatial and temporal MVDR filters. While RTF estimation errors do degrade speech enhancement performance of the MVDR beamformer, the relative stationarity of the RTF vector allows for an effective application of mitigation procedures such as temporal averaging or recursive smoothing. In contrast, the temporal MVDR filter is much more susceptible to estimation errors due to the high time-variance of the speech TCV and the interference TCM–despite its potential to achieve large

noise reduction performance with little speech distortion. Such estimation errors can manifest themselves as reduced speech enhancement performance and increased speech distortion (particularly time-varying distortion called musical noise, which is often even perceived as more disturbing than the original noise) [54].

3.1.3 *Binaural Spatial Wiener Filter*

In [56], [59] the binaural spatial Wiener filter was proposed, which minimizes the MSE between the binaural output signals and the target speech components at the left and right hearing device reference microphones with indices L and R. The optimization problem for the binaural spatial filter vectors $\check{\mathbf{w}}_t^{\mathrm{WF},L}$ and $\check{\mathbf{w}}_t^{\mathrm{WF},R}$ in (2.50) is given by

$$J(\check{\mathbf{w}}_t^L, \check{\mathbf{w}}_t^R) = \mathcal{E}\left\{ \left\| \begin{matrix} x_{t,L} - \check{\mathbf{w}}_t^{L,\mathsf{H}} \check{\mathbf{y}}_t \\ x_{t,R} - \check{\mathbf{w}}_t^{R,\mathsf{H}} \check{\mathbf{y}}_t \end{matrix} \right\|_2^2 \right\}, \tag{3.29}$$

yielding

$$\check{\mathbf{w}}_t^{\mathrm{WF},\nu} = \check{\mathbf{\Phi}}_{y,t}^{-1} \check{\mathbf{\Phi}}_{x,t} \check{\mathbf{e}}_\nu, \quad \nu \in \{L, R\}, \tag{3.30}$$

Hence, to implement the binaural spatial Wiener filter, estimates of the inverse noisy SCM $\check{\mathbf{\Phi}}_{y,t}^{-1}$ and the speech SCM $\check{\mathbf{\Phi}}_{x,t}$ are required. Using (2.16) and the matrix inversion lemma, it can be easily shown that the binaural spatial Wiener filter in (3.30) can be decomposed as an MVDR beamformer and a real-valued scalar postfilter [5], [192], i.e.,

$$\boxed{\check{\mathbf{w}}_t^{\mathrm{WF},\nu} = \underbrace{\frac{\check{\mathbf{\Phi}}_{n,t}^{-1} \check{\mathbf{h}}_t^\nu}{\check{\mathbf{h}}_t^{\nu,\mathsf{H}} \check{\mathbf{\Phi}}_{n,t}^{-1} \check{\mathbf{h}}_t^\nu}}_{\text{MVDR beamformer}} \underbrace{\frac{\phi_{x,t}^\nu}{\phi_{x,t}^\nu + \check{\mathbf{h}}_t^{\nu,\mathsf{H}} \check{\mathbf{\Phi}}_{n,t}^{-1} \check{\mathbf{h}}_t^\nu}}_{\text{postfilter}},} \tag{3.31}$$

where the MVDR beamformer minimizes the output noise PSD while preserving the spatial correlation of the speech component, while the postfilter provides additional noise reduction at the cost of allowing for some speech distortion. This alternative formulation relies on estimates of the inverse noise SCM $\check{\mathbf{\Phi}}_{n,t}^{-1}$, the RTF vectors $\check{\mathbf{h}}_t^L$ and $\check{\mathbf{h}}_t^R$, as well as the speech PSDs $\phi_{x,t}^L$ and $\phi_{x,t}^R$. This formulation is common in spatial Wiener filters [5], [192], because it allows the separate estimation of the MVDR beamformer in (3.23) (which depends on the often more slowly varying noise SCM and the RTF vector) and the spectro-temporal postfilter (which depends on the typically more rapidly varying speech PSD).

Assuming accurate estimates of the required quantities, it can be shown that the binaural spatial Wiener filter preserves the binaural cues of the target speech component but changes the binaural cues of the noise component to the cues of the target speech component [196].

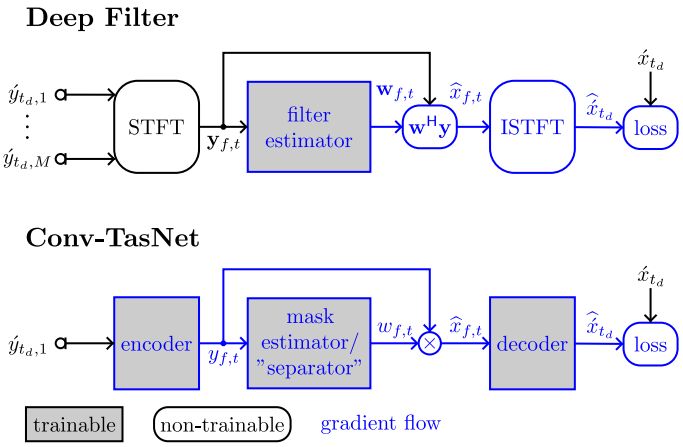


Figure 3.1: Overview of the deep filter (DF) algorithm [129] and the Conv-TasNet algorithm [89].

3.2 Learning-Based Speech Enhancement Algorithms

In the previous section, we have introduced purely model-based optimal filters for spatio-temporal filtering and the special cases of spatial filtering and temporal filtering, requiring the estimation of quantities related to speech and interference correlation. As mentioned in Section 1.2.2, purely learning-based approaches avoid the estimation of such quantities. Instead, they employ DNNs to estimate the filters or even the target speech signal directly. In this section, we review the DF algorithm and the Conv-TasNet algorithm as examples of purely learning-based algorithms (Fig. 3.1).

3.2.1 *Deep Filter*

In [129], the DF algorithm was proposed, which is a purely learning-based approach that employs a DNN to estimate complex-valued STFT-domain filters for speech enhancement (Fig. 3.1, top). Unlike learning-based speech enhancement algorithms that apply real-valued or complex-valued masks to individual noisy STFT coefficients, the DF algorithm estimates a set of complex-valued filter coefficients that are applied in a filter-and-sum procedure to a local region in the STFT domain, i.e., to multiple frequency bins and multiple time frames. To avoid having to define target filter coefficients, the DF algorithm is trained using a signal approximation loss function defined at the output of the filter. This approach allows the DF algorithm to address destructive interference, which is a usually overlooked issue that can occur if the speech and noise component cancel each other, which—in the case of complete destructive interference—would render masking-based approaches inef-

fective. Results in [129] showed a benefit of including multiple time frames in the filter but no considerable benefit of including multiple frequency bins.

In the context of this dissertation, we adopt the general idea of the DF algorithm as our default purely learning-based baseline. The flexibility of this approach allows it to be readily adapted from its original single-microphone configuration to the various scenarios investigated in this thesis. By simply changing the input signal vector and the output of the filter estimator, we can replace the spectro-temporal filter proposed in [129] with the spatial filter in (2.47), the temporal filter in (2.48), the (monaural) spatio-temporal filter in (2.49), or the binaural spatio-temporal filters in (2.50). This adaptability enables fair comparisons between purely learning-based approaches and the hybrid approaches investigated throughout this thesis, as we can maintain consistent training data, input features, DNN architectures, loss functions, and training procedures across all algorithms. By controlling for these factors as well as computational complexity considerations (such as number of parameters and operations), we can draw conclusions about the differences between purely learning-based versus hybrid approaches to speech enhancement.

3.2.2 *Conv-TasNet*

While the motivation for including the DF algorithm introduced in the previous section was to include a purely learning-based algorithm that allows deriving conclusions about the model-based component in hybrid approaches investigated in this thesis, the motivation for including the Conv-TasNet algorithm proposed in [89] is to provide a broader perspective on state-of-the-art purely learning-based approaches, as well as to introduce the DNN architecture that is used for most of the learning-based modules throughout this thesis. In contrast to many model-based speech enhancement algorithms, Conv-TasNet learns its transformation directly from data rather than relying on a predefined transformation such as the STFT. Among purely learning-based approaches, Conv-TasNet has emerged as one of the most popular (though already slightly outdated) algorithms due to its strong single-microphone speech enhancement performance and public availability, and it has been studied extensively [15], [65], [90], [105], [128], [131], [174], [188], [197]–[199] as well as being the basis of further improvements [78], [115], [133], [165], [172], [173], [189], [200]–[202]. The use of a learned transformation circumvents the phase estimation problem—which has plagued early STFT-domain learning-based algorithms—while also enabling low-latency processing. Given its strong performance and widespread use, Conv-TasNet serves as a key baseline algorithm in this work, representing the class of purely learning-based approaches that do not rely on predefined signal representations.

3.2.2.1 *Overview*

Conv-TasNet is a fully convolutional, end-to-end single-microphone speech enhancement algorithm that operates in the time domain. It comprises three main processing

stages, i.e., an encoder, a separator, and a decoder (Fig. 3.1, bottom). Note that notation throughout this section may slightly differ from the notation in the rest of this thesis. First, the encoder transforms the time-domain noisy microphone signal into a high-dimensional representation with feature index $f \in \{1, \ldots, F\}$ using a 1-dimensional convolutional layer. This learned transformation serves as a replacement for the STFT and is optimized jointly with the rest of the network. Second, the separator estimates multiplicative masks $w_{f,t}$ that are applied element-wise to the representation of the noisy microphone signal in order to estimate the target speech component. Third, the decoder transforms the masked representations back to the time domain via a 1-dimensional transposed convolutional layer.

3.2.2.2 *Encoder and Decoder*

A key distinguishing feature of Conv-TasNet compared to many model-based speech enhancement algorithms, which apply a fixed STFT to represent the time-domain noisy microphone signals, is its learned transformation. More in particular, the encoder first segments the time-domain signal $\acute{y}_{t_d}$ into T overlapping frames $\acute{\mathbf{y}}_t \in \mathbb{R}^{N_d}$ of length N_d samples, with frame index t, and applies a rectangular analysis window. Each frame $\acute{\mathbf{y}}_t$ is then transformed into an F-dimensional representation using a 1-dimensional convolutional layer parameterized by a learned matrix $\mathbf{U} \in \mathbb{R}^{F \times N_d}$, where F denotes the number of learned basis functions, i.e.,

$$\mathbf{y}_t = \mathcal{A}(\mathbf{U}\acute{\mathbf{y}}_t), \tag{3.32}$$

where $\mathbf{y}_t \in \mathbb{R}^F$ denotes the encoded coefficients for frame t, and $\mathcal{A}(\cdot)$ is an optional nonlinear activation function. In contrast to the STFT, where the transformation is predefined using sinusoidal basis functions, the basis functions in $\mathbf{U}$ are optimized jointly with the rest of the network. Typically, this learned representation is designed to be overcomplete by choosing a large value for F, meaning that the number of basis functions F exceeds the frame length N_d. Empirical analysis of the learned basis functions has shown that these functions tend to result in a higher resolution at lower frequencies [89], [90], where a lot of energy of the speech signal resides (Section 1.1.1), versus the uniform frequency resolution of the STFT. It should be emphasized that Conv-TasNet inherently preserves phase because its learned representation is derived directly from the time-domain waveform, without decomposing the signal into separate magnitude and phase components as the STFT does.

The masked encoded representation $\widehat{\mathbf{x}}_t$ in (3.46) is mapped back to the time domain using a 1-dimensional transposed convolutional layer parameterized by a learned matrix $\mathbf{V} \in \mathbb{R}^{N_d \times F}$, i.e.,[1]

$$\widehat{\acute{\mathbf{x}}}_t = \mathbf{V}\widehat{\mathbf{x}}_t, \tag{3.33}$$

1 Note that, for simplicity, we omitted stride and padding as the mechanism to combine multiple frames.

where $\widehat{\mathbf{x}}_t$ denotes the t-th estimated frame of the speech component in the time domain. Finally, the estimated speech signal is reconstructed using OLA processing.

Unlike the STFT, where the fixed analysis and synthesis transformations are constrained to enable perfect reconstruction, the analysis and synthesis transformations in Conv-TasNet can be independently learned. In the anechoic environment evaluated in [89], independent analysis and synthesis transformations resulted in a better speech enhancement performance than choosing the synthesis transformation as the pseudoinverse of the analysis transformation. However, [90] highlighted robustness issues in reverberant environments. To address this issue, [202] imposed an analytic constraint that couples the analysis and synthesis transformations based on the Hilbert transform, outperforming the unconstrained variant and even the STFT in moderately reverberant conditions.

The frame length $N_d \triangleq 1\,\text{ms}$ used in Conv-TasNet is typically much shorter than in typical STFT configurations, enabling straightforward low-latency processing—although, as mentioned before, approaches to achieve low latency with STFT-domain algorithms such as the use of temporal filters, asymmetric analysis and synthesis windows, or the use of predicted future STFT time frames [15], [16], also exist.

For Conv-TasNet, both a causal and a non-causal version exist, which differ in how the temporal convolution is applied. In this thesis, we only consider the causal version, as it is suitable for real-time applications and enables a fairer comparison to the (also causal) algorithms proposed in this thesis.

3.2.2.3 *Temporal Convolutional Network (TCN)-Based Separator*

OVERVIEW The separator takes the encoded representation of the noisy microphone signal $\mathbf{y}_t \in \mathbb{R}^F$ as input and estimates the multiplicative mask $\mathbf{W} \in \mathbb{R}^{F \times T}$. A detailed overview of the separator is provided in Fig. 3.2. Conv-TasNet [89] employs the TCN architecture for the separator, allowing for more efficient parallel processing, improved gradient flow, and reduced computational complexity compared with the original, LSTM-based TasNet [88]. More in particular, the TCN in Conv-TasNet comprises stacked 1-dimensional dilated temporal convolutional (1D-Conv) blocks, designed to capture long-range temporal context [98]. Each 1D-Conv block employs depthwise separable convolutions, which reduce the number of trainable weights while retaining most of the representation capacity.

INPUT NORMALIZATION AND BOTTLENECK Each input frame $\mathbf{y}_t \in \mathbb{R}^F$ is first normalized using conventional layer normalization, i.e.,

$$\mathbf{y}_t \leftarrow \frac{\mathbf{y}_t - \mu_{y,t}}{\sqrt{\sigma_{y,t}^2 + \epsilon}}, \tag{3.34}$$

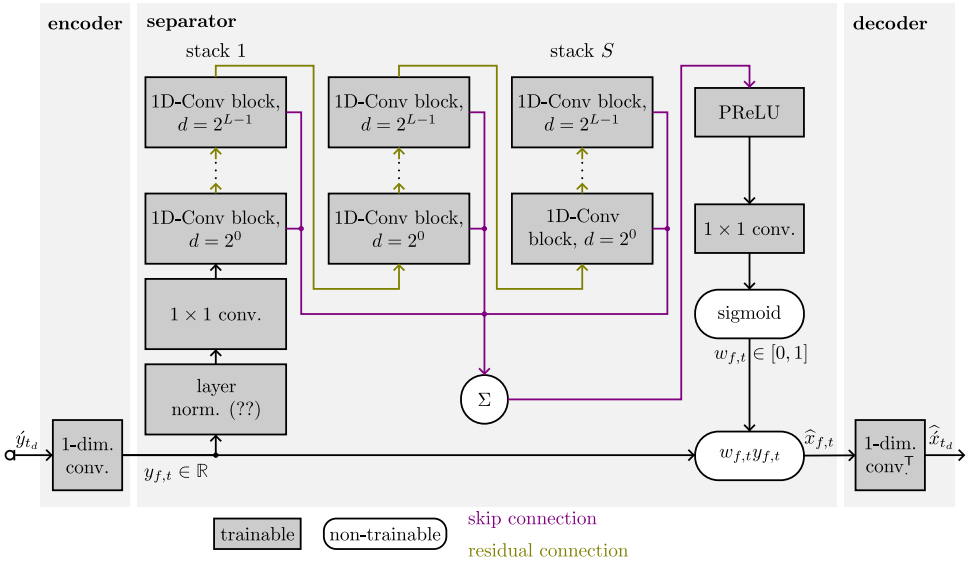


Figure 3.2: Detailed overview of the Conv-TasNet algorithm. Adapted from [89].

where ϵ is a small constant to avoid division by zero and $\mu_{y,t} \in \mathbb{R}$ as well as $\sigma^2_{y,t} \in \mathbb{R}$ denote the mean and variance computed along the feature dimension as

$$\mu_{y,t} = \frac{1}{F} \sum_{f=1}^{F} y_{f,t}, \tag{3.35}$$

$$\sigma^2_{y,t} = \frac{1}{F} \sum_{f=1}^{F} (y_{f,t} - \mu_{y,t})^2. \tag{3.36}$$

A pointwise convolution is then applied to reduce the number of features from F to F^{BN} (resulting in a so-called bottleneck (BN)), i.e.,

$$\mathbf{y}_t^{\mathrm{BN}} = \mathbf{W}^{\mathrm{BN}} \mathbf{y}_t \in \mathbb{R}^{F^{\mathrm{BN}}}, \tag{3.37}$$

where $\mathbf{W}^{\mathrm{BN}} \in \mathbb{R}^{F^{\mathrm{BN}} \times F}$ is a learned weight matrix and $F^{\mathrm{BN}} < F$. Afterwards, $\mathbf{y}_t^{\mathrm{BN}}$ is passed through S stacks of L 1D-Conv blocks with increasing dilation factor, resulting in an exponentially increasing temporal receptive field within each stack.

DILATED TEMPORAL CONVOLUTIONAL (1D-CONV) BLOCKS In each 1D-Conv block, the input $\mathbf{z}_t^{\mathrm{BN}}$ is first transformed using a pointwise convolution to

increase the number of features from F^{BN} to F^{Conv} before being passed through a parametric rectified linear unit (PReLU) activation function, i.e.,[2]

$$\tilde{\mathbf{h}}^{1\times1} = \text{PReLU}\left(\mathbf{W}^{1\times1}\mathbf{z}_t^{\text{BN}}\right) \in \mathbb{R}^{F^{\text{Conv}}\times T}, \tag{3.38}$$

where $\mathbf{W}^{1\times1} \in \mathbb{R}^{F^{\text{Conv}}\times F^{\text{BN}}}$ is a learned weight matrix, followed by a cumulative layer normalization layer, which transforms each time frame t cumulatively based on the mean and variance of current and past time frames (thereby enabling causal processing), i.e.,

$$\mathbf{h}_t^{1\times1} = \frac{\tilde{\mathbf{h}}_t^{1\times1} - \boldsymbol{\mu}_{h,t}}{\sqrt{\boldsymbol{\sigma}_{h,t}^2 + \epsilon}} \odot \boldsymbol{\gamma} + \boldsymbol{\beta}, \tag{3.39}$$

where $\odot$ denotes the Hadamard (elementwise) product, the mean $\boldsymbol{\mu}_{h,t}$ and variance $\boldsymbol{\sigma}_{h,t}^2$ are computed as

$$\boldsymbol{\mu}_{h,t} = \frac{1}{t}\sum_{\tau=1}^{t} \tilde{\mathbf{h}}_\tau^{1\times1} \tag{3.40}$$

$$\boldsymbol{\sigma}_{h,t}^2 = \frac{1}{t}\sum_{\tau=1}^{t} \left(\tilde{\mathbf{h}}_\tau^{1\times1} - \boldsymbol{\mu}_{h,t}\right)^2, \tag{3.41}$$

and $\boldsymbol{\gamma} \in \mathbb{R}^{F^{\text{Conv}}}$ and $\boldsymbol{\beta} \in \mathbb{R}^{F^{\text{Conv}}}$ are learned parameters.

Afterwards, Conv-TasNet employs depthwise separable convolutions, which decomposes the conventional convolution operation into two separate steps, i.e., a depthwise convolution (which applies independent kernels to each feature dimension), followed by a pointwise convolution (which actually corresponds to a simple linear combination of features). More in particular, $\mathbf{h}_t^{1\times1}$ is first transformed for each frequency as

$$h_{f,t}^{\text{DConv}} = \left(\mathbf{h}_F^{1\times1} \circledast \mathbf{k}_f^{\text{DConv}}\right)_t \tag{3.42}$$

where $\mathbf{h}_F^{1\times1} \in \mathbb{R}^T$ and $\mathbf{k}_f^{\text{DConv}} \in \mathbb{R}^P$ denote the f-th input feature and its corresponding depthwise convolutional kernel of size P, respectively. The operator $\circledast$ denotes causal convolution, defined as

$$(\mathbf{h}_F^{1\times1} \circledast \mathbf{k}_f^{\text{DConv}})_t = \sum_{p=0}^{P-1} k_{f,p}^{\text{DConv}} h_{f,t-p}^{1\times1}, \quad \text{for } t \geq P-1, \tag{3.43}$$

where the summation only considers past and current values ($h_{f,t-p}$) to enable causal processing, and zero-padding is applied for the cases $t-p<0$ (i.e., accessing past values before the signal starts).

2 In the first 1D-Conv block, the input corresponds to $\mathbf{y}_t^{\text{BN}}$ from (3.37).

After the depthwise convolution, the intermediate feature $\mathbf{h}^{\text{DConv}} \in \mathbb{R}^{F^{\text{Conv}}}$ is passed through another cumulative layer normalization as in (3.39) and a PReLU activation function. Finally, it is linearly transformed using another pointwise convolution to result in the skip connection and residual connection outputs, i.e.,

$$\tilde{\mathbf{z}}_t^{\text{skip}} = \mathbf{W}^{\text{skip}} \mathbf{h}_t^{\text{DConv}} \in \mathbb{R}^{F^{\text{skip}}} \tag{3.44}$$

$$\tilde{\mathbf{z}}_t^{\text{res}} = \mathbf{W}^{\text{res}} \mathbf{h}_t^{\text{DConv}} \in \mathbb{R}^{F^{\text{res}}}, \tag{3.45}$$

where $\mathbf{W}^{\text{skip}} \in \mathbb{R}^{F^{\text{skip}} \times F^{\text{Conv}}}$ and $\mathbf{W}^{\text{res}} \in \mathbb{R}^{F^{\text{res}} \times F^{\text{Conv}}}$ are the pointwise convolutional weight matrices producing the output for the skip and residual connections, respectively, aggregating information across input features. The decomposition of standard convolution into depthwise and pointwise convolutions reduces the number of trainable weights by a factor of ${}^{F^{\text{Conv}}P}/_{F^{\text{Conv}}+P}$, which is approximately equal to P if $F^{\text{Conv}} \gg P$ [89].

The residual output $\tilde{\mathbf{z}}_t^{\text{skip}}$ is used as the input to the next 1D-Conv block. The skip connection outputs from all 1D-Conv blocks are summed, passed through a PReLU activation function, transformed using a final pointwise convolution to increase the number of features from F^{BN} to F, and passed through a sigmoid activation function, yielding the final mask estimate $w_{f,t} \in [0, 1]$. The skip connection improves gradient flow particularly for deep TCN configurations, since—during backpropagation—there is a direct path from the estimated mask to each 1D-Conv block. Finally, $w_{f,t}$ is applied to the encoded representation of the noisy microphone signal $y_{f,t}$ as

$$\widehat{x}_{f,t} = w_{f,t} y_{f,t}. \tag{3.46}$$

DILATED CONVOLUTIONS To effectively model long-term speech dependencies, the TCN employs dilated depthwise separable convolutions in (3.42), each with a kernel size of P that is applied to each feature dimension independently. The dilation factor d increases exponentially across layers indexed by l as

$$d = 2^l, \quad l \in \{0, 1, \ldots, L-1\}. \tag{3.47}$$

Dilated convolutions introduce gaps (or "dilations") between the input frames, effectively skipping d input frames between each kernel element. This allows the convolution to capture a larger temporal context without increasing the number of trainable weights. Given a kernel size P and a layer depth L, the receptive field of the TCN is roughly equal to $P(2^L - 1)$ [98].

3.3 Hybrid Speech Enhancement Algorithm

Hybrid approaches aim at combining the interpretability of model-based approaches with the strong representation capacity of learning-based approaches. As mentioned before, we define hybrid approaches as those that use a model-based enhancement stage, and where the estimation of quantities is performed both during training (for

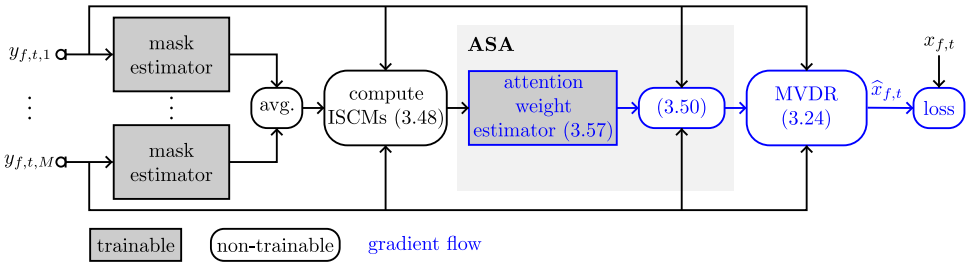


Figure 3.3: Overview of mask-based MVDR beamformer with attention-based spatial covariance matrix aggregator (ASA). The mask estimators are applied per microphone and share trainable weights, which are frozen during the training of the ASA.

the DNN) and during inference (for the model-based enhancement stage). In this section, we review the mask-based MVDR beamformer with attention-based spatial covariance matrix aggregator (ASA) [172] as an example of a coupled, structured estimation hybrid approach.

3.3.1 *Mask-Based Beamformer with ASA*

The mask-based beamformer with ASA considers the MVDR beamformer in (3.24), which relies on estimates of the speech and inverse noise SCMs $\check{\mathbf{\Phi}}_{x,t}$ and $\check{\mathbf{\Phi}}_{n,t}^{-1}$. As depicted in Fig. 3.3, to obtain these estimates, the algorithm employs a DNN to estimate time-frequency speech and noise masks for each microphone, which are then averaged across microphones and used to compute instantaneous estimates of the speech and noise SCMs. Instead of using a heuristic approach to temporally aggregate these instantaneous estimates, such as temporal averaging or recursive smoothing, the algorithm employs a attention-based spatial covariance matrix aggregator (ASA).

3.3.1.1 *Estimation of Spatial Covariance Matrices*

Mask-based beamformers rely on estimates of speech and noise masks, which are typically obtained using a spatial clustering-based approach [203] or a DNN [116], [162], [163], [172]–[175], [177], [204]–[206]. Assuming that the acoustic scenario is static and that estimates of the speech and noise masks are available, the speech and noise SCMs can be estimated using temporal averaging [163], [203], [204], i.e.,

$$\widehat{\check{\mathbf{\Phi}}}_{\nu} = \frac{1}{\sum_{\tau'=1}^{T} m_{\nu,\tau'}^{\mathbb{R}}} \sum_{\tau=1}^{T} \underbrace{m_{\nu,\tau}^{\mathbb{R}} \bar{\mathbf{y}}_{\tau} \bar{\mathbf{y}}_{\tau}^{\mathsf{H}}}_{=:\widehat{\check{\mathbf{\Phi}}}_{\nu,\tau}}, \tag{3.48}$$

where $\nu \in \{x, n\}$ indicates the speech or noise component, $m_{\nu,t}^{\mathbb{R}} \in [0, 1]$ denotes the real-valued speech or noise mask, and $\widehat{\check{\mathbf{\Phi}}}_{\nu,t}$ denotes the instantaneous SCM (ISCM)

estimate. The speech and noise masks are typically obtained by applying a DNN-based mask estimator independently for each channel, followed by averaging across microphones, i.e., $m^{\mathbb{R}}_{\nu,t} = \frac{1}{M}\sum_{m=1}^{M} m^{\mathbb{R}}_{\nu,t,m}$. Since $\widehat{\tilde{\mathbf{\Phi}}}_{\nu}$ in (3.48) is time-independent, the MVDR beamformer coefficients in (3.24) will also be time-independent, hence limiting speech enhancement performance in acoustic scenarios with moving sources.

To deal with this issue, heuristic approaches can be applied to result in time-varying SCM estimates. A popular approach is to apply recursive smoothing [175], [203], i.e.,

$$\widehat{\tilde{\mathbf{\Phi}}}_{\nu,t} = \sum_{\tau=1}^{T} \lambda_{\nu}^{t-\tau} \widehat{\tilde{\mathbf{\Psi}}}_{\nu,\tau}, \quad \nu \in \{x, n\}, \tag{3.49}$$

where $\lambda_{\nu} \in [0, 1]$ denotes the (time- and frequency-invariant) smoothing factor for the speech or noise SCM. Recursive smoothing applies exponentially decreasing weight to past estimates, allowing the SCM to adapt to changes in the acoustic scenario. However, the choice of the smoothing factor λ_{ν} can significantly affect the performance of the algorithm.

The temporal averaging and recursive smoothing approaches in (3.48) and (3.49) can be generalized as

$$\boxed{\widehat{\tilde{\mathbf{\Phi}}}_{\nu,t} = \sum_{\tau=1}^{T} a_{\nu,t,\tau} \widehat{\tilde{\mathbf{\Psi}}}_{\nu,\tau},} \tag{3.50}$$

where the (frequency-independent) attention weights $a_{\nu,t,\tau}$ control how the ISCM estimates at time frames $\tau \in \{1, \ldots, T\}$ are temporally aggregated to yield estimates of the speech and noise SCMs at time frame t. In other words, the attention weights $a_{\nu,t,\tau}$ determine the influence of frame τ on the estimate of the speech or noise SCM at time frame t. For each time frame t, the attention weights can be collected as

$$\mathbf{a}_{\nu,t} = \begin{bmatrix} a_{\nu,t,\tau=1} & \ldots & a_{\nu,t,\tau=T} \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^{T}. \tag{3.51}$$

Using (3.50), temporal averaging in (3.48) can be expressed by setting the attention weights as

$$a_{\nu,t,\tau} = \frac{1}{\sum_{\tau'=1}^{T} m^{\mathbb{R}}_{\nu,\tau'}} \quad \nu \in \{x, n\}, \tag{3.52}$$

and recursive smoothing in (3.49) can be expressed by setting the attention weights as

$$a_{\nu,t,\tau} = \begin{cases} \lambda_{\nu}^{t-\tau} & \text{if } \tau \leq t, \\ 0 & \text{otherwise} \end{cases}, \quad \nu \in \{x, n\}. \tag{3.53}$$

In the mask-based beamformer with ASA [172], instead of manually designing the attention weights as in temporal averaging in (3.52) or recursive smoothing in (3.53), the attention weights are estimated using a self-attention mechanism as explained in the following section, which allows for a more flexible temporal aggregation of the ISCM estimates.

As noted in [172], the generalized temporal aggregation formulation in (3.50) has a similar structure as the self-attention module in transformers [102]. Given a sequence of T query vectors $\mathbf{Q} = \{\mathbf{q}_1, \ldots, \mathbf{q}_T\} \in \mathbb{R}^{T \times F^{\mathrm{KQ}}}$, a sequence of key vectors $\mathbf{K} = \{\mathbf{k}_1, \ldots, \mathbf{k}_T\} \in \mathbb{R}^{T \times F^{\mathrm{KQ}}}$, and a sequence of value vectors $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_T\} \in \mathbb{R}^{T \times F^{\mathrm{V}}}$, the output of the self-attention module is computed as

$$\mathbf{A} = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathsf{T}}}{\sqrt{F^{\mathrm{KQ}}}}\right)\mathbf{V} \in \mathbb{R}^{T \times T} \tag{3.54}$$

$$\mathbf{Z} = \mathbf{A}\mathbf{V} \in \mathbf{R}^{T \times F^{\mathrm{V}}}, \tag{3.55}$$

where $\mathbf{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_T\}$ denotes the sequence of attention weights in (3.51), $\mathrm{softmax}(\cdot)$ denotes the softmax activation function applied per row, and $\mathbf{Z} \in \mathbb{R}^{T \times F^{\mathrm{V}}}$ denotes the output of the self-attention module. Hence, the output of the self-attention module in (3.55) at frame t can be written as

$$\mathbf{z}_t = \sum_{\tau=1}^{T} a_{t,\tau}\mathbf{v}_\tau, \tag{3.56}$$

which is equivalent to (3.50) by choosing the temporally aggregated SCM estimate as the output ($\mathbf{z}_t = \widehat{\tilde{\mathbf{\Phi}}}_{\nu,t}$) and the ISCM estimate as the value ($\mathbf{v}_\tau = \widehat{\tilde{\mathbf{\Psi}}}_{\nu,\tau}$).

3.3.1.2 *Estimation of Attention Weights*

To obtain the speech and noise attention weights in (3.50), separate self-attention-based DNNs (more specifically, transformer encoders [102]) were employed, i.e.,

$$\mathbf{a}_{\nu,t} = \mathbf{f}_\nu\left(\{\boldsymbol{\chi}_{\nu,t}\}_{t=1}^{T}; \boldsymbol{\theta}_\nu\right), \tag{3.57}$$

where $\mathbf{f}_\nu$ denotes the DNN to estimate the speech or noise attention weights, $\boldsymbol{\chi}_{\nu,t}$ denotes the input feature, and $\boldsymbol{\theta}_\nu$ denotes the trainable weights of the DNN. The vectorized estimated ISCMs defined in (3.48), concatenated along the frequency dimension, were used as the speech and noise input features, i.e.,

$$\boldsymbol{\chi}_{\nu,f,t} = \left[\Re\left(\mathrm{vec}(\widehat{\tilde{\mathbf{\Psi}}}_{\nu,f,\tau})^{\mathsf{T}}\right) \quad \Im\left(\mathrm{vec}(\widehat{\tilde{\mathbf{\Psi}}}_{\nu,f,\tau})^{\mathsf{T}}\right)\right]^{\mathsf{T}} \in \mathbb{R}^{2M^2} \tag{3.58}$$

$$\boldsymbol{\chi}_{\nu,t} = \left[\boldsymbol{\chi}_{\nu,f=1,t}^{\mathsf{T}} \quad \ldots \quad \boldsymbol{\chi}_{\nu,f=F,t}^{\mathsf{T}}\right]^{\mathsf{T}} \in \mathbb{R}^{2FM^2}, \tag{3.59}$$

where $\Re(\cdot)$ and $\Im(\cdot)$ denote the real and imaginary part, respectively, and $\mathrm{vec}(\cdot)$ denotes the vectorization operator that concatenates the columns of a matrix into a single column vector.

To avoid having to define target attention weights, the DNNs were trained using a signal approximation loss function, defined on the estimated speech signal at the output of the MVDR beamformer (Fig. 3.3). Nonetheless, simulation results in [172]

showed that the estimated attention weights indeed correspond to some extent to expected behavior. For example, to estimate the noise SCM, the corresponding attention weights tended to be high during speech inactivity, where noise statistics can be estimated more reliably. To estimate the speech SCM, corresponding attention weights tended to be high around the current time frame index, giving little weight to temporally distant frames. Furthermore, an analysis of the beampatterns resulting from the MVDR beamformer coefficients demonstrated that the ASA enables to accurately track the moving target speaker. However, the utilized estimated ISCM input features highly depend on the microphone array configuration, making the algorithm specific for the configuration used during training. Consequently, integrating the ASA into the mask-based MVDR beamformer enables adaptation to dynamic scenarios but sacrifices the microphone array configuration independence of most mask-based beamformers.

3.4 Summary

In this chapter, we reviewed several state-of-the-art speech enhancement algorithms that are used throughout this thesis, categorized into model-based, learning-based, and hybrid approaches. As examples of model-based speech enhancement algorithms, we reviewed the spatio-temporal MVDR filter, which minimizes the output interference PSD while preserving the speech component, and the STWF, which minimizes the MSE between the output signal and the target speech component—optimal filters that form the basis of many model-based speech enhancement approaches. For both filters, performance generally improves for more microphones and time frames, with the STWF yielding a better noise reduction performance than the spatio-temporal MVDR filter at the cost of introducing speech distortion. We also discussed spatial and temporal MVDR filters as special cases of the spatio-temporal MVDR filter and highlighted the challenges of estimating the required quantities for the temporal MVDR filter due to the highly time-varying nature of speech. Further, we discussed the binaural spatial Wiener filter, which minimizes the MSE between the binaural output signals and the target speech components at the left and right hearing device reference microphones, and mentioned that it preserves the binaural cues of the target speech component but changes the binaural cues of the noise component to the cues of the target speech component. As examples of learning-based speech enhancement algorithms, we reviewed the DF algorithm and the Conv-TasNet algorithm. First, avoiding explicit quantity estimation, the DF algorithm employs a DNN to directly estimate complex-valued filters and is easily adaptable to various microphone configurations, including binaural configurations, making it a flexible baseline for comparisons with the hybrid approaches investigated in this thesis. Second, the Conv-TasNet algorithm utilizes a learned transformation instead of a fixed STFT and estimates real-valued masks that are applied in this learned transform-domain. We detailed the architecture of the Conv-TasNet, including its encoder, TCN-based separator, and decoder, and described important aspects of this architecture, such as depthwise-separable convolutions, dilated convolutions, and cumulative layer normalization. Although arguably a bit

outdated, the strong performance and widespread use of Conv-TasNet, as well as its use of the TCN architecture (which will also be utilized in most of our proposed algorithms) make it a valuable baseline algorithm, representing purely learning-based approaches that do not rely on predefined signal transforms. Finally, as an example of hybrid speech enhancement algorithms, we reviewed the mask-based MVDR beamformer with ASA, combining the interpretability of model-based beamforming with the representation capacity of learning-based quantity estimation. This algorithm employs a DNN to estimate time-frequency masks and uses a self-attention mechanism to temporally aggregate instantaneous SCM estimates, enabling adaptation to dynamic acoustic scenarios. While this algorithm demonstrates the potential of combining model-based and learning-based approaches, it also highlights a challenge of designing hybrid algorithms, i.e., losing microphone-array configuration independence due to the design of the learning-based quantity estimation stage.

The algorithms reviewed in this chapter form the foundation for the algorithms proposed in the remainder of this thesis. In Chapter 4, we propose a coupled, structured estimation hybrid speech enhancement approach by embedding the temporal MVDR filter within a deep learning framework utilizing the TCN architecture and show that the coupled estimation procedure substantially outperforms an uncoupled estimation procedure as well as purely learning-based algorithms. In Chapter 5, we extend the hybrid approach to binaural speech enhancement by embedding the binaural spatio-temporal Wiener filter within a deep learning framework and show that the resulting algorithm outperforms two purely learning-based algorithms. In Chapter 6, we propose a spatial regularization procedure for the estimated RTF vector required by the spatio-temporal MVDR filter and show that this procedure improves interpretability of the RTF vector while maintaining speech enhancement performance and not introducing any additional computational complexity. Finally, in Chapter 7, we propose three procedures to improve robustness against varying microphone array configurations for the mask-based beamformer with ASA and show that the resulting algorithm can be applied to microphone array configurations not seen during training.

4

DEEP MULTI-FRAME FILTER FOR SINGLE-MICROPHONE SPEECH ENHANCEMENT

As discussed in Chapter 3, the temporal MVDR filter yields very good noise reduction with little speech distortion when accurate estimates of the required quantities are available. However, its performance is rather sensitive to estimation errors, especially in the speech temporal correlation vector [54]. None of the currently available model-based approaches is able to yield sufficiently accurate estimates mainly due to the highly time-varying nature of the required quantities, and hence the potential of multi-frame algorithms has not been fully exploited.

Aiming at better exploiting the potential of multi-frame algorithms, in this chapter we propose a coupled, structured estimation hybrid speech enhancement approach, where all required quantities of the temporal MVDR filter are estimated by embedding the fully differentiable temporal MVDR filter within a deep learning framework using TCNs. Instead of using a loss defined on the quantities, the TCNs are trained by minimizing the signal approximation SI-SDR loss function at the output of the temporal MVDR filter. We investigate different matrix structures for the TCMs, namely Hermitian positive-definite, Hermitian positive-definite Toeplitz, and rank-1. For the Hermitian positive-definite matrix structure, we consider an estimation procedure based on recursive smoothing, where only the smoothing factors are estimated using TCNs, as well as an estimation procedure based on the Cholesky decomposition. The main differences between the considered TCM estimation procedures lie in the number of parameters that need to be estimated by the TCNs

This chapter is partly based on:

[175] M. Tammen and S. Doclo, "Deep Multi-Frame MVDR Filtering for Single-Microphone Speech Enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, Jun. 2021, pp. 8443–8447.

[176] M. Tammen and S. Doclo, "Parameter Estimation Procedures for Deep Multi-Frame MVDR Filtering for Single-Microphone Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3237–3248, Aug. 2023.

as well as the required linear algebra operations, yielding a different computational complexity. In the case of rank-1 TCMs, we show that the temporal MVDR filter can be written as a linear combination of the TCN outputs, significantly reducing the computational complexity. Experimental results on the DNS 1 challenge dataset demonstrate that the proposed coupled hybrid approach of estimating the TCMs yields a substantial improvement in speech enhancement performance compared with a decoupled SPP-driven hybrid approach as well as consistent improvements compared with a purely learning-based approach that directly estimates the temporal filter coefficients. Furthermore, the TCM estimation procedure using the Hermitian positive-definite matrix structure based on the Cholesky decomposition yields the best performance. Interestingly, the estimation procedure using the rank-1 matrix structure yields only a slightly lower performance, with the advantage of being computationally less demanding.

The remainder of this chapter is organized as follows. In Section 4.2, the conventional quantity approximation-based SPP-driven supervised learning approach to estimate the quantities of the temporal MVDR filter is reviewed. In Section 4.3, we propose a signal approximation-based approach to estimate the quantities of the temporal MVDR filter, including multiple procedures to estimate the required TCMs. The simulation setup is discussed in Section 4.4 and the corresponding simulation results are presented in Section 4.5.

4.1 Temporal Covariance Matrix Structures

In this chapter, we consider different matrix structures for the $N \times N$-dimensional noisy and interference TCMs, which differ in the number of parameters required to determine these matrices. As we consider a single-microphone configuration, we omit out the reference microphone index r. First, by definition, TCMs are Hermitian. We assume that the considered TCMs are full-rank (rank-N), such that they are positive-definite, i.e., all eigenvalues are real-valued and larger than zero. Hence, the noisy and interference TCMs can be decomposed using the Cholesky decomposition [207] as

$$\bar{\mathbf{\Phi}}_{\nu,t} = \mathbf{L}_{\nu,t}\mathbf{L}_{\nu,t}^{\mathsf{H}}, \quad \nu \in \{y, i\}, \tag{4.1}$$

where the Cholesky factor $\mathbf{L}_{\nu,t}$ is an $N \times N$-dimensional complex-valued lower-triangular matrix with real and positive diagonal elements determined by N^2 real-valued parameters and ν indicates the noisy or interference component.

Assuming the signals to be stationary over N frames, the TCMs also exhibit a Toeplitz structure, i.e., the elements on all diagonals are equal. It has been shown in [208] that Hermitian positive-definite Toeplitz (PDT) matrices can be decomposed using their so-called balanced Vandermonde factorization as

$$\bar{\mathbf{\Phi}}_{\nu,t} = \mathbf{V}_{\nu,t}\mathbf{D}_{\nu,t}\mathbf{V}_{\nu,t}^{\mathsf{H}}, \quad \nu \in \{y, i\}, \tag{4.2}$$

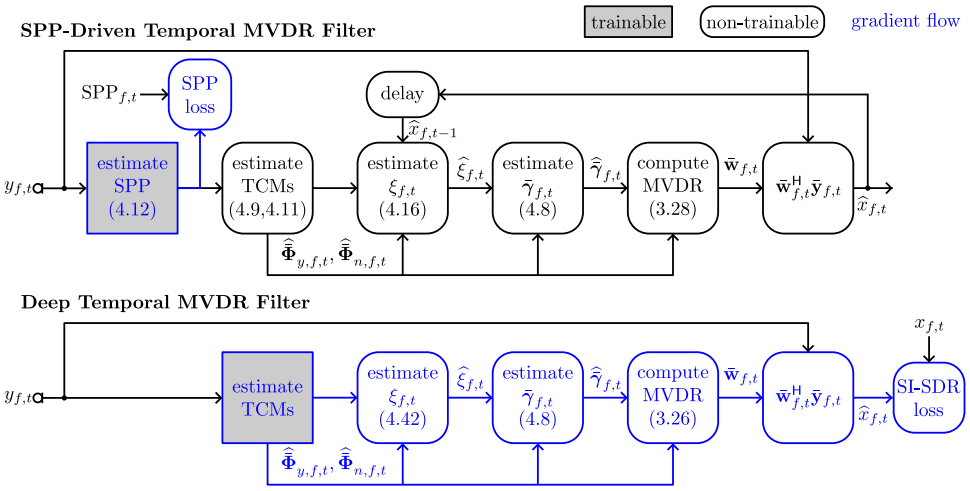


Figure 4.1: Block diagrams of the baseline SPP-driven temporal MVDR filter [161] and the proposed deep temporal MVDR filter. For the baseline SPP-driven temporal MVDR filter, the SPP estimator is trained decoupled from the temporal MVDR filter, and the a-priori SNR is estimated using the decision-directed approach (DDA), which utilizes the speech estimate $\widehat{x}_{t-1}$ from the previous frame. For the proposed signal approximation-based deep temporal MVDR filter, the TCM estimators and the a-priori SNR estimator are jointly trained taking into account the temporal MVDR filter using the SI-SDR loss function.

with $\mathbf{D}_{\nu,t}$ an $N \times N$-dimensional diagonal matrix with real and positive elements and $\mathbf{V}_{\nu,t}$ an $N \times N$-dimensional balanced Vandermonde matrix, defined as

$$\mathbf{V}_{\nu,t} = \begin{bmatrix} 1 & \zeta_{\nu,t,0}^{1} & \zeta_{\nu,t,0}^{2} & \cdots & \zeta_{\nu,t,0}^{N-1} \\ 1 & \zeta_{\nu,t,1}^{1} & \zeta_{\nu,t,1}^{2} & \cdots & \zeta_{\nu,t,1}^{N-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \zeta_{\nu,t,N-1}^{1} & \zeta_{\nu,t,N-1}^{2} & \cdots & \zeta_{\nu,t,N-1}^{N-1} \end{bmatrix}, \quad \nu \in \{y, i\}, \tag{4.3}$$

with $\zeta_{\nu,t,\mu}$ a complex number on the unit circle, i.e., $\zeta_{\nu,t,\mu} = \exp(j\theta_{\nu,t,\mu})$ $\forall \mu \in \{0, \ldots, N-1\}$. Hence, since a balanced Vandermonde matrix can be fully described by the angles $\theta_{\nu,t,\mu}$, the matrices $\mathbf{V}_{\nu,t}$ and $\mathbf{D}_{\nu,t}$ are described by N real-valued parameters each. It should be noted that this assumption presumably holds better for the noise component than for the speech and interference components, which tend to be quite non-stationary.

4.2 SPP-Based Deep MFMVDR Filter

In this section, we review the SPP-driven approach presented in [161] to estimate the required quantities of the temporal MVDR filter in (3.26), depicted in Fig. 4.1 (top). We employ this quantity approximation-based approach for a baseline algorithm in

order to investigate the impact of embedding the temporal MVDR filter in a deep learning framework, hence employing a signal approximation approach. Neglecting the uncorrelated speech component (such that the interference TCM $\bar{\mathbf{\Phi}}_{i,t}$ reduces to the noise TCM $\bar{\mathbf{\Phi}}_{n,t}$ and resulting in the simplified temporal MVDR filter in (3.27)), estimates of the inverse noise TCM $\bar{\mathbf{\Phi}}_{n,t}^{-1}$ and the speech TCV $\bar{\boldsymbol{\gamma}}_{t,r}^{r}$ are required. Substituting (2.20) in (2.28), it can be shown that

$$\bar{\mathbf{\Phi}}_{y,t}\bar{\mathbf{e}} = \phi_{x,t}\bar{\boldsymbol{\gamma}}_t + \bar{\mathbf{\Phi}}_{n,t}\bar{\mathbf{e}}, \tag{4.4}$$

such that the speech TCV can be written as

$$\bar{\boldsymbol{\gamma}}_t = \frac{\bar{\mathbf{\Phi}}_{y,t}\bar{\mathbf{e}}}{\phi_{x,t}} - \frac{\bar{\mathbf{\Phi}}_{n,t}\bar{\mathbf{e}}}{\phi_{x,t}}. \tag{4.5}$$

By defining the a-priori SNR as

$$\xi_t = \frac{\phi_{x,t}}{\phi_{n,t}}, \tag{4.6}$$

with $\phi_{n,t} = \bar{\mathbf{e}}^{\mathsf{T}}\bar{\mathbf{\Phi}}_{n,t}\bar{\mathbf{e}}$ the noise PSD, and using

$$\phi_{y,t} = \bar{\mathbf{e}}^{\mathsf{T}}\bar{\mathbf{\Phi}}_{y,t}\bar{\mathbf{e}} = \phi_{x,t} + \phi_{n,t}, \tag{4.7}$$

the speech TCV in (4.5) can be written in terms of the noisy TCM $\bar{\mathbf{\Phi}}_{y,t}$, the noise TCM $\bar{\mathbf{\Phi}}_{n,t}$, and the a-priori SNR ξ_t as

$$\bar{\boldsymbol{\gamma}}_t = \frac{1+\xi_t}{\xi_t}\frac{\bar{\mathbf{\Phi}}_{y,t}\bar{\mathbf{e}}}{\bar{\mathbf{e}}^{\mathsf{T}}\bar{\mathbf{\Phi}}_{y,t}\bar{\mathbf{e}}} - \frac{1}{\xi_t}\frac{\bar{\mathbf{\Phi}}_{n,t}\bar{\mathbf{e}}}{\bar{\mathbf{e}}^{\mathsf{T}}\bar{\mathbf{\Phi}}_{n,t}\bar{\mathbf{e}}}. \tag{4.8}$$

Hence, an estimate of the speech TCV can be obtained based on an estimate of the a-priori SNR.

To estimate the a-priori SNR, the noise TCM $\bar{\mathbf{\Phi}}_{n,t}$ is first estimated using recursive smoothing with time- and frequency-dependent smoothing factors $\lambda_{n,t}^{\mathrm{SPP}}$. These smoothing factors are computed based on an estimate of the SPP [48], allowing the noise estimate to adapt to speech activity, i.e,

$$\widehat{\bar{\mathbf{\Phi}}}_{n,t}^{\mathrm{SPP}} = \lambda_{n,t}^{\mathrm{SPP}}\widehat{\bar{\mathbf{\Phi}}}_{n,t-1}^{\mathrm{SPP}} + (1-\lambda_{n,t}^{\mathrm{SPP}})\bar{\mathbf{y}}_t\bar{\mathbf{y}}_t^{\mathsf{H}}, \tag{4.9}$$

$$\lambda_{n,t}^{\mathrm{SPP}} = \alpha_n^{\mathrm{SPP}} + (1-\alpha_n^{\mathrm{SPP}})\widehat{\mathrm{SPP}}_t, \tag{4.10}$$

where α_n^{SPP} serves as a lower bound for $\lambda_{n,t}^{\mathrm{SPP}}$, ensuring a minimum level of smoothing, and $\widehat{\mathrm{SPP}}_t$ denotes the estimated SPP. When speech is likely present, $\lambda_{n,t}^{\mathrm{SPP}}$ is high, and the current estimate is strongly influenced by the past noise estimate. In contrast, when speech is likely absent, $\lambda_{n,t}^{\mathrm{SPP}}$ is low, and the current estimate updates rapidly. The noisy TCM $\bar{\mathbf{\Phi}}_{y,t}$ is also estimated using recursive smoothing but with a fixed smoothing factor λ_y^{SPP}, since it captures both speech and noise, i.e.,

$$\widehat{\bar{\mathbf{\Phi}}}_{y,t}^{\mathrm{SPP}} = \lambda_y^{\mathrm{SPP}}\widehat{\bar{\mathbf{\Phi}}}_{y,t-1}^{\mathrm{SPP}} + (1-\lambda_y^{\mathrm{SPP}})\bar{\mathbf{y}}_t\bar{\mathbf{y}}_t^{\mathsf{H}}. \tag{4.11}$$

The SPP in each frequency bin is estimated using a TCN $\mathbf{f}^{\text{SPP}}$ with trainable weights $\boldsymbol{\theta}^{\text{SPP}}$, i.e.,[0]

$$\widehat{\text{SPP}}_{f,t} = \left[\mathbf{f}^{\text{SPP}}\left(\{\boldsymbol{\chi}_{f,t}\}_{t=1}^{T}; \boldsymbol{\theta}^{\text{SPP}}\right)\right]_f, \tag{4.12}$$

with $[\cdot]_f$ denoting the output for the f-th frequency bin, and with the sequence of logarithmic magnitudes of the noisy STFT coefficients, concatenated along the frequency dimension, chosen as the input features, i.e.,

$$\boldsymbol{\chi}_t = \begin{bmatrix}\log_{10}|y_{f=1,t}| & \log_{10}|y_{f=2,t}| & \cdots & \log_{10}|y_{f=F,t}|\end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^F. \tag{4.13}$$

The training target of the TCN was the SPP defined in [49], which assumes independent complex Gaussian-distributed speech and noise STFT coefficients.

To ensure invertibility of the estimated noise TCM before using it in (3.28), diagonal loading is applied, i.e.,

$$\widehat{\bar{\boldsymbol{\Phi}}}^{\text{SPP}}_{n,f,t} \leftarrow \widehat{\bar{\boldsymbol{\Phi}}}^{\text{SPP}}_{n,f,t} + \rho^{\text{SPP}}_{n,f,t}\mathbf{I}_N, \tag{4.14}$$

where $\mathbf{I}_N$ denotes the $N \times N$-dimensional identity matrix, and $\rho^{\text{SPP}}_{n,f,t}$ denotes the regularization factor defined as [53], [209]

$$\rho^{\text{SPP}}_{n,f,t} = \frac{\rho}{N}\text{trace}\left(\widehat{\bar{\boldsymbol{\Phi}}}^{\text{SPP}}_{n,f,t}\right), \tag{4.15}$$

with ρ a small constant. The a-priori SNR $\xi_{f,t}$ is estimated using the decision-directed approach (DDA) [39], i.e.,

$$\widehat{\xi}^{\text{SPP}}_{f,t} = \lambda^{\text{DDA}}\frac{\left|\widehat{x}_{f,t-1}\right|^2}{\widehat{\phi}^{\text{SPP}}_{n,f,t-1}} + \left(1-\lambda^{\text{DDA}}\right)\max\left(\frac{|y_{f,t}|^2}{\widehat{\phi}^{\text{SPP}}_{n,f,t}} - 1, 0\right), \tag{4.16}$$

where λ^{DDA} denotes a smoothing factor, $\widehat{\phi}^{\text{SPP}}_{n,f,t} = \bar{\mathbf{e}}^{\mathsf{T}}\widehat{\bar{\boldsymbol{\Phi}}}^{\text{SPP}}_{n,f,t}\bar{\mathbf{e}}$ is an estimate of the noise PSD based on the estimated noise TCM in (4.9), and $\widehat{x}_{f,t-1}$ is the estimated speech component in the previous time frame. In [161], employing the estimated quantities obtained using this SPP-driven estimation approach in a temporal MVDR filter resulted in a higher speech enhancement performance compared with employing them in a (single-frame) Wiener filter. However, especially if the SNR is low or if the noise statistics are changing rapidly, estimation errors can lead to reduced speech enhancement performance.

4.3 Signal Approximation-Based Deep MFMVDR Filter

Contrary to the SPP-driven approach described in the previous section, in this section we propose the deep temporal MVDR filter, which relies on a signal approximation-based approach to estimate all quantities using DNNs that are jointly

0 For the remainder of this chapter, we reintroduce the frequency bin f for precise notation.

trained with a loss function at the output of the temporal MVDR filter, depicted in Fig. 4.1 (bottom). In other words, the training of the DNNs is guided by the speech estimate obtained at the output of the deep temporal MVDR filter and hence no ground-truth quantities are required. As already mentioned, the quantities required to compute the speech TCV $\bar{\boldsymbol{\gamma}}_{f,t}$ in (4.8) and the temporal MVDR filter vector in (3.26) are the noisy TCM $\bar{\boldsymbol{\Phi}}_{y,f,t}$, the inverse interference TCM $\bar{\boldsymbol{\Phi}}_{i,f,t}^{-1}$, and the a-priori SNR $\xi_{f,t}$. A separate TCN is used per quantity, with different input features for the TCNs estimating the TCMs (Section 4.3.1) and the a-priori SNR (Section 4.3.2).

4.3.1 *Covariance Matrices*

In this section we propose different estimation procedures for the noisy and interference TCMs $\bar{\boldsymbol{\Phi}}_{y,f,t}$ and $\bar{\boldsymbol{\Phi}}_{i,f,t}$. All estimation procedures have in common that the TCN estimating $\bar{\boldsymbol{\Phi}}_{y,f,t}$ and the TCN estimating $\bar{\boldsymbol{\Phi}}_{i,f,t}$ are jointly trained using a signal approximation loss function (Fig. 4.1), i.e., without the need for defining target TCMs. In the following, we will consider Hermitian positive-definite, Hermitian PDT, and rank-1 matrix structures, where the main difference lies in the number of parameters that need to be estimated as well as in the required linear algebra operations. It should be noted that similarly to (4.14), diagonal loading is applied to the estimated interference TCM before using it in (3.26). Since the TCMs capture phase relationships, we use a concatenation of the logarithmic magnitude as well as the cosine and sine of the phase of the noisy STFT coefficients at all frequency bins as input features $\boldsymbol{\chi}_t$ for both DNNs[1], i.e.,

$$\boldsymbol{\chi}_{f,t} = \begin{bmatrix} \log_{10}(|y_{f,t}| + \epsilon) & \cos\left(\angle y_{f,t}\right) & \sin\left(\angle y_{f,t}\right) \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^3 \tag{4.17}$$

$$\boldsymbol{\chi}_t = \begin{bmatrix} \boldsymbol{\chi}_{f=1,t}^{\mathsf{T}} & \boldsymbol{\chi}_{f=2,t}^{\mathsf{T}} & \cdots & \boldsymbol{\chi}_{f=F,t}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^{3F}, \tag{4.18}$$

where ϵ denotes a small positive constant to avoid numerical issues and $\angle\cdot$ denotes the phase. Both the cosine and sine of the phase are used to obtain an unambiguous and smooth phase representation [83].

4.3.1.1 *Hermitian Positive-Definite*

We propose two different estimation procedures that result in the assumed Hermitian positive-definite structure for TCMs. The first procedure is based on recursive smoothing and only requires one parameter to be estimated for each time-frequency bin and by each DNN, while the second procedure is based on the Cholesky decomposition and requires N^2 parameters to be estimated for each time-frequency bin and by each DNN.

1 In preliminary experiments, this feature choice outperformed the use of the real and imaginary parts of the STFT coefficients as input features.

RECURSIVE SMOOTHING (RS) Similarly to (4.9), the noisy and interference TCMs are estimated as

$$\widehat{\bar{\mathbf{\Phi}}}^{\mathrm{RS}}_{\nu,f,t} = \lambda^{\mathrm{RS}}_{\nu,f,t}\widehat{\bar{\mathbf{\Phi}}}^{\mathrm{RS}}_{\nu,f,t-1} + (1-\lambda^{\mathrm{RS}}_{\nu,f,t})\mathbf{y}_{f,t}\mathbf{y}^{\mathsf{H}}_{f,t}, \quad \nu \in \{y,i\}, \tag{4.19}$$

where the recursive smoothing factors are obtained using DNNs $\mathbf{f}^{\mathrm{RS}}_{\nu}$ with trainable weights $\boldsymbol{\theta}^{\mathrm{RS}}_{\nu}$, i.e.,

$$\lambda^{\mathrm{RS}}_{\nu,f,t} = \left[\mathbf{f}^{\mathrm{RS}}_{\nu}\left(\left\{\boldsymbol{\chi}_{f,t};\boldsymbol{\theta}^{\mathrm{RS}}_{\nu}\right\}_{t=1}^{T}\right)\right]_{f}, \quad \nu \in \{y,i\}, \tag{4.20}$$

where a sigmoid activation function is used to ensure that the recursive smoothing factors are bounded to $[0,1]$. The proposed recursive smoothing procedure differs from the conventional recursive smoothing procedure in Section 4.2 by allowing a time-varying smoothing factor for both TCMs and by jointly training the DNNs with a signal approximation loss function instead of a quantity approximation SPP-driven loss function.

CHOLESKY DECOMPOSITION (CD) As already mentioned in Section 4.1, the $N \times N$-dimensional lower-triangular Cholesky factors $\mathbf{L}_{y,f,t}$ and $\mathbf{L}_{i,f,t}$ of the noisy and interference TCMs are fully determined by N^2 real-valued parameters each, which can be stacked in the vectors $\boldsymbol{o}^{\mathrm{CD}}_{y,f,t}$, $\boldsymbol{o}^{\mathrm{CD}}_{i,f,t} \in \mathbb{R}^{N^2}$. We propose to estimate these vectors using separate TCNs $\mathbf{f}^{\mathrm{CD}}_{y}$ and $\mathbf{f}^{\mathrm{CD}}_{i}$ with trainable weights $\boldsymbol{\theta}^{\mathrm{CD}}_{y}$ and $\boldsymbol{\theta}^{\mathrm{CD}}_{i}$, i.e.,

$$\widehat{\boldsymbol{o}}^{\mathrm{CD}}_{y,f,t} = \left[\mathbf{f}^{\mathrm{CD}}_{y}\left(\{\boldsymbol{\chi}_t\}_{t=1}^{T};\boldsymbol{\theta}^{\mathrm{CD}}_{y}\right)\right]_{f} \tag{4.21}$$

$$\widehat{\boldsymbol{o}}^{\mathrm{CD}}_{i,f,t} = \left[\mathbf{f}^{\mathrm{CD}}_{i}\left(\{\boldsymbol{\chi}_t\}_{t=1}^{T};\boldsymbol{\theta}^{\mathrm{CD}}_{i}\right)\right]_{f}. \tag{4.22}$$

The estimated Cholesky factors $\widehat{\mathbf{L}}_{y,f,t}$ and $\widehat{\mathbf{L}}_{i,f,t}$ are assembled from $\widehat{\boldsymbol{o}}^{\mathrm{CD}}_{y,f,t}$ and $\widehat{\boldsymbol{o}}^{\mathrm{CD}}_{i,f,t}$ as described in Algorithm 1, where separate subsets of the real-valued vector elements are used for the real strictly lower triangular part, the imaginary strictly lower triangular part, and the real positive diagonal part. Positivity of the diagonal part is ensured by using a softplus activation function. Finally, estimates of the TCMs are obtained as $\widehat{\bar{\mathbf{\Phi}}}^{\mathrm{CD}}_{y,f,t} = \widehat{\mathbf{L}}_{y,f,t}\widehat{\mathbf{L}}^{\mathsf{H}}_{y,f,t}$ and $\widehat{\bar{\mathbf{\Phi}}}^{\mathrm{CD}}_{i,f,t} = \widehat{\mathbf{L}}_{i,f,t}\widehat{\mathbf{L}}^{\mathsf{H}}_{i,f,t}$.

4.3.1.2 *Hermitian Positive-Definite Toeplitz (PDT)*

When assuming stationarity of the noisy and interference components over N frames, the respective TCMs $\bar{\mathbf{\Phi}}_{y,f,t}$ and $\bar{\mathbf{\Phi}}_{i,f,t}$ exhibit a Hermitian PDT structure. As mentioned in Section 4.1, PDT matrices can be decomposed as $\bar{\mathbf{\Phi}}_{\nu,f,t} = \mathbf{V}_{\nu,f,t}\mathbf{D}_{\nu,f,t}\mathbf{V}^{\mathsf{H}}_{\nu,f,t}$, where the balanced Vandermonde matrix $\mathbf{V}_{\nu,f,t}$ in (4.3) is fully determined by N angles $\theta_{\nu,f,t,\mu}$, and the (positive-definite) diagonal matrix $\mathbf{D}_{\nu,f,t}$ is fully determined by N real-valued parameters. These angles and parameters can

Algorithm 1 Construct a Hermitian positive-definite matrix $\widehat{\hat{\mathbf{\Phi}}}$ from a vector of real-valued parameters $\widehat{\boldsymbol{o}}$ using its Cholesky decomposition. $\mathrm{stril}(\cdot)$ assembles a strictly lower triangular matrix.

procedure CHOLESKY($\widehat{\boldsymbol{h}} \in \mathbb{R}^{N^2}$)
  construct strictly lower triangular matrices:
  $\widehat{\mathbf{O}}^{\Re} = \mathrm{stril}\left([\widehat{\boldsymbol{o}}]_{1:\frac{1}{2}(N^2-N)}\right)$
  $\widehat{\mathbf{O}}^{\Im} = \mathrm{stril}\left([\widehat{\boldsymbol{o}}]_{\frac{1}{2}(N^2-N)+1:N^2-N}\right)$
  construct real-valued positive diagonal matrix:
  $\widehat{\mathbf{B}} = \mathrm{diag}\left(\mathrm{softplus}\left([\widehat{\boldsymbol{o}}]_{N^2-N+1:N^2}\right)\right)$
  construct Cholesky factor:
  $\widehat{\mathbf{L}} = \widehat{\mathbf{O}}^{\Re} + j\widehat{\mathbf{O}}^{\Im} + \widehat{\mathbf{B}}, \quad j^2 = -1$
  return $\widehat{\hat{\mathbf{\Phi}}} = \widehat{\mathbf{L}}\widehat{\mathbf{L}}^{\mathsf{H}}$
end procedure

Algorithm 2 Construct a Hermitian PDT matrix $\widehat{\hat{\mathbf{\Phi}}}$ from a vector of real-valued parameters $\widehat{\mathbf{a}}$ using its Vandermonde factorization.

procedure VANDERMONDE($\widehat{\mathbf{a}} \in \mathbb{R}^{2N}$)
  $\widehat{\boldsymbol{\zeta}} = \begin{bmatrix}\widehat{\zeta}_1 & \cdots & \widehat{\zeta}_N\end{bmatrix}^{\mathsf{T}} = \exp(j\pi \tanh([\widehat{\mathbf{a}}]_{1:N}))$
  using $\widehat{\boldsymbol{\zeta}}$, assemble $\widehat{\mathbf{V}}$ as in (4.3)
  $\widehat{\mathbf{D}} = \mathrm{diag}\left(\mathrm{softplus}\left([\widehat{\mathbf{a}}]_{N+1:2N}\right)\right)$
  return $\widehat{\hat{\mathbf{\Phi}}} = \widehat{\mathbf{V}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}^{\mathsf{H}}$
end procedure

be stacked in the vectors $\boldsymbol{o}^{\mathrm{PDT}}_{y,f,t}$, $\boldsymbol{o}^{\mathrm{PDT}}_{i,f,t} \in \mathbb{R}^{2N}$. We propose to estimate these vectors using separate TCNs $\mathbf{f}^{\mathrm{PDT}}_y$ and $\mathbf{f}^{\mathrm{PDT}}_i$ with trainable weights $\boldsymbol{\theta}^{\mathrm{PDT}}_y$ and $\boldsymbol{\theta}^{\mathrm{PDT}}_i$, i.e.,

$$\widehat{\boldsymbol{o}}^{\mathrm{PDT}}_{y,f,t} = \left[\mathbf{f}^{\mathrm{PDT}}_y\left(\{\boldsymbol{\chi}_t\}_{t=1}^{T}; \boldsymbol{\theta}^{\mathrm{PDT}}_y\right)\right]_f \tag{4.23}$$

$$\widehat{\boldsymbol{o}}^{\mathrm{PDT}}_{i,f,t} = \left[\mathbf{f}^{\mathrm{PDT}}_i\left(\{\boldsymbol{\chi}_t\}_{t=1}^{T}; \boldsymbol{\theta}^{\mathrm{PDT}}_i\right)\right]_f. \tag{4.24}$$

The estimated Hermitian PDT matrices $\widehat{\hat{\mathbf{\Phi}}}^{\mathrm{PDT}}_{y,f,t}$ and $\widehat{\hat{\mathbf{\Phi}}}^{\mathrm{PDT}}_{i,f,t}$ are assembled from $\widehat{\boldsymbol{o}}^{\mathrm{PDT}}_{y,f,t}$ and $\widehat{\boldsymbol{o}}^{\mathrm{PDT}}_{i,f,t}$ as described in Algorithm 2. The angles are computed from the first N elements of $\widehat{\boldsymbol{o}}^{\mathrm{PDT}}_{\nu,f,t}$, bounded to $[-\pi, \pi]$ using a scaled hyperbolic tangent activation function, i.e., $\widehat{\theta}_{\nu,f,t,\mu} = \pi \tanh\left([\widehat{\boldsymbol{o}}^{\mathrm{PDT}}_{\nu,f,t}]_\mu\right)$. The diagonal matrix $\widehat{\mathbf{D}}_{\nu,f,t}$ is computed from the next N elements of $\widehat{\boldsymbol{o}}^{\mathrm{PDT}}_{\nu,f,t}$, where positivity of the diagonal elements is ensured by using a softplus activation function. Finally, estimates of the TCMs are obtained as $\widehat{\mathbf{\Phi}}^{\mathrm{PDT}}_{y,f,t} = \widehat{\mathbf{V}}_{y,f,t}\widehat{\mathbf{D}}_{y,f,t}\widehat{\mathbf{V}}^{\mathsf{H}}_{y,f,t}$ and $\widehat{\mathbf{\Phi}}^{\mathrm{PDT}}_{i,f,t} = \widehat{\mathbf{V}}_{i,f,t}\widehat{\mathbf{D}}_{i,f,t}\widehat{\mathbf{V}}^{\mathsf{H}}_{i,f,t}$.

4.3.1.3 *Rank-1 (R1)*

Although the noisy and interference TCMs are full-rank, here we assume that these matrices can be approximated using a rank-1 structure,[2] i.e.,

$$\bar{\mathbf{\Phi}}^{\mathrm{R1}}_{y,f,t} = \boldsymbol{o}^{\mathrm{R1},\mathbb{C}}_{y,f,t}\left(\boldsymbol{o}^{\mathrm{R1},\mathbb{C}}_{y,f,t}\right)^{\mathsf{H}}, \tag{4.25}$$

$$\bar{\mathbf{\Phi}}^{\mathrm{R1}}_{i,f,t} = \boldsymbol{o}^{\mathrm{R1},\mathbb{C}}_{i,f,t}\left(\boldsymbol{o}^{\mathrm{R1},\mathbb{C}}_{i,f,t}\right)^{\mathsf{H}}, \tag{4.26}$$

where $\boldsymbol{o}^{\mathrm{R1},\mathbb{C}}_{y,f,t}$ and $\boldsymbol{o}^{\mathrm{R1},\mathbb{C}}_{i,f,t}$ denote N-dimensional complex-valued vectors fully determined by $2N$ real-valued parameters each, which can be stacked in the vectors $\boldsymbol{o}^{\mathrm{R1}}_{y,f,t}$, $\boldsymbol{o}^{\mathrm{R1}}_{i,f,t} \in \mathbb{R}^{2N}$. We propose to estimate these vectors using separate TCNs $\mathbf{f}^{\mathrm{R1}}_{y}$ and $\mathbf{f}^{\mathrm{R1}}_{i}$ with trainable weights $\boldsymbol{\theta}^{\mathrm{R1}}_{y}$ and $\boldsymbol{\theta}^{\mathrm{R1}}_{i}$, i.e.,

$$\widehat{\boldsymbol{o}}^{\mathrm{R1}}_{y,f,t} = \left[\mathbf{f}^{\mathrm{R1}}_{y}\left(\{\boldsymbol{\chi}_t\}_{t=1}^{T}; \boldsymbol{\theta}^{\mathrm{R1}}_{y}\right)\right]_f \tag{4.27}$$

$$\widehat{\boldsymbol{o}}^{\mathrm{R1}}_{i,f,t} = \left[\mathbf{f}^{\mathrm{R1}}_{i}\left(\{\boldsymbol{\chi}_t\}_{t=1}^{T}; \boldsymbol{\theta}^{\mathrm{R1}}_{i}\right)\right]_f. \tag{4.28}$$

The vectors $\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{y,f,t}$ and $\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t}$ are then obtained as

$$\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{\nu,f,t} = [\widehat{\boldsymbol{o}}^{\mathrm{R1}}_{\nu,f,t}]_{1:N} + j\ [\widehat{\boldsymbol{o}}^{\mathrm{R1}}_{\nu,f,t}]_{N+1:2N}. \tag{4.29}$$

Since the rank-1 matrix $\widehat{\bar{\mathbf{\Phi}}}^{\mathrm{R1}}_{i,f,t} = \widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t}\left(\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t}\right)^{\mathsf{H}}$ is not invertible and hence cannot be directly used in (3.26), diagonal loading is applied, i.e.,

$$\widehat{\bar{\mathbf{\Phi}}}^{\mathrm{R1}}_{i,f,t} \leftarrow \widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t}\left(\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t}\right)^{\mathsf{H}} + \rho^{\mathrm{R1}}_{i,f,t}\mathbf{I}_N, \tag{4.30}$$

where the regularization factor $\rho^{\mathrm{R1}}_{i,f,t}$ is defined as in (4.15), i.e.,

$$\rho^{\mathrm{R1}}_{i,f,t} = \frac{\rho}{N}\mathrm{trace}\left(\widehat{\bar{\mathbf{\Phi}}}^{\mathrm{R1}}_{i,f,t}\right) = \frac{\rho}{N}\left\|\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t}\right\|_2^2, \tag{4.31}$$

with ρ a small constant and $\|\cdot\|_2$ denoting the ℓ^2-norm. Using the (regularized) rank-1 TCMs in (4.25) and (4.30), it can be shown that the temporal MVDR filter in (3.26) can be directly computed as a linear combination of the TCN outputs $\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{y,f,t}$ and $\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t}$ without requiring computationally complex matrix inversions. First, using (4.25) and (4.30) in (4.8), the speech TCV can be written as a linear combination of the TCN outputs, i.e.,

$$\widehat{\boldsymbol{\gamma}}^{\mathrm{R1}}_{t} = \alpha_{y,f,t}\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{y,f,t} + \alpha_{i,f,t}\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t} + \alpha_{e,f,t}\mathbf{e}, \tag{4.32}$$

2 Although we realize that this approximation lacks a theoretical motivation, especially for the noisy TCM, it will result in a lower-complexity estimation procedure with a very good speech enhancement performance (Section 4.5).

with

$$\alpha_{y,f,t} = \frac{1+\widehat{\xi}_{f,t}}{\widehat{\xi}_{f,t}} \frac{1}{[\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{y,f,t}]_0} \tag{4.33}$$

$$\alpha_{i,f,t} = -\frac{1}{\widehat{\xi}_{f,t}} \frac{[\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t}]^*_0}{|[\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t}]_0|^2 + \rho^{\mathrm{R1}}_{i,f,t}} \tag{4.34}$$

$$\alpha_{e,f,t} = \alpha_{i,f,t} \frac{\rho^{\mathrm{R1}}_{i,f,t}}{[\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t}]^*_0}. \tag{4.35}$$

By using the matrix inversion lemma, it can be easily shown that the inverse of the interference TCM in (4.30) is equal to

$$\left(\widehat{\tilde{\boldsymbol{\Phi}}}^{\mathrm{R1}}_{i,f,t}\right)^{-1} = \frac{1}{\rho^{\mathrm{R1}}_{i,f,t}} \left(\mathbf{I}_N - \eta_{f,t} \widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t} \left(\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t}\right)^{\mathsf{H}}\right), \tag{4.36}$$

with

$$\eta_{f,t} = \frac{1}{\rho^{\mathrm{R1}}_{i,f,t} + \left\|\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t}\right\|_2^2} \in \mathbb{R}. \tag{4.37}$$

By substituting (4.36) in (3.26), the temporal MVDR filter vector can be written as

$$\mathbf{w}^{\mathrm{R1}}_{f,t} = \frac{\left(\mathbf{I}_N - \eta_{f,t} \widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t} \left(\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t}\right)^{\mathsf{H}}\right)}{\left(\widehat{\boldsymbol{\gamma}}^{\mathrm{R1}}_t\right)^{\mathsf{H}} \left(\mathbf{I}_N - \eta_{f,t} \widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t} \left(\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t}\right)^{\mathsf{H}}\right) \widehat{\boldsymbol{\gamma}}^{\mathrm{R1}}_t} \widehat{\boldsymbol{\gamma}}^{\mathrm{R1}}_t \tag{4.38}$$

$$= \frac{1}{\kappa_{f,t}} \left(\mathbf{I}_N - \eta_{f,t} \widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t} \left(\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t}\right)^{\mathsf{H}}\right) \widehat{\boldsymbol{\gamma}}^{\mathrm{R1}}_t, \tag{4.39}$$

with

$$\kappa_{f,t} = \left\|\widehat{\boldsymbol{\gamma}}^{\mathrm{R1}}_t\right\|_2^2 - \eta_{f,t} \left|\left(\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t}\right)^{\mathsf{H}} \widehat{\boldsymbol{\gamma}}^{\mathrm{R1}}_t\right|^2 \in \mathbb{R}. \tag{4.40}$$

Finally, by substituting (4.32) in (4.39), the temporal MVDR filter vector can be written as a linear combination of the TCN outputs, i.e.,

$$\mathbf{w}^{\mathrm{R1}}_{f,t} = \frac{1}{\kappa_{f,t}} \left[\alpha_{y,f,t} \widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{y,f,t} + \left(\alpha_{i,f,t} - \eta_{f,t} \left(\widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t}\right)^{\mathsf{H}} \widehat{\boldsymbol{\gamma}}^{\mathrm{R1}}_t\right) \widehat{\boldsymbol{o}}^{\mathrm{R1},\mathbb{C}}_{i,f,t} + \alpha_{e,f,t} \mathbf{e}\right]. \tag{4.41}$$

4.3.2 *A-Priori SNR*

To estimate the a-priori SNR $\xi_{f,t}$, a similar approach is used as for estimating the TCMs. Instead of training a DNN to map input features to an a-priori SNR target, a DNN is trained by minimizing a signal approximation loss function at the output of the temporal MVDR filter (Fig. 4.1, bottom), where it should be noted that the DNN $\mathbf{f}^{\xi}$ with trainable weights $\boldsymbol{\theta}^{\xi}$ estimating the a-priori SNR is trained jointly

with the DNNs estimating the TCMs (Section 4.3.1). Since the a-priori SNR is a quantity relating signal powers, magnitude information is assumed to be sufficient for its estimation. Hence, similarly as for the SPP in (4.12), we use the input features $\boldsymbol{\chi}_t$ defined in (4.13), i.e.,

$$\widehat{\xi}_{f,t} = \text{softplus}\left(\left[\mathbf{f}^{\xi}\left(\{\boldsymbol{\chi}_t\}_{t=1}^{T}; \boldsymbol{\theta}^{\xi}\right)\right]_f\right), \tag{4.42}$$

where positivity of $\widehat{\xi}_{f,t}$ is ensured by using a softplus activation function.

4.4 Simulation Setup

In this section, we present our simulation setup, more in particular the used datasets (Section 4.4.1), the baseline speech enhancement algorithms (Section 4.4.2), and the settings of the considered algorithms (Section 4.4.3).

4.4.1 *Datasets*

To train, validate, and evaluate all algorithms, we constructed datasets using diverse speech and noise source material from the DNS 1 challenge [66].

4.4.1.1 *Training & Validation*

The training and validation datasets were generated using the official dataset generator of the DNS 1 challenge [66], i.e., 500 h of English speech from 2150 speakers from the LibriSpeech dataset [210], 60 000 noise clips from the Audioset dataset [211], and 10 000 noise clips from the Freesound and DEMAND datasets [212], [213]. Noisy utterances were generated by randomly choosing speech and noise source signals and mixing them at fullband SNRs ranging from 0 dB to 19 dB, with each utterance of length 4 s.

4.4.1.2 *Testing*

To evaluate the considered speech enhancement algorithms, we used the official DNS 1 challenge evaluation dataset, comprising English speech from the Graz University dataset [214] and noise from the Audioset and Freesound datasets, totaling 150 utterances at SNRs ranging from 0 dB to 19 dB. All evaluation utterances had a duration of 10 s, and the training, validation, and test datasets were disjoint.

4.4.2 *Baseline Algorithms*

In addition to comparing the different proposed estimation procedures for the signal approximation-based deep temporal MVDR filter (Section 4.3), we compare the signal approximation-based deep temporal MVDR filter to several baseline algorithms:

1. SPP-driven temporal MVDR filter (Section 4.2), aiming at investigating the difference between using the SPP as a training target and using a signal approximation loss function defined at the output of the temporal MVDR filter.
2. Masking: Aiming at investigating the benefit of temporal filtering ($N > 1$), we also consider (single-frame) masking algorithms, either using a real-valued mask (i.e., $\widehat{x}_{f,t} = m^{\Re}_{f,t} y_{f,t}$) or a complex-valued mask (i.e., $\widehat{x}_{f,t} = m^{\mathbb{C}}_{f,t} y_{f,t}$). The real-valued mask is estimated using a TCN and input features in (4.17) as
$$m^{\mathbb{R}}_{f,t} = \text{sigmoid}\left(\left[\mathbf{f}^{\mathbb{R}}\left(\{\boldsymbol{\chi}_t\}_{t=1}^{T}; \boldsymbol{\theta}^{\mathbb{R}}\right)\right]_f\right). \tag{4.43}$$
where a sigmoid activation function is used to ensure that the mask is bounded to $[0, 1]$. Similarly, the complex-valued mask is estimated using a TCN as
$$\begin{bmatrix} \Re\left(m^{\mathbb{C}}_{f,t}\right) \\ \Im\left(m^{\mathbb{C}}_{f,t}\right) \end{bmatrix} = \mathbf{tanh}\left(\left[\mathbf{f}^{\mathbb{C}}\left(\{\boldsymbol{\chi}_t\}_{t=1}^{T}; \boldsymbol{\theta}^{\mathbb{C}}\right)\right]_f\right). \tag{4.44}$$
where a hyperbolic tangent activation function is used to ensure that both parts are bounded to $[-1, 1]$.
3. Deep filter (DF) algorithm: Aiming at investigating the benefit of imposing the temporal MVDR filter structure in (3.26), we also consider directly estimating the complex-valued elements of the N-dimensional temporal filter in (2.48) using a TCN, similarly to the deep filter (DF) algorithm [129] described in more detail in Section 3.2.1. The real and imaginary parts of the temporal filter $\mathbf{w}^{\text{DF}}_{f,t}$ are estimated as
$$\begin{bmatrix} \Re\left(\mathbf{w}^{\text{DF}}_{f,t}\right) \\ \Im\left(\mathbf{w}^{\text{DF}}_{f,t}\right) \end{bmatrix} = \mathbf{tanh}\left(\left[\mathbf{f}^{\text{DF}}\left(\{\boldsymbol{\chi}_t\}_{t=1}^{T}; \boldsymbol{\theta}^{\text{DF}}\right)\right]_f\right). \tag{4.45}$$
where a hyperbolic tangent activation function is used to ensure that the real and imaginary parts of all filter coefficients are bounded to $[-1, 1]$ as in [129].

4.4.3 *Algorithmic Settings*

For all considered algorithms, the same STFT framework was used. To increase speech correlation across consecutive STFT frames, a high temporal resolution was utilized, i.e., a frame length of 8 ms and a frame overlap of 75 % (see Section 1.1.1), similarly as in [45]. A $\sqrt{\text{Hann}}$ window was used both as analysis and synthesis

window. All algorithms using temporal filters, i.e., all deep temporal MVDR filters and the DF algorithm, used a filter length of $N = 5$, enabling the algorithms to exploit temporal correlations within 16 ms (motivated by the drop of correlation at a frame lag of $\Delta = 4$ in Fig. 1.4). Similarly as in [161], for the SPP-driven temporal MVDR filter we used a smoothing factor $\alpha_n^{\text{SPP}} = 0.9694$ (corresponding to about 50 ms) for the noise TCM in (4.10), a smoothing factor $\lambda_y^{\text{SPP}} = 0.8464$ (corresponding to about 12 ms) for the noisy TCM in (4.11), and a smoothing factor $\lambda_{\text{DDA}} = 0.9408$ (corresponding to about 33 ms) for the DDA in (4.16). The target SPP was defined as [49]

$$\text{SPP}_{f,t} = \left(1 + (1 + \xi_{\mathcal{H}_1}) \exp\left(-\frac{|Y_{f,t}|^2}{\widehat{\phi}_{n,f,t}} \frac{\xi_{\mathcal{H}_1}}{1 + \xi_{\mathcal{H}_1}}\right)\right)^{-1} \tag{4.46}$$

$$\widehat{\phi}_{n,f,t} = \lambda_{f,t}^{\text{SPP}} \widehat{\phi}_{n,f,t-1} + (1 - \lambda_{f,t}^{\text{SPP}})|n_{f,t}|^2, \tag{4.47}$$

with $\lambda_{f,t}^{\text{SPP}}$ defined in (4.10). For the deep temporal MVDR filter, diagonal loading with a fixed regularization constant $\rho = 10^{-3}$ was applied to all estimated interference or noise TCMs. In (4.17), we used $\epsilon = 1 \times 10^{-8}$ to avoid numerical issues. No recursive smoothing was applied in the case of the Cholesky decomposition (CD), PDT, or rank-1 (R1) deep temporal MVDR filters.

To limit speech distortion, a smooth minimum gain was applied to the output $\widehat{x}_{f,t}$ of all considered algorithms, i.e., the final speech estimate $\widehat{x}_{f,t}^{\text{fin}}$ used in the evaluation was obtained as

$$\widehat{x}_{f,t}^{\text{fin}} = \beta_{f,t} \widehat{x}_{f,t} + (1 - \beta_{f,t}) g_{\min} y_{f,t}, \tag{4.48}$$

where the factor $\beta_{f,t}$ interpolates between the estimated speech component $\widehat{x}_{f,t}$ and a hard minimum gain $g_{\min}$ applied to the noisy STFT coefficients. We propose to use a smooth (time-varying) interpolation factor

$$\beta_{f,t} = \left(1 + \exp(-2s(|\widehat{x}_{f,t}| - |g_{\min} y_{f,t}|))\right)^{-1}, \tag{4.49}$$

where s controls how accurately the hard minimum gain $g_{\min}$ is approximated. The advantage of using the minimum gain approximation in (4.48) over a hard minimum gain (corresponding to $s = \infty$) is the fact that the former is smoothly differentiable, allowing for it to be used during backpropagation. For all algorithms, we used $g_{\min} = -17\,\text{dB}$ and $s = 10$.

The TCNs were trained using the AdamW optimizer [215] with an initial learning rate of 3×10^{-4}. The learning rate was halved after the validation loss did not decrease for 3 consecutive epochs, and training was stopped if either 50 epochs were completed or the validation loss did not decrease for 10 consecutive epochs. The gradient ℓ^2-norms were clipped to 5, and a batch size of 4 was used. As loss function for the SPP-driven temporal MVDR filter we used the mean square error between the estimated SPP in (4.12) and the target SPP defined in (4.46). For all other algorithms, we used the signal approximation-based SI-SDR defined in (2.54) as loss function. All algorithms were implemented in PyTorch 1.10.1 [216], and training and evaluation were executed on NVIDIA GeForce® RTX 2080 Ti graphics

cards. As the DNN architecture for all estimators, we used temporal convolutional networks (TCNs), which exhibit strong temporal and spectral representation capacity [98] and have been shown to be quite effective in the context of speech enhancement [89] (see Section 3.2.2.3 for a more detailed description). We used the official Conv-TasNet TCN implementation,[3] excluding the encoder and decoder modules, since the proposed algorithms operate in the STFT domain. We fixed the hyperparameters of all TCN modules to 2 stacks of 4 layers each, with a kernel size of 3, resulting in a temporal receptive field size of 61 frames (corresponding to 128 ms). Aiming at a fair comparison, the number of channels in the 1D-Conv blocks and the number of channels in the bottleneck were varied to obtain a similar total number of trainable weights for all considered algorithms (Table 4.2). The ratio between these numbers was kept fixed at 4 as proposed in [89]. The motivation behind varying these numbers as opposed to the number of stacks/layers or the kernel size is to keep the temporal receptive field size fixed, which might otherwise result in an unfair comparison. Frequency bins were treated as channels in the context of the TCN architecture, resulting in the following input-to-output mapping: $\mathbb{R}^{B\times 3\,\mathrm{F}\times\mathrm{T}} \rightarrow \mathbb{R}^{B\times(n_{\mathrm{param}}\,\mathrm{F})\times\mathrm{T}}$, with B the batch size, F the number of frequency bins, T the number of time frames, and n_{param} the number of parameters per frequency bin required by the specific TCM estimation procedure.

4.5 Simulation Results

In this section, we evaluate the speech enhancement performance and the computational complexity of the signal approximation-based deep temporal MVDR filters using the different proposed TCM estimation procedures proposed in Section 4.3. In Section 4.5.1, the SPP-driven temporal MVDR filter and the signal approximation-based deep temporal MVDR filters using the different proposed TCM estimation procedures are compared. in Section 4.5.2, the best-performing deep temporal MVDR filters are compared with several baseline algorithms.

4.5.1 *Comparison of Deep MFMVDR Filters*

Table 4.1 shows the speech enhancement performance on the DNS 1 evaluation dataset for the SPP-driven temporal MVDR filter and the signal approximation-based deep temporal MVDR filters using the different TCM estimation procedures (RS, CD, PDT, R1). In this section, we focus on all considered deep temporal MVDR filters. The speech enhancement performance is presented in terms of SI-SDR [141], the narrowband and wideband PESQ metrics [183], the STOI metric [218], as well as the DNSMOS metric [149], using the speech signal as the reference signal. As can be observed, all considered signal approximation-based temporal MVDR filters

3 https://github.com/naplab/Conv-TasNet

Table 4.1: Speech enhancement performance on the DNS 1 evaluation dataset, presented in terms of SI-SDR, narrowband PESQ (PESQ-NB), wideband PESQ (PESQ-WB), STOI, and DNSMOS for deep temporal MVDR filters using different TCM estimation procedures based on speech presence probability (SPP), recursive smoothing (RS), Cholesky decomposition (CD), positive-definite Toeplitz (PDT), and rank-1 (R1), the real- and complex-valued masking algorithms, the deep filter (DF) algorithm, as well as the DCCRN-MC and DCUNET-MC algorithms.

Algorithm	SI-SDR / dB	PESQ-NB	PESQ-WB	STOI	DNSMOS
Noisy	9.23	2.45	1.58	0.915	3.15
Deep MFMVDR (SPP)	10.87	2.57	1.71	0.907	3.24
Deep MFMVDR (RS)	15.03	2.97	2.24	0.942	3.45
Deep MFMVDR (CD)	**17.60**	3.28	**2.71**	**0.961**	**3.75**
Deep MFMVDR (PDT)	15.20	2.98	2.29	0.945	3.37
Deep MFMVDR (R1)	17.31	3.25	2.67	0.959	3.74
Masking (real) (4.43)	16.24	3.11	2.44	0.949	3.57
Masking (complex) (4.44)	17.25	3.20	2.59	0.956	3.74
DF (4.45)	17.37	3.20	2.61	0.955	3.73
DCCRN-MC [217]	16.50	3.21	—	0.951	—
DCUNET-MC [217]	17.46	**3.30**	—	**0.961**	—

yield a significant improvement in terms of all performance metrics. Comparing the SPP-driven temporal MVDR filter and the signal approximation-based deep temporal MVDR filter using the recursive smoothing (RS) TCM estimation procedure, which both utilize a variant of recursive smoothing, a clear benefit of defining the loss function on the signal level can be observed. A possible explanation is the fact that the SPP-driven temporal MVDR filter requires a number of smoothing parameters to be chosen by hand, i.e., α_n^{SPP} for the noise TCM in (4.10) (also affecting the target used in the loss function), λ_y^{SPP} for the noisy TCM in (4.11), and λ_{DDA} for the DDA in (4.16). The combination of these smoothing parameters critically affects the resulting speech enhancement performance, and different parameter choices may be more suitable for different acoustic scenarios. In contrast, in the signal approximation-based deep temporal MVDR filter approach the required smoothing parameters, i.e., $\lambda_{y,l}^{\text{RS}}$ for the noisy TCM and $\lambda_{i,l}^{\text{RS}}$ for the interference TCM in (4.19) are determined by the TCNs. Relating these results to the classification of hybrid speech enhancement approaches in Section 1.2.3, these results demonstrate a performance benefit of employing a *coupled* hybrid approach instead of a *decoupled* hybrid approach.

Comparing the signal approximation-based deep temporal MVDR filters, it can be observed that the CD TCM estimation procedure consistently yields the highest improvement in terms of all performance metrics, closely followed by the R1 TCM

Table 4.2: Network size, presented in terms of trainable weights, bottleneck dimension size, and computational complexity, presented in terms of the relative transfer function (RTF), the contribution of the temporal MVDR linear algebra operations to the RF, the number of FLOPS, and the number of estimated parameters per time frame.

Algorithm	Trainable Weights / M	Bottleneck Dimension	RTF	RTF MFMVDR / %	FLOPS / M	Estimated Parameters
Deep MFMVDR (SPP)	5.3	231	0.176	54.9	22.9	65
Deep MFMVDR (RS)	4.9	128	0.167	47.9	26.9	130
Deep MFMVDR (CD)	5.3	128	0.139	39.0	46.2	3250
Deep MFMVDR (PDT)	5.1	128	0.170	43.4	65.4	1300
Deep MFMVDR (R1)	5.1	128	0.100	7.5	32.5	1300
Masking (real)	5.0	226	0.075	0.0	16.3	65
Masking (complex)	5.0	226	0.077	0.0	16.4	130
DF	5.2	226	0.079	0.0	17.1	650
DCCRN-MC [217]	20.0	—	—	—	—	—
DCUNET-MC [217]	3.5	—	—	—	—	—

estimation procedure. The RS and PDT TCM estimation procedures yield the lowest improvements.

As discussed in Section 4.1, the CD TCM estimation procedure merely imposes a Hermitian positive-definite structure on the estimated TCMs, i.e., it actually does not restrict the estimated TCMs more than mathematically required. While the RS, PDT, and R1 TCM estimation procedures also yield Hermitian positive-definite TCMs, they impose further structure. More specifically, from (4.19) it can be seen that the RS TCM estimation procedure imposes rank-1 updates using the noisy vector. By imposing a Toeplitz structure, the PDT TCM estimation procedure imposes stationarity over N frames on the noisy and interference components, leading to a significant reduction in speech enhancement performance compared with the CD TCM estimation procedure. Hence, a Toeplitz structure does not seem to be a viable choice for modeling the noisy and interference TCMs. Interestingly, the R1 TCM estimation procedure, which imposes a dominant principal subspace on the noisy and interference TCMs, yields quite a high speech enhancement performance.

For the deep temporal MVDR filters, Table 4.2 shows the network size in terms of trainable weights, bottleneck dimension size, and the computational complexity in terms of the RF, defined as the ratio between processing duration and signal duration, as well as the contribution of the temporal MVDR linear algebra operations to the RF, the number of FLOPS, and the number of estimated parameters. Any operation between obtaining the TCN outputs and applying the temporal filter to the noisy vector is counted as a linear algebra operation. All metrics were computed using the PyTorch profiler for the DNS 1 evaluation dataset, i.e., 100 signals of length 10 s, using a single core of an AMD EPYC 7443P CPU clocked at 3.8 GHz. First, it can be observed that all deep temporal MVDR filters exhibit an RF that is

significantly smaller than 1. Second, the signal approximation-based deep temporal MVDR filter using the R1 estimation procedure exhibits a significantly smaller RF than the other deep temporal MVDR filters. This can be explained by the fact that the SPP-driven, RS, CD and PDT estimation procedures require relatively computationally complex operations to construct the TCM estimates from the TCN outputs (see, e.g., Algorithms 1 and 2) and require a matrix inversion to compute the temporal MVDR filter in (3.26). Hence, as can be observed in Table 4.2, the RF is primarily determined by the temporal MVDR linear algebra operations and not by the number of parameters that need to be estimated by the TCNs.

Relating the speech enhancement performance in Table 4.1 and the computational complexity in Table 4.2, a benefit of the proposed R1 TCM estimation procedure can be identified, since it yields a speech enhancement performance that is only slightly lower than the CD TCM estimation procedure, while reducing the RF and the number of FLOPS by approximately 30 %.

4.5.2 *Comparison with Baseline Algorithms*

In this section, we focus on the best-performing deep temporal MVDR filters, i.e., the signal approximation-based deep temporal MVDR filter using the CD and R1 TCM estimation procedures, and compare their speech enhancement performance (Table 4.1) and computational complexity (Table 4.2) with several baseline algorithms. As the first set of baseline algorithms, we consider the real- and complex-valued masking as well as the DF algorithms discussed in Section 4.4.2, which are based on the same TCN architecture and trained using the same procedure as the proposed deep temporal MVDR filters, thus allowing to investigate the effect of imposing the deep temporal MVDR structure. As additional state-of-the-art baseline algorithms, we consider the deep complex convolutional recurrent network (DCCRN-MC) and the deep complex U-Net (DCUNET-MC) algorithms proposed in [217]. These algorithms integrate a complex convolutional recurrent-based or a complex U-Net-based encoder-decoder structure with complex convolutional block attention modules to estimate single-frame complex-valued masks, which are applied to the noisy STFT coefficients. The DCCRN-MC and DCUNET-MC were trained using a mixed loss function, comprising an SI-SDR term as well as a term consisting of the squared complex-valued mask error. Note that the DCCRN-MC and DCUNET-MC algorithms were not retrained in the context of these simulations, and instead the official published results on the DNS 1 challenge evaluation dataset are presented, which is the reason for missing values in Table 4.1 and Table 4.2.

First, comparing the speech enhancement performance of the real- and complex-valued masking algorithms, it can be confirmed that adding the potential of phase enhancement yields an increased performance in terms of all considered metrics. Second, comparing the complex-valued masking and DF algorithms, it can be observed that the speech enhancement performance is hardly improved by increasing the number of filter coefficients from $N = 1$ to $N = 5$ (with only minor improvements in terms of SI-SDR and wideband PESQ). Third, comparing the complex-

valued masking algorithms that employ either the TCN architecture, the DCCRN architecture, or the DCUNET architecture, it can be observed that the TCN-based complex-valued masking algorithm performs slightly better than the DCCRN-MC algorithm, while both algorithms are consistently outperformed by the DCUNET-MC algorithm. Fourth, we compare the best-performing proposed deep temporal MVDR filters, i.e., using the R1 and the CD estimation procedure, with the best-performing baseline algorithms, i.e., the DF algorithm and the DCUNET-MC algorithm. The proposed R1 deep temporal MVDR filter outperforms the DF algorithm (except in terms of SI-SDR) while being outperformed by the DCUNET-MC algorithm. In contrast, the proposed CD deep temporal MVDR filter consistently outperforms the DF algorithm and outperforms the DCUNET-MC algorithm in terms of SI-SDR while yielding a similar performance in terms of narrowband PESQ and STOI. Exemplary audio examples are available online.[4]

In terms of computational complexity, it can be observed in Table 4.2 that the RF of the real- and complex-valued masking algorithms as well as the DF algorithm is lower than the RF of the R1 and CD deep temporal MVDR filters, due to the absence of temporal MVDR linear algebra operations. Hence, for the considered algorithms employing a TCN architecture, a trade-off exists between speech enhancement performance and computational complexity.

4.6 Summary

In this chapter, we proposed to embed the temporal MVDR filter for single-microphone speech enhancement within a deep learning framework. We proposed to estimate the required quantities of the temporal MVDR filter, i.e., the noisy and interference TCMs as well as the a-priori SNR, using a signal approximation loss function defined at the output of the temporal MVDR filter. For the TCMs, we investigated different matrix structures, namely Hermitian positive-definite, Hermitian positive-definite Toeplitz (assuming stationarity of the noisy and interference components over N frames) and rank-1 (assuming a dominant principal subspace). For the Hermitian positive-definite matrix structure, we proposed an estimation procedure based on recursive smoothing, requiring one real-valued parameter for each TCM, and an estimation procedure based on the Cholesky decomposition, requiring N^2 real-valued parameters for each TCM. For the Hermitian positive-definite Toeplitz matrix structure, we proposed an estimation procedure that involves balanced Vandermonde matrices, requiring $2N$ real-valued parameters for each TCM. For the rank-1 matrix structure, we showed that the temporal MVDR filter can be written as a linear combination of the DNN outputs, circumventing computationally complex matrix inversions, and proposed an estimation procedure requiring $2N$ real-valued parameters for each TCM.

4 https://uol.de/en/sigproc/research/audio-demos/multi-frame-speech-enhancement/deep-mfmvdr-journal

Experimental results on the DNS 1 challenge dataset demonstrate that the coupled hybrid approach of estimating the TCMs using a signal approximation loss function defined at the output of the temporal MVDR filter yields a substantial improvement in speech enhancement performance compared with the decoupled hybrid approach of estimating the TCMs using an SPP-driven loss function. Comparing the signal approximation-based estimation procedures, the TCM estimation procedure based on the Cholesky decomposition yields the best speech enhancement performance, closely followed by the computationally less complex rank-1 TCM estimation procedure. The TCM estimation procedures based on recursive smoothing and the positive-definite Toeplitz matrix structure yield the lowest speech enhancement performance, hinting that updating the TCMs based on recursive smoothing or assuming stationarity is not as appropriate. In addition, the simulation results show that the best-performing signal approximation-based deep temporal MVDR filter outperforms real- and complex-valued masking as well as directly estimating the temporal filter coefficients, demonstrating the benefit of imposing structure on the temporal filters in a hybrid approach as opposed to a purely learning-based approach. Finally, a competitive performance compared with state-of-the-art algorithms is demonstrated.

5

DEEP MULTI-FRAME FILTER FOR BINAURAL SPEECH ENHANCEMENT

In Chapter 4, we proposed a coupled, structured estimation hybrid approach for single-microphone speech enhancement, imposing structure on the interference temporal covariance matrix. In this chapter, we extend this approach to binaural speech enhancement by embedding the binaural spatio-temporal Wiener filter (STWF) within an end-to-end deep learning framework, imposing structure not only on the interference spatio-temporal covariance matrices (STCMs), but also on the speech spatio-temporal correlation vectors (STCVs). Specifically, we consider the decomposition of the binaural STWF into a binaural spatio-temporal MVDR filter and a spectral postfilter, hence requiring estimates of the speech STCVs, the speech PSDs, and the inverse interference STCMs. Please note that in principle these quantities need to be estimated both for the filter estimating the target speech component at the left device as well as for the filter estimating the target speech component at the right device. In addition to imposing a Hermitian positive-definite structure on the inverse interference STCMs, the main focus of this chapter is to investigate the potential of imposing spatio-temporal correlation structure on the speech STCVs and the inverse interference STCMs. We propose several procedures which mainly differ in terms of the relation between the microphones, particularly between the left and the right hearing device, and the number of parameters that need to be estimated. First, assuming that the spatial correlation of the speech component is stationary over the length of the temporal filter, the speech STCVs can be decomposed as the Kronecker product of a relative transfer function vector and a temporal correlation vector. We either consider a single "global" reference microphone, requiring

This chapter is partly based on:

[177] M. Tammen and S. Doclo, "Deep Multi-Frame MVDR Filtering for Binaural Noise Reduction," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022, pp. 1–5.

[178] M. Tammen and S. Doclo, "Imposing Correlation Structures for Deep Binaural Spatio-Temporal Wiener Filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1278–1292, Mar. 2025.

the speech temporal correlation vector to be estimated only for this microphone, or a reference microphone for each hearing device, requiring (left and right) speech temporal correlation vectors to be estimated for both reference microphones. The STCV structure considering two reference microphones involves more parameters than the STCV structure considering a single reference microphone, but it allows for more degrees of freedom. Second, we propose to replace the left and right interference STCMs by a common interference STCM, as the difference between both STCMs can be assumed to be negligible. In addition, we consider a bilateral STWF by assuming no spatio-temporal correlation between both hearing devices, both for the speech STCVs and for the interference STCM. To train and evaluate the deep bilateral STWF and the deep binaural STWF using the proposed spatio-temporal correlation structures, we constructed matched datasets using diverse speech and noise sources from the DNS 1 and DNS 2 challenges [66], [68] as well as simulated binaural room impulse responses from the CEC 1 [69]. In addition, to evaluate the generalization capabilities of the considered algorithms, we considered a mismatched evaluation dataset from CEC 3 that comprises noise backgrounds and RIRs recorded in complex environments as well as simulated head rotation. Simulation results show that the binaural STWF using a combination of the speech STCV structure considering two reference microphones and a common interference STCM significantly reduces the computational complexity while yielding a similar speech enhancement and binaural cue preservation performance compared to not imposing any spatio-temporal correlation structure. Furthermore, simulation results demonstrate that this deep binaural STWF outperforms the deep bilateral STWF as well as two state-of-the-art binaural speech enhancement algorithms, namely the deep filter algorithm [129] (which directly estimates the binaural temporal filter coefficients using a purely learning-based approach) and the binaural Conv-TasNet algorithm [189], while approaching the performance of the non-causal BCCTN algorithm [188].

The remainder of this chapter is organized as follows. In Section 5.2, we propose several spatio-temporal correlation structures for the speech STCVs and the inverse interference STCMs. In Section 5.3, we describe the signal approximation-based approach to estimate the required quantities, taking into account these correlation structures. The simulation setup, the approach to validate the proposed correlation structures, and the simulation results for the proposed deep binaural and bilateral STWFs as well as for several baseline algorithms are presented and discussed in Sections Section 5.4, Section 5.5, and Section 5.6.

5.1 Binaural Spatio-Temporal Wiener Filter

In this chapter, we consider an acoustic scenario with a single speech source and background noise in a reverberant room, captured using binaural hearing devices with M_L and M_R microphones on the left and right devices, respectively, totaling $M = M_L + M_R$ microphones (Fig. 2.1). We assume that all microphone signals are synchronized and transmitted (e.g., via a wireless link) between the hearing devices

without transmission delay and quantization errors. The target speech components $\widehat{x}_{t,L}$ and $\widehat{x}_{t,R}$ at both hearing devices are defined (without loss of generality) at reference microphones $L = 1$ and $R = M_L + 1$.

In [56], [59] the binaural spatial Wiener filter was proposed, which aims at minimizing the mean square error between the binaural output signals and the target speech components at both reference microphones, only considering spatial correlations (described in more detail in Section 3.1.3). In this section, we introduce a multi-frame extension of the binaural spatial Wiener filter, termed binaural spatio-temporal Wiener filter (STWF), which considers both spatial as well as temporal correlations. The binaural STWF can also be considered as a binaural extension of the (monaural) STWF proposed in [3], which is reviewed in Section 3.1.1.2. The binaural spatio-temporal filters $\mathbf{w}_t^L$ and $\mathbf{w}_t^R$ in (2.50) are computed by minimizing the cost function

$$J(\mathbf{w}_t^L, \mathbf{w}_t^R) = \mathcal{E}\left\{ \left\| \begin{matrix} x_{t,L} - \mathbf{w}_t^{L,\mathsf{H}} \mathbf{y}_t \\ x_{t,R} - \mathbf{w}_t^{R,\mathsf{H}} \mathbf{y}_t \end{matrix} \right\|_2^2 \right\}, \tag{5.1}$$

yielding

$$\mathbf{w}_t^\nu = \mathbf{\Phi}_{y,t}^{-1} \mathbf{\Phi}_{x,t} \mathbf{e}_\nu, \quad \nu \in \{L, R\}, \tag{5.2}$$

where ν indicates the left or right reference microphone. Similarly as for the monaural STWF in (3.5), using (2.53) and the fact that $\mathbf{\Phi}_{x',t}^\nu \mathbf{e}_\nu = \mathbf{0}$, it can be easily shown that both STWF vectors in (5.2) can be decomposed as a spatio-temporal MVDR filter [219] and a real-valued scalar postfilter, i.e.,

$$\boxed{\mathbf{w}_t^\nu = \underbrace{\frac{\left(\mathbf{\Phi}_{i,t}^\nu\right)^{-1} \boldsymbol{\gamma}_t^\nu}{\boldsymbol{\gamma}_t^{\nu,\mathsf{H}} \left(\mathbf{\Phi}_{i,t}^\nu\right)^{-1} \boldsymbol{\gamma}_t^\nu}}_{\text{spatio-temporal MVDR}} \underbrace{\frac{\phi_{x,t}^\nu}{\phi_{x,t}^\nu + \boldsymbol{\gamma}_t^{\nu,\mathsf{H}} \left(\mathbf{\Phi}_{i,t}^\nu\right)^{-1} \boldsymbol{\gamma}_t^\nu}}_{\text{postfilter}}} \tag{5.3}$$

with $\nu \in \{L, R\}$. The spatio-temporal MVDR filter minimizes the output interference PSD while preserving the spatio-temporal correlation of the speech component, with the postfilter providing additional noise reduction at the cost of allowing for some speech distortion. The performance of the binaural STWF strongly depends on how well the required quantities, i.e., the left and right inverse interference STCMs $(\mathbf{\Phi}_{i,t}^L)^{-1}$ and $(\mathbf{\Phi}_{i,t}^R)^{-1}$, the left and right speech STCVs $\boldsymbol{\gamma}_t^L$ and $\boldsymbol{\gamma}_t^R$, as well as the left and right speech PSDs $\phi_{x,t}^L$ and $\phi_{x,t}^R$ are estimated from the noisy STFT coefficients. To benefit from the representation capacity of DNNs for this estimation task, similarly as in Chapter 4 and [206], [219], in this chapter we propose a coupled, structured estimation hybrid speech enhancement approach, where all required quantities of the binaural STWF are estimated by embedding the fully differentiable STWF within a deep learning framework using TCNs (see Section 5.3). In addition, we investigate imposing different spatio-temporal structures on the speech STCVs and the interference STCMs, as described in the following section.

5.2 Spatio-Temporal Correlation Structures

The quantities required by the binaural STWF are determined by both temporal and spatial correlations. On the one hand, the temporal correlations of the speech and interference components can vary drastically across a small number of time frames. On the other hand, the spatial correlations of the speech and interference components mainly depend on the acoustic scene, i.e., the positions of the listener and the speech and noise sources, which can be assumed to be stationary across a small number of time frames. In this section, we propose several spatio-temporal structures for the speech STCVs and the interference STCMs, relating these quantities between the left and the right hearing device. First, assuming spatial stationarity of the speech component over a small number of time frames, in Section 5.2.1 we impose spatial structure on the speech STCVs. Second, assuming that the uncorrelated speech components are negligible, in Section 5.2.2 we set the left and the right interference STCM equal to each other. Third, in Section 5.2.3 we assume no correlation between the left and right hearing devices for both the speech and interference components. The considered structures greatly differ in the number of parameters that need to be estimated. As will be demonstrated by the simulation results in Section 5.6, imposing structure on the speech STCVs and the interference STCMs is beneficial in terms of computational complexity.

5.2.1 *Speech Correlation Vectors*

As described in Section 2.1.5, assuming the microphone with index ν as the reference microphone and further assuming that the RTFs are constant over N frames (i.e., $h^{\nu}_{t,m}$, $h^{\nu}_{t-1,m}$, …, $h^{\nu}_{t-N+1,m}$ are equal), the multi-frame speech vector in (2.17) can be written as [3], [47]

$$\bar{\mathbf{x}}_{t,m} = h^{\nu}_{t,m}\bar{\mathbf{x}}_{t,\nu}, \tag{5.4}$$

where the RTF between the reference microphone with index ν and the m-th microphone is defined as

$$h^{\nu}_{t,m} = \frac{\mathbb{E}\left(x_{t,m}x^{*}_{t,\nu}\right)}{\mathbb{E}\left(|x_{t,\nu}|^{2}\right)} = \frac{\mathbf{e}_m^{\mathsf{T}}\mathbf{\Phi}_{x,t}\mathbf{e}_{\nu}}{\phi^{\nu}_{x,t}}. \tag{5.5}$$

In the following, we will either consider a single "global" reference microphone for both hearing devices or a reference microphone for each hearing device.

5.2.1.1 *Global Relative Transfer Function*

Without loss of generality, we choose the reference microphone on the left hearing device (with index $L = 1$) as the global reference microphone. Using (5.4), the multi-microphone multi-frame speech vector in (2.29) can then be modeled as

$$\mathbf{x}_t = \begin{bmatrix} \bar{\mathbf{x}}_{t,1} \\ \bar{\mathbf{x}}_{t,2} \\ \vdots \\ \bar{\mathbf{x}}_{t,2M} \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{x}}_{t,1} \\ h^1_{t,2}\bar{\mathbf{x}}_{t,1} \\ \vdots \\ h^1_{t,2M}\bar{\mathbf{x}}_{t,1} \end{bmatrix} = \check{\mathbf{h}}^{\text{glob}}_t \otimes \bar{\mathbf{x}}_{t,1}, \tag{5.6}$$

where the global RTF vector $\check{\mathbf{h}}^{\text{glob}}_t$ contains the RTFs between all microphones and the global reference microphone, i.e.,

$$\check{\mathbf{h}}^{\text{glob}}_t = \begin{bmatrix} 1 & h^1_{t,2} & \dots & h^1_{t,2M} \end{bmatrix}^{\mathsf{T}} \in \mathbb{C}^{2M}, \tag{5.7}$$

and $\otimes$ denotes the Kronecker product. Using (5.6), the left and right speech STCVs in (2.34) can be written as

$$\boldsymbol{\gamma}^{\text{glob},\nu}_t = \frac{\mathcal{E}\left\{\check{\mathbf{h}}^{\text{glob}}_t \otimes \bar{\mathbf{x}}_{t,1}x^*_{t,\nu}\right\}}{\phi^{\nu}_{x,t}} = \check{\mathbf{h}}^{\text{glob}}_t \otimes \bar{\boldsymbol{\gamma}}^{\nu}_{t,1}, \tag{5.8}$$

with $\nu \in \{L, R\}$, and where the speech temporal correlation vector $\bar{\boldsymbol{\gamma}}^{\nu}_{t,1}$ describes the correlation between the N most recent speech STFT coefficients at the global reference microphone (with index $L = 1$) and the current target speech STFT coefficient ($x_{t,L}$ or $x_{t,R}$), i.e.,

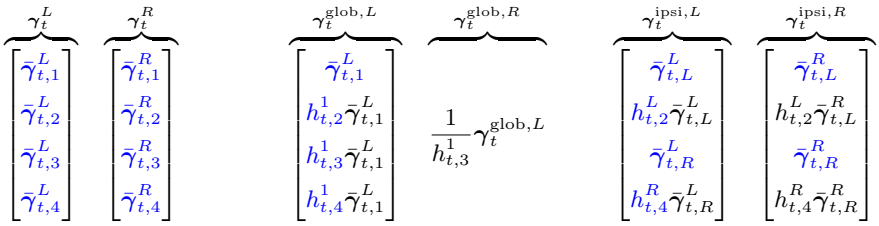
$$\bar{\boldsymbol{\gamma}}^{\nu}_{t,1} = \frac{\mathcal{E}\left\{\bar{\mathbf{x}}_{t,1}x^*_{t,\nu}\right\}}{\phi^{\nu}_{x,t}} \in \mathbb{C}^N, \quad \nu \in \{L, R\}. \tag{5.9}$$

When imposing the global RTF structure on the speech STCVs in (5.8), the speech STCVs can be interpreted as being decomposed into a spatial factor ($\check{\mathbf{h}}^{\text{glob}}_t$) and a temporal factor ($\bar{\boldsymbol{\gamma}}^{\nu}_{t,1}$). Furthermore, for two of the quantities to be estimated, the global RTF structure yields an explicit relation between the left and the right hearing device: First, since the reference microphones on both hearing devices are related as $x_{t,R} = h^1_{t,R}x_{t,L}$, the left and right speech PSDs are related as

$$\phi^{R}_{t,\text{glob}} = |h^1_{t,R}|^2\phi^{\text{glob},L}_{x,t}, \tag{5.10}$$

where the global RTF $h^1_{t,R}$ is an element of $\check{\mathbf{h}}^{\text{glob}}_t$ in (5.7). Second, the left and right speech STCVs in (5.8) are related as

$$\boldsymbol{\gamma}^{\text{glob},R}_t = \frac{\mathcal{E}\left\{\mathbf{x}_t x^*_{t,R}\right\}}{\mathcal{E}\left\{|x_{t,R}|^2\right\}} = \frac{1}{h^1_{t,R}}\frac{\mathcal{E}\left\{\mathbf{x}_t x^*_{t,L}\right\}}{\mathcal{E}\left\{|x_{t,L}|^2\right\}} = \frac{1}{h^1_{t,R}}\boldsymbol{\gamma}^{\text{glob},L}_t, \tag{5.11}$$



(a) no structure (b) global RTF structure (c) ipsilateral RTF structure

Figure 5.1: Illustration of the proposed spatial structures imposed on the speech STCVs, assuming $M_L = M_R = 2$ microphones per hearing device (with reference indices $L = 1$ and $R = 3$). The parameters to be estimated are highlighted in blue once per structure in order to emphasize the parameter reuse achieved by the global RTF structure and the ipsilateral RTF structure.

i.e., the left and right speech STCVs are related by a complex-valued scalar (so they are parallel). It should be realized that—although both vectors may differ in amplitude and phase—the relation between microphones is the same for both hearing devices. The global RTF structure is visualized in Fig. 5.1b, assuming $M_L = M_R = 2$ microphones per hearing device.

The model in (5.6) assumes fully correlated speech components between the global reference microphone and all microphones of both hearing devices, which is a common assumption in binaural speech enhancement algorithms [24], [59]. However, depending on the STFT frame length, this assumption may be violated in practice due to large inter-microphone distances and reverberation, especially when considering the correlation between the global reference microphone and the contralateral microphones. This motivates the investigation of an alternative structure in the next section.

5.2.1.2 *Ipsilateral Relative Transfer Function*

As a less restrictive alternative to the global RTF structure in (5.6), we propose to use this model for each hearing device independently, i.e., using the microphone with index L as the reference for the microphones of the left hearing device and the microphone with index R as the reference for the microphones of the right hearing device. The multi-microphone multi-frame speech vector can then be modeled as

$$\mathbf{x}_t = \begin{bmatrix} \check{\mathbf{h}}_t^{\text{ipsi},L} \otimes \bar{\mathbf{x}}_{t,L} \\ \check{\mathbf{h}}_t^{\text{ipsi},R} \otimes \bar{\mathbf{x}}_{t,R} \end{bmatrix}, \tag{5.12}$$

where the ipsilateral RTF vectors, defined as

$$\check{\mathbf{h}}_t^{\text{ipsi},L} = \begin{bmatrix} 1 & h_{t,2}^L & \dots & h_{t,M_L}^L \end{bmatrix}^{\mathsf{T}} \in \mathbb{C}^M \tag{5.13}$$

$$\check{\mathbf{h}}_t^{\text{ipsi},R} = \begin{bmatrix} 1 & h_{t,M_L+2}^R & \dots & h_{t,M}^R \end{bmatrix}^{\mathsf{T}} \in \mathbb{C}^M, \tag{5.14}$$

relate the speech component at the microphones of the left hearing device to the left reference microphone and the speech component at the microphones of the right hearing device to the right reference microphone. Using (5.12) in (2.51), the left and right speech STCVs can be written as

$$\boldsymbol{\gamma}_{t,\text{ipsi}}^{\nu} = \frac{1}{\phi_{x,t}^{\nu}} \mathcal{E}\left\{ \begin{bmatrix} \check{\mathbf{h}}_t^{\text{ipsi},L} \otimes \bar{\mathbf{x}}_{t,L} \\ \check{\mathbf{h}}_t^{\text{ipsi},R} \otimes \bar{\mathbf{x}}_{t,R} \end{bmatrix} x_{t,\nu}^* \right\} = \begin{bmatrix} \check{\mathbf{h}}_t^{\text{ipsi},L} \otimes \bar{\boldsymbol{\gamma}}_{t,L}^{\nu} \\ \check{\mathbf{h}}_t^{\text{ipsi},R} \otimes \bar{\boldsymbol{\gamma}}_{t,R}^{\nu} \end{bmatrix}, \tag{5.15}$$

with $\nu \in \{L, R\}$, and where the speech temporal correlation vectors $\bar{\boldsymbol{\gamma}}_{t,L}^{\nu}$ and $\bar{\boldsymbol{\gamma}}_{t,R}^{\nu}$ are defined similarly as in (5.9), i.e.,

$$\bar{\boldsymbol{\gamma}}_{t,L}^{\nu} = \frac{\mathcal{E}\left\{\bar{\mathbf{x}}_{t,L} x_{t,\nu}^*\right\}}{\phi_{x,t}^{\nu}}, \quad \bar{\boldsymbol{\gamma}}_{t,R}^{\nu} = \frac{\mathcal{E}\left\{\bar{\mathbf{x}}_{t,R} x_{t,\nu}^*\right\}}{\phi_{x,t}^{\nu}}. \tag{5.16}$$

It should be noted that the ipsilateral RTF structure comprises four speech temporal correlation vectors, whereas the global RTF structure only comprises one speech temporal correlation vector in (5.9). The ipsilateral RTF structure is visualized in Fig. 5.1c.

5.2.2 *Interference Covariance Matrices*

The left and right interference STCMs are defined as

$$\boldsymbol{\Phi}_{i,t}^{\nu} = \boldsymbol{\Phi}_{x',t}^{\nu} + \boldsymbol{\Phi}_{n,t}, \quad \nu \in \{L, R\}. \tag{5.17}$$

Assuming the uncorrelated speech STCMs $\boldsymbol{\Phi}_{x',t}^{\nu}$ to be negligible compared to the noise STCM $\boldsymbol{\Phi}_{n,t}$, the left and right interference STCMs $\boldsymbol{\Phi}_{i,t}^{L}$ and $\boldsymbol{\Phi}_{i,t}^{R}$ can be replaced by a common STCM $\boldsymbol{\Phi}_{i,t}$, i.e.,

$$\boldsymbol{\Phi}_{i,t}^{L} = \boldsymbol{\Phi}_{i,t}^{R} =: \boldsymbol{\Phi}_{i,t}. \tag{5.18}$$

This assumption is generally more valid at lower SNRs, where the noise STCM becomes more dominant relative to the uncorrelated speech STCMs. As will be demonstrated in Section 5.5, this assumption holds quite well in practice. It should be noted that when combining this assumption with the global RTF structure—where the left and right speech STCVs are parallel—the resulting binaural STWF filter vectors are parallel as well.

5.2.3 *Bilateral Correlation*

The binaural STWF using the proposed speech STCV and interference STCM structures in Sections Section 5.2.1 and Section 5.2.2 exploits spatio-temporal correlation

between both hearing devices, requiring the microphone signals to be transmitted between the left and right hearing devices. In order to investigate the performance benefit achieved by binaural processing, we will also consider bilateral processing, where both hearing devices operate independently. This corresponds to assuming no correlation between the left and right hearing devices for both the speech and the interference components such that the left and right speech STCVs can be modeled using a non-zero subvector for ipsilateral correlations and a zero subvector for contralateral correlations, i.e.,

$$\boldsymbol{\gamma}_{t,\text{bil}}^{L} = \begin{bmatrix} \bar{\boldsymbol{\gamma}}_{t,1}^{L} \\ \vdots \\ \bar{\boldsymbol{\gamma}}_{t,M_L}^{L} \\ \mathbf{0}_{M_R N \times 1} \end{bmatrix}, \quad \boldsymbol{\gamma}_{t,\text{bil}}^{R} = \begin{bmatrix} \mathbf{0}_{M_L N \times 1} \\ \bar{\boldsymbol{\gamma}}_{t,M_L+1}^{R} \\ \vdots \\ \bar{\boldsymbol{\gamma}}_{t,M}^{R} \end{bmatrix}, \tag{5.19}$$

and, similarly, the common interference STCM in (5.18) can be modeled using non-zero submatrices for ipsilateral correlations and zero submatrices for the contralateral correlations, i.e.,

$$\boldsymbol{\Phi}_{i,t,\text{bil}} = \begin{bmatrix} \tilde{\boldsymbol{\Phi}}_{i,t}^{LL} & \mathbf{0}_{M_L N \times M_R N} \\ \mathbf{0}_{M_R N \times M_L N} & \tilde{\boldsymbol{\Phi}}_{i,t}^{RR} \end{bmatrix}. \tag{5.20}$$

Since

$$\boldsymbol{\Phi}_{i,t,\text{bil}}^{-1} \boldsymbol{\gamma}_{t,\text{bil}}^{L} = \begin{bmatrix} \left(\tilde{\boldsymbol{\Phi}}_{i,t}^{LL}\right)^{-1} [\boldsymbol{\gamma}_{t}^{L}]_{1:M_L N} \\ \mathbf{0}_{M_R N \times 1} \end{bmatrix} \tag{5.21}$$

$$\boldsymbol{\Phi}_{i,t,\text{bil}}^{-1} \boldsymbol{\gamma}_{t,\text{bil}}^{R} = \begin{bmatrix} \mathbf{0}_{M_L N \times 1} \\ \left(\tilde{\boldsymbol{\Phi}}_{i,t}^{RR}\right)^{-1} [\boldsymbol{\gamma}_{t}^{R}]_{M_L N+1:MN} \end{bmatrix}, \tag{5.22}$$

the binaural STWF in (5.3) reduces to a set of bilateral filters, i.e.,

$$\mathbf{w}_{t}^{\text{bilat},L} = \begin{bmatrix} [\mathbf{w}_{t}^{L}]_{1:M_L N} \\ \mathbf{0}_{M_R N \times 1} \end{bmatrix}, \ \mathbf{w}_{t}^{\text{bilat},R} = \begin{bmatrix} \mathbf{0}_{M_L N \times 1} \\ [\mathbf{w}_{t}^{R}]_{M_L N+1:MN} \end{bmatrix}, \tag{5.23}$$

where $\mathbf{w}_{t}^{\text{bilat},L}$ only depends on quantities related to the left hearing device and $\mathbf{w}_{t}^{\text{bilat},R}$ only depends on quantities related to the right hearing device.

For the bilateral STWF, it is also possible to consider the ipsilateral RTF structure for the speech STCV in (5.19), i.e.,

$$\boldsymbol{\gamma}_{t}^{\text{bil,ipsi},L} = \begin{bmatrix} \check{\mathbf{h}}_{t}^{\text{ipsi},L} \otimes \bar{\boldsymbol{\gamma}}_{t,L}^{L} \\ \mathbf{0}_{M_R N \times 1} \end{bmatrix}, \ \boldsymbol{\gamma}_{t}^{\text{bil,ipsi},R} = \begin{bmatrix} \mathbf{0}_{M_L N \times 1} \\ \check{\mathbf{h}}_{t}^{\text{ipsi},R} \otimes \bar{\boldsymbol{\gamma}}_{t,R}^{R} \end{bmatrix}, \tag{5.24}$$

with $\check{\mathbf{h}}_{t}^{\text{ipsi},L}$ and $\check{\mathbf{h}}_{t}^{\text{ipsi},R}$ defined in (5.13) and (5.14). As will be demonstrated in Section 5.5, these bilateral structures introduce relatively large estimation errors in practice (see also Fig. 2.2), suggesting that the correlation between the left and right hearing devices for both the speech and the interference components is quite relevant.

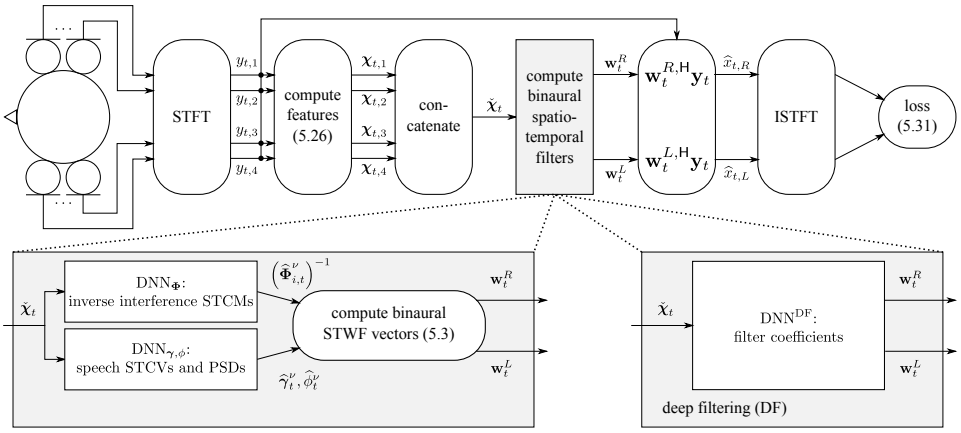


Figure 5.2: Block diagram of the proposed deep binaural STWF and the baseline deep filter algorithm, assuming $M_L = M_R = 2$ microphones per hearing device.

5.3 Deep Binaural Spatio-Temporal Wiener Filter

To estimate all required quantities, the binaural STWF is embedded into a supervised learning framework (see Fig. 5.2). Specifically, the speech STCVs $\boldsymbol{\gamma}_t^L$ and $\boldsymbol{\gamma}_t^R$, the speech PSDs $\phi_{x,t}^L$ and $\phi_{x,t}^R$, as well as the inverse STCMs $(\mathbf{\Phi}_{i,t}^L)^{-1}$ and $(\mathbf{\Phi}_{i,t}^R)^{-1}$ are estimated using TCNs. Note that this is different from Chapter 4, where we estimated the speech TCV in (4.8) indirectly using estimates of the noisy and interference TCMs as well as the a-priori SNR. Separate TCNs are used for the quantities related to the speech component (i.e., STCVs and PSDs) and for the quantities related to the interference component (i.e., inverse STCMs), where both TCNs are jointly trained using a loss function that compares the ground-truth binaural speech components to the estimated components obtained at the output of the deep binaural STWF. As input features, similarly as in Chapter 4, we utilized the concatenation of the logarithmic magnitude, as well as the cosine and sine of the phase, of the noisy STFT coefficients at all frequency bins and microphones[0], i.e.,

$$\boldsymbol{\chi}_{f,t,m} = \Big[\log_{10}(\big|y_{f,t,m}\big| + \epsilon) \quad \cos\big(\angle y_{f,t,m}\big) \quad \sin\big(\angle y_{f,t,m}\big)\Big]^{\mathsf{T}} \in \mathbb{R}^3 \tag{5.25}$$

$$\boldsymbol{\chi}_{t,m} = \Big[\boldsymbol{\chi}_{f=1,t,m}^{\mathsf{T}} \quad \boldsymbol{\chi}_{f=2,t,m}^{\mathsf{T}} \quad \cdots \quad \boldsymbol{\chi}_{f=F,t,m}^{\mathsf{T}}\Big]^{\mathsf{T}} \in \mathbb{R}^{3F}, \tag{5.26}$$

$$\boldsymbol{\chi}_{t} = \Big[\boldsymbol{\chi}_{t,m=1}^{\mathsf{T}} \quad \boldsymbol{\chi}_{t,m=2}^{\mathsf{T}} \quad \cdots \quad \boldsymbol{\chi}_{t,m=M}^{\mathsf{T}}\Big]^{\mathsf{T}} \in \mathbb{R}^{3FM}. \tag{5.27}$$

We chose both the cosine and the sine of the phase to obtain an unambiguous and smooth phase representation [83]. For more details about the TCNs and the loss

0 In preliminary experiments, this feature choice outperformed the use of the real and imaginary parts of the STFT coefficients as input features.

Table 5.1: The number of required parameters per frequency bin to estimate the speech STCVs and the inverse interference STCMs for the proposed spatio-temporal correlation structures (assuming $N = 5$ and $M_L = M_R = 2$, i.e., $M = 4$; $\nu \in \{L, R\}$), as well as the model mismatch on the evaluation dataset in terms of the mean relative ℓ_2 norm ϵ_{ℓ_2}, the mean Hermitian angle ϵ_θ, the mean relative Frobenius norm ϵ_{Fro}, and the mean correlation matrix distance ϵ_{CMD} in (5.42).

quantity	structure	required parameters	ϵ_{ℓ_2}/dB (↓)	$\epsilon_\theta/^\circ$ (↓)	ϵ_{Fro}/dB (↓)	ϵ_{CMD} (↓)
$\boldsymbol{\gamma}_t^\nu$	—	$4(MN-1) \triangleq$ 76	$-\infty$	0.0	—	—
	global RTF	$2(M+N-2) \triangleq$ **14**	−4.5	30.1	—	—
	ipsilateral RTF	$2(M+4N-4) \triangleq$ 40	**−15.6**	**9.9**	—	—
	bilateral	$2(MN-2) \triangleq$ 36	−3.2	43.7	—	—
	bilateral & ipsilateral RTF	$2(M+2N-4) \triangleq$ 20	−3.0	44.9	—	—
$\boldsymbol{\Phi}_{i,t}^\nu$	—	$2(MN)^2 \triangleq$ 800	—	—	$-\infty$	0.000
	common STCM ($\boldsymbol{\Phi}_{i,t}^L = \boldsymbol{\Phi}_{i,t}^R$)	$(MN)^2 \triangleq$ 400	—	—	**−20.2**	**0.004**
	bilateral	$1/2(MN)^2 \triangleq$ **200**	—	—	−4.0	0.220

function, we refer to Section 5.4.3. For the spatio-temporal correlation structures proposed in Section 5.2, in the following we explain in more detail the parameters that are estimated by the TCNs. Table 5.1 provides an overview of the number of parameters per time-frequency bin for the speech STCVs and the inverse interference STCMs, which greatly differ between the different structures.

5.3.1 *Speech Correlation Vectors and Power Spectral Densities*

NO STRUCTURE When not imposing any structure on the speech STCVs, estimates of both complex-valued vectors $\boldsymbol{\gamma}_t^L$ and $\boldsymbol{\gamma}_t^R$ and both PSDs $\phi_{x,t}^L$ and $\phi_{x,t}^R$ are required. Since one element of each speech STCV is equal to 1 (cf. (2.38)) each speech STCV is determined by $2(2MN-1)$ real-valued parameters. Similarly to [220], the speech PSDs are estimated by applying real-valued masks $m_{t,L}^{\mathbb{R}}$ and $m_{t,R}^{\mathbb{R}}$ to the noisy STFT coefficients at the reference microphones, i.e.,

$$\widehat{\phi}_{x,t}^\nu = |m_{t,\nu}^{\mathbb{R}} y_{t,\nu}|^2, \quad m_{t,\nu}^{\mathbb{R}} \in [0,1], \quad \nu \in \{L, R\}, \tag{5.28}$$

with each mask determined by a single real-valued parameter, ensuring the range $[0, 1]$ with a sigmoid activation function. A single TCN uses the features in (5.25) at all microphones to estimate the undetermined parameters of the speech STCVs and the PSD masks.

GLOBAL RTF STRUCTURE When imposing the global RTF structure on the speech STCVs, estimates of the global RTF vector $\tilde{\mathbf{h}}_t^{\text{glob}}$ in (5.7) and the speech temporal correlation vector $\bar{\boldsymbol{\gamma}}_{t,1}^L$ in (5.9) are required, separating the estimation process into a spatial factor and a temporal factor. The global RTF vector is determined by $2(M-1)$ real-valued parameters, while the speech temporal correlation vector

is determined by $2(N-1)$ real-valued parameters. Since the left and right speech PSDs are directly related by the squared magnitude of the global RTF (cf. (5.10)), only one PSD mask needs to be estimated for the global reference microphone. A single TCN uses the features in (5.25) at all microphones to estimate the undetermined parameters of the global RTF vector, the speech temporal correlation vector, and the PSD mask.

IPSILATERAL RTF STRUCTURE When imposing the ipsilateral RTF structure on the speech STCVs, estimates of the ipsilateral RTF vectors $\check{\mathbf{h}}_{t,L}^{\text{ipsi}}$ and $\check{\mathbf{h}}_{t,R}^{\text{ipsi}}$ in (5.13) and (5.14) as well as the speech temporal correlation vectors $\bar{\boldsymbol{\gamma}}_{t,L}^{L}$, $\bar{\boldsymbol{\gamma}}_{t,R}^{R}$, $\bar{\boldsymbol{\gamma}}_{t,L}^{R}$ and $\bar{\boldsymbol{\gamma}}_{t,R}^{L}$ in (5.16) are required. The vectors $\check{\mathbf{h}}_{t,L}^{\text{ipsi}}$ and $\check{\mathbf{h}}_{t,R}^{\text{ipsi}}$ are determined by $2(M_L-1)$ and $2(M_R-1)$ real-valued parameters, the vectors $\bar{\boldsymbol{\gamma}}_{t,L}^{L}$ and $\bar{\boldsymbol{\gamma}}_{t,R}^{R}$ are determined by $2(N-1)$ real-valued parameters each, and the vectors $\bar{\boldsymbol{\gamma}}_{t,L}^{R}$ and $\bar{\boldsymbol{\gamma}}_{t,R}^{L}$ are determined by $2N$ real-valued parameters each (since none of the elements needs to be equal to 1). The ipsilateral RTF structure does not impose an explicit relationship between the left and right PSDs. A single TCN uses the features in (5.25) at all microphones to estimate the undetermined parameters of the ipsilateral RTF vectors, the speech TCVs, and the PSD masks.

5.3.2 *Interference Covariance Matrices*

NO STRUCTURE In Chapter 4, we showed that, among Hermitian positive-definite, Hermitian positive-definite Toeplitz, and rank-1 TCM structures, the Hermitian positive-definite structure based on the Cholesky decomposition resulted in the highest speech enhancement performance. Assuming that the interference STCMs are full-rank, such that they are positive-definite, also their inverses are positive-definite, i.e., they can also be decomposed using the Cholesky decomposition [206], [207] as

$$\left(\boldsymbol{\Phi}_{i,t}^{\nu}\right)^{-1} = \mathbf{L}_{i,t}^{\nu}\mathbf{L}_{i,t}^{\nu,\mathsf{H}}, \quad \nu \in \{L,R\}, \tag{5.29}$$

where the Cholesky factor $\mathbf{L}_{i,t}^{\nu} \in \mathbb{C}^{MN\times MN}$ is a lower-triangular matrix with real-valued and positive diagonal elements, determined by $(MN)^2$ real-valued parameters. In this chapter, we utilize the Cholesky decomposition of the *inverse* interference STCM in (5.29) rather than the Cholesky decomposition of the interference STCM in (4.1). In preliminary experiments, this choice improved both speech enhancement performance (as no explicit matrix regularization with manually chosen regularization constant was required) and computational complexity (because no explicit matrix inverse was computed). A single TCN uses the features in (5.25) at all microphones to estimate these parameters. Similarly as in Chapter 4, we then construct the Cholesky factors $\widehat{\mathbf{L}}_{i,t}^{\nu}$ by using disjoint subsets of the parameters for the real strictly lower triangular part, the imaginary strictly lower triangular part, and the real positive diagonal part, ensuring positivity of the diagonal part with a softplus activation function. Finally, we construct the inverse interference STCMs

as in (5.29), i.e., without explicitly computing a computationally complex matrix inverse.

COMMON INTERFERENCE COVARIANCE MATRIX When assuming the left and right interference STCMs to be equal, a single Cholesky factor and inverse interference STCM is estimated using the procedure described above, reducing the number of parameters by half to $(MN)^2$.

5.3.3 *Bilateral Correlation*

When imposing a bilateral structure on the speech STCVs and the interference STCMs, only the ipsilateral correlations need to be estimated. The left and right ipsilateral speech STCVs in (5.19) are determined by $2(M_L N-1)$ and $2(M_R N-1)$ real-valued parameters, respectively. When additionally imposing the ipsilateral RTF structure, the speech STCVs in (5.24) are determined by $2(M_L-1)$ and $2(M_R-1)$ real-valued parameters for the ipsilateral RTF vectors $\check{\mathbf{h}}_t^{\text{ipsi},L}$ and $\check{\mathbf{h}}_t^{\text{ipsi},R}$, respectively, as well as $2(N-1)$ real-valued parameters for each of the speech temporal correlation vectors $\bar{\boldsymbol{\gamma}}_{t,L}^{L}$ and $\bar{\boldsymbol{\gamma}}_{t,R}^{R}$. The Cholesky factors of the inverse submatrices $(\tilde{\boldsymbol{\Phi}}_{i,t}^{LL})^{-1}$ and $(\tilde{\boldsymbol{\Phi}}_{i,t}^{RR})^{-1}$ in (5.21) and (5.22) are determined by $(M_L N)^2$ and $(M_R N)^2$ real-valued parameters, respectively. In contrast to the estimation procedures for the binaural STWF, separate TCNs are used for the left and right filters of the bilateral STWF. More specifically, one TCN uses the features in (5.25) at the microphones of the left hearing device to estimate the undetermined parameters of the left ipsilateral speech STCV and the left PSD mask, while another TCN uses the features at the microphones of the right hearing device to estimate the undetermined parameters of the right ipsilateral speech STCV and the right PSD mask. Similarly, two separate TCNs are used to estimate the undetermined parameters of the left and right inverse interference STCMs.

5.4 Simulation Setup

In this section, we present our simulation setup, consisting of the used datasets (Section 5.4.1), the baseline binaural speech enhancement algorithms (Section 5.4.2), and the settings of all algorithms (Section 5.4.3).

5.4.1 *Datasets*

To train, validate, and evaluate all supervised learning-based binaural speech enhancement algorithms, we constructed datasets using diverse speech and noise source material from the DNS 1 and DNS 2 challenge datasets [66], [68] and simulated binaural room impulse responses (BRIRs) from the CEC 1 dataset [69]. In

addition, to test the generalization capabilities of the considered algorithms in more realistic scenarios, we considered the CEC 3 dataset (Task 3)[1]. All datasets were used at a sampling rate of 16 kHz.

5.4.1.1 *Training and Validation*

For the training and validation datasets, we used the speech and noise source material from the DNS 2 challenge dataset, consisting of English sentences from 11 350 speakers and 600 noise classes. We chose not to use the speech and noise source material from the CEC 1 dataset in order to increase speaker and noise diversity. For the BRIRs, we used the CEC 1 training dataset, consisting of 6000 different room configurations. These BRIRs were simulated for a randomly positioned directional speech source and a randomly positioned omnidirectional noise source captured by binaural behind-the-ear hearing aids mounted on an artificial head in randomly sized rooms with reverberation time T_{60} ranging from 0.2 s to 0.4 s (i.e., low to moderate reverberation). The hearing aids consisted of three microphones each (front, mid, and rear), with a microphone spacing of about 7.6 mm. The front and mid microphones were chosen for all simulations, i.e., $M_L = M_R = 2$ microphones were used per hearing device. The speech source was located at an angle within $\pm 30^\circ$ w.r.t. the listener, while the noise source could be positioned anywhere in the room except for less than 1 m from the walls or the listener. Surface absorption coefficients were varied to simulate various room characteristics such as doors, windows, curtains, rugs, or furniture. Random speech and noise sources were convolved with BRIRs corresponding to a randomly chosen room configuration before being mixed at better-ear SNRs ranging from 0 dB to 15 dB (considering the reference microphones at both hearing aids). In total, the training and validation datasets have a length of 80 h and 20 h, respectively, with each utterance of length 4 s.

5.4.1.2 *Evaluation*

For evaluation, we considered a matched dataset resembling the training dataset and a mismatched dataset designed to test generalization capabilities. For the matched evaluation dataset, we used speech and noise source material from the DNS 1 challenge evaluation dataset and BRIRs from the CEC 1 validation dataset. Random speech and noise sources were convolved with BRIRs corresponding to a randomly chosen room configuration before being mixed at better-ear SNRs from -5 dB to 20 dB in steps of 5 dB. The training/validation and matched evaluation datasets were disjoint in terms of speakers, noise sources, and BRIRs. In total, 100 utterances were considered per SNR, with each utterance of length 10 s.

For the mismatched evaluation dataset, we used a subset of the CEC 3 development dataset (Task 3), comprising real noise backgrounds and higher-order ambisonic

1 https://claritychallenge.org/docs/cec3/task_3/cec3_task3_overview

RIRs recorded in complex environments (busy roads, railway platforms, and social gatherings) as well as simulated head rotation. Note that we omitted the social gatherings environment, which would have required a target speech extraction approach and is thus incompatible with the approach in this chapter. Random speech sources were convolved with higher-order ambisonic RIRs, rotated to simulate head movement, and binauralized with measured head-related transfer functions. The resulting speech components were mixed with real noise backgrounds at better-ear SNRs from -5 dB to 6 dB. This SNR range represents a subset of the full -12 dB to 6 dB range, chosen to provide a reasonable mismatch with our training dataset (which contained the range from 0 dB to 15 dB). The considered evaluation dataset presents mismatches in terms of speakers, noise types, noise spatial coherence, acoustic conditions (as reflected in the recorded RIRs), and the dynamic aspect of head rotation. In total, 997 utterances were selected, with each utterance lasting about 5 s.

5.4.2 *Baseline Algorithms*

We consider three baseline binaural speech enhancement algorithms, where two algorithms are causal and one algorithm is non-causal. The first baseline algorithm is causal and employs a purely learning-based approach that directly estimates the binaural temporal filter coefficients in the STFT domain (see Fig. 5.2), which can be viewed as a binaural extension of the DF algorithm proposed in [129]. More specifically, rather than estimating the speech STCVs, speech PSDs, and inverse interference STCMs to compute the binaural STWF vectors using (5.3), the DF algorithm directly estimates the binaural spatio-temporal filter vectors $\mathbf{w}_t^L$ and $\mathbf{w}_t^R \in \mathbb{C}^{MN}$ in (2.50), each determined by $2MN$ real-valued parameters. A single TCN uses the features in (5.25) at all microphones to estimate these parameters, ensuring the range $[-1, 1]$ with a hyperbolic tangent activation function as proposed in [129]. For more details about the DF algorithm, we refer to Section 3.2.1.

As second state-of-the-art baseline algorithm, we consider the causal binaural Conv-TasNet algorithm, which uses a learned transform instead of the STFT and a TCN-based separator that estimates real-valued masks employed in a mask-and-sum approach [189], an extension of the Conv-TasNet algorithm [89] introduced in Section 3.2.2.

As third state-of-the-art baseline algorithm, we use the non-causal BCCTN algorithm [188], which uses a complex-valued convolutional transformer network that estimates complex-valued time-frequency masks for the left and right reference channels. Note that the BCCTN algorithm uses a non-causal multi-head attention implementation and is thus not suitable for real-time processing—in contrast to all other considered algorithms.

5.4.3 *Algorithmic Settings*

The algorithmic settings mostly followed the settings in Chapter 4. We used the same STFT framework and input features for the STWF algorithms and the DF algorithm. To increase speech correlation across successive STFT frames, we used a high temporal resolution, i.e., a frame length of 8 ms and 75 % overlap, resulting in a low input-output latency. A $\sqrt{\text{Hann}}$ window was used both as analysis and synthesis window. As temporal filter length, the STWF algorithms and the DF algorithm used $N = 5$ frames, such that temporal correlations within 16 ms could be exploited. This choice represents a trade-off between capturing sufficient temporal context to exploit speech correlations and maintaining reasonable computational complexity. To limit speech distortion, a minimum gain of $g_{\min} = -20\,\text{dB}$ was applied to the binaural output signals of the STWF algorithms and the DF algorithm during evaluation. The final estimated binaural target speech components were thus obtained as

$$\widehat{x}_{t,\nu} = \begin{cases} g_{\min} y_{t,\nu}, & \text{if } |\mathbf{w}_t^{\nu,\mathsf{H}} \mathbf{y}_t| < |g_{\min} y_{t,\nu}| \\ \mathbf{w}_t^{\nu,\mathsf{H}} \mathbf{y}_t, & \text{else} \end{cases}. \tag{5.30}$$

To train all algorithms except the BCCTN algorithm, we used the following STFT-domain loss function proposed in [120]:

$$\begin{aligned} \mathcal{L}_{f,t,\nu} &= \beta \left| x_{f,t,\nu} - \widehat{x}_{f,t,\nu} \right| + (1-\beta) \left| |x_{f,t,\nu}| - |\widehat{x}_{f,t,\nu}| \right| \\ \mathcal{L} &= \frac{1}{2FT} \sum_{f=1}^{F} \sum_{t=1}^{T} \sum_{\nu \in \{L,R\}} \mathcal{L}_{f,t,\nu}, \end{aligned} \tag{5.31}$$

where $\beta = 0.4$ is a hyperparameter chosen as in [120]. The magnitude term helps preserve spectral shape and formant structure, while the complex-valued difference term helps preserve phase relationships. Note that this loss was computed after the estimated binaural target speech components in (2.50) were transformed back to the time domain using an inverse STFT (see Fig. 5.2), followed by an additional transformation to the STFT domain with a frame length of 32 ms and 50 % overlap. For the BCCTN algorithm, we used the STFT framework and loss function proposed in [188], which incorporates terms reflecting noise reduction, intelligibility improvement, and binaural cue preservation.

We used TCNs as the DNN architecture for the STWF algorithms and the DF algorithm, implemented based on the official (monaural) Conv-TasNet implementation[2]. We fixed the number of stacks to two, the number of layers to six, the kernel size to three, and the bottleneck size to 32, yielding a temporal receptive field size of 512 ms. All STWF algorithms (imposing different correlation structures) and the DF algorithm were trained separately, with the TCN architecture remaining identical and only adapting its final layer to match the required parameter count for each algorithm (presented in Table 5.1 for the STWF algorithms).

2 `https://github.com/naplab/Conv-TasNet`

For both the binaural Conv-TasNet algorithm and the BCCTN algorithm, we used the code provided by the authors and the DNN hyperparameters proposed in [189] and [188], respectively.

All algorithms were trained for a maximum of 100 epochs with early stopping using the AdamW optimizer [215] and with gradient ℓ_2 norms clipped to 5. The learning rate was initialized at 10^{-3} and halved after three epochs without validation loss improvement. Early stopping was applied after ten epochs without validation loss improvement. For the STWF algorithms and the DF algorithm, complex-valued numbers were constructed from the (real-valued) TCN outputs by assigning separate output elements for the real and imaginary parts. The simulations were implemented using PyTorch 2.0.1 [216] and executed on NVIDIA GeForce RTX A5000 graphics cards.

5.5 Validity of Spatio-Temporal Correlation Structures

In this section, we describe an approach to validate the proposed spatio-temporal correlation structures for the speech STCVs and interference STCMs in Section 5.2. We first discuss how to compute ground-truth STCVs and STCMs, impose spatio-temporal structure on these quantities, introduce several metrics to evaluate the incurred model mismatch, and finally present the validation results.

5.5.1 *Ground-Truth STCVs and STCMs*

To compute the ground-truth speech STCVs and interference STCMs, we first apply recursive smoothing on the instantaneous oracle speech and noise STCMs, i.e.,

$$\mathbf{\Phi}_{x,t} = \lambda \mathbf{\Phi}_{x,t-1} + (1-\lambda)\mathbf{x}_t \mathbf{x}_t^{\mathsf{H}} \tag{5.32}$$

$$\mathbf{\Phi}_{n,t} = \lambda \mathbf{\Phi}_{n,t-1} + (1-\lambda)\mathbf{n}_t \mathbf{n}_t^{\mathsf{H}}. \tag{5.33}$$

To track rapidly varying speech and noise statistics, we set the smoothing factor λ to match the frame shift of the employed STFT (equal to 2 ms), using the relation $\lambda = \exp(-T_s/\tau)$, where $T_s = 2\,\text{ms}$ denotes the frame shift used in the simulations and $\tau = 2\,\text{ms}$ denotes the smoothing time constant. Using (5.32), the left and right speech STCVs and PSDs are computed using (2.45), i.e.,

$$\boldsymbol{\gamma}_{t,\nu} = \frac{\mathbf{\Phi}_{x,t}\mathbf{e}_\nu}{\phi_{x,t}^{\nu}}, \qquad \nu \in \{L, R\} \tag{5.34}$$

$$\phi_{x,t}^{\nu} = \mathbf{e}_\nu^{\mathsf{T}} \mathbf{\Phi}_{x,t} \mathbf{e}_\nu, \quad \nu \in \{L, R\}. \tag{5.35}$$

The left and right interference STCMs are computed using (2.53), i.e.,

$$\mathbf{\Phi}_{i,t}^{\nu} = \mathbf{\Phi}_{x,t} - \phi_{x,t}^{\nu} \boldsymbol{\gamma}_t^{\nu} \boldsymbol{\gamma}_t^{\nu,\mathsf{H}} + \mathbf{\Phi}_{n,t}, \quad \nu \in \{L, R\}. \tag{5.36}$$

5.5.2 *Spatio-Temporal Structures*

To impose the global RTF structure or the ipsilateral RTF structure (see Section 5.2.1) on the ground-truth speech STCVs, the ground-truth RTFs are first computed according to (5.5) using the oracle speech STCM (5.32) and the oracle speech PSDs (5.35). For the global RTF structure, the speech STCVs are constructed using (5.8) and (5.11), where the global RTF vector $\check{\mathbf{h}}_t^{\text{glob}}$ in (5.7) is constructed using the ground-truth RTFs and the speech temporal correlation vector $\bar{\boldsymbol{\gamma}}_{t,1}^{L}$ is extracted from the speech STCV in (5.34). Similarly, for the ipsilateral RTF structure, the speech STCVs are constructed using (5.15), where the ipsilateral RTF vectors $\check{\mathbf{h}}_t^{\text{ipsi},L}$ and $\check{\mathbf{h}}_t^{\text{ipsi},R}$ in (5.13) and (5.14) are constructed using the ground-truth RTFs and the speech temporal correlation vectors $\bar{\boldsymbol{\gamma}}_{t,L}^{\nu}$ and $\bar{\boldsymbol{\gamma}}_{t,R}^{\nu}$ are extracted from the speech STCV in (5.34).

The common interference STCM (see Section 5.2.2) is constructed as the matrix minimizing the squared Frobenius norm of the difference with the ground-truth interference STCMs $\boldsymbol{\Phi}_{i,t}^{L}$ and $\boldsymbol{\Phi}_{i,t}^{R}$ in (5.36), i.e.,

$$\tilde{\boldsymbol{\Phi}}_{i,t} = \underset{\boldsymbol{\Phi}}{\operatorname{argmin}} \left(\left\| \boldsymbol{\Phi} - \boldsymbol{\Phi}_{i,t}^{L} \right\|_F^2 + \left\| \boldsymbol{\Phi} - \boldsymbol{\Phi}_{i,t}^{R} \right\|_F^2 \right) \tag{5.37}$$

$$= \frac{1}{2} \left(\boldsymbol{\Phi}_{i,t}^{L} + \boldsymbol{\Phi}_{i,t}^{R} \right). \tag{5.38}$$

For the bilateral correlation structures (see Section 5.2.3), the coefficients of the speech STCVs and the interference STCMs corresponding to the contralateral correlations are simply set to zero.

5.5.3 *Results*

We evaluate the model mismatch incurred by imposing spatio-temporal correlation structures on the ground-truth speech STCVs and interference STCMs in terms of several metrics on the evaluation dataset. To evaluate model mismatch for the speech STCVs, we consider the mean relative ℓ_2 norm and the mean Hermitian angle, defined as

$$\epsilon_{\ell_2} = \frac{1}{2FT} \sum_{f=1}^{F} \sum_{t=1}^{T} \sum_{\nu \in \{L,R\}} \frac{\left\| \tilde{\boldsymbol{\gamma}}_{f,t}^{\nu} - \boldsymbol{\gamma}_{f,t}^{\nu} \right\|_2}{\left\| \boldsymbol{\gamma}_{f,t}^{\nu} \right\|_2} \tag{5.39}$$

$$\epsilon_{\theta} = \frac{1}{2FT} \sum_{f=1}^{F} \sum_{t=1}^{T} \sum_{\nu \in \{L,R\}} \arccos \left(\frac{|\tilde{\boldsymbol{\gamma}}_{f,t}^{\nu,\mathsf{H}} \boldsymbol{\gamma}_{f,t}^{\nu}|}{\left\| \tilde{\boldsymbol{\gamma}}_{f,t}^{\nu} \right\|_2 \left\| \boldsymbol{\gamma}_{f,t}^{\nu} \right\|_2} \right), \tag{5.40}$$

where $\boldsymbol{\gamma}_{f,t}^{\nu}$ corresponds to the ground-truth speech STCVs in (5.34) and $\tilde{\boldsymbol{\gamma}}_{f,t}^{\nu}$ corresponds to the speech STCVs with imposed structure.

To evaluate model mismatch for the interference STCMs, we consider the mean relative Frobenius norm and the mean correlation matrix distance [221], defined as

$$\epsilon_{\text{Fro}} = \frac{1}{2FT} \sum_{f=1}^{F} \sum_{t=1}^{T} \sum_{\nu \in \{L,R\}} \frac{\left\| \tilde{\mathbf{\Phi}}_{i,f,t} - \mathbf{\Phi}^{\nu}_{i,f,t} \right\|_F}{\left\| \mathbf{\Phi}^{\nu}_{i,f,t} \right\|_F} \tag{5.41}$$

$$\epsilon_{\text{CMD}} = \frac{1}{2FT} \sum_{f=1}^{F} \sum_{t=1}^{T} \sum_{\nu \in \{L,R\}} \left(1 - \frac{\text{trace}\{\tilde{\mathbf{\Phi}}_{i,f,t} \mathbf{\Phi}^{\nu,\mathsf{H}}_{i,f,t}\}}{\|\tilde{\mathbf{\Phi}}_{i,f,t}\|_F \|\mathbf{\Phi}^{\nu}_{i,f,t}\|_F} \right), \tag{5.42}$$

where $\mathbf{\Phi}^{\nu}_{i,f,t}$ corresponds to the ground-truth interference STCMs in (5.36) and $\tilde{\mathbf{\Phi}}_{i,f,t}$ corresponds to the common interference STCM in (5.37). The correlation matrix distance ϵ_{CMD} can be interpreted as the angle between the vectorized STCMs $\tilde{\mathbf{\Phi}}_{i,f,t}$ and $\mathbf{\Phi}_{i,f,t}$ in $(2MN)^2$-dimensional space, yielding values between 0 and 1, where smaller values denote higher similarity.

For all considered spatio-temporal correlation structures, Table 5.1 shows the number of required parameters to estimate the speech STCVs and the interference STCMs, both as a function of the number of microphones M and the temporal filter length N as well as specifically for $M_L = M_R = 2$ (i.e., $M = 4$) and $N = 5$ (used in the simulations). It also shows the model mismatch by imposing spatio-temporal correlation structures on the ground-truth speech STCVs and interference STCMs in terms of the metrics introduced in the previous subsection.

For the speech STCVs, it can be observed on the one hand that all spatio-temporal correlation structures significantly reduce the number of required parameters. The global RTF structure yields the largest reduction (factor 5.4), while the ipsilateral RTF structure yields the smallest reduction (factor 1.9). On the other hand, both the global RTF structure and the bilateral structures incur a relatively large model mismatch in terms of both metrics. Although the global RTF structure is commonly used in single-frame algorithms (typically utilizing longer STFT frames), the global RTF structure may be less suitable for the considered multi-frame algorithms, which rely on short STFT frames to exploit the temporal correlations of speech signals. In addition, due to reverberation and the relatively large distance between the microphones on the left and right hearing devices, the speech STFT coefficients of the global reference microphone (on the left hearing device) may not be fully correlated with the speech STFT coefficients of the contralateral microphones (on the right hearing device). In contrast, the ipsilateral RTF structure incurs a smaller model mismatch while reducing the number of required parameters with a similar factor as the bilateral RTF structure.

For the interference STCMs, it can be observed on the one hand that assuming a common STCM reduces the number of required parameters by a factor 2, while the bilateral structure further reduces the number of required parameters by an additional factor 2. In absolute numbers, the parameter reductions for the interference STCMs are much larger than the reductions for the speech STCVs. On the other hand, the bilateral interference correlation structure incurs a relatively large model mismatch in terms of both metrics, consistent with the model mismatch for the

Table 5.2: Computational complexity in terms of the average real-time factor (RF), the number of multiply-accumulate operations per second (MACS), and the number of trainable weights for the deep binaural STWF and the deep bilateral STWF (imposing different correlation structures) as well as the binaural deep filter (DF) algorithm, the binaural Conv-TasNet algorithm, and the non-causal binaural complex convolutional transformer network (BCCTN) algorithm.

	$\mathbf{\Phi}_{i,t}^{L} = \mathbf{\Phi}_{i,t}^{R}$	$\boldsymbol{\gamma}_{t}^{\nu}$ structure	RF / % (↓)	MACS / M (↓)	trainable weights / M (↓)
noisy	—	—	—	—	—
binaural STWF	✗	✗	54.6	539	2.10
	✓	✗	36.2	287	1.30
	✓	global RTF	25.1	273	1.19
	✓	ipsilateral RTF	35.0	287	1.24
bilateral STWF	—	✗	12.1	76	0.42
	—	ipsilateral RTF	12.4	76	0.41
binaural DF	—	—	**4.8**	**6**	**0.34**
binaural Conv-TasNet [189]	—	—	36.0	75	1.67
BCCTN (non-causal) [188]	—	—	—	5730	11.09

bilateral speech correlation structure. In contrast, assuming a common interference STCM incurs a smaller model mismatch. The observations in terms of model mismatch for the considered spatio-temporal correlation structures will be confirmed by the simulation results in the next section.

5.6 Simulation Results

In this section, we investigate the speech enhancement performance and the computational complexity of the deep binaural STWF and the deep bilateral STWF using the proposed correlation structures. Furthermore, we compare their performance with three baseline algorithms, i.e., the binaural DF algorithm, the binaural Conv-TasNet algorithm, and the BCCTN algorithm. As mentioned before, all considered algorithms except the BCCTN algorithm are causal. To evaluate computational complexity, we consider the RF, defined as processing duration vs. utterance duration using a single thread on an AMD EPYC 7443P CPU, as well as the number of multiply-accumulate operations per second (MACS) (determined using the PyTorch profiler) and the number of trainable weights. To evaluate speech enhancement performance, we consider the wideband PESQ [183] metric and the hearing aid speech quality and speech intelligibility index (HASQI) [185] as objective metrics of speech quality, as well as the hearing aid speech perception index (HASPI) [187] as an objective metric of speech intelligibility. To evaluate binaural cue preservation, we consider the ILD and IPD errors between the output signals and the target speech components of the input signals described in Section 2.2. More in particular, we use the ILD error defined in (2.55) and the IPD error defined in (2.57), where both errors are computed only in STFT bins with active speech [188].

For all speech enhancement and binaural cue preservation metrics, we used the left and right reverberant speech signals at the reference microphones as the reference signals. For PESQ, we averaged the values across the left and right output signals, whereas HASQI and HASPI inherently consider only the better ear. For HASQI and HASPI, we assumed normal hearing (i.e., a flat hearing loss of 0 dB). Audio demos for matched and mismatched conditions (including a condition with a moving source entirely mismatched from the training dataset) can be found online[3].

5.6.1 *Computational Complexity*

Table 5.2 shows the computational complexity for all considered binaural speech enhancement algorithms, where all metrics were averaged across all utterances and SNRs of the matched evaluation dataset. The DF algorithm achieves the lowest computational complexity, with an RF of 4.8 %, 6 M MACS, and 0.34 M trainable weights. In contrast, the non-causal BCCTN algorithm shows the highest computational complexity with 5730 M MACS and 11.09 M trainable weights, highlighting its substantial computational demand. The deep binaural STWF not imposing any correlation structure results in an RF of 54.6 %, 539 M MACS, and 2.10 M trainable weights. Imposing a common interference STCM reduces the RF to 36.2 %, while also decreasing the MACS to 287 M and the trainable weights to 1.30 M. Imposing the global RTF structure further reduces the RF to 25.1 %, the MACS to 273 M, and the trainable weights to 1.19 M. In contrast, imposing the ipsilateral RTF structure yields only minimal additional computational savings over the common interference STCM. Compared to the deep binaural STWF algorithms, the deep bilateral STWF algorithms result in lower computational complexity, with RFs around 12 %, 76 M MACS, and between 0.41 M and 0.42 M trainable weights.

5.6.2 *Matched Evaluation Dataset*

For the matched evaluation dataset, Table 5.3 shows the speech enhancement and binaural cue preservation performance for all considered binaural speech enhancement algorithms, averaged across all utterances and SNRs. First, it can be observed that all algorithms yield improvements in terms of all considered speech enhancement metrics. The deep binaural STWF not imposing any correlation structure achieves a high PESQ value (2.40) and the highest HASQI and HASPI values (0.50 and 0.95, respectively), while the non-causal BCCTN algorithm achieves the highest PESQ value (2.48), but only moderate HASQI and HASPI values (0.48, and 0.91). Compared to the unstructured variant, imposing a common interference STCM results in only minor PESQ reductions (2.38) while achieving the same HASQI and HASPI values. Further imposing the global RTF structure slightly reduces speech

3 https://uol.de/en/sigproc/research/audio-demos/binaural-noise-reduction/stwf

Table 5.3: Speech enhancement performance in terms of average PESQ, HASQI, and HASPI values and binaural cue preservation in terms of average ILD and IPD errors for the deep binaural STWF and the deep bilateral STWF (imposing different correlation structures) as well as the binaural deep filter (DF) algorithm, the binaural Conv-TasNet algorithm, and the non-causal binaural complex convolutional transformer network (BCCTN) algorithm on the matched evaluation dataset.

	$\mathbf{\Phi}_{i,t}^{L} = \mathbf{\Phi}_{i,t}^{R}$	$\boldsymbol{\gamma}_{t}^{\nu}$ structure	PESQ (↑)	HASQI (↑)	HASPI (↑)	ΔILD / dB (↓)	ΔIPD / rad (↓)
noisy	—	—	1.62	0.39	0.90	—	—
binaural STWF	✗	✗	2.40	**0.50**	**0.95**	3.26	0.72
	✓	✗	2.38	**0.50**	**0.95**	3.27	0.73
	✓	global RTF	2.34	0.49	0.94	4.03	0.90
	✓	ipsilateral RTF	2.39	**0.50**	**0.95**	3.29	0.73
bilateral STWF	—	✗	2.19	0.47	0.94	4.14	0.78
	—	ipsilateral RTF	2.18	0.47	0.94	4.13	0.78
binaural DF	—	—	2.23	0.47	0.94	3.42	0.75
binaural Conv-TasNet [189]	—	—	2.19	0.49	0.94	3.86	0.81
BCCTN (non-causal) [188]	—	—	**2.48**	0.48	0.91	**2.62**	**0.66**

enhancement performance (PESQ of 2.34). In contrast, imposing the ipsilateral RTF structure preserves the speech enhancement performance of the unstructured variant (PESQ of 2.39). The deep bilateral STWF algorithms, the binaural DF algorithm, and the binaural Conv-TasNet algorithm result in lower speech enhancement performance (PESQ around 2.20). The lower performance of the deep bilateral STWF algorithms is presumably caused by the lack of information exchange between the left and right hearing devices, which limits both the spatial diversity the TCNs can exploit as well as the number of microphones available for filtering.

In terms of binaural cue preservation, it can be observed that the non-causal BCCTN algorithm achieves the best performance, while the deep bilateral STWF algorithms and the deep binaural STWF algorithm imposing the global RTF structure exhibit the worst performance. Notably, imposing the ipsilateral RTF structure preserves binaural cues as effectively as the unstructured variant. The good binaural cue preservation of the BCCTN algorithm is presumably caused by the inclusion of loss terms that penalize binaural cue distortion. However, it should be noted that, from a perceptual perspective, all algorithms except the deep bilateral STWF algorithms preserve the binaural cues quite well.

5.6.3 *Mismatched Evaluation Dataset*

For the mismatched evaluation dataset, Table 5.4 shows the speech enhancement and binaural cue preservation performance for all considered binaural speech enhancement algorithms, averaged across all utterances. Compared to the matched evaluation dataset, the speech enhancement performance of all algorithms is reduced, which is expected due to the strong mismatch between training and evaluation conditions. Nevertheless, all algorithms still yield improvements in terms

Table 5.4: Speech enhancement performance in terms of average PESQ, HASQI, and HASPI values and binaural cue preservation in terms of average ILD and IPD errors for the deep binaural STWF and the deep bilateral STWF (imposing different correlation structures) as well as the binaural deep filter (DF) algorithm, the binaural Conv-TasNet algorithm, and the non-causal binaural complex convolutional transformer network (BCCTN) algorithm on the mismatched evaluation dataset.

	$\mathbf{\Phi}^{L}_{i,t} = \mathbf{\Phi}^{R}_{i,t}$	$\boldsymbol{\gamma}^{\nu}_{t}$ structure	PESQ (↑)	HASQI (↑)	HASPI (↑)	ΔILD / dB (↓)	ΔIPD / rad (↓)
noisy	—	—	1.09	0.07	0.42	—	—
binaural STWF	✗	✗	1.21	**0.12**	**0.55**	1.33	6.35
	✓	✗	1.20	0.11	0.54	1.33	6.35
	✓	global RTF	1.19	0.11	0.52	1.42	6.84
	✓	ipsilateral RTF	1.19	0.11	0.54	1.33	6.32
bilateral STWF	—	✗	1.18	0.11	0.54	1.40	7.93
	—	ipsilateral RTF	1.16	0.11	0.54	1.44	8.39
binaural DF	—	—	1.19	**0.12**	0.53	1.31	6.41
binaural Conv-TasNet [189]	—	—	1.14	0.09	0.44	1.44	7.45
BCCTN (non-causal) [188]	—	—	**1.23**	0.10	0.42	**1.07**	**4.69**

of all considered speech enhancement metrics (except for the BCCTN algorithm in terms of HASPI). The speech enhancement and binaural cue preservation tendencies across algorithms are similar to those observed on the matched evaluation dataset, however with smaller differences between algorithms.

To experimentally evaluate the robustness of the considered binaural speech enhancement algorithms, we have also included an audio example of an acoustic scene that is entirely mismatched from the training dataset (featuring a moving target speaker in an unseen room, with unseen quasi-diffuse noise, and a hearing aid configuration with unseen inter-microphone spacings) on the webpage[5].

5.7 Summary

In this chapter, we proposed several procedures to impose spatio-temporal correlation structures on the speech STCVs and interference STCMs, required to implement the binaural STWF. These procedures mainly differ in terms of the relation between the microphones, particularly between the left and the right hearing device, as well as the number of parameters to be estimated. First, assuming that the spatial correlation of the speech component is stationary over the length of the temporal filter, we proposed to decompose the speech STCV as the Kronecker product of a spatial RTF vector and a temporal correlation vector. We either considered a single global reference microphone or a reference microphone for each hearing device. Second, we proposed to replace the left and right interference STCMs by a common interference STCM. In addition, we considered a bilateral STWF by neglecting all spatio-temporal correlations between both hearing devices. All required parameters were estimated by embedding the binaural STWF into a supervised learning framework.

Simulation results using both simulated and measured BRIRs as well as diverse speech and noise sources demonstrate that the combination of the speech STCV structure considering two reference microphones and a common interference STCM yields the best overall performance, reducing the real-time factor by around 36 % while maintaining speech enhancement and binaural cue preservation performance compared to not imposing any spatio-temporal correlation structure. These results are consistent with a validation based on ground-truth quantities. Furthermore, the best deep binaural STWF algorithm outperforms two state-of-the-art binaural speech enhancement algorithms based on supervised learning, namely the deep filter algorithm and the binaural Conv-TasNet algorithm, while approaching the performance of the (non-causal) BCCTN algorithm.

6

SPATIAL REGULARIZATION FOR IMPROVED INTERPRETABILITY

In Chapter 5, we decomposed the speech STCV into the Kronecker product of a spatial factor (corresponding to the RTF vector) and a temporal factor (the TCV), using a signal approximation loss function defined on the output of the deep binaural STWF. While this deep binaural STWF yielded a high speech enhancement performance, in this chapter, we investigate the acoustic interpretability of the estimated RTF vector—specifically whether it accurately reflects the underlying spatial characteristics of the acoustic scenario. This investigation is motivated by one of the key reasons for employing a hybrid speech enhancement approach such as the deep binaural STWF: interpretability. Since we find the estimated RTF vector to be acoustically implausible, we propose a spatial regularization procedure that incorporates an additional loss term defined on the estimated RTF vector. This loss term incentivizes the DNN to output estimates that are not only effective for speech enhancement but also reflect the underlying spatial characteristics of the acoustic scenario. To automatically balance the individual loss terms, we employ an adaptive weighting method based on homoscedastic uncertainty [222]. Using the DNS 1 and DNS 2 challenge datasets and simulated RIRs, simulation results show that the combined loss function yields accurate estimates of the RTF vector even in reverberant environments, without increasing computational complexity or sacrificing speech enhancement performance. We hypothesize that this regularization procedure can be extended to other coupled hybrid speech enhancement approaches to improve the acoustic interpretability of estimated quantities.

The remainder of this chapter is organized as follows. In Section 6.1, we briefly review the deep spatio-temporal MVDR filter and its required quantities. In Section 6.2, we discuss the plausibility of the estimated RTF vectors. In Section 6.3, we introduce the proposed spatial regularization procedure, which includes RTF loss terms and an adaptive weighting method based on homoscedastic uncertainty. In Section 6.4, we introduce our approach to evaluate the plausibility of the estimated RTF vectors by analyzing their beampatterns. The simulation setup and the corresponding results for the proposed spatial regularization procedure are presented in Section 6.5 and Section 6.6, respectively.

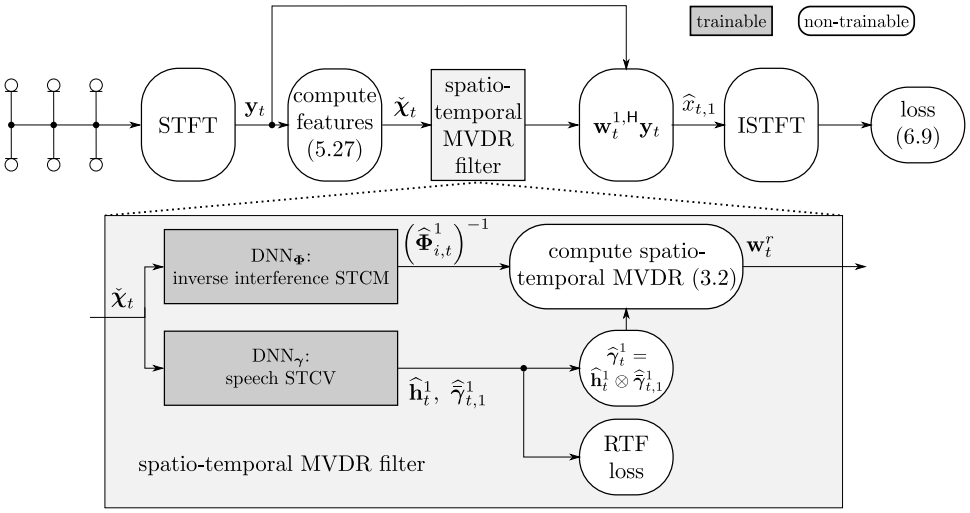


Figure 6.1: Block diagram of the deep spatio-temporal MVDR filter, without or with prior factorization of the speech STCV into the RTF vector and the speech TCV, and with the optional loss term defined on the estimated RTF vector.

6.1 Deep Spatio-Temporal MVDR Filter

Aiming at minimizing the output interference PSD while preserving the spatio-temporally correlated speech component at the reference microphone with index $r = 1$, the spatio-temporal MVDR filter in (3.2) requires estimates of the inverse interference STCM $(\boldsymbol{\Phi}_{i,t}^1)^{-1}$ and the speech STCV $\boldsymbol{\gamma}_t^1$. Assuming that the RTFs are constant over N frames and selecting the microphone with index $r = 1$ as the reference microphone, the speech STCV, which describes the correlation between the N most recent speech STFT coefficients at each microphone and the current speech STFT coefficient at the reference microphone, can be factorized into a spatial factor, corresponding to the RTF vector $\check{\mathbf{h}}_t^1$, and a temporal factor, corresponding to the speech TCV $\bar{\boldsymbol{\gamma}}_{t,1}^1$ (cf. (2.37)).

Similarly as in the previous chapter, we embed the spatio-temporal MVDR filter into a deep learning framework, using DNNs to estimate the required quantities. More in particular, as illustrated in Fig. 6.1, we use one DNN to estimate both factors $\check{\mathbf{h}}_t^1$ and $\bar{\boldsymbol{\gamma}}_{t,1}^1$, and we use one DNN for estimating the inverse interference STCM $(\boldsymbol{\Phi}_{i,t}^1)^{-1}$.

6.2 Plausibility of Estimated RTF Vectors

Since the DNNs are trained with a signal approximation loss function defined at the output of the model-based enhancement stage, they are free to estimate the RTF vector and the speech TCV in a way that leads to good speech enhancement per-

formance, without necessarily reflecting the true underlying quantities. This lack of consistency undermines one of the key motivations for employing a hybrid speech enhancement approach, which is to provide interpretable quantity estimates. By interpretability, we mean that the estimated RTF vectors should reflect the spatial characteristics of the acoustic scenario, including the relative positioning of the microphones and the target source, as well as the influence of the acoustic environment. Instead of analyzing the RTF vectors directly, we propose to evaluate their plausibility by deriving spatial filters from them and inspecting their corresponding beampatterns. Traditional spatial filters have often been designed based on a steering vector [191], which assumes free-field and far-field propagation as well as only modeling the direct path from the source to the microphones. In contrast, the RTF vector captures additional acoustic effects, including reverberation and acoustic shadowing (Section 1.1.3). By incorporating the RTF vector in the spatial filter design, not only the direct-path sound but also reflections can be accounted for, which can improve speech intelligibility [22]. Hence, to assess the plausibility of the estimated RTF vectors, we propose to analyze their beampatterns under different acoustic environments, comparing them to expected beampatterns.

6.3 Proposed Spatial Regularization Procedure

The main cause of the lack of acoustic interpretability of the RTF vectors estimated using a DNN is the reliance on a loss function that only contains a signal approximation term. This does not incentivize the DNN to output accurate RTF vector estimates if the goal of minimizing its loss function is achievable otherwise. To address this issue, we propose to augment the loss function by considering both a signal approximation term and a term that penalizes estimation errors in the RTF vectors.

The concept of incorporating RTF-based loss terms has been explored in prior research [101], [197]. For example, [197] proposed to imprint binaural RTFs estimated using a dual-path recurrent neural network (DPRNN) onto the outputs of a DPRNN-based binaural source separation algorithm, improving binaural cue preservation and slightly improving speech enhancement performance. The DPRNN-based RTF estimator was trained using a weighted MSE loss. Furthermore, [101] proposed a simultaneous speech enhancement and localization algorithm that employs a CRNN to output a spatial filter. More in particular, in [101] an additional loss term inspired by the distortionless response constraint of MVDR beamformers in (3.22) was used, incentivizing the CRNN to output a spatial filter with a distortionless response towards the target source. Localization was then performed by finding the direction with the highest response, similarly as in the well-known steered response power with phase transform algorithm [223]. The inclusion of the distortionless response term improved both speech enhancement performance and localization accuracy compared to not using it. Notably, their loss term was not defined on the estimated RTF; the RTF was merely used to establish the distortionless response term. We propose to define loss terms on the estimated RTFs (Section 6.3.1), with the dis-

tance between the target and estimated RTFs computed using either the MSE or the Hermitian angle [224] (Section 6.3.1). Our proposed spatial regularization procedure aims at improving the interpretability of the estimated quantities, promoting estimates that accurately reflect the acoustic scenario. To avoid the need for manual tuning of the RTF loss term's contributions to the overall loss function, we employ the method proposed in [222], which adaptively balances the individual loss terms based on an estimate of the homoscedastic uncertainty of each task (Appendix A).

6.3.1 *RTF Loss Terms*

To improve the accuracy of the estimated RTF vectors, we propose to use one of two loss terms: the MSE RTF loss term or the Hermitian angle RTF loss term. The MSE RTF loss term is defined as

$$\boxed{\mathcal{L}_{\text{RTF-MSE}} = \frac{1}{FTM} \sum_{f=1}^{F} \sum_{t=1}^{T} b_{f,t} \sum_{m=1}^{M} \left| \widehat{h}^{1}_{f,t,m} - h^{1}_{f,t,m} \right|^2,} \tag{6.1}$$

where $b_{f,t}$ denotes a time- and frequency-dependent weight given by

$$b_{f,t} = \frac{\left| x_{f,t,1} \right|^2}{\text{percentile}_{0.95}\left(\left\{ \left| x_{f,t,1} \right|^2 \right\}_{f,t} \right)}, \tag{6.2}$$

and where $\text{percentile}_{0.95}(\cdot)$ denotes the 95 % percentile. The weight $b_{f,t}$ emphasizes estimates of the RTFs in STFT bins with high target speech power while preventing the inclusion of RTF estimates in STFT bins where no target speech is active. The Hermitian angle RTF loss term is defined as

$$\boxed{\mathcal{L}_{\text{RTF-HA}} = \frac{1}{FT} \sum_{f=1}^{F} \sum_{t=1}^{T} b_{f,t} \arccos\left(\frac{\left| \widehat{\mathbf{h}}^{1,\mathsf{H}}_{f,t} \check{\mathbf{h}}_{f,t} \right|}{\left\| \widehat{\mathbf{h}}_{f,t} \right\| \left\| \check{\mathbf{h}}_{f,t} \right\|} \right).} \tag{6.3}$$

In comparison to the MSE RTF loss term, the Hermitian angle RTF loss term considers only the angle between the estimated and target RTF vectors.

6.4 Beampattern Evaluation Procedure

As mentioned before, instead of analyzing the RTF vectors directly, we assess their interpretability by deriving spatial filters from them and inspecting their corresponding beampatterns. Similarly as in [225], we employ the matched filter (MF), which aims at maximizing the white noise gain or, equivalently, minimizing the output

noise PSD under the assumption of spatially white noise subject to a unit response for the RTF vector,[1]. The MF is given by

$$\check{\mathbf{w}}_{\mathrm{MF}}^{1} = \underset{\check{\mathbf{w}}^{1}}{\operatorname{argmin}}\, \check{\mathbf{w}}^{1,\mathsf{H}}\check{\mathbf{w}}^{1} \text{ subject to } \check{\mathbf{w}}^{1,\mathsf{H}}\check{\mathbf{h}}^{1} = 1 \tag{6.4}$$

$$= \frac{\check{\mathbf{h}}^{1}}{\left\|\check{\mathbf{h}}^{1}\right\|^{2}}, \tag{6.5}$$

where we consider a spatially stationary acoustic scenario for simplicity, resulting in a time-invariant RTF vector.

ANECHOIC BEAMPATTERNS Beampatterns describe the directional response of a spatial filter $\check{\mathbf{w}}_{f}^{1}$ to sources arriving from different directions, typically assuming free-field propagation without reflections and hence parameterized by the direction of arrival (DOA). The subband beampattern for the f-th frequency bin is defined as [225]

$$B_{f,\theta} = \left|\check{\mathbf{w}}_{f}^{1,\mathsf{H}}\check{\mathbf{d}}_{f,\theta}\right|, \tag{6.6}$$

where $\check{\mathbf{d}}_{f,\theta} = \begin{bmatrix}1 & \exp\left(-j\omega_F d_2 \tau_{2,\theta}\right) & \dots & \exp\left(-j\omega_F d_M \tau_{M,\theta}\right)\end{bmatrix}^{\mathsf{T}}$ denotes the (anechoic) steering vector for a source at DOA θ, ω_F denotes the angular center frequency of the f-th frequency bin, and $\tau_{m,\theta}$ denotes the DOA-dependent time difference of arrival between the m-th microphone and the reference microphone. The fullband beampattern aggregates the beampower across all frequencies, i.e., [225]

$$B_{\theta} = \sum_{f=1}^{F} \left|B_{f,\theta}\right|^{2}. \tag{6.7}$$

While other definitions of the fullband beampattern exist (such as in [3], which weighs subband beampatterns with the speech PSD), we follow the evaluation procedure in [225].

REVERBERANT BEAMPATTERNS In conventional beampattern analysis, the steering vectors at each frequency $\check{\mathbf{d}}_{f,\theta}$ depend solely on the DOA. However, in reverberant environments, reflections and room acoustics affect the observed sound field. An alternative analysis approach proposed in [225] replaces anechoic steering vectors in (6.6) with measured or simulated ATFs. This approach requires generating RIRs for sources placed at regularly spaced positions on a circle centered at the microphone array, transforming the RIRs into the frequency domain, and using the resulting ATFs to compute beampatterns. Unlike anechoic beampatterns, which depend only on the DOA, reverberant beampatterns are influenced by the absolute

1 This can be seen as a special case of the MVDR beamformer, where the noise spatial coherence matrix is assumed to be equal to the identity matrix.

source and microphone positions within the room, including source-microphone distance. While objective evaluation metrics such as beamwidth can be used, this chapter primarily focuses on visually inspecting beampatterns to assess the interpretability of estimated RTFs.

BEAMPATTERN INSPECTION SETUP To assess the impact of the proposed spatial regularization procedure, we compute the subband and fullband beampatterns in (6.6) and (6.7) using the MF for three acoustic environments: an anechoic environment, a moderately reverberant environment with $T_{60} = 0.4\,\mathrm{s}$, and a highly reverberant environment with $T_{60} = 1\,\mathrm{s}$ (Section 6.5.1). The time-invariant MFs are computed using (6.5) from the target RTF vector and from the time-averaged RTF vectors estimated by the deep spatio-temporal MVDR filter using either no RTF loss term, the MSE RTF loss term in (6.1), or the Hermitian angle RTF loss term in (6.3).

6.5 Simulation Setup

In this section, we present our simulation setup, consisting of the used datasets (Section 6.5.1), the baseline algorithm (Section 6.5.2), and algorithmic settings (Section 6.5.3).

6.5.1 *Datasets*

To train, validate, and evaluate the proposed spatial regularization procedure, we constructed datasets using diverse speech and noise source material from the DNS 1, DNS 3, and CHiME-3 datasets [66], [68], [70] as well as simulated RIRs. All datasets were used at a sampling rate of 16 kHz.

6.5.1.1 *Training and Validation*

For training and validation, we used the speech and noise source material from the DNS 3 training dataset, consisting of English sentences from 11 350 speakers and 150 noise classes. Additionally, we used multi-microphone noise recordings from the CHiME-3 dataset, which includes real-world acoustic environments such as cafés, street junctions, public transport, and pedestrian areas, representing both quasi-diffuse and localized noise. We simulated 6000 different acoustic scenarios using the `pyroomacoustics` Python package [74], with room length drawn from $\mathcal{U}(3, 8)\,\mathrm{m}$, area drawn from $\mathcal{N}(17.7, 5.5)\,\mathrm{m}^2$ (ensuring a minimum width of 2.5 m), and height drawn from $\mathcal{N}(2.7, 0.8)\,\mathrm{m}$, with $\mathcal{U}(\cdot)$ denoting the uniform distribution. The microphone array geometry was based on the CHiME-3 dataset [70], excluding the rear-facing second channel, resulting in $M = 5$ microphones (Fig. 6.2). The microphone array

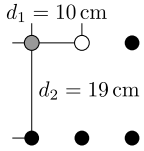


Figure 6.2: Considered CHiME-3 microphone array geometry. Grey circles denote the reference and white circles denote unused microphones.

position was randomized but constrained to be at least 1 m away from walls, with its height drawn from $\mathcal{U}(1, 1.5)$ m.

The positions of the target source and noise sources were randomized as follows. The target source was placed at least 1 m away from the walls, with the height drawn from $\mathcal{U}(1.4, 1.8)$ m. Between 1 and 3 noise sources were placed at least 1 m from the walls and the target source, with their heights drawn from $\mathcal{U}(1.4, 1.8)$ m. The target speech source signal was drawn from the DNS 3 training dataset, while the noise source signals were drawn from the DNS 3 training dataset and the CHiME-3 dataset as follows. First, random noise source signals from the DNS 3 training dataset were assigned to each noise source. With a 50 % probability, an additional random CHiME-3 noise signal was included. With a 10 % probability, only the CHiME-3 noise signal was used; otherwise, the CHiME-3 noise signal was mixed with the DNS 3 noise signals at a ratio drawn from $\mathcal{U}(-10, 10)$ dB. The reverberation times for training ranged from mild ($T_{60} \approx 0.2$ s) to high ($T_{60} \approx 1.0$ s), with a median value of $T_{60} = 0.44$ s (Fig. 6.3). The target speech and DNS 3 noise signals were convolved with the respective simulated RIRs and, including the CHiME-3 noise signal, mixed at input SNRs ranging from -6 dB to 9 dB (selecting the first microphone as the reference microphone). The noisy multi-microphone signals were generated by convolving the target speech and DNS 3 noise source signals with the corresponding simulated RIRs, followed by mixing with CHiME-3 noise signals, with the input SNRs at the first microphone ranging from -6 dB to 9 dB. The total dataset duration was 80 h for training and 20 h for validation, with each utterance lasting 4 s.

6.5.1.2 *Evaluation*

For evaluation, we simulated two types of datasets, namely for objective evaluation (including randomized spatial configurations similar to the training and validation datasets) and for inspection of beampatterns (with a more controlled spatial configuration to facilitate interpretability analysis).

OBJECTIVE EVALUATION For objective evaluation, we simulated two datasets that differ only in terms of the reverberation time: moderately reverberant (median value of $T_{60} = 0.48$ s) and highly reverberant (median value of $T_{60} = 1.01$ s) (Fig. 6.3). The dataset simulation followed the same procedure as for the training

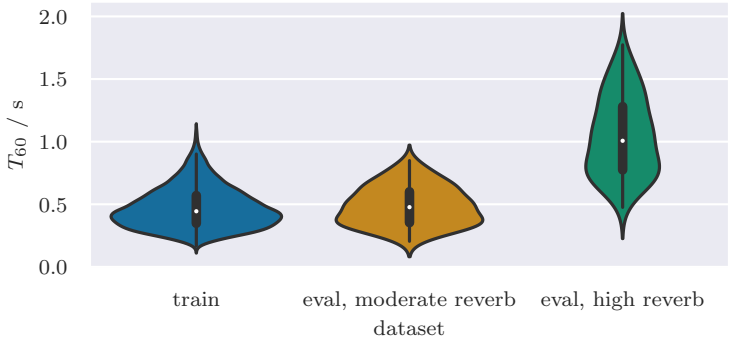


Figure 6.3: Distributions of reverberation times (T_{60}) for the training dataset, the moderately reverberant evaluation dataset, and the highly reverberant evaluation dataset.

and validation datasets, except for the following differences. To avoid overlap regarding the speech and noise source material, we used source material from the DNS 1 evaluation dataset, and we used multi-microphone noise recordings from different recording sessions of the CHiME-3 dataset than for the training and validation datasets. We simulated 100 different acoustic scenarios with the same CHiME-3 microphone array geometry as well as the same distributions for the room configuration and source locations, except that the target source and the noise sources were constrained to the same horizontal plane as the microphone array. Compared to the training dataset, we extended the SNR range of $-9\,\mathrm{dB}$ to $12\,\mathrm{dB}$. The training/validation and evaluation datasets were disjoint in terms of speakers, noise signals, and rooms.

BEAMPATTERN INSPECTION For beampattern inspection, we simulated three acoustic scenarios with different reverberation times: anechoic, moderate ($T_{60} = 0.4\,\mathrm{s}$), and high ($T_{60} = 1.0\,\mathrm{s}$). Unlike the objective evaluation datasets, which used randomized spatial configurations, this dataset was designed with a fixed spatial configuration to allow a controlled inspection of beampatterns (Fig. 6.4). We used a shoe-box room of dimensions $8 \times 8 \times 3\,\mathrm{m}^3$, with the microphone array placed at the center and randomly rotated within the horizontal plane. The target source was placed 2.5 m from the array center at a DOA of 281°, while two noise sources were placed at distances of 2.1 m and 2.4 m at DOAs of 266° and 195°, respectively. All sources were placed at a height of 1.1 m. The target speech and DNS 3 noise signals were convolved with simulated RIRs and mixed at 0 dB input SNR, i.e., no CHiME-3 noise was considered. Following [225], we generated additional RIRs for beampattern evaluation using sources placed on a circle of radius 2.5 m around the array center, with DOAs ranging from 0° to 360° in steps of 2.5°. ATFs computed from these RIRs replaced conventional anechoic steering vectors in the beampattern analysis in (6.6).

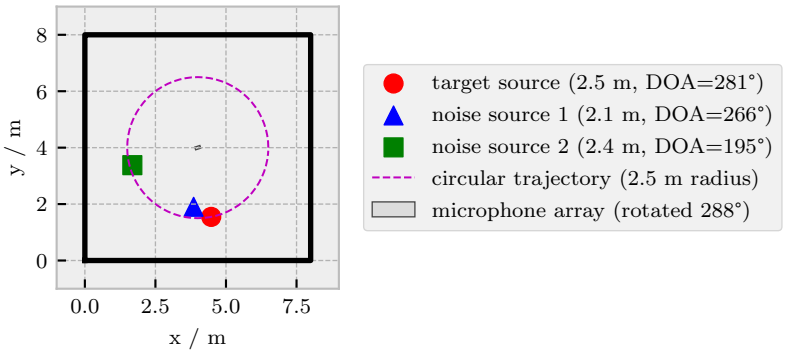


Figure 6.4: Acoustic scenario for beampattern inspection, used at different reverberation times: anechoic, moderate ($T_{60} = 0.4\,\mathrm{s}$), and high ($T_{60} = 1.0\,\mathrm{s}$).

6.5.2 *Baseline Algorithm*

To assess the impact of spatial regularization on the accuracy of RTF vector estimation and the resulting beampatterns, we consider the deep spatio-temporal MVDR filter that does not include an RTF loss term as the baseline multi-microphone speech enhancement algorithm.

6.5.3 *Algorithmic Settings*

STFT PARAMETERS We used an STFT with high temporal resolution, i.e., a frame length of 8 ms and 75 % overlap, with a $\sqrt{\text{Hann}}$ window as analysis and synthesis window. All algorithms used $N = 5$ frames as multi-frame filter length, such that temporal correlations within 16 ms could be exploited. The first microphone was selected as the reference microphone. To limit speech distortion, a minimum gain of $g_{\min} = -20\,\mathrm{dB}$ was applied to the output STFT coefficients of all algorithms during evaluation. The final estimated target speech component was thus obtained as

$$\widehat{x}_t^{\text{fin}} = \begin{cases} g_{\min} y_t, & \text{if } |\mathbf{w}_t^{\mathsf{H}} \mathbf{y}_t| < |g_{\min} y_t| \\ \mathbf{w}_t^{\mathsf{H}} \mathbf{y}_t, & \text{else} \end{cases} . \tag{6.8}$$

DNN TRAINING The considered algorithms were trained using an optional loss term defined on the estimated RTF vector (Section 6.3.1) as well as the speech

enhancement loss term proposed in [108], which comprises time-domain and STFT-domain ℓ_1 terms, i.e.,

$$\mathcal{L} = \frac{\beta}{T_d} \sum_{t_d=1}^{T_d} \left| \widehat{x}_{t_d} - \acute{x}_{t_d} \right| + \frac{1}{FT} \sum_{f=1}^{F} \sum_{t=1}^{T} \left| |\widehat{x}_{f,t}| - |x_{f,t}| \right|, \tag{6.9}$$

where T_d denotes the number of time-domain samples, and $\beta = 10$ is a weight chosen as in [108].[2] The loss was computed after transforming the estimated target speech components back to the time domain using an inverse STFT. For the STFT domain term, an additional transformation was applied with a frame length of 32 ms and 50 % overlap. The reverberant speech component at the reference microphone was selected as the target signal.

DNN ARCHITECTURE We re-implemented and used the full- and subband long short-term memory (FSB-LSTM) architecture and corresponding hyperparameter configuration proposed in [198] for all algorithms. This architecture was selected over the TCN architecture used in previous chapters due to its relatively simple real-time implementation—requiring a buffer of only a single frame—and computational complexity, while still resulting in a strong speech enhancement performance. The FSB-LSTM architecture first applies a 2-dimensional convolutional layer to the input features in (5.25) to obtain a D-dimensional embedding for each STFT bin. These embeddings are then refined through multiple FSB-LSTM blocks, before the subsequent 2-dimensional transposed convolutional layer produces the output. Each of the FSB-LSTM blocks consists of a fullband block and a subband block. The fullband block captures spectro-temporal patterns across all frequencies, while the subband block captures frequency-specific patterns. Residual connections help preserve fine-grained details across each block.

Training was conducted for 15 epochs using the AdamW optimizer [215] with gradient ℓ_2 norms clipped to 5. The one-cycle learning rate schedule [226] was employed, with the learning rate ranging from 10^{-5} to 10^{-3}. Complex-valued numbers were constructed from the (real-valued) FSB-LSTM outputs by assigning separate output elements for the real and imaginary part. The simulations were implemented using PyTorch 2.5.1 [216] and executed on NVIDIA GeForce RTX A5000 graphics cards.

TARGET RTF COMPUTATION The RTF loss terms proposed in Section 6.3.1 rely on a target RTF vector, which is computed using oracle knowledge of the speech component. This oracle knowledge is used only during training to compute the RTF

2 Although the weight β could have been optimized using the uncertainty-based adaptive loss weighting method described in Appendix A, we retained $\beta = 10$ since it was already shown to be effective in [108], eliminating the need for tuning.

loss terms and is not needed during inference. Under the rank-1 assumption of the speech SCM in (2.13), the target RTF vector can be extracted as

$$\check{\mathbf{h}}_t^r = \frac{\check{\mathbf{\Phi}}_{x,t}\check{\mathbf{e}}_r}{\check{\mathbf{e}}_r^{\mathsf{T}}\check{\mathbf{\Phi}}_{x,t}\check{\mathbf{e}}_r}, \tag{6.10}$$

where $\check{\mathbf{\Phi}}_{x,t}$ is computed using recursive smoothing similarly as in (5.32) with a time constant of 2 ms. A more accurate alternative to compute the target RTF vector that does not rely on the rank-1 assumption is the covariance whitening (CW) procedure [227], which computes the RTF vector as the principal eigenvector of the speech SCM $\check{\mathbf{\Phi}}_{x,t}$. While the rank-1 assumption may not perfectly hold for the short STFT frame length used in this chapter, preliminary experiments showed that the rank-1 assumption-based procedure and the CW procedure produced similar target RTF vectors. Given its lower computational complexity, which is particularly advantageous in the deep learning framework used in this chapter, the rank-1 assumption-based procedure was chosen.

6.6 Simulation Results

In this section, we evaluate the speech enhancement performance and RTF vector estimation accuracy of the proposed deep spatio-temporal MVDR filters with prior factorization into the RTF vector and the speech TCV, either using no RTF loss term, the MSE RTF loss term, or the Hermitian angle RTF loss term. To evaluate speech enhancement performance, we consider the wideband PESQ metric [183] and the STOI metric [218], using the reverberant speech component at the reference microphone as the reference signal. To objectively evaluate RTF estimation accuracy, we consider the MSE and the Hermitian angle between the target and estimated RTF vectors, with the target RTF vectors computed using (6.10). Moreover, to subjectively evaluate the plausibility of the estimated RTF vectors, we inspect beampatterns following the procedure in [225] (Section 6.4). It should be noted that the RTF vector estimation accuracy does not necessarily correlate with speech enhancement performance, as the RTF vector is only used to compute the speech STCV required by the spatio-temporal MVDR filter. The DNN might compensate for estimation errors in the estimated RTF vector by adjustments in the other estimated quantities.

6.6.1 *Objective Evaluation*

For the moderately and highly reverberant evaluation datasets, Table 6.1 shows the speech enhancement performance and the RTF vector estimation accuracy, averaged across all utterances. First, it can be observed that all algorithms yield similar and considerable improvements on both datasets in terms of PESQ and STOI compared to the noisy microphone signal. Second, it can be observed that the use of an RTF loss term substantially improves RTF estimation accuracy, as evidenced by lower

Table 6.1: Speech enhancement performance in terms of average wideband PESQ and STOI values and RTF vector estimation accuracy in terms of average MSE and Hermitian angle values for the deep spatio-temporal MVDR filters with prior factorization into the RTF vector and the speech TCV, using no RTF loss term, the MSE RTF loss term, or the Hermitian angle RTF loss term on the moderately and highly reverberant evaluation datasets.

	PESQ-WB		STOI		MSE / dB		Hermitian angle / rad	
algorithm	moderate	high	moderate	high	moderate	high	moderate	high
noisy	1.40	1.55	0.76	0.73	—	—	—	—
no RTF loss	**2.13**	**2.10**	**0.86**	**0.82**	0.80	0.90	1.17	1.17
MSE loss	2.12	2.07	**0.86**	**0.82**	**-17.38**	**-18.46**	**0.21**	**0.21**
Hermitian angle loss	2.12	2.08	**0.86**	**0.82**	-17.24	-18.36	**0.21**	**0.21**

MSE and Hermitian angle values, compared to not using an RTF loss term. The MSE RTF loss term achieves slightly better MSE values than the Hermitian angle RTF loss term, while both loss terms yield similar results in terms of the Hermitian angle. Notably, the use of either the MSE or the Hermitian angle as the RTF loss term does not lead to a significant degradation when using the other as an evaluation metric, suggesting that both loss terms guide the model towards robust improvements in RTF estimation accuracy. Overall, incorporating an RTF loss term improves RTF vector estimation accuracy without sacrificing speech enhancement performance, with the MSE RTF loss term and the Hermitian angle RTF loss term yielding a similar performance.

6.6.2 *Beampattern Inspection*

ANECHOIC SCENARIO For the anechoic scenario, Fig. 6.5 and Fig. 6.6 show the subband and fullband beampatterns, respectively, for matched filters computed from the target RTF vector or from the estimated RTF vector obtained using no RTF loss term, the MSE RTF loss term, or the Hermitian angle RTF loss term.[3] The subband beampatterns reveal that without an RTF loss term, the estimated RTF vector does not exhibit a consistently high response at the target direction. This is acoustically implausible, as the RTF vector should reduce to the steering vector in anechoic conditions (up to near-field scaling effects), which inherently maintains a consistently high response at the target direction. In contrast, if an RTF loss term is used, the estimated RTF vectors yield consistently high responses at the target direction. Additionally, the subband beampatterns computed from the target RTF vector and from the estimated RTF vectors obtained using an RTF loss term exhibit the same general shape, except for scaling differences. This is confirmed by

3 As is commonly done, fullband beampatterns are visualized as polar plots, with the maximum value normalized to 0 dB.

the fullband beampatterns, which show a near-perfect match between the target (orange) and estimated (blue) RTF vectors when an RTF loss term is used. No clear difference is observed between the MSE RTF loss and the Hermitian angle RTF loss.

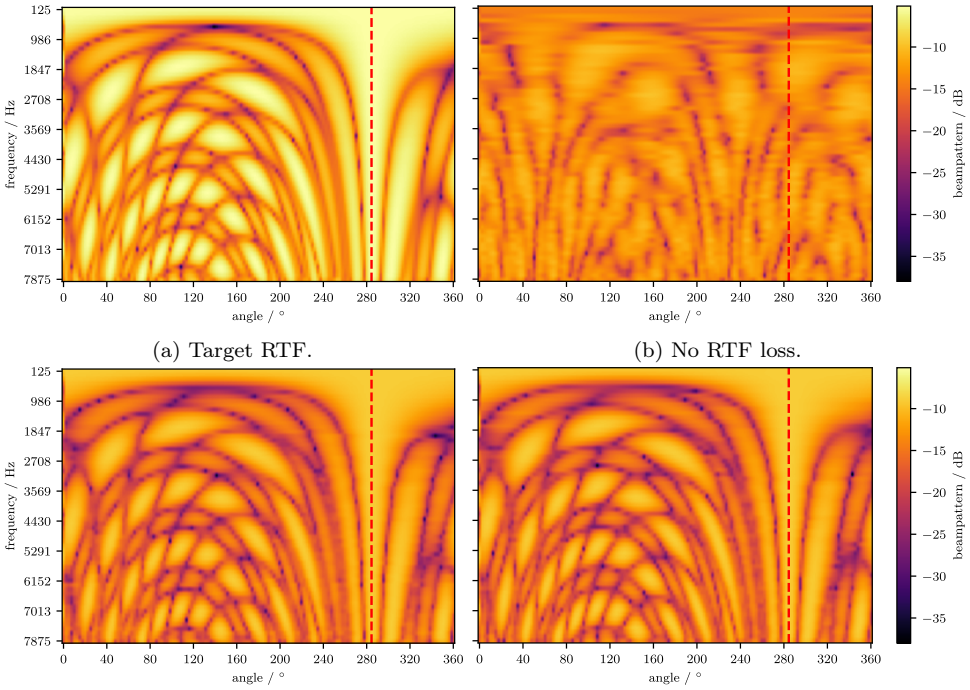
Relating the visual impression with the metrics used to evaluate RTF vector estimation accuracy, the achieved MSE values were 3.84 dB, −12.77 dB, and −12.66 dB, and the achieved Hermitian angle values were 1.16 rad, 0.14 rad, and 0.13 rad for the deep spatio-temporal MVDR filter trained using no RTF loss term, the MSE RTF loss term, and the Hermitian angle RTF loss term, respectively.

MODERATELY REVERBERANT SCENARIO For the moderately reverberant scenario, Fig. 6.7 and Fig. 6.8 show the subband and fullband beampatterns, respectively. Compared to the anechoic scenario, reverberation reduces sharpness and spatial selectivity for the beampatterns computed from both the target and estimated RTF vectors. Nevertheless, the RTF vectors estimated using an RTF loss term still result in a consistently high response at the target direction, whereas the RTF vector estimated without an RTF loss term does not. The overall shape of the subband beampatterns remains similar between the target and estimated RTF vectors if an RTF loss term is used, except for scaling differences, which is again confirmed by the fullband beampatterns. The achieved MSE values were 2.17 dB, −17.36 dB, and −16.25 dB, and the achieved Hermitian angle values were 1.18 rad, 0.21 rad, and 0.21 rad for the deep spatio-temporal MVDR filter trained using no RTF loss, the MSE RTF loss, and the Hermitian angle RTF loss, respectively.

HIGHLY REVERBERANT SCENARIO For the highly reverberant scenario, Fig. 6.9 and Fig. 6.10 show the subband and fullband beampatterns, respectively. The trends observed in the moderately reverberant scenario persist, with even greater reduction in sharpness and spatial selectivity. While the beampatterns computed from the target and estimated RTF vectors obtained using an RTF loss term remain somewhat similar, the discrepancies are more pronounced than in the scenarios with lower reverberation. The achieved MSE values were 1.76 dB, −17.78 dB, and −17.26 dB, and the achieved Hermitian angle values were 1.20 rad, 0.24 rad, and 0.24 rad for the deep spatio-temporal MVDR filter trained using no RTF loss term, the MSE RTF loss term, and the Hermitian angle RTF loss term, respectively.

6.7 Summary

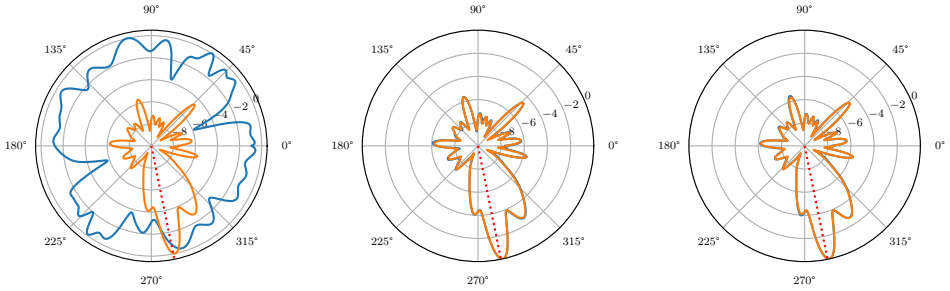
In this chapter, we investigated the acoustic interpretability of estimated RTF vectors in the deep spatio-temporal MVDR filter, focusing on whether they reflect the underlying characteristics of the acoustic scenario. Similarly as for the binaural STWF in Chapter 5, we decomposed the speech STCV into the Kronecker product of an RTF vector and a TCV. While the use of a loss function that incorporates



(a) Target RTF. (b) No RTF loss.

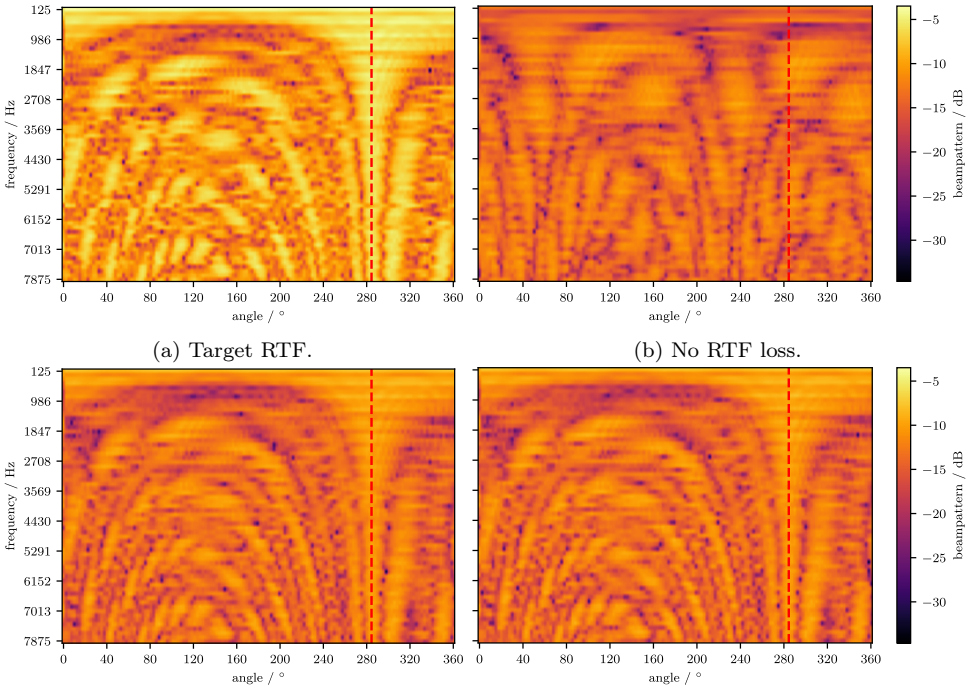
(c) MSE RTF loss. (d) Hermitian angle RTF loss.

Figure 6.5: Beampatterns for the anechoic scenario as a function of frequency and angle for the matched filters computed from the target RTFs or computed from the RTF vector estimated by the deep spatio-temporal MVDR filter trained using no loss function on the RTFs, the MSE RTF loss, or the Hermitian angle RTF loss. The red dashed line indicates the target direction.



(a) No RTF loss. (b) MSE RTF loss. (c) Hermitian angle RTF loss.

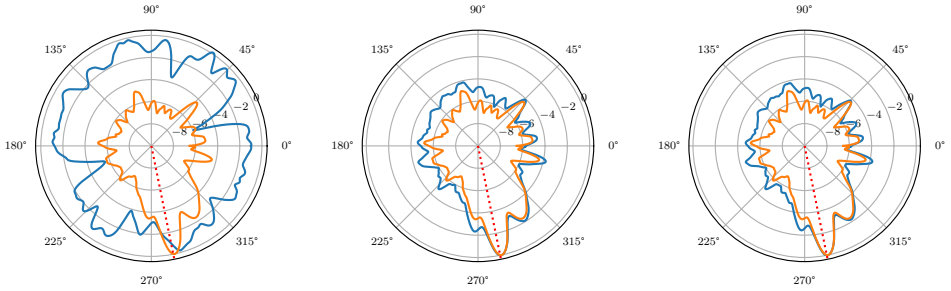
Figure 6.6: Polar plots of beampatterns for the anechoic scenario, averaged across frequencies for the matched filters computed from the RTFs estimated by the deep spatio-temporal MVDR filter trained using no loss function on the RTF, the MSE loss, or the Hermitian angle loss (blue), or computed from the target RTFs (orange). The red dashed line indicates the target direction.



(a) Target RTF. (b) No RTF loss.

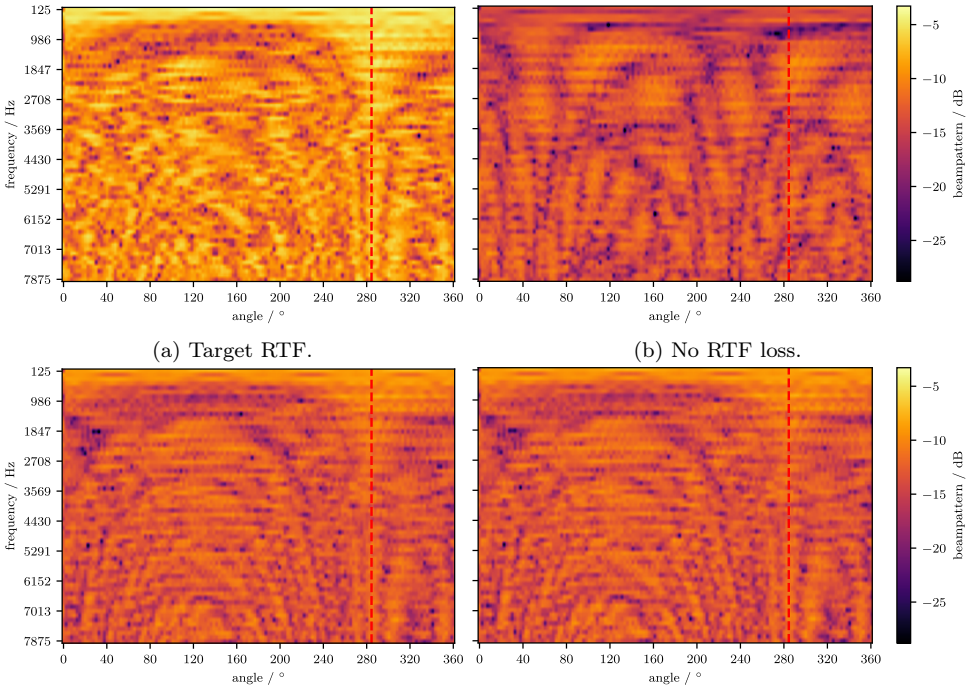
(c) MSE RTF loss. (d) Hermitian angle RTF loss.

Figure 6.7: Beampatterns for the moderately reverberant scenario ($T_{60} \approx 0.4\,\mathrm{s}$) as a function of frequency and angle for the matched filters computed from the target RTFs or computed from the RTF vector estimated by the deep spatio-temporal MVDR filter trained using no loss function on the RTFs, the MSE RTF loss, or the Hermitian angle RTF loss.



(a) No RTF loss. (b) MSE RTF loss. (c) Hermitian angle RTF loss.

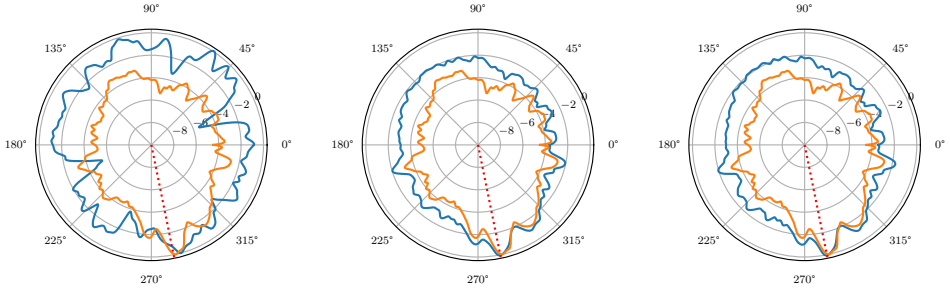
Figure 6.8: Polar plots of beampatterns for the moderately reverberant scenario ($T_{60} = 0.4\,\mathrm{s}$), averaged across frequencies for the matched filters computed from the RTFs estimated by the deep spatio-temporal MVDR filter trained using no loss function on the RTF, the MSE loss, or the Hermitian angle loss (blue), or computed from the target RTFs (orange).



(a) Target RTF. (b) No RTF loss.

(c) MSE RTF loss. (d) Hermitian angle RTF loss.

Figure 6.9: Beampatterns for the highly reverberant scenario ($T_{60} = 1.0\,\mathrm{s}$) as a function of frequency and angle for the matched filters computed from the target RTFs or computed from the RTF vector estimated by the deep spatio-temporal MVDR filter trained using no loss function on the RTFs, the MSE RTF loss, or the Hermitian angle RTF loss.



(a) No RTF loss. (b) MSE RTF loss. (c) Hermitian angle RTF loss.

Figure 6.10: Polar plots of beampatterns for the highly reverberant scenario ($T_{60} = 1.0\,\mathrm{s}$), averaged across frequencies for the matched filters computed from the RTFs estimated by the deep spatio-temporal MVDR filter trained using no loss function on the RTF, the MSE loss, or the Hermitian angle loss (blue), or computed from the target RTFs (orange).

only a signal approximation term resulted in strong speech enhancement performance, the estimated RTF vectors were found to be acoustically implausible, as they did not result in beampatterns with consistently high response towards the target direction. To address this issue and improve interpretability, we proposed a spatial regularization procedure designed to incentivize the DNN to output RTF vector estimates that better reflect the spatial characteristics of the acoustic scenario. This was achieved by incorporating an additional loss term that penalizes discrepancies between the estimated and target RTF vectors. We considered two formulations of this loss term, namely based on the MSE and on the Hermitian angle. To balance the signal approximation and RTF loss terms, we adopted an adaptive weighting method based on homoscedastic uncertainty. We evaluated the proposed spatial regularization procedure in terms of speech enhancement, RTF estimation accuracy and subjective plausibility through the inspection of beampatterns. Simulation results demonstrated that incorporating an RTF-based loss term significantly improved RTF vector estimation accuracy without sacrificing speech enhancement performance. Beampattern inspection further confirmed that the proposed procedure improved the interpretability of the estimated RTF vectors, even in reverberant environments. The proposed approach may be extended to other coupled hybrid speech enhancement approaches to improve the interpretability of estimated quantities.

7

MASK-BASED BEAMFORMER FOR ARBITRARY MICROPHONE ARRAY GEOMETRIES

The mask-based beamformer with ASA [172] can accurately track moving speakers by employing a self-attention mechanism to temporally aggregate instantaneous estimates of the SCMs, allowing to compute time-varying speech and noise SCMs without manual tuning of heuristic temporal aggregation approaches such as recursive smoothing. However, since the employed training procedure, DNN architecture, and input features depend on the channel configuration, the mask-based beamformer with ASA lost the ability to operate with arbitrary configurations, one of the key benefits of conventional mask-based beamformers. Aiming at restoring this key benefit and thereby realizing a mask-based beamformer with ASA for arbitrary microphone arrays, in this chapter we propose three procedures extending the prior work in [172]. First, we investigate incorporating random channel configurations in the training procedure to prevent the DNN from overfitting to specific channel permutations and channel numbers. Second, we propose to employ the TAC method [115] in the ASA module to process multi-microphone features, allowing for any channel number and enabling permutation invariance. The TAC method was originally proposed for channel permutation-invariant multi-microphone source separation and has been successfully employed, e.g., in time-frequency masking algorithms [228], [229] and stationary mask-based beamformers [116]. Third, we investigate utilizing input features that are less sensitive to variations of the channel configuration than the input features in [172]. Through experiments on the CHiME-3 [70] and DEMAND [230] datasets including moving speakers, we demonstrate the benefit of jointly integrating the three proposed approaches into the ASA module. Notably, our proposed approaches not only maintain high performance under matched con-

This chapter is partly based on:

[179] M. Tammen, T. Ochiai, M. Delcroix, T. Nakatani, S. Araki, and S. Doclo, "Array Geometry-Robust Attention-Based Neural Beamformer for Moving Speakers," in *Proc. Interspeech*, Kos, Greece: ISCA, Sep. 2024, pp. 3345–3349.

ditions but also yield a good speech enhancement performance even for microphone arrays unseen during training, consistently outperforming a baseline mask-based beamformer with recursive smoothing and the mask-based beamformer with the original ASA in [172].

The remainder of this chapter is organized as follows. In Section 7.1, the conventional mask-based beamformer with ASA is briefly reviewed and the reason for its channel configuration dependence is explained. In Section 7.2, we propose three procedures to improve the robustness of the mask-based beamformer with ASA against channel configuration variations. The simulation setup and the corresponding results are presented in Section 7.3.

7.1 Conventional Attention Weight Estimation

To compute the MVDR beamformer in (3.24), estimates of the inverse noise SCM $\check{\widehat{\mathbf{\Phi}}}_{n,t}^{-1}$ and the speech SCM $\check{\widehat{\mathbf{\Phi}}}_{x,t}$ are required. To obtain these estimates, the mask-based MVDR beamformer with ASA [172] employs a self-attention mechanism, which allows for a flexible temporal aggregation of instantaneous SCM (ISCM) estimates (see Section 3.3 for a detailed description), i.e.,

$$\boxed{\check{\widehat{\mathbf{\Phi}}}_{\nu,t} = \sum_{\tau=1}^{T} a_{\nu,t,\tau} \check{\widehat{\mathbf{\Psi}}}_{\nu,\tau},} \tag{7.1}$$

where the (frequency-independent) attention weights $a_{\nu,t,\tau}$ control how the ISCM estimates $\check{\widehat{\mathbf{\Psi}}}_{\nu,\tau}$ at time frames $\tau \in 1,\dots,T$ are temporally aggregated to yield estimates of the speech and noise SCMs at time frame t. To obtain these attention weights, a self-attention-based DNN $\mathbf{f}$ with trainable weights $\boldsymbol{\theta}$ (a transformer encoder [102]) is employed, i.e.,

$$\mathbf{a}_t = \mathbf{f}\left(\left\{\boldsymbol{\chi}_t^{\mathrm{ISCM}}\right\}_{t=1}^{T}; \boldsymbol{\theta}\right), \tag{7.2}$$

where $\mathbf{a}_t = [\mathbf{a}_{x,t}^{\mathsf{T}}\ \mathbf{a}_{n,t}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{R}^{2T}$ denotes the vector of attention weights for the speech and noise SCMs at the t-th time frame and $\boldsymbol{\chi}_t^{\mathrm{ISCM}} = \left[\boldsymbol{\chi}_{x,t}^{\mathrm{ISCM},\mathsf{T}}\ \boldsymbol{\chi}_{n,t}^{\mathrm{ISCM},\mathsf{T}}\right]^{T} \in \mathbb{R}^{4FM^2}$ denotes the ISCM-based input features defined in (3.59).[0] As illustrated in Fig. 7.1 (top), the input features are first transformed into a time-varying embedding vector via a fully connected layer. This embedding vector then passes through several MHA encoder blocks, each comprising multi-head self-attention layers and position-wise feedforward layers, which are all interconnected through residual connections.

0 In [172], separate DNNs $\mathbf{f}_x$ and $\mathbf{f}_n$ were used for the speech and noise components. Our preliminary experiments showed a similar or better performance at a lower computational complexity when using a single DNN $\mathbf{f}$.

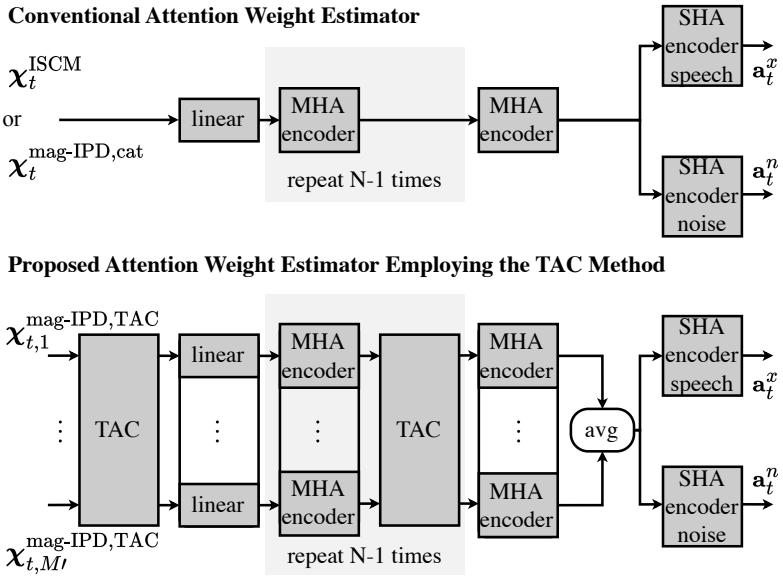


Figure 7.1: Attention weight estimator employing different approaches to process multi-microphone features. The vertically stacked linear and multi-head attention (MHA) encoder modules share trainable weights.

Finally, the attention weights $\mathbf{a}_t$ are extracted from two separate speech and noise single-head attention (SHA) layers. Although the formulation in (7.1) in principle allows for a potential application across different channel configurations, it should be noted that the approach proposed in [172] does depend on the channel configuration. More specifically,

- the training procedure used a fixed channel configuration, not accounting for channel configuration variability.
- the DNN architecture used a fixed input layer size.
- the ISCM features in (3.59) simultaneously incorporate spatial and spectro-temporal information, making them sensitive to the channel configuration considered during training, and the TAC method cannot be applied to the ISCM features.

We define the term channel configuration to include the permutation of channels, the number of channels, and the microphone array geometry.

7.2 Improving Robustness Against Channel Configuration Variations

In this section, we propose three procedures to improve the robustness of the mask-based beamformer with ASA against channel configuration variations, namely training with random channel configurations, employing the TAC method to process multi-microphone features, and using input features that are robust to the channel configuration.

7.2.1 *Training With Random Channel Configurations*

To prevent the DNN from overfitting to specific channel permutations, channel numbers, and microphone array geometries, a straightforward approach is to integrate random channel configurations into the training procedure. Assuming that a single microphone array with $M_{\max}$ channels is available for training, for each minibatch a channel number M' is drawn from the uniform random distribution $\mathcal{U}(2,\ M_{\max})$. From the available $M_{\max}$ channels, M' channels are then selected in random permutation, resulting in random microphone subarrays.

7.2.2 *TAC Method to Process Multi-Channel Features*

To accommodate a variable number of microphones in the training of the DNN with fixed input layer size, zero-padding up to $M_{\max}$ channels can be applied. However, this approach may sacrifice upper bound speech enhancement performance for robustness, since the DNN needs to learn to deal with zero-padded input features, while also being limited to $M' \leq M_{\max}$ channels. To deal with this issue, we propose to employ the TAC method [115] to process multi-microphone features in the attention weight estimator, as depicted in Fig. 7.1 (bottom). A TAC block takes as input a sequence of feature streams $\{\mathbf{z}_{t,m} \in \mathbb{R}^D\}_{m=1}^{M'}$ with variable M' and a channel-independent feature dimension D, shares information across the streams in a non-linearly transformed space, and outputs a sequence of modified feature streams $\{\tilde{\mathbf{z}}_{t,m} \in \mathbb{R}^D\}_{m=1}^{M'}$. We adopt the efficient TAC implementation from [228], obtaining the modified feature stream at time frame t and channel m as

$$\tilde{\mathbf{z}}_{t,m} = \left[\mathrm{ReLU}(\mathbf{L}_1 \mathbf{z}_{t,m})^\mathsf{T} \ \frac{1}{M'} \sum_{\mu=1}^{M'} \mathrm{ReLU}(\mathbf{L}_2 \mathbf{z}_{t,\mu})^\mathsf{T} \right]^\mathsf{T}, \tag{7.3}$$

where ReLU denotes the rectified linear unit activation function and $\mathbf{L}_1, \mathbf{L}_2 \in \mathbb{R}^{(D/2)\times D}$ denote trainable linear transforms shared across streams. $\mathbf{L}_1$ is responsible for extracting stream-specific features, while $\mathbf{L}_2$ is responsible for extracting global features via permutation-invariant averaging across streams.

The TAC method is integrated into the attention weight estimator by interleaving TAC blocks with M' parallel MHA encoder blocks sharing the same parameters

(see Fig. 7.1, bottom). After N stacks of parallel interleaved TAC blocks and MHA encoder blocks, the streams are averaged and passed to the final SHA speech and noise encoder blocks. This integration enables handling a varying channel number M' (even $M' > M_{\max}$) and ensures invariance to the channel permutation. This significantly enhances the flexibility and applicability of the attention weight estimator across diverse channel configurations, without necessitating modifications in the DNN architecture or hyperparameters.

7.2.3 *Input Features*

Due to the definition of the ISCM in (3.48), the input features in (3.59) simultaneously encode inter-microphone level differences as well as inter-microphone phase differences and hence strongly depend on the microphone array geometry. In addition, these features are incompatible with the TAC method, since it requires channel-wise feature streams with a channel number-independent feature dimension. To address these issues, we propose to adopt alternative channel-wise feature streams (denoted as mag-IPD features), defined as

$$\boldsymbol{\chi}_{\nu,f,t,m}^{\text{mag-IPD}} = \left[|\widehat{\nu}_{f,t,m}|^2 \ \cos\left(\widehat{\delta}_{f,t,m}\right) \ \sin\left(\widehat{\delta}_{f,t,m}\right) \right]^{\mathsf{T}} \in \mathbb{R}^3, \tag{7.4}$$

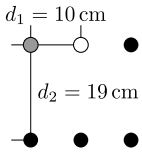
where $\widehat{\delta}_{f,t,m}$ denotes the difference between the unwrapped phases of the masked STFT coefficients $\widehat{\nu}_{f,t,m}$ and the channel-averaged masked STFT coefficients $\widehat{\sigma}_{f,t}$, i.e.,

$$\widehat{\delta}_{f,t,m} = \angle\widehat{\nu}_{f,t,m} - \angle\widehat{\sigma}_{f,t} \tag{7.5}$$

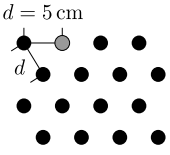
$$\widehat{\nu}_{f,t,m} = m_{\nu,f,t}^{\mathbb{R}} y_{f,t,m} \tag{7.6}$$

$$\widehat{\sigma}_{f,t} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\nu}_{f,t,m}, \tag{7.7}$$

and cos and sin have been applied to result in a smooth phase representation, similarly to the input features used in Chapters 4 and 5. The features in (7.4) differ from those in [228], which used $|\widehat{\sigma}_{f,t}|^2$ instead of $|\widehat{\nu}_{f,t,m}|^2$ and the phase component instead of its cosine and sine, yielding a worse performance in our preliminary experiments. Concatenating the speech and noise features along the frequency dimension yields M streams of $6F$-dimensional features $\boldsymbol{\chi}_{t,m}^{\text{mag-IPD,TAC}}$ (see Fig. 7.1, bottom). We hypothesize that these features are less sensitive to the channel configuration than the features in (3.59) because they do not explicitly depend on channel pairs and they effectively separate the channel configuration-dependent IPD information from magnitude information, which is less influenced by the channel configuration. In addition, we employ the proposed features with the conventional attention weight estimator, in which case we concatenate the speech and noise features along the frequency and channel dimensions, yielding $6FM$-dimensional features $\boldsymbol{\chi}_{t,m}^{\text{mag-IPD,cat}}$ (see Fig. 7.1, top).



(a) CHiME-3 [70]



(b) DEMAND [230]

Figure 7.2: Considered microphone array geometries. Grey circles denote the reference microphone and white circles denote unused microphones.

7.3 Simulations

7.3.1 *Datasets*

To evaluate the effectiveness of the proposed approaches, we constructed datasets of simulated moving speakers in noisy conditions using speech source material from the Wall Street Journal 0 (WSJ0) corpus [231] and noise recordings from the CHiME-3 [70] and DEMAND [230] datasets. We constructed two datasets with different microphone array geometries (illustrated in Fig. 7.2), both with a sampling frequency of 16 kHz. Similarly to [172], we simulated speakers moving on a linear trajectory with constant speed using the gpuRIR tool [75] by generating RIRs at 128 positions on a line, with room width and depth uniformly drawn from 3.0 m, 3.5 m, 4.0 m, 4.5 m, and 5.0 m, room height equal to 2.5 m, reverberation time T_{60} drawn uniformly between 0.1 s to 0.3 s, and the microphone array randomly placed in the room. We added the speech source signal convolved with the simulated RIRs and the recorded noise signals at 2 dB and 8 dB input SNRs.

The first dataset consists of simulated utterances based on the WSJ0 speech and CHiME-3 noise signals, resulting in a maximum number of $M_{\text{max}} = 5$ channels available for training (excluding the rear-facing second channel). This dataset was used for training, development, and evaluation. The second dataset consists of simulated utterances based on the WSJ0 speech and DEMAND noise signals, resulting in 16 available channels. This dataset was only used for evaluation.

During evaluation, we considered a matched condition and several mismatched conditions. "matched" represents the CHiME-3-based evaluation dataset with a fixed channel permutation and the channel number $M' = M_{\text{max}} = 5$, similar to the fixed training condition. To evaluate a mismatch in terms of the channel permutation, we randomly permuted the channels from the CHiME-3-based evaluation dataset. To evaluate a mismatch in terms of the channel number, we selected the first $M' = 3$ channels from the CHiME-3-based evaluation dataset. To evaluate a mismatch in terms of the microphone array geometry, we randomly selected $M' = 5$ channels from the DEMAND-based evaluation dataset. This procedure allows for diverse microphone array geometries, e.g., including linear, triangular, rectangular, and trapezoidal shapes, some of which are not realizable with the CHiME-3 microphone array used for training (see Fig. 7.2). To evaluate a mismatch in terms of both the

channel number and the microphone array geometry, we randomly selected $M' = 3$ channels from the DEMAND-based evaluation dataset. In all evaluation conditions, the reference channel was chosen as depicted in Fig. 7.2. We created 30 000, 2000, and 2000 noisy utterances for training, development, and each evaluation dataset, respectively.

7.3.2 *Settings*

We mostly followed the experimental settings presented in [172] to increase comparability with the associated results. We trained the attention weight estimator using a signal approximation loss function, in particular the scale-dependent SNR loss function [141] at the output of the mask-based beamformer (Fig. 3.3), with the reverberant speech component at the reference microphone as the target signal. During training, we used oracle Wiener-like time-frequency masks [126] to compute the ISCMs in (3.48) and optimized only the trainable weights of the attention weight estimator. During evaluation, we used a time-frequency mask estimator based on a temporal convolutional network architecture [89] (Section 3.2.2.3). For the attention weight and time-frequency mask estimators, we adopted the DNN and training hyperparameters in [172], except for using a single DNN for both the speech and noise components (Section 7.1). For the TAC blocks, we adopted the implementation proposed in [228], consisting of fully connected layer layers and ReLU activation functions (cf. (7.3)).

In addition to the mask-based MVDR beamformer with the original ASA in [172], we considered a mask-based MVDR beamformer with recursive smoothing using a fixed (frequency-independent) smoothing factor that corresponds to a time constant of 1.6 s (tuned according to the highest SDR values under the matched evaluation condition) as a baseline algorithm [172], [203]. Since we rely on the mask-based MVDR beamformer in this chapter, which does not exploit temporal correlations of speech and noise, we used a Hann window with a frame length of 64 ms and 16 ms shift for the STFT, i.e., longer frames and larger frame shift than in all other chapters of this thesis.

We evaluated the speech enhancement performance in terms of PESQ [183] and SDR [182] (allowing for distortions caused by time-invariant filters), with the reverberant speech component at the reference microphone as the reference signal.

7.3.3 *Results*

Table 7.1 shows the averaged PESQ and SDR values for the noisy mixtures, the mask-based beamformer with recursive smoothing (described in the previous section), and for the mask-based beamformer with ASA employing different attention weight estimators (baseline estimator in [172] and proposed estimators). In this table, "config" indicates whether the channel permutation and number were fixed or randomized during training (Section 7.2.1); "features" represents the utilized in-

Table 7.1: Average PESQ and SDR values for the noisy mixtures, a mask-based MVDR beamformer with recursive smoothing using a fixed forgetting factor, and the mask-based MVDR beamformer with ASA employing different attention weight estimators, evaluated on datasets corresponding to a matched condition and various mismatched conditions.

					matched		mismatched in terms of							
							Permutation		Number		Geometry		Number & Geom.	
		config.	features	use TAC	PESQ	SDR	PESQ	SDR	PESQ	SDR	PESQ	SDR	PESQ	SDR
1	mixture	—	—	—	1.37	5.19	1.37	5.19	1.37	5.19	1.38	3.12	1.38	3.12
2	recursive MVDR	—	—	—	2.04	10.18	2.04	10.18	1.73	9.05	2.00	8.94	1.73	7.40
3	baseline [172]	fixed	ISCM	False	**2.64**	16.34	2.31	13.72	1.84	10.66	2.19	11.32	1.71	7.39
4	proposed	fixed	mag-IPD	False	2.57	16.27	2.40	14.70	1.84	10.71	2.15	11.01	1.77	8.34
5	proposed	fixed	mag-IPD	True	2.62	**16.39**	**2.62**	**16.39**	2.05	12.55	2.20	11.84	1.87	9.25
6		random	ISCM	False	2.42	14.37	2.42	14.36	1.96	11.86	2.18	11.55	1.93	10.02
7	proposed	random	mag-IPD	False	2.53	15.85	2.52	15.86	2.32	14.07	2.18	11.82	1.99	10.54
8		random	mag-IPD	True	2.59	16.02	2.59	16.02	**2.34**	**14.15**	**2.21**	**12.35**	**2.03**	**11.34**

put features, either the ISCM features in (3.59) or the proposed mag-IPD features in (7.4); "use TAC" indicates whether TAC was employed or not. We evaluated these beamformers both under a matched condition as well as under various mismatched conditions described in Section 7.3.1.

The results in Table 7.1 show that under all conditions both the mask-based beamformer with recursive smoothing as well as the mask-based beamformer with ASA (for all attention weight estimators) substantially improve the PESQ and SDR values compared to the noisy mixtures. Under the matched condition, it can be observed that models trained with a fixed channel configuration (rows 3–5) achieve the highest PESQ and SDR values. This is expected as these models can exploit the specific spatial information seen during training, representing an upper bound in performance.

Under mismatched conditions, the baseline model (row 3) shows notable performance degradation, particularly in terms of channel number and microphone array geometry. The model employing mag-IPD features (row 4) exhibits a similar performance as the baseline model in most conditions, except for a reduced performance drop under the channel permutation mismatch. The model employing ISCM features with randomized training configurations (row 6) demonstrates similar robustness across mismatched conditions as the model in row 4, albeit with a worse performance under the matched condition, highlighting a trade-off between robustness and upper bound performance. The incorporation of mag-IPD features and the TAC method (row 5) further mitigates performance drops across all mismatch conditions, completely alleviating the degradation under the channel permutation mismatch while maintaining strong matched condition performance. The model combining mag-IPD features, TAC, and randomized training configurations (row 8) achieves the most consistent high performance, performing similarly as the best model under the matched condition and the channel permutation mismatch conditions (row 5), as well as outperforming all models under channel number and microphone array geometry mismatches. The results clearly show that the combina-

tion of training with random channel configurations, employing the TAC method, and using the mag-IPD-based input features resulted in a significantly higher speech enhancement performance compared to the baseline model [172] (significance determined using a two-sided T-test with Bonferroni correction).

It should be emphasized that the evaluation included diverse microphone array geometries by randomly selecting channels from the DEMAND-based evaluation dataset, i.e., "Geometry" and "Number & Geom." in Table 7.1. Hence, the results show that the mask-based beamformer with ASA using the combination of all proposed approaches can perform noise reduction for moving speakers and arbitrary microphone arrays, consistently outperforming the mask-based beamformer with recursive smoothing and the baseline mask-based beamformer with the original ASA.

7.4 Summary

In this chapter, we proposed several approaches to improve the robustness of the mask-based beamformer with ASA against channel configuration variations. These approaches include the integration of random channel configurations during training, employing the TAC method to process multi-microphone features (allowing for any channel number and enabling permutation invariance), as well as using mag-IPD features that are robust against channel configuration variations. Experiments using the CHiME-3 and DEMAND datasets suggest that the mask-based beamformer with ASA integrating the proposed approaches can perform noise reduction for moving speakers and arbitrary microphone arrays. Future research will extend this investigation to explore more diverse channel configurations during training and evaluation as well as address the computational complexity of the proposed TAC integration.

8

CONCLUSIONS AND FURTHER RESEARCH

8.1 Conclusions

Speech enhancement plays a crucial role in modern speech communication applications, improving speech quality and intelligibility across a wide range of acoustic scenarios and devices, including smartphones, smart speakers, and hearing devices. Although model-based speech enhancement approaches offer interpretability and theoretical guarantees, they often struggle in complex, real-world acoustic scenarios where their assumptions are violated. In contrast, learning-based approaches generally achieve higher performance in such scenarios due to their strong representation capacity but may lack interpretability, theoretical guarantees, and robustness when the data observed during inference does not match the training data. Motivated by the potential to combine the interpretability of model-based approaches with the strong representation capacity of learning-based approaches, the primary objective of this thesis was to develop and evaluate hybrid single- and multi-microphone speech enhancement algorithms that employ deep neural networks to estimate the quantities required by a model-based enhancement stage. The main focus was on investigating whether imposing structure on estimated quantities—such as correlation matrix structure, correlation vector structure, or spatial structure—improves speech enhancement performance, interpretability, and computational complexity. Another focus was on developing geometry-robust hybrid speech enhancement algorithms that can operate with arbitrary microphone array configurations. While the proposed algorithms can be used for various speech enhancement applications, the main focus was on hearing devices, where low latency is crucial. To this end, we mainly considered causal multi-frame filters in the STFT domain as the model-based enhancement stage, leveraging their inherent low-latency capabilities and applicability to dynamic acoustic scenarios.

In Chapter 2, we introduced the notation and STFT-domain signal models and described the objective performance measures used to evaluate the speech enhancement algorithms considered throughout this thesis. Specifically, we presented single-frame and multi-frame signal models in both single-microphone and multi-microphone configurations, as well as an extension for binaural hearing devices. Speech enhancement approaches based on single-frame signal models typically as-

sume that consecutive speech STFT coefficients are uncorrelated, which holds approximately when using sufficiently long time frames and large frame shifts. Under these conditions, independent (real-valued) masks can be applied to each STFT coefficient to suppress noise, but at the cost of introducing speech distortion. Mitigating this issue, multi-frame algorithms leverage the fact that speech and noise STFT coefficients indeed do exhibit temporal correlation, especially if using a small frame shift. By jointly processing multiple consecutive STFT frames, these algorithms can achieve noise reduction while preventing speech distortion. This is particularly beneficial in single-microphone configurations, where spatial filtering and distortionless constraints are not available. However, multi-frame signal models can be advantageous also in multi-microphone configurations, allowing to exploit both temporal and spatial correlations for improved speech enhancement performance.

In Chapter 3, we reviewed model-based, learning-based, and hybrid speech enhancement algorithms used throughout this thesis. As examples of model-based speech enhancement algorithms, we reviewed the spatio-temporal MVDR filter, which minimizes the output interference PSD while preserving the speech component, and the STWF, which minimizes the MSE between the output signal and the target speech component. For both filters, performance generally improves for more microphones and time frames, with the STWF yielding a better noise reduction performance than the spatio-temporal MVDR filter at the cost of introducing speech distortion. We also discussed spatial and temporal MVDR filters as special cases of the spatio-temporal MVDR filter and highlighted the challenges of estimating the required quantities for the temporal MVDR filter due to the highly time-varying nature of speech. Further, we discussed the binaural spatial Wiener filter, which minimizes the MSE between the binaural output signals and the target speech components at the left and right hearing device reference microphones, and mentioned that it preserves the binaural cues of the target speech component but changes the binaural cues of the noise component to the cues of the target speech component. As examples of learning-based speech enhancement algorithms, we reviewed the DF algorithm and the Conv-TasNet algorithm. First, avoiding explicit quantity estimation, the DF algorithm employs a DNN to directly estimate complex-valued filters and is readily adaptable to various microphone configurations, including binaural configurations. Second, the Conv-TasNet algorithm utilizes a learned transformation instead of a fixed STFT and estimates real-valued masks that are applied in this learned transform-domain. We detailed the architecture of the Conv-TasNet, including its encoder, TCN-based separator, and decoder, and described important aspects of this architecture, such as depthwise-separable convolutions, dilated convolutions, and cumulative layer normalization. Although arguably a bit outdated, the still strong performance and widespread use of Conv-TasNet made it a valuable baseline algorithm in the context of this thesis, representing purely learning-based approaches that do not rely on predefined signal transforms. Finally, as an example of hybrid speech enhancement algorithms, we reviewed the mask-based MVDR beamformer with ASA, combining the interpretability of model-based beamforming with the representation capacity of learning-based quantity estimation. This algorithm employs a DNN to estimate time-frequency masks and uses a self-attention mechanism to temporally aggregate instantaneous SCM estimates, enabling adapta-

tion to dynamic acoustic scenarios. While this algorithm demonstrates the potential of combining model-based and learning-based approaches, it also highlights a challenge of designing hybrid algorithms, i.e., losing microphone-array configuration independence due to the design of the learning-based quantity estimation stage.

To address the first focus of the thesis, we proposed coupled, structured estimation hybrid approaches for both single-microphone and multi-microphone speech enhancement, which combine a model-based enhancement stage with a learning-based estimation stage that imposes structure on the estimated quantities. In Chapter 4, we first proposed to embed the single-microphone temporal MVDR filter within a deep learning framework, explicitly imposing various structures on the required interference temporal covariance matrix and enabling to exploit the representation capacity of temporal convolutional networks (TCNs) for the difficult estimation task. In Chapter 5, we then extended this coupled, structured estimation approach to the binaural STWF, which can exploit both temporal and spatial correlations, explicitly imposing structure on both the interference spatio-temporal covariance matrices and the speech spatio-temporal correlation vectors. In Chapter 6, we investigated and improved the acoustic plausibility of the estimated RTF vector by imposing spatial structure, achieving accurate RTF vector estimation at no cost to speech enhancement performance or computational complexity. To address the second focus of the thesis, we proposed three procedures to improve the geometry robustness of the mask-based beamformer with attention-based spatial covariance matrix aggregator (ASA), enabling its application to arbitrary microphone array geometries.

In Chapter 4, we proposed a coupled, structured estimation hybrid speech enhancement approach by embedding the fully differentiable single-microphone temporal MVDR filter within an end-to-end deep learning framework. Aiming at exploiting the representation capacity of TCNs, we trained them to estimate the noisy and interference TCMs as well as the a-priori SNR from the noisy speech STFT coefficients, minimizing the signal approximation SI-SDR loss function at the output of the temporal MVDR filter. Since this estimation procedure resulted in strong speech enhancement performance, it was also adopted in Chapters 5 and 6. For the noisy and interference TCMs, we investigated imposing different matrix structures: Hermitian positive-definite (which is a requirement for any covariance matrix), Hermitian positive-definite Toeplitz, and rank-1. For the Hermitian positive-definite structure, we considered estimation procedures based on both recursive smoothing and the Cholesky decomposition. The main differences between the investigated procedures lie in the number of parameters that need to be estimated by the TCNs and the required linear algebra operations, yielding a different computational complexity. We showed that with a rank-1 structure, the temporal MVDR filter can be reformulated as a linear combination of the TCN outputs, avoiding computationally complex matrix inversions and thereby significantly reducing computational complexity. Using the DNS 1 challenge dataset, simulation results demonstrated that

- the TCM estimation procedure using the Hermitian positive-definite structure based on the Cholesky decomposition yields the best performance, while the

rank-1 structure achieves almost the same performance at a lower computational complexity.

- using the proposed coupled hybrid approach led to a substantial improvement in speech enhancement performance compared to a decoupled SPP-driven hybrid approach, demonstrating the benefit of coupling the learning-based estimation stage and the model-based enhancement stage.
- the proposed hybrid approach outperformed a purely learning-based approach that does not impose structure on the multi-frame filter coefficients, demonstrating the benefit of including the model-based enhancement stage in the first place.

In Chapter 5, we extended the coupled, structured estimation hybrid approach from Chapter 4 to binaural speech enhancement by embedding the binaural STWF within an end-to-end deep learning framework. In contrast to the temporal MVDR filter investigated in Chapter 4, which can exploit temporal correlations of speech and noise, the binaural STWF can exploit both temporal and spatial correlations. Aiming at reducing computational complexity while preserving speech enhancement performance and binaural cues, we proposed various procedures to impose spatio-temporal correlation structures on the required interference STCMs and speech STCVs at both hearing devices, which mainly differ in the assumed relationship between microphones (particularly between the left and right device) and the number of parameters that need to be estimated. First, assuming that the spatial correlation of the speech component is stationary over the length of the multi-frame filter, we decomposed the speech STCVs as the Kronecker product of an RTF vector and a TCV, separating the estimation process into a spatial factor and a temporal factor. We either considered a single "global" reference microphone, requiring the speech TCV to be estimated only for this microphone, or a reference microphone for each hearing device, requiring speech TCVs to be estimated for both (left and right) reference microphones. Second, we proposed to replace the left and right interference STCMs with a common interference STCM, as the difference between both STCMs can be assumed to be negligible. Additionally, we considered a bilateral STWF by assuming no spatio-temporal correlation between both hearing devices, both for the speech STCVs and for the interference STCM. Using the DNS 2 and CEC 1 datasets, we first performed validation simulations using oracle speech and noise components to determine the mismatch introduced by the proposed spatio-temporal correlation structures. The validation results showed that

- the speech STCV structure using a single reference microphone incurs a relatively large model mismatch, while the speech STCV structure using a reference microphone for each hearing device incurs a small model mismatch. Although the global RTF structure is commonly used in single-frame algorithms (typically utilizing longer STFT frames), the global RTF structure may be less suitable for the considered multi-frame algorithms, which rely on short STFT frames to exploit the temporal correlations of speech signals.
- the common interference STCM structure incurs a small model mismatch.

- the bilateral speech STCV and interference STCM structures incur a large model mismatch, indicating the importance of contralateral correlations.

Using the DNS 1, DNS 2, CEC 1, and CEC 3 datasets, we then performed an extensive evaluation of the proposed correlation without access to oracle speech and noise components. The simulation results demonstrated that

- the binaural STWF using a combination of the speech STCV structure with two reference microphones and a common interference STCM significantly reduces computational complexity while achieving similar speech enhancement and binaural cue preservation performance compared to not imposing any spatio-temporal correlation structure.
- the proposed binaural STWF outperforms the deep bilateral STWF, the binaural Conv-TasNet algorithm, and the purely learning-based binaural DF algorithm, again confirming the benefit of including the model-based enhancement stage.

While end-to-end training with a signal approximation loss function as in Chapters 4 and 5 is effective for speech enhancement, it does not incentivize the DNN to output quantity estimates that reflect the underlying characteristics of the ground truth quantities. However, one of the key reasons for employing a hybrid speech enhancement approach such as the deep temporal MVDR filter or the deep binaural STWF is interpretability—which is compromised if the estimated quantities are not acoustically plausible. Hence, in Chapter 6, we investigated the acoustic interpretability of the estimated RTF vector in the deep spatio-temporal MVDR algorithm, using the same Kronecker factorization of the speech STCVs as in Chapter 5. The focus on the RTF vector was motivated by the fact that it reflects the underlying spatial characteristics of the acoustic scenario, which are presumably easier to interpret than the temporal correlations of the speech and interference. Instead of analyzing the RTF vectors directly, we assessed their interpretability by deriving matched filters from them and inspecting their corresponding beampatterns. We found the estimated RTF vectors to be acoustically implausible, as they did not result in beampatterns with consistently high response towards the target direction. Hence, we proposed a spatial regularization procedure that incorporates an additional loss term defined on the estimated RTF vector: either based on the MSE or based on the Hermitian angle. This loss term incentivizes the DNN to output estimates that are not only effective for speech enhancement but also reflect the spatial characteristics of the acoustic scenario. To automatically balance the individual loss terms, we employed an adaptive weighting method based on homoscedastic uncertainty. Using the DNS 1 and DNS 2 challenge datasets and simulated RIRs, simulation results demonstrated that the proposed spatial regularization procedure yields accurate estimates of the RTF vector with consistently high response towards the target direction—even in reverberant environments—without sacrificing speech enhancement performance or increasing computational complexity. We hypothesize that this regularization procedure can be extended to other coupled hybrid speech enhancement approaches to improve the acoustic interpretability of estimated quantities.

Finally, in Chapter 7, we proposed three procedures to improve the robustness of the mask-based MVDR beamformer with attention-based spatial covariance matrix aggregator (ASA) against varying microphone array configurations. First, we incorporated random channel configurations during training to prevent the DNN from overfitting to specific channel permutations and channel numbers. Second, we employed the TAC method to process multi-microphone features, allowing the algorithm to adapt to different channel numbers and enabling permutation invariance. Third, we utilized input features that are relatively insensitive to variations in channel configuration. Using the CHiME-3 and DEMAND datasets with simulated moving speakers, simulation results demonstrated that

- combining all three procedures improves generalization to unseen microphone arrays, while speech enhancement performance under matched conditions—where the spatial information provided by a specific microphone array geometry can be exploited—is maintained.
- the mask-based MVDR beamformer combining all three procedures consistently outperforms both a baseline mask-based beamformer with recursive smoothing and the mask-based MVDR beamformer with the original ASA.

8.2 Suggestions for Further Research

While the hybrid single- and multi-microphone speech enhancement approaches proposed in this thesis have demonstrated the benefits of combining a model-based enhancement stage with a learning-based estimation stage, several open research questions remain. One important avenue for further research concerns the relationship between the structure imposed on estimated quantities and the representation capacity of the DNN. While the results in Chapters 4 and 5 consistently showed the benefits of hybrid approaches over purely learning-based methods, the extent of these benefits likely depends on the specific DNN architecture, DNN size, and dataset choice. Future work could systematically investigate how these choices influence the trade-off between interpretability, computational efficiency, and enhancement performance.

Another potential research direction concerns computational complexity, particularly in the context of real-time speech enhancement for resource-constrained devices. While the proposed algorithms (with the exception of the mask-based beamformer with ASA) were designed for low-latency processing and have achieved real-time factors smaller than 1 on a single CPU core, further optimizations could improve feasibility for devices such as hearing devices. Several strategies could be explored to mitigate this issue. On the model-based enhancement side, avoiding computationally complex matrix inversions, as discussed in Chapters 4 and 5, is crucial. On the learning-based estimation side, reducing complexity could involve employing more efficient DNN architectures, such as tiny LSTM networks [96], applying model quantization to lower precision requirements, utilizing model pruning techniques to remove redundant trainable weights and operations, and leveraging

knowledge distillation to train smaller networks that retain the performance of larger models.

A further potential research direction lies in refining the alignment between the model-based enhancement stage and the learning-based estimation stage, as also noted in [32]. In this thesis, the optimization goals of the model-based MVDR and Wiener filters were not directly aligned with the signal approximation loss functions of the learning-based components. Addressing this inconsistency could potentially improve the interpretability of estimated quantities (similar to what was achieved by the spatial regularization procedure proposed in Chapter 6 for the RTF vector). One potentially promising approach could be to refine hybrid approaches by augmenting the signal approximation loss function with loss terms motivated by the model-based component, such as the distortionless constraint of the MVDR beamformer. This approach has shown promise in [101], where the constraint resulted in beamformer coefficients that could be exploited for downstream tasks such as speaker localization.

Improving the robustness of the proposed algorithms against varying microphone array configurations is another promising direction. The deep binaural STWF in Chapter 5 and the deep spatio-temporal MVDR filter in Chapter 6 require the same configuration to be used during training and inference, limiting their applicability in ad-hoc microphone configurations, which are gaining attractiveness with the improving availability of microphones in consumer devices. One possible solution is to follow the approach proposed in Chapter 7, in particular employing the TAC method. More recently, attention-based methods [112], [116] have demonstrated strong potential, as they generalize the fixed averaging operation in the TAC method to a learned attention weighting. This improvement is analogoous to how the mask-based beamformer with attention-based spatial covariance matrix aggregator employs attention for the temporal aggregation of ISCMs instead of applying heuristics.

Another potential research direction is extending the coupled, structured estimation hybrid approach beyond the MVDR and Wiener filters to other optimal filters. As a generalization of the MVDR beamformer, the linearly constrained minimum variance (LCMV) beamformer, for example, allows for the integration of multiple constraints, including the suppression of localized noise sources. While the LCMV beamformer has been extensively studied in model-based speech enhancement [24] and has been applied in a decoupled hybrid approach [232], its integration into a coupled hybrid approach remains unexplored, potentially employing the spatial regularization procedure from Chapter 6 to improve interpretability.

To conclude, this thesis has demonstrated the vast potential of combining the interpretability of model-based approaches with the representation capacity of learning-based approaches for single- and multi-microphone speech enhancement. Future research should continue to refine hybrid approaches, improve computational efficiency, better align optimization objectives, explore alternative optimal filters, and enhance robustness to varying microphone configurations. By addressing these open questions, hybrid speech enhancement approaches could become even more effective and applicable across a wider range of real-world scenarios.

A

APPENDIX TO CHAPTER 6

In multi-task learning, balancing multiple loss terms corresponding to multiple tasks effectively (such as a speech enhancement loss term and an RTF loss term) is crucial for stable training and robust performance for each of the tasks.[1] Traditional approaches often rely on manually tuned weights to combine multiple loss terms, requiring extensive hyperparameter tuning. To address this issue, we adopt the method proposed in [222], which adaptively balances multiple loss terms based on the estimated homoscedastic uncertainty of each task—a form of task-specific uncertainty that remains constant across input samples. In the context of speech enhancement, [170] used this method to balance an SPP loss term with other loss terms defined on the target magnitude spectrum, the Wiener gain, the noise PSD, or the SNR. Although they did not present their results, they claimed that the homoscedastic uncertainty-based weighting method outperformed manual tuning.

Before introducing the homoscedastic uncertainty-based weighting method, it is useful to understand the broader categorization of uncertainty in the training of DNNs:

- Epistemic uncertainty (DNN uncertainty) arises from limited training data or insufficient model capacity. For example, if a DNN is trained primarily on American English speakers, it may struggle to accurately process speech from a speaker with a Scottish accent due to the lack of such examples in its training dataset. Collecting more diverse data or improving the DNN can reduce this uncertainty.
- Aleatoric uncertainty (data uncertainty) is inherent in the data and cannot be reduced by adding more data or improving the model. For instance, in a noisy environment like a busy café, the unpredictable overlap of babble noise with the target speech introduces uncertainty in distinguishing speech from noise. Aleatoric uncertainty can be further categorized into:
  - heteroscedastic uncertainty, which varies with the input signal. For example, sudden loud noises like a door slamming or rapidly changing back-

1 For simplicity, we consider the case of one loss term per task.

ground sounds such as a passing siren can cause spikes in uncertainty during those moments. In contrast, a quiet environment with stable background noise results in lower uncertainty.

– homoscedastic uncertainty, which remains constant across all inputs but varies between tasks. An example is the inherent variability in human speech production—subtle, unpredictable differences in vocal fold vibrations and articulation—that exist in all speech signals, regardless of the acoustic environment. This inherent variability introduces uncertainty in every utterance.

Consider a DNN with trainable weights $\boldsymbol{\theta}$, designed to perform I distinct tasks. The i-th task is associated with an output $\widehat{\mathbf{o}}_i = \mathbf{f}_i(\boldsymbol{\chi};\boldsymbol{\theta})$, where $\mathbf{f}_i$ represents the task-specific mapping between input $\boldsymbol{\chi}$ and output $\widehat{\mathbf{o}}_i$ (e.g., given by different output layers or by a single layer that bundles all outputs in a DNN). A major challenge in multi-task learning is how to effectively balance each loss term's contribution to the overall loss function. Instead of manually tuning weights, a likelihood-based approach can be employed, where each loss term's contribution is based on the estimated homoscedastic uncertainty. For the i-th regression task with target $\mathbf{o}_i \in \mathbb{R}^D$ (e.g., the real and imaginary parts of the target speech STFT coefficients or the RTFs), which is assumed to be drawn from a Gaussian distribution, the likelihood function can be written as

$$p(\mathbf{o}_i|\boldsymbol{\chi},\boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{o}_i; \mathbf{f}_i(\boldsymbol{\chi};\boldsymbol{\theta}), \sigma_i^2\right), \tag{A.1}$$

where $\mathcal{N}$ denotes the Gaussian distribution with task-specific mean $\mathbf{f}_i(\boldsymbol{\chi};\boldsymbol{\theta})$ and variance σ_i^2, termed "uncertainty" in this context. The corresponding log-likelihood function can be written as

$$\log p(\mathbf{o}_i \mid \boldsymbol{\chi},\boldsymbol{\theta}) \propto -\frac{1}{2\sigma_i^2}\left|\mathbf{o}_i - \mathbf{f}_i(\boldsymbol{\chi};\boldsymbol{\theta})\right|^2 - \log(\sigma_i^2). \tag{A.2}$$

where the first term $\left|\mathbf{o}_i - \mathbf{f}_i(\boldsymbol{\chi};\boldsymbol{\theta})\right|^2$ corresponds to the regression loss term, scaled by $1/2\sigma_i^2$, and the second term $\log(\sigma_i^2)$ penalizes σ_i^2 from growing arbitrarily large (which would trivialize the regression term).

In the case of I independent tasks with individual outputs $\mathbf{o}_i$, the overall likelihood function can be factorized as

$$p(\mathbf{o}_1,\ldots,\mathbf{o}_I|\boldsymbol{\chi},\boldsymbol{\theta}) = \prod_{i=1}^{I} p(\mathbf{o}_i|\boldsymbol{\chi},\boldsymbol{\theta}) \tag{A.3}$$

$$= \prod_{i=1}^{I} \mathcal{N}\left(\mathbf{o}_i; \mathbf{f}_i(\boldsymbol{\chi};\boldsymbol{\theta}), \sigma_i^2\right). \tag{A.4}$$

Assuming that the target $\mathbf{o}_i$ for each task follows a Gaussian distribution with mean given by the DNN output $\mathbf{f}_i(\boldsymbol{\chi};\boldsymbol{\theta})$ and variance σ_i^2, training can be interpreted as

maximizing the likelihood of the observed data. Specifically, minimizing the negative log-likelihood of this distribution leads to the loss function

$$\mathcal{L} = -\log(p(\mathbf{o}_1, \ldots, \mathbf{o}_I | \boldsymbol{\chi}, \boldsymbol{\theta})) \tag{A.5}$$

$$\propto \sum_{i=1}^{I} \left[\frac{1}{2\sigma_i^2} \mathcal{L}_i + \log(\sigma_i) \right], \tag{A.6}$$

where $\mathcal{L}_i = \|\mathbf{o}_i - \mathbf{f}_i(\boldsymbol{\chi}; \boldsymbol{\theta})\|^2$ represents the loss term for the i-th task. This formulation ensures that tasks with higher uncertainty σ_i^2 contribute less to the overall loss, while the $\log(\sigma_i^2)$ term prevents arbitrarily large uncertainty values. Crucially, this interpretation holds under the assumption of Gaussian-distributed regression targets; if a different distribution is assumed, the corresponding loss function changes. For example, assuming a Laplace distribution results in an ℓ_1 loss instead of an ℓ_2 loss (see also Section 1.2.2.5), for which the adaptive weighting method was also empirically shown to work well [222].

To allow the loss weights to be tuned automatically, σ_i^2 is optimized jointly with the trainable DNN weights $\boldsymbol{\theta}$. Since directly optimizing σ_i^2 can be numerically unstable, σ_i^2 is reparameterized as

$$s_i = \log\left(\sigma_i^2\right) \Rightarrow \sigma_i^2 = \exp(s_i), \tag{A.7}$$

ensuring that the variance is strictly positive. Substituting (A.7) in (A.6), the overall loss function can be written as

$$\mathcal{L} = \sum_{i=1}^{I} \left[\frac{1}{2\exp(s_i)} \mathcal{L}_i + \frac{1}{2} s_i \right]. \tag{A.8}$$

Since s_i is included in the overall loss function, it is automatically updated via backpropagation, allowing the model to adapt the relative importance of each loss term. Notably, s_i is not an output of the DNN but a directly trainable weight.

Although s_i can be interpreted as homoscedastic uncertainty, it is not explicitly optimized to match a ground-truth uncertainty value. This is similar to the motivation behind the proposed spatial regularization procedure: just as the signal approximation loss term alone does not yield acoustically interpretable RTF estimates, the learned loss weights do not necessarily reflect ground-truth homoscedastic uncertainty.

In the context of Chapter 6, the multi-task learning framework consists of a speech enhancement loss term $\mathcal{L}_{\text{SE}}$ and an RTF estimation loss term $\mathcal{L}_{\text{RTF}}$. Hence, the overall loss function is given by

$$\mathcal{L} = \frac{1}{2\exp(s_{\text{SE}})} \mathcal{L}_{\text{SE}}(\boldsymbol{\theta}) + \frac{1}{2\exp(s_{\text{RTF}})} \mathcal{L}_{\text{RTF}}(\boldsymbol{\theta}) + \frac{1}{2}(s_{\text{SE}} + s_{\text{RTF}}), \tag{A.9}$$

where s_{SE} and s_{RTF} denote reparameterized uncertainty weights in (A.7), which are automatically tuned during training to adjust the relative importance of the speech enhancement and RTF estimation loss terms.

BIBLIOGRAPHY

[1] J. Peissig and B. Kollmeier, "Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners," *The Journal of the Acoustical Society of America*, vol. 101, no. 3, pp. 1660–1670, Mar. 1997.

[2] B. Rutherford, K. K. Brewster, J. S. Golub, A. Kim, and S. Roose, "Sensation and Psychiatry: Linking Age-Related Hearing Loss to Late-Life Depression and Cognitive Decline.," *The American Journal of Psychiatry*, vol. 175 3, pp. 215–224, 2017.

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

[4] 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 Publishers, 2013.

[5] 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.

[6] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, 2018.

[7] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[8] C. Zheng, H. Zhang, W. Liu, X. Luo, A. Li, X. Li, and B. C. J. Moore, "Sixty Years of Frequency-Domain Monaural Speech Enhancement: From Traditional to Deep Learning Methods," *Trends in Hearing*, vol. 27, pp. 1–52, Jan. 2023.

[9] P. Ochieng, "Deep neural network techniques for monaural speech enhancement and separation: State of the art analysis," *Artificial Intelligence Review*, vol. 56, no. S3, pp. 3651–3703, Dec. 2023.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[11] M. Mauch and S. Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 659–663.

[12] L. R. Rabiner, *Digital Processing of Speech Signals*. Englewood Cliffs, N.J: Pearson, Jun. 1978, 1–528.

[13] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, May 2002, pp. I-529-I–532.

[14] P. C. Loizou, *Speech Enhancement: Theory and Practice*. New York, USA: CRC Press, 2007.

[15] Z.-Q. Wang, G. Wichern, S. Watanabe, and J. Le Roux, "STFT-Domain Neural Speech Enhancement With Very Low Algorithmic Latency," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 397–410, 2023.

[16] H. Wu and S. Braun, "Ultra-Low Latency Speech Enhancement—A Comprehensive Study," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India: IEEE, Apr. 2025, pp. 1–5.

[17] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, May 2012.

[18] A. Aroudi and S. Doclo, "Cognitive-Driven Binaural Beamforming Using EEG-Based Auditory Attention Decoding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 862–875, Jan. 2020.

[19] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigne, E. Lalor, B. T. Meyer, S. Miran, T. Francart, and A. Bertrand, "Electroencephalography-Based Auditory Attention Decoding: Toward Neurosteered Hearing Devices," *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 89–102, Jul. 2021.

[20] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, Dec. 2016.

[21] W. Reichardt, O. Alim, and W. Schmidt, "Abhängigkeit der Grenzen zwischen Brauchbarer und Unbrauchbarer Durchsichtigkeit von der Art des Musikmotives, der Nachhallzeit und der Nachhalleinsatzzeit," *Applied Acoustics*, vol. 7, no. 4, pp. 243–264, 1974.

[22] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 3233–3244, Jun. 2003.

[23] H. Kuttruff, *Room Acoustics*. New York, USA: Taylor & Francis, 2000.

[24] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[25] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, 1997.

[26] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1648–1661, Mar. 1992.

[27] D. D. Dirks and R. H. Wilson, "The Effect of Spatially Separated Sound Sources on Speech Intelligibility," *Journal of Speech and Hearing Research*, vol. 12, no. 1, pp. 5–38, Mar. 1969.

[28] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

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

[30] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Chichester, England; Hoboken, NJ: John Wiley, 2006, 625 pp.

[31] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase Processing for Single-Channel Speech Enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, Mar. 2015.

[32] R. Haeb-Umbach, T. Nakatani, M. Delcroix, C. Boeddeker, and T. Ochiai, "Microphone Array Signal Processing and Deep Learning for Speech Enhancement: Combining model-based and data-driven approaches to parameter estimation and filtering," *IEEE Signal Processing Magazine*, vol. 41, no. 6, pp. 12–23, Nov. 2024.

[33] N. L. Westhausen and B. T. Meyer, "Dual-Signal Transformation LSTM Network for Real-Time Noise Suppression," in *Proc. Interspeech*, ISCA, Oct. 2020, pp. 2477–2481.

[34] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Semi-Supervised Multichannel Speech Enhancement With a Deep Speech Prior," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2197–2212, Dec. 2019.

[35] Z. Huang, S. Watanabe, S.-w. Yang, P. Garcia, and S. Khudanpur, "Investigating Self-Supervised Learning for Speech Enhancement and Separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022, pp. 6837–6841.

[36] E. Tzinis, Y. Adi, V. K. Ithapu, B. Xu, P. Smaragdis, and A. Kumar, "RemixIT: Continual Self-Training of Speech Enhancement Models via Bootstrapped Remixing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1329–1341, Oct. 2022.

[37] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[38] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. The MIT Press, Aug. 1949.

[39] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[40] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[41] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, USA, May 2002, pp. 253–256.

[42] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[43] P. C. Loizou and G. Kim, "Reasons why Current Speech-Enhancement Algorithms do not Improve Speech Intelligibility and Suggested Solutions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 47–56, Jan. 2011.

[44] J. Benesty and Y. Huang, "A single-channel noise reduction MVDR filter," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 273–276.

[45] Y. A. Huang and J. Benesty, "A Multi-Frame Approach to the Frequency-Domain Single-Channel Noise Reduction Problem," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1256–1269, May 2012.

[46] K. T. Andersen and M. Moonen, "Robust Speech-Distortion Weighted Interframe Wiener Filters for Single-Channel Noise Reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 97–107, Jan. 2018.

[47] E. A. P. Habets, J. Benesty, and J. Chen, "Multi-microphone noise reduction using interchannel and interframe correlations," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 305–308.

[48] I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[49] T. Gerkmann and R. Hendriks, "Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[50] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, USA, Apr. 2008, pp. 4897–4900.

[51] D. Malah, R. Cox, and A. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Phoenix, AZ, USA, Mar. 1999, 789–792 vol.2.

[52] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved *A Posteriori* Speech Presence Probability Estimation Based on a Likelihood Ratio With Fixed Priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 910–919, Jul. 2008.

[53] A. Schasse and R. Martin, "Estimation of Subband Speech Correlations for Noise Reduction via MVDR Processing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1355–1365, Sep. 2014.

[54] D. Fischer and S. Doclo, "Sensitivity Analysis of the Multi-Frame MVDR Filter for Single-Microphone Speech Enhancement," in *Proc. European Signal Processing Conference (EUSIPCO)*, Kos, Greece, Aug. 2017, pp. 603–607.

[55] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications* (Digital Signal Processing). Berlin, Germany: Springer, 2001.

[56] B. Cornelis, S. Doclo, T. Van dan Bogaert, M. Moonen, and J. Wouters, "Theoretical Analysis of Binaural Multimicrophone Noise Reduction Techniques," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 342–355, Feb. 2010.

[57] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. J. Ray Liu, Eds., John Wiley & Sons, 2010, pp. 269–302.

[58] J. Benesty, J. Chen, and Y. A. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer, 2008.

[59] S. Doclo, S. Gannot, D. Marquardt, and E. Hadad, "Binaural Speech Processing with Application to Hearing Devices," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds., John Wiley & Sons, 2018, pp. 413–442.

[60] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "A Convex Approximation of the Relaxed Binaural Beamforming Optimization Problem," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 321–331, Feb. 2019.

[61] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, no. 2, pp. 251–257, 1991.

[62] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, Nov. 2017.

[63] K. Zmolikova, M. S. Pedersen, and J. Jensen, "Masked Spectrogram Prediction for Unsupervised Domain Adaptation in Speech Enhancement," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 274–283, 2024.

[64] A. Pandey and D. Wang, "On Cross-Corpus Generalization of Deep Learning Based Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2489–2499, Aug. 2020.

[65] P. Gonzalez, T. S. Alstrøm, and T. May, "Assessing the Generalization Gap of Learning-Based Speech Enhancement Systems in Noisy and Reverberant Environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3390–3403, Sep. 2023.

[66] C. K. A. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 2492–2496.

[67] C. K. A. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "ICASSP 2021 Deep Noise Suppression Challenge," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 6623–6627.

[68] C. K. A. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "INTERSPEECH 2021 Deep Noise Suppression Challenge," in *Proc. Interspeech*, Brno, Czech Republic, Aug. 2021, pp. 2796–2800.

[69] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. V. Muñoz, "Clarity-2021 Challenges: Machine Learning Challenges for Advancing Hearing Aid Processing," in *Proc. Interspeech*, Brno, Czech Republic, Aug. 2021, pp. 686–690.

[70] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, USA, Dec. 2015, pp. 504–511.

[71] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Antibes, France, Sep. 2014, pp. 313–317.

[72] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ACE challenge - Corpus description and performance evaluation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA, Oct. 2015, pp. 1–5.

[73] D. Fejgin, W. Middelberg, and S. Doclo, "BRUDEX Database: Binaural Room Impulse Responses with Uniformly Distributed External Microphones," in *Proc. Speech Communication (ITG)*, Aachen, Germany: VDE VERLAG GMBH, Sep. 2023.

[74] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A Python Package for Audio Room Simulation and Array Processing Algorithms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 351–355.

[75] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpuRIR: A python library for room impulse response simulation with GPU acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, Feb. 2021.

[76] T. Wendt, S. Van De Par, and S. Ewert, "A Computationally-Efficient and Perceptually-Plausible Algorithm for Binaural Room Impulse Response Simulation," *Journal of the Audio Engineering Society*, vol. 62, no. 11, pp. 748–766, Dec. 2014.

[77] A. Ratnarajah, S.-X. Zhang, M. Yu, Z. Tang, D. Manocha, and D. Yu, "Fast-Rir: Fast Neural Diffuse Room Impulse Response Generator," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 571–575.

[78] Z.-Q. Wang, H. Erdogan, S. Wisdom, K. Wilson, D. Raj, S. Watanabe, Z. Chen, and J. R. Hershey, "Sequential Multi-Frame Neural Beamforming for Speech Separation and Enhancement," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China, Jan. 2021, pp. 905–911.

[79] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "WHAMR!: Noisy and Reverberant Single-Channel Speech Separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 696–700.

[80] K. Tan and D. Wang, "A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement," in *Proc. Interspeech*, ISCA, Sep. 2018, pp. 3229–3233.

[81] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, Nov. 2013.

[82] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Monaural Speech Dereverberation Using Temporal Convolutional Networks with Self Attention," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1–10, 2020.

[83] J. Pak and J. W. Shin, "Sound Localization Based on Phase Difference Enhancement Using Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1335–1345, May 2019.

[84] K. Tan and D. Wang, "Learning Complex Spectral Mapping with Gated Convolutional Recurrent Networks for Monaural Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 380–390, Nov. 2019.

[85] M. Delfarah and D. Wang, "Features for Masking-Based Monaural Speech Separation in Reverberant Conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1085–1094, May 2017.

[86] J.-M. Valin, U. Isik, N. Phansalkar, R. Giri, K. Helwani, and A. Krishnaswamy, "A Perceptually-Motivated Approach for Low-Complexity, Real-Time Enhancement of Fullband Speech," in *Proc. Interspeech*, ISCA, Oct. 2020, pp. 2482–2486.

[87] H. Schröter, T. Rosenkranz, A.-N. Escalante-B, and A. Maier, "Low Latency Speech Enhancement for Hearing Aids Using Deep Filtering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2716–2728, 2022.

[88] Y. Luo and N. Mesgarani, "TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 696–700.

[89] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[90] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying TasNet: A Dissecting Approach," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain: IEEE, May 2020, pp. 6359–6363.

[91] T. Cord-Landwehr, C. Boeddeker, T. Von Neumann, C. Zorilă, R. Doddipatla, and R. Haeb-Umbach, "Monaural Source Separation: From Anechoic To Reverberant Environments," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022, pp. 1–5.

[92] K. Tesch, N.-H. Mohrmann, and T. Gerkmann, "On the Role of Spatial, Spectral, and Temporal Processing for DNN-based Non-linear Multi-channel Speech Enhancement," in *Proc. Interspeech*, Incheon, Korea: ISCA, Sep. 2022, pp. 2908–2912.

[93] A. Briegleb and W. Kellermann, "Analysis of spatial filtering in neural spatiospectral filters and its dependence on training target characteristics," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, pp. 2–20, Nov. 2024.

[94] J.-M. Valin, "A Hybrid DSP/Deep Learning Approach to Real-Time Full-Band Speech Enhancement," in *Proc. IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, Vancouver, BC: IEEE, Aug. 2018, pp. 1–5.

[95] Y. Xia, S. Braun, C. K. A. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 871–875.

[96] I. Fedorov, M. Stamenovic, C. Jensen, L.-C. Yang, A. Mandell, Y. Gan, M. Mattina, and P. N. Whatmough, "TinyLSTMs: Efficient Neural Speech Enhancement for Hearing Aids," in *Proc. Interspeech*, Shanghai, China: ISCA, Oct. 2020, pp. 4054–4058.

[97] R. Sinha, C. Rollwage, and S. Doclo, "Low-Complexity Real-Time Single-Channel Speech Enhancement Based on Skip-GRUs," in *Proc. ITG Conference on Speech Communication*, Aachen, Germany, Sep. 2023.

[98] S. Bai, J. Z. Kolter, and V. Koltun. "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling." (Apr. 2018), [Online]. Available: `http://arxiv.org/abs/1803.01271`, pre-published.

[99] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 2472–2476.

[100] S. Braun, H. Gamper, C. K. Reddy, and I. Tashev, "Towards Efficient Models for Real-Time Deep Noise Suppression," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada: IEEE, Jun. 2021, pp. 656–660.

[101] H. Chang, Y. Hsu, and M. R. Bai, "Deep Beamforming for Speech Enhancement and Speaker Localization with an Array Response-Aware Loss Function," *Frontiers in Signal Processing*, vol. 4, pp. 1–6, Sep. 2024.

[102] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, Long Beach, USA, Dec. 2017, pp. 5998–6008.

[103] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating Full- and Sub-Band Modeling for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, Aug. 2023.

[104] V. Srinivas, M. Itani, T. Chen, E. S. Eskimez, T. Yoshioka, and S. Gollakota, "Knowledge boosting during low-latency inference," in *Proc. Interspeech*, Kos, Greece: ISCA, Sep. 2024, pp. 4338–4342.

[105] C. Subakan, M. Ravanelli, S. Cornell, F. Grondin, and M. Bronzi, "Exploring Self-Attention Mechanisms for Speech Separation," version 2, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2169–2180, Jun. 2023.

[106] T. Grzywalski and S. Drgas, "Using Recurrences in Time and Frequency within U-net Architecture for Speech Enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 6970–6974.

[107] A. E. Bulut and K. Koishida, "Low-Latency Single Channel Speech Enhancement Using U-Net Convolutional Neural Networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 6214–6218.

[108] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-Attention Dense U-Net for Multichannel Speech Enhancement," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain: IEEE, May 2020, pp. 836–840.

[109] E. J. Nustede and J. Anemüller, "On The Generalization Ability Of Complex-Valued Variational U-Networks For Single-Channel Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1–12, Aug. 2024.

[110] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A Full-Band and Sub-Band Fusion Model for Real-Time Single-Channel Speech Enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, Jun. 2021, pp. 6633–6637.

[111] X. Hao and X. Li. "Fast FullSubNet: Accelerate Full-band and Sub-band Fusion Model for Single-channel Speech Enhancement." (Dec. 2022), [Online]. Available: `http://arxiv.org/abs/2212.09019`, pre-published.

[112] D. Lee and J.-W. Choi, "DeFTAN-AA: Array Geometry Agnostic Multichannel Speech Enhancement," in *Proc. Interspeech*, ISCA, Sep. 2024, pp. 3360–3364.

[113] L. Drude, B. Raj, and R. Haeb-Umbach, "On the Appropriateness of Complex-Valued Neural Networks for Speech Enhancement," in *Proc. Interspeech*, San Francisco, CA, USA: ISCA, Sep. 2016, pp. 1745–1749.

[114] H. Wu, K. Tan, B. Xu, A. Kumar, and D. Wong, "Rethinking Complex-Valued Deep Neural Networks for Monaural Speech Enhancement," in *Proc. Interspeech*, Dublin, Ireland: ISCA, Aug. 2023, pp. 3889–3893.

[115] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end Microphone Permutation and Number Invariant Multi-channel Speech Separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 6394–6398.

[116] A. Jukić, J. Balam, and B. Ginsburg, "Flexible Multichannel Speech Enhancement for Noise-Robust Frontend," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2023, pp. 1–5.

[117] D. Rethage, J. Pons, and X. Serra, "A Wavenet for Speech Denoising," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 5069–5073.

[118] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Tokyo, Japan: IEEE, Sep. 2017, pp. 1–6.

[119] A. Pandey and D. Wang, "Learning Complex Spectral Mapping for Speech Enhancement with Improved Cross-Corpus Generalization," in *Proc. Interspeech*, Shanghai, China: ISCA, Oct. 2020, pp. 4511–4515.

[120] Z.-Q. Wang, P. Wang, and D. Wang, "Complex Spectral Mapping for Single- and Multi-Channel Speech Enhancement and Robust ASR," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1778–1787, Jan. 2020.

[121] A. Pandey, B. Xu, A. Kumar, J. Donley, P. Calamia, and D. Wang, "Multichannel Speech Enhancement Without Beamforming," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore: IEEE, May 2022, pp. 6502–6506.

[122] K. Tan, Z.-Q. Wang, and D. Wang, "Neural Spectrospatial Filtering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 605–621, 2022.

[123] H. Taherian, A. Pandey, D. Wong, B. Xu, and D. Wang, "Multi-input Multi-output Complex Spectral Mapping for Speaker Separation," in *Proc. Interspeech*, ISCA, Aug. 2023, pp. 1070–1074.

[124] Yuxuan Wang and DeLiang Wang, "Towards Scaling Up Classification-Based Speech Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.

[125] A. Narayanan and D. Wang, "Ideal Ratio Mask Estimation Using Deep Neural Networks for Robust Speech Recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada: IEEE, May 2013, pp. 7092–7096.

[126] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-Sensitive and Recognition-Boosted Speech Separation Using Deep Recurrent Neural Networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Queensland, Australia, Apr. 2015, pp. 708–712.

[127] D. Williamson, Y. Wang, and D. Wang, "Complex Ratio Masking for Monaural Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

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

[129] W. Mack and E. A. P. Habets, "Deep Filtering: Signal Extraction and Reconstruction Using Complex Time-Frequency Filters," *IEEE Signal Processing Letters*, vol. 27, pp. 61–65, Nov. 2020.

[130] Y. Luo and N. Mesgarani, "Implicit Filter-and-Sum Network for End-to-End Multi-Channel Speech Separation," in *Proc. Interspeech*, Brno, Czech Republic: ISCA, Aug. 2021, pp. 3071–3075.

[131] A. Li, W. Liu, C. Zheng, and X. Li, "Embedding and Beamforming: All-Neural Causal Beamformer for Multichannel Speech Enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022, pp. 6487–6491.

[132] K. Tesch and T. Gerkmann, "Insights Into Deep Non-Linear Filters for Improved Multi-Channel Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 563–575, 2023.

[133] R. Gu, S.-X. Zhang, Y. Zou, and D. Yu, "Towards Unified All-Neural Beamforming for Time and Frequency Domain Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–14, 2022.

[134] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai: IEEE, Mar. 2016, pp. 5745–5749.

[135] Y. Xu, Z. Zhang, M. Yu, S.-X. Zhang, and D. Yu, "Generalized Spatio-Temporal RNN Beamformer for Target Speech Separation," in *Proc. Interspeech*, Brno, Czech Republic, Aug. 2021, pp. 3076–3080.

[136] W. Meng, X. Li, A. Li, X. Luo, S. Yan, X. Li, and C. Zheng, "Deep Kronecker Product Beamforming for Large-Scale Microphone Arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4537–4553, 2024.

[137] Y. Wang, A. Narayanan, and D. Wang, "On Training Targets for Supervised Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[138] M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. Jensen, "On Loss Functions for Supervised Monaural Time-Domain Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 825–838, 2020.

[139] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Separated Noise Suppression and Speech Restoration: LSTM-Based Speech Enhancement in Two Stages," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2019, pp. 239–243.

[140] C. Boeddeker, W. Zhang, T. Nakatani, K. Kinoshita, T. Ochiai, M. Delcroix, N. Kamo, Y. Qian, and R. Haeb-Umbach, "Convolutive Transfer Function Invariant SDR Training Criteria for Multi-Channel Reverberant Speech Separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada: IEEE, Jun. 2021, pp. 8428–8432.

[141] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?" In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 626–630.

[142] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). New York: Springer, 2006, 738 pp.

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

[144] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," in *Proc. International Conference on Telecommunications and Signal Processing (TSP)*, Jul. 2021, pp. 72–76.

[145] J. M. Martín-Doñas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A Deep Learning Loss Function Based on the Perceptual Evaluation of the Speech Quality," *IEEE Signal Processing Letters*, vol. 25, no. 11, pp. 1680–1684, Nov. 2018.

[146] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative Adversarial Networks based Black-box Metric Scores Optimization for Speech Enhancement," in *Proc. International Conference on Machine Learning (ICML)*, Long Beach, USA: PMLR, May 2019, pp. 2031–2041.

[147] S.-W. Fu, C.-F. Liao, and Y. Tsao, "Learning With Learned Loss Function: Speech Enhancement With Quality-Net to Improve Perceptual Evaluation of Speech Quality," *IEEE Signal Processing Letters*, vol. 27, pp. 26–30, 2020.

[148] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "DNN-Based Source Enhancement to Increase Objective Sound Quality Assessment Score," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1780–1792, Oct. 2018.

[149] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, Jun. 2021, pp. 6493–6497.

[150] C. K. A. Reddy, V. Gopal, and R. Cutler, "Dnsmos P.835: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022, pp. 886–890.

[151] S. Leglaive, M. Fraticelli, H. ElGhazaly, L. Borne, M. Sadeghi, S. Wisdom, M. Pariente, J. R. Hershey, D. Pressnitzer, and J. P. Barker, "Objective and subjective evaluation of speech enhancement methods in the UDASE task of the 7th CHiME challenge," *Computer Speech & Language*, vol. 89, p. 101 685, Jan. 2025.

[152] D. De Oliveira, S. Welker, J. Richter, and T. Gerkmann, "The PESQetarian: On the Relevance of Goodhart's Law for Speech Enhancement," in *Proc. Interspeech*, Kos, Greece: ISCA, Sep. 2024, pp. 3854–3858.

[153] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., Oct. 2020, pp. 12 449–12 460.

[154] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

[155] S. Braun and H. Gamper, "Effect of Noise Suppression Losses on Speech Distortion and ASR Performance," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022, pp. 996–1000.

[156] G. Close, W. Ravenscroft, T. Hain, and S. Goetze, "Perceive and Predict: Self-Supervised Speech Representation Based Loss Functions for Speech Enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.

[157] R. Sutherland, G. Close, T. Hain, S. Goetze, and J. Barker, "Using Speech Foundational Models in Loss Functions for Hearing Aid Speech Enhancement," in *Proc. European Signal Processing Conference (EUSIPCO)*, Lyon, France, Aug. 2024, pp. 421–425.

[158] X. Wang, S. Takaki, and J. Yamagishi, "Neural Source-Filter Waveform Models for Statistical Parametric Speech Synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 402–415, 2020.

[159] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively Trained Recurrent Neural Networks for Single-Channel Speech Separation," in *Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Atlanta, GA, USA, Dec. 2014, pp. 577–581.

[160] H. Erdogan and T. Yoshioka, "Investigations on Data Augmentation and Loss Functions for Deep Learning Based Speech-Background Separation," in *Proc. Interspeech*, Hyderabad, India: ISCA, Sep. 2018, pp. 3499–3503.

[161] 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 191–195.

[162] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3RD CHiME challenge," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, USA, Dec. 2015, pp. 444–451.

[163] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. L. Roux, "Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 1981–1985.

[164] T. Nakatani, R. Takahashi, T. Ochiai, K. Kinoshita, R. Ikeshita, M. Delcroix, and S. Araki, "DNN-supported Mask-based Convolutional Beamforming for Simultaneous Denoising, Dereverberation, and Source Separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 6399–6403.

[165] T. Ochiai, M. Delcroix, R. Ikeshita, K. Kinoshita, T. Nakatani, and S. Araki, "Beam-TasNet: Time-domain Audio Separation Network Meets Frequency-domain Beamformer," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 6384–6388.

[166] Q. Zhang, A. M. Nicolson, M. Wang, K. Paliwal, and C.-X. Wang, "DeepMMSE: A Deep Learning Approach to MMSE-based Noise Power Spectral Density Estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1404–1415, Apr. 2020.

[167] A. Nicolson and K. K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," *Speech Communication*, vol. 111, pp. 44–55, Aug. 2019.

[168] A. Nicolson and K. K. Paliwal, "On training targets for deep learning approaches to clean speech magnitude spectrum estimation," *The Journal of the Acoustical Society of America*, vol. 149, no. 5, pp. 3273–3293, May 2021.

[169] H. Li, D. Wang, X. Zhang, and G. Gao, "Recurrent Neural Networks and Acoustic Features for Frame-Level Signal-to-Noise Ratio Estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2878–2887, 2021.

[170] L. Wang, J. Zhu, and I. Kodrasi, "Multi-task Single Channel Speech Enhancement Using Speech Presence Probability As A Secondary Task Training Target," in *Proc. European Signal Processing Conference (EUSIPCO)*, Dublin, Ireland, Aug. 2021, pp. 296–300.

[171] S. Tao, P. Mowlaee, J. R. Jensen, and M. Græsbøll Christensen, "Learning-Based Multi-Channel Speech Presence Probability Estimation using A Low-Parameter Model and Integration with MVDR Beamforming for Multi-Channel Speech Enhancement," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aalborg, Denmark, Sep. 2024, pp. 100–104.

[172] T. Ochiai, M. Delcroix, T. Nakatani, and S. Araki, "Mask-Based Neural Beamforming for Moving Speakers With Self-Attention-Based Tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 835–848, 2023.

[173] Y. Wang, A. Politis, and T. Virtanen, "Attention-Driven Multichannel Speech Enhancement in Moving Sound Source Scenarios," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Apr. 2024, pp. 11 221–11 225.

[174] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "ADL-MVDR: All Deep Learning MVDR Beamformer for Target Speech Separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 6089–6093.

[175] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, "Exploring Practical Aspects of Neural Mask-Based Beamforming for Far-Field Speech Recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 6697–6701.

[176] M. Tammen and S. Doclo, "Deep Multi-Frame MVDR Filtering for Single-Microphone Speech Enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, Jun. 2021, pp. 8443–8447.

[177] M. Tammen and S. Doclo, "Parameter Estimation Procedures for Deep Multi-Frame MVDR Filtering for Single-Microphone Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3237–3248, Aug. 2023.

[178] M. Tammen and S. Doclo, "Deep Multi-Frame MVDR Filtering for Binaural Noise Reduction," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022, pp. 1–5.

[179] M. Tammen and S. Doclo, "Imposing Correlation Structures for Deep Binaural Spatio-Temporal Wiener Filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 33, pp. 1278–1292, Mar. 2025.

[180] M. Tammen, T. Ochiai, M. Delcroix, T. Nakatani, S. Araki, and S. Doclo, "Array Geometry-Robust Attention-Based Neural Beamformer for Moving Speakers," in *Proc. Interspeech*, Kos, Greece: ISCA, Sep. 2024, pp. 3345–3349.

[181] T. C. Lawin-Ore and S. Doclo, "Reference Microphone Selection for MWF-based Noise Reduction Using Distributed Microphone Arrays," in *Proc. ITG Conference on Speech Communication*, Braunschweig, Germany, Sep. 2012, pp. 31–34.

[182] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[183] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ) - A New Method for Speech Quality Assessment of Telephone Networks and Codecs," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, USA, May 2001, pp. 749–752.

[184] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment," *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 366–384, 2013.

[185] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Quality Index (HASQI)," *Journal of the Audio Engineering Society*, vol. 58, no. 5, pp. 363–381, Jun. 2010.

[186] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. H. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Texas, USA, Mar. 2010, pp. 4214–4217.

[187] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Perception Index (HASPI) Version 2," *Speech Communication*, vol. 131, pp. 35–46, Jul. 2021.

[188] V. Tokala, E. Grinstein, M. Brookes, S. Doclo, J. Jensen, and P. A. Naylor, "Binaural Speech Enhancement Using Deep Complex Convolutional Transformer Networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Apr. 2024, pp. 681–685.

[189] C. Han, Y. Luo, and N. Mesgarani, "Real-Time Binaural Speech Separation with Preserved Spatial Cues," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 6404–6408.

[190] F. P. Itturriet and M. H. Costa, "Perceptually Relevant Preservation of Interaural Time Differences in Binaural Hearing Aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 753–764, Apr. 2019.

[191] B. Van Veen and K. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[192] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., Springer, 2001, pp. 39–60.

[193] M. Souden, J. Benesty, and S. Affes, "New insights into non-causal multichannel linear filtering for noise reduction," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, Apr. 2009, pp. 141–144.

[194] D. Fischer and S. Doclo, "Subspace-Based Speech Correlation Vector Estimation for Single-Microphone Multi-Frame MVDR Filtering," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 856–860.

[195] D. Fischer, K. Brümann, and S. Doclo, "Comparison of Parameter Estimation Methods for Single-Microphone Multi-Frame Wiener Filtering," in *Proc. European Signal Processing Conference (EUSIPCO)*, A Coruña, Spain, Sep. 2019, pp. 1809–1813.

[196] S. Doclo, T. J. Klasen, T. Van den Bogaert, J. Wouters, and M. Moonen, "Theoretical analysis of binaural cue preservation using multi-channel Wiener filtering and interaural transfer functions," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2006, pp. 1–4.

[197] Z. Feng, Y. Tsao, and F. Chen, "Estimation and Correction of Relative Transfer Function for Binaural Speech Separation Networks to Preserve Spatial Cues," in *Proc. APSIPA Annual Summit and Conference (ASC)*, 2021.

[198] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "Neural Speech Enhancement with Very Low Algorithmic Latency and Complexity via Integrated full- and sub-band Modeling," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece: IEEE, Jun. 2023, pp. 1–5.

[199] N. L. Westhausen and B. T. Meyer, "Binaural Multichannel Blind Speaker Separation With a Causal Low-Latency and Low-Complexity Approach," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 238–247, Dec. 2024.

[200] D. Ditter and T. Gerkmann, "A Multi-Phase Gammatone Filterbank for Speech Separation Via Tasnet," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 36–40.

[201] B. J. Borgström, M. S. Brandstein, G. A. Ciccarelli, T. F. Quatieri, and C. J. Smalt, "Speaker separation in realistic noise environments with applications to a cognitively-controlled hearing aid," *Neural Networks*, vol. 140, pp. 136–147, Aug. 2021.

[202] S. Cornell, M. Pariente, F. Grondin, and S. Squartini, "Learning Filterbanks for End-to-End Acoustic Beamforming," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore: IEEE, May 2022, pp. 6507–6511.

[203] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5210–5214.

[204] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 196–200.

[205] Z.-Q. Wang, P. Wang, and D. Wang, "Multi-microphone Complex Spectral Mapping for Utterance-wise and Continuous Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2001–2014, 2021.

[206] J. Casebeer, J. Donley, D. Wong, B. Xu, and A. Kumar. "NICE-Beam: Neural Integrated Covariance Estimators for Time-Varying Beamformers." (Dec. 2021), [Online]. Available: `http://arxiv.org/abs/2112.04613`, prepublished.

[207] G. H. Golub and C. F. Van Loan, *Matrix Computations*. The Johns Hopkins University Press, 1996.

[208] T. Bäckström, "Vandermonde Factorization of Toeplitz Matrices and Applications in Filtering and Warping," *IEEE Transactions on Signal Processing*, vol. 61, no. 24, pp. 6257–6263, Dec. 2013.

[209] F. Vincent and O. Besson, "Steering vector uncertainties and diagonal loading," in *Proc. Sensor Array and Multichannel Signal Processing Workshop*, Barcelona, Spain, Jul. 2004, pp. 327–331.

[210] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Queensland, Australia, Apr. 2015, pp. 5206–5210.

[211] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An Ontology and Human-Labeled Dataset for Audio Events," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, Mar. 2017, pp. 776–780.

[212] F. Font, G. Roma, and X. Serra, "Freesound Technical Demo," in *Proc. ACM International Conference on Multimedia*, New York, NY, USA, Oct. 2013, pp. 411–412.

[213] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-Channel Acoustic Noise database: A Database of Multichannel Environmental Noise Recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, May 2013.

[214] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A Pitch Tracking Corpus with Evaluation on Multipitch Tracking Scenario," in *Proc. Interspeech*, Florence, Italy, Aug. 2011, pp. 1509–1512.

[215] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Proc. International Conference on Learning Representations (ICLR)*, Vancouver, Canada, Sep. 2018, pp. 1–18.

[216] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *Advances in neural information processing systems*, vol. 32, pp. 1–12, Dec. 2019.

[217] S. Zhao, T. H. Nguyen, and B. Ma, "Monaural Speech Enhancement with Complex Convolutional Block Attention Module and Joint Time Frequency Losses," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, Jun. 2021, pp. 6648–6652.

[218] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[219] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, D. S. Williamson, and D. Yu, "Multi-Channel Multi-Frame ADL-MVDR for Target Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3526–3540, Nov. 2021.

[220] J.-M. Lemercier, J. Thiemann, R. Koning, and T. Gerkmann, "A neural network-supported two-stage algorithm for lightweight dereverberation on hearing devices," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, pp. 1–12, May 2023.

[221] M. Herdin, N. Czink, H. Ozcelik, and E. Bonek, "Correlation Matrix Distance, a Meaningful Measure for Evaluation of Non-Stationary MIMO Channels," in *Proc. IEEE Vehicular Technology Conference (VTC)*, Stockholm, Sweden, May 2005, pp. 136–140.

[222] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 7482–7491.

[223] J. H. DiBiase, "A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays," Ph.D. dissertation, Brown University, Providence, Rhode Island, USA, May 2000.

[224] R. Varzandeh, M. Taseska, and E. A. P. Habets, "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation," in *Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Arrays*, San Francisco, USA, Mar. 2017, pp. 11–15.

[225] O. Shmaryahu and S. Gannot, "On The Importance of Acoustic Reflections in Beamforming," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany: IEEE, Sep. 2022, pp. 1–5.

[226] L. N. Smith and N. Topin, "Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006, SPIE, May 2019, pp. 369–386.

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

[228] T. Yoshioka, X. Wang, D. Wang, M. Tang, Z. Zhu, Z. Chen, and N. Kanda, "VarArray: Array-Geometry-Agnostic Continuous Speech Separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 6027–6031.

[229] D. Wang, Z. Chen, and T. Yoshioka, "Neural Speech Separation Using Spatially Distributed Microphones," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 339–343.

[230] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multichannel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. Meetings on Acoustics (ICA)*, Montreal, Canada, 2013, pp. 1–6.

[231] D. B. Paul and J. M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," in *Proc. Workshop on Speech and Natural Language*, New York, USA, Feb. 1992, pp. 357–362.

[232] S. E. Chazan, J. Goldberger, and S. Gannot, "DNN-Based Concurrent Speakers Detector and its Application to Speaker Extraction with LCMV Beamforming," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 6712–6716.

LIST OF PUBLICATIONS

The following publications are related to the work in this thesis.

Peer-Reviewed Journal Papers

[J1] M. Tammen and S. Doclo, "Parameter Estimation Procedures for Deep Multi-Frame MVDR Filtering for Single-Microphone Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3237–3248, Aug. 2023.

[J2] M. Tammen and S. Doclo, "Imposing Correlation Structures for Deep Binaural Spatio-Temporal Wiener Filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 33, pp. 1278–1292, Mar. 2025.

Peer-Reviewed Conference Papers

[C1] M. Tammen, S. Doclo, and I. Kodrasi, "Joint Estimation of RETF Vector and Power Spectral Densities for Speech Enhancement Based on Alternating Least Squares," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 795–799.

[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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 191–195.

[C3] M. Tammen, H. Gode, H. Kayser, E. J. Nustede, N. L. Westhausen, J. Anemuller, and S. Doclo, "Combining weighted binaural LCMP beamforming and deep multi-frame filtering for joint dereverberation and interferer reduction in the Clarity-2021 Challenge," in *Proc. International Clarity Workshop on Machine Learning Challenges for Hearing Aids*, Sep. 2021, pp. 1–6.

[C4] M. Tammen and S. Doclo, "Deep Multi-Frame MVDR Filtering for Single-Microphone Speech Enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, Jun. 2021, pp. 8443–8447.

[C5] H. Gode, M. Tammen, and S. Doclo, "Joint Multi-Channel Dereverberation and Noise Reduction Using a Unified Convolutional Beamformer With Sparse Priors," in *Proc. Speech Communication (ITG)*, 2021, pp. 144–148.

[C6] R. Sinha, M. Tammen, C. Rollwage, and S. Doclo, "Speaker-conditioned Target Speaker Extraction Based on Customized LSTM Cells," in *Proc. Speech Communication (ITG)*, online, Sep. 2021, pp. 89–93.

[C7] M. Tammen and S. Doclo, "Deep Multi-Frame MVDR Filtering for Binaural Noise Reduction," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022, pp. 1–5.

[C8] R. Sinha, M. Tammen, C. Rollwage, and S. Doclo, "Speaker-Conditioning Single-Channel Target Speaker Extraction using Conformer-based Architectures," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022, pp. 1–5.

[C9] M. Tammen, T. Ochiai, M. Delcroix, T. Nakatani, S. Araki, and S. Doclo, "Array Geometry-Robust Attention-Based Neural Beamformer for Moving Speakers," in *Proc. Interspeech*, Kos, Greece: ISCA, Sep. 2024, pp. 3345–3349.

Conference Abstracts

[A1] M. Tammen, I. Kodrasi, and S. Doclo, "Alternating Least Squares-Based Joint Estimation of RETFs and PSDs for Multi-Channel Speech Enhancement," in *International Congress on Acoustics (ICA)*, Aachen, Germany, Sep. 2019.

[A2] M. Tammen and S. Doclo, "Deep Learning-Based Multi-Frame Filtering for Binaural Speech Enhancement," in *International Congress on Acoustics (ICA)*, Gyeongju, South Korea, Oct. 2022.

[A3] M. Tammen and S. Doclo, "Supervised Learning-Based Multi-Frame Filtering for Binaural Speech Enhancement," in *International Hearing Aid Research Conference (IHCON)*, Lake Tahoe, USA, Aug. 2022.

[A4] T. Jansen, N. L. Westhausen, M. Tammen, T. Herzke, V. Hohmann, and H. Kayser, "Evaluation of Deep-Learning-Based Signal Enhancement in Hearing Aids in Complex Acoustic Scenarios," in *International Hearing Aid Research Conference (IHCON)*, Lake Tahoe, USA, Aug. 2024.