

FAU

Friedrich-Alexander-Universität
Faculty of Engineering



Advances in Audio Decorrelation and Rendering of Spatially Extended Sound Sources

**Fortschritte im Bereich der Audio-Dekorrelation und der
Wiedergabe von räumlich ausgedehnten Schallquellen**

Der Technischen Fakultät
der Friedrich-Alexander-Universität
Erlangen-Nürnberg
zur
Erlangung des Doktorgrades Dr.-Ing.
vorgelegt von

Carlotta Anemüller
aus Braunschweig, Deutschland

Als Dissertation genehmigt
von der Technischen Fakultät
der Friedrich-Alexander-Universität Erlangen-Nürnberg
Tag der mündlichen Prüfung: 19.05.2025
Gutachter/in: Prof. Dr. ir. Emanuel Habets
Prof. Dr. Jens Ahrens

Acknowledgments

At the end of this challenging yet rewarding Ph.D. journey, it is time to thank all the colleagues and friends who have accompanied me during this chapter of my life.

First and foremost, I would like to express my gratitude to my supervisor Prof. Emanuël Habets for his support and guidance throughout my research endeavors. I am extremely grateful that he agreed to supervise me when I approached him during the third year of my Ph.D. My sincere thanks also go to Prof. Jürgen Herre for guiding me through the first years of my Ph.D. and for laying the foundation for my further research work by introducing me to the field of spatial audio rendering. I am also very thankful to Oliver Thiergart, who co-supervised most of the work included in this thesis, for his detailed feedback and many fruitful discussions.

Over the years, I have had the pleasure of working with many knowledgeable and friendly people at the AudioLabs and Fraunhofer IIS. I sincerely thank all my current and former colleagues for their support, the constructive discussions, and most importantly, for making the AudioLabs such a great place to work. A special thanks goes to everyone who so willingly participated in my countless listening experiments. I also want to thank the secretarial team for handling all the administrative issues, and Stefan Turowski for managing the IT infrastructure. Thank you, Alexander Adami, for being such a great office mate all these years and for helping me settle in quickly when I started my Ph.D. A very special thank you goes to all my fellow Ph.D. colleagues for their support and all the fun times we shared, both inside and outside the office.

Last but not least, I want to thank my family and friends for always having my back, encouraging me, and cheering me up. I am especially grateful to my parents, Susanne and Rolf, for their continuous support in every way possible, despite the distance between us. Thank you, Patrick, for always believing in me and standing by my side through all the ups and downs of this Ph.D. journey.

Abstract

The aim of immersive spatial audio technologies, as used, e.g., in virtual and augmented reality applications, is to provide the user with an immersive and plausible listening experience. The overall goal is to render the presented three-dimensional sound scenes realistically in a perceptual sense, either over headphones or using a multi-channel loudspeaker setup. Besides a good sound quality, it is essential to consider relevant spatial attributes of the presented sound scenes. One important aspect is the localization of individual sound sources. Additionally, other perceptual aspects of the presented sound scenes need to be considered, including the perceived spatial extent (i.e., “size”) of a sound source and the perceptual impression of the surrounding environment.

From a perceptual point of view, the degree of correlation between the sounds received by the ears is an important factor influencing both the perceived spatial extent of a sound source and the perceptual impression of the surrounding environment. A low correlation is typically associated with an increased size of the auditory event and an enhanced sense of envelopment. This perceptual relevance makes audio decorrelation an important tool within the field of spatial audio rendering to help control the spatial perception of the sound image. This thesis investigates the suitability of neural networks for the task of audio decorrelation. Additionally, methods for binaural rendering of spatially extended sound sources (SESSs) are developed, which employ audio decorrelation techniques.

The first part of this thesis deals with neural network-based approaches to audio decorrelation. Since neural network-based approaches have not previously been applied with the goal of audio decorrelation, we first provide a proof of concept. We therefore propose a convolutional neural network (CNN) architecture, which is trained to mimic the behavior of a state-of-the-art reference decorrelator. By means of a formal listening test, we show that the output of the proposed method is perceptually similar to the output of the reference decorrelator. As a next step, a reference-free method based on generative adversarial networks (GANs) is developed. For the generator network, the same CNN architecture is employed. The training objective is defined directly w.r.t. the input signal and consists of a number of individual loss terms to control both the input-output correlation and the output signal quality. The proposed reference-free approach allows to specifically tailor the training procedure to the desired output signal properties. Finally, the proposed GAN-based audio decorrelation method is extended to provide a multi-channel output signal, which is required in the context of multi-channel spatial audio rendering. A separate generator network is employed for each output channel. All generator networks are optimized jointly to obtain output channels that are mutually uncorrelated and exhibit both a low correlation and a high perceptual similarity to the input signal.

The second part of this thesis introduces methods for binaural rendering of SESSs that are based on audio decorrelation techniques. SESSs can be characterized by means of their radiation behavior. While homogeneous SESSs emit sound with constant radiation characteristics over their extent, heterogeneous

SESSs exhibit a position-dependent radiation behavior. First, a method for efficient rendering of homogeneous SESSs is introduced. By modeling the homogeneous SESS as an incoherently extended sound source with position-independent energy and spectral content, a number of target auditory cues are determined. A binaural output signal with the desired properties is then synthesized by mixing two decorrelated input signals, which can be generated from a single-channel input signal using a single decorrelation filter. Compared to a direct implementation of the rendering model, the proposed approach comes with the advantage of reduced computational complexity and relaxed requirements for the employed decorrelation filters. Second, a method for binaural rendering of heterogeneous SESSs is proposed. Input to the algorithm is a two-channel signal, which provides information about the position-dependent radiation characteristics of the sound source. By extending the homogeneous SESS rendering model to take the position-dependent energy of the sound source into account, a heterogeneous SESS rendering model is defined. Based on this rendering model, the target covariance matrix of the binaural output signal is determined. Using an optimal mixing approach previously proposed in the literature, a binaural output signal with the desired properties is obtained while preserving the spatial characteristics encoded in the two-channel input signal. A formal listening test demonstrates that the output of the proposed method comes close to the simulated binaural reference signal in terms of spatial impression and overall audio quality.

Finally, the suitability of the proposed GAN-based audio decorrelation method for the developed homogeneous SESS rendering method is investigated. To improve the overall audio quality of the decorrelated stereo signal, which serves as the basis for the homogeneous SESS rendering method, an additional loss term is introduced to minimize spectral magnitude differences between the channels of the decorrelated stereo signal.

Kurzfassung

Das Ziel immersiver räumlicher Audiotechnologien, wie sie beispielsweise in Anwendungen der virtuellen und erweiterten Realität zum Einsatz kommen, besteht in der Schaffung eines immersiven und plausiblen Hörerlebnisses für den Nutzer. Das übergeordnete Ziel ist eine im Sinne der Wahrnehmung realistische Wiedergabe der dargestellten dreidimensionalen Klangszenen, entweder über Kopfhörer oder über ein mehrkanaliges Lautsprechersystem. Neben einer guten Klangqualität ist die Berücksichtigung relevanter räumlicher Attribute der dargestellten Klangszenen essentiell. Ein wichtiger Aspekt ist dabei die Lokalisierung einzelner Schallquellen. Darüber hinaus müssen weitere perzeptuelle Aspekte der dargestellten Klangszenen berücksichtigt werden, wie beispielsweise die wahrgenommene räumliche Ausdehnung (d. h. die „Größe“) einer Schallquelle und der wahrgenommene Umgebungseindruck.

In Hinblick auf die Wahrnehmung spielt der Grad der Korrelation zwischen den von den Ohren empfangenen Schallwellen eine wesentliche Rolle. Dieser beeinflusst sowohl die wahrgenommene räumliche Ausdehnung einer Schallquelle als auch den wahrgenommenen Umgebungseindruck. Eine geringe Korrelation geht in der Regel mit einer größeren Ausdehnung des Hörereignisses sowie einem stärkeren Gefühl der Umhüllung einher. Diese Wahrnehmungsrelevanz macht Audio-Dekorrelation zu einem wichtigen Werkzeug im Bereich der räumlichen Audiowiedergabe, insbesondere zur Steuerung der räumlichen Wahrnehmung des Klangbildes. Im Rahmen dieser Arbeit wird die Eignung neuronaler Netze zur Audio-Dekorrelation untersucht. Des Weiteren werden Methoden zur binauralen Wiedergabe von räumlich ausgedehnten Schallquellen (engl.: spatially extended sound sources (SESSs)) entwickelt, welche Audio-Dekorrelationstechniken einsetzen.

Der erste Teil dieser Arbeit beschäftigt sich mit Ansätzen zur Audio-Dekorrelation basierend auf neuronalen Netzen. Da diese zuvor noch nicht mit dem Ziel der Audio-Dekorrelation eingesetzt wurden, erfolgt zunächst ein Eignungsnachweis. Zu diesem Zweck wird ein faltendes neuronales Netzwerk (engl.: convolutional neural network (CNN)) vorgeschlagen, welches dahingehend optimiert wird, das Verhalten eines modernen Referenzdekorrelators nachzuahmen. Die Ergebnisse eines formalen Hörversuchs zeigen, dass das Ausgangssignal der vorgeschlagenen Methode und das des Referenzdekorrelators eine hohe perzeptuelle Ähnlichkeit aufweisen. Im nächsten Schritt wird eine referenzfreie Methode entwickelt, welche auf generativen gegnerischen Netzwerken (engl.: generative adversarial networks (GANs)) basiert. Für das Generator-Netzwerk wird die gleiche CNN-Architektur verwendet. Die Zielfunktion wird direkt in Bezug auf das Eingangssignal definiert und besteht aus einer Reihe von individuellen Komponenten, welche darauf abzielen, sowohl die Eingangs-Ausgangs-Korrelation als auch die Qualität des Ausgangssignals zu kontrollieren. Der referenzfreie Ansatz ermöglicht eine gezielte Anpassung des Trainingsverfahrens an die gewünschten Eigenschaften des Ausgangssignals. Schließlich erfolgt eine Erweiterung der vorgeschlagenen GAN-basierten Audio-Dekorrelationsmethode auf ein mehrkanaliges Ausgangssignal. Dies wird insbesondere im Kontext einer mehrkanaligen räumlichen Audiowiedergabe

gabe benötigt. Für jeden Ausgangskanal wird ein separates Generator-Netzwerk verwendet, wobei eine gemeinsame Optimierung aller Generator-Netzwerke erfolgt. Dadurch wird erreicht, dass die Ausgangskanäle untereinander unkorreliert sind und sowohl eine geringe Korrelation als auch eine hohe perzeptuelle Ähnlichkeit mit dem Eingangssignal aufweisen.

Im zweiten Teil dieser Arbeit werden Methoden zur binauralen Wiedergabe von SESSs vorgestellt, welche auf Audio-Dekorrelationstechniken basieren. In Abhängigkeit des Abstrahlverhaltens wird dabei zwischen homogenen und heterogenen SESSs unterschieden. Während homogene SESSs Schall mit konstanter Abstrahlcharakteristik emittieren, weisen heterogene SESSs ein positionsabhängiges Abstrahlverhalten auf. Zunächst wird eine Methode zur effizienten Wiedergabe von homogenen SESSs vorgestellt. Durch die Modellierung der homogenen SESS als inkohärent ausgedehnte Schallquelle mit positionsunabhängiger Energie und spektralem Inhalt erfolgt die Definition einer Reihe von Ziel-auditiven Cues. Ein binaurales Ausgangssignal mit den gewünschten Eigenschaften wird anschließend durch Mischen zweier dekorrelierter Eingangssignale generiert. Die zwei dekorrelierten Eingangssignale werden dabei mit Hilfe eines einzigen Dekorrelationsfilters aus einem einkanaligen Eingangssignal erzeugt. Im Vergleich zu einer direkten Implementierung des Wiedergabemodells weist der vorgeschlagene Ansatz eine niedrigere Berechnungskomplexität sowie geringere Anforderungen an die verwendeten Dekorrelationsfilter auf. Im Anschluss wird eine Methode zur binauralen Wiedergabe von heterogenen SESSs vorgeschlagen. Die Methode basiert auf einem zweikanaligen Eingangssignal, welches Informationen über die positionsabhängigen Abstrahlcharakteristika der Schallquelle liefert. Auf Basis eines heterogenen SESS-Wiedergabemodells wird die Ziel-Kovarianzmatrix des binauralen Ausgangssignals bestimmt. Dazu wird das homogene SESS-Wiedergabemodell dahingehend erweitert, dass die position-sabhängige Energie der Schallquelle Berücksichtigung findet. Unter Verwendung eines in der Literatur vorgeschlagenen optimalen Mixing-Ansatzes wird ein binaurales Ausgangssignal mit den gewünschten Eigenschaften bestimmt. Dabei wird sichergestellt, dass die im zweikanaligen Eingangssignal kodierten räumlichen Charakteristika erhalten bleiben. Die Ergebnisse eines formalen Hörversuchs zeigen, dass das Ausgangssignal der vorgeschlagenen Methode der binauralen Referenz hinsichtlich des räumlichen Eindrucks und der Audioqualität nahekommt.

Im Rahmen einer abschließenden Untersuchung erfolgt eine Evaluierung der vorgeschlagenen GAN-basierten Audio-Dekorrelationsmethode hinsichtlich ihrer Eignung für den Einsatz in der entwickelten homogenen SESS-Wiedergabemethode. Zur Optimierung der Audioqualität des dekorrelierten Stereosignals, welches als Grundlage für die homogene SESS-Wiedergabemethode dient, wird eine zusätzliche Komponente in die Zielfunktion integriert, welche die spektralen Unterschiede zwischen den Kanälen des dekorrelierten Stereosignals minimiert.

Contents

Acknowledgments	i
Abstract	iv
Kurzfassung	vi
Glossary of Acronyms, Symbols, and Notation	xi
List of Figures	xviii
List of Tables	xix
1 Introduction	1
1.1 Scope and Motivation	1
1.2 Main Contributions and Outline	2
1.3 List of Publications and Contribution Statements	3
2 Background and State of the Art	7
2.1 Spatial Hearing	7
2.1.1 Head-Related Spherical Coordinate System	7
2.1.2 Auditory Cues	7
2.1.3 Horizontal Localization	8
2.1.4 Vertical Localization and Front-Back Differentiation	9
2.1.5 Distance Perception	10
2.1.6 Spatial Impression	11
2.1.7 Just-Noticeable Differences of Binaural Cues	13
2.1.8 Head-Related Transfer Functions	14
2.2 Audio Decorrelation	15
2.2.1 Time-Domain Methods	15
2.2.2 Frequency-Domain Methods	16
2.2.3 Challenges and Limitations	18
2.3 Rendering of Spatially Extended Sound Sources	18
2.3.1 Homogeneously Extended Sound Sources	19
2.3.2 Heterogeneously Extended Sound Sources	21
2.3.3 Challenges and Limitations	22

3	Reference-Based Neural Audio Decorrelation	23
3.1	Problem Statement	24
3.2	Proposed Method	24
3.2.1	Reference Method	25
3.2.2	Model Architecture	25
3.2.3	Loss Function	26
3.3	Datasets and Training	26
3.4	Performance Evaluation	27
3.4.1	Objective Evaluation	27
3.4.2	Perceptual Evaluation	29
3.5	Summary	30
4	Single-Channel Audio Decorrelation Using Generative Adversarial Networks	33
4.1	Problem Statement	34
4.2	Proposed Method	34
4.2.1	Generator Architecture	35
4.2.2	Discriminator Architecture	35
4.2.3	Generator Loss Functions	36
4.2.4	Discriminator Loss Function	37
4.3	Datasets and Training	37
4.4	Performance Evaluation	38
4.4.1	Objective Evaluation	38
4.4.2	Perceptual Evaluation	40
4.4.3	Discussion on Stereo Signal Quality	42
4.5	Summary	44
5	Multi-Channel Audio Decorrelation Using Generative Adversarial Networks	45
5.1	Problem Statement	46
5.2	Proposed Method	46
5.2.1	Generator Architecture	47
5.2.2	Discriminator Architecture	47
5.2.3	Generator Loss Functions	48
5.2.4	Discriminator Loss Function	50
5.3	Datasets and Training	50
5.4	Performance Evaluation	51
5.4.1	Comparison Methods	51
5.4.2	Independent Channel Optimization	52
5.4.3	Objective Evaluation	52
5.4.4	Perceptual Evaluation	57
5.4.5	Discussion	63
5.5	Summary	63

6	Binaural Rendering of Homogeneously Extended Sound Sources	65
6.1	Rendering Model	66
6.2	Proposed Rendering Method	68
6.2.1	Processing Steps	69
6.2.2	Time-Domain Implementation	71
6.3	Performance Evaluation	73
6.3.1	Evaluation Setup	73
6.3.2	Influence of Phase Smoothing	74
6.3.3	Objective Evaluation	75
6.3.4	Perceptual Evaluation	76
6.3.5	Discussion	80
6.4	Application in VR/AR Standardization	81
6.5	Summary	82
7	Binaural Rendering of Heterogeneously Extended Sound Sources	83
7.1	Rendering Model	84
7.2	Proposed Rendering Method	85
7.3	Performance Evaluation	87
7.3.1	Evaluation Setup	87
7.3.2	Comparison Methods	88
7.3.3	Objective Evaluation	89
7.3.4	Perceptual Evaluation	91
7.4	Summary	93
8	Application of GAN-based Audio Decorrelation to Binaural Rendering of Homogeneously Extended Sound Sources	95
8.1	Single-Channel GAN-Based Audio Decorrelation	96
8.1.1	Overview	96
8.1.2	Extended Generator Loss	96
8.2	Performance Evaluation	97
8.2.1	Evaluation Setup	97
8.2.2	Considered Decorrelation Methods	98
8.2.3	Objective Evaluation	98
8.2.4	Perceptual Evaluation	101
8.3	Summary	105
9	Conclusions and Outlook	107
9.1	Conclusions	107
9.2	Future Research	109
A	Supplemental Material	115
A.1	Decorrelation Tree Structure	115
A.1.1	Single Decorrelator Instance	115
A.1.2	Multiple Decorrelator Instances	117

A.2 Objective Evaluation Plots Complementing Section 7.3.3	119
Bibliography	121

Acronyms, Symbols, and Notation

Acronyms

3DoF	three-degrees-of-freedom
6DoF	six-degrees-of-freedom
AR	augmented reality
ASW	apparent source width
BRIR	binaural room impulse response
CI	confidence interval
CNN	convolutional neural network
DFT	discrete Fourier transform
DirAC	directional audio coding
FD-SDR	frequency-domain signal-to-distortion ratio
FIR	finite impulse response
GAN	generative adversarial network
HRIR	head-related impulse response
HRTF	head-related transfer function
IC	interaural coherence
ICC	interaural cross-correlation
IDFT	inverse discrete Fourier transform
IIR	infinite impulse response
ILD	interaural level difference
IPD	interaural phase difference
ITD	interaural time difference
JND	just-noticeable difference
LSD	log-spectral distance
ODG	objective difference grade
OTT	one-to-two
PEAQ	perceptual evaluation of audio quality
PSD	power spectral density
QMF	quadrature mirror filter
ReLU	rectified linear unit
RMSE	root-mean-square error
SDG	subjective difference grade

SESS	spatially extended sound source
STFT	short-time Fourier transform
TD-SDR	time-domain signal-to-distortion ratio
VR	virtual reality

Symbols

C	number of channels
c	channel index
\mathbf{C}_x	covariance matrix of frequency-domain multi-channel signal \mathbf{x}
c_{xy}	correlation between two time-domain signals x and y
d	discrete time delay of correlation function
e	Euler's number
φ	azimuth angle
f_s	sampling rate
G	ear gain
H, \mathbf{h}	head-related transfer function
j	imaginary unit
K	number of frequency bins
k	discrete frequency index
L	number of time frames
l	discrete time frame index
\mathcal{L}	loss term
λ	loss term weight
\mathbf{M}	mixing matrix
M	number of mel frequency bands
m	mel-scale frequency index
N	number of time samples
n	discrete time index
P	power spectral density
q	filter length
ρ_{xy}	coherence between two frequency-domain signals X and Y
s, S	source signal
t	continuous time index
θ	elevation angle
τ	continuous time delay of correlation function
$\mathbf{u} = [\varphi, \theta]^T$	direction, comprising azimuth and elevation angles
w, W	processing filter
x, X, \mathbf{x}	input signal
y, Y, \mathbf{y}	output signal

Mathematical Notation

a, b, c	time-domain signals
A, B, C	frequency-domain signals
$\mathbf{a}, \mathbf{b}, \mathbf{c}$	column vectors
$\mathbf{A}, \mathbf{B}, \mathbf{C}$	matrices
$(\cdot)^*$	complex conjugate
$(\cdot)^H$	complex (Hermitian) transpose
$(\cdot)^T$	transpose
$\angle(\cdot)$	phase of a complex number
$*$	convolution operator
\odot	element-wise multiplication (Hadamard product)
$\bar{\cdot}$	average value
$ \cdot $	absolute value, magnitude of a complex number
$\text{circshift}(\cdot)_{q/2}$	circular shift of $q/2$ samples
$\text{dec}\{\cdot\}$	decorrelation operation
$\mathcal{D}_i(\cdot)$	processing with i -th discriminator
$\mathcal{E}\{\cdot\}$	statistical expectation
$\log_{10}(\cdot)$	logarithm to the base 10
$\text{movmean}(\cdot)$	moving average filter of length ν
$\mathcal{T}_m(\cdot)^\nu$	transformation from linear to mel frequency scale
$\text{unwrap}\{\cdot\}$	phase unwrapping operation

List of Figures

1.1	Overview of thesis outline.	3
2.1	Head-related spherical coordinate system.	8
2.2	Illustration of acoustic propagation effects important for horizontal localization.	9
2.3	Illustration of a cone of confusion. Points on the surface of the cone represent positions with a constant distance difference to the two ears. Points on any circular cross-section of the cone represent positions that are equidistant from the left ear and equidistant from the right ear.	10
3.1	Block diagram of proposed reference-based neural audio decorrelation method.	24
3.2	Boxplots of objective evaluation metrics for all three training datasets and both test datasets. The horizontal lines represent the median values, and the boxes represent the first to third quartiles. The whiskers are within 1.5 times the inter-quartile range; outliers are depicted by diamonds.	28
3.3	Relative frequency of listening test scores grouped by signal type, ratings of proposed method w.r.t. reference method.	30
4.1	Block diagram of proposed single-channel GAN-based audio decorrelation method.	34
4.2	Mean and standard deviation of absolute coherence over frequency between input and output signal for proposed method (trained on music signals) and MPEG-I decorrelator for all three test datasets.	38
4.3	Mean and standard deviation of absolute coherence over frequency between channels of stereo signal for proposed method (trained on music signals) and MPEG-I decorrelator for all three test datasets.	39
4.4	Relative frequency of stereo signal listening test scores for all items in terms of perceived envelopment, ratings of proposed method w.r.t. MPEG-I decorrelator.	41
4.5	Mean values and bootstrapped 95 % CIs of mono signal listening test scores per item in terms of overall audio quality.	42
5.1	Block diagram of proposed multi-channel GAN-based audio decorrelation method.	46
5.2	Mean and standard deviation of absolute coherence over frequency evaluated on the music test dataset. The coherence values were averaged over all input-output pairs resp. all output channel pairs, with $C = 4$	53

5.3	Boxplot of mel-spectrogram loss and ODG averaged over all output channels for the music test dataset, with $C = 4$. The horizontal lines represent the median values, and the boxes represent the first to third quartiles. The whiskers are within 1.5 times the inter-quartile range; outliers are depicted by diamonds.	54
5.4	Scatter plot of average coherence loss over average mel-spectrogram loss evaluated on the music test dataset for the proposed method trained using different values of λ_{coh} , with $C = 4$	55
5.5	Boxplot of average coherence loss and average mel-spectrogram loss evaluated on the music test dataset for the proposed method trained for different number of output channels C , with $\lambda_{\text{coh}} = 2.5$	55
5.6	Mean and standard deviation of absolute coherence over frequency evaluated on the music test dataset for independent channel optimization, with $C = 4$ and $\lambda_{\text{coh}} = 4.0$. The coherence values were averaged over all input-output pairs resp. all output channel pairs.	56
5.7	Mean and standard deviation of absolute coherence over frequency for the proposed method using different datasets for training and evaluation, with $C = 4$ and $\lambda_{\text{coh}} = 2.5$. The coherence values were averaged over all relevant signal pairs.	56
5.8	Boxplot of mel-spectrogram loss averaged over all output channels for the proposed method using different datasets for training and evaluation, with $C = 4$ and $\lambda_{\text{coh}} = 2.5$	57
5.9	Four-channel loudspeaker setup used for the listening experiments.	58
5.10	Mean values and bootstrapped 95 % CIs of general performance evaluation listening test scores per item as well as aggregated over all items in terms of perceived envelopment.	60
5.11	Mean values and bootstrapped 95 % CIs of general performance evaluation listening test scores per item as well as aggregated over all items in terms of overall audio quality.	61
5.12	Mean values and bootstrapped 95 % CIs of training dataset evaluation listening test scores, grouped by signal type, in terms of perceived envelopment and overall audio quality, respectively. Both conditions were rated relative to the model trained on the music dataset.	62
6.1	Schematic overview of binaural SESS rendering model. For visualization purposes, the special case of horizontal spatial extent only is considered.	67
6.2	Block diagram of proposed homogeneous SESS rendering method.	69
6.3	Block diagram of IC adjustment.	70
6.4	LSD between implemented filters and ideal filters without magnitude distortion depending on the target IC value averaged over frequency. A large variety of spatial extent ranges was considered, covering all areas of the sphere.	74
6.5	Mean and standard deviation of RMSE between objective metrics of proposed method and binaural reference depending on the extent $\Delta\varphi$, averaged over all values for the center $\bar{\varphi}$ of the SESS, for white noise input signal. Two variants of the proposed method were considered: applying decorrelation to generate the two decorrelated input signals and using ideal incoherent noise sequences.	75
6.6	Maximum correlation between all decorrelated source signal pairs for all three input signals.	77

6.7	LSD between all decorrelated source signal pairs for all three input signals.	77
6.8	Mean values and bootstrapped 95 % CIs of listening test scores for proposed method output, grouped by input signal.	80
6.9	Exemplary MPEG-I VR test scene, including a fountain as sound source with extent.	81
7.1	Block diagram of proposed heterogeneous SESS rendering method.	85
7.2	Schematic overview of feasible SESS two-channel input signal recording setup.	85
7.3	Schematic overview of point source reproduction procedure. For visualization purposes, the special case of horizontal spatial extent only is considered.	86
7.4	Schematic overview of two-channel input signal simulation setup employed for the performance evaluation.	88
7.5	Schematic overview of comparison methods considered for the performance evaluation. For visualization purposes, the special case of horizontal spatial extent only is considered.	89
7.6	Mean and standard deviation of RMSE between objective metrics of considered processing methods and binaural reference depending on the extent $\Delta\varphi$, averaged over all considered values for the center $\bar{\varphi}$ of the SESS, for the dense applause signal.	90
7.7	Mean values and bootstrapped 95 % CIs of listening test scores per extent range.	92
8.1	Ocean waves signal: Mean and standard deviation of RMSE between objective metrics of homogeneous SESS rendering output and their target values according to the rendering model, for different decorrelation methods. The mean and standard deviation of the RMSE values were calculated depending on the extent $\Delta\varphi$, averaged over all considered values for the center $\bar{\varphi}$ of the SESS.	99
8.2	Music1 signal: Mean and standard deviation of RMSE between objective metrics of homogeneous SESS rendering output and their target values according to the rendering model, for different decorrelation methods. The mean and standard deviation of the RMSE values were calculated depending on the extent $\Delta\varphi$, averaged over all considered values for the center $\bar{\varphi}$ of the SESS.	100
8.3	Mean and bootstrapped 95 % CIs of listening test scores aggregated per source signal and per extent range, respectively, in terms of overall audio quality.	102
8.4	Visual representations of the considered target extent ranges provided to the listeners in the spatial extent listening test.	103
8.5	Mean and bootstrapped 95 % CIs of listening test scores aggregated per source signal and per extent range, respectively, in terms of perceived spatial extent rated relative to a visual representation of the considered target extent range.	104
8.6	Mean and bootstrapped 95 % CIs of listening test difference scores w.r.t. $\lambda_{\text{mel,st}} = 0.15$ aggregated per source signal and per extent range, respectively, in terms of perceived spatial extent rated relative to a visual representation of the considered target extent range.	104
A.1	Block diagram to obtain two-channel decorrelated output signal according to (A.3).	116
A.2	Block diagram of tree structure used to generated multi-channel decorrelated output signal, for $I = 3$ individual single-channel decorrelator instances.	117

- A.3 Mean and standard deviation of RMSE between objective metrics of considered processing methods and binaural reference depending on the extent $\Delta\varphi$, averaged over all considered values for the center $\bar{\varphi}$ of the SESS, for the sparse applause signal. 119
- A.4 Mean and standard deviation of RMSE between objective metrics of considered processing methods and binaural reference depending on the extent $\Delta\varphi$, averaged over all considered values for the center $\bar{\varphi}$ of the SESS, for the speech signal. 120
-

List of Tables

3.1	Overview of CNN architecture, with a total of roughly 2.97 M trainable parameters for $K = 129$, where K and L denote the number of frequency bins and time frames of $X(k, l)$, respectively.	25
4.1	Mean and standard deviation of \mathcal{L}_{mel} [dB] for proposed method (trained on music signals) and MPEG-I decorrelator for all three test datasets.	40
4.2	Mean and standard deviation of $D_{\text{st,in}}$ [dB] for proposed method (trained on music signals) and MPEG-I decorrelator for all three test datasets.	43
4.3	Mean and standard deviation of $D_{\text{st,ch}}$ [dB] for proposed method (trained on music signals) and MPEG-I decorrelator for all three test datasets.	43
6.1	Extent ranges considered in the listening test. Values indicated in bold were employed exclusively with the pink noise signal.	78
9.1	Links to audio examples for the proposed neural network-based audio decorrelation and binaural SESS rendering methods, organized by chapter.	109

CHAPTER 1

Introduction

1.1 Scope and Motivation

Immersive spatial audio technologies have become more and more commonly available, not only to professionals but also to general consumers. In the context of home entertainment, immersive home cinema setups are readily available by means of surround loudspeaker setups or 3D soundbars. Nowadays, many car manufacturers integrate high-end surround loudspeaker setups into their higher-priced models with the aim to enable an immersive spatial audio experience. Another emerging field concerns virtual reality (VR) and augmented reality (AR) applications, which enable interactive rendering of immersive sound scenes either over headphones or using a multi-channel loudspeaker setup.

The overall aim of all these spatial audio rendering applications is to achieve a realistic rendering of the presented sound scenes in a perceptual sense. For interactive applications such as VR and AR, the rendering needs to be adjusted in real time according to the listener's position and orientation. Besides good sound quality, it is particularly important to consider relevant spatial attributes of the presented sound scenes. One important aspect is the localization of individual sound sources. Additionally, other perceptual aspects of the reproduced sound scenes need to be considered, including the perceived spatial extent (i.e., "size") of a sound source and the perceptual impression of the surrounding environment.

The perceived size of an auditory event as well as the perceived envelopment, which describes the perceptual impression of being surrounded by a sound field [1], are significantly influenced by the degree of correlation between the sounds received by the ears. A low correlation is typically associated with an increased size of the auditory event [2] and an enhanced sense of envelopment [3]. This perceptual relevance makes audio decorrelation an important tool within the field of spatial audio rendering to help control the spatial perception of the sound image.

Given an audio input signal, the goal of audio decorrelation methods is to generate one or more output signals that are at the same time statistically uncorrelated and perceptually as close as possible to the original input signal [4, 5]. Audio decorrelation methods have been employed for various tasks within the field of spatial audio rendering. Typical applications include parametric spatial audio coding [6] and reproduction [7], headphone externalization [8], and artificial reverberation [9]. Another application of audio decorrelation methods concerns rendering of spatially extended sound sources (SESSs) [10], which is particularly relevant in the context of VR and AR applications. A distinction is made between methods that target rendering of homogeneous and heterogeneous SESSs. While

homogeneous SESSs emit sound with constant radiation characteristics over the extent, heterogeneous SESSs exhibit a position-dependent radiation behavior.

A significant challenge of existing audio decorrelation techniques concerns the degradation of the output audio quality. In particular, existing methods often introduce temporal smearing and coloration artifacts [4,11], and their perceptual performance is typically highly signal-dependent. Another challenge presents the design of a large number of mutually independent decorrelation filters that simultaneously offer a high degree of decorrelation and a good perceptual quality. This is, for example, required in the context of multi-channel spatial audio reproduction [7].

SESS rendering methods that are based on audio decorrelation techniques naturally suffer from signal artifacts introduced by the employed audio decorrelation methods. A further challenge concerns the computational complexity, which is particularly critical for interactive applications such as VR and AR, where the rendering must be adjusted in real time according to the listener’s position and orientation. The computational complexity of most existing methods increases significantly with increasing spatial extent of the SESS, which is undesirable. While many methods exist that aim at realistically rendering homogeneous SESSs (see, e.g., [12–14]), only a few works can be found specifically targeting heterogeneous SESSs (see, e.g., [15]). Depending on the source’s characteristics, accounting for the position-dependent radiation behavior of the SESS can be crucial for achieving a realistic rendering.

This thesis deals with methods for audio decorrelation and binaural rendering of SESSs, with the aim to address some of the common limitations of state-of-the-art approaches. In the first part of this thesis, the use of neural networks for the task of audio decorrelation is investigated. The second part of this thesis proposes methods for binaural rendering of both homogeneous and heterogeneous SESSs, which employ audio decorrelation techniques. Finally, both parts are combined by investigating the suitability of the proposed neural audio decorrelation approach for the binaural homogeneous SESS rendering method that has been developed within the scope of this thesis.

1.2 Main Contributions and Outline

In this section, an outline of this thesis is provided and the main contributions are summarized per chapter. A conceptual overview of the thesis outline is depicted in Figure 1.1.

Chapter 2 provides an overview of the relevant background related to spatial hearing. Additionally, state-of-the-art approaches concerning both audio decorrelation and rendering of SESSs are discussed, including a further discussion on related challenges and limitations.

Chapters 3, 4, and 5 introduce neural network-based audio decorrelation methods that have been developed within the scope of this thesis. In Chapter 3, a reference-based approach to audio decorrelation using a convolutional neural network (CNN) architecture is introduced. The neural network is trained to mimic the behavior of a state-of-the-art frequency-domain decorrelation method. Chapter 3 primarily serves as a proof of concept regarding the use of neural networks for the task of audio decorrelation. Using the same CNN architecture, in Chapter 4, a reference-free approach to audio decorrelation based on generative adversarial networks (GANs) is proposed. The generator’s loss function consists of a number of individual loss terms to control both the correlation between the output and the input signal and the output signal quality. Finally, Chapter 5 provides a multi-channel extension of the method discussed in Chapter 4. For each output channel, a separate generator network is employed. The different generator networks are optimized jointly to ensure both a low input-output and inter-channel

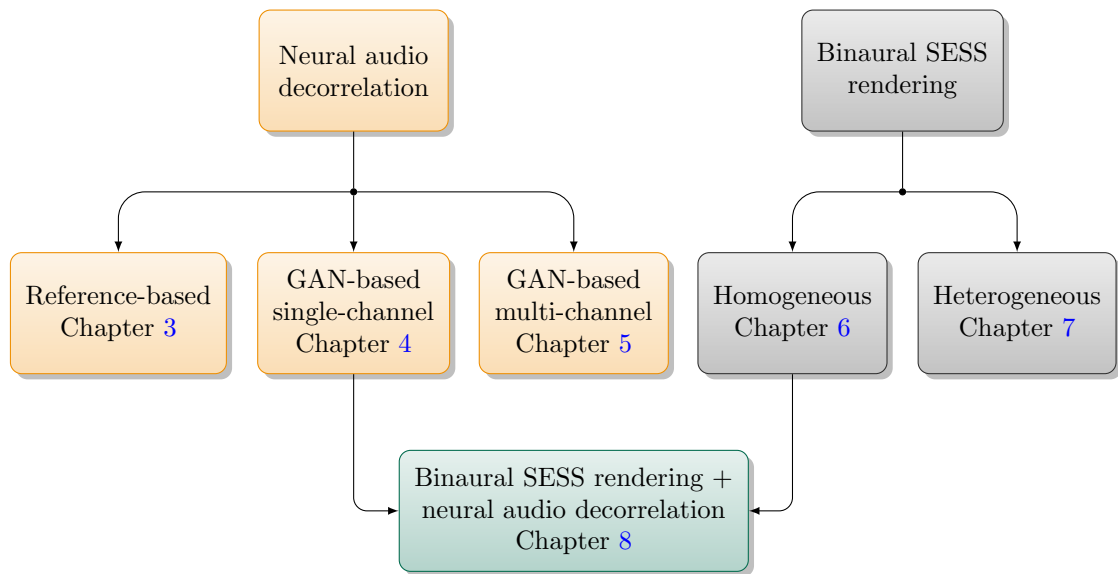


Figure 1.1: Overview of thesis outline.

correlation. Systematic evaluations of the proposed multi-channel decorrelation method are conducted considering a variety of network configurations.

Chapters 6 and 7 deal with methods for binaural rendering of SESSs developed within the scope of this thesis. Chapter 6 introduces an efficient approach to binaural rendering of homogeneous SESSs. Based on the model of an incoherently extended sound source with position-independent energy, the target auditory cues of the binaural output signal are synthesized directly using two decorrelated input signals only. Compared to direct rendering of a number of decorrelated point sources distributed over the entire extent range, this approach comes at the advantage of a reduced computational complexity and fewer decorrelation artifacts. In Chapter 7, a binaural rendering method particularly suited for heterogeneous SESSs is introduced. For this purpose, the rendering model used in Chapter 6 is extended to take the position-dependent energy of the SESS into account. Given a two-channel input signal, a binaural output signal with the desired inter-channel properties is determined based on this rendering model. As an additional criterion, it is ensured that the directional characteristics encoded in the two-channel input signal are preserved.

Chapter 8 investigates the suitability of the single-channel GAN-based audio decorrelation method introduced in Chapter 4 for the binaural homogeneous SESS rendering method introduced in Chapter 6 in particular. To improve the overall audio quality of the decorrelated stereo signal, which serves as the basis for the homogeneous SESS rendering method, an extended version of the decorrelator’s generator loss is proposed.

Chapter 9 summarizes the main contributions of this thesis and provides an outlook into possible future research directions.

1.3 List of Publications and Contribution Statements

The main content of this thesis is based on a number of journal and conference publications. A list of all relevant publications is provided below, along with a contribution statement and the respective

chapters they contribute to. All publications are published in international peer-reviewed journals or conference proceedings.

Publications Related to Neural Audio Decorrelation

- C. Anemüller, O. Thiergart, and E. A. P. Habets, “A data-driven approach to audio decorrelation,” *IEEE Signal Process. Lett.*, vol. 29, pp. 2477–2481, Nov. 2022, [16] (Chapter 3).
Contribution statement: The present author developed and implemented the proposed method, conducted the experiments, and wrote the manuscript.
Scientific contributions: First approach to audio decorrelation using neural networks.
- C. Anemüller, O. Thiergart, and E. A. P. Habets, “Neural audio decorrelation using generative adversarial networks,” in *Proc. WASPAA*, Oct. 2023, [17] (Chapter 4).
Contribution statement: The present author developed and implemented the proposed method, conducted the experiments, and wrote the manuscript.
Scientific contributions: First reference-free neural audio decorrelation method.
- C. Anemüller, O. Thiergart, and E. A. P. Habets, “Multi-channel neural audio decorrelation using generative adversarial networks,” *EURASIP J. Audio, Speech Music. Process.*, vol. 2024, no. 58, Nov. 2024, [18] (Chapter 5).
Contribution statement: The present author developed and implemented the proposed method, conducted the experiments, and wrote the manuscript.
Scientific contributions: Extension of [17] to provide multi-channel output signals. Systematic evaluation of the proposed method for different network configurations.

Publications Related to Binaural SESS Rendering

- C. Anemüller, A. Adami, and J. Herre, “Efficient binaural rendering of spatially extended sound sources,” *J. Audio Eng. Soc.*, vol. 71, no. 5, pp. 281–292, May 2023, [19] (Chapter 6).
Contribution statement: The present author developed and implemented the proposed method, conducted the experiments, and wrote the manuscript. Parts of the statistical evaluation regarding the listening experiment were performed by A. Adami.
Scientific contributions: Particularly efficient method for binaural rendering of homogeneous SESSs, with the advantage of reduced computational complexity and fewer decorrelation artifacts compared to a direct implementation of the rendering model.
- C. Anemüller, O. Thiergart, and E. A. P. Habets, “Binaural rendering of heterogeneous sound sources with extent,” in *Proc. ICASSP*, Apr. 2024, pp. 471–475, [20] (Chapter 7).
Contribution statement: The present author developed and implemented the proposed method, conducted the experiments, and wrote the manuscript.
Scientific contributions: Binaural rendering method specifically suited for heterogeneous SESSs.

Additional Publications not Included in this Thesis

- C. Anemüller, O. Thiergart, and E. A. P. Habets, “Sector-based parametric sound field reproduction in the circular harmonic domain using covariance based rendering,” in *Proc. IWAENC*, Sep. 2022, [21].
-

- C. Anemüller and J. Herre, “Calculation of directivity patterns from spherical microphone array recordings,” in Proc. AES 147th Conv., Oct. 2019, [22].
-

CHAPTER 2

Background and State of the Art

In this chapter, an overview of the relevant background related to spatial hearing is provided. Additionally, state-of-the-art approaches regarding both audio decorrelation and rendering of spatially extended sound sources (SESSs) are discussed.

2.1 Spatial Hearing

The term spatial hearing describes the human ability to analyze the spatial aspects of sound. An important aspect of spatial hearing concerns sound source localization, which refers to the judgment of direction and distance of a sound source. Another closely related aspect is the perceived spatial extent of a sound source. Additionally, spatial hearing provides a perceptual impression of the surrounding environment, which depends on the environment's acoustic properties.

This section provides an overview of the fundamentals of spatial hearing that are relevant within the scope of this thesis. Further details can be found, e.g., in [23–25].

2.1.1 Head-Related Spherical Coordinate System

In the field of spatial hearing, a spherical coordinate system is typically employed to describe a position in three-dimensional space relative to the listener.

In spherical coordinates, a position in three-dimensional space is described by the distance $r \in [0, \infty)$, the azimuth angle $\varphi \in (-180^\circ, 180^\circ]$, and the elevation angle $\theta \in [-90^\circ, 90^\circ]$. The azimuth angle φ and the elevation angle θ together form the direction $\mathbf{u} = [\varphi, \theta]^T$, where $(\cdot)^T$ denotes the transpose operation. In this thesis, the head-related spherical coordinate system is aligned as depicted in Figure 2.1. The origin of the coordinate system is located midway between the entrances to the two ear canals. For the frontal direction, both the azimuth angle φ and the elevation angle θ are equal to 0° .

2.1.2 Auditory Cues

The two sound waves received by the ears depend on the position of a sound source in a characteristic way. The sound propagation from a sound source to the ears is influenced by a number of physical phenomena related to the human's pinnae, head, and torso. These include, e.g., head shadowing,

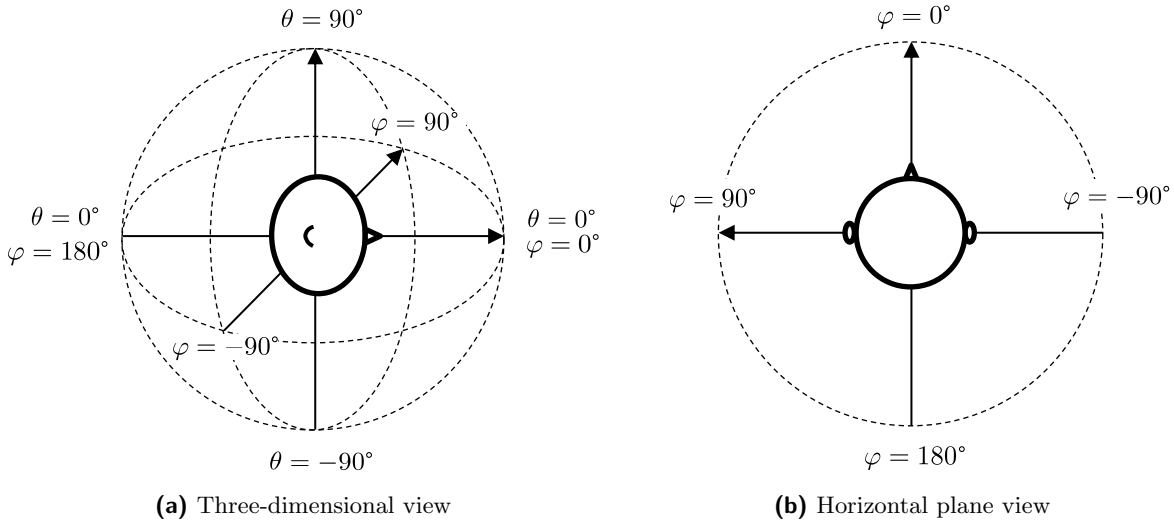


Figure 2.1: Head-related spherical coordinate system.

as well as reflections and diffractions associated with the head, pinnae, and torso. These acoustic propagation effects result in a sound source position- and frequency-dependent linear filtering of the sound waves received at the two ears. The human auditory system utilizes the associated attributes of the sound waves arriving at the ears, so-called auditory cues, to draw conclusions about the spatial sound properties. A distinction is made between binaural cues, which describe the relationship between the sound pressure at both ears, and monaural cues, which are based on the sound pressure at one ear only. The most important binaural cues are the interaural level difference (ILD), the interaural time difference (ITD), the interaural phase difference (IPD), and the interaural cross-correlation (ICC). In terms of monaural cues, spectral cues are particularly relevant.

2.1.3 Horizontal Localization

Horizontal (i.e., left-right) localization is mostly dominated by binaural localization cues, namely the ILD, the ITD, and the IPD. The acoustic propagation effects that primarily influence these binaural localization cues are illustrated in Figure 2.2.

The ILD is predominantly attributed to head shadowing, resulting in a reduced sound pressure level at the far ear (see Figure 2.2a). Head shadowing is most pronounced at high frequencies due to the relatively short wavelengths compared to the dimensions of the head. At low frequencies, where the wavelength is long compared to the dimensions of the head, the sound is diffracted around the head and little or no shadowing occurs. The ILD takes values between approximately 0 dB and 20 dB, depending on the incident sound direction and frequency [25].

The ITD and the closely related IPD result from path differences from the source to the two ears (see Figure 2.2b). At low frequencies, where the wavelength is longer than the distance around the head, the ITD corresponds to a unique IPD. At higher frequencies, however, the IPD becomes ambiguous. Consequently, the IPD provides only an effective localization cue at low frequencies. Nevertheless, the ITD of the sound waves' envelopes can still be evaluated by the human auditory system at high frequencies. The physically possible range of the ITD is approximately ± 1 ms, which follows from the

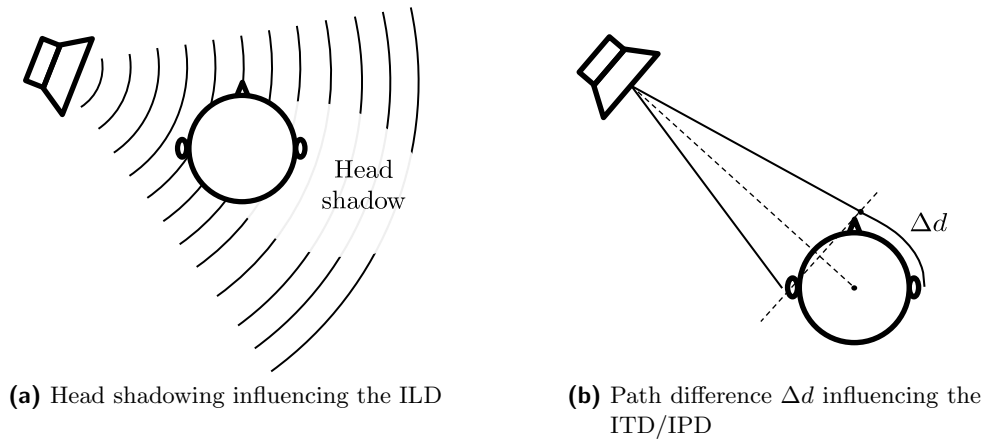


Figure 2.2: Illustration of acoustic propagation effects important for horizontal localization.

typical dimensions of the human head [26].

The frequency dependency of the localization cues motivated the duplex theory [27], which states that sound source localization in the horizontal plane is dominated by the IPD and ITD at low frequencies (below about 1500 Hz) and by the ILD at higher frequencies (above about 1500 Hz). However, the fact that the human auditory system utilizes the ITD of the sound waves' envelopes at high frequencies shows that the duplex theory is not entirely correct.

The localization accuracy in the horizontal direction is best for sources located in front of the listener, for which values of the localization blur as low as 1° have been reported [23]. With increasing displacement from the frontal direction toward the left or the right, the localization accuracy deteriorates, and the localization blur can be up to 3 to 10 times as large as for the frontal direction [23].

2.1.4 Vertical Localization and Front-Back Differentiation

While the ILD and the ITD/IPD are effective localization cues for horizontal localization, they cannot be used reliably for vertical localization or front-back differentiation. Due to symmetries in the propagation paths, these positions result in nearly identical binaural cues. More generally, positions with the same distance difference to the two ears and therefore with very similar binaural cues form so-called cones of confusion [28]. These are imaginary cones extending outward from each ear along the interaural axis, as visualized in Figure 2.3. Any circular cross-section of a cone of confusion corresponds to positions that are equidistant from the left ear and equidistant from the right ear.

In the absence of binaural localization cues, vertical localization and front-back differentiation primarily rely on monaural spectral cues. Monaural spectral cues describe direction-dependent spectral changes of the sound introduced by the pinnae, head, and torso. Spectral changes caused by the pinnae predominantly occur at high frequencies, whereas head diffraction and torso reflections also affect the spectrum at lower frequencies. It has been demonstrated that a boost in sound energy in certain frequency ranges, so-called directional bands, corresponds to sound source localization at specific positions in the vertical plane (e.g., in front, behind, or above the listener) [28]. Especially at high frequencies above approximately 8 kHz, spectral differences of up to 30 dB can occur depending on the elevation angle of a sound source [26].

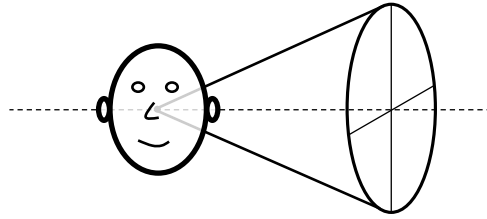


Figure 2.3: Illustration of a cone of confusion. Points on the surface of the cone represent positions with a constant distance difference to the two ears. Points on any circular cross-section of the cone represent positions that are equidistant from the left ear and equidistant from the right ear.

Generally, monaural localization cues are much weaker than binaural localization cues. Front-back confusion may occur, and the localization accuracy in the vertical direction is considerably lower than in the horizontal direction. The best localization accuracy has been reported for noise stimuli located in front of the listener, with a vertical localization blur of about 4° [23], compared to 1° for horizontal localization, as discussed in Section 2.1.3. Moreover, the vertical localization accuracy is highly dependent on the characteristics of the sound source. For very narrowband sounds, it has been demonstrated that the perceived direction does not depend on the direction of the sound source, but solely on the sound's frequency [23].

2.1.5 Distance Perception

In this section, the different aspects related to the perceived distance of a sound source are briefly discussed. A comprehensive review regarding distance perception can, e.g., be found in [29].

In a free field, several acoustic propagation effects influence the properties of the sound waves arriving at the ears depending on the source distance. Under ideal free-field conditions, the sound pressure level at the ears decreases by 6 dB when the source distance is doubled. At large distances above approximately 15 m, air absorption becomes significant, resulting in a dampening of high-frequency components [23]. Considering binaural cues, both the ILD and the ITD/IPD are nearly distance-independent for distances larger than approximately 1 m. In contrast, for nearby sources, the ILD increases considerably with decreasing distance [30].

Several studies have investigated the influence of the corresponding attributes of the sound waves arriving at the ears on the perceived distance of a sound source. It has been demonstrated that the sound pressure level can indeed serve as an effective auditory cue for distance perception, see, e.g., [31, 32]. For familiar sources, such as speech, the sound pressure level can serve as an absolute distance cue [31]. For unfamiliar sources, however, only relative distance judgments are possible [32]. A number of studies have demonstrated that the perceived distance can be modulated by changing the amount of high-frequency sound energy [33, 34], which is related to the physical phenomenon of air absorption. The results reported in [35] show that near-field distance perception in anechoic environments is dominated by the low-frequency ILD.

In reverberant environments, the ratio of direct-to-reverberant sound energy, which is inversely related to the source distance, provides an additional distance cue. Several authors have shown that the direct-to-reverberant energy ratio can serve as a reliable distance cue, see, e.g., [36, 37]. The results reported in [36] furthermore indicate that the direct-to-reverberant energy ratio provides an absolute distance cue, rather than allowing relative distance judgments only.

2.1.6 Spatial Impression

The previous sections discussed the human ability to localize sound sources in terms of both direction and distance. Further important aspects of spatial sound perception include the perceived size of a sound source and the perceptual impression of the surrounding environment. For both, the ICC is an important auditory cue.

2.1.6.1 Interaural Cross-Correlation

The ICC is a measure of the correlation between the two sound waves arriving at the ears, quantifying the similarity of the two waveforms. Let the ear signals $y_l(t)$ and $y_r(t)$, with continuous time index t , represent the time-dependent sound pressures at the two ears. The ICC is then defined as the maximum absolute value of the normalized cross-correlation function $c_{lr}(t, \tau)$ between $y_l(t)$ and $y_r(t)$ [23, 26]:

$$c_{lr}(t, \tau) = \frac{\mathcal{E}\{y_l(t)y_r(t+\tau)\}}{\sqrt{\mathcal{E}\{y_l^2(t)\}\mathcal{E}\{y_r^2(t+\tau)\}}}, \quad (2.1)$$

$$\text{ICC}(t) = \max_{\tau} |c_{lr}(t, \tau)|, \quad (2.2)$$

where $\mathcal{E}\{\cdot\}$ denotes the statistical expectation. Delays τ in the range of ± 1 ms are considered, which corresponds to the physically possible range of the ITD as discussed in Section 2.1.3. As a result of the normalization, $c_{lr}(t, \tau)$ takes values between -1 and 1 , resulting in ICC values between 0 and 1 . Alternatively, the ICC can be defined as the signed value of the normalized cross-correlation function $c_{lr}(t, \tau)$ with maximum absolute value, resulting in ICC values between -1 and 1 .

The frequency-domain equivalent of the ICC is typically referred to as interaural coherence (IC). Let $Y_l(k, l)$ and $Y_r(k, l)$ denote time-frequency domain representations of $y_l(t)$ and $y_r(t)$, respectively, with discrete frequency index k and discrete time frame index l . The IC is then defined as follows [38]:

$$\text{IC}(k, l) = \frac{|\mathcal{E}\{Y_l(k, l)Y_r^*(k, l)\}|}{\sqrt{\mathcal{E}\{|Y_l(k, l)|^2\}\mathcal{E}\{|Y_r(k, l)|^2\}}}. \quad (2.3)$$

For a point source in an anechoic environment, the ICC (and thus also the IC) takes values close to 1 . In a reverberant environment, the ICC decreases as the relative amount of reflected sound energy increases. In an ideal diffuse sound field, the frequency-dependent ICC can be approximated by a modified sinc function [39, 40], resulting in high ICC values at low frequencies with a steep decrease with increasing frequency.

2.1.6.2 Concert Hall Acoustics

Given that the value of the ICC depends on the relative amount of reflected sound energy, it is not surprising that the ICC is related to the perceptual impression of the surrounding environment. In a reverberant environment, the ear signals can be simulated by convolving a source signal with a set of binaural room impulse responses (BRIRs). A more detailed description of BRIRs will be provided in Section 2.1.8. The ICC of the ear signals, as defined in (2.2), is thus closely related to the ICC of the BRIRs. In the context of concert hall acoustics, the ICC of the BRIRs is frequently regarded as an objective measure related to the perceptual attributes of an enclosure. In this context, two perceptual attributes related to the spatial impression of an enclosure are typically considered: listener

envelopment and apparent source width (ASW) [41].

Although slightly different definitions of envelopment have been used in the literature, a unifying factor is that it describes the perceptual impression of being surrounded by a sound field [1]. A number of studies have reported a clear correlation between the ICC of the late part of the BRIRs in particular and the perceived envelopment [3, 42, 43]. A lower ICC is typically associated with an enhanced sense of envelopment, i.e., the sound field appears to surround the listener more. Another objective measure that was found to correlate strongly with the perceived envelopment is the late lateral energy fraction, i.e., the amount of late reverberation arriving from the side at the listener relative to the total sound energy [3]. Although these two measures appear to be very different, they are usually similarly influenced by lateral reflections [44].

The term ASW, as used in the context of concert hall acoustics, refers to the perceived width of an auditory event due to the influence of early reflections. While listener envelopment is primarily associated with the ICC of the late part of the BRIRs, the ASW is particularly related to the ICC of the early part of the BRIRs [45]. A lower ICC is typically associated with an increased ASW. Furthermore, the early lateral energy fraction, which describes the ratio between lateral and omnidirectional early energy, was found to be closely related to the ASW [46]. The results reported in [47] indicate that the ICC of the early part of the BRIRs and the early lateral energy fraction are strongly correlated, especially in the lower frequency range.

2.1.6.3 Spatial Sound Reproduction

As discussed in the previous section, early and late reflections were found to influence the ASW and the listener envelopment, respectively, in a reverberant environment. Considering headphone or multi-channel loudspeaker reproduction in an anechoic environment, similar perceptual effects can be evoked artificially.

Instead of ASW, which refers solely to the perceived width of an auditory event, the more general term perceived spatial extent, as proposed in [48], will be used in this thesis to describe the perceived size of an auditory event in the context of spatial sound reproduction. The term spatial extent refers to horizontal extent, vertical extent, and potentially depth.

Similar to the influence of natural decorrelation in a reverberant environment on the ASW, the perceived horizontal spatial extent in the context of headphone or loudspeaker reproduction can be controlled by the ICC. The results reported in [2], obtained through controlled headphone experiments using noise stimuli, demonstrated a negative correlation between the two quantities. An ICC value of 1 thus ideally indicates a point source, whereas decreasing the ICC value enlarges the perceived extent. For very low ICC values close to 0, two separate auditory events may be perceived near each ear [2]. In [49], similar results were obtained using a two-channel horizontal loudspeaker setup in an anechoic environment, again considering noise stimuli. The ICC was controlled indirectly by modifying the loudspeakers' inter-channel correlation, and it was shown that a decrease in inter-channel correlation, and thus in ICC, results in an increase in the perceived spatial extent. At higher frequencies, the ICC of the sound waves' envelopes seems to be the relevant descriptor of the perceived spatial extent of a sound source [50]. Note that an analogous high-frequency behavior can be observed for the ITD, as discussed in Section 2.1.3. In contrast to the results reported for horizontal loudspeaker setups, Gribben et al. [51] showed that vertical inter-channel decorrelation has only a slight effect on the perceived vertical spatial extent, by reproducing incoherent noise stimuli over a two-channel vertical loudspeaker setup.

In addition to the ICC, the spectral content, duration, and loudness of a sound were found to influence the perceived spatial extent. Considering headphone reproduction, the results reported in [52] show that a lower frequency, an increased sound level, and an increased sound duration are associated with a larger spatial extent. In [53], similar results regarding the effect of spectral content and loudness of a sound on the perceived horizontal and vertical spatial extent were reported using a vertical loudspeaker setup.

An artificial experience related to listener envelopment can be evoked by reproducing multiple mutually incoherent noise stimuli over a set of loudspeakers surrounding the listener [26]. In [54], it was shown that a reduced inter-channel correlation results in an enhanced sense of (diffuse) envelopment, considering the reproduction of noise stimuli over varying loudspeaker setups. Moreover, several studies have investigated the number of loudspeakers required to reproduce the spatial impression of a diffuse sound field [54–56]. It was concluded that, depending on the loudspeaker arrangement and stimulus type, as few as four loudspeakers can be sufficient to reproduce the spatial impression of a diffuse sound field.

2.1.7 Just-Noticeable Differences of Binaural Cues

This section briefly discusses the just-noticeable differences (JNDs) of the different binaural cues as reported in the literature. The JNDs of the binaural cues are of interest as they enable to judge the performance of spatial audio rendering methods based on objective measures. The JND is defined as the smallest perceivable change of a given quantity. Typically, the JND depends on the magnitude of the reference measurement. The reported values were all obtained through controlled headphone experiments, using either pure tones or narrowband/broadband noise signals as stimuli.

For a low reference ILD, JND values as low as approximately 0.5 dB to 1 dB have been reported [57, 58]. The reported JND values exhibit a slight frequency dependency, with the JND being largest for frequencies around 1 kHz. The results reported in [59] furthermore indicate an increase in JND for frequencies above around 10 kHz. Several studies have reported a slight decrease in JND with increasing stimulus level [60, 61]. The JND of the ILD furthermore shows a moderate dependency on the reference ILD. If the reference ILD increases, the JND increases as well. In [58], an average increase in JND of 1.9 dB has been reported for a reference ILD of 15 dB compared to a reference ILD of 0 dB.

The JND of the IPD was found to be approximately constant for frequencies below around 1 kHz, with a value of 3° for a reference IPD of 0° [62]. Similarly, for frequencies below around 1 kHz, the reported values for the JND of the ITD correspond approximately to a constant threshold in IPD of about 3° for a reference ITD of 0 ms [63]. An increase in the reference ITD/IPD was shown to result in an increase in the respective JNDs [61, 62]. Moreover, the JNDs of both the ITD and the IPD demonstrate a slight dependency on the stimulus level, reaching a minimum at comfortable listening levels [61, 64]. At higher frequencies, the JND of the IPD increases substantially, and above around 1300 kHz listeners were unable to detect any changes in the IPD [64]. Similarly, time differences in the fine-structure waveforms cannot be detected at higher frequencies. However, the ITD of the signals' envelopes can be processed. The results reported in [65] demonstrate that for complex waveforms, the auditory system can be as sensitive to changes in the ITD at high frequencies as it is at low frequencies.

Considering the ICC, the JND strongly depends on the reference value. In this context, not only the absolute value of the normalized cross-correlation function is relevant, but also its sign. Consequently, the following discussion assumes the signed alternative of the ICC definition provided in Section 2.1.6.1.

For broadband noise stimuli, JND values as low as 0.04 have been reported for a reference ICC of 1 [66,67]. In contrast, for a reference ICC of 0, the reported JND values are in the range of 0.3 to 0.5. The results reported in [66] demonstrate that listeners are more sensitive to changes toward positive correlation than to changes toward negative correlation for a reference ICC of 0. Additionally, slightly higher JND values were observed for a reference ICC of -1 in comparison to a reference ICC of 1. It was shown that the sensitivity to changes in ICC is relatively independent of the stimulus level, with a slight increase in sensitivity at comfortable listening levels [68]. In [69], the impact of the noise bandwidth on the JND was investigated for narrowband noise stimuli centered around 500 Hz. For a reference ICC of 1, the JND was found to be approximately constant for bandwidths up to 115 Hz, with a value of 0.004, and to increase toward approximately 0.04 for higher bandwidths. For a reference ICC of 0, the JND was observed to decrease from approximately 0.7 to 0.35 as the bandwidth increased from 3 Hz to 115 Hz, and to remain at a constant value of approximately 0.35 for bandwidths exceeding 115 Hz. In [70], the sensitivity to changes in ICC was investigated for narrowband noise stimuli with varying frequency, considering different values for the reference ICC. The results indicate a decrease in sensitivity for frequencies above 750 Hz. Similar to the ITD, at higher frequencies, the ICC of the signals' envelopes is particularly relevant from a perceptual point of view [50]. Considering the ICC of the signals' envelopes for narrowband noise stimuli centered around 4 kHz, in [71], JND values of around 0.1 were reported for a reference ICC of 1.

2.1.8 Head-Related Transfer Functions

Head-related transfer functions (HRTFs) describe the free-field transfer functions for sound traveling from a specific position in space to the two ears. Consequently, they capture the effects of the human's pinnae, head, and torso on the two ear signals. HRTFs can be measured by recording the direction-dependent transfer functions from a loudspeaker to two microphones placed at the entrances of the ear canal. Additionally, a reference measurement is taken at the position corresponding to the center of the head without the head actually being present. This reference measurement allows for the compensation of the influence of the employed loudspeaker and microphones. While the HRTFs are approximately distance-independent in the far field, a clear distance dependency exists for distances below approximately 1 m [30] (see also Section 2.1.5).

Given a set of HRTFs, the ear signals for different directions of sound incidence can be simulated. In the time domain, the simulated ear signals are obtained by convolving a given source signal with the impulse responses corresponding to the HRTFs. Alternatively, time-frequency-domain processing can be performed. Given a time-frequency-domain single-channel source signal $S(k, l)$ and a set of HRTFs $\mathbf{h}(k, \mathbf{u}) = [H_l(k, \mathbf{u}), H_r(k, \mathbf{u})]^T$ for direction $\mathbf{u} = [\varphi, \theta]^T$, the time-frequency-domain binaural output signal $\mathbf{y}(k, l) = [Y_l(k, l), Y_r(k, l)]^T$ can be determined as follows:

$$\mathbf{y}(k, l) = \mathbf{h}(k, \mathbf{u})S(k, l). \quad (2.4)$$

Note that (2.4) represents a time-frequency-domain approximation of the time-domain convolution, the accuracy of which depends on the employed time-frequency-domain conversion parameters [72]. Multiple point sources can be easily incorporated by linear superposition. By reproducing the simulated ear signals over headphones, the listener can experience spatialized binaural sound.

A general distinction is made between individualized and generic HRTFs. Individualized HRTFs are

measured with the listener’s own head and ears, whereas generic HRTFs are obtained for representative human subjects or using so-called artificial heads or head and torso simulators. Since HRTF acquisition is a time-consuming process which requires a dedicated measurement setup, the use of generic HRTFs is often preferred. However, the employment of generic HRTFs may result in an increased rate of front-back confusion and a decrease in vertical localization performance when compared to individualized HRTFs [73, 74].

A commonly used head and torso simulator is the KEMAR manikin [75], for which multiple HRTF datasets with varying spatial resolution have been published, see, e.g., [76, 77]. Considering multiple head-above-torso orientations, in [78], a high-resolution HRTF dataset of the FABIAN head and torso simulator has been released. Additionally, multiple HRTF datasets of human subjects have been measured, see, e.g., [79, 80].

In reverberant environments, the two sound waves received by the ears are influenced not only by the HRTFs, but also by the environment’s acoustic properties. To account for the acoustic properties of the room, BRIRs can be used. These include both the influence of the dry HRTFs and the sound propagation through the room.

2.2 Audio Decorrelation

The overall aim of audio decorrelation methods is to generate one or more statistically uncorrelated versions of an audio input signal that are otherwise perceptually as close as possible to the input signal [4, 5].

As discussed in Section 2.1.6, the degree of ICC strongly influences the spatial perception of the sound image. Most importantly, a low ICC is associated with an increased size of the auditory event and an enhanced sense of envelopment. Given the perceptual relevance of the ICC, audio decorrelation is a valuable tool in the field of spatial audio rendering. In parametric spatial audio coding and reproduction, audio decorrelation techniques are frequently utilized to render the diffuse or ambient sound field components [6, 7]. Another common application concerns rendering of SESSs, as further discussed in Section 2.3. Moreover, audio decorrelation techniques have been employed for headphone externalization [8], artificial reverberation [9], and to reduce comb-filtering artifacts in multi-channel loudspeaker reproduction systems [4].

Aside from applications in the field of spatial audio rendering, decorrelation techniques have been successfully employed to improve the performance of multi-channel acoustic echo cancellation systems [81, 82]. However, due to the differing objectives of acoustic echo cancellation systems compared to spatial audio rendering applications, decorrelation methods used in this context have different underlying constraints. Consequently, they are not considered further in this thesis.

This section presents a review of approaches to audio decorrelation proposed in the literature, focusing on spatial audio rendering applications. The majority of approaches aim to achieve a low correlation by altering the signal’s phase while largely preserving its temporal and spectral envelopes. Given that the human ear is relatively insensitive to phase variations, this is a promising approach [10]. Existing methods can be broadly divided into time-domain and frequency-domain methods.

2.2.1 Time-Domain Methods

Time-domain audio decorrelation methods typically process the input signal with some kind of finite impulse response (FIR) or infinite impulse response (IIR) filter.

Perhaps the first approach to audio decorrelation has been proposed by Lauridsen in 1954 [83]. Two decorrelated signals are obtained by adding/subtracting a time-delayed and scaled version of the input signal to/from the unprocessed input signal. A reduced correlation is achieved by complementary comb-filter peaks and notches in the two output signals. The delay length is an important design parameter of the Lauridsen decorrelator. A too short delay may result in significant comb-filtering artifacts, whereas a too long delay may result in audible echoes, especially for high-frequency components. While this approach is particularly computationally efficient, the resulting comb-filtering artifacts are generally undesirable.

Another straightforward approach to audio decorrelation is to convolve the input signal with a (decaying) noise sequence. The longer the noise sequence, the lower the degree of input-output correlation that can be achieved. However, if the length of the employed noise sequence exceeds 20 ms to 30 ms, the temporal smearing of the input signal becomes audible [84]. To minimize temporal smearing artifacts, the use of exponentially decaying noise sequences is generally beneficial [85]. Audio decorrelation based on exponentially decaying white noise sequences is, e.g., considered in [6]. In [86], Alary et al. proposed a method based on decaying velvet-noise sequences. Due to the sparse nature of velvet-noise sequences, the proposed method is particularly efficient in terms of computational complexity. In [87], this method was further optimized to minimize spectral coloration and improve the decorrelation performance.

A further common approach is to apply all-pass filters to the input signal, realized as either FIR or IIR filters. In [4], Kendall proposed the design of FIR all-pass filters for the task of audio decorrelation by calculating the inverse discrete Fourier transform (DFT) of a frequency-domain signal with unit magnitude and random phase. While this approach ensures a unit magnitude response at the DFT frequency bins, the magnitude response at intermediate frequencies is not well controlled. Also considering FIR filters, Xie et al. [88] proposed the design of all-pass filters based on pseudo-random binary sequences. In terms of computational complexity, IIR all-pass filters are preferred over FIR all-pass filters, as they require fewer filter coefficients. In [89], a method based on cascaded biquad all-pass filters has been proposed. The magnitudes and frequencies of the all-pass filters' poles are chosen randomly within perceptually motivated constraints. Informal listening tests demonstrated that the proposed method resulted in fewer signal distortions relative to the method introduced in [4], which was accompanied by a slight reduction in decorrelation performance. Another method based on cascaded biquad all-pass filters has been proposed in [84]. In this method, the all-pass filters are designed by specifying the group delay as a function of frequency, which allows for the minimization of perceptual artifacts.

2.2.2 Frequency-Domain Methods

As an alternative to time-domain methods, more advanced audio decorrelation methods often operate in the time-frequency domain. This facilitates frequency-dependent processing, which is beneficial since the signal's autocorrelation properties depend on the wavelength.

In [5], Boueri et al. proposed a method that applies random time shifts in perceptually-motivated critical bands. The maximum allowable time shift is frequency-dependent. Due to the longer wavelengths, larger time shifts are permitted at lower frequencies in comparison to higher frequencies. Furthermore, the exact time shift values are constrained in a manner that avoids destructive interference at the edges of the critical bands. A similar approach has been employed in [90] to render the diffuse/ambient sound field components for the task of spatial audio reproduction. Pseudo-random delays were applied in the

quadrature mirror filter (QMF) domain, with the employed delays ranging between 20 and 80 ms at low frequencies and between 5 and 15 ms at high frequencies. The exact delay ranges were obtained through manual adjustment to ensure sufficient decorrelation while avoiding reverberation artifacts. Furthermore, a recursive onset suppressor was implemented prior to decorrelation to minimize temporal smearing artifacts.

In [91], Penniman proposed a method based on [5] that incorporates separate transient handling. First, the transient and non-transient signal components are separated. Subsequently, the non-transient signal components are decorrelated using [5]. For the transient signal components, amplitude panning is performed to rapidly changing random directions. The proposed method was evaluated by considering diffuse sound field reproduction over a five-channel loudspeaker setup, comparing it to the original method described in [5]. However, the presented listening test results are inconclusive, and the benefit over [5], if any, seems to depend heavily on the considered signal type.

As part of MPEG Surround, a decorrelation method using lattice all-pass filters and (frequency-dependent) delays operating in the QMF domain has been proposed [26, 92]. To reduce temporal smearing and coloration artifacts, a QMF-domain energy comparison stage is employed, which adjusts the time-averaged energy of the output signal to match that of the input signal. Additionally, tools are available to control the fine temporal envelope of the output signal, thereby enhancing the signal quality for particularly transient signal components. By carefully selecting the all-pass filter coefficients, multiple mutually uncorrelated output signals can be obtained.

More recently, a similar method has been proposed as part of the MPEG-I Immersive Audio standard [93]. Processing is performed in the short-time Fourier transform (STFT) domain, using Schroeder all-pass filters [94] and a frequency-independent pre-delay. Furthermore, temporal envelope shaping in the STFT domain is incorporated to match the temporal and spectral envelopes of the decorrelated output signal to those of the input signal. Additionally, transients are detected and excluded from the decorrelation processing to prevent the impairment of transients due to phase dispersion. Compared to the decorrelator employed in MPEG Surround, this method offers a reduced processing delay and a lower computational complexity.

As a post-processing step to a classical signal processing-based decorrelator, an approach to temporal envelope shaping based on convolutional neural networks (CNNs) was proposed in [95]. The work focused on transient input signals, considering a QMF-domain decorrelator based on all-pass filters and frequency-dependent delays. The neural network receives as input the QMF-domain decorrelator output signal and the time-domain input signal. Its objective is to reshape the temporal envelope of the decorrelator output signal to align with that of the input signal. The training target was obtained by applying a classical signal processing-based time-domain envelope reshaping method to the decorrelator output signal.

In [96], a decorrelation method specifically designed for applause signals has been proposed as an alternative to generic decorrelation techniques. The transient and non-transient signal components are first separated. The non-transient signal components are processed using the decorrelator employed in MPEG Surround, whereas the transient signal components are decorrelated by applying simple phase shifts. In [96], a spatial audio coding application is considered, and the phase shifts are estimated based on the original stereo input signal. If such estimates are unavailable, random phase shifts may be employed instead [97]. A listening test demonstrated that the proposed decorrelation method improves the quality of applause items for the considered spatial audio coding application. A drawback of signal-

specific decorrelation methods is that a signal classification is required, which causes computational overhead.

2.2.3 Challenges and Limitations

Common limitations of existing decorrelation techniques include the presence of temporal smearing and coloration artifacts [4, 11]. Temporal smearing artifacts may include both artifacts resulting from phase dispersion and added reverberation or echoes. While transient signals predominantly suffer from temporal smearing artifacts, coloration artifacts are particularly problematic for noise-like signals. As a result, the perceptual performance of decorrelation techniques is generally signal-dependent. Although several measures have been taken to improve the perceptual quality of decorrelation methods, such as temporal/spectral envelope shaping and separate transient handling, certain artifacts persist. Furthermore, a tradeoff between the degree of decorrelation and the perceptual quality is typically observed.

Another challenge concerns the design of a large number of mutually independent decorrelation filters that simultaneously provide a high degree of decorrelation and a good perceptual quality. This is, for example, required in the context of multi-channel spatial audio reproduction [7]. Considering deterministic approaches, such as those based on IIR all-pass filters, a specific design of mutually independent filter pairs is typically required. For approaches that involve random components, such as convolution with decaying noise sequences or the application of random time shifts, multiple decorrelated output signals are readily available. However, there is no guarantee that the degree of mutual decorrelation achieved is satisfactory.

In many spatial audio rendering applications, the decorrelated signals are mixed with signals that are (partly) correlated with the original input signal. Therefore, a high degree of decorrelation is required not only with regard to the spatial perception but also to avoid coloration when mixing the decorrelated and non-decorrelated signals.

2.3 Rendering of Spatially Extended Sound Sources

Various spatial audio rendering applications require the reproduction of sound sources over either loudspeakers or headphones. Relevant applications include, e.g., auralization of virtual or augmented reality (VR/AR) environments [15, 98]. A general distinction can be made between six-degrees-of-freedom (6DoF) and three-degrees-of-freedom (3DoF) applications. While 6DoF applications allow rotational and translational user movement in all three axes, 3DoF applications consider only rotational user movement. The most straightforward method for reproducing sound sources over such setups is to render them as point sources. However, when aiming at reproducing physical sound sources with non-negligible auditory spatial extent, this model is not sufficient. Examples of such sound sources include a grand piano, a choir, or a waterfall, all of which have a certain “size” (i.e., geometric and perceived extent). To render such sound sources in a realistic manner, methods specifically designed for rendering of SESSs are required. In general, there is a distinction made between approaches to rendering of homogeneous and heterogeneous SESSs. While homogeneous SESSs emit sound with constant radiation characteristics over the extent, heterogeneous SESSs exhibit a position-dependent radiation behavior.

In this section, approaches to rendering of homogeneous and heterogeneous SESSs proposed in the literature are reviewed, considering both loudspeaker and headphone reproduction. The majority of these approaches consider a two-dimensional spatial extent, i.e., on a sphere surrounding the listener. However, any two-dimensional approach can be extended to three-dimensional source shapes by projecting the SESS geometry onto an imaginary sphere surrounding the listener, similar to [99, 100].

2.3.1 Homogeneously Extended Sound Sources

Since a position-independent radiation behavior is assumed, approaches to rendering of homogeneous SESSs typically operate on a single-channel input signal. In a wider sense, the methods proposed in the literature all rely on spatially distributing a number of individual source components, determined from the single-channel input signal, over the extent range.

2.3.1.1 Coherent Rendering

A straightforward approach is to create multiple coherent (i.e., identical) copies of the input signal and render them from different directions depending on the extent range.

Considering a three-dimensional loudspeaker setup, in [101], this approach has been employed with the objective of stabilizing the direction-dependent spatial extent of an amplitude panned source. Note that the objective of the method proposed in [101] is to obtain a (relatively small) uniform spatial extent that is independent of the panning direction, rather than to render SESSs with arbitrary spatial extents. Following a similar approach, Schissler et al. [100] proposed a method for efficiently rendering SESSs for headphone reproduction. Under the assumption that the same source signal is emitted by the SESS throughout the entire spatial extent range, all HRTFs within that range are combined to form a single spatial filter. The dry source signal is then convolved with the combined spatial filter to obtain the binaural output signal.

The coherent SESS rendering approach offers a particularly low computational complexity and a high signal fidelity. However, with regard to the spatial impression, coherently extended sound sources tend to collapse into relatively narrow auditory events as a result of the summing localization phenomenon [23].

2.3.1.2 Frequency-Dependent Rendering

A further common approach is to spatially distribute the individual frequency bands of the single-channel source signal over the extent range.

In [12], Pulkki et al. proposed such a method based on directional audio coding (DirAC). The individual frequency bands of the source signal are randomly distributed to different directions within the source's extent range in a time-dependent manner. By using the assigned time- and frequency-dependent directions as directional metadata for a DirAC decoding block, binaural or loudspeaker output signals are generated. The diffuseness value, which is required as additional metadata in the DirAC decoding block, is set to 0. This work was further extended by Laitinen et al. [102]. Rather than assigning the individual directions in a completely random manner, the direction assignment was optimized to minimize audible artifacts and improve the spatial extent perception. Additionally, the diffuseness value was increased for larger spatial extents. As a result, the source is perceived as being more enveloped and possible temporal artifacts are reduced.

The general suitability of the frequency-dependent SESS rendering approach has been investigated further by other authors, with consideration given to both headphone and loudspeaker reproduction [103–108]. With regard to horizontal spatial extents, the results reported in the different works indicate that this approach can be used successfully to systematically create the perception of an SESS. However, the absolute size of the perceived spatial extent was found to be often narrower than intended [103,105]. Generally, the output signal quality and the perceived spatial extent are highly signal-dependent. Depending on the signal type, artifacts related to temporal smearing, comb filtering, and coloration may occur [103]. In particular, the approach was found to be ill-suited for signals with predominant transient signal components due to the poor output signal quality [103]. Additionally, broadband frequency content is preferred. Naturally, the approach is not suited if a frequency-dependent spatial extent range is desired.

2.3.1.3 Time-Dependent Rendering

In [13], a method for synthesizing SESSs based on the spatial distribution of the individual temporal grains of a granular synthesis-based stimulus was proposed, with a focus on loudspeaker reproduction. In granular synthesis, sounds are generated by combining multiple short signals, which are typically referred to as grains. Note that this approach requires access to the individual temporal grains of a source signal, and therefore is not applicable to arbitrary signal types. The granular synthesis-based method was compared to a frequency-dependent method based on the spatial distribution of individual frequency bands. Considering a rain and an impulse train stimulus, the granular synthesis-based method was found to result in significantly larger perceived spatial extents than the frequency-dependent method. It should be noted that this may be due to the considered signal types, which are suboptimal for the frequency-dependent SESS rendering approach as discussed in [103]. For both methods, judgments of horizontal spatial extent were found to vary systematically with the physical extent, whereas the correspondence between the perceived and the physical vertical extent was poor. Furthermore, the ICC was considered as a predictor for the perceived horizontal spatial extent evoked by the employed processing methods. A strong correspondence was found between the ICC and the perceived horizontal spatial extent, which aligns with the findings of other studies as discussed in Section 2.1.6.

2.3.1.4 Incoherent Rendering

Another prominent group of approaches is based on the spatial distribution of multiple mutually incoherent copies of the single-channel input signal over the extent range. Given the importance of the ICC for spatial extent perception, as discussed in Section 2.1.6, it is not surprising that this is a promising approach for homogeneous SESS synthesis.

In [109], a series of formal listening experiments were conducted to investigate the suitability of the incoherent SESS rendering approach for evoking the auditory perception of an SESS within a 3DoF VR environment. The participants were presented with a visual stimulus representing an SESS. As a visual representation, ellipsoids of varying sizes and shapes were employed, representing either a point source, a horizontal SESS, a vertical SESS, or a horizontal and vertical SESS. In accordance with the visual stimulus, the audio was rendered by distributing a number of decorrelated point sources over the entire spatial extent range. The participants were then asked to rate how well the perceived spatial

extent matched the presented visual representation. For horizontal SESSs, results indicated a high degree of correspondence between the perceived spatial extent and the visual representation. However, for vertical SESSs, the performance was found to be limited.

In [14], Potard et al. proposed a method for synthesizing different source shapes. By using decorrelation techniques, multiple decorrelated point sources are generated from a single-channel source signal. Subsequently, the individual point sources are placed in different spatial locations according to the desired source shape. The effectiveness of the proposed method was investigated considering loudspeaker reproduction, and it was demonstrated that listeners were indeed able to distinguish between sources with varying horizontal and vertical spatial extents. The relative judgments of perceived spatial extent were slightly more accurate for sources with varying horizontal spatial extent than for sources with varying vertical spatial extent. In general, the identification of the specific source shape was found to be less reliable and highly signal-dependent. This method is also employed in MPEG-4 Advanced AudioBIFS [110], where volumetric shapes can be filled with a number of equally distributed and decorrelated point sources to achieve three-dimensional spatial extent.

Several other authors have proposed similar methods, all based on the basic concept of spatially distributing multiple decorrelated point sources generated from the original single-channel input signal over the extent range [111–115]. The decorrelation techniques applied to generate the required decorrelated versions of the original input signal typically deteriorate the signal quality (see Section 2.2), which represents a significant limitation of the incoherent SESS rendering approach. McCormack et al. proposed a method for rendering of SESSs over arbitrary playback setups [116], with the aim to reduce decorrelation artifacts by using the covariance domain framework introduced in [117]. First, signals corresponding to the center of the SESS are generated. These signals, as well as decorrelated variants thereof, are then mixed to obtain output signals with specific target inter-channel relationships based on the model of an incoherently extended sound source. Considering binaural reproduction, the proposed method was shown to reduce decorrelation artifacts in comparison to an unconstrained method which directly mixes two decorrelated input signals.

As previously discussed in Section 2.1.6, horizontal inter-channel decorrelation has a significantly larger effect on the perceived spatial extent than vertical inter-channel decorrelation. This aligns with the findings of the listening experiments reported in [14, 109], which demonstrated that the incoherent SESS rendering approach is more effective for rendering horizontal spatial extents than for rendering vertical spatial extents.

2.3.2 Heterogeneously Extended Sound Sources

While numerous methods exist that aim at realistically rendering homogeneous SESSs, only a few works have been published that specifically target heterogeneous SESSs. To obtain information about the position-dependent radiation behavior of the SESS, methods for rendering of heterogeneous SESSs generally require a multi-channel input signal.

A method for rendering of heterogeneous SESSs inside a 6DoF environment has recently been proposed as part of the MPEG-I Immersive Audio standard [15, 118]. The method accepts as input either a higher-order Ambisonics or multi-channel signal describing the sound source. The rendering is performed based on virtual loudspeakers, using a so-called interior representation when the listener is inside of the extent and a so-called exterior representation when the listener is outside of the extent. For the interior representation, up to 12 virtual loudspeakers are used, arranged in layers of varying

heights surrounding the listener. The virtual loudspeaker signals are derived directly from the input signal, using mixing and decorrelation when necessary. For the exterior representation, up to five virtual loudspeakers are used, whose positions depend on the relative position between source and listener. The virtual loudspeaker signals are derived from the interior representation using a set of virtual directional microphones, which are directed toward different spatial regions of the SESS.

2.3.3 Challenges and Limitations

As discussed in Section 2.3.1, the different approaches proposed for homogeneous SESS rendering all come with their own limitations. To achieve a convincing spatial extent perception for a wide range of signal types, the incoherent SESS rendering approach appears to be the most promising. A significant limitation of this approach is the introduction of signal artifacts resulting from the employed decorrelation techniques. Therefore, it is essential to optimize the utilized decorrelation method to minimize signal degradation. Furthermore, the amount of decorrelated signal energy introduced should be minimized. For all proposed methods, rendering of vertical spatial extent is generally less effective than rendering of horizontal spatial extent.

Considering heterogeneous SESS rendering methods, only a limited number of investigations have been conducted. A potentially challenging scenario arises when only a small number of input signals are available, resulting in comparably little information about the position-dependent radiation behavior of the SESS. Consequently, achieving a realistic rendering may prove to be particularly difficult in such scenarios. When decorrelation techniques are incorporated, similar considerations apply as for the homogeneous SESS rendering methods.

A general challenge concerns the computational complexity, which is particularly critical for interactive applications such as VR and AR, where the rendering must be adjusted in real time according to the listener's position and orientation. For the majority of SESS rendering methods, the computational complexity increases significantly with increasing spatial extent of the SESS, which is undesirable.

CHAPTER 3

Reference-Based Neural Audio Decorrelation

*The main content of this chapter is based on: C. Anemüller, O. Thierngart, and E. A. P. Habets, “A data-driven approach to audio decorrelation,” *IEEE Signal Process. Lett.*, vol. 29, pp. 2477–2481, Nov. 2022, [16].*

Audio decorrelation is a widely used tool in the field of spatial audio processing, employed for a variety of tasks including spatial audio coding and reproduction [6, 7], as well as rendering of spatially extended sound sources (SESSs) [10]. As discussed in Section 2.2, various approaches to audio decorrelation have been proposed in the literature, most of which aim to achieve a low correlation by altering the signal’s phase while largely preserving its temporal and spectral envelopes. Advanced audio decorrelation methods often operate in the time-frequency domain, which allows to easily exploit the signal’s frequency-dependent autocorrelation properties.

As discussed in Section 2.2.3, a notable drawback of existing decorrelation techniques is the degradation of the output signal quality. In particular, existing methods often introduce temporal smearing and coloration artifacts. While transient signals mostly suffer from temporal smearing artifacts, coloration artifacts are particularly problematic for noise-like signals. Additionally, the design of a large number of mutually independent decorrelation filters that simultaneously offer a high degree of decorrelation and a good perceptual quality presents a significant challenge. This is particularly relevant in the context of multi-channel spatial audio reproduction (see, e.g., [7]).

Data-driven approaches to audio decorrelation might be capable of mitigating (some of) these limitations. Since deep learning-based approaches have not previously been applied with the goal of audio decorrelation, the objective of this chapter is to present a general proof of concept regarding the use of such approaches for audio decorrelation. The methods introduced in Chapters 4 and 5 build on the methodology presented in this chapter with the aim of addressing the aforementioned limitations of existing decorrelation techniques. We propose a convolutional neural network (CNN) that is trained to mimic the behavior of the decorrelator proposed in [93], a state-of-the-art single-channel decorrelation method operating in the short-time Fourier transform (STFT) domain. The neural network’s performance is evaluated by comparing its output to the output of the reference decorrelator by means of an objective evaluation as well as through a formal listening test.

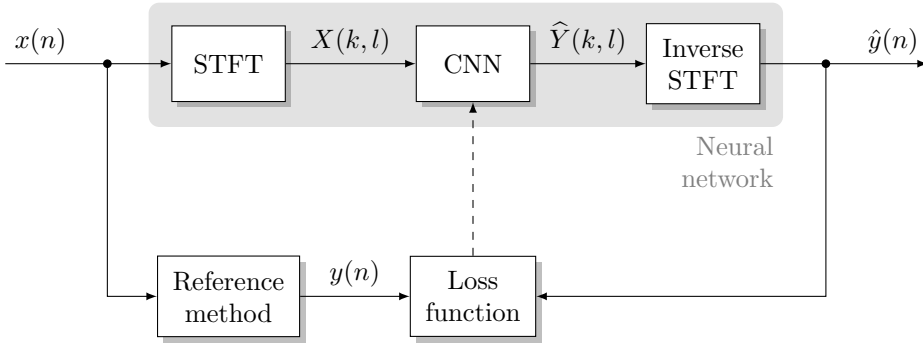


Figure 3.1: Block diagram of proposed reference-based neural audio decorrelation method.

The remainder of this chapter is structured as follows. Section 3.1 provides the problem statement concerning single-channel decorrelation methods. In Section 3.2, the proposed method and the employed network architecture are described. Section 3.3 covers the datasets used for training and evaluation as well as details regarding the training procedure, and in Section 3.4, the objective and perceptual evaluations are discussed. Finally, Section 3.5 concludes this chapter by providing a summary.

3.1 Problem Statement

Given a time-domain single-channel input signal $x(n)$, with discrete time index n , the overall goal is to obtain a single-channel output signal $y(n)$ that is:

1. Statistically uncorrelated from the input signal $x(n)$.
2. Perceptually transparent w.r.t. the input signal $x(n)$, i.e., the two audio signals are perceptually indistinguishable [119]. While perceptual transparency is mathematically difficult to define, a necessary requirement is that the input signal’s temporal and spectral envelopes are not (significantly) altered [92].

Processing methods that aim at fulfilling the two specified requirements are referred to as decorrelation methods. The output signal $y(n)$ is then called decorrelated w.r.t. the input signal $x(n)$.

3.2 Proposed Method

Figure 3.1 shows a block diagram of the proposed reference-based neural decorrelation method. Input to the method is a time-domain single-channel input signal $x(n)$. By processing $x(n)$ using a neural network, the goal is to obtain a decorrelated version of the input signal $x(n)$ as output of the neural network, denoted here as $\hat{y}(n)$. To achieve this, we employ a complex spectral mapping approach. In this chapter, the network is trained w.r.t. a reference signal $y(n)$, which is obtained by processing $x(n)$ using a state-of-the-art reference decorrelator [93]. In the following sections, the employed reference method, the model architecture, and the corresponding loss function are described.

Table 3.1: Overview of CNN architecture, with a total of roughly 2.97 M trainable parameters for $K = 129$, where K and L denote the number of frequency bins and time frames of $X(k, l)$, respectively.

Layer	Input size	Output size	# Parameters
C2R	$K \times L$ (complex)	$(2 \cdot K) \times L$	–
conv1d_1	$(2 \cdot K) \times L$	$(16 \cdot K) \times L$	$2 \cdot 16 \cdot K \cdot 40$
ReLU_1	$(16 \cdot K) \times L$	$(16 \cdot K) \times L$	–
conv1d_2	$(16 \cdot K) \times L$	$(16 \cdot K) \times L$	$16^2 \cdot K \cdot 40$
ReLU_2	$(16 \cdot K) \times L$	$(16 \cdot K) \times L$	–
conv1d_3	$(16 \cdot K) \times L$	$(16 \cdot K) \times L$	$16^2 \cdot K \cdot 40$
ReLU_3	$(16 \cdot K) \times L$	$(16 \cdot K) \times L$	–
conv1d_4	$(16 \cdot K) \times L$	$(2 \cdot K) \times L$	$16 \cdot 2 \cdot K \cdot 40$
R2C	$(2 \cdot K) \times L$	$K \times L$ (complex)	–

3.2.1 Reference Method

As reference method, the decorrelator described in [93] is used, which is part of the MPEG-I Immersive Audio standard.

The input signal $x(n)$ is first transformed to the time-frequency domain using an STFT with a frame length of 256 samples and a hop size of 128 samples at a sampling rate of 48 kHz. To ensure sufficient decorrelation, frequency-dependent delays are applied in each frequency band, followed by four cascaded Schroeder all-pass filters [94]. An envelope shaper is subsequently incorporated, which adjusts the temporal envelope of the output signal to that of the input signal, thereby reducing temporal smearing artifacts. In a final step, the output signal is transformed back to the time domain, resulting in the reference signal $y(n)$. Note that the transient detection described in [93] was not included in this study. Additionally, a slightly different parameterization of the all-pass filters and frequency-dependent delays was used. Specifically, the frequency-dependent delay was monotonically decreased from 16 time frames at low frequencies to 1 time frame at high frequencies. The gain γ for each of the Schroeder all-pass filters was set to 0.7, and the delays of the four Schroeder all-pass filters were set to 1, 3, 5, and 7 time frames, respectively.

3.2.2 Model Architecture

The input signal $x(n)$ is first transformed to the time-frequency domain using an STFT. The time-frequency-domain signal $X(k, l)$, with discrete frequency index k and discrete time frame index l , is subsequently processed by a CNN, resulting in the time-frequency-domain output signal $\hat{Y}(k, l)$. After applying an inverse STFT, the time-domain output signal $\hat{y}(n)$ is obtained. In accordance with the reference method described in Section 3.2.1, an STFT frame length of 256 samples and a hop size of 128 samples at a sampling rate of 48 kHz are employed. As a result, a single-sided spectrum with $K = 129$ frequency bins and L time frames is obtained.

An overview of the CNN architecture with its layers and the corresponding input sizes, output sizes, and number of parameters can be found in Table 3.1. The core elements of the CNN architecture are four 1D convolutional layers (conv1d_x, $x \in \{1, 2, 3, 4\}$). After each convolutional layer, except the last one, a rectified linear unit (ReLU) activation function is applied (ReLU_x, $x \in \{1, 2, 3\}$). Before the first convolutional layer, a complex-to-real operation is performed (C2R), which stacks the real and the imaginary part of the time-frequency-domain input signal $X(k, l)$ in an alternating fashion.

The real-to-complex operation after the last convolutional layer realizes the reverse process (R2C). The convolutional layers perform a convolution in the time dimension only. This choice was made since there is also no interaction between the frequency bins in the reference decorrelator. For all convolutional layers, a kernel size of 40 and a dilation and stride of 1 are used. We use causal convolutions [120] and introduce zero padding to ensure the same number of output as input time frames. The number of channel groups of each convolutional layer equals K , each frequency bin therefore has its own set of parameters. The number of channels of the intermediate layers is increased by a factor of 8 to model more complex input-to-output relationships. This factor was determined empirically, considering a tradeoff between performance and model size.

The employed network architecture is inspired by the concept of time-delay neural networks [121] as well as the more recently proposed WaveNet architecture [120]. Since processing is performed in the sub-sampled time-frequency domain, employing dilated convolutions was not found to be necessary. The required receptive field was determined by analyzing the reference decorrelator. The impulse response of the concatenated all-pass filters used in the reference decorrelator drops by 40 dB after 123 time frames (328 ms), which is considered sufficient here. The maximum incorporated delay of the reference decorrelator equals 16 time frames (43 ms). A receptive field of at least 139 time frames (371 ms) is therefore required. Based on this requirement, we select the aforementioned configuration of four convolutional layers, each with a kernel size of 40 (107 ms), which results in a receptive field of 157 time frames (419 ms).

3.2.3 Loss Function

As the loss function, a frequency-domain variant of the time-domain signal-to-distortion ratio (TD-SDR) [122] is employed. To this end, the SDR is calculated per frequency band and averaged subsequently. The frequency-domain signal-to-distortion ratio (FD-SDR) is preferred here over the TD-SDR (as defined in (3.2)) since it was found to result in a more uniform convergence behavior across frequencies.

Starting from the time-frequency-domain neural network output signal $\hat{Y}(k, l)$ and the time-frequency-domain reference signal $Y(k, l)$, the FD-SDR is defined as follows:

$$\text{FD-SDR} = \frac{1}{K} \sum_{k=1}^K 10 \log_{10} \left(\frac{\sum_l |Y(k, l)|^2}{\sum_l |Y(k, l) - \hat{Y}(k, l)|^2} \right), \quad (3.1)$$

where $\log_{10}(\cdot)$ denotes the logarithm to the base 10, and $Y(k, l)$ is obtained by transforming $y(n)$ to the STFT domain, using the same STFT parameters as employed for the neural network (see Section 3.2.2).

3.3 Datasets and Training

For training and evaluation of the proposed method, we focused on two specific signal types: music and applause signals. To represent the class of music signals, the mixture signals of the MUSDB18-HQ dataset were used [123], which consists of 150 full-length music tracks with a total length of about 10 h. A dataset of applause signals was generated using the applause generator described in [124]. Starting from 24 dry studio recordings of individual persons clapping, items with varying applause density were generated by virtual placement of clapping people. The final applause dataset consisted

of 200 applause signals with a length of 18 s each and a total length of about 1 h. The number of clapping people was randomly varied between 15 and 100. The reference signals required for training were generated using the reference decorrelator described in Section 3.2.1.

The neural network was trained using the combined music/applause dataset as well as using both datasets individually. The datasets were randomly split into a training (60%), a validation (20%), and a test (20%) subset. The signals were split into chunks of 144 000 samples (3 s) each. A context of 20 480 previous samples was added to each chunk, which corresponds to 160 STFT time frames and thus exceeds the receptive field of the CNN. The context is only used to determine the CNN output and is removed before calculating the loss function and other metrics. A batch size of 32 was used during training. Training was performed using the Adam optimizer with a learning rate and a weight decay of 5×10^{-4} .

3.4 Performance Evaluation

The performance of the proposed method was evaluated by comparing its output to the output of the employed reference method. Since the proposed method aims to mimic the behavior of the reference method, ideally, there should not be a significant difference between the proposed method's output and the reference signal. An objective evaluation was conducted based on a number of relevant objective metrics, considering the two requirements specified in Section 3.2.2. Additionally, a perceptual evaluation was performed by means of a formal listening test. The neural network was trained as described in Sections 3.2 and 3.3.

3.4.1 Objective Evaluation

3.4.1.1 Evaluation Metrics

Three distinct evaluation metrics were considered to objectively evaluate the proposed method's performance. To assess general deviations of the neural network output signal $\hat{y}(n)$ from the reference signal $y(n)$, the TD-SDR is determined [122]:

$$\text{TD-SDR} = 10 \log_{10} \left(\frac{\sum_n |y(n)|^2}{\sum_n |y(n) - \hat{y}(n)|^2} \right). \quad (3.2)$$

To specifically focus on differences in terms of the signals' magnitude spectra, the log-spectral distance (LSD) between the time-frequency-domain neural network output signal $\hat{Y}(k, l)$ and the reference signal $Y(k, l)$ is considered [125, 126]:

$$\text{LSD} = \frac{1}{L} \sum_{l=1}^L \sqrt{\frac{1}{K} \sum_{k=1}^K \left[10 \log_{10} \left(\frac{|\hat{Y}(k, l)|^2}{|Y(k, l)|^2} \right) \right]^2}. \quad (3.3)$$

Finally, we evaluate the maximum correlation $c_{x\hat{y}}$ between the input signal $x(n)$ and the neural network output signal $\hat{y}(n)$:

$$c_{x\hat{y}} = \max_d \frac{|\sum_n x(n)\hat{y}(n+d)|}{\sqrt{\sum_n x^2(n) \sum_n \hat{y}^2(n)}}. \quad (3.4)$$

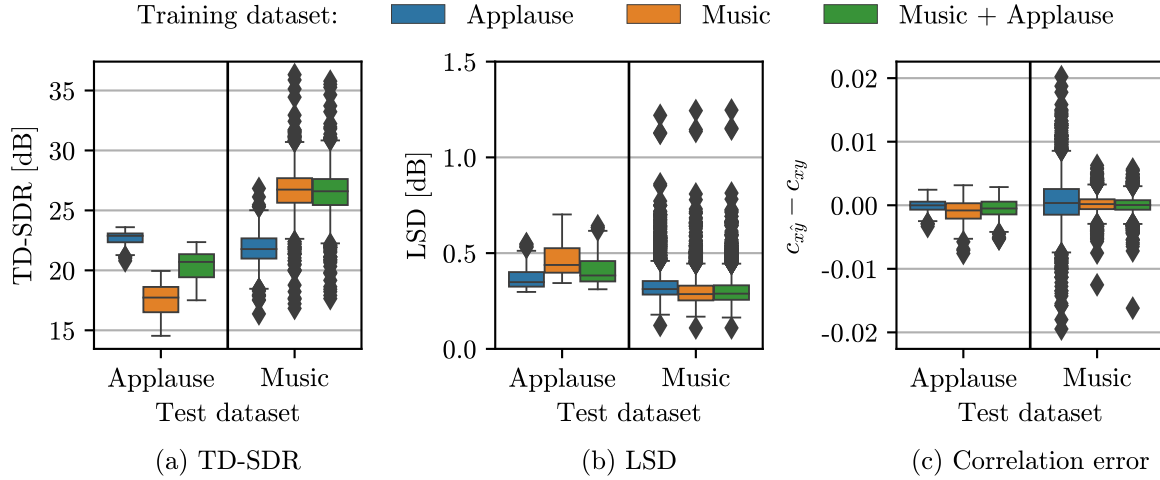


Figure 3.2: Boxplots of objective evaluation metrics for all three training datasets and both test datasets. The horizontal lines represent the median values, and the boxes represent the first to third quartiles. The whiskers are within 1.5 times the inter-quartile range; outliers are depicted by diamonds.

In accordance with Section 2.1.6.1, delays d within the perceptually relevant range of ± 1 ms are considered. As a result of the normalization, $c_{x\hat{y}}$ takes values between 0 and 1. The obtained value can be compared to the maximum correlation c_{xy} between the input signal $x(n)$ and the reference signal $y(n)$, to judge whether a comparable amount of decorrelation is achieved.

3.4.1.2 Results

For the objective evaluation, the neural network was trained separately on all three considered datasets: the combined music/applause dataset, the individual music dataset, and the individual applause dataset. The evaluation was performed on both the music and the applause test datasets.

Figure 3.2 depicts boxplots of the considered evaluation metrics for all three training datasets and both test datasets. Considering the TD-SDR (Figure 3.2a), the performance is best when training and testing on the same signal type. For the music test dataset, the performance degradation is only minimal when training on the combined music/applause dataset. At the same time, a considerable performance increase is observed when training on the combined music/applause dataset as opposed to training on the applause dataset only. For the applause test dataset, similar considerations hold. However, the performance degradation for the combined music/applause training dataset is considerably more pronounced in comparison to the music test dataset. This observation may be attributed to the fact that the applause dataset is considerably smaller than the music dataset. The overall trend of the LSD (Figure 3.2b) is similar to that of the TD-SDR. However, the differences between the individual training datasets are less pronounced. As a final metric, we consider the maximum correlation $c_{x\hat{y}}$ between the input signal and the neural network output signal. Figure 3.2c depicts boxplots of $c_{x\hat{y}} - c_{xy}$, i.e., the error of $c_{x\hat{y}}$ when compared to the maximum correlation c_{xy} between the input signal and the reference signal. The correlation error is generally very small, indicating that the neural network achieves a similar amount of decorrelation as the reference method (the maximum possible error equals 1). As for the TD-SDR and the LSD, the performance is best when training and testing on the same signal type. This is particularly evident when considering the music test dataset,

for which the correlation error is clearly larger when training on the applause dataset only.

3.4.2 Perceptual Evaluation

To supplement the objective evaluation results, a perceptual evaluation of the proposed method was performed. In a formal listening test, the proposed method’s output was compared to the reference signal in terms of preference. Only the neural network trained on the combined music/applause dataset was considered, and a variety of music and applause signals were employed as test items. A selection of the items used in the listening test is available at www.audiolabs-erlangen.de/resources/2022-SPL-DD-Decorrelation.

3.4.2.1 Listening Test Setup

A total of 15 test items were incorporated in the listening test, including 11 music and four applause stimuli. All selected test items are independent of the datasets used during training and for the objective evaluation. The applause stimuli all originated from the evaluation subset of the FSD50K dataset [127], whereas the music stimuli were partly taken from the EBU SQAM CD [128] and the evaluation subset of the FSD50K dataset. The test items covered a large variety of musical instruments and applause densities.

The conditions under test included the proposed method (see Sections 3.2 and 3.3) and the reference method (see Section 3.2.1). For the proposed method, the neural network was trained on the combined music/applause dataset.

Instead of reproducing the single-channel output signals directly, corresponding stereo signals were compared. This allowed to not only judge the overall signal quality but also the amount of decorrelation introduced by the respective processing method, which is related to the spatial impression of the presented stereo signals. Instead of simply reproducing $x(n)$ and either $y(n)$ or $\hat{y}(n)$ on one channel each, the stereo signals were generated by following the procedure outlined in Appendix A.1, which ensures similar temporal characteristics of the left and right channel signals. According to (A.3), the stereo signals $\mathbf{y}_s(n)$ and $\hat{\mathbf{y}}_s(n)$ were obtained as follows:

$$\mathbf{y}_s(n) = \frac{1}{\sqrt{2}} \begin{bmatrix} x(n) + y(n) \\ x(n) - y(n) \end{bmatrix}, \quad (3.5)$$

$$\hat{\mathbf{y}}_s(n) = \frac{1}{\sqrt{2}} \begin{bmatrix} x(n) + \hat{y}(n) \\ x(n) - \hat{y}(n) \end{bmatrix}. \quad (3.6)$$

The participants were asked to rate one condition w.r.t. the other in terms of their preference on a discrete seven-point scale: “much worse” (−3), “worse” (−2), “slightly worse” (−1), “the same as” (0), “slightly better” (1), “better” (2), “much better” (3), following the ITU-T P.800 recommendation [129]. The order of the two conditions (proposed method and reference method) and of the items was randomized. The listeners were instructed to pay attention to attributes related to both, the overall signal quality and the spatial impression. To ensure consistent judgments regarding aspects related to the spatial impression, the listeners were informed about the target application of decorrelation, and instructed that a more enveloped stimulus is considered to be preferred. Additionally, prior to conducting the listening test, the listeners were presented with examples of a coherent (less enveloped, i.e., worse) and an incoherent (more enveloped, i.e., better) white noise stimulus. The items were reproduced over

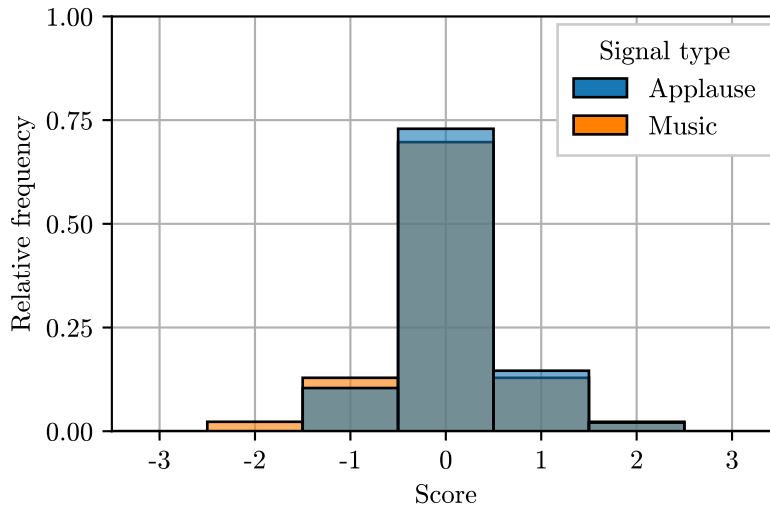


Figure 3.3: Relative frequency of listening test scores grouped by signal type, ratings of proposed method w.r.t. reference method.

Beyerdynamic DT-770 Pro headphones using a customized version of the webMUSHRA software [130]. The participants conducted the listening experiment at home, and were instructed to ensure a quiet environment. A total of 12 subjects participated in the listening test, 11 male and one female. The average age of the participants was 31 years, and all of them had prior listening test experience.

3.4.2.2 Results

Figure 3.3 shows the relative frequency of all listening test scores grouped by signal type for ratings of the proposed method w.r.t. the reference method. The distribution of the scores is nearly symmetrical around 0, which corresponds to the rating “the same as.” A large majority of the scores was given for a rating of exactly 0. The participants did thus most often not perceive a difference between the proposed method and the reference method.

To determine whether the median value of the listening test scores statistically significantly differs from 0, we determined the bootstrapped 99 % confidence interval (CI) of the median value, as proposed in [131]. The median values and their 99 % CIs were calculated for all items as well as for the applause and music items separately. In all cases, the median value as well as the upper and lower bounds of its 99 % CI were found to be equal to 0. It can thus be concluded that the proposed method and the reference method are perceptually equivalent for the considered test items.

3.5 Summary

This chapter explored the use of a data-driven approach for the task of audio decorrelation. We proposed a CNN architecture that was trained to mimic the behavior of the state-of-the-art decorrelation method described in [93]. Although not optimized for real-time processing, the proposed architecture is causal by design. An objective evaluation based on music and applause signals demonstrated that the proposed method matches well with the reference method in terms of TD-SDR, LSD, and decorrelation properties. The signal types incorporated in the training data were found to have a considerable influence on the performance of the proposed method when evaluated on the respective signal type.

The objective findings have been verified by a listening test, which showed that no perceptual difference exists between the proposed method and the reference method in terms of their stereo output signal. Consequently, it can be concluded that data-driven approaches are generally capable of performing audio signal decorrelation.

While the proposed method succeeds at providing a general proof of concept w.r.t. the use of data-driven approaches for audio decorrelation, it was not yet designed to mitigate the limitations of traditional decorrelation techniques discussed in Section 2.2.3. A possibility for improvement would be to generate optimized reference signals for training. We could, for example, use different reference methods or parameter settings depending on the input signal type. Further room for improvement remains w.r.t. the model architecture. It is expected that the number of network parameters and, consequently, the computational complexity can be reduced by exploiting analogies across the frequency bins.

CHAPTER 4

Single-Channel Audio Decorrelation Using Generative Adversarial Networks

The main content of this chapter is based on: C. Anemüller, O. Thiergart, and E. A. P. Habets, “Neural audio decorrelation using generative adversarial networks,” in Proc. WASPAA, Oct. 2023, [17].

In Chapter 3, we proposed a reference-based neural audio decorrelation method based on a convolutional neural network (CNN), which is trained with the help of a decorrelated reference signal. A major limitation of this training procedure is that the achievable performance is inherently constrained by the performance of the employed reference method. To overcome this limitation, a reference-free training procedure is highly desirable.

This chapter proposes a reference-free approach to audio decorrelation based on generative adversarial networks (GANs) [132]. Although not previously applied with the goal of audio decorrelation, GANs have been used successfully for several other audio processing tasks, including speech synthesis (e.g., [133–135]) and bandwidth extension (e.g., [136, 137]). The training network consists of a generator and a discriminator. Given a single-channel input signal, the generator’s task is to output a decorrelated version of the input signal, while the discriminator’s task is to distinguish between the original input signal and the generated decorrelated output signal. As generator, we employ the CNN architecture proposed in Chapter 3, which was designed for the specific task of audio decorrelation. Furthermore, we make use of the HiFi-GAN discriminators [133], which were originally applied to speech synthesis. The generator’s training objective comprises a number of individual loss terms to control both the input-output correlation and the output signal quality. This includes loss components related to the input signal, as well as adversarial training, which aims to improve the overall output signal quality. During the adversarial training, the discriminator is optimized to correctly classify the original input signal and the generated decorrelated output signal, whereas the generator is optimized to fool the discriminator. The proposed reference-free approach enables specifically tailoring the training procedure to the desired output signal properties, with the potential to outperform conventional decorrelation techniques in terms of performance and flexibility. Throughout this chapter, processing is performed at a sampling rate of 22.05 kHz to maintain a reasonable size for both the generator and

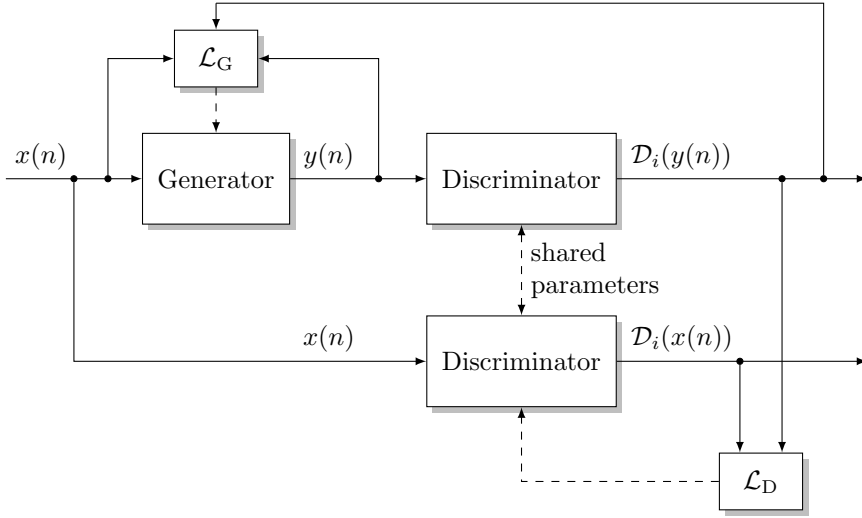


Figure 4.1: Block diagram of proposed single-channel GAN-based audio decorrelation method.

the discriminator models. However, it has been demonstrated that GANs are generally capable of generating audio at higher sampling rates (see, e.g., [138]).

The remainder of this chapter is structured as follows. In Section 4.1, the problem statement regarding single-channel audio decorrelation methods is revisited. Section 4.2 describes the proposed method, and Section 4.3 covers the dataset used for training as well as details regarding the training procedure. Subsequently, the objective and perceptual evaluations are described in Section 4.4. Section 4.5 concludes this chapter by providing a summary.

4.1 Problem Statement

Following the definition provided in Section 3.1, we define audio decorrelation as the task to generate an output signal $y(n)$ from a single-channel input signal $x(n)$, with discrete time index n , that is:

1. Statistically uncorrelated from the input signal $x(n)$. In general, there are various ways to mathematically express the correlation between two signals. In this chapter, the frequency-domain coherence is considered (see Section 4.2.3.1).
2. Perceptually transparent w.r.t. the input signal $x(n)$, i.e., the two audio signals are perceptually indistinguishable [119]. A necessary requirement for perceptual transparency is that the temporal and spectral envelopes of the input signal are not (significantly) altered [92].

4.2 Proposed Method

Figure 4.1 depicts a block diagram of the proposed single-channel GAN-based audio decorrelation method. Given a time-domain single-channel input signal $x(n)$, the goal is to obtain a decorrelated version of this input signal as output of the generator network, denoted here as $y(n)$. The generator's training objective is defined w.r.t. the input signal $x(n)$, based on the two requirements discussed in Section 4.1. Additionally, adversarial training is employed to improve the overall output signal quality.

For the adversarial training, an additional discriminator network is used whose task is to distinguish between the original input signal and the generated output signal. The discriminator network is optimized to correctly classify the input signal as unprocessed and the decorrelated output signal as processed, whereas the generator network is optimized to fool the discriminator into believing that the decorrelated output signal is actually unprocessed. The total generator and discriminator losses are denoted by \mathcal{L}_G and \mathcal{L}_D , respectively. In the following sections, the generator and discriminator architecture and the respective loss functions are described.

4.2.1 Generator Architecture

As generator, we employ the CNN architecture introduced in Section 3.2.2. First, the input signal $x(n)$ is transformed to the short-time Fourier transform (STFT) domain. After processing the frequency-domain signal by a CNN, an inverse STFT is applied, resulting in the output signal $y(n)$. The CNN architecture consists of four 1D causal convolutional layers with a total receptive field of 157 time frames.

While in this chapter processing is performed at a sampling rate of 22.05 kHz, a sampling rate of 48 kHz was used in Chapter 3. Accordingly, an STFT frame length of 116 samples and a hop size of 58 samples are selected to ensure a comparable time and frequency resolution. As a result, the generator model has a total of approximately 1.36 M trainable parameters. All remaining hyperparameters of the generator model are chosen as discussed in Section 3.2.2.

4.2.2 Discriminator Architecture

For the discriminator, we employ the HiFi-GAN discriminators proposed in [133]. In HiFi-GAN, two discriminators are used: a multi-period and a multi-scale discriminator. The multi-period discriminator aims to identify periodic patterns present in the audio signals, whereas the multi-scale discriminator aims to assess consecutive patterns and long-term dependencies. Both discriminators are based on CNN architectures.

The multi-period discriminator consists of five sub-discriminators which operate on reshaped versions of the input signal. Given the sub-discriminator’s period $p \in \{2, 3, 5, 7, 11\}$ and a single-channel input signal with a length of N samples, the input signal is reshaped to a two-dimensional signal with a height of N/p samples and a width of p samples. The multi-scale discriminator consists of three sub-discriminators which operate on different input scales: raw input audio, two-times average-pooled audio, and four-times average-pooled audio. Consequently, the employed discriminator comprises a total of eight sub-discriminators.

The process to determine the output of the i -th sub-discriminator is denoted by the operator $\mathcal{D}_i(\cdot), i = 1, \dots, 8$. An output value of 1 corresponds to classifying the discriminator’s input signal as unprocessed, whereas a value of 0 corresponds to classifying the input signal as processed (i.e., decorrelated). Note that the sub-discriminators’ output is not strictly restricted to values in the range $[0, 1]$. Instead, output values in this range are obtained only after initial convergence as a result of the employed adversarial loss functions defined by (4.4) and (4.5). In total, the discriminator model has approximately 70.7 M trainable parameters.

4.2.3 Generator Loss Functions

The generator loss \mathcal{L}_G consists of three individual loss terms to control the properties of the decorrelated output signal based on the two requirements specified in Section 4.1: a coherence loss \mathcal{L}_{coh} , a mel-spectrogram loss \mathcal{L}_{mel} , and an adversarial loss \mathcal{L}_{adv} . The coherence loss \mathcal{L}_{coh} aims to minimize the correlation between the input and the output signal, the mel-spectrogram loss \mathcal{L}_{mel} aims to minimize spectral magnitude differences between the input and the output signal, and the adversarial loss \mathcal{L}_{adv} realizes the adversarial training with the aim of improving the overall output signal quality.

By weighting \mathcal{L}_{coh} and \mathcal{L}_{mel} with λ_{coh} and λ_{mel} , respectively, the following expression for \mathcal{L}_G is obtained:

$$\mathcal{L}_G = \lambda_{\text{coh}}\mathcal{L}_{\text{coh}} + \lambda_{\text{mel}}\mathcal{L}_{\text{mel}} + \mathcal{L}_{\text{adv}}. \quad (4.1)$$

The weights λ_{coh} and λ_{mel} allow controlling the tradeoff between the degree of input-output correlation and the output signal quality. Consequently, they are key parameters of the proposed method. Since the individual loss terms are all defined in different domains, suitable values for λ_{coh} and λ_{mel} can only be determined empirically. Moreover, they depend heavily on the specific loss term definitions, including any nonlinear mapping or averaging that is employed.

In the following sections, the individual loss term components of the generator loss are described.

4.2.3.1 Coherence Loss

To assess the correlation between $y(n)$ and $x(n)$ according to the first requirement specified in Section 4.1, the frequency-domain coherence is considered. To account for the frequency resolution of the human auditory system, we define the coherence loss \mathcal{L}_{coh} as the mel-scale frequency mean coherence:

$$\mathcal{L}_{\text{coh}} = \frac{1}{M} \sum_{m=1}^M \frac{\mathcal{T}_m(|\sum_l (X(k, l)Y^*(k, l))|)}{\mathcal{T}_m(\sqrt{\sum_l |X(k, l)|^2} \sqrt{\sum_l |Y(k, l)|^2})}, \quad (4.2)$$

where $Y(k, l)$ and $X(k, l)$ are obtained by transforming $y(n)$ and $x(n)$ to the time-frequency domain using an STFT, with k and l denoting the discrete frequency and time frame indices, respectively. For the STFT, a frame length of 1024 samples and a hop size of 256 samples are used. Compared to the generator model (see Section 4.2.1), a longer STFT frame length is used to ensure a sufficiently high frequency resolution in the low-frequency region for the transformation to the mel frequency scale. The operator $\mathcal{T}_m(\cdot)$ denotes the transformation from linear to mel frequency scale, m the mel-scale frequency index, and $(\cdot)^*$ the complex conjugate. The number of mel frequency bands M is set to 80. The coherence loss \mathcal{L}_{coh} takes values in the range $[0, 1]$.

4.2.3.2 Mel-Spectrogram Loss

The mel-spectrogram loss \mathcal{L}_{mel} intends to minimize spectral magnitude differences between $y(n)$ and $x(n)$ and therefore relates to the second requirement specified in Section 4.1. We again employ a mel frequency scale to take the frequency resolution of the human auditory system into account. In particular, \mathcal{L}_{mel} is defined as the ℓ_1 -norm between the mel-spectrogram of $y(n)$ and the mel-spectrogram

of $x(n)$, both expressed in dB:

$$\mathcal{L}_{\text{mel}} = \frac{1}{ML} \sum_{m=1}^M \sum_{l=1}^L \left| 10 \log_{10} \left(\frac{\mathcal{T}_m(|Y(k, l)|^2)}{\mathcal{T}_m(|X(k, l)|^2)} \right) \right|, \quad (4.3)$$

where L describes the number of STFT time frames. The STFT parameters and the number of mel frequency bands are chosen as for the coherence loss \mathcal{L}_{coh} . Magnitude differences in dB between the input and output spectra are more perceptually relevant than differences between them on a linear scale. Given that the objective is to minimize the (perceptual) differences between both spectra on a time-frequency bin basis, the dB values are calculated prior to taking the arithmetic mean. The mel-spectrogram loss \mathcal{L}_{mel} takes values in the range $[0 \text{ dB}, \infty)$.

4.2.3.3 Adversarial Loss

Furthermore, the adversarial loss \mathcal{L}_{adv} is included to enable the adversarial training aimed at improving the overall output signal quality. In particular, the least-squares loss proposed in [139] is used:

$$\mathcal{L}_{\text{adv}} = \frac{1}{8} \sum_{i=1}^8 \left[(\mathcal{D}_i(y(n)) - 1)^2 \right]. \quad (4.4)$$

The adversarial loss \mathcal{L}_{adv} is minimal when $\mathcal{D}_i(y(n)) = 1, \forall i$. This corresponds to the output signal $y(n)$ being classified as unprocessed by all sub-discriminators. After initial convergence, \mathcal{L}_{adv} takes values in the range $[0, 1]$.

4.2.4 Discriminator Loss Function

The discriminator loss \mathcal{L}_{D} contains solely an adversarial loss term, for which the least-squares loss is once again employed [139]:

$$\mathcal{L}_{\text{D}} = \frac{1}{8} \sum_{i=1}^8 \left[(\mathcal{D}_i(x(n)) - 1)^2 + (\mathcal{D}_i(y(n)))^2 \right]. \quad (4.5)$$

The discriminator loss \mathcal{L}_{D} is minimal when $\mathcal{D}_i(x(n)) = 1, \forall i$ and $\mathcal{D}_i(y(n)) = 0, \forall i$. This corresponds to the input signal $x(n)$ being classified as unprocessed and the output signal $y(n)$ being classified as processed by all sub-discriminators. After initial convergence, \mathcal{L}_{D} takes values in the range $[0, 1]$.

4.3 Datasets and Training

The proposed model was trained on music signals. We used the MUSDB18-HQ dataset [123], which contains 150 full-length music tracks with a total length of about 10 h. Only the mixture signals of the dataset were used.

The dataset was randomly split into a training (60%), a validation (20%), and a test (20%) subset. The signals were downsampled to 22.05 kHz and split into segments of 66 120 samples (~ 3 s) each. We added a context of 9280 previous samples (160 STFT time frames) to each segment, which exceeds the receptive field of the generator's CNN. The context is used to determine the generator output

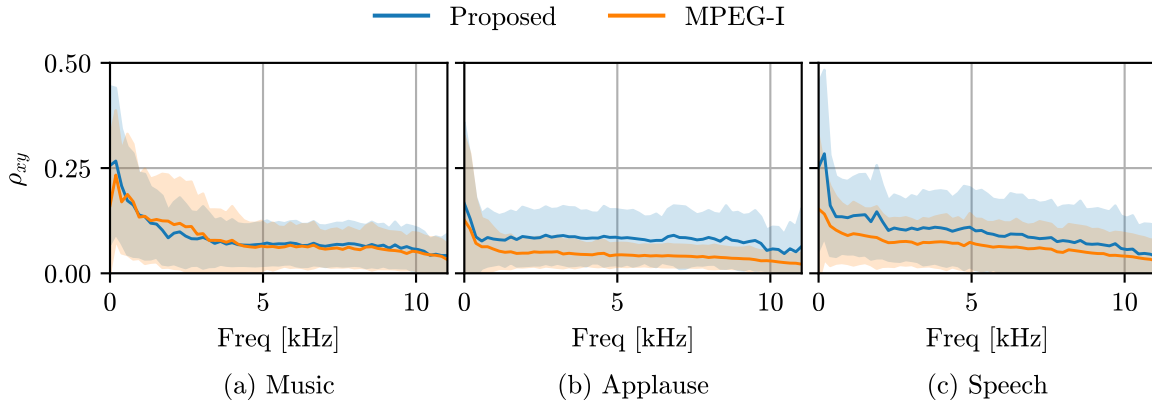


Figure 4.2: Mean and standard deviation of absolute coherence over frequency between input and output signal for proposed method (trained on music signals) and MPEG-I decorrelator for all three test datasets.

but is removed before calculating the loss function and other metrics. A batch size of 16 was used during training. Training was performed using the AdamW optimizer [140] with $\beta = \{0.8, 0.99\}$. The generator and discriminator learning rates were set to 1.6×10^{-3} . Training was performed for 50 K steps; beyond that, no significant improvement in terms of output signal quality could be observed. The discriminator loss \mathcal{L}_D reached a value of around 0.2 at the end of training.

4.4 Performance Evaluation

The proposed method was evaluated both objectively and perceptually. The neural network was trained as described in Sections 4.2 and 4.3, using a coherence loss weight of $\lambda_{\text{coh}} = 2.5$ and a mel-spectrogram loss weight of $\lambda_{\text{mel}} = 0.65$. These values for λ_{coh} and λ_{mel} were selected empirically to achieve both a sufficiently low input-output correlation and an acceptable output signal quality. The proposed method was compared to the conventional decorrelation method described in [93], a state-of-the-art decorrelation technique based on Schroeder all-pass filters and frequency-dependent delays operating in the STFT domain. This method will be referred to as the MPEG-I decorrelator. Note that the transient detection described in [93] was not included in the present study. Furthermore, a slightly different parameterization of the all-pass filters and frequency-dependent delays was used. Specifically, the parameterization described in Section 3.2.1 was used. Since the MPEG-I decorrelator performs processing at 48 kHz, its output was downsampled to 22.05 kHz for a fair comparison. A comparison to the reference-based neural audio decorrelation method proposed in Chapter 3 was not performed, since it was shown to be equivalent to the MPEG-I decorrelator in a perceptual sense.

4.4.1 Objective Evaluation

For the objective evaluation, metrics related to both requirements specified in Section 4.1 were considered. To evaluate the degree of input-output decorrelation achieved by the respective decorrelation method, the frequency-dependent absolute coherence was evaluated. Given two STFT-domain signals $X(k, l)$ and $Y(k, l)$, the frequency-dependent absolute coherence $\rho_{xy}(k)$ between both is defined as

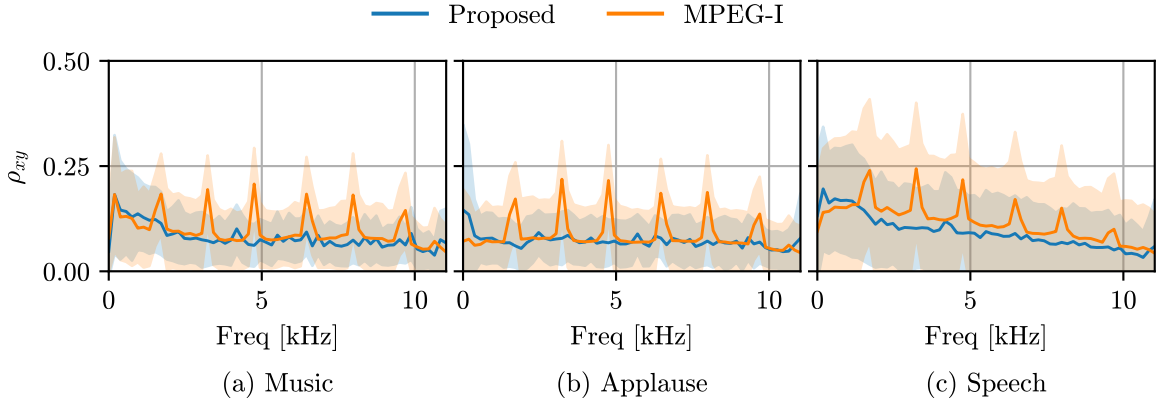


Figure 4.3: Mean and standard deviation of absolute coherence over frequency between channels of stereo signal for proposed method (trained on music signals) and MPEG-I decorrelator for all three test datasets.

follows:

$$\rho_{xy}(k) = \frac{|\sum_l X(k, l)Y^*(k, l)|}{\sqrt{\sum_l |X(k, l)|^2 \sum_l |Y(k, l)|^2}}. \quad (4.6)$$

As for the generator model (see Section 4.2.1), an STFT frame length of 116 samples and a hop size of 58 samples were used. To evaluate the (perceptual) similarity between the output and the input signal, we consider differences in terms of the signals’ magnitude spectra by evaluating the mel-spectrogram loss \mathcal{L}_{mel} , calculated as defined in (4.3). In addition to evaluating the performance on the test subset of the music dataset used for training, we evaluated the model on applause and speech signals to gain information about the generalization properties of the network. For this purpose, we made use of the FSD50K dataset [127]. Items labeled as “applause” were used to represent the class of applause signals, and items labeled as “speech” were used to represent the class of speech signals.

Figure 4.2 shows the mean and standard deviation of the absolute coherence over frequency between the decorrelator’s input and output signals for the proposed method and the MPEG-I decorrelator for all three test datasets. For the music test dataset, comparable coherence values are obtained for the proposed method and the MPEG-I decorrelator, indicating that a similar amount of decorrelation is achieved. Although the coherence values are also relatively low for the applause and speech test datasets, the proposed method yields slightly higher coherence values than the MPEG-I decorrelator. A possible reason for this is that these signal types were not included in the training data.

In the perceptual evaluation discussed in Section 4.4.2, the effectiveness of the decorrelation is assessed using the stereo signal defined by (4.7). Therefore, we additionally evaluate the absolute coherence between the channels of the stereo signal. Figure 4.3 depicts the mean and standard deviation of the absolute coherence over frequency between the channels of the stereo signal for the proposed method and the MPEG-I decorrelator for all three test datasets. With the exception of frequency-selective peaks in the coherence of the MPEG-I decorrelator, comparable coherence values are obtained for the proposed method and the MPEG-I decorrelator for the music test dataset. For all three test datasets, the differences in terms of absolute coherence between the proposed method and the MPEG-I decorrelator are clearly below the just-noticeable difference (JND) of the interaural cross-correlation (ICC), which, according to several studies [66, 67], lies between 0.3 and 0.5 given a reference correlation

Table 4.1: Mean and standard deviation of \mathcal{L}_{mel} [dB] for proposed method (trained on music signals) and MPEG-I decorrelator for all three test datasets.

Test dataset	Proposed method	MPEG-I decorrelator
Music	1.94 ± 0.31	3.25 ± 0.50
Applause	2.02 ± 0.33	3.28 ± 0.60
Speech	2.71 ± 1.20	3.93 ± 1.22

of zero.

Table 4.1 shows the mean and standard deviation of \mathcal{L}_{mel} for the proposed method and the MPEG-I decorrelator for all three test datasets. The proposed method considerably outperforms the MPEG-I decorrelator in terms of \mathcal{L}_{mel} for all considered test datasets. Consequently, the input signal’s magnitude spectrum is better preserved. Furthermore, in terms of \mathcal{L}_{mel} , the proposed method generalizes well w.r.t. applause and speech signals, which were not included in the training data.

4.4.2 Perceptual Evaluation

Two formal listening tests were conducted to perform a perceptual evaluation of the proposed method, considering both requirements specified in Section 4.1. The first test evaluated the effectiveness of the decorrelation in terms of perceived stereo signal envelopment. Note that the perceived envelopment is closely related to the ICC, as discussed in Section 2.1.6. The second test compared the proposed method’s output to the input signal in terms of overall mono signal quality, thereby evaluating the perceptual similarity between the output and the input signal. All items used in the listening tests are available at www.audiolabs-erlangen.de/resources/2023-WASPAA-Decorrelation-GAN.

A total of seven test items were incorporated in each listening test. The stimuli were partly taken from the EBU SQAM CD [128] and the FSD50K dataset [127]. Four music stimuli were included: two pop songs (similar to the training data), one particularly transient music item (castanets), and one particularly tonal music item (violin). Additionally, one applause, one speech, and one ocean waves item were included. The items were reproduced over headphones using a customized version of the webMUSHRA software [130]. A total of 14 subjects participated in each listening test, 12 male and two female. The average age of the participants was 29 years, and all of them had prior listening test experience.

4.4.2.1 Stereo Signal Decorrelation Test

The stereo signal decorrelation test evaluated the effectiveness of the decorrelation in terms of perceived envelopment.

The stereo signals were generated by following the procedure outlined in Appendix A.1. According to (A.3), the stereo signal $\mathbf{y}_s(n)$ was generated from the input signal $x(n)$ and the decorrelated output signal $y(n)$ as follows:

$$\mathbf{y}_s(n) = \frac{1}{\sqrt{2}} \begin{bmatrix} x(n) + y(n) \\ x(n) - y(n) \end{bmatrix}. \quad (4.7)$$

Assessing the resulting stereo signals allows judging the effectiveness of the decorrelation for the respective processing method while ensuring similar temporal characteristics of the left and right channel

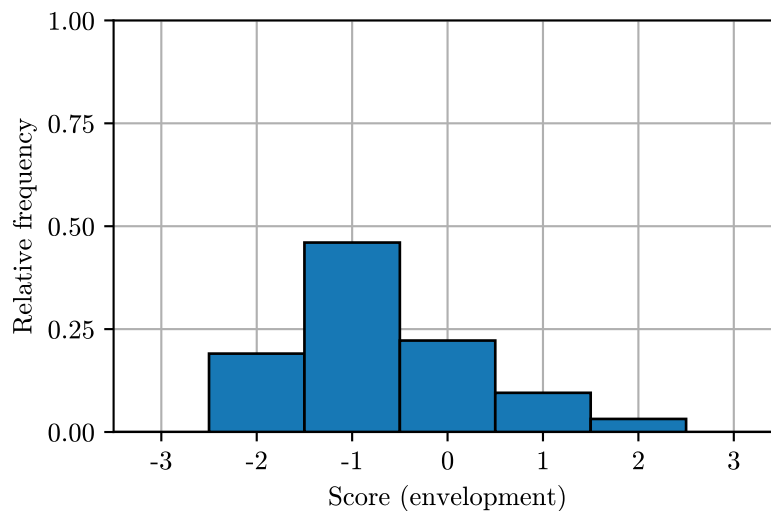


Figure 4.4: Relative frequency of stereo signal listening test scores for all items in terms of perceived envelopment, ratings of proposed method w.r.t. MPEG-I decorrelator.

signals.

The conditions under test included the proposed method and the MPEG-I decorrelator. Since no ideal reference could be defined, the unprocessed mono input signal was provided as guidance by reproducing it over the left and right headphones. The participants were asked to rate one condition w.r.t. the other in terms of perceived envelopment on a discrete seven-point scale: “much less enveloped” (−3), “less enveloped” (−2), “slightly less enveloped” (−1), “the same as” (0), “slightly more enveloped” (1), “more enveloped” (2), “much more enveloped” (3), based on the comparison category rating method described in the ITU-T P.800 recommendation [129]. To ensure a common understanding of the term envelopment, the listeners were presented with examples of a coherent (less enveloped) and an incoherent (more enveloped) white noise stimulus prior to conducting the listening test.

Figure 4.4 shows the relative frequency of all listening test scores for ratings of the proposed method w.r.t. the MPEG-I decorrelator. Most scores were given for a rating of −1, which corresponds to “slightly less enveloped.” The median value of the listening test scores as well as the upper and lower bounds of its bootstrapped 95% confidence interval (CI) were found to be equal to −1. It can thus be concluded that, on average, the proposed method is perceived as “slightly less enveloped” compared to the MPEG-I decorrelator for the considered items. This is not fully in accordance with the objective evaluation results (see Section 4.4.1), which indicates that the coherence measure considered in the objective evaluation does not capture the perceptual impression to its full extent.

4.4.2.2 Mono Signal Quality Test

The mono signal quality test evaluated the perceptual similarity between the output signal of the respective processing method and the input signal. A MUSHRA listening test [141] was conducted, in which the unprocessed input signal corresponded to the reference signal. The conditions under test included the proposed method and the MPEG-I decorrelator. Additionally, a 3.5 kHz low-pass anchor and a hidden reference were included. The respective mono signals were reproduced over the left and right headphones. The participants were asked to rate the presented conditions w.r.t. the reference signal (i.e., the input signal) in terms of overall audio quality on a continuous scale ranging from 0

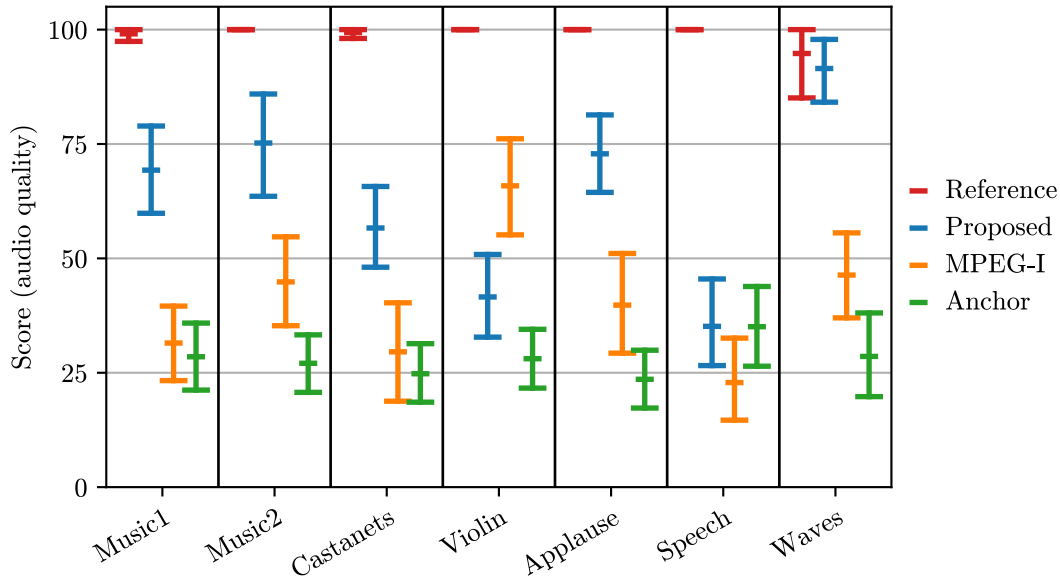


Figure 4.5: Mean values and bootstrapped 95 % CIs of mono signal listening test scores per item in terms of overall audio quality.

(bad) to 100 (excellent).

Figure 4.5 shows the mean values and bootstrapped 95 % CIs of the listening test scores for all items and conditions under test. The proposed method significantly outperforms the MPEG-I decorrelator for all items except the violin item. This is in accordance with the objective evaluation (Section 4.4.1), which showed that the proposed method better preserves the input signal’s magnitude spectrum. For the violin item, the MPEG-I decorrelator outperforms the proposed method, which indicates that there is room for improvement of the proposed method w.r.t. the mono signal quality of particularly tonal signals.

4.4.3 Discussion on Stereo Signal Quality

In Section 4.4.2, the overall audio quality of the proposed method’s output was assessed by directly comparing it to the input signal. Similarly, Section 4.4.1 evaluated the similarity between the output and input signals’ magnitude spectra by means of the mel-spectrogram loss \mathcal{L}_{mel} . However, the audio quality of the stereo signal $\mathbf{y}_s(n)$, as defined in (4.7), was not explicitly assessed. Given that certain applications, such as the homogeneous spatially extended sound source (SESS) rendering method proposed in Chapter 6, directly use the stereo signal $\mathbf{y}_s(n)$, this section provides a brief discussion on informal perceptual observations concerning the stereo signal quality.

Depending on the signal type, the proposed method’s stereo signal is affected by fine-granular time-frequency-dependent attenuation/amplification of the individual channels relative to the input signal, which can be observed by inspecting the stereo signals’ magnitude spectrum. Informal listening revealed that this results in noticeable artifacts, particularly for noise-like signals that densely cover the time-frequency spectrum. To support the perceptual observations, we evaluated the spectral magnitude differences between the individual channels of the stereo signal and the input signal, as well as between the channels of the stereo signal. Accordingly, we determine $D_{\text{st},\text{in}}$, which assesses spectral magnitude differences w.r.t. the input signal, and $D_{\text{st},\text{ch}}$, which assesses spectral magnitude differences between

Table 4.2: Mean and standard deviation of $D_{\text{st,in}}$ [dB] for proposed method (trained on music signals) and MPEG-I decorrelator for all three test datasets.

Test dataset	Proposed method	MPEG-I decorrelator
Music	2.25 ± 0.17	2.44 ± 0.23
Applause	2.22 ± 0.18	2.28 ± 0.25
Speech	2.60 ± 0.84	2.61 ± 0.65

Table 4.3: Mean and standard deviation of $D_{\text{st,ch}}$ [dB] for proposed method (trained on music signals) and MPEG-I decorrelator for all three test datasets.

Test dataset	Proposed method	MPEG-I decorrelator
Music	3.71 ± 0.18	3.08 ± 0.25
Applause	3.63 ± 0.12	2.85 ± 0.11
Speech	3.33 ± 0.41	2.68 ± 0.36

the two stereo signal channels. Similarly to \mathcal{L}_{mel} , the definitions of $D_{\text{st,in}}$ and $D_{\text{st,ch}}$ are as follows:

$$D_{\text{st,in}} = \frac{1}{ML} \sum_{m=1}^M \sum_{l=1}^L \frac{1}{2} \left(\left| 10 \log_{10} \left(\frac{\mathcal{T}_m(|Y_{s,1}(k,l)|^2)}{\mathcal{T}_m(|X(k,l)|^2)} \right) \right| + \left| 10 \log_{10} \left(\frac{\mathcal{T}_m(|Y_{s,2}(k,l)|^2)}{\mathcal{T}_m(|X(k,l)|^2)} \right) \right| \right), \quad (4.8)$$

$$D_{\text{st,ch}} = \frac{1}{ML} \sum_{m=1}^M \sum_{l=1}^L \left| 10 \log_{10} \left(\frac{\mathcal{T}_m(|Y_{s,1}(k,l)|^2)}{\mathcal{T}_m(|Y_{s,2}(k,l)|^2)} \right) \right|, \quad (4.9)$$

where $Y_{s,1}(k,l)$ and $Y_{s,2}(k,l)$ are obtained by transforming the two channels of $\mathbf{y}_s(n)$ to the time-frequency domain using an STFT. The STFT parameters and the number of mel frequency bands are chosen as for the mel-spectrogram loss \mathcal{L}_{mel} . Tables 4.2 and 4.3 present the mean and standard deviation of $D_{\text{st,in}}$ and $D_{\text{st,ch}}$, respectively, for the proposed method and the MPEG-I decorrelator for all three test datasets. For the proposed method, the mean values of $D_{\text{st,in}}$ and $D_{\text{st,ch}}$ are mostly higher than those of \mathcal{L}_{mel} (see Table 4.1). With regard to $D_{\text{st,in}}$, the mean values for the proposed method are marginally lower than those for the MPEG-I decorrelator (between 0.01 dB and 0.19 dB, depending on the test dataset). In contrast, the mean values of $D_{\text{st,ch}}$ are clearly lower for the MPEG-I decorrelator than for the proposed method (between 0.63 dB and 0.78 dB, depending on the test dataset). The clearly higher mean values of $D_{\text{st,ch}}$ in comparison to $D_{\text{st,in}}$ for the proposed method (between 0.73 dB and 1.46 dB, depending on the test dataset) indicate the presence of complementary constructive and destructive interference in the stereo signal channels, which may contribute to the observed signal artifacts.

In light of the discussed observations, Chapter 8 proposes an extended version of the generator loss with the aim to improve the overall audio quality of the stereo signal, targeting the homogeneous SESS rendering method proposed in Chapter 6 as a specific application.

4.5 Summary

In this chapter, we proposed a novel approach to audio decorrelation using GANs. As generator, we employed the CNN model proposed in Chapter 3, which was designed for the specific task of audio decorrelation. For the discriminator network, we made use of the HiFi-GAN discriminators [133], which were originally proposed for speech synthesis. In contrast to the training procedure employed in Chapter 3, the approach proposed in this chapter does not rely on a decorrelated reference signal. The training objective is defined directly w.r.t. the input signal and consists of a number of individual loss terms to control both the input-output correlation and the signal quality of the output signal.

The model was trained on music signals, selecting a specific weighting of the respective loss terms that offers a reasonable tradeoff between the input-output correlation and the output signal quality. An objective evaluation showed that the proposed method outperforms the state-of-the-art conventional decorrelation method described in [93] in terms of spectral magnitude reconstruction considering the ℓ_1 -norm between the input and output signals' mel-spectrograms. Moreover, a comparable performance is achieved in terms of the resulting input-output coherence. The objective evaluation furthermore showed that the proposed method generalizes well to applause and speech signals, which were not included in the training dataset. Two formal listening tests were conducted to verify the objective findings. In terms of mono signal quality, the proposed method significantly outperformed the conventional method for the majority of items. Furthermore, the stereo signal decorrelation test showed that the proposed method's output is perceived as only "slightly less enveloped" compared to the output of the conventional decorrelation method.

CHAPTER 5

Multi-Channel Audio Decorrelation Using Generative Adversarial Networks

The main content of this chapter is based on: C. Anemüller, O. Thiergart, and E. A. P. Habets, “Multi-channel neural audio decorrelation using generative adversarial networks,” EURASIP J. Audio, Speech Music. Process., vol. 2024, no. 58, Nov. 2024, [18].

In Chapter 4, we proposed a single-channel neural audio decorrelation method based on generative adversarial networks (GANs). As generator, the convolutional neural network (CNN) architecture introduced in Chapter 3 was employed, which was designed for the specific task of audio decorrelation. While in Chapter 3, the CNN was trained with the help of a decorrelated reference signal, in Chapter 4, a reference-free training procedure was proposed. The proposed training objective is defined w.r.t. the audio input signal and consists of a number of individual loss terms that control both the input-output correlation and the output signal quality. Although the method proposed in Chapter 4 was shown to reach a good performance compared to the state-of-the-art short-time Fourier transform (STFT)-domain decorrelator proposed in [93], it is limited to single-channel output signals. Moreover, the evaluation provided in Chapter 4 is restricted to one specific network configuration.

In this chapter, we extend the single-channel GAN-based audio decorrelation method proposed in Chapter 4 to provide a multi-channel output signal. A separate generator network is employed for each output channel. All generator networks are optimized jointly to obtain output channels that are mutually uncorrelated and exhibit both a low correlation and a high perceptual similarity to the input signal. As in Chapter 4, processing is performed at a sampling rate of 22.05 kHz. The performance of the proposed method is evaluated objectively as well as through formal listening tests. To this end, a comparison with two classical signal processing-based multi-channel decorrelators is performed. Furthermore, a naive approach is considered in which the generator networks for the different output channels are optimized independently, using different seeds for their parameter initialization. Additionally, we investigate the influence of the number of output channels, the individual loss term weightings, and the employed training data on the neural network’s performance.

The remainder of this chapter is structured as follows. In Section 5.1 the problem statement regarding multi-channel audio decorrelation methods is provided. Section 5.2 describes the proposed multi-

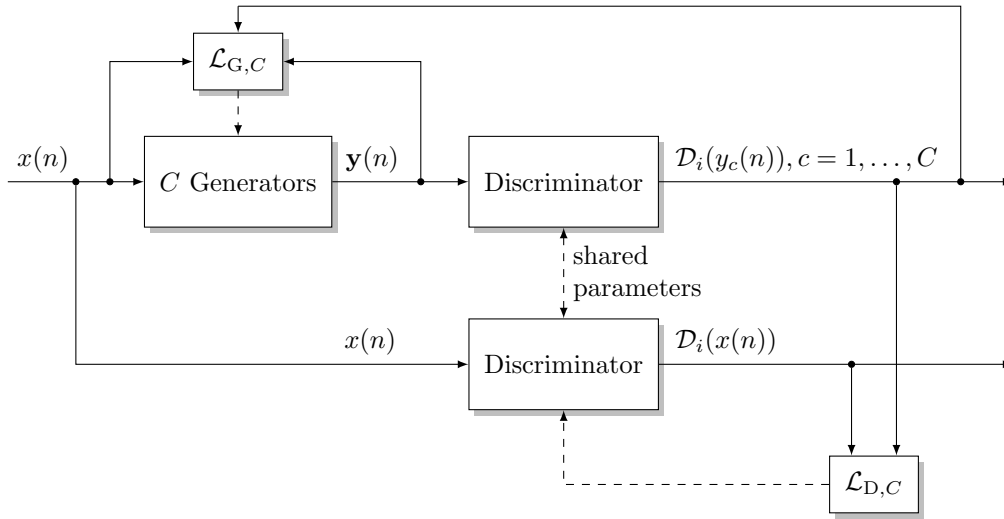


Figure 5.1: Block diagram of proposed multi-channel GAN-based audio decorrelation method.

channel decorrelation method, and Section 5.3 covers the datasets used for training and evaluation as well as details regarding the training procedure. Subsequently, the performance evaluation is described in Section 5.4. Section 5.5 concludes this chapter by providing a summary.

5.1 Problem Statement

In this chapter, we consider the task of generating a decorrelated multi-channel output signal $\mathbf{y}(n) = [y_1(n), y_2(n), \dots, y_C(n)]^T$ given a single-channel input signal $x(n)$, with n denoting the discrete time index and C denoting the number of output channels. The operator $(\cdot)^T$ denotes the transpose operation. By extending the requirements of decorrelation methods specified in Section 3.1 to the considered multi-channel scenario, we define the following two requirements that should ideally be met by the decorrelated output signal $\mathbf{y}(n)$:

1. Each channel of $\mathbf{y}(n)$ is uncorrelated from the input signal $x(n)$. Additionally, all output channels are mutually uncorrelated. The individual output channels are required to be uncorrelated from the input signal for applications where the output signals are used in conjunction with the original input signal, e.g., in the context of spatial audio reproduction [7]. In general, there are various ways to mathematically express the correlation between two signals. In this chapter, the frequency-domain coherence is considered (see Section 5.2.3.1).
2. Each channel of $\mathbf{y}(n)$ is perceptually transparent w.r.t. the input signal $x(n)$, i.e., the two audio signals are perceptually indistinguishable [119]. A necessary requirement for perceptual transparency is that the input signal's temporal and spectral envelopes are not (significantly) altered [92].

5.2 Proposed Method

Figure 5.1 shows a block diagram of the proposed multi-channel GAN-based audio decorrelation

method. The input to the method is a single-channel signal $x(n)$. By processing $x(n)$ using C individual generator networks, the goal is to obtain a decorrelated multi-channel output signal $\mathbf{y}(n)$, where the element $y_c(n)$, $c \in \{1, 2, \dots, C\}$, corresponds to the output of the c -th generator network. The individual generator networks are optimized jointly, considering the two requirements specified in Section 5.1. During the optimization process, loss components related to the input signal are incorporated, and adversarial training is employed. For the adversarial training, a single discriminator network is used whose task is to distinguish between the original input signal and the channels of the generated multi-channel output signal. The discriminator network is optimized to correctly classify the input signal as unprocessed and each individual channel of the decorrelated output signal as processed, whereas the generator networks are optimized to fool the discriminator into believing that each individual channel of the decorrelated output signal is actually unprocessed. The total multi-channel generator and discriminator losses are denoted by $\mathcal{L}_{G,C}$ and $\mathcal{L}_{D,C}$, respectively. The proposed multi-channel decorrelation method is a generalization of the single-channel decorrelation method introduced in Chapter 4. In the following sections, the generator and discriminator architecture and the employed loss functions are described.

5.2.1 Generator Architecture

As generator, we use the STFT-domain CNN architecture introduced in Section 3.2.2, which was also employed in the single-channel method proposed in Chapter 4. The CNN architecture consists of four 1D causal convolutional layers with a total receptive field of 157 time frames. As in Chapter 4, an STFT frame length of 116 samples and a hop size of 58 samples are used. The STFT parameters are chosen identically to Chapter 4, given that processing is performed at the same sampling rate of 22.05 kHz.

For each output channel, an individual generator network is employed. Except for independent parameter initialization, the C generator networks are identical. In total, the generator model has approximately $C \times 1.36$ M trainable parameters.

5.2.2 Discriminator Architecture

For the discriminator, we use the HiFi-GAN discriminator proposed in [133], which was also employed in Chapter 4. The HiFi-GAN discriminator consists of eight sub-discriminators: five multi-period and three multi-scale discriminators, all based on CNN architectures. The multi-period discriminators aim to identify periodic patterns present in the audio signals, whereas the multi-scale discriminators aim to assess consecutive patterns and long-term dependencies.

The process to determine the output of the i -th sub-discriminator is denoted by the operator $\mathcal{D}_i(\cdot)$, with $i = 1, \dots, 8$. An output value of 1 corresponds to classifying the discriminator’s input signal as unprocessed, whereas a value of 0 corresponds to classifying the input signal as processed (i.e., decorrelated). Note that the sub-discriminators’ output is not strictly restricted to values in the range $[0, 1]$. Instead, output values in this range are obtained only after initial convergence as a result of the employed adversarial loss functions described in Sections 5.2.3.3 and 5.2.4. In total, the discriminator model has roughly 70.7 M trainable parameters. Since only a single discriminator model is employed $C + 1$ times, the number of trainable parameters is independent of the number of output channels C .

5.2.3 Generator Loss Functions

The multi-channel generator loss $\mathcal{L}_{G,C}$ consists of three individual loss terms to control the properties of the decorrelated multi-channel output signal based on the two requirements specified in Section 5.1: a multi-channel coherence loss $\mathcal{L}_{\text{coh},C}$, a multi-channel mel-spectrogram loss $\mathcal{L}_{\text{mel},C}$, and a multi-channel adversarial loss $\mathcal{L}_{\text{adv},C}$. The multi-channel coherence loss $\mathcal{L}_{\text{coh},C}$ aims to minimize the correlation between the input signal and each individual output channel as well as between all output channel pairs, the multi-channel mel-spectrogram loss $\mathcal{L}_{\text{mel},C}$ aims to minimize spectral magnitude differences between the individual output channels and the input signal, and the multi-channel adversarial loss $\mathcal{L}_{\text{adv},C}$ realizes the adversarial training with the aim of improving the overall quality of each individual output channel.

The total multi-channel generator loss $\mathcal{L}_{G,C}$ is obtained as a weighted linear combination of the individual loss terms:

$$\mathcal{L}_{G,C} = \lambda_{\text{coh}}\mathcal{L}_{\text{coh},C} + \lambda_{\text{mel}}\mathcal{L}_{\text{mel},C} + \mathcal{L}_{\text{adv},C}, \quad (5.1)$$

where λ_{coh} and λ_{mel} describe the coherence and mel-spectrogram loss weights, respectively. Since the individual loss terms are all defined in different domains, suitable values for λ_{coh} and λ_{mel} can only be determined empirically. Moreover, they depend heavily on the specific loss term definitions, including any nonlinear mapping or averaging that is employed. In the following sections, the individual loss term components of the multi-channel generator loss are described.

5.2.3.1 Coherence Loss

To assess the correlation between the individual signal pairs according to the first requirement specified in Section 5.1, the frequency-domain coherence is considered. Analogously to (4.2), the single-channel coherence loss \mathcal{L}_{coh} , between two STFT-domain signals $Y_c(k, l)$ and $X(k, l)$, is defined as the mel-scale frequency mean coherence:

$$\mathcal{L}_{\text{coh}}(Y_c, X) = \frac{1}{M} \sum_{m=1}^M \frac{\mathcal{T}_m(|\sum_l (Y_c(k, l)X^*(k, l))|)}{\mathcal{T}_m(\sqrt{\sum_l |Y_c(k, l)|^2 \sum_l |X(k, l)|^2})}, \quad (5.2)$$

where $Y_c(k, l)$ and $X(k, l)$ are obtained by transforming $y_c(n)$ and $x(n)$ to the time-frequency domain using an STFT, with k and l denoting the discrete frequency and time frame indices, respectively. The operator $\mathcal{T}_m(\cdot)$ denotes the transformation from linear to mel frequency scale, m the mel-scale frequency index, and $(\cdot)^*$ the complex conjugate. For the coherence loss, an STFT frame length of 128 samples (5.8 ms) and a hop size of 64 samples (2.9 ms) are used. The number of mel frequency bands M is set to 10.

In accordance with the first requirement specified in Section 5.1, the multi-channel coherence loss $\mathcal{L}_{\text{coh},C}$ aims to minimize the input-output coherence between each individual channel of $\mathbf{y}(n)$ and $x(n)$, as well as the inter-channel coherence between all channels of $\mathbf{y}(n)$. To achieve this, $\mathcal{L}_{\text{coh},C}$ is defined as the maximum of the single-channel coherence loss \mathcal{L}_{coh} over all relevant signal pairs:

$$\mathcal{L}_{\text{coh},C} = \max \left(\max_c \mathcal{L}_{\text{coh}}(Y_c, X), \max_{\{c,z\} | c \neq z} \mathcal{L}_{\text{coh}}(Y_c, Y_z) \right). \quad (5.3)$$

Note that the indices k and l have been omitted for the sake of notational brevity. \mathcal{L}_{coh} and, consequently, $\mathcal{L}_{\text{coh},C}$ take values in the range $[0, 1]$.

The time-frequency-domain conversion parameters of the coherence loss (with an STFT frame length of 5.8 ms and a hop size of 2.9 ms) are chosen differently than in Chapter 4, where significantly longer STFT frames were employed. For the considered multi-channel scenario, the selected time-frequency-domain conversion parameters were found to result in a favorable convergence behavior with regard to the output signal quality. From a perceptual point of view, the correlation between two signals is of interest for signal delays in the range of ± 1 ms [26]. The STFT parameters chosen in this chapter result in a time resolution that more closely aligns with the perceptually relevant delay range, while maintaining a sufficiently high frequency resolution.

While it may seem more intuitive to choose the multi-channel coherence loss as the average value over all signal pairs, minimizing its maximum value was found to result in a more uniform convergence behavior across the different output channels, especially for larger values of C . More specifically, minimizing its average value may cause one of the output channels to converge toward the input signal. The reasoning for this is as follows. If one of the output channels is equal to the input signal, we obtain a single low mel-spectrogram and adversarial loss term at the cost of a single high coherence loss term. While for the coherence loss, the number of individual loss terms increases quadratically with the number of output channels, for the multi-channel mel-spectrogram and adversarial losses, the number of individual loss terms increases only linearly with the number of output channels (see (5.4) and (5.6)). Therefore, a single coherence loss term weighs significantly less than a single mel-spectrogram/adversarial loss term for large values of C . This makes the convergence of a single output channel toward the input signal a preferred solution in terms of the resulting loss value. Opting for the maximum-based coherence loss calculation avoids this solution, as it would result in a very high multi-channel coherence loss.

5.2.3.2 Mel-Spectrogram Loss

The multi-channel mel-spectrogram loss $\mathcal{L}_{\text{mel},C}$ aims to minimize spectral magnitude differences between the individual channels of $\mathbf{y}(n)$ and the input signal $x(n)$. Therefore, it relates to the second requirement specified in Section 5.1. $\mathcal{L}_{\text{mel},C}$ is defined as the average over the single-channel mel-spectrogram loss \mathcal{L}_{mel} of the individual output channels:

$$\mathcal{L}_{\text{mel},C} = \frac{1}{C} \sum_{c=1}^C \mathcal{L}_{\text{mel}}(Y_c, X), \quad (5.4)$$

with \mathcal{L}_{mel} defined analogously to (4.3):

$$\mathcal{L}_{\text{mel}}(Y_c, X) = \frac{1}{ML} \sum_{m=1}^M \sum_{l=1}^L \left| 10 \log_{10} \left(\frac{\mathcal{T}_m(|Y_c(k, l)|^2)}{\mathcal{T}_m(|X(k, l)|^2)} \right) \right|, \quad (5.5)$$

where L describes the number of STFT time frames. For the mel-spectrogram loss, an STFT frame length of 1024 samples, an STFT hop size of 256 samples, and $M = 80$ mel frequency bands are used. Compared to the coherence loss, a larger STFT frame length is selected to ensure a sufficiently high frequency resolution. \mathcal{L}_{mel} and, consequently, $\mathcal{L}_{\text{mel},C}$ take values in the range $[0 \text{ dB}, \infty)$.

5.2.3.3 Adversarial Loss

Furthermore, a multi-channel adversarial loss is included to enable the adversarial training aimed at improving the overall quality of each individual output channel. Similarly to $\mathcal{L}_{\text{mel},C}$, the multi-channel adversarial loss $\mathcal{L}_{\text{adv},C}$ is defined as the average over the single-channel adversarial loss \mathcal{L}_{adv} of the individual output channels:

$$\mathcal{L}_{\text{adv},C} = \frac{1}{C} \sum_{c=1}^C \mathcal{L}_{\text{adv}}(y_c), \quad (5.6)$$

where \mathcal{L}_{adv} is defined as the least-squares loss [139]:

$$\mathcal{L}_{\text{adv}}(y_c) = \frac{1}{8} \sum_{i=1}^8 \left[(\mathcal{D}_i(y_c(n)) - 1)^2 \right]. \quad (5.7)$$

The multi-channel adversarial loss $\mathcal{L}_{\text{adv},C}$ is minimal when $\mathcal{D}_i(y_c(n)) = 1, \forall i, c$. This corresponds to each individual output channel $y_c(n)$ being classified as unprocessed by all sub-discriminators. After initial convergence, \mathcal{L}_{adv} and, consequently, $\mathcal{L}_{\text{adv},C}$ take values in the range $[0, 1]$.

5.2.4 Discriminator Loss Function

The multi-channel discriminator loss $\mathcal{L}_{D,C}$ only includes an adversarial loss term. To determine $\mathcal{L}_{D,C}$, the single-channel discriminator loss \mathcal{L}_D is averaged over all output channels:

$$\mathcal{L}_{D,C} = \frac{1}{C} \sum_{c=1}^C \mathcal{L}_D(x, y_c). \quad (5.8)$$

Analogously to the generator's adversarial loss described by (5.7), \mathcal{L}_D is defined as the least-squares loss [139]:

$$\mathcal{L}_D(x, y_c) = \frac{1}{8} \sum_{i=1}^8 \left[(\mathcal{D}_i(x(n)) - 1)^2 + (\mathcal{D}_i(y_c(n)))^2 \right]. \quad (5.9)$$

The multi-channel discriminator loss $\mathcal{L}_{D,C}$ is minimal when $\mathcal{D}_i(x(n)) = 1, \forall i$ and $\mathcal{D}_i(y_c(n)) = 0, \forall i, c$. This corresponds to the input signal $x(n)$ being classified as unprocessed, and each individual output channel $y_c(n)$ being classified as processed by all sub-discriminators. After initial convergence, \mathcal{L}_D and, consequently, $\mathcal{L}_{D,C}$ take values in the range $[0, 1]$.

5.3 Datasets and Training

For training and evaluation of the proposed method, we focused on the class of music signals. To represent general music signals, we used the mixture signals of the MUSDB18-HQ dataset [123], which consists of 150 full-length music tracks with a total length of about 10 h. This dataset will be referred to as the music dataset. Additionally, the MAESTRO dataset [142] was used to represent the more specific class of piano signals. To obtain a dataset with a size comparable to the MUSDB18-HQ dataset, a 10 h subset of the MAESTRO dataset was randomly selected. This dataset will be referred

to as the piano dataset.

Both datasets were randomly split into a training (60%), a validation (20%), and a test (20%) subset. The signals were downsampled to 22.05 kHz and split into segments of 66 120 samples (~ 3 s) each. A context of 9280 previous samples (160 STFT time frames) was added to each segment, which exceeds the receptive field of the CNN of the generator networks. The context is used to determine the output of the generator networks but is removed before calculating the loss function and other metrics. A batch size of 8 was used during training. Training was performed using the AdamW optimizer [140] with $\beta = \{0.8, 0.99\}$. The generator and discriminator learning rates were set to 1.6×10^{-3} . Training was performed for 50 K steps.

5.4 Performance Evaluation

The performance of the proposed method was evaluated objectively, considering metrics such as coherence and mel-spectrogram loss, as well as perceptually through formal listening tests. The proposed method was compared to two classical signal processing-based multi-channel decorrelators. A naive independent channel optimization alternative to the proposed method was furthermore considered. Additionally, the influence of the coherence loss weight λ_{coh} , the number of output channels C , and the employed training dataset on the proposed method's performance was investigated. For all experiments, the neural network was trained as described in Sections 5.2 and 5.3 using a mel-spectrogram loss weight of $\lambda_{\text{mel}} = 0.65$.

5.4.1 Comparison Methods

As the first comparison method, we considered the lattice all-pass filter-based quadrature mirror filter (QMF)-domain decorrelator employed in MPEG Surround [26, 92], which will be referred to as the MPS decorrelator. In MPEG Surround, a multi-channel output signal is obtained from a single-channel input signal by concatenating multiple so-called one-to-two (OTT) decoding blocks in a tree structure. Each decoding block uses an individual single-channel MPS decorrelator instance. In particular, a decorrelated C -channel output signal can be obtained by concatenating $C - 1$ OTT decoding blocks, setting both the inter-channel coherence and the inter-channel level difference parameters to 0. Using this tree structure, for $C = 4$ output channels, the following expressions are obtained for the individual channels of the decorrelated multi-channel output signal $\mathbf{y}(n)$:

$$\begin{aligned} y_1(n) &= \frac{1}{\sqrt{2}} \left(\frac{1}{\sqrt{2}} (x(n) + \text{dec}_1\{x(n)\}) + \text{dec}_2\{x(n)\} \right), \\ y_2(n) &= \frac{1}{\sqrt{2}} \left(\frac{1}{\sqrt{2}} (x(n) + \text{dec}_1\{x(n)\}) - \text{dec}_2\{x(n)\} \right), \\ y_3(n) &= \frac{1}{\sqrt{2}} \left(\frac{1}{\sqrt{2}} (x(n) - \text{dec}_1\{x(n)\}) + \text{dec}_3\{x(n)\} \right), \\ y_4(n) &= \frac{1}{\sqrt{2}} \left(\frac{1}{\sqrt{2}} (x(n) - \text{dec}_1\{x(n)\}) - \text{dec}_3\{x(n)\} \right), \end{aligned} \tag{5.10}$$

where $\text{dec}_i\{\cdot\}$, $i \in \{1, 2, 3\}$, describes the process to determine the output of the i -th single-channel decorrelator instance. More details on the employed tree structure can be found in Appendix A.1. Note that by calculating $\mathbf{y}(n)$ according to (5.10), the final decorrelated multi-channel output signal

is obtained by mixing the original input signal with the outputs of the individual decorrelators. As a consequence, the decorrelated multi-channel output signal is partially correlated with the original input signal.

As the second comparison method, we considered the QMF-domain decorrelator employed in [90], which applies frequency-dependent pseudo-random delays to the input signal. Therefore, this method will be referred to as the delay decorrelator. The applied delays were varied between approximately 20 and 80 ms. In contrast to [90], no onset suppression was applied to ensure a fair comparison between the different decorrelation methods. As in [90], the output signals of the individual decorrelators were used directly to form the decorrelated multi-channel output signal.

Note that a comparison to the MPEG-I decorrelator [93], which was considered for the evaluation of the single-channel method discussed in Section 4.4, could not be performed since only a single-channel variant is currently available.

5.4.2 Independent Channel Optimization

In addition to the two classical signal processing-based comparison methods described in Section 5.4.1, an independent channel optimization alternative to the proposed method was considered. In this method, the C generator networks for the different output channels are optimized independently, using different seeds for their parameter initialization. Consequently, the loss function for each individual generator network is defined as described in Section 5.2 for the special case of a single output channel.

5.4.3 Objective Evaluation

5.4.3.1 Influence of Coherence Loss Weight

First, we evaluated the influence of the coherence loss weight λ_{coh} on the properties of the decorrelated output signal. To this end, four separate neural networks were trained with $\lambda_{\text{coh}} \in \{0.75, 1.5, 2.5, 4.0\}$. Additionally, a comparison was performed to the two classical signal processing-based methods described in Section 5.4.1. The number of output channels was set to $C = 4$, and all networks were trained on the music dataset.

As a first metric, we evaluated the frequency-dependent absolute coherence. Given two STFT-domain signals $X(k, l)$ and $Y(k, l)$, the frequency-dependent absolute coherence $\rho_{xy}(k)$ between both is defined as follows:

$$\rho_{xy}(k) = \frac{|\sum_l X(k, l)Y^*(k, l)|}{\sqrt{\sum_l |X(k, l)|^2 \sum_l |Y(k, l)|^2}}. \quad (5.11)$$

As for the generator model (see Section 5.2.1), an STFT frame length of 116 samples and a hop size of 58 samples were used. The coherence was calculated between each individual output channel and the input signal, as well as between all output channel pairs. Two separate frequency-dependent coherence values were then determined: averaged over all input-output pairs and averaged over all output channel pairs.

Figure 5.2 shows the mean and standard deviation of the absolute coherence over frequency evaluated on the music test dataset. For all values of λ_{coh} , the input-output coherence values are fairly similar to the inter-channel coherence values. This can be expected since all relevant signal pairs are treated equally in the multi-channel coherence loss definition, see (5.3). Naturally, the absolute coherence is

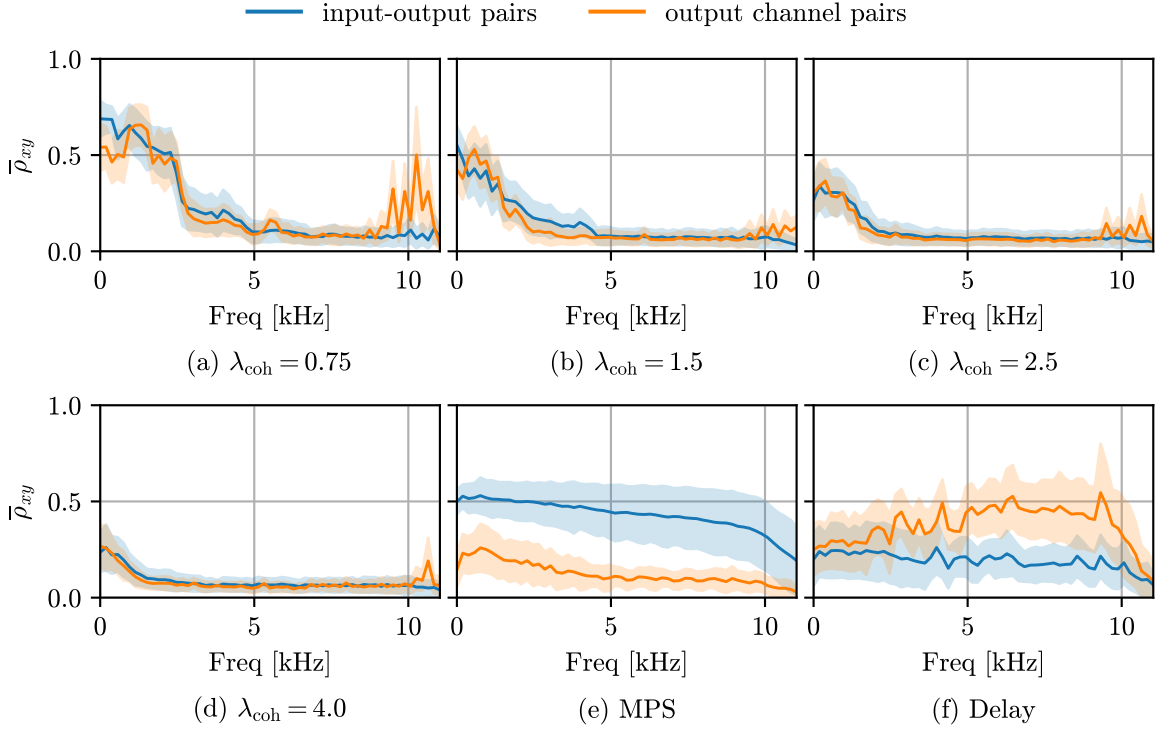


Figure 5.2: Mean and standard deviation of absolute coherence over frequency evaluated on the music test dataset. The coherence values were averaged over all input-output pairs resp. all output channel pairs, with $C = 4$.

high for small values of λ_{coh} . Furthermore, larger values are obtained at lower frequencies compared to higher frequencies. One reason for this may be that most of the music items used for training exhibit most energy in the lower frequency range. While the coherence and the mel-spectrogram loss do not depend on the signal energy due to their normalization, the adversarial loss will primarily focus on high-energy regions. This may cause an increase in coherence in the lower frequency range. For the MPS decorrelator, the input-output coherence values are considerably larger than the inter-channel coherence values. This can be expected since the decorrelated multi-channel output signal is obtained by mixing the outputs of the individual decorrelators with the input signal according to (5.10). For the delay decorrelator, the inter-channel coherence values are considerably larger than the input-output coherence values, especially at higher frequencies. Considering the inter-channel coherence values, which are most relevant for direct loudspeaker reproduction of the multi-channel output signals, the performance of the proposed method is most similar to the MPS and delay decorrelators for a coherence loss weight of $\lambda_{\text{coh}} = 2.5$, at least in the low-frequency range.

Furthermore, we evaluated the (perceptual) similarity between the individual output channels and the input signal. Therefore, we considered two different measures: the mel-spectrogram loss, calculated as defined in (5.5), to assess spectral differences between both, and the perceptual evaluation of audio quality (PEAQ) method [143]. The output of the PEAQ method is the objective difference grade (ODG), which aims to predict the subjective difference grade (SDG) between a reference signal and a degraded version of the reference signal. The SDG takes values between 0 and -4 , where 0 corresponds to an imperceptible impairment, and -4 corresponds to an impairment judged as very annoying. Note

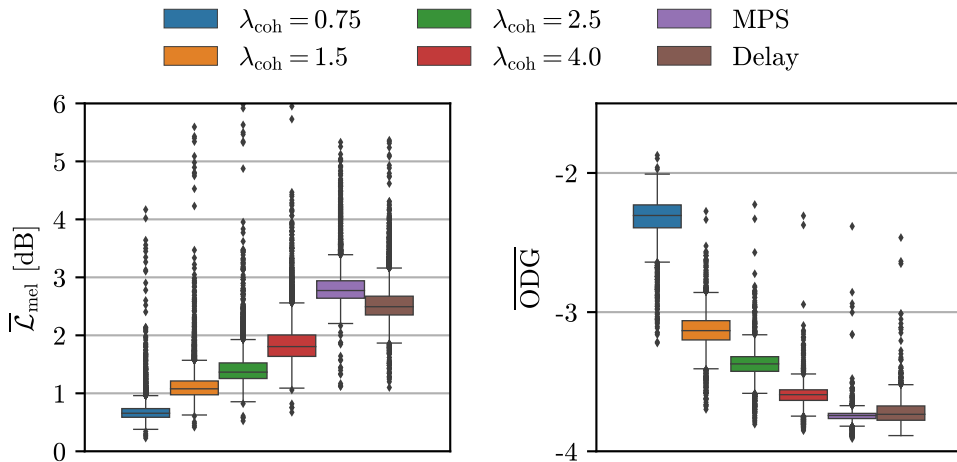


Figure 5.3: Boxplot of mel-spectrogram loss and ODG averaged over all output channels for the music test dataset, with $C = 4$. The horizontal lines represent the median values, and the boxes represent the first to third quartiles. The whiskers are within 1.5 times the inter-quartile range; outliers are depicted by diamonds.

that PEAQ was originally designed to assess the quality of audio codecs and may only be of limited reliability for the decorrelation methods considered here. Figure 5.3 depicts boxplots of both the mel-spectrogram loss and the ODG, evaluated on the music test dataset and averaged over all output channels. The ODG values were determined using an open-source implementation of the PEAQ basic model [144]. In terms of both the mel-spectrogram loss and the ODG, the performance improves for decreasing values of λ_{coh} . The proposed method furthermore outperforms the MPS and delay decorrelators both in terms of the mel-spectrogram loss and the ODG for all values of λ_{coh} . As an alternative to PEAQ, we considered ViSQOLAudio [145] as a perceptually motivated similarity metric. The general trend between the different decorrelation methods for ViSQOLAudio was found to be highly similar to that of PEAQ; therefore, a plot of the results was not included here.

In addition to an independent evaluation of the proposed method’s coherence properties and spectral differences, Figure 5.4 shows a scatter plot between the coherence loss and the mel-spectrogram loss to assess their interaction. The coherence loss was determined as defined in (5.2) and averaged over all signal pairs considered in (5.3). The mel-spectrogram loss was once more determined as defined in (5.5) and averaged over all output channels. The scatter plot shows a clear trend of increase in mel-spectrogram loss when the coherence loss decreases. Despite the fact that only the coherence loss is explicitly controlled during training by adjusting the coherence loss weight λ_{coh} , the resulting mel-spectrogram loss changes inherently. This indicates that there exists a certain tradeoff between the degree of decorrelation and the amount of spectral magnitude differences between the individual output channels and the input signal for the proposed method.

5.4.3.2 Influence of Number of Output Channels

In the second experiment, we evaluated the influence of the number of output channels on the properties of the decorrelated output signal. Five separate neural networks were trained, varying the number of output channels $C \in \{1, 2, 3, 4, 5\}$. The networks were trained on the music dataset, using a coherence loss weight of $\lambda_{\text{coh}} = 2.5$.

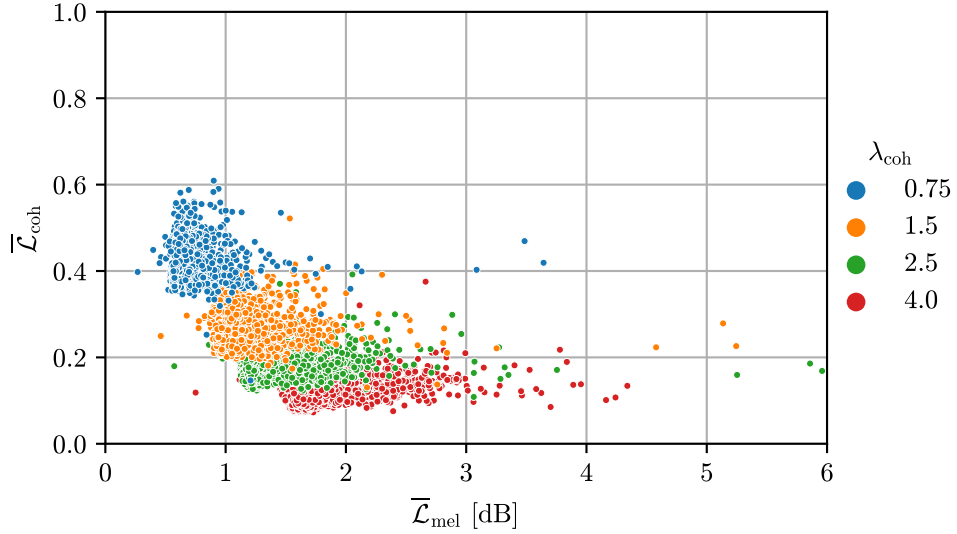


Figure 5.4: Scatter plot of average coherence loss over average mel-spectrogram loss evaluated on the music test dataset for the proposed method trained using different values of λ_{coh} , with $C = 4$.

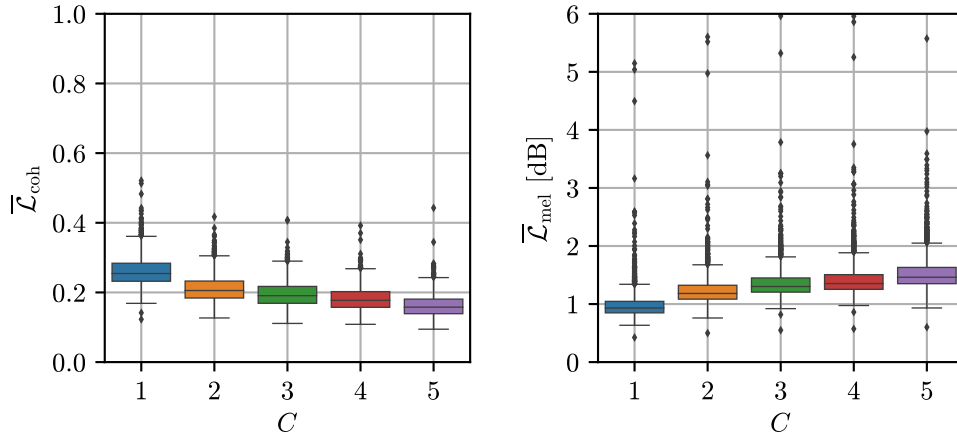


Figure 5.5: Boxplot of average coherence loss and average mel-spectrogram loss evaluated on the music test dataset for the proposed method trained for different number of output channels C , with $\lambda_{\text{coh}} = 2.5$.

Figure 5.5 shows boxplots of the coherence loss and the mel-spectrogram loss, again determined as defined in (5.2) and (5.5), respectively, both averaged over all relevant signal pairs. The coherence loss exhibits a negative correlation with the number of output channels, whereas the mel-spectrogram loss exhibits a positive correlation. To ensure that the decorrelated output signal exhibits comparable properties in terms of decorrelation and spectral magnitude differences when the number of output channels is varied, it is necessary to adjust the coherence loss weight λ_{coh} accordingly.

5.4.3.3 Independent Channel Optimization

In the third experiment, we considered the independent channel optimization method described in Section 5.4.2. The total number of output channels was set to $C = 4$. Each generator network was trained on the music dataset using a coherence loss weight of $\lambda_{\text{coh}} = 4.0$. As demonstrated in Section 5.4.3.2, a coherence loss weight of $\lambda_{\text{coh}} = 2.5$ resulted in slightly higher coherence values for

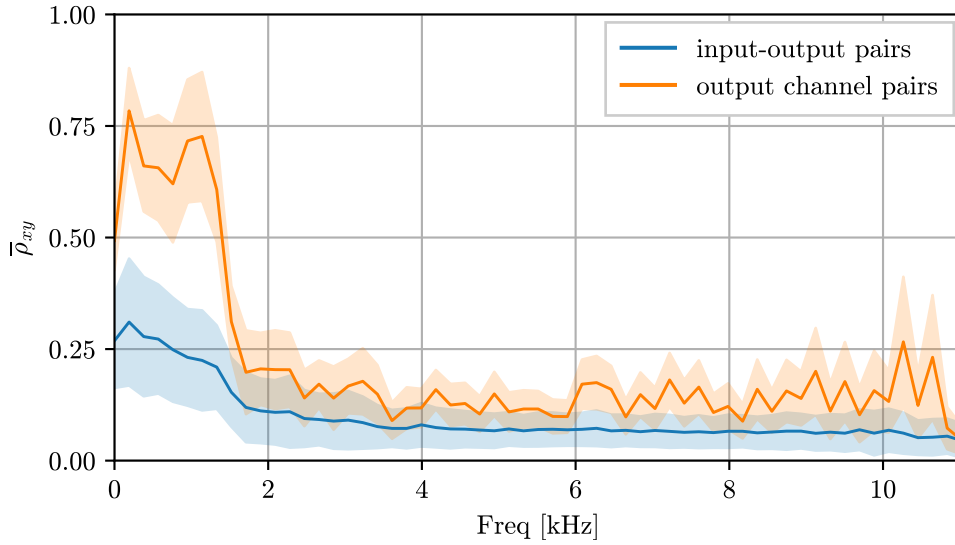


Figure 5.6: Mean and standard deviation of absolute coherence over frequency evaluated on the music test dataset for independent channel optimization, with $C = 4$ and $\lambda_{\text{coh}} = 4.0$. The coherence values were averaged over all input-output pairs resp. all output channel pairs.

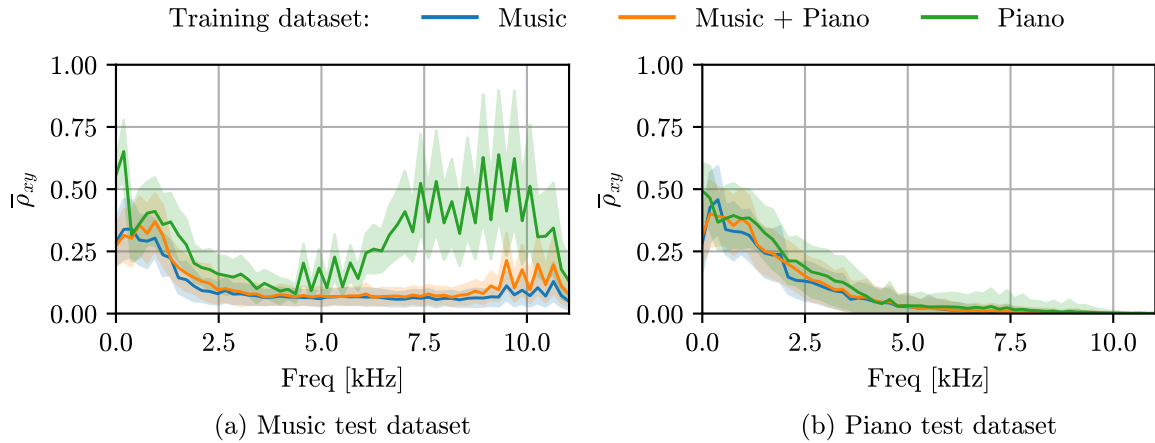


Figure 5.7: Mean and standard deviation of absolute coherence over frequency for the proposed method using different datasets for training and evaluation, with $C = 4$ and $\lambda_{\text{coh}} = 2.5$. The coherence values were averaged over all relevant signal pairs.

$C = 1$ than for $C = 4$. Consequently, a larger value for λ_{coh} was chosen here.

Figure 5.6 shows the mean and standard deviation of the absolute coherence over frequency evaluated on the music test dataset, averaged over all input-output pairs and all output channel pairs, respectively. The absolute coherence for each signal pair was again determined as defined in (5.11). The absolute coherence between the output channel pairs is considerably higher than the absolute coherence between the input-output pairs, particularly in the lower frequency range. This is to be expected, given that the coherence between the output channel pairs is not explicitly minimized during training. These results indicate that joint optimization of the generator networks is beneficial for controlling not only the input-output coherence but also the inter-channel coherence.

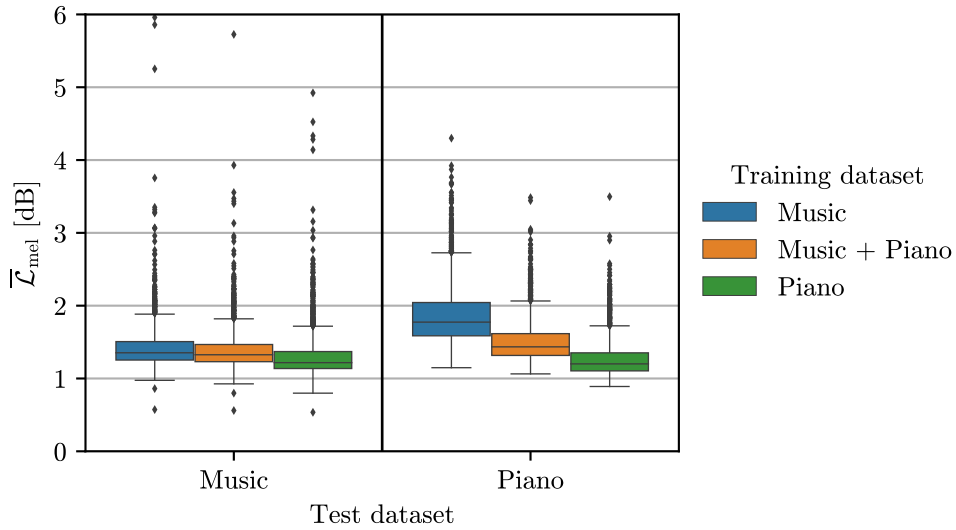


Figure 5.8: Boxplot of mel-spectrogram loss averaged over all output channels for the proposed method using different datasets for training and evaluation, with $C = 4$ and $\lambda_{\text{coh}} = 2.5$.

5.4.3.4 Influence of Training Dataset

Finally, we evaluated the influence of the employed training dataset on the resulting output signal properties using the example of piano signals. Therefore, the neural network was trained on three different datasets: the music dataset, the combined music/piano dataset, and the piano dataset. The number of output channels was set to $C = 4$, and a coherence loss weight of $\lambda_{\text{coh}} = 2.5$ was used.

Figure 5.7 depicts the mean and standard deviation of the absolute coherence over frequency, averaged over all signal pairs considered in (5.3), for all three training datasets evaluated on both the music and the piano test datasets. The absolute coherence for each signal pair was again determined as defined in (5.11). For the networks trained on the music and the combined music/piano datasets, the obtained coherence values are comparable for both the music and the piano test datasets. However, for the network trained on the piano dataset, considerably larger coherence values are obtained for the music test dataset, especially in the high-frequency range. This discrepancy may be attributed to the lack of high-frequency content in the piano training dataset.

Figure 5.8 shows boxplots of the mel-spectrogram loss for all three training datasets, evaluated on both the music and the piano test datasets. The mel-spectrogram loss values were again determined as defined in (5.5) and averaged over all output channels. For the music test dataset, only minor differences in the resulting mel-spectrogram loss can be observed for the different training datasets. The resulting mel-spectrogram loss for both the combined music/piano training dataset and the piano training dataset is slightly lower than that for the music training dataset. However, this is accompanied by an increase in coherence, as can be observed in Figure 5.7. For the piano test dataset, the differences in the resulting mel-spectrogram loss are more pronounced. The inclusion of the piano dataset during training clearly reduces the resulting mel-spectrogram loss for the piano test dataset.

5.4.4 Perceptual Evaluation

Two formal listening tests were conducted to assess the perceptual performance of the proposed method, evaluating a subset of the networks considered in the objective evaluation discussed in Section 5.4.3.

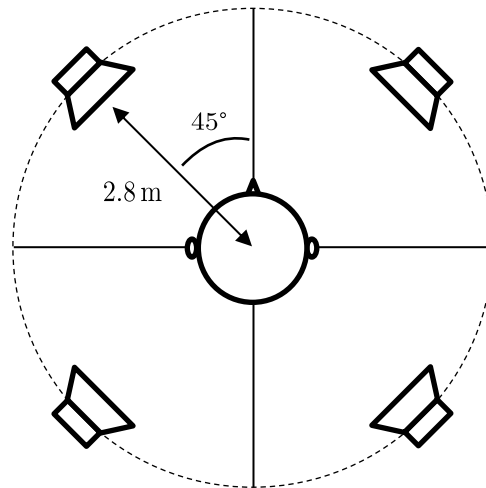


Figure 5.9: Four-channel loudspeaker setup used for the listening experiments.

First, the general performance was evaluated by comparing the proposed method to the two classical signal processing-based methods described in Section 5.4.1, which were also considered in the objective evaluation (Section 5.4.3.1). Furthermore, the independent channel optimization method, as described in Section 5.4.2 and objectively evaluated in Section 5.4.3.3, was included. In a second listening test, the influence of the employed training dataset on the proposed method’s performance was assessed for an otherwise fixed configuration. As in Section 5.4.3.4 of the objective evaluation, this was done using the example of piano signals. Both listening tests were split into two parts related to the two requirements specified in Section 5.1: an envelopment and a quality subtest. Note that the perceived envelopment is closely related to the degree of inter-channel correlation of the multi-channel output signals [54].

The perceptual evaluation was limited to $C = 4$ output channels. The decorrelated multi-channel output signals were reproduced over a four-channel loudspeaker setup, with loudspeakers positioned at ear height at $\pm 45^\circ$ and $\pm 135^\circ$ azimuth on a circle with a radius of 2.8 m. A visual representation of the employed loudspeaker setup is provided in Figure 5.9. The listening tests were conducted in the Mozart listening test room at Fraunhofer IIS [146]. A total of 11 subjects participated in each listening test, eight male and three female. The average age of the participants was 29 years, and all of them had prior listening test experience. The listeners were sitting in the center/sweet spot but were allowed to rotate and move their heads as long as they remained seated.

Note that the direct loudspeaker reproduction considered in the conducted listening experiments concerns only one specific use case of the decorrelated multi-channel output signals. Depending on the target application, the decorrelated output signals may be employed differently, and the results discussed in this section may not be directly transferable.

5.4.4.1 General Performance Evaluation

Listening test setup Two variants of the proposed method were included in the general performance evaluation listening test, using a coherence loss weight of $\lambda_{\text{coh}} = 0.75$ and $\lambda_{\text{coh}} = 2.5$, respectively. Both networks were trained on the music dataset and are identical to the respective networks considered in Section 5.4.3.1 of the objective evaluation. The independent channel optimization method (“ICO”)

described in Section 5.4.2 was also included in the listening test. As in Section 5.4.3.3, the individual networks were trained on the music dataset using a coherence loss weight of $\lambda_{\text{coh}} = 4.0$. Additionally, both the MPS and the delay decorrelators described in Section 5.4.1 were included for comparison.

The listening test comprised a total of seven test items. The items were partly taken from the FSD50K dataset [127] and the EBU SQAM CD [128]. Four music stimuli were incorporated: two pop songs, one particularly transient music item (castanets), and one particularly tonal music item (piano). Additionally, one applause, one speech, and one ocean waves item were included.

For the envelopment subtest, no ideal reference could be defined. The original input signal, reproduced over the center speaker (0° azimuth), was provided as lower anchor. As a result, the envelopment subtest included a total of six conditions. The participants were asked to rate the presented conditions in terms of perceived envelopment on a continuous scale ranging from 0 (lowest possible envelopment, i.e., the sound field is perceived as a point source) to 100 (maximum possible envelopment, i.e., the sound field is perceived as if it fully surrounds the listener, with no localization of individual sound sources possible). The listeners were instructed to focus exclusively on the perceived envelopment, disregarding any differences related to the audio quality. To ensure a common understanding of the term envelopment, the listeners were presented with two examples prior to conducting the listening test. The first example was a single-channel white noise stimulus reproduced over the center speaker, which corresponds to the lowest possible envelopment. The second example was an incoherent four-channel white noise stimulus, which corresponds to the maximum possible envelopment.

For the quality subtest, the original input signal reproduced over the center speaker was provided as reference. Additionally, a hidden reference was included, and a 3.5 kHz low-pass filtered version of the original input signal reproduced over the center speaker was provided as lower anchor. The quality subtest thus included a total of seven conditions. The participants were asked to rate the presented conditions in terms of overall audio quality w.r.t. the reference on a continuous scale ranging from 0 (bad) to 100 (excellent). The listeners were instructed to focus exclusively on the audio quality and to disregard any differences related to the perceived envelopment.

Results Figures 5.10 and 5.11 show the mean values and bootstrapped 95% confidence intervals (CIs) of the listening test scores for the envelopment and quality subtests, respectively, for all items and conditions under test. In addition to assessing the scores for each item individually, the mean values and bootstrapped 95% CIs aggregated over all items are evaluated.

Considering the listening test scores aggregated over all items, the proposed method with $\lambda_{\text{coh}} = 2.5$ outperforms the proposed method with $\lambda_{\text{coh}} = 0.75$ in terms of perceived envelopment. However, in terms of overall audio quality, the proposed method with $\lambda_{\text{coh}} = 0.75$ outperforms the proposed method with $\lambda_{\text{coh}} = 2.5$. This finding aligns with the objective evaluation results discussed in Section 5.4.3.1, which demonstrated that for the proposed method, a certain tradeoff exists between the degree of decorrelation and the amount of spectral magnitude differences between the individual output channels and the input signal. The ICO condition exhibits a comparable performance to the $\lambda_{\text{coh}} = 0.75$ condition in terms of perceived envelopment, while it performs significantly worse in terms of overall audio quality. In comparison to the $\lambda_{\text{coh}} = 2.5$ condition, the ICO condition demonstrates a similar performance in terms of overall audio quality, while it performs significantly worse in terms of perceived envelopment. It can thus be concluded that joint optimization of the generator networks is necessary to achieve the optimal tradeoff between perceived envelopment and overall audio quality. This finding

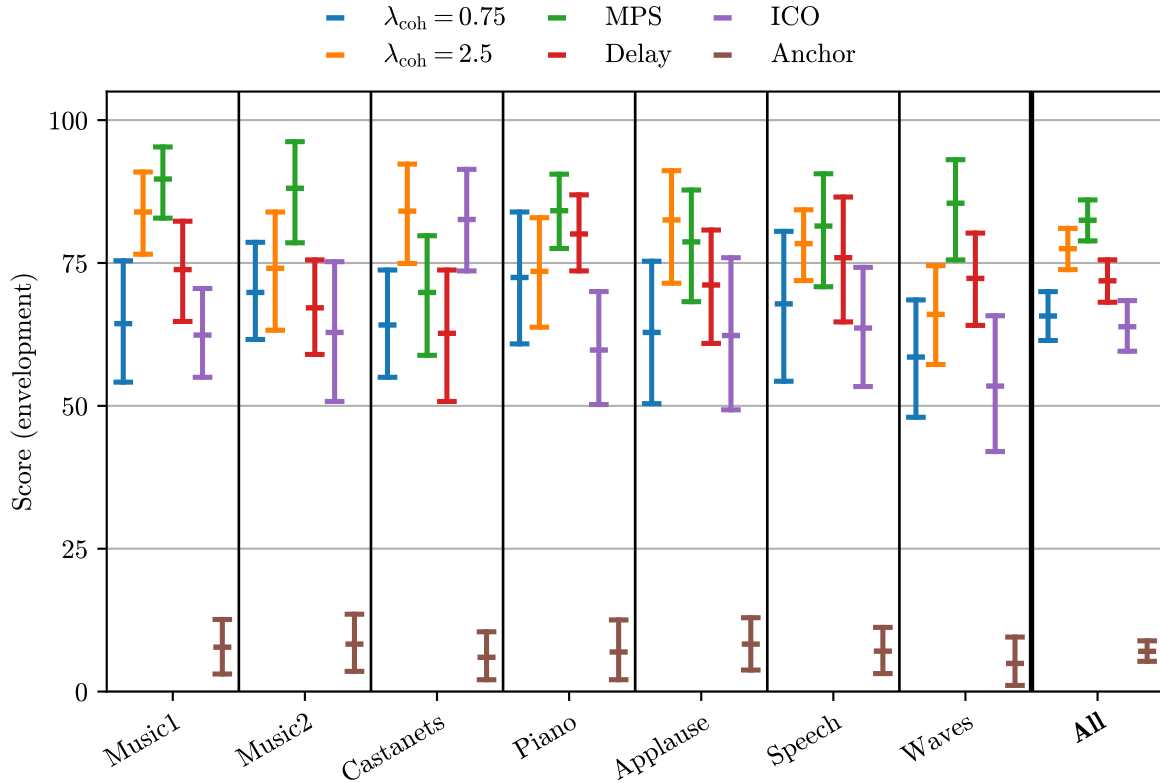


Figure 5.10: Mean values and bootstrapped 95% CIs of general performance evaluation listening test scores per item as well as aggregated over all items in terms of perceived envelopment.

aligns with the objective evaluation results discussed in Section 5.4.3.3. Overall, the MPS decorrelator and the $\lambda_{\text{coh}} = 0.75$ condition demonstrate the best performance in terms of overall audio quality. As previously discussed in Section 5.4.3.1, the MPS decorrelator performed worst in terms of \mathcal{L}_{mel} and ODG. This indicates that the two considered objective metrics are insufficient for predicting the overall audio quality of multi-channel decorrelation methods in general. One reason for this discrepancy may be that the objective metrics are determined by comparing each output channel to the input signal individually, whereas during the listening test, all output channels are played back and rated jointly. Considering both the perceived envelopment and the overall audio quality, the $\lambda_{\text{coh}} = 2.5$ condition reaches a comparable performance as the delay decorrelator. However, the proposed method fails to reach the performance of the MPS decorrelator. While the $\lambda_{\text{coh}} = 0.75$ condition reaches a comparable performance in terms of overall audio quality, it performs significantly worse in terms of perceived envelopment. For the $\lambda_{\text{coh}} = 2.5$ condition, approximately the opposite holds true.

In terms of perceived envelopment, the performance of the individual conditions is relatively consistent across the different items. In terms of overall audio quality, the performance is considerably more item-dependent. The proposed method with $\lambda_{\text{coh}} = 2.5$ performs particularly poorly for the piano item. Similar results were obtained for the perceptual evaluation of the single-channel variant of the proposed method discussed in Section 4.4.2, which showed room for improvement w.r.t. the mono signal quality of the considered violin item. Both the piano item considered in the present listening experiment and the violin item considered in the listening experiment discussed in Section 4.4.2 can

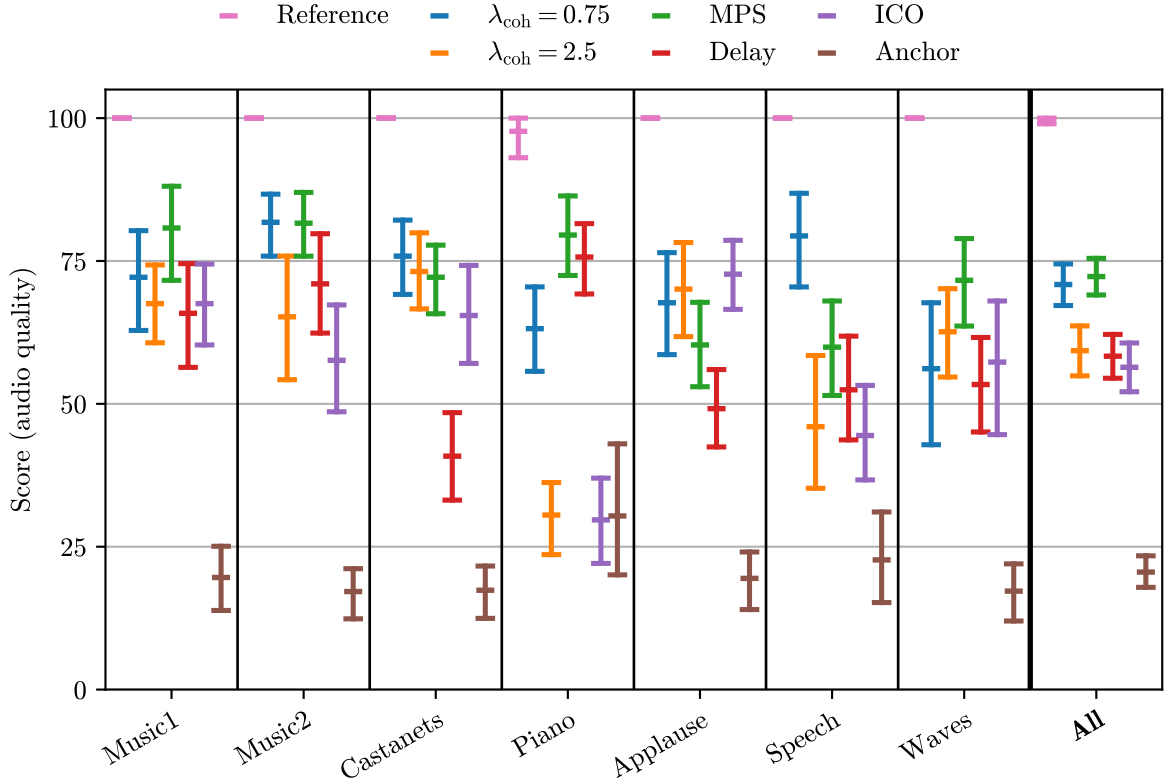


Figure 5.11: Mean values and bootstrapped 95% CIs of general performance evaluation listening test scores per item as well as aggregated over all items in terms of overall audio quality.

be categorized as rather tonal signals.

5.4.4.2 Training Dataset Evaluation

Listening test setup In the second listening test, the influence of the employed training dataset on the perceptual performance of the proposed method was evaluated. Since the general performance evaluation listening test discussed in Section 5.4.4.1 showed room for improvement of the proposed method w.r.t. the overall audio quality for the piano item in particular, we focused on the class of piano signals in this experiment. Three variants of the proposed method were considered in the listening test, each trained on a different dataset: the music dataset, the combined music/piano dataset, and the piano dataset. The coherence loss weight λ_{coh} was set to 2.5 for all networks. The three networks are identical to those considered in Section 5.4.3.4 of the objective evaluation. In both the envelopment and quality subtests, the participants were asked to rate the networks trained on the combined music/piano dataset and the piano dataset, respectively, w.r.t. the network trained on the music dataset. This choice was made because the network trained on the music dataset corresponds to the $\lambda_{\text{coh}} = 2.5$ condition of the general performance evaluation listening test discussed in Section 5.4.4.1.

Four test items were incorporated in the listening test: the two pop songs and the piano item also included in the general performance evaluation listening test, as well as an additional piano item from the piano test dataset.

In the envelopment subtest, the participants were asked to rate the two conditions w.r.t. the model

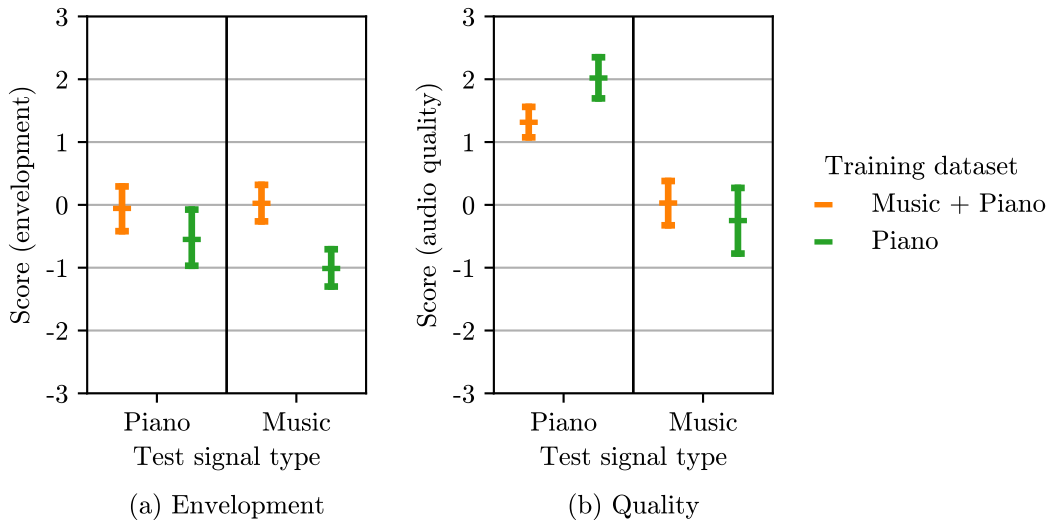


Figure 5.12: Mean values and bootstrapped 95% CIs of training dataset evaluation listening test scores, grouped by signal type, in terms of perceived envelopment and overall audio quality, respectively. Both conditions were rated relative to the model trained on the music dataset.

trained on the music dataset in terms of perceived envelopment on a continuous scale ranging from -3 (“much less enveloped”) to 3 (“much more enveloped”), using the following intermediate anchor points: -2 (“less enveloped”), -1 (“slightly less enveloped”), 0 (“the same as”), 1 (“slightly more enveloped”), 2 (“more enveloped”). The grading scale had a resolution of one decimal place. As in the general performance evaluation envelopment subtest, the listeners were instructed to focus exclusively on the perceived envelopment and to disregard any differences related to the audio quality.

In the quality subtest, the participants were asked to rate the two conditions w.r.t. the model trained on the music dataset in terms of overall audio quality on a continuous scale ranging from -3 (“much worse”) to 3 (“much better”), using the following intermediate anchor points: -2 (“worse”), -1 (“slightly worse”), 0 (“the same as”), 1 (“slightly better”), 2 (“better”). The grading scale again had a resolution of one decimal place. As in the general performance evaluation quality subtest, the listeners were instructed to focus exclusively on the audio quality and to disregard any differences related to the perceived envelopment.

Results Figure 5.12 depicts the mean values and bootstrapped 95% CIs of the listening test scores for the envelopment and quality subtests, for all conditions under test grouped by signal type.

In terms of perceived envelopment, the network trained on the combined music/piano dataset does not exhibit significant differences from the network trained only on the music dataset, for both the piano and the music test items. However, the network trained on the piano dataset only is perceived as significantly less enveloped for both signal types. This finding aligns with the objective evaluation results discussed in Section 5.4.3.4, which demonstrated that the resulting absolute coherence is considerably higher for the network trained exclusively on the piano dataset.

In terms of overall audio quality, the networks trained on the combined music/piano dataset and the piano dataset both show a significant improvement for the piano test items. For the music test items, neither network shows significant differences w.r.t. the network trained only on the music dataset. These findings are consistent with the objective evaluation results discussed in Section 5.4.3.4, which demonstrated that the networks trained on the combined music/piano dataset and the piano dataset

outperform the network trained on the music dataset in terms of spectral magnitude reconstruction, particularly for the piano test dataset.

In summary, incorporating the piano dataset during training enhances the overall audio quality of the piano items while maintaining comparable performance for the music items both in terms of overall audio quality and perceived envelopment. While the overall audio quality of the piano items can be improved further by training exclusively on the piano dataset, this comes at the cost of a reduced perceived envelopment for both the piano and the music test items.

5.4.5 Discussion

Based on the objective and perceptual evaluation results discussed in Sections 5.4.3 and 5.4.4, a number of conclusions can be drawn.

The coherence loss weight significantly influences the properties of the proposed method’s multi-channel output signal. The perceptual evaluation showed that a higher coherence loss weight results in an improved performance in terms of perceived envelopment, but also leads to a performance decrease in terms of overall audio quality. Consequently, the coherence loss weight allows controlling the existing tradeoff between the degree of decorrelation and the output signal quality of the proposed method’s output.

In comparison to the proposed method, the independent channel optimization alternative was found to result in a lower performance in terms of perceived envelopment when a comparable performance in terms of overall audio quality is reached. It can thus be concluded that joint optimization of the generator networks is essential to achieve the optimal tradeoff between perceived envelopment and overall audio quality.

Furthermore, it was demonstrated that by explicitly including piano data during training, the overall audio quality of the piano items can be improved while maintaining comparable performance for the music items, both in terms of overall audio quality and perceived envelopment. This indicates that the explicit inclusion of specific training data can be beneficial to improve the overall audio quality of the proposed method’s output for specific signal types.

Depending on the configuration, the proposed method reached a comparable performance to the classical signal processing-based decorrelator employed in [90], which is based on frequency-dependent pseudo-random delays, both in terms of perceived envelopment and overall audio quality. However, the proposed method did not succeed in reaching the perceptual performance of the classical signal processing-based decorrelator proposed in [92], which is based on lattice all-pass filters. While the perceptual evaluation results mostly align with the considered objective metrics, the mel-spectrogram loss and the ODG, calculated using PEAQ, were found not to be well suited to predict the overall audio quality of the two classical signal processing-based decorrelation methods relative to the proposed method. This highlights the need for better-suited objective quality metrics, both for optimizing the proposed method and evaluating decorrelation methods in general.

5.5 Summary

In this chapter, we proposed a multi-channel extension of the single-channel GAN-based audio decorrelation method introduced in Chapter 4. A separate generator network is used for each output

channel. All generator networks are optimized jointly based on a number of individual loss terms that aim to control the relevant properties of the decorrelated multi-channel output signal, namely, the inter-channel correlation, the input-output correlation, and the output signal quality.

The proposed model was trained on music signals. A number of experiments were performed to investigate the influence of the coherence loss weight, the number of output channels, and the employed training dataset on the proposed method's performance. Additionally, the proposed method was compared to an independent channel optimization alternative as well as to two classical signal processing-based multi-channel decorrelators. In addition to objective evaluations, two formal listening tests were conducted to assess the perceptual performance of the proposed method, considering reproduction over a four-channel loudspeaker setup.

The objective and perceptual evaluations showed that the coherence loss weight is a key parameter of the proposed method to control the existing tradeoff between the degree of decorrelation and the output signal quality. Furthermore, the perceptual evaluation demonstrated that joint optimization of the generator networks is important to reach the best possible tradeoff between perceived envelopment and overall audio quality. Additionally, it was observed that explicitly including certain training data can be beneficial to improve the overall audio quality of the proposed method's output for specific signal types. While the proposed method was found to reach a comparable performance to the classical signal processing-based decorrelator employed in [90], it did not succeed in reaching the perceptual performance of the classical signal processing-based decorrelator proposed in [92].

The mel-spectrogram loss was found not to be well suited to predict the overall audio quality of decorrelation methods in general. Therefore, the availability of better-suited objective quality metrics is expected to be a key factor in improving the performance of the proposed method. The perceptual evaluation furthermore showed that the performance of the proposed method in terms of overall audio quality is rather signal-dependent. Consequently, adjusting the loss function in a signal-dependent manner may be beneficial to control the discussed tradeoff between decorrelation and output signal quality per signal type. Since a separate generator network is used for each output channel, the model complexity increases linearly with the number of output channels. It is anticipated that the model complexity can be reduced by sharing part of the networks' parameters across the different output channels.

CHAPTER 6

Binaural Rendering of Homogeneously Extended Sound Sources

The main content of this chapter is based on: C. Anemüller, A. Adami, and J. Herre, “Efficient binaural rendering of spatially extended sound sources,” J. Audio Eng. Soc., vol. 71, no. 5, pp. 281–292, May 2023, [19].

In virtual and augmented reality (VR/AR) or 3D applications with binaural audio, it is often desired to render sound sources with a certain spatial extent in a realistic way. Relevant sound sources include, e.g., a grand piano, a choir, or a waterfall, all of which have a certain “size” (i.e., geometric and perceived extent). Spatially extended sound sources (SESSs) can be characterized further by means of their radiation behavior. While homogeneous SESSs emit sound with constant radiation characteristics over the extent, heterogeneous SESSs exhibit a position-dependent radiation behavior. This chapter focuses on rendering of homogeneous SESSs, based on a single-channel input signal characterizing the sound source.

As discussed in Section 2.3, a prominent group of approaches to rendering of homogeneous SESSs is based on distributing a number of incoherent point sources over the desired spatial extent range. Decorrelated versions of the single-channel input signal, which are required for these types of rendering methods, are typically generated by applying suitable decorrelation filters. A notable drawback of these approaches is that the employed decorrelation techniques typically significantly deteriorate the signal quality, as discussed in Section 2.2. To minimize signal artifacts, careful design of the decorrelation filters is important. Additionally, it is beneficial to minimize the amount of decorrelated signal energy introduced. A further challenge concerns the computational complexity, which typically increases with the number of individual point sources and, consequently, depends on the desired spatial extent range. In particular, for interactive applications such as VR and AR, the computational complexity is an essential factor, as the rendering must be adjusted in real time according to the listener’s position and orientation.

In this chapter, we propose a method for efficient and realistic binaural rendering of homogeneous SESSs. Based on the model of an incoherently extended sound source with position-independent energy and spectral content, a number of target auditory cues are determined. Subsequently, a binaural output signal with the desired target auditory cues is synthesized by mixing two decorrelated input signals.

Given a single-channel input signal, the two decorrelated input signals can be generated using a single decorrelation filter. The proposed rendering method depends only on the desired spatial extent range and is thus signal-independent. Compared to direct rendering of a number of decorrelated point sources distributed over the extent range, the proposed method has the advantage of reduced computational complexity and relaxed requirements for the employed decorrelation filters.

Note that in [116]¹, a similar method was proposed. Using the same model of an incoherently extended sound source with position-independent energy and spectral content, the covariance domain framework introduced in [117] is employed to obtain a binaural output signal with the desired inter-channel properties. First, a linear mixing operation is applied to the binaural signal corresponding to point source reproduction at the extent center. Second, if necessary, an additional decorrelated signal is introduced. Compared to the proposed method, two mutually independent decorrelation filters are required instead of one. Furthermore, a larger number of filtering operations are employed. Additionally, an alternative unconstrained method is described in [116], which directly mixes two decorrelated input signals. This method is conceptually closer to the proposed method; however, it was found to exhibit considerable signal distortions. The method proposed in this chapter aims to avoid signal distortions through a perceptually motivated design of the employed processing filters.

The remainder of this chapter is structured as follows. In Section 6.1, a mathematical formulation of the rendering model is provided. In Section 6.2, the proposed rendering method is described, and its performance is evaluated both objectively and perceptually in Section 6.3 by comparing it against a direct implementation of the rendering model. Section 6.4 discusses the application of the proposed method in the context of VR/AR standardization efforts. Finally, Section 6.5 concludes this chapter by providing a summary.

6.1 Rendering Model

In this chapter, we use a model of an incoherently extended sound source with position-independent energy and spectral content to represent a homogeneous SESS. The model is described in the discrete time-frequency domain.

Let $S(k, l, \mathbf{u})$ denote the source signal emitted by the SESS arriving from direction $\mathbf{u} = [\varphi, \theta]^T$ at a receiver, where φ describes the incident azimuth angle, θ the incident elevation angle, and k, l the discrete frequency and time frame indices, respectively. The operator $(\cdot)^T$ denotes the transpose operation. Since incoherent sound radiation for each direction is assumed, the following condition holds for two arbitrary directions \mathbf{u}_i and \mathbf{u}_j :

$$\mathcal{E}\{S(k, l, \mathbf{u}_i)S^*(k, l, \mathbf{u}_j)\} = \begin{cases} 0, & \mathbf{u}_i \neq \mathbf{u}_j \\ P(k, l, \mathbf{u}_i), & \mathbf{u}_i = \mathbf{u}_j, \end{cases} \quad (6.1)$$

where $P(k, l, \mathbf{u})$ denotes the incident power spectral density (PSD) for direction \mathbf{u} , $\mathcal{E}\{\cdot\}$ the statistical expectation, and $(\cdot)^*$ the complex conjugate. Generally, $P(k, l, \mathbf{u})$ reflects both the energy of the SESS and the distance attenuation, which depends on the geometry of the SESS and its relative position to the receiver. Here, we consider the more general case in which no specific information about the

¹This work was published subsequent to the completion of the performance evaluation presented in this chapter. Consequently, a direct comparison between the proposed method and [116] was not conducted.

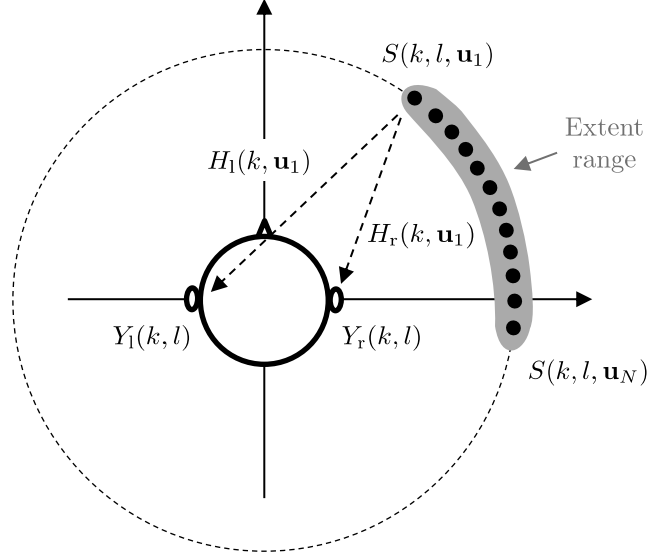


Figure 6.1: Schematic overview of binaural SESS rendering model. For visualization purposes, the special case of horizontal spatial extent only is considered.

geometry of the SESS is available. We, therefore, assume that the distance from the receiver to the SESS is direction-independent and do not model the distance attenuation. Since the energy of the SESS is also assumed to be direction-independent, $P(k, l, \mathbf{u}) = P(k, l)$ holds.

In this chapter, binaural rendering of the SESS is considered. Given a set of head-related transfer functions (HRTFs) $\mathbf{h}(k, \mathbf{u}_n) = [H_1(k, \mathbf{u}_n), H_r(k, \mathbf{u}_n)]^T$ for a number of discrete and uniformly spaced directions $\mathbf{u}_n, n = 1, \dots, N$, covering the extent of the SESS, the binaural output signal $\mathbf{y}(k, l) = [Y_l(k, l), Y_r(k, l)]^T$ can be determined:

$$\mathbf{y}(k, l) = \sqrt{\frac{1}{N}} \sum_{n=1}^N \mathbf{h}(k, \mathbf{u}_n) S(k, l, \mathbf{u}_n). \quad (6.2)$$

In Figure 6.1, a schematic overview of the binaural rendering model described by (6.2) is provided. For visualization purposes, the special case of horizontal spatial extent only is considered. Using (6.1) and (6.2), the output covariance matrix $\mathbf{C}_y(k, l)$ follows:

$$\begin{aligned} \mathbf{C}_y(k, l) &= \mathcal{E}\{\mathbf{y}(k, l)\mathbf{y}^H(k, l)\} \\ &= P(k, l) \cdot \frac{1}{N} \sum_{n=1}^N \mathbf{h}(k, \mathbf{u}_n)\mathbf{h}^H(k, \mathbf{u}_n) \\ &= P(k, l) \cdot \frac{1}{N} \sum_{n=1}^N \begin{bmatrix} |H_1(k, \mathbf{u}_n)|^2 & H_1(k, \mathbf{u}_n)H_r^*(k, \mathbf{u}_n) \\ H_r(k, \mathbf{u}_n)H_1^*(k, \mathbf{u}_n) & |H_r(k, \mathbf{u}_n)|^2 \end{bmatrix} \\ &= \begin{bmatrix} C_{y,ll}(k, l) & C_{y,lr}(k, l) \\ C_{y,rl}(k, l) & C_{y,rr}(k, l) \end{bmatrix}, \end{aligned} \quad (6.3)$$

where $(\cdot)^H$ denotes the Hermitian transpose, and $|\cdot|$ the magnitude of a complex number.

Based on the output covariance matrix $\mathbf{C}_y(k, l)$, a number of relevant auditory cues can be determined. In particular, the interaural coherence (IC), the interaural phase difference (IPD), and the left

and right ear gains are considered. The interaural coherence $IC(k)$ is essentially a frequency-domain representation of the interaural cross-correlation (ICC) and can be determined as follows using (6.3):

$$\begin{aligned} IC(k) &= \frac{C_{y,lr}(k, l)}{\sqrt{C_{y,ll}(k, l)C_{y,rr}(k, l)}} \\ &= \frac{\sum_{n=1}^N H_l(k, \mathbf{u}_n)H_r^*(k, \mathbf{u}_n)}{\sqrt{\sum_{n=1}^N |H_l(k, \mathbf{u}_n)|^2 \sum_{n=1}^N |H_r(k, \mathbf{u}_n)|^2}}. \end{aligned} \quad (6.4)$$

By calculating the phase of $IC(k)$, the interaural phase difference $IPD(k)$ is obtained, using the fact that $P(k, l)$ has zero phase:

$$\begin{aligned} IPD(k) &= \angle(C_{y,lr}(k, l)) = \angle(IC(k)) \\ &= \angle\left(\sum_{n=1}^N H_l(k, \mathbf{u}_n)H_r^*(k, \mathbf{u}_n)\right), \end{aligned} \quad (6.5)$$

where $\angle(\cdot)$ denotes the phase of a complex number. Finally, the left and right ear gains, $G_l(k)$ and $G_r(k)$, are determined, again using (6.3):

$$\begin{aligned} G_l(k) &= \sqrt{\frac{C_{y,ll}(k, l)}{P(k, l)}} = \sqrt{\frac{\sum_{n=1}^N |H_l(k, \mathbf{u}_n)|^2}{N}}, \\ G_r(k) &= \sqrt{\frac{C_{y,rr}(k, l)}{P(k, l)}} = \sqrt{\frac{\sum_{n=1}^N |H_r(k, \mathbf{u}_n)|^2}{N}}. \end{aligned} \quad (6.6)$$

6.2 Proposed Rendering Method

Figure 6.2 shows a block diagram of the proposed homogeneous SESS rendering method, considering a time-frequency-domain representation. Input to the algorithm is a single-channel signal $X(k, l)$, from which a two-channel decorrelated input signal $\mathbf{x}_d(k, l) = [X_{d,1}(k, l), X_{d,2}(k, l)]^T$ is derived using decorrelation techniques. For a given spatial extent range, a binaural output signal $\mathbf{y}_p(k, l)$ with the desired target auditory cues according to the rendering model discussed in Section 6.1 is determined by processing $\mathbf{x}_d(k, l)$.

During rendering, the following binaural cues are considered: the IC, the IPD, and the interaural level difference (ILD). Besides that, monaural spectral cues are reproduced. As discussed in Section 2.1.4, monaural spectral cues are mainly important for sound source localization in the vertical plane and for correct sound timbre. While the IPD and ILD are mainly important for horizontal localization (see Section 2.1.3), the ICC, and consequently the IC, is known to be a crucial cue for spatial extent perception in the horizontal plane (see Section 2.1.6).

The desired spatial extent is described in terms of azimuth and elevation angle ranges. The desired azimuth angle range is given by $[\varphi_1, \varphi_2] = \left[\bar{\varphi} - \frac{\Delta\varphi}{2}, \bar{\varphi} + \frac{\Delta\varphi}{2}\right]$, the desired elevation angle range by $[\theta_1, \theta_2] = \left[\bar{\theta} - \frac{\Delta\theta}{2}, \bar{\theta} + \frac{\Delta\theta}{2}\right]$, where $\bar{\varphi}$ and $\bar{\theta}$ describe the center of the SESS, and $\Delta\varphi$ and $\Delta\theta$ describe the extent of the SESS in terms of azimuth and elevation angle, respectively. Note that, in general, every other representation describing a two-dimensional spatial extent would be suitable as well.

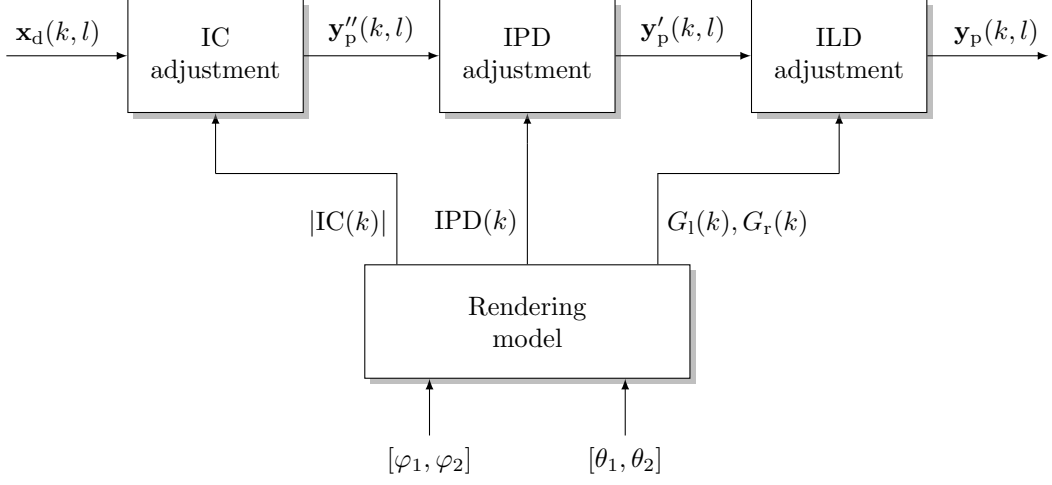


Figure 6.2: Block diagram of proposed homogeneous SESS rendering method.

6.2.1 Processing Steps

Given the single-channel input signal $X(k, l)$, the two channels of the decorrelated input signal $\mathbf{x}_d(k, l) = [X_{d,1}(k, l), X_{d,2}(k, l)]^T$ can be determined using a single decorrelation filter as discussed in Appendix A.1:

$$\begin{aligned} X_{d,1}(k, l) &= \frac{1}{\sqrt{2}} (X(k, l) + \text{dec}\{X(k, l)\}), \\ X_{d,2}(k, l) &= \frac{1}{\sqrt{2}} (X(k, l) - \text{dec}\{X(k, l)\}), \end{aligned} \quad (6.7)$$

where $\text{dec}\{\cdot\}$ describes a decorrelation operation. Under the assumption of an ideal decorrelation operation, $X(k, l)$ and $\text{dec}\{X(k, l)\}$ are uncorrelated and do have the same PSD. Therefore, the following equations hold:

$$\mathcal{E}\{\text{dec}\{X(k, l)\}X^*(k, l)\} = 0, \quad (6.8)$$

$$\mathcal{E}\{|\text{dec}\{X(k, l)\}|^2\} = \mathcal{E}\{|X(k, l)|^2\}. \quad (6.9)$$

Given (6.8) and (6.9), $X_{d,1}(k, l)$ and $X_{d,2}(k, l)$ are also uncorrelated and do have the same PSD as $X(k, l)$:

$$\mathcal{E}\{X_{d,1}(k, l)X_{d,2}^*(k, l)\} = 0, \quad (6.10)$$

$$\mathcal{E}\{|X_{d,1}(k, l)|^2\} = \mathcal{E}\{|X_{d,2}(k, l)|^2\} = \mathcal{E}\{|X(k, l)|^2\}. \quad (6.11)$$

Given a desired spatial extent, first, $\text{IC}(k)$, $\text{IPD}(k)$, $G_1(k)$, and $G_r(k)$ are obtained according to (6.4) – (6.6). Given $\mathbf{x}_d(k, l)$, the SESS is then rendered by successively adjusting the IC, the IPD and the ILD to match the corresponding target auditory cues. By this, the auditory properties of an SESS with the desired extent are resembled. The resulting binaural output signal $\mathbf{y}_p(k, l)$ can be played back via headphones to render the SESS.

In the IC adjustment block, a signal $\mathbf{y}_p''(k, l) = [Y_{p,1}''(k, l), Y_{p,2}''(k, l)]^T$ with inter-channel coherence

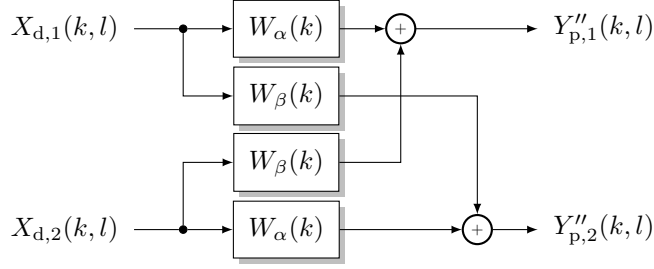


Figure 6.3: Block diagram of IC adjustment.

properties according to the target value $|\text{IC}(k)|$ is obtained by applying a real-valued mixing matrix to $\mathbf{x}_d(k, l)$ [38]:

$$\mathbf{y}_p''(k, l) = \begin{bmatrix} W_\alpha(k) & W_\beta(k) \\ W_\beta(k) & W_\alpha(k) \end{bmatrix} \mathbf{x}_d(k, l), \quad (6.12)$$

with

$$W_\beta(k) = \sqrt{\frac{1}{2} \left(1 - \sqrt{1 - |\text{IC}(k)|^2} \right)}, \quad (6.13)$$

$$W_\alpha(k) = \sqrt{1 - W_\beta^2(k)}. \quad (6.14)$$

Applying (6.12) results in the desired IC, given that (6.10) and (6.11) hold. Since $W_\alpha^2(k) + W_\beta^2(k) = 1$, the mixing process is furthermore energy preserving. The block diagram of the IC adjustment is shown in Figure 6.3.

Generally, an infinite number of real-valued 2×2 mixing matrices exist, that result in an output signal with the desired IC, as discussed in [147, 148]. However, using (6.12) – (6.14) in conjunction with (6.7) results in an output signal with a number of favorable properties. For all values of $|\text{IC}(k)|$, the same amount of decorrelated signal energy is mixed into each output channel. Both channels are thus equally influenced by potential decorrelation artifacts. Furthermore, the larger $|\text{IC}(k)|$, the less decorrelated signal energy is introduced. For the extreme case of $|\text{IC}(k)| = 1$, which is obtained for point source rendering (i.e., $\Delta\varphi = \Delta\theta = 0$), $Y_{p,1}''(k, l) = Y_{p,2}''(k, l) = X(k, l)$ holds. The influence of the decorrelation operation is thus eliminated, and potential decorrelation artifacts are avoided.

Subsequently, IPD adjustment is performed to obtain $\mathbf{y}'_p(k, l)$:

$$\mathbf{y}'_p(k, l) = \begin{bmatrix} e^{j\frac{\text{IPD}(k)}{2}} \\ e^{-j\frac{\text{IPD}(k)}{2}} \end{bmatrix} \odot \mathbf{y}_p''(k, l), \quad (6.15)$$

where \odot denotes the element-wise multiplication. The target IPD is equally distributed to both output channels to minimize potential artifacts introduced by this processing step.

Finally, the ILD adjustment is performed and the binaural output signal $\mathbf{y}_p(k, l)$ is obtained:

$$\mathbf{y}_p(k, l) = \begin{bmatrix} G_l(k) \\ G_r(k) \end{bmatrix} \odot \mathbf{y}'_p(k, l). \quad (6.16)$$

Applying (6.16) results in the desired ILD as long as the two channels of $\mathbf{y}'_p(k, l)$ do have the same PSD. This criterion is fulfilled when (6.10) and (6.11) hold. As the left and right ear gains are used directly, monaural spectral cues (i.e., the desired target spectral envelope) are reproduced correctly in addition to the ILD.

6.2.2 Time-Domain Implementation

Based on the processing steps described in Section 6.2.1, the actual processing can be performed either directly in the time-frequency domain (e.g., using a short-time Fourier transform (STFT)), or in the time domain, by applying an equivalent set of time-domain filters. In this chapter, a time-domain implementation is considered. The corresponding implementation details are discussed in this section.

We now consider a time-domain single-channel input signal $x(n)$, with discrete time index n . From $x(n)$, two decorrelated time-domain input signals $x_{d,1}(n)$ and $x_{d,2}(n)$ are determined analogously to (6.7):

$$\begin{aligned} x_{d,1}(n) &= \frac{1}{\sqrt{2}} (x(n) + \text{dec}\{x(n)\}), \\ x_{d,2}(n) &= \frac{1}{\sqrt{2}} (x(n) - \text{dec}\{x(n)\}). \end{aligned} \quad (6.17)$$

Furthermore, a set of head-related impulse responses (HRIRs), each with a length of q samples, is available. The sampling rate f_s of the input signal and the HRIRs is assumed to be identical. The corresponding HRTFs, needed to calculate the target auditory cues as described in Section 6.1, are determined from the HRIRs by applying a discrete Fourier transform (DFT) of length q . The frequency resolution of the HRTFs and hence also of the calculated target auditory cues thus equals f_s/q .

6.2.2.1 Filter Design

To derive a set of time-domain filters, first, four frequency-domain filters are determined by combining the processing steps described in Section 6.2.1:

$$\mathbf{y}_p(k, l) = \begin{bmatrix} W_{11}(k) & W_{12}(k) \\ W_{r1}(k) & W_{r2}(k) \end{bmatrix} \mathbf{x}_d(k, l), \quad (6.18)$$

with

$$\begin{aligned} W_{11}(k) &= W_\alpha(k) e^{j\frac{\text{IPD}(k)}{2}} G_1(k), \\ W_{12}(k) &= W_\beta(k) e^{j\frac{\text{IPD}(k)}{2}} G_1(k), \\ W_{r1}(k) &= W_\beta(k) e^{-j\frac{\text{IPD}(k)}{2}} G_r(k), \\ W_{r2}(k) &= W_\alpha(k) e^{-j\frac{\text{IPD}(k)}{2}} G_r(k). \end{aligned} \quad (6.19)$$

The frequency-domain filters are used to derive corresponding time-domain finite impulse response (FIR) filters, $w_{11}(n)$, $w_{12}(n)$, $w_{r1}(n)$, and $w_{r2}(n)$, in a manner similar to the FIR filter design procedure

proposed in [149]:

$$w_{ij}(n) = \underset{q/2}{\text{circshift}}(\text{IDFT}\{W_{ij}(k)\}), \quad (6.20)$$

with $i \in \{l, r\}$ and $j \in \{1, 2\}$. First, an inverse discrete Fourier transform (IDFT) is performed. Afterwards, a circular shift of $q/2$ samples is applied, denoted by $\underset{q/2}{\text{circshift}}(\cdot)$, to ensure that the resulting filters are causal. Note that while processing with these filters is causal, they do introduce a delay and may furthermore introduce pre-echoes. The resulting FIR filters have a length of q samples, i.e., the same length as the original HRIRs. The left and right channel output signals, $y_{p,l}(n)$ and $y_{p,r}(n)$, are obtained by convolving the decorrelated input signals with the FIR filters:

$$\begin{aligned} y_{p,l}(n) &= w_{l1}(n) * x_{d,1}(n) + w_{l2}(n) * x_{d,2}(n), \\ y_{p,r}(n) &= w_{r1}(n) * x_{d,1}(n) + w_{r2}(n) * x_{d,2}(n), \end{aligned} \quad (6.21)$$

where the operator $*$ denotes the convolution operation.

6.2.2.2 Phase Smoothing

Following the FIR filter design procedure outlined in (6.20), the magnitude responses of the filters can be precisely controlled solely at the specified frequency bins. At intermediate frequencies, a smooth interpolation behavior is desirable. However, if the filter's phase response exhibits significant changes between adjacent frequency bins, the interpolation behavior of the filter's magnitude response becomes highly non-smooth. Ripples in the filter's magnitude response emerge, resulting in perceptible magnitude distortions of the filtered output signals. These observations are similar to those reported in [4], where FIR all-pass filters with random phase responses were designed for the task of audio decorrelation. This phenomenon is mainly observed at relatively large spatial extents. In such cases, a multitude of incoherent point sources from different directions are superimposed, resulting in a somewhat random phase response. Note that this problem is not unique to the employed time-domain implementation; similar issues arise when directly applying the filters in the time-frequency domain.

In general, different approaches can be pursued to reduce the amount of magnitude distortion. Here, the method employed is to perform a smoothing of $\text{IPD}(k)$ over the frequency axis prior to calculating the frequency-domain filters. Another feasible option would be to directly use the phase response of a set of HRTFs for a single direction that is representative of the considered spatial extent range.

The phase smoothing is realized by applying a moving average filter of length ν , denoted by $\underset{\nu}{\text{movmean}}(\cdot)$, to the unwrapped target IPD:

$$\text{IPD}'(k) = \underset{\nu}{\text{movmean}}(\text{unwrap}\{\text{IPD}(k)\}), \quad (6.22)$$

where $\text{unwrap}\{\cdot\}$ represents the phase unwrapping operation. Subsequently, $\text{IPD}(k)$ is replaced by $\text{IPD}'(k)$ in (6.19).

The length ν of the moving average filter is selected based on the target IC:

$$\nu = \begin{cases} 3, & |\overline{\text{IC}(k)}| \geq 0.1 \\ 5, & |\overline{\text{IC}(k)}| < 0.1, \end{cases} \quad (6.23)$$

where $\bar{\cdot}$ denotes the average over the frequency axis. Given that more severe phase discontinuities occur for large spatial extents and thus at small IC values, a larger ν is applied in order to achieve a greater degree of phase smoothing. Since the IPD is of lesser perceptual importance at lower IC values (i.e., between largely uncorrelated ear signals) [150, 151], it can be expected that applying a higher amount of smoothing will not be problematic. At small spatial extents, for which no or only minimal artifacts arise, the (unwrapped) target IPD is approximately linear, and the moving average filter will thus have only a minimal influence. The employed values for ν were determined empirically, considering the tradeoff between the amount of magnitude distortion and the degree of phase smoothing based on the HRTF dataset used for the performance evaluation described in Section 6.3.1 ($q = 256$ samples, $f_s = 44.1$ kHz). Note that suitable values for ν depend on the filters' frequency resolution and thus on the length q and the sampling rate f_s of the HRIRs.

6.3 Performance Evaluation

The performance of the proposed rendering method was evaluated by comparing its output to a binaural reference signal, which was generated according to the rendering model outlined in Section 6.1. An objective evaluation based on a number of perceptually relevant objective metrics was performed, as well as a perceptual evaluation based on a formal listening test. Additionally, the influence of the employed phase smoothing on the filters' magnitude response was assessed.

Since the main objective of the proposed method is to recreate the perception of the rendering model, ideally, there should not be a significant difference between the proposed method's output and the binaural reference signal. The evaluation thus aimed at verifying the proposed method against the rendering model rather than verifying the rendering model itself in an absolute sense. Given that the perception of vertical spatial extent is rather limited, as discussed in Section 2.3.1.4, and that there is no difference in treatment between horizontal and vertical extent, the evaluation focused mainly on the practically more relevant case of horizontal extent.

6.3.1 Evaluation Setup

The binaural reference signal was generated by distributing a number of incoherent point sources over the considered spatial extent range and filtering them with the corresponding HRTFs, according to (6.2). For the required incoherent source signals, either incoherent noise sequences were used, or they were generated from the input signal $x(n)$ by applying decorrelation techniques.

Throughout the evaluation, the neutral head orientation HRIRs of the FABIAN HRTF dataset [78, 152] were employed. The HRIRs all have a length of $q = 256$ samples at a sampling rate of $f_s = 44.1$ kHz. A diffuse field equalization was performed using the provided equalization filter. The FABIAN HRTF dataset was selected primarily because of its dense spatial resolution of 2° in elevation and about 2° great circle distance in azimuth, resulting in 11 950 source positions. The HRTF dataset positions used to calculate the target auditory cues for the proposed method were selected to align with the positions of the point sources employed to render the binaural reference signal, thereby ensuring direct comparability between the two methods. Unless stated otherwise, all HRTF dataset positions within the considered extent range were used during rendering.

Decorrelated versions of the input signal $x(n)$, needed both for the proposed method and to render the binaural reference signal, were generated using the lattice all-pass filter-based quadrature mirror

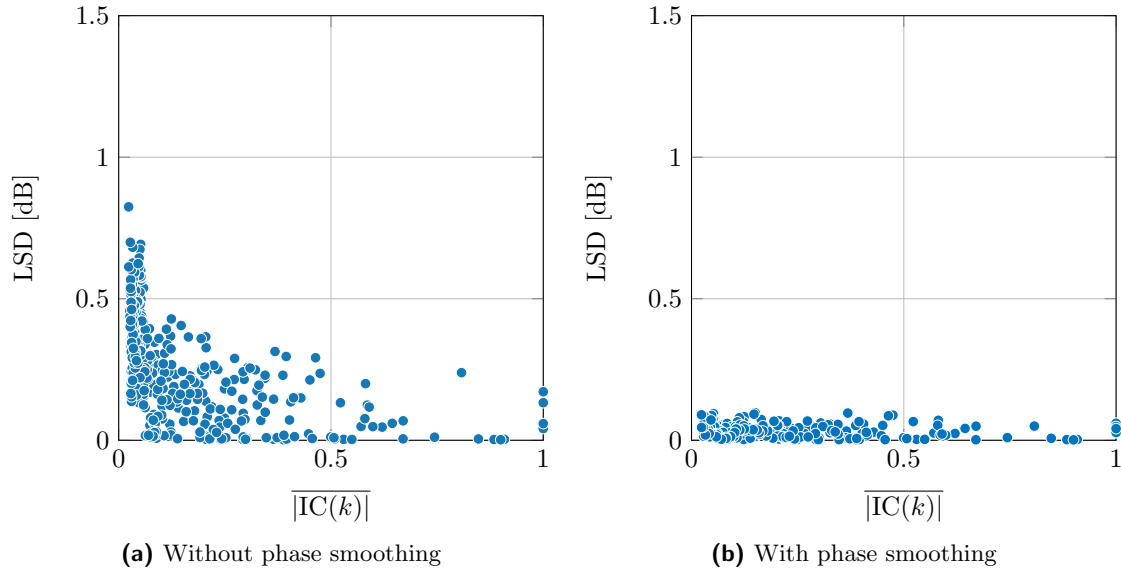


Figure 6.4: LSD between implemented filters and ideal filters without magnitude distortion depending on the target IC value averaged over frequency. A large variety of spatial extent ranges was considered, covering all areas of the sphere.

filter (QMF)-domain decorrelator employed in MPEG Surround [26, 92]. 10 mutually independent sets of filter coefficients were available, which allowed for the generation of 10 decorrelated versions of $x(n)$. To determine the decorrelated source signals used to render the binaural reference signal, the tree structure described in Appendix A.1 was employed, which is a multi-channel generalization of (6.17). As a result, 11 decorrelated source signals were obtained.

6.3.2 Influence of Phase Smoothing

First, the influence of the employed phase smoothing described in Section 6.2.2.2 on the filters' magnitude response was evaluated. To quantify the amount of magnitude distortion, we considered the log-spectral distance (LSD) between ideal filters without magnitude distortion and the implemented filters, both with and without phase smoothing. The LSD quantifies differences between the filters in terms of their magnitude characteristics and is, therefore, directly related to the amount of magnitude distortion introduced to the resulting output signals.

The ideal filters without magnitude distortion were determined as described in Section 6.2.2.1, with $\text{IPD}(k)$ set to 0. This ensures a smooth interpolation behavior of the filters' magnitude response in between the specified frequency bins. All time-domain FIR filters $w_{ij}(n)$ were zero-padded to a length of $q' = 1024$ samples, and a DFT was applied to obtain zero-padded frequency-domain filters $W'_{ij}(k)$, $k = 1, \dots, q'/2 + 1$. Subsequently, the power spectra of the four filters were summed up to obtain one overall power spectrum $P(k)$:

$$P(k) = |W'_{11}(k)|^2 + |W'_{12}(k)|^2 + |W'_{r1}(k)|^2 + |W'_{r2}(k)|^2. \quad (6.24)$$

This overall power spectrum was calculated for both the implemented and the undistorted filters,

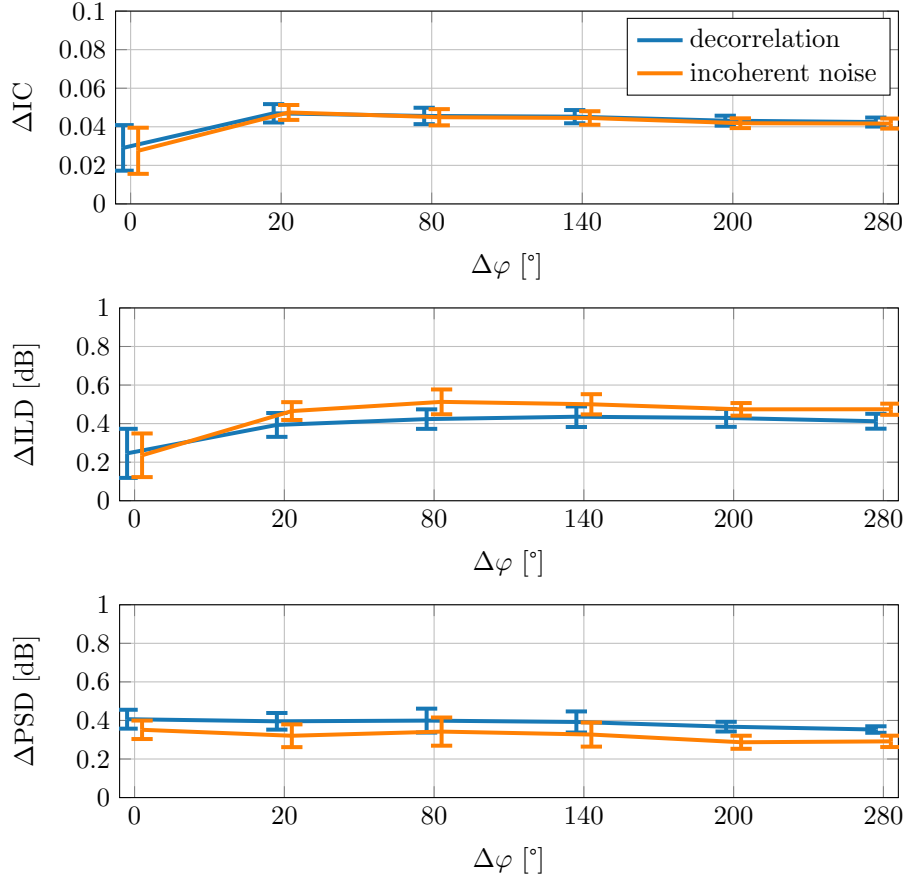


Figure 6.5: Mean and standard deviation of RMSE between objective metrics of proposed method and binaural reference depending on the extent $\Delta\varphi$, averaged over all values for the center $\bar{\varphi}$ of the SESS, for white noise input signal. Two variants of the proposed method were considered: applying decorrelation to generate the two decorrelated input signals and using ideal incoherent noise sequences.

denoted by $P_i(k)$ and $P_u(k)$, respectively. The LSD was then calculated as follows:

$$\text{LSD} = \sqrt{\frac{1}{q'/2+1} \sum_{k=1}^{q'/2+1} \left[10 \log_{10} \left(\frac{P_u(k)}{P_i(k)} \right) \right]^2}, \quad (6.25)$$

where $\log_{10}(\cdot)$ denotes the logarithm to the base 10.

Figure 6.4 shows the LSD over $|\text{IC}(k)|$, considering a large variety of spatial extent ranges that cover all areas of the sphere. From the plots, it is evident that the LSD and, thus, the amount of magnitude distortion is considerably reduced when phase smoothing is applied.

6.3.3 Objective Evaluation

For the objective evaluation, a white noise input signal with a duration of 1 s was employed. To avoid decorrelation artifacts, ideal incoherent white noise sequences were used to generate the binaural reference signal instead of applying decorrelation. Regarding the proposed method, two variants were considered: the application of decorrelation to generate the two decorrelated input signals according

to (6.17) and the use of ideal incoherent noise sequences. This allows quantifying the influence of the employed decorrelation on the considered objective metrics.

Three perceptually relevant objective metrics were calculated from the binaural output signals for the different processing methods: the IC, the ILD, and the PSD. Given a time-frequency-domain binaural output signal $\mathbf{y}(k, l) = [Y_1(k, l), Y_r(k, l)]^T$, the objective metrics are defined as follows:

$$\text{IC}(k, l) = \frac{|\mathcal{E}\{Y_1(k, l)Y_r^*(k, l)\}|}{\sqrt{\mathcal{E}\{|Y_1(k, l)|^2\}\mathcal{E}\{|Y_r(k, l)|^2\}}}, \quad (6.26)$$

$$\text{ILD}(k, l) = 10 \log_{10} \left(\frac{\mathcal{E}\{|Y_1(k, l)|^2\}}{\mathcal{E}\{|Y_r(k, l)|^2\}} \right), \quad (6.27)$$

$$\text{PSD}(k, l) = 10 \log_{10} (\mathcal{E}\{|Y_1(k, l)|^2\} + \mathcal{E}\{|Y_r(k, l)|^2\}). \quad (6.28)$$

To compute the objective metrics, both the proposed method output and the binaural reference signal were transformed to the time-frequency domain using an STFT with a frame length of 256 samples and a hop size of 128 samples. The statistical expectation was approximated by calculating the average value over the entire signal length. As a result, the final considered objective metrics are independent of l . The objective metrics were calculated using all 180 HRTF dataset positions in the horizontal plane as the center $\bar{\varphi}$ of the SESS (i.e., $\bar{\theta} = 0^\circ$). Only horizontal spatial extents with $\Delta\varphi \in \{0^\circ, 20^\circ, 80^\circ, 140^\circ, 200^\circ, 280^\circ\}$ and $\Delta\theta = 0^\circ$ were considered.

For each extent range, the root-mean-square error (RMSE) between the objective metrics of the proposed method output and those of the binaural reference signal was calculated by averaging over the frequency dimension. Figure 6.5 depicts the mean and standard deviation of the RMSE for each objective metric as a function of the extent $\Delta\varphi$, averaged over all considered values for the center $\bar{\varphi}$ of the SESS. In terms of the considered objective metrics, no significant degradation can be observed when decorrelation is applied. Furthermore, the RMSE values exhibit only a slight dependency on the horizontal extent $\Delta\varphi$. In general, the objective metrics of the proposed method output come very close to those of the binaural reference signal, with an RMSE value clearly below 1 dB for the ILD and the PSD and an RMSE value clearly below 0.1 for the IC. This demonstrates the high precision of the proposed rendering method.

6.3.4 Perceptual Evaluation

To perform a perceptual evaluation of the proposed method, a formal listening test was conducted. In the listening test, the output of the proposed method was compared to the binaural reference signal. A selection of the items used in the listening test is available at www.audiolabs-erlangen.de/resources/2022-JAES-SESS.

6.3.4.1 Listening Test Setup

Items A total of 73 test items were incorporated in the listening test, covering a wide range of spatial extent ranges as well as different input signals.

Three different input signals were considered: a pink noise, a vocal quartet, and a piano signal. The pink noise signal and decorrelated versions thereof, which are required for the proposed method and for the generation of the binaural reference signal, were obtained by spectrally shaping incoherent

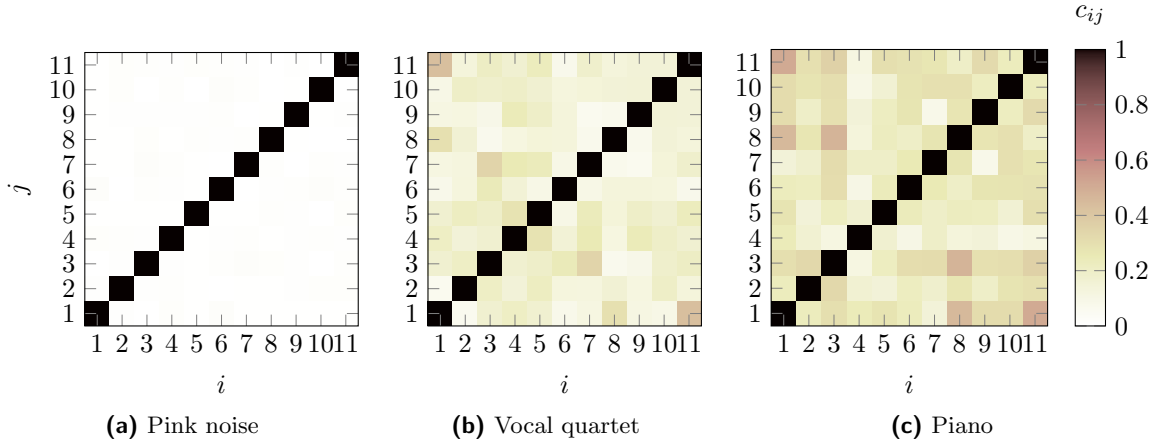


Figure 6.6: Maximum correlation between all decorrelated source signal pairs for all three input signals.

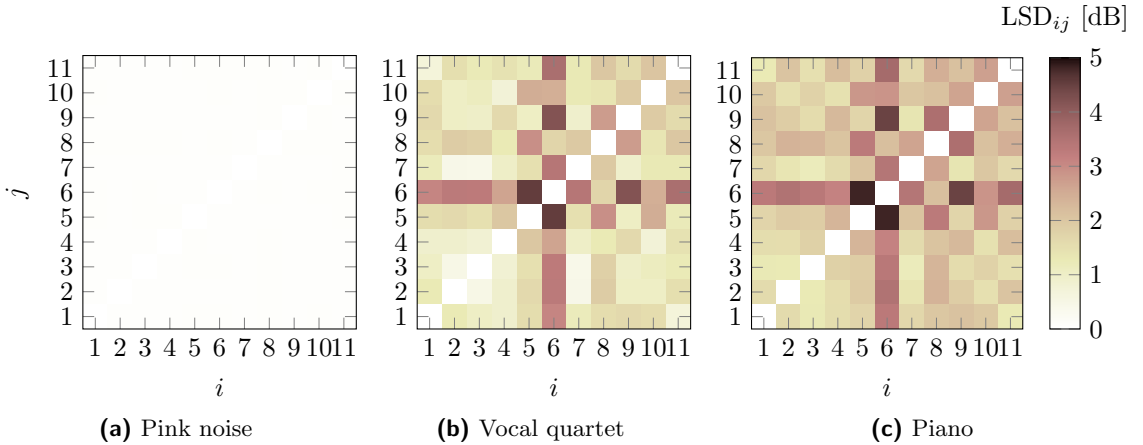


Figure 6.7: LSD between all decorrelated source signal pairs for all three input signals.

white noise sequences. The vocal quartet and piano signals were taken from the EBU SQAM CD [128] (track numbers 48 and 60, respectively). Only the first channel was used, and excerpts of about 10 s were extracted. The required decorrelated versions of both signals were generated using decorrelation techniques as described in Section 6.3.1. For the pink noise signal, all HRTF dataset positions within the considered extent range were used. For the music signals, a significantly coarser spacing of 20° in azimuth and 40° in elevation was employed, as only 11 decorrelated source signals were available to generate the binaural reference signal.

Under ideal conditions, the employed decorrelated signals are mutually uncorrelated and do not possess significant differences in terms of their spectral magnitude characteristics. To assess the effectiveness of the decorrelation, we evaluated the correlation and the LSD between the decorrelated source signals used to generate the binaural reference signal, denoted by $s_i(n)$, $i = 1, \dots, 11$. Given two time-domain signals $s_i(n)$ and $s_j(n)$, the maximum correlation c_{ij} between both is defined as follows:

$$c_{ij} = \max_d \frac{|\sum_n s_i(n)s_j(n+d)|}{\sqrt{\sum_n s_i^2(n)\sum_n s_j^2(n)}}. \quad (6.29)$$

Table 6.1: Extent ranges considered in the listening test. Values indicated in bold were employed exclusively with the pink noise signal.

(a) Horizontal extent		(b) Horizontal and vertical extent	
$\Delta\varphi$	20°, 80°, 140°, 200°, 280°	$\Delta\varphi$	20°, 80°, 140°
$\Delta\theta$	0°	$\Delta\theta$	40°, 80°
$\bar{\varphi}$	0°, 40°, 90°	$\bar{\varphi}$	0°, 40°, 90°
$\bar{\theta}$	0°	$\bar{\theta}$	0°

(c) Additional cases	
$[\varphi_1, \varphi_2]$	$[\theta_1, \theta_2]$
$[-180^\circ, 180^\circ]$	$[0^\circ, 0^\circ]$
$[-90^\circ, 90^\circ]$	$[-90^\circ, 90^\circ]$
$[-180^\circ, 180^\circ]$	$[0^\circ, 90^\circ]$
$[-180^\circ, 180^\circ]$	$[-90^\circ, 90^\circ]$

In accordance with Section 2.1.6.1, delays d within the perceptually relevant range of ± 1 ms are considered. Figure 6.6 depicts the maximum correlation between all decorrelated signal pairs for each input signal. The mean value of c_{ij} (for $i \neq j$) equals 0.0094 for the pink noise, 0.1652 for the vocal quartet, and 0.2464 for the piano signal. Moreover, spectral magnitude differences between the decorrelated signal pairs were evaluated by means of the LSD, which is defined analogously to (6.25):

$$\text{LSD}_{ij} = \sqrt{\frac{1}{K} \sum_{k=1}^K \left[10 \log_{10} \left(\frac{P_i(k)}{P_j(k)} \right) \right]^2}, \quad (6.30)$$

with $P_i(k) = \frac{1}{L} \sum_l |S_i(k, l)|^2$, where $S_i(k, l)$ denotes a time-frequency-domain representation of $s_i(n)$ which is obtained by applying an STFT with a frame length of 256 samples and a hop size of 128 samples, resulting in a single-sided spectrum with $K = 129$ frequency bins and L time frames. Figure 6.7 depicts the LSD between all decorrelated signal pairs for each input signal. The mean value of LSD_{ij} (for $i \neq j$) equals 0.0762 dB for the pink noise, 1.7196 dB for the vocal quartet and 2.2258 dB for the piano signal. In conclusion, it can be stated that the output of the decorrelator for both the vocal quartet and the piano signal exhibits clear deviations from the ideal case, both in terms of the resulting correlation and spectral magnitude characteristics.

The considered spatial extent ranges can be classified into three categories: horizontal extent only, horizontal and vertical extent, and additional cases. Table 6.1 provides an overview of all considered extent ranges. For the horizontal as well as horizontal and vertical extent cases, all possible combinations of the variables $\Delta\varphi$, $\Delta\theta$, $\bar{\varphi}$, and $\bar{\theta}$ were examined. The values indicated in bold were employed exclusively with the pink noise signal. For the music signals, an insufficient number of decorrelated source signals were available to generate the binaural reference signal corresponding to these extent ranges. Four additional, rather extreme use cases were considered for the pink noise signal. This resulted in a total of 37 test items for the pink noise signal and 18 test items for each of the music signals, adding up to 73 test items in total.

Procedure The conditions under test included the proposed method output, generated as described in Section 6.2, and a hidden binaural reference, generated as described in Section 6.3.1. For each item, the participants were asked to rate the presented conditions w.r.t. the binaural reference in terms of overall perceived differences on a continuous scale ranging from 1 (“huge differences”) to 5 (“no differences”), using the following intermediate anchor points: 2 (“clear differences”), 3 (“small differences”), 4 (“very small differences”). The employed grading scale had a resolution of two decimal places. The listening test was designed in accordance with the ITU-R BS.1116 recommendation [153], using different scale labels. This approach was taken to obtain a judgment about the amount of difference between the conditions, rather than about which condition sounds better or worse.

In order to avoid listener fatigue, the test was divided into two sessions which were carried out on different days. Before the main test, a training phase was included to familiarize the participants with the test procedure and the nature of the signals to be expected. The training consisted of three test items, one for each input signal. In order to implement the test methodology, a customized version of the webMUSHRA software was used [130].

The listening test was conducted in a quiet listening room. Stimuli were presented via open electrostatic Stax SR-404 headphones with an SRM-006tII driver unit.

Subjects A total number of 14 subjects participated in the listening test. All subjects were employees of either the International Audio Laboratories Erlangen or Fraunhofer IIS and were working within the field of audio. One subject reported having no experience, six subjects reported having little experience, and six subjects reported having considerable experience in performing listening tests. In a post-screening step, participants who missed the hidden reference by more than 0.5 points in more than 15% of the cases were excluded. After post-screening, data of 13 participants was included in the statistical analysis, 10 male and 3 female. The average age of the participants was 30 years.

6.3.4.2 Results

Figure 6.8 shows the mean values and bootstrapped 95% confidence intervals (CIs) of the listening test scores for the proposed method output, grouped by input signal. The bootstrapped CIs were determined based on 1000 iterations. A Shapiro-Wilk test revealed that the listening test scores in none of the input signal groups are normally distributed (pink noise: $W = 0.56, p < .001$, piano: $W = 0.93, p < .001$, vocal quartet: $W = 0.88, p < .001$). Consequently, non-parametric tests were employed for the statistical analysis.

The mean ratings for the vocal quartet and piano items are 3.97 (CI 3.83 to 4.12) and 3.79 (CI 3.66 to 3.93), respectively. These values are close to 4.0, which corresponds to very small differences. This indicates that there is only a very small perceptual difference between the proposed method output and the binaural reference signal for the piano and vocal quartet items. However, from Figure 6.8 and a pairwise Wilcoxon test, it is evident that the vocal quartet ($p < 0.01$) and piano ($p < 0.01$) items are rated significantly lower than the pink noise items. In terms of processing, the only difference between the different input signals is how the corresponding decorrelated versions were obtained, both for generating the binaural reference signal and the proposed method output. For the music signals, imperfect decorrelation filters were used, both in terms of the resulting correlation and spectral magnitude characteristics, as discussed in Section 6.3.4.1. This causes the binaural reference signal as well as the proposed method output to deviate from the ideal case and can explain the larger differences

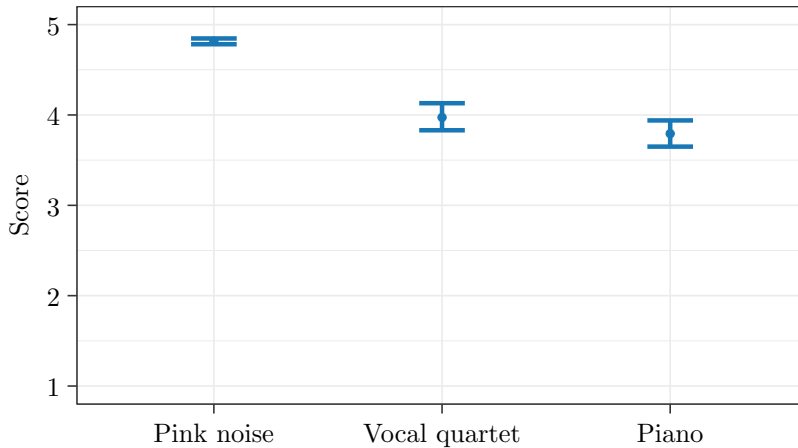


Figure 6.8: Mean values and bootstrapped 95% CIs of listening test scores for proposed method output, grouped by input signal.

observed in comparison to the pink noise items.

Since the performance for the pink noise items is not impaired by suboptimal decorrelation filters, these results are most suited to evaluate differences caused by the processing method itself. A paired Wilcoxon test comparing the scores of the proposed method output and the corresponding hidden binaural reference signal for the pink noise items shows statistically significant differences ($W = 12459.5, p < .001$). However, the ratings of the pink noise items are very close to 5.0 with a mean of 4.82 (CI 4.78 to 4.85), which corresponds to no perceived differences. For the pink noise items, the proposed method output is thus perceptually nearly indistinguishable from the binaural reference signal.

6.3.5 Discussion

The objective evaluation provided in Section 6.3.3 showed that the proposed method output matches the binaural reference signal well in terms of IC, ILD, and PSD. Furthermore, the listening test discussed in Section 6.3.4 has demonstrated that the proposed method output is perceptually nearly indistinguishable from the binaural reference signal. In comparison to a direct implementation of the rendering model, which the binaural reference signal corresponds to, the proposed method offers a number of advantages.

A significant advantage is the reduction of computational complexity. Starting from a single-channel input signal, only one decorrelation filter is applied to generate the two decorrelated input signals according to (6.17). Subsequently, each of the two signals is filtered with two FIR filters of the same length as the original HRIRs. In contrast, a direct implementation of the rendering model requires the computation of a potentially large number of decorrelated input signals, each of which must be filtered with the corresponding HRIRs individually. Given that high-quality audio decorrelation is a computationally expensive operation, this results in a significant difference in complexity in favor of the proposed method. Furthermore, the savings increase in proportion to the desired spatial extent range.

The run-time complexity of the proposed method can be further reduced by precomputing the



Figure 6.9: Exemplary MPEG-I VR test scene, including a fountain as sound source with extent.

target auditory cues and storing them in a lookup table, from which they can be retrieved at run time. Alternatively, since the employed processing filters are signal-independent, the resulting FIR filters can be stored directly. As an additional advantage, precomputing this data ensures that the run-time complexity is independent of the desired spatial extent range.

A second advantage of the proposed method is that there is no need for a large number of decorrelation filters that produce mutually incoherent output signals, as is required for a direct implementation of the rendering model. Identifying such filters that simultaneously achieve a low degree of correlation and good perceptual quality is generally a challenging task, which becomes increasingly complex when a growing number of incoherent output signals are required. This makes a high-quality implementation of the rendering model impractical. Conversely, the need for only one incoherent output signal provides flexibility when selecting an appropriate decorrelation filter.

6.4 Application in VR/AR Standardization

The proposed method for binaural rendering of homogeneous SESSs has been employed successfully in the submission of Fraunhofer IIS and its partners to the ISO/MPEG standardization effort of MPEG-I Immersive Audio [15]. The goal of this standardization effort is to develop a comprehensive specification for a compressed representation and high-quality audio rendering of six-degrees-of-freedom (6DoF) VR and AR applications. This includes a multitude of sophisticated acoustic effects, such as sources with extent and directivity, accurate modeling of acoustic propagation effects including Doppler, and diffraction and occlusion by acoustic obstacles. Specifically, the technology described in this chapter was part of the winning proposal in the Call for Proposals. To this end, it was evaluated against 13 competing proposals as part of a large-scale test effort conducted in 12 test laboratories on a worldwide scale, which assessed the real-time VR audio rendering quality. The VR part of the test included several test scenes that demand realistic rendering of SESSs, such as fountains and waterfalls. An example of an MPEG-I VR test scene is presented in Figure 6.9. Similarly, the technology was evaluated for AR use cases. As a result of the perceptual evaluation, the technology was adopted as part of the MPEG-I

6DoF Audio Reference Model, i.e., the baseline for all further technology development of the MPEG-I Immersive Audio standard.

6.5 Summary

In this chapter, we proposed a method for efficient and realistic binaural rendering of homogeneous SESSs. Given a desired spatial extent range, a number of target auditory cues are determined based on the model of an incoherently extended sound source with position-independent energy and spectral content. A binaural output signal with the desired target auditory cues is generated by processing two decorrelated input signals, which are obtained from a single-channel input signal using decorrelation techniques. The employed processing filters are signal-independent and thus depend solely on the desired spatial extent range. The proposed method exhibits two important advantages over a direct implementation of the rendering model:

1. The computational complexity is reduced significantly. When precomputed target auditory cues are employed, the run-time complexity is furthermore independent of the desired spatial extent range.
2. There is no need for a large number of decorrelation filters that produce mutually incoherent output signals, which are generally challenging to find.

The proposed method output was compared to a binaural reference signal resulting from a direct implementation of the rendering model. An objective evaluation showed that the proposed method output closely matches the binaural reference signal in terms of IC, ILD, and PSD. Complementing the objective evaluation results, a formal listening test showed that the output of the proposed method is perceptually nearly indistinguishable from the binaural reference signal. The proposed method thus succeeds in replicating the rendering model in a perceptual sense and can be used instead without performance degradation.

Currently, the proposed method acts on the assumption that the two decorrelated input signals are perfectly uncorrelated and possess the same PSD. However, in practice, deviations from these ideal conditions may occur. One potential cause for this is an imperfect decorrelation filter used to generate the two decorrelated input signals. Furthermore, recorded stereo signals that do not fully meet the specified requirements may be used directly as the two decorrelated input signals. A possibility for improvement would thus be to adjust the processing steps by taking the input signal properties into account.

The proposed method was adopted as part of the MPEG-I 6DoF Audio Reference Model, which serves as the baseline for further technology development of the MPEG-I Immersive Audio standard.

CHAPTER 7

Binaural Rendering of Heterogeneously Extended Sound Sources

The main content of this chapter is based on: C. Anemüller, O. Thiergart, and E. A. P. Habets, “Binaural rendering of heterogeneous sound sources with extent,” in Proc. ICASSP, Apr. 2024, pp. 471–475, [20].

In spatial audio rendering applications, it is often desired to realistically render sound sources with a certain spatial extent. While existing methods for rendering of spatially extended sound sources (SESSs) mainly consider rendering of homogeneous SESSs (i.e., with constant radiation characteristics over the extent), rendering of heterogeneous SESSs (i.e., with position-dependent radiation characteristics) has barely been discussed in the literature. However, depending on the source’s characteristics, accounting for the position-dependent radiation behavior of the SESS can be crucial for achieving a realistic rendering. Examples of relevant heterogeneous SESSs are an applauding audience, for which the individual people’s claps may possess different temporal or spectral characteristics, or a choir, for which the vocal range of the singers is usually position-dependent.

To the best of our knowledge, the method described in [15, 118], which is based on virtual loudspeaker rendering, is the only method proposed in the literature that specifically addresses rendering of heterogeneous SESSs. In contrast to homogeneous SESS rendering methods, rendering of heterogeneous SESSs generally requires a multi-channel input signal to obtain information about the position-dependent radiation behavior.

In this chapter, we propose a method for binaural rendering of heterogeneous SESSs given a two-channel input signal. As opposed to the method described in [15, 118]¹, the proposed method does not rely on virtual loudspeaker rendering. Instead, a binaural output signal is generated directly based on a heterogeneous SESS rendering model. Therefore, the homogeneous SESS rendering model used in Chapter 6 is extended by taking the position-dependent radiation characteristics of the SESS into account. Based on this rendering model, a binaural output signal with the desired inter-channel properties is obtained using the covariance domain framework proposed in [117], while ensuring that the directional characteristics encoded in the two-channel input signal are preserved.

¹As of the submission date of this thesis, neither a sufficiently detailed description nor an implementation of the method described in [15, 118] was available. Consequently, no direct comparison between the proposed method and the method described in [15, 118] was conducted.

The remainder of this chapter is structured as follows. Section 7.1 presents a mathematical formulation of the employed heterogeneous SESS rendering model. In Section 7.2, the proposed rendering method is described, and its performance is evaluated both objectively and perceptually in Section 7.3. This is achieved by comparing the proposed method's output against a simulated binaural reference signal. Finally, Section 7.4 concludes this chapter by providing a summary.

7.1 Rendering Model

In this chapter, we use a model of an incoherently extended sound source with position-dependent energy and spectral content to describe a heterogeneous SESS. Therefore, we extend the homogeneous SESS rendering model described in Section 6.1 by taking the position-dependent radiation characteristics of the SESS into account. While the heterogeneous SESS rendering model largely aligns with the homogeneous SESS rendering model discussed in Section 6.1, a comprehensive description is given here for the sake of completeness. The model is described in the discrete time-frequency domain.

Let $S(k, l, \mathbf{u})$ denote the source signal emitted by the SESS arriving from direction $\mathbf{u} = [\varphi, \theta]^T$ at a receiver, where φ describes the incident azimuth angle, θ the incident elevation angle, and k, l the discrete frequency and time frame indices, respectively. The operator $(\cdot)^T$ denotes the transpose operation. Since incoherent sound radiation for each direction is assumed, the following condition holds for two arbitrary directions \mathbf{u}_i and \mathbf{u}_j :

$$\mathcal{E}\{S(k, l, \mathbf{u}_i)S^*(k, l, \mathbf{u}_j)\} = \begin{cases} 0, & \mathbf{u}_i \neq \mathbf{u}_j \\ P(k, l, \mathbf{u}_i), & \mathbf{u}_i = \mathbf{u}_j, \end{cases} \quad (7.1)$$

where $P(k, l, \mathbf{u})$ denotes the incident power spectral density (PSD) for direction \mathbf{u} , $\mathcal{E}\{\cdot\}$ the statistical expectation, and $(\cdot)^*$ the complex conjugate. Generally, $P(k, l, \mathbf{u})$ reflects both the position-dependent radiation behavior of the SESS and the distance attenuation, which depends on the geometry of the SESS and its relative position to the receiver. Here, we consider the more general case in which no specific information about the geometry of the SESS is available. We, therefore, assume that the distance from the receiver to the SESS is angle-independent and do not model the distance attenuation.

In this chapter, binaural rendering of the SESS is considered. Given a set of head-related transfer functions (HRTFs) $\mathbf{h}(k, \mathbf{u}_n) = [H_1(k, \mathbf{u}_n), H_r(k, \mathbf{u}_n)]^T$ for a number of discrete and uniformly spaced directions $\mathbf{u}_n, n = 1, \dots, N$, covering the extent of the SESS, the binaural output signal $\mathbf{y}(k, l) = [Y_1(k, l), Y_r(k, l)]^T$ can be determined:

$$\mathbf{y}(k, l) = \sqrt{\frac{1}{N}} \sum_{n=1}^N \mathbf{h}(k, \mathbf{u}_n) S(k, l, \mathbf{u}_n). \quad (7.2)$$

Using (7.1) and (7.2), the output covariance matrix $\mathbf{C}_y(k, l)$ follows:

$$\begin{aligned} \mathbf{C}_y(k, l) &= \mathcal{E}\{\mathbf{y}(k, l)\mathbf{y}^H(k, l)\} \\ &= \frac{1}{N} \sum_{n=1}^N P(k, l, \mathbf{u}_n) \mathbf{h}(k, \mathbf{u}_n) \mathbf{h}^H(k, \mathbf{u}_n), \end{aligned} \quad (7.3)$$

where $(\cdot)^H$ denotes the Hermitian transpose.

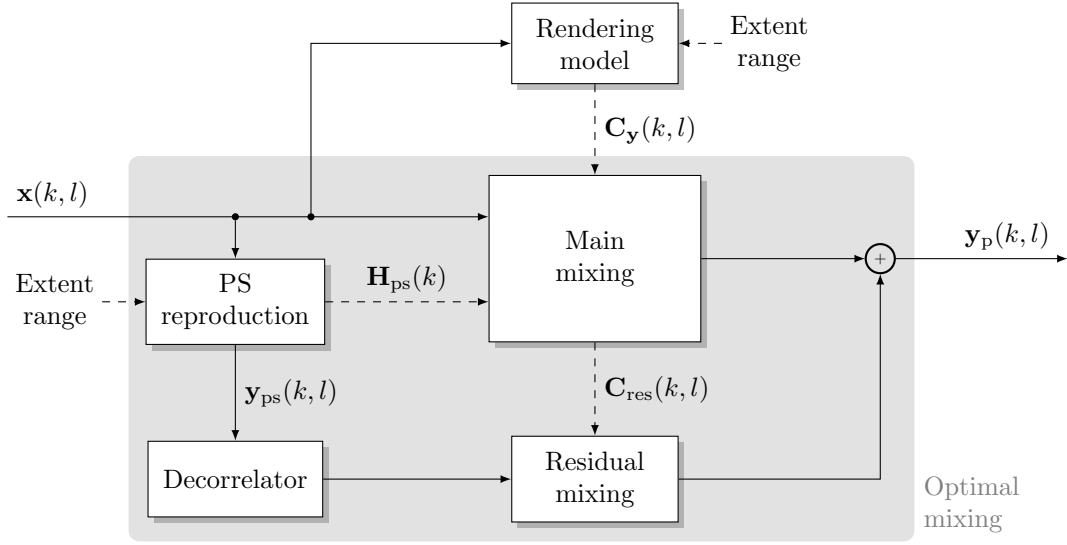


Figure 7.1: Block diagram of proposed heterogeneous SESS rendering method.

By assuming direction-independent energy, i.e., $P(k, l, \mathbf{u}) = P(k, l)$, expressions for the binaural output signal and its covariance matrix corresponding to the homogeneous SESS rendering model follow, given by (6.2) and (6.3), respectively. The schematic overview of the homogeneous SESS binaural rendering model depicted in Figure 6.1 is thus also applicable to the heterogeneous SESS binaural rendering model considered here.

7.2 Proposed Rendering Method

Figure 7.1 shows a block diagram of the proposed heterogeneous SESS rendering method. Input to the rendering method is a time-frequency-domain multi-channel signal $\mathbf{x}(k, l)$. The channels of $\mathbf{x}(k, l)$ represent different spatial regions of the heterogeneous SESS and can be obtained by recording the SESS using multiple directional microphones. In the following, the simplest case of a two-channel input signal $\mathbf{x}(k, l) = [X_l(k, l), X_r(k, l)]^T$ is considered, for which the first channel represents the left part of the SESS, and the second channel represents the right part of the SESS. Figure 7.2 depicts a schematic overview of a feasible two-channel input signal recording setup.

For a given extent range, a binaural output signal $\mathbf{y}_p(k, l)$ with inter-channel properties that match

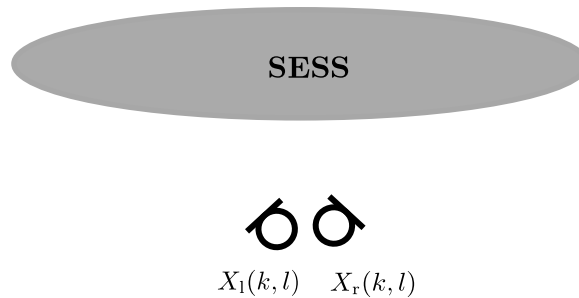


Figure 7.2: Schematic overview of feasible SESS two-channel input signal recording setup.

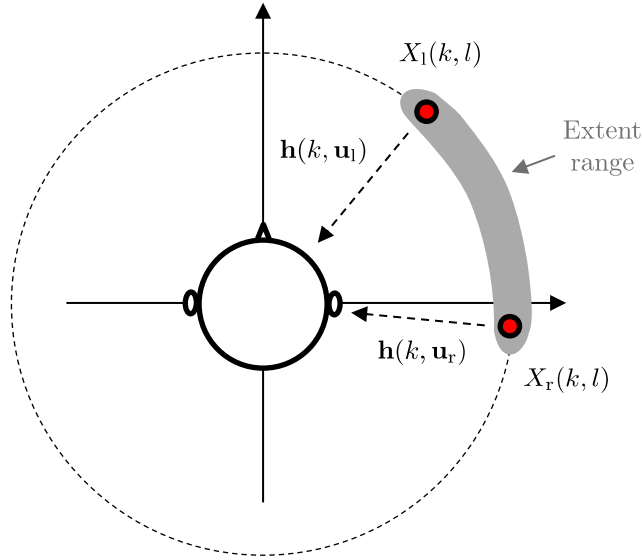


Figure 7.3: Schematic overview of point source reproduction procedure. For visualization purposes, the special case of horizontal spatial extent only is considered.

the rendering model discussed in Section 7.1 is determined using the covariance domain framework proposed in [117]. To ensure that the directional characteristics encoded in the two-channel input signal are preserved, the solution is constrained to be as close as possible to $\mathbf{y}_{\text{ps}}(k, l)$, the binaural output signal for point source reproduction of $X_1(k, l)$ and $X_r(k, l)$ at the respective outer edges of the extent. Given the directions \mathbf{u}_l and \mathbf{u}_r corresponding to the outer left and right edges of the extent relative to the listener's orientation, respectively, $\mathbf{y}_{\text{ps}}(k, l)$ is defined as follows:

$$\begin{aligned} \mathbf{y}_{\text{ps}}(k, l) &= [\mathbf{h}(k, \mathbf{u}_l), \mathbf{h}(k, \mathbf{u}_r)] \mathbf{x}(k, l) \\ &= \mathbf{H}_{\text{ps}}(k) \mathbf{x}(k, l). \end{aligned} \quad (7.4)$$

In Figure 7.3, a schematic overview of the point source reproduction procedure is depicted. For visualization purposes, the special case of horizontal spatial extent only is considered.

Given a desired spatial extent, first, the target output covariance matrix $\mathbf{C}_{\mathbf{y}}(k, l)$ is determined according to (7.3). The direction-dependent energy $P(k, l, \mathbf{u}_n)$ is approximated based on the PSD of the two-channel input signal. At the directions \mathbf{u}_l and \mathbf{u}_r , the PSD is set to $P(k, l, \mathbf{u}_l) = \mathcal{E}\{|X_1(k, l)|^2\}$ and $P(k, l, \mathbf{u}_r) = \mathcal{E}\{|X_r(k, l)|^2\}$. Since a one-dimensional extent is considered, linear interpolation is performed to obtain the PSD at intermediate directions.

Given the target covariance matrix $\mathbf{C}_{\mathbf{y}}(k, l)$ and an estimate of the input covariance matrix $\mathbf{C}_{\mathbf{x}}(k, l)$:

$$\mathbf{C}_{\mathbf{x}}(k, l) = \mathcal{E}\{\mathbf{x}(k, l) \mathbf{x}^{\text{H}}(k, l)\}, \quad (7.5)$$

the optimal mixing procedure described in [117] is employed to determine a main mixing matrix $\mathbf{M}(k, l)$ and a residual mixing matrix $\mathbf{M}_{\text{res}}(k, l)$ such that:

$$\mathbf{y}_{\text{p}}(k, l) = \mathbf{M}(k, l) \mathbf{x}(k, l) + \mathbf{M}_{\text{res}}(k, l) \text{dec}\{\mathbf{y}_{\text{ps}}(k, l)\}, \quad (7.6)$$

where $\text{dec}\{\cdot\}$ describes a decorrelation operation.

First, $\mathbf{M}(k, l)$ is determined such that output signals with covariance matrix $\mathbf{C}_y(k, l)$ are obtained when mixing $\mathbf{x}(k, l)$. Since the optimization problem is under-determined, as an additional constraint, the least-squares error between $\mathbf{M}(k, l)\mathbf{x}(k, l)$ and an energy-normalized version of $\mathbf{y}_{ps}(k, l)$ is minimized following the procedure described in [117]. The energy normalization is performed by matching the per-channel energies of $\mathbf{y}_{ps}(k, l)$ to those of $\mathbf{M}(k, l)\mathbf{x}(k, l)$. This additional optimization constraint ensures that the directional characteristics encoded in the two-channel input signal are preserved. Secondly, residual mixing is performed if mixing with $\mathbf{M}(k, l)$ fails to reach the target covariance matrix $\mathbf{C}_y(k, l)$ sufficiently. The residual mixing matrix $\mathbf{M}_{res}(k, l)$ is determined analogously to $\mathbf{M}(k, l)$, using the residual target covariance matrix $\mathbf{C}_{res}(k, l)$:

$$\mathbf{C}_{res}(k, l) = \mathbf{C}_y(k, l) - \mathbf{M}(k, l)\mathbf{C}_x(k, l)\mathbf{M}^H(k, l). \quad (7.7)$$

As an additional optimization constraint, the least-squares error between $\mathbf{M}_{res}(k, l) \text{dec}\{\mathbf{y}_{ps}(k, l)\}$ and an energy-normalized version of $\text{dec}\{\mathbf{y}_{ps}(k, l)\}$ is now minimized. Note that residual mixing is only required if the channels of $\mathbf{x}(k, l)$ are not sufficiently uncorrelated.

7.3 Performance Evaluation

The performance of the proposed rendering method was evaluated objectively, based on a number of perceptually relevant objective metrics, and perceptually, by means of a formal listening test. Simulated two-channel input signals were used, and the proposed method's output was compared to a simulated binaural reference signal. Additionally, the point source reproduction signal $\mathbf{y}_{ps}(k, l)$ and two homogeneous extent rendering baselines were considered.

7.3.1 Evaluation Setup

For the evaluation of the proposed method, simulated two-channel input signals were used. This enables the generation of a binaural reference signal for all considered extent ranges. Because of the greater perceptual relevance of horizontal spatial extent compared to vertical spatial extent, as discussed in Section 2.3.1.4, only extents in azimuth direction at $\theta = 0^\circ$ elevation were included in the evaluation. The considered spatial extent is thus expressed in terms of the azimuth angle range $[\varphi_1, \varphi_2] = \left[\bar{\varphi} - \frac{\Delta\varphi}{2}, \bar{\varphi} + \frac{\Delta\varphi}{2}\right]$, where $\bar{\varphi}$ describes the extent center, and $\Delta\varphi$ describes the extent in terms of azimuth angle.

A schematic overview of the employed input signal simulation setup is depicted in Figure 7.4. To simulate the two-channel input signals, a number of point sources were uniformly distributed between -55° and 55° azimuth at a distance of 2 m from the receivers. The source signals of the individual point sources are denoted by $S_m(k, l)$, $m = 1, \dots, M$. Only direct sound was simulated. Two recording setups were considered: the ORTF and the XY setup [154]. For each recording setup, two cardioid microphones were used as receivers, pointing toward $\pm 55^\circ$ azimuth, which corresponds to the outer edges of the SESS. For the ORTF setup, the two receivers have a spacing of 17 cm, while for the XY setup, the receivers are not spatially separated.

For the evaluation, three different signals were used: a speech signal, a sparse applause signal, and a dense applause signal. For the speech signal, a freely available text-to-speech synthesis tool was used to generate excerpts of eight different people speaking the same text (i.e., $M = 8$). For the applause

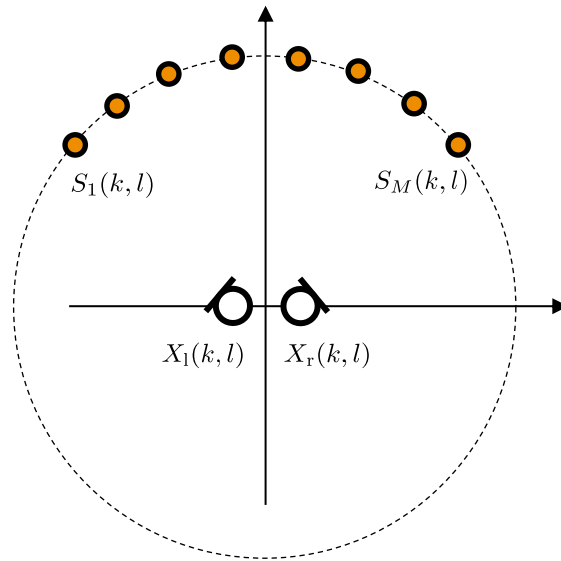


Figure 7.4: Schematic overview of two-channel input signal simulation setup employed for the performance evaluation.

signals, dry studio recordings of individual persons clapping were used. Two applause signal variants were generated: a sparse applause signal, using $M = 48$ individual point sources, and a dense applause signal, using $M = 192$ individual point sources. Note that the degree of heterogeneity of the sources was chosen such that the two-channel approach can provide a reasonable approximation of the source’s direction-dependent energy using linear interpolation. For SESSs with a considerably higher degree of heterogeneity, more input channels are required to ensure a good match between the rendering model and the source’s characteristics.

Processing was performed in the short-time Fourier transform (STFT) domain, using a frame length of 512 samples and a hop size of 256 samples at 44.1 kHz sampling rate. Since the two-channel input signals were simulated under time-invariant conditions, batch processing was performed, i.e., all signal-dependent parameters were estimated over the entire signal length. Note that this is no limitation of the proposed method; realizing online processing would be straightforward. For the proposed method, no residual mixing was employed since it was not found to be necessary for the considered configurations. Consequently, also no decorrelation was required. Throughout the evaluation, the neutral head orientation HRTFs of the FABIAN HRTF dataset were used [78, 152], which offers a high spatial resolution of 2° azimuth in the horizontal plane. A diffuse field equalization was conducted using the provided equalization filter. For the considered homogeneous SESS rendering baselines, decorrelation was performed using the quadrature mirror filter (QMF)-domain decorrelator employed in MPEG Surround [26, 92], which is based on cascaded lattice all-pass filters.

7.3.2 Comparison Methods

The output of the proposed method was compared to a simulated binaural reference signal as well as to the point source reproduction signal $\mathbf{y}_{\text{ps}}(k, l)$, calculated according to (7.4). Additionally, a comparison to two homogeneous SESS rendering baselines was performed.

The binaural reference signal was generated by uniformly distributing the individual point sources

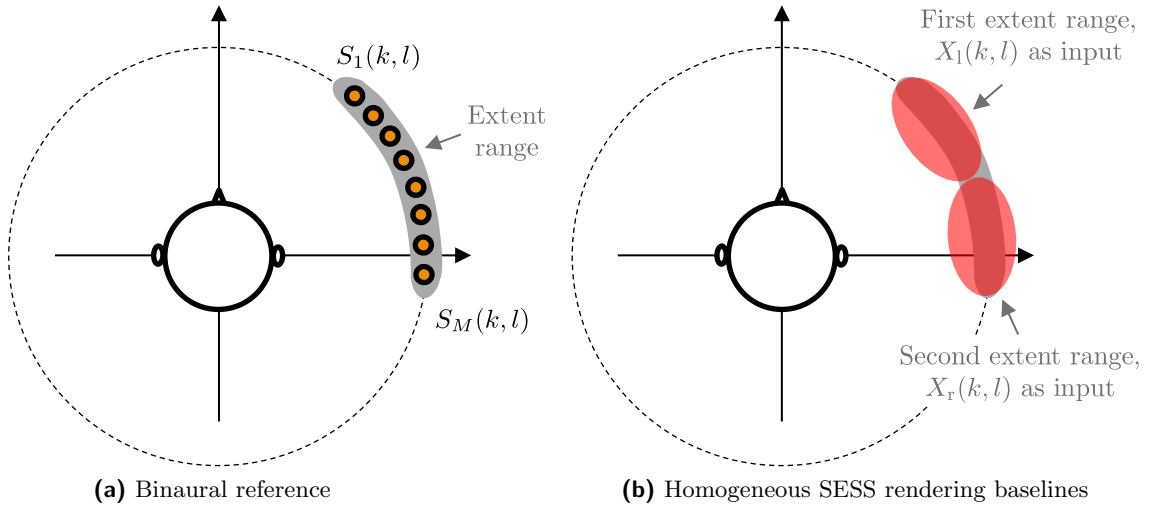


Figure 7.5: Schematic overview of comparison methods considered for the performance evaluation. For visualization purposes, the special case of horizontal spatial extent only is considered.

$S_m(k, l)$, which were used for the input signal simulation, over the considered extent range. By involving the individual point sources with the corresponding HRTFs, the binaural reference signal is obtained. A schematic overview of the binaural reference signal generation is depicted in Figure 7.5a.

For the homogeneous SESS rendering baselines, the homogeneous SESS rendering method described in Chapter 6, as well as the method proposed in [116], were considered. For both, the heterogeneous SESS was rendered as two homogeneous SESSs, covering the left and right parts of the desired extent, respectively. For each of the two homogeneous SESSs, the corresponding channel of the two-channel input signal was used as input. A schematic overview of the binaural rendering procedure for the homogeneous SESS rendering baselines is depicted in Figure 7.5b.

For the homogeneous SESS rendering method described in Chapter 6, two variants were considered: using one and two decorrelation filters per extent range. In the case of the one decorrelation filter variant, the two decorrelated input signals were determined according to (6.17). In the case of the two decorrelation filter variant, the two decorrelated input signals were obtained directly as the output of two mutually incoherent decorrelation filters. Since the channels of the two-channel input signal are partly correlated in a realistic scenario, only the two decorrelation filter variant ensures correct processing in terms of the resulting output covariance matrix when combining the binaural output signals of the two homogeneous SESSs. However, the use of two decorrelation filters per extent range results in a significant increase in decorrelation artifacts. Consequently, we additionally evaluated a variant with only one decorrelation filter per extent range, which corresponds to the original implementation described in Chapter 6.

7.3.3 Objective Evaluation

The objective evaluation focused primarily on the dense applause signal, using both the ORTF and the XY recording setup for the input signal simulation. Three perceptually relevant objective metrics were calculated from the binaural output signals for the different processing methods: the interaural coherence (IC), the interaural level difference (ILD), and the PSD. Given a time-frequency-domain

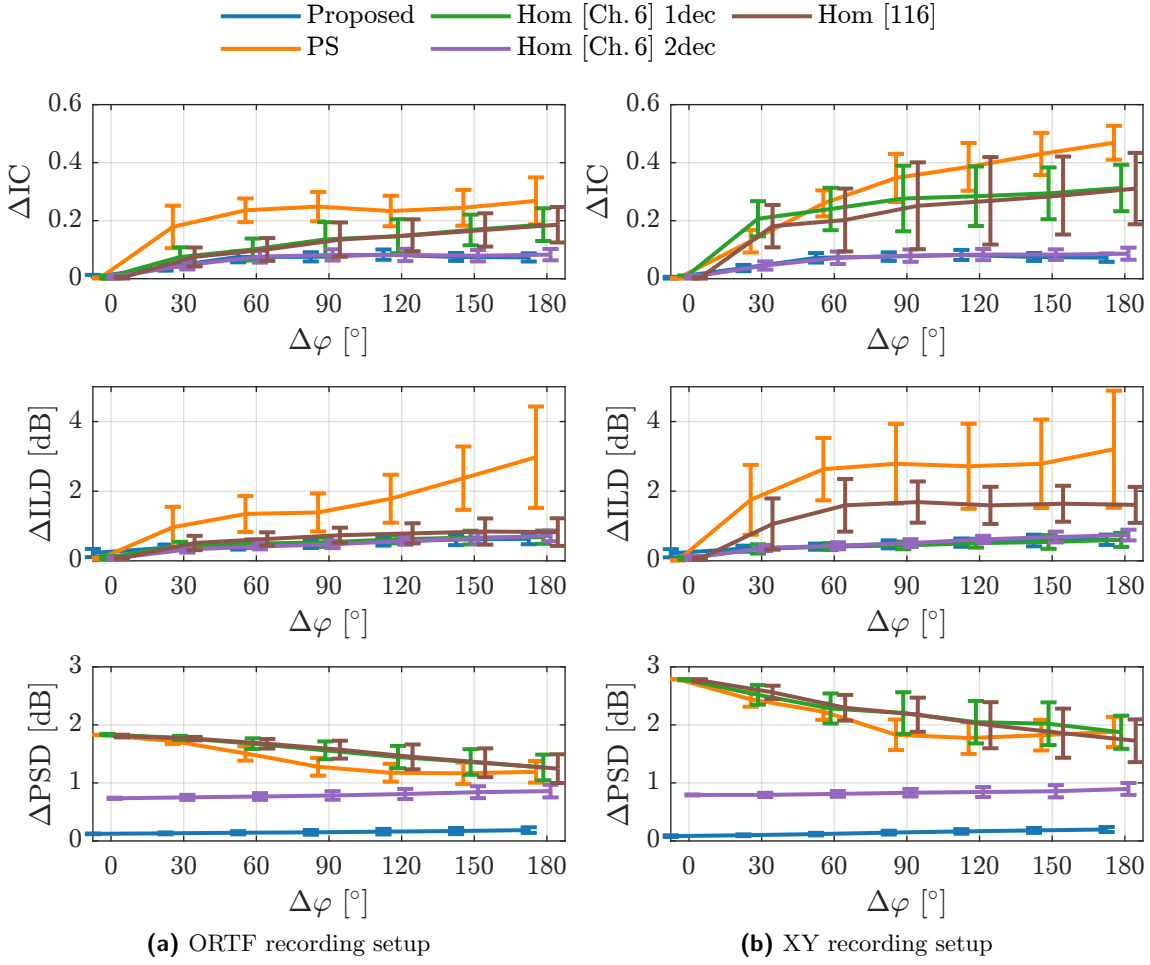


Figure 7.6: Mean and standard deviation of RMSE between objective metrics of considered processing methods and binaural reference depending on the extent $\Delta\varphi$, averaged over all considered values for the center $\bar{\varphi}$ of the SESS, for the dense applause signal.

binaural output signal $\mathbf{y}(k, l) = [Y_1(k, l), Y_r(k, l)]^T$, the objective metrics are defined as follows:

$$\text{IC}(k, l) = \frac{|\mathcal{E}\{Y_1(k, l)Y_r^*(k, l)\}|}{\sqrt{\mathcal{E}\{|Y_1(k, l)|^2\}\mathcal{E}\{|Y_r(k, l)|^2\}}}, \quad (7.8)$$

$$\text{ILD}(k, l) = 10 \log_{10} \left(\frac{\mathcal{E}\{|Y_1(k, l)|^2\}}{\mathcal{E}\{|Y_r(k, l)|^2\}} \right), \quad (7.9)$$

$$\text{PSD}(k, l) = 10 \log_{10} (\mathcal{E}\{|Y_1(k, l)|^2\} + \mathcal{E}\{|Y_r(k, l)|^2\}). \quad (7.10)$$

The objective metrics were calculated in third-octave bands, and the statistical expectation was approximated by calculating the average value over the entire signal length. As a result, the final considered objective metrics are independent of l . A variety of extents $\Delta\varphi \in \{0^\circ, 30^\circ, 60^\circ, 90^\circ, 120^\circ, 150^\circ, 180^\circ\}$ was considered for the objective evaluation, using all 180 HRTF dataset positions in the horizontal plane as center $\bar{\varphi}$ of the SESS.

For each extent range, the root-mean-square error (RMSE) between the objective metrics of each processing method and those of the binaural reference was calculated by averaging over the frequency

dimension. Figure 7.6 shows mean and standard deviation of the RMSE for each of the objective metrics and processing methods depending on the extent $\Delta\varphi$, averaged over all considered values for the center $\bar{\varphi}$ of the SESS, for the dense applause signal. The RMSE values for the proposed method are comparably low. It, therefore, approximates the binaural reference well in terms of the considered objective measures for both recording setups. The homogeneous extent rendering variant based on Chapter 6 with two decorrelation filters (“Hom [Ch. 6] 2dec”) reaches similar RMSE values as the proposed method for both the IC and the ILD. For the PSD, slightly higher RMSE values are obtained. One reason for this may be the employed decorrelation filters. The homogeneous extent rendering variants based on Chapter 6 with one decorrelation filter (“Hom [Ch. 6] 1dec”) and [116] (“Hom [116]”) both show a comparable performance. Their RMSE values are considerably larger than those for “Hom [Ch. 6] 2dec”; this is primarily attributed to the non-negligible correlation between the channels of the input signal. This finding aligns with the observation that their performance is considerably better for the ORTF recording setup compared to the XY setup, for which the inter-channel correlation is higher due to the lack of spatial separation between the microphones. Point source reproduction (“PS”) yields the highest RMSE values for both the IC and the ILD. For the PSD, the values are only marginally lower than those for “Hom [Ch. 6] 1dec” and “Hom [116].” In general, all considered processing methods demonstrate either improved or comparable performance for the ORTF recording setup relative to the XY setup.

Figures A.3 and A.4 of the appendix depict equivalent objective evaluation plots for the sparse applause and the speech signal, respectively. The objective evaluation results for both signals are generally comparable to the discussed results for the dense applause signal. However, it is notable that for “Hom [Ch. 6] 2dec,” the RMSE values of the PSD are considerably higher for the sparse applause signal compared to the two other signal types. This is predominantly attributed to the employed decorrelation filters, given that the decorrelation of transient signals is known to be particularly challenging. Furthermore, of all processing methods, “Hom [Ch. 6] 2dec” is inherently most affected by decorrelation.

7.3.4 Perceptual Evaluation

A formal listening test was conducted to perform a perceptual evaluation of the proposed method. In the listening test, the different processing methods were compared to the binaural reference in terms of both spatial impression and overall signal quality. All items used in the listening test are available at www.audiolabs-erlangen.de/resources/2024-ICASSP-SESS-Heterogeneous.

7.3.4.1 Listening Test Setup

The speech signal, the sparse applause signal, and the dense applause signal were all included in the listening test. Since the considered processing methods all perform either better or comparably well for the ORTF recording setup compared to the XY setup in terms of the objective metrics evaluated in Section 7.3.3, only the ORTF recording setup was considered in the listening test. For all three source signals, four horizontal extent ranges were included, examining all combinations of $\bar{\varphi} \in \{0^\circ, 60^\circ\}$ and $\Delta\varphi \in \{60^\circ, 120^\circ\}$. Consequently, the listening test comprised a total of 12 test items.

The conditions under test included the proposed method (“Proposed”), point source reproduction (“PS”), homogeneous extent rendering using [116] (“Hom [116]”), and homogeneous extent rendering

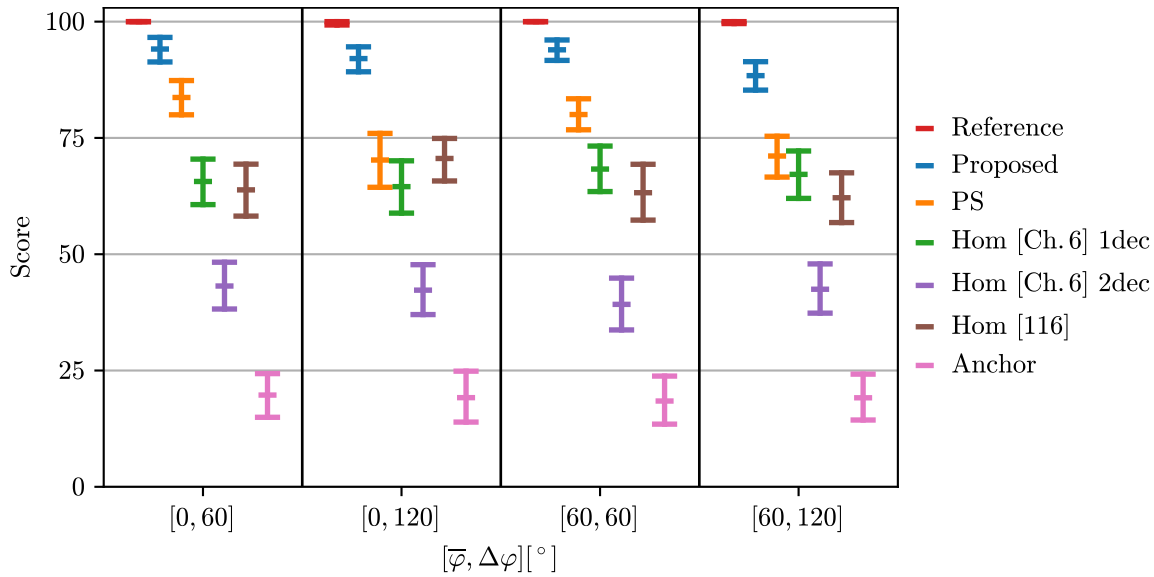


Figure 7.7: Mean values and bootstrapped 95% CIs of listening test scores per extent range.

based on Chapter 6 using either one (“Hom [Ch. 6] 1dec”) or two (“Hom [Ch. 6] 2dec”) decorrelation filters per extent range. Additionally, a hidden binaural reference (“Reference”) and a lower anchor (“Anchor”) were included. The lower anchor was generated by reproducing both channels of the input signal as point sources at the center of the extent, followed by a 3.5 kHz low-pass filter.

The participants were asked to rate the presented conditions w.r.t. the binaural reference in terms of spatial impression and overall signal quality on a continuous scale ranging from 0 (bad) to 100 (excellent). The items were reproduced over headphones using a customized version of the webMUSHRA software [130]. A total of 12 subjects participated in the listening test, eight male and four female. The average age of the participants was 32 years, and all of them had prior listening test experience.

7.3.4.2 Results

Figure 7.7 shows the mean values and bootstrapped 95% confidence intervals (CIs) of the listening test scores per extent range. For all extent ranges, the proposed method significantly outperforms all other considered processing methods. For the proposed method, the mean values of the listening test scores furthermore only depend slightly on the extent range. For most extent ranges, the second-best performing method is point source reproduction (“PS”). For point source reproduction, the performance is significantly worse for the larger extents with $\Delta\varphi = 120^\circ$ than for the smaller extents with $\Delta\varphi = 60^\circ$, which makes intuitively sense since the gap between the two rendered point sources is increased. This indicates that the proposed method offers a greater advantage over point source reproduction for larger extents. For point source reproduction of the smaller extents with $\Delta\varphi = 60^\circ$, a trend of decreased performance can furthermore be seen for the extent centered around $\bar{\varphi} = 60^\circ$, compared to the extent centered around $\bar{\varphi} = 0^\circ$. Although “Hom [Ch. 6] 2dec” performed second-best in the objective evaluation as discussed in Section 7.3.3, except for the anchor, the lowest listening test scores are obtained. The main reason for this is presumably the presence of significant decorrelation artifacts. The two other homogeneous extent rendering variants, “Hom [Ch. 6] 1dec” and “Hom [116],” mostly exhibit a similar performance. For all considered extent ranges, they perform either comparably well or significantly

worse than point source reproduction, which is preferred since it is significantly less complex.

7.4 Summary

In this chapter, we proposed a method for binaural rendering of heterogeneous SESSs using a two-channel input signal. Based on a model of an incoherently extended sound source with position-dependent energy and spectral content, a binaural output signal with the desired inter-channel properties is obtained using the covariance domain framework proposed in [117]. By constraining the solution to be as close as possible to the point source reproduction of the individual input channels, it is ensured that the directional characteristics encoded in the two-channel input signal are preserved. The proposed method was evaluated both objectively and perceptually using simulated input signals, comparing it to the two homogeneous extent rendering baselines proposed in Chapter 6 and [116] as well as to point source reproduction of the individual input channels. The objective evaluation showed that the proposed method matches the simulated binaural reference best in terms of IC, ILD, and PSD. A formal listening test furthermore showed that the proposed method comes closest to the binaural reference in terms of spatial impression and overall signal quality.

Future work could evaluate the proposed method considering more than two input channels as well as taking distance attenuation into account. The former can be particularly beneficial for sources whose direction-dependent energy is not approximated well by linear interpolation, as it can improve the degree of agreement between the rendering model and the source's characteristics. It would furthermore be possible to adapt the proposed method to loudspeaker reproduction by replacing the HRTFs with vector base amplitude panning gains [155], similar to what has been discussed in [116].

CHAPTER 8

Application of GAN-based Audio Decorrelation to Binaural Rendering of Homogeneously Extended Sound Sources

In the preceding chapters, the proposed methods for neural network-based audio decorrelation and binaural rendering of spatially extended sound sources (SESSs) have been evaluated independently. Given that the considered methods for binaural rendering of SESSs rely on audio decorrelation techniques, it is of particular interest to assess the performance of the proposed generative adversarial network (GAN)-based audio decorrelation approach for this specific application. For the homogeneous SESS rendering method proposed in Chapter 6, the amount of introduced decorrelated signal energy is independent of the single-channel input signal. In contrast, the heterogeneous SESS rendering method proposed in Chapter 7 takes a two-channel input signal and can exploit the inherent decorrelation between the input signal channels, necessitating the introduction of additional decorrelated signal energy only when the channels of the input signal are not sufficiently uncorrelated. The impact of decorrelation on the resulting binaural output signal is thus considerably more pronounced for the homogeneous SESS rendering method than for the heterogeneous SESS rendering method. Consequently, the homogeneous SESS rendering method, which employs a single-channel decorrelator, is of particular relevance for the considered evaluation.

Accordingly, this chapter investigates the suitability of the single-channel GAN-based audio decorrelation method introduced in Chapter 4 for the homogeneous SESS rendering method introduced in Chapter 6. To improve the overall audio quality of the decorrelated stereo signal, which serves as the basis for the homogeneous SESS rendering method, we propose an extended version of the decorrelator's generator loss given by (4.1). Specifically, an additional loss term is introduced, which aims to minimize spectral magnitude differences between the channels of the decorrelated stereo signal. The influence of the employed decorrelation technique on the performance of the homogeneous SESS rendering method is evaluated both objectively and perceptually. In addition to the proposed GAN-based audio decorrelation method, two classical signal processing-based audio decorrelation methods are considered during the evaluation.

The remainder of this chapter is structured as follows. In Section 8.1, the considered configuration of the single-channel GAN-based audio decorrelation method is described, including the proposed modifications to the generator loss. Section 8.2 describes the performance evaluation of the homogeneous SESS rendering method for the considered decorrelation techniques. Finally, Section 8.3 concludes this chapter by providing a summary.

8.1 Single-Channel GAN-Based Audio Decorrelation

8.1.1 Overview

Given a single-channel input signal $x(n)$, with discrete time index n , a single-channel decorrelated output signal $y(n)$ is obtained using the GAN-based audio decorrelation method described in Section 4.2. By mixing $y(n)$ with $x(n)$, a decorrelated stereo signal is derived according to (6.17), which serves as the basis for the homogeneous SESS rendering method. The generator and discriminator architecture as well as the discriminator loss of the GAN-based audio decorrelation method are chosen identical to Section 4.2. To improve the overall audio quality of the decorrelated stereo signal, an extended version of the generator loss is proposed. The following section provides a detailed description of the employed extended generator loss.

8.1.2 Extended Generator Loss

As described in Section 6.2, the homogeneous SESS rendering method uses the decorrelated stereo signal $\mathbf{x}_d(n) = [x_{d,1}(n), x_{d,2}(n)]^T$ as the basis for the processing, where $(\cdot)^T$ denotes the transpose operation. Given the input signal $x(n)$ and the decorrelated output signal $y(n)$, the two channels of $\mathbf{x}_d(n)$ are obtained as follows according to (6.17):

$$\begin{aligned} x_{d,1}(n) &= \frac{1}{\sqrt{2}}(x(n) + y(n)), \\ x_{d,2}(n) &= \frac{1}{\sqrt{2}}(x(n) - y(n)). \end{aligned} \quad (8.1)$$

The objective of the original generator loss, as defined in (4.1), is to regulate the correlation between $x(n)$ and $y(n)$, as well as the audio quality of $y(n)$. However, the properties of the decorrelated stereo signal $\mathbf{x}_d(n)$ are not directly controlled. As discussed in Section 4.4.3, time-frequency-dependent complementary constructive and destructive interference may occur in the stereo signal channels in (8.1), resulting in a degradation of the audio quality. To reduce the observed signal artifacts, the mel-spectrogram stereo loss $\mathcal{L}_{\text{mel,st}}$ is introduced, which aims to minimize spectral magnitude differences between the two channels of the stereo signal.

Analogously to the mel-spectrogram loss \mathcal{L}_{mel} (4.3), the mel-spectrogram stereo loss $\mathcal{L}_{\text{mel,st}}$ is defined as the ℓ_1 -norm between the mel-spectrogram of $x_{d,1}(n)$ and the mel-spectrogram of $x_{d,2}(n)$, both expressed in dB:

$$\mathcal{L}_{\text{mel,st}} = \frac{1}{ML} \sum_{m=1}^M \sum_{l=1}^L \left| 10 \log_{10} \left(\frac{\mathcal{T}_m(|X_{d,1}(k,l)|^2)}{\mathcal{T}_m(|X_{d,2}(k,l)|^2)} \right) \right|, \quad (8.2)$$

where $X_{d,1}(k,l)$ and $X_{d,2}(k,l)$ are obtained by transforming $x_{d,1}(n)$ and $x_{d,2}(n)$ to the time-frequency domain using a short-time Fourier transform (STFT), with k and l denoting the discrete frequency and time frame indices, respectively. The number of STFT time frames is denoted by L . As for the coherence loss \mathcal{L}_{coh} and the mel-spectrogram loss \mathcal{L}_{mel} (see Sections 4.2.3.1 and 4.2.3.2), an STFT frame length of 1024 samples and a hop size of 256 samples are used at a sampling rate of 22.05 kHz. The operator $\mathcal{T}_m(\cdot)$ denotes the transformation from linear to mel frequency scale, where m denotes the

mel-scale frequency index. The number of mel frequency bands M is set to 80. The mel-spectrogram stereo loss $\mathcal{L}_{\text{mel,st}}$ takes values in the range $[0 \text{ dB}, \infty)$.

The extended generator loss $\mathcal{L}_{\text{G,ext}}$ is obtained by adding a weighted version of $\mathcal{L}_{\text{mel,st}}$ to the original generator loss \mathcal{L}_{G} given by (4.1):

$$\mathcal{L}_{\text{G,ext}} = \lambda_{\text{coh}}\mathcal{L}_{\text{coh}} + \lambda_{\text{mel}}\mathcal{L}_{\text{mel}} + \lambda_{\text{mel,st}}\mathcal{L}_{\text{mel,st}} + \mathcal{L}_{\text{adv}}, \quad (8.3)$$

where $\lambda_{\text{mel,st}}$ denotes the mel-spectrogram stereo loss weight. The loss terms \mathcal{L}_{coh} , \mathcal{L}_{mel} , and \mathcal{L}_{adv} , as well as the loss weights λ_{coh} and λ_{mel} , are defined in Section 4.2.3.

8.2 Performance Evaluation

The aim of this chapter is to assess the suitability of the single-channel GAN-based audio decorrelation method proposed in Chapter 4 for the binaural homogeneous SESS rendering method introduced in Chapter 6. Therefore, the influence of the employed decorrelation method on the performance of the homogeneous SESS rendering method was evaluated both objectively and perceptually. Two variants of the single-channel GAN-based audio decorrelation method were considered: one including and one excluding the mel-spectrogram stereo loss introduced in Section 8.1. Additionally, two classical signal processing-based audio decorrelation methods were included in the evaluation.

8.2.1 Evaluation Setup

Given a time-domain single-channel input signal $x(n)$ as well as a desired spatial extent range, a binaural output signal representing the homogeneous SESS was obtained using the time-domain implementation of the homogeneous SESS rendering method described in Section 6.2.2. The phase smoothing discussed in 6.2.2.2 was not applied in the present study. Instead, the phase responses of the head-related transfer functions (HRTFs) corresponding to the extent center were used directly as the phase responses of the processing filters given by (6.19). This approach was selected on the basis of informal listening, which revealed that the phase smoothing technique may result in an undesirable localization offset for certain signal types and extent ranges.

Because of the greater perceptual relevance of horizontal spatial extent compared to vertical spatial extent, as discussed in Section 2.3.1.4, only extents in azimuth direction at $\theta = 0^\circ$ elevation were included in the evaluation. Accordingly, the considered spatial extent is expressed in terms of the azimuth angle range $[\varphi_1, \varphi_2] = \left[\bar{\varphi} - \frac{\Delta\varphi}{2}, \bar{\varphi} + \frac{\Delta\varphi}{2} \right]$, where $\bar{\varphi}$ describes the extent center, and $\Delta\varphi$ describes the extent in terms of azimuth angle.

A total of six source signals were considered during the evaluation, partly taken from the EBU SQAM CD [128] and the FSD50K dataset [127]. Four music signals were included: two pop songs (denoted by music1 and music2), one particularly transient music signal (castanets), and one particularly tonal music signal (violin). Additionally, one applause and one ocean waves signal were included. Of the source signals considered, the music1 and music2 signals are most similar to the training data used for the GAN-based audio decorrelation method. Note that the source signals employed in this study represent a subset of those used for the perceptual evaluation discussed in Section 4.4.2.

Throughout the evaluation, the neutral head orientation head-related impulse responses (HRIRs) of the FABIAN HRTF dataset [78, 152] were employed. All HRIRs have a length of 256 samples at a

sampling rate of 44.1 kHz. A diffuse field equalization was conducted using the provided equalization filter. Although the homogeneous SESS rendering method performed processing at the sampling rate of the HRIRs (i.e., 44.1 kHz), its input signals were bandlimited to an upper frequency of 11.025 kHz. This was done to ensure a fair comparison with the single-channel GAN-based audio decorrelation method, which operates at a sampling rate of 22.05 kHz.

8.2.2 Considered Decorrelation Methods

Two variants of the proposed single-channel GAN-based audio decorrelation method were considered during the evaluation, both using the extended generator loss defined in (8.3). The first variant did not incorporate the mel-spectrogram stereo loss, i.e., $\lambda_{\text{mel,st}} = 0$. In contrast, the second variant did include the mel-spectrogram stereo loss, with a mel-spectrogram stereo loss weight of $\lambda_{\text{mel,st}} = 0.15$. This value for $\lambda_{\text{mel,st}}$ was selected empirically to ensure a sufficient reduction of the observed stereo signal artifacts. As for the independent evaluation of the single-channel GAN-based audio decorrelation method discussed in Section 4.4, a coherence loss weight of $\lambda_{\text{coh}} = 2.5$ and a mel-spectrogram loss weight of $\lambda_{\text{mel}} = 0.65$ were used. Both networks were trained on the MUSDB18-HQ music dataset [123], using the training configuration described in Section 4.3. Note that the network with $\lambda_{\text{mel,st}} = 0$ is identical to the network used for the independent evaluation described in Section 4.4.

In addition to the proposed single-channel GAN-based audio decorrelation method, two classical signal processing-based decorrelators were considered during the evaluation. As the first comparison method, the state-of-the-art decorrelation technique described in [93] was used, which is based on Schroeder all-pass filters and frequency-dependent delays operating in the STFT domain. This method will be referred to as the MPEG-I decorrelator. The transient detection described in [93] was not included in the present study, and a slightly different parameterization of the all-pass filters and frequency-dependent delays was used. Specifically, the parameterization described in Section 3.2.1 was used. As the second comparison method, we used the quadrature mirror filter (QMF)-domain decorrelator employed in [90], which applies frequency-dependent pseudo-random delays to the input signal and will thus be referred to as the delay decorrelator. The applied delays were varied between approximately 20 and 80 ms. In contrast to [90], no onset suppression was applied to ensure a fair comparison between the different decorrelation methods.

8.2.3 Objective Evaluation

Three perceptually relevant objective metrics were calculated from the binaural output signals for the different decorrelation methods: the interaural coherence (IC), the interaural level difference (ILD), and the power spectral density (PSD). Given a time-frequency-domain binaural output signal $\mathbf{y}(k, l) = [Y_1(k, l), Y_r(k, l)]^T$, the objective metrics are defined as follows:

$$\widehat{\text{IC}}(k, l) = \frac{|\mathcal{E}\{Y_1(k, l)Y_r^*(k, l)\}|}{\sqrt{\mathcal{E}\{|Y_1(k, l)|^2\}\mathcal{E}\{|Y_r(k, l)|^2\}}}, \quad (8.4)$$

$$\widehat{\text{ILD}}(k, l) = 10 \log_{10} \left(\frac{\mathcal{E}\{|Y_1(k, l)|^2\}}{\mathcal{E}\{|Y_r(k, l)|^2\}} \right), \quad (8.5)$$

$$\widehat{\text{PSD}}(k, l) = 10 \log_{10} (\mathcal{E}\{|Y_1(k, l)|^2\} + \mathcal{E}\{|Y_r(k, l)|^2\}), \quad (8.6)$$

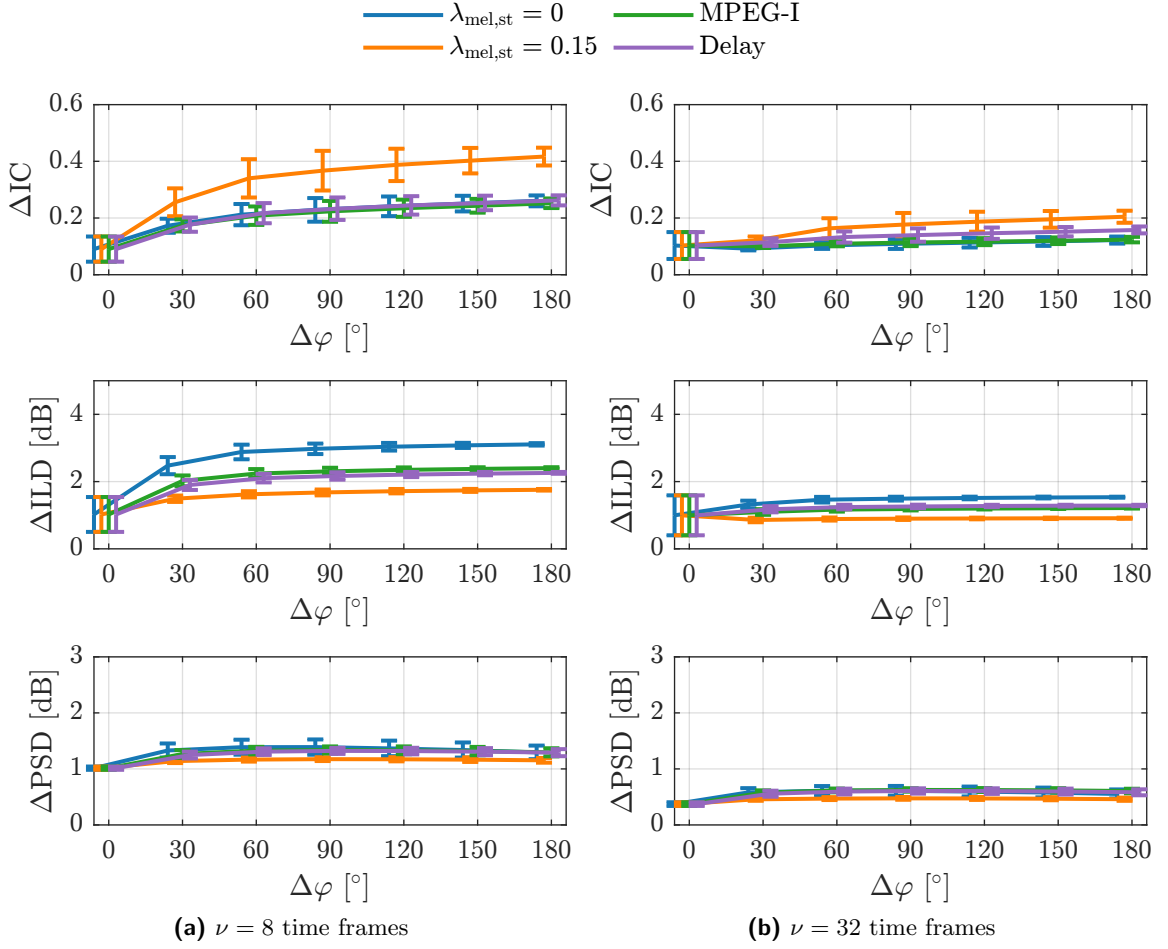


Figure 8.1: Ocean waves signal: Mean and standard deviation of RMSE between objective metrics of homogeneous SESS rendering output and their target values according to the rendering model, for different decorrelation methods. The mean and standard deviation of the RMSE values were calculated depending on the extent $\Delta\varphi$, averaged over all considered values for the center $\bar{\varphi}$ of the SESS.

where $\mathcal{E}\{\cdot\}$ denotes the statistical expectation. To compute the objective metrics, the time-domain binaural output signals were transformed to the time-frequency domain using an STFT with a frame length of 256 samples and a hop size of 128 samples at a sampling rate of 44.1 kHz. The statistical expectation was approximated by means of a moving average filter of length ν , considering two values for the averaging length ν : 8 time frames (≈ 23 ms) and 32 time frames (≈ 93 ms).

The objective metrics of the binaural output signals were compared to their target values according to the rendering model described in Section 6.1. The target IC was calculated as defined in (6.4). The target ILD and PSD were calculated based on the target values for the left and right ear gains $G_l(k)$ and $G_r(k)$, given by (6.6), as follows:

$$\text{ILD}(k) = 10 \log_{10} \left(\frac{|G_l(k)|^2}{|G_r(k)|^2} \right), \quad (8.7)$$

$$\text{PSD}(k, l) = 10 \log_{10} (|G_l(k)|^2 \mathcal{E}\{|X(k, l)|^2\} + |G_r(k)|^2 \mathcal{E}\{|X(k, l)|^2\}), \quad (8.8)$$

where $X(k, l)$ was obtained by transforming the input signal $x(n)$ to the time-frequency domain, again

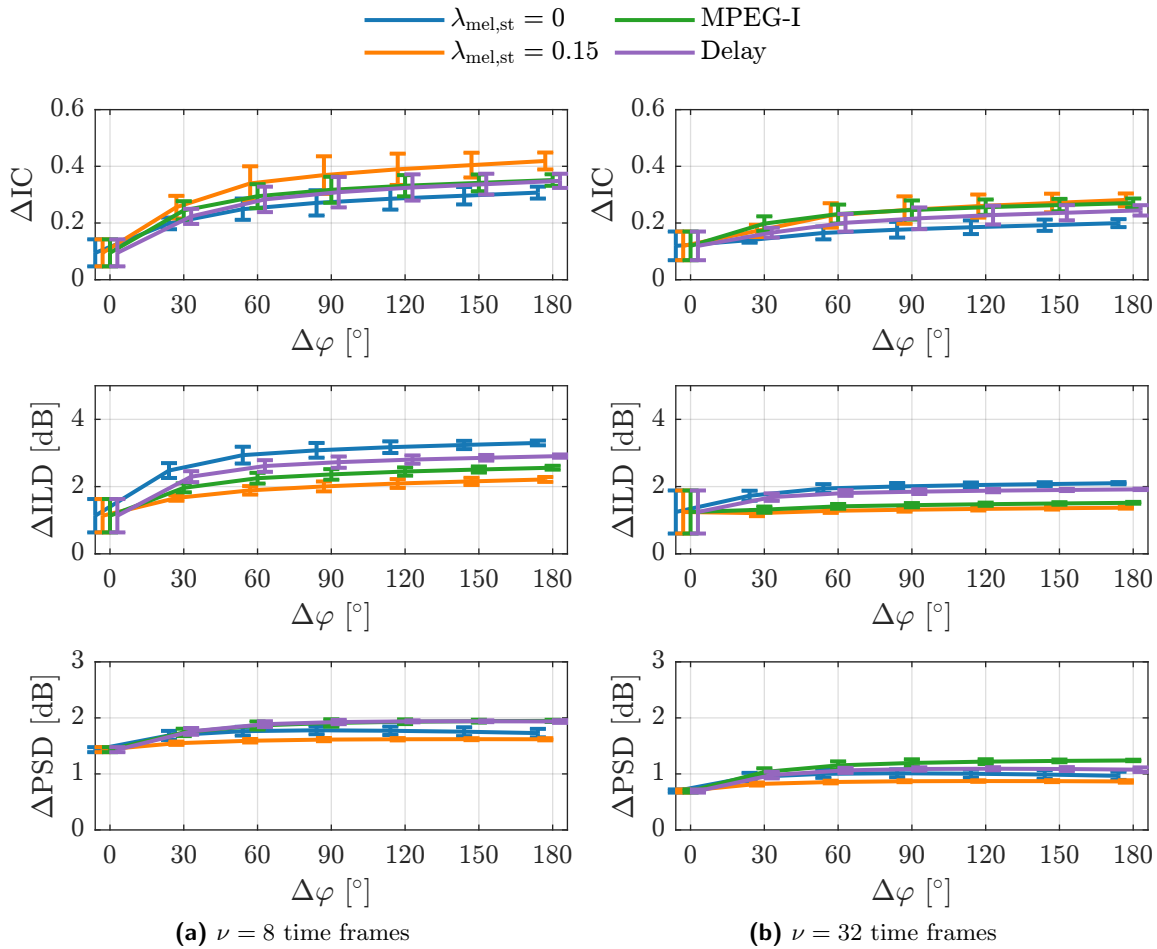


Figure 8.2: Music1 signal: Mean and standard deviation of RMSE between objective metrics of homogeneous SESS rendering output and their target values according to the rendering model, for different decorrelation methods. The mean and standard deviation of the RMSE values were calculated depending on the extent $\Delta\varphi$, averaged over all considered values for the center $\bar{\varphi}$ of the SESS.

using an STFT with a frame length of 256 samples and a hop size of 128 samples. The statistical expectation was again approximated by means of a moving average filter. To ensure a fair comparison, the moving average length was set to the same value used to compute the objective metrics of the binaural output signals.

A variety of extents $\Delta\varphi \in \{0^\circ, 30^\circ, 60^\circ, 90^\circ, 120^\circ, 150^\circ, 180^\circ\}$ was considered for the objective evaluation, using all 180 HRTF dataset positions in the horizontal plane as center $\bar{\varphi}$ of the SESS. For each extent range, the root-mean-square error (RMSE) between the objective metrics of the binaural output signals and their target values according to the rendering model was calculated by averaging over the time and frequency dimensions. Time-frequency bins where the input signal power was more than 40 dB below the peak value were excluded from the RMSE calculation. Only the ocean waves and the music1 pop song signal were considered for the objective evaluation. The music1 pop song signal was selected due to its similarity to the training data used for the GAN-based audio decorrelation method. The ocean waves signal was additionally included as it demonstrated the most pronounced differences between the decorrelation methods.

Figures 8.1 and 8.2 depict the mean and standard deviation of the RMSE for each objective metric

and decorrelation method as a function of the extent $\Delta\varphi$, averaged over all considered values for the center $\bar{\varphi}$ of the SESS, for the ocean waves and the music1 pop song signal, respectively. For both signals and all decorrelation methods, the RMSE values of the objective metrics are lower for $\nu = 32$ than for $\nu = 8$, i.e., when using a longer averaging length to approximate the statistical expectation. Additionally, the relative differences between the different decorrelation methods are reduced when a longer averaging length is employed. For the ocean waves signal and $\nu = 8$ in particular, the $\lambda_{\text{mel,st}} = 0$ condition results in considerably higher RMSE values for the ILD relative to the other decorrelation methods. This finding aligns with the results reported in Section 4.4.3, which showed that the spectral magnitude differences between the channels of the decorrelated stereo signal, which serves as the basis for the homogeneous SESS rendering method, are comparably large for this decorrelation method. By including the mel-spectrogram stereo loss, the $\lambda_{\text{mel,st}} = 0.15$ condition succeeds in reducing the RMSE values for the ILD. However, this is accompanied by an increase in the RMSE values for the IC. A similar trend can be observed for the music1 pop song signal, with less pronounced differences between the considered decorrelation methods compared to the ocean waves signal.

8.2.4 Perceptual Evaluation

Two formal listening tests were conducted to perceptually evaluate the influence of the employed decorrelation technique on the performance of the homogeneous SESS rendering method. The first test evaluated the overall audio quality by comparing the binaural output signals to a point source reproduced at the extent center. The second test evaluated the perceived spatial extent of the binaural output signals relative to a visual representation of the considered target extent range. Note that no ideal binaural reference representing the homogeneous SESS can be defined. All items used in the listening tests are available at www.audiolabs-erlangen.de/resources/2024-SESS-DDD.

For all six source signals, two horizontal extent ranges were included, with $\Delta\varphi \in \{60^\circ, 120^\circ\}$ and $\bar{\varphi} = 0^\circ$. Consequently, each listening test comprised a total of 12 test items. The items were reproduced over headphones using a customized version of the webMUSHRA software [130]. A total of 12 subjects participated in each listening test, 10 male and two female. The average age of the participants was 31 years, and all of them had prior listening test experience.

8.2.4.1 Quality Test

For the quality test, the original input signal reproduced as a point source at the extent center ($\varphi = 0^\circ$) was provided as reference. The conditions under test included the binaural output signals of the homogeneous SESS rendering method for the four considered decorrelation techniques described in Section 8.2.2. Additionally, a hidden reference was included, and a 3.5 kHz low-pass filtered version of the reference was provided as lower anchor. The participants were asked to rate the presented conditions in terms of overall audio quality w.r.t. the reference on a continuous scale ranging from 0 (bad) to 100 (excellent). The listeners were instructed to focus exclusively on the audio quality, disregarding any differences related to the spatial impression.

Figure 8.3 shows the mean values and bootstrapped 95% confidence intervals (CIs) of the listening test scores per source signal and per extent range. For all considered decorrelation methods, the performance is only mildly dependent on the extent range. Conversely, a clear dependency of the performance on the considered source signal can be observed. With the exception of the castanets

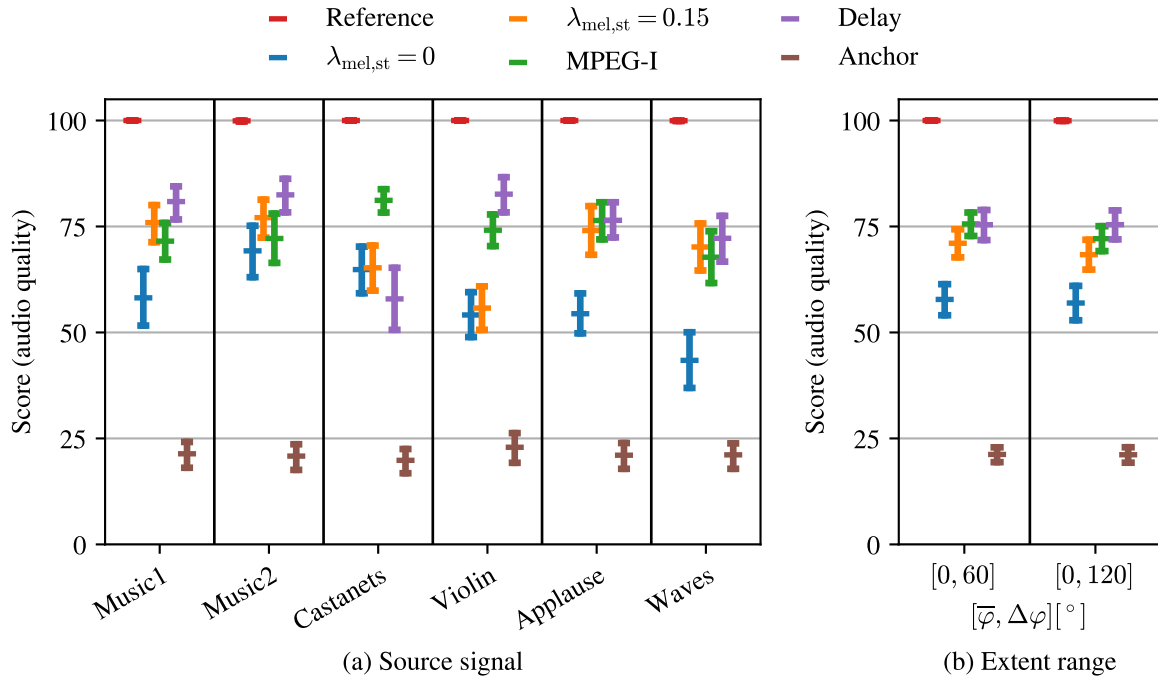


Figure 8.3: Mean and bootstrapped 95% CIs of listening test scores aggregated per source signal and per extent range, respectively, in terms of overall audio quality.

and violin signals, the $\lambda_{\text{mel,st}} = 0.15$ condition reaches a comparable performance to the MPEG-I and delay decorrelators. Moreover, the $\lambda_{\text{mel,st}} = 0.15$ condition significantly outperforms the $\lambda_{\text{mel,st}} = 0$ condition for the music1, applause, and ocean waves signals. By including the mel-spectrogram stereo loss, the $\lambda_{\text{mel,st}} = 0.15$ condition thus succeeded in improving the overall audio quality for certain signals, particularly those that more densely cover the time-frequency spectrum. It is noteworthy that the observed results differ significantly from those obtained in the independent evaluation of the mono signal quality discussed in Section 4.4.2.2, where the $\lambda_{\text{mel,st}} = 0$ condition mostly outperformed the MPEG-I decorrelator. For the delay decorrelator, a similar effect can be observed when comparing the present results with those of the listening experiment discussed in Section 5.4.4.1, which considered a multi-channel variant of the delay decorrelator. Compared to the results presented in this section, the performance of the delay decorrelator in terms of overall audio quality was considerably worse. These observations show that the performance of decorrelation techniques can be highly application-dependent.

8.2.4.2 Spatial Extent Test

For the spatial extent test, no ideal reference could be defined. Instead, a visual representation of the considered target extent range was provided to the listeners alongside each item. Figure 8.4 depicts the provided visual representations for the two considered extent ranges. The conditions under test included the binaural output signals of the homogeneous SESS rendering method for the four considered decorrelation techniques described in Section 8.2.2. Additionally, the original input signal reproduced as a point source at the extent center ($\varphi = 0^\circ$) was provided as lower anchor. The participants were asked to rate the presented conditions in terms of the perceived spatial extent relative to the considered

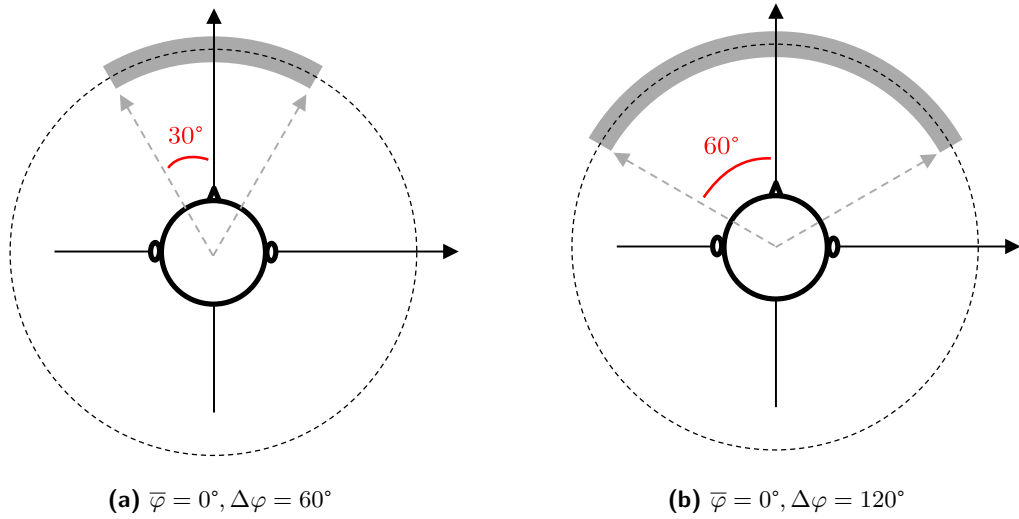


Figure 8.4: Visual representations of the considered target extent ranges provided to the listeners in the spatial extent listening test.

target extent range on a continuous scale ranging from -3 (“much smaller”) to 3 (“much larger”), using the following intermediate anchor points: -2 (“smaller”), -1 (“slightly smaller”), 0 (“the same as”), 1 (“slightly larger”), 2 (“larger”). The grading scale had a resolution of one decimal place. The listeners were instructed to focus exclusively on the perceived spatial extent, disregarding any differences related to the audio quality.

To provide the listeners with an understanding of the perceptual aspects related to the spatial extent of a sound source, three ideal examples were presented to the listeners prior to conducting the listening test, representing an SESS with a small ($\bar{\varphi} = 0^\circ, \Delta\varphi = 0^\circ$), medium ($\bar{\varphi} = 0^\circ, \Delta\varphi = 30^\circ$), and large ($\bar{\varphi} = 0^\circ, \Delta\varphi = 90^\circ$) spatial extent. The examples were generated by distributing a number of mutually incoherent white noise stimuli as point sources over the considered extent range according to the rendering model described in Section 6.1.

Figure 8.5 shows the mean values and bootstrapped 95 % CIs of the listening test scores per source signal and per extent range. In contrast to the quality test, the results show a clear dependency on the extent range. While the relative ratings between the conditions are similar for the two extent ranges, the smaller extent range with $\Delta\varphi = 60^\circ$ generally yields larger ratings, indicating a greater deviation from the visual representation. For both extent ranges, the perceived spatial extent is rated as being larger than the provided visual representation for most decorrelation methods. This may indicate that the employed rendering model does not fully align with the auditory perception. Nevertheless, compared to the lower anchor (i.e., point source reproduction), the homogeneous SESS rendering method achieves a good spatialization effect for all considered decorrelation methods.

Since the absolute ratings show a clear dependency on the extent range, the relative ratings w.r.t. the $\lambda_{\text{mel,st}} = 0.15$ condition are considered for further analysis. Figure 8.6 shows the mean values and bootstrapped 95 % CIs of the listening test difference scores w.r.t. the $\lambda_{\text{mel,st}} = 0.15$ condition, per source signal and per extent range. In contrast to the absolute scores, the difference scores are relatively independent of the extent range. Depending on the source signal, the MPEG-I and $\lambda_{\text{mel,st}} = 0$ conditions are rated similarly or slightly higher than the $\lambda_{\text{mel,st}} = 0.15$ condition. This indicates that

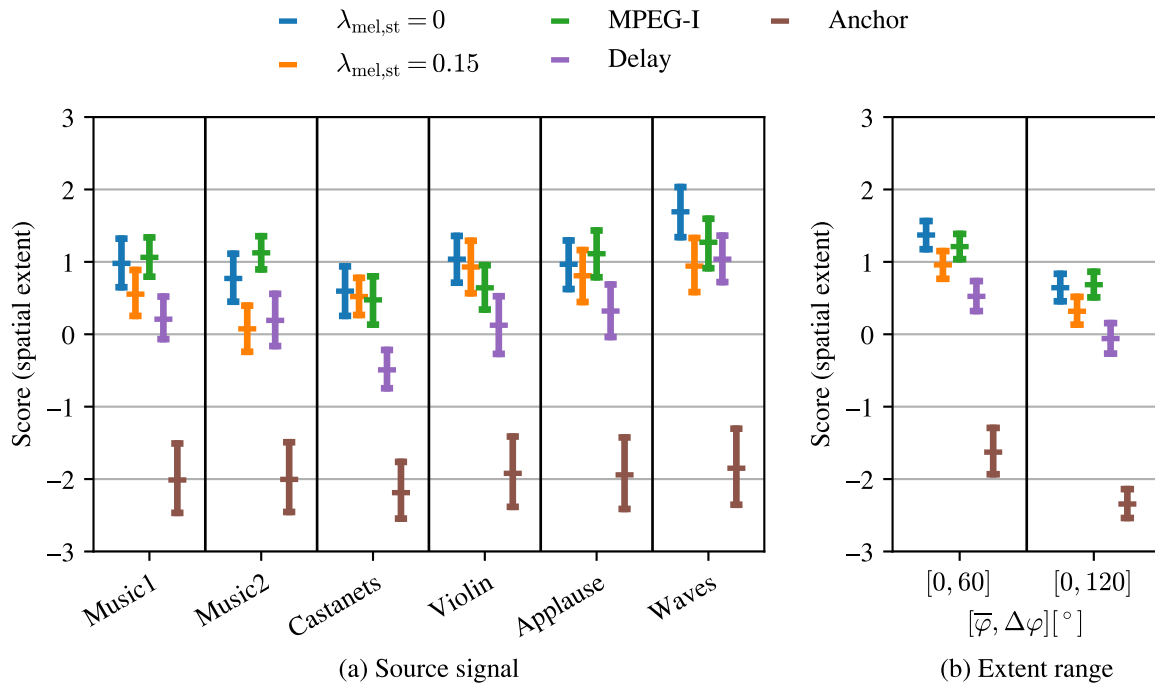


Figure 8.5: Mean and bootstrapped 95 % CIs of listening test scores aggregated per source signal and per extent range, respectively, in terms of perceived spatial extent rated relative to a visual representation of the considered target extent range.

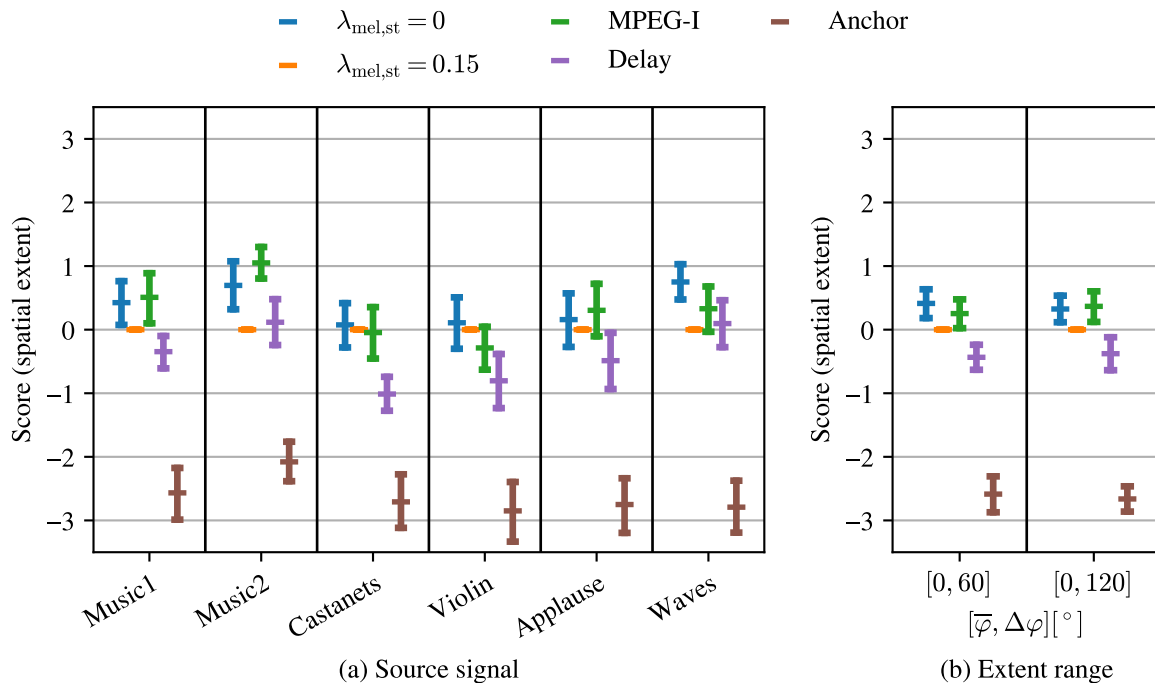


Figure 8.6: Mean and bootstrapped 95 % CIs of listening test difference scores w.r.t. $\lambda_{mel,st} = 0.15$ aggregated per source signal and per extent range, respectively, in terms of perceived spatial extent rated relative to a visual representation of the considered target extent range.

decorrelation with the MPEG-I and $\lambda_{\text{mel,st}} = 0$ decorrelators is more effective. This finding aligns with the objective evaluation results discussed in Section 8.2.3, which showed that the $\lambda_{\text{mel,st}} = 0.15$ condition results in a larger IC error for the ocean waves and music1 signals. The delay condition is rated similarly or slightly lower than the $\lambda_{\text{mel,st}} = 0.15$ condition, depending on the source signal. This indicates that decorrelation with the delay decorrelator is less effective, which cannot be explained directly by the objective evaluation results.

After conducting the listening test, some of the participants have reported that they found it difficult to provide confident absolute judgments of the perceived spatial extent relative to the provided visual representation of the target extent range. While this should not significantly affect the relative ratings between the different conditions, the absolute ratings may be of limited reliability. More reliable absolute ratings may be obtained by conducting the evaluation in a more interactive setting and providing the listeners with a direct visual representation of the SESS, e.g., inside a virtual reality (VR) environment.

8.3 Summary

This chapter investigated the suitability of the single-channel GAN-based audio decorrelation method introduced in Chapter 4 for the homogeneous SESS rendering method introduced in Chapter 6. To improve the overall audio quality of the decorrelated stereo signal, which serves as the basis for the homogeneous SESS rendering method, an extended version of the generator loss was proposed. Specifically, the mel-spectrogram stereo loss was introduced, which aims to minimize spectral magnitude differences between the channels of the decorrelated stereo signal. The influence of the employed decorrelation technique on the performance of the homogeneous SESS rendering method was evaluated both objectively and perceptually. Two variants of the single-channel GAN-based audio decorrelation method were considered: one including and one excluding the newly proposed mel-spectrogram stereo loss. Additionally, a comparison was performed with two classical signal processing-based audio decorrelation methods.

The objective evaluation showed that including the mel-spectrogram stereo loss in the single-channel GAN-based audio decorrelation method leads to an improved agreement between the output of the homogeneous SESS rendering method and the rendering model in terms of ILD. However, this is accompanied by an increased mismatch in terms of IC. These findings align with the perceptual evaluation results, which showed that including the mel-spectrogram stereo loss results in improved performance in terms of overall audio quality, while the perceived spatial extent was rated as being slightly smaller. Except for the castanets and violin signals, the condition including the mel-spectrogram stereo loss reached a similar performance as the two classical signal processing-based audio decorrelation methods in terms of overall audio quality. For one of the two classical signal processing-based audio decorrelation methods, the perceived spatial extent was rated as being similar or slightly larger compared to the condition including the mel-spectrogram stereo loss, depending on the source signal. In contrast, for the other method, the perceived spatial extent was rated as being similar or slightly smaller compared to the condition including the mel-spectrogram stereo loss. The perceived spatial extent was mostly rated as being larger than the provided visual representation. While further investigations in a more interactive setting are required to verify the validity of the absolute ratings for the perceived spatial extent, this may indicate that the employed rendering model does not fully align with the auditory perception.

CHAPTER 9

Conclusions and Outlook

In this thesis, neural network-based approaches to audio decorrelation have been introduced. Additionally, methods for binaural rendering of spatially extended sound sources (SESSs) have been proposed, which employ audio decorrelation techniques. The presented tools and methods are highly relevant for various spatial audio rendering applications, such as auralization of virtual or augmented reality (VR/AR) environments, contributing toward the overarching goal of providing the user with an immersive and plausible listening experience. In this chapter, the main contributions of this thesis are summarized, and possible future research directions are highlighted.

9.1 Conclusions

The first part of this thesis investigated, for the first time, the use of neural networks for the task of audio decorrelation. As a preliminary step, a proof of concept was provided in Chapter 3. A convolutional neural network (CNN) architecture was proposed, and trained to mimic the behavior of a state-of-the-art reference decorrelator, focusing on music and applause signals. A formal listening test demonstrated that the proposed neural network is perceptually equivalent to the reference decorrelator in terms of their stereo output signals.

Building on the proof of concept presented in Chapter 3, Chapter 4 proposed a reference-free approach to audio decorrelation based on generative adversarial networks (GANs), which provides a single-channel output signal. For the generator network, the CNN architecture introduced in Chapter 3 was employed. Instead of relying on a decorrelated reference signal, the training objective was defined directly w.r.t. the audio input signal. The generator loss consisted of a number of individual loss terms to control both the input-output correlation and the output signal quality, allowing the training procedure to be specifically tailored to the desired output signal properties. The model was trained on music signals, selecting a specific weighting of the respective loss terms that offers a reasonable tradeoff between the input-output correlation and the output signal quality. Two formal listening tests demonstrated that the proposed GAN-based audio decorrelation method significantly outperforms a state-of-the-art classical signal processing-based decorrelator in terms of mono output signal quality for the majority of considered items, while its stereo output signal is perceived as only “slightly less enveloped.”

In Chapter 5, the single-channel GAN-based audio decorrelation method introduced in Chapter 4

was extended to provide a multi-channel output signal. A separate generator network was used for each output channel. All generator networks were optimized jointly based on a number of individual loss terms, which control the inter-channel and input-output correlation, as well as the output signal quality. The proposed model was trained on music signals, and a number of experiments were conducted based on objective evaluations and formal listening tests considering reproduction over a four-channel loudspeaker setup. It was demonstrated that joint optimization of the generator networks is essential in reaching the optimal tradeoff between perceived envelopment and overall audio quality. Moreover, it was shown that the coherence loss weight represents a key parameter of the proposed method to control the existing tradeoff between the degree of decorrelation and the output signal quality. Using piano signals as an example, it was demonstrated that explicitly including certain training data can be beneficial to improve the overall audio quality of the proposed method's output for specific signal types. The proposed method only succeeded in reaching the perceptual performance of some of the considered classical signal processing-based comparison methods. The objective evaluations showed that the mel-spectrogram loss used for training is not well suited to predict the overall audio quality of decorrelation methods in general. Thus, even though the proposed GAN-based audio decorrelation method did not outperform the considered classical signal processing-based comparison methods in terms of perceptual performance, the availability of better-suited objective quality metrics is expected to be a key component in enhancing its performance.

The second part of this thesis considered methods for binaural rendering of SESSs that are based on audio decorrelation methods. Chapter 6 proposed a method for efficient and realistic binaural rendering of homogeneous SESSs given a single-channel input signal, i.e., targeting SESSs that emit sound with constant radiation characteristics over their spatial extent. The homogeneous SESS was modeled as an incoherently extended sound source with position-independent energy and spectral content. Based on this model, a set of perceptually relevant target auditory cues was determined. A binaural output signal with the desired target auditory cues was then synthesized by mixing two decorrelated input signals, which were obtained from the single-channel input signal using a single decorrelation filter. A formal listening test showed that the proposed method's output is perceptually nearly indistinguishable from the binaural output signal resulting from a direct implementation of the rendering model. Compared to a direct implementation of the rendering model, the proposed homogeneous SESS rendering method comes with the advantage of reduced computational complexity and relaxed requirements for the employed decorrelation filters.

Chapter 7 proposed a binaural rendering method particularly suited for heterogeneous SESSs, i.e., targeting SESSs that exhibit a position-dependent radiation behavior. Accordingly, the rendering model used in Chapter 6 was extended by taking the position-dependent energy of the SESS into account. Input to the rendering method is a two-channel input signal. The two channels represent the left and right parts of the SESS, respectively, and can be obtained by recording the SESS using directional microphones. Based on the aforementioned extended rendering model, a binaural output signal with the desired inter-channel properties is obtained by processing the two-channel input signal using an optimal mixing approach previously proposed in the literature. By constraining the solution to be as close as possible to point source reproduction of the individual input channels, it is ensured that the directional characteristics encoded in the two-channel input signal are preserved. The proposed method was evaluated based on simulated input signals, comparing it to point source reproduction of the individual input channels as well as to two homogeneous SESS rendering baselines, one of which

Table 9.1: Links to audio examples for the proposed neural network-based audio decorrelation and binaural SESS rendering methods, organized by chapter.

Chapter	Web page
3	www.audiolabs-erlangen.de/resources/2022-SPL-DD-Decorrelation
4	www.audiolabs-erlangen.de/resources/2023-WASPAA-Decorrelation-GAN
6	www.audiolabs-erlangen.de/resources/2022-JAES-SESS
7	www.audiolabs-erlangen.de/resources/2024-ICASSP-SESS-Heterogeneous
8	www.audiolabs-erlangen.de/resources/2024-SESS-DDD

was the method introduced in Chapter 6. A formal listening test demonstrated that the output of the proposed heterogeneous SESS rendering method comes closest to the simulated binaural reference signal in terms of spatial impression and overall signal quality.

In Chapter 8, the two parts of this thesis were combined by investigating the suitability of the single-channel GAN-based audio decorrelation method introduced in Chapter 3 for the binaural homogeneous SESS rendering method introduced in Chapter 6. To improve the overall audio quality of the decorrelated stereo signal, which serves as the basis for the homogeneous SESS rendering method, an extended version of the generator loss was proposed for the GAN-based audio decorrelation method. Specifically, the mel-spectrogram stereo loss was introduced, which aims to minimize spectral magnitude differences between the channels of the decorrelated stereo signal. Two formal listening tests were conducted to evaluate the impact of the employed decorrelation technique on the performance of the homogeneous SESS rendering method. The results demonstrated that the inclusion of the mel-spectrogram stereo loss in the GAN-based audio decorrelation method results in an improvement of the overall audio quality for noise-like signals in particular, while the perceived spatial extent is only slightly reduced.

Audio examples for the proposed neural network-based audio decorrelation and binaural SESS rendering methods can be accessed via the web pages listed in Table 9.1.

9.2 Future Research

Based on the conclusions provided in the previous section, several open questions and limitations of the methods developed within the scope of this thesis can be identified.

Neural Audio Decorrelation The proposed GAN-based audio decorrelation method demonstrated a rather signal-dependent performance. Similarly to signal processing-based approaches that perform signal-dependent processing, the loss function could be adapted in a signal-dependent manner to control the desired output signal properties per signal type. In order to realize this, a signal classification would be required during training, but not necessarily during inference.

The perceptual experiments conducted demonstrated that the performance of the considered decorrelation techniques can be highly dependent on the specific application. Consequently, it would be valuable to evaluate and optimize the proposed decorrelation methods considering further applications. Specifically, adaptations of the employed loss function based on application-specific requirements could be beneficial.

It was shown that the mel-spectrogram loss employed for the GAN-based audio decorrelation method

is not a very good predictor of the overall audio quality of multi-channel decorrelation methods in general. To improve the overall audio quality of the proposed GAN-based audio decorrelation method, it is essential to identify better-suited objective quality metrics that can be utilized for training. Objective quality metrics previously proposed in the literature could be considered. In particular, objective quality metrics based on perceptual models of human hearing may be promising. Potentially, the development of novel objective quality metrics specifically tailored to audio decorrelation methods could be beneficial. Any objective quality metric selected for use as a loss function must be differentiable and should possess a reasonable computational complexity.

Further room for improvement remains with regard to the computational complexity of the generator network. The employed generator network performs independent processing for each frequency bin. It is anticipated that the number of parameters, and thus the computational complexity, can be reduced by leveraging analogies across the frequency bins. Similarly, for the multi-channel variant, analogies across the channels could be exploited to reduce the computational complexity.

As an alternative to GANs, other generative models may be considered. These may include diffusion probabilistic models [156, 157], variational autoencoders [158], or normalizing flows [159].

Finally, it may be beneficial to consider alternative processing structures. The proposed neural network-based approaches to audio decorrelation all employ end-to-end processing using a CNN. Although certain hyperparameters of the proposed generator architecture were set based on a classical signal processing-based audio decorrelation method, the employed end-to-end neural network-based processing does not allow for the incorporation of any further domain knowledge. As an alternative, hybrid approaches combining neural networks and classical signal processing techniques could be considered. By exploiting specific domain knowledge, it may be possible to reduce the computational complexity and/or enhance the performance. A straightforward approach would be to perform neural network-based post-processing for a classical signal processing-based decorrelator, with the aim of improving the output signal quality. In [95], first investigations into such post-processing methods have been conducted, proposing an approach to temporal envelope shaping. While [95] specifically addresses temporal envelope shaping, it may be of interest to consider more general training targets for post-processing. Another option would be to consider approaches based on differentiable digital signal processing techniques [160]. The fundamental concept would be to construct the decorrelator from a number of basic signal processing building blocks, such as delays or all-pass filters. Based on the input signal, the optimal parameters of each signal processing block could be estimated using a neural network. As for the methods proposed in this thesis, end-to-end training could be applied by defining the training objective based on the decorrelated output signal.

Binaural SESS Rendering For both the homogeneous and heterogeneous SESS rendering methods, a number of extensions to the employed rendering model could be considered to improve its agreement with a real-world sound source. In particular, the specific geometry of the SESS can be taken into account, resulting in a direction-dependent distance attenuation. Furthermore, instead of assuming an incoherently extended sound source, the rendering model could be extended to take the degree of spatial coherence of the SESS as an additional parameter. The computational complexity of the rendering model could be reduced by employing a sparser spatial sampling of the head-related transfer function (HRTF) dataset positions used to compute the target output covariance matrix. The minimal required spatial resolution can be determined by conducting a series of perceptual experiments.

Furthermore, investigating the suitability of the proposed methods for loudspeaker rendering would be of interest. As a first step, the HRTFs employed in the rendering model need to be replaced by feasible directional loudspeaker rendering functions, such as vector base amplitude panning gains [155]. Further rendering method-specific adaptations would be necessary, particularly given the typical multi-channel nature of the loudspeaker output signal.

The perceptual evaluations of both the homogeneous and heterogeneous binaural SESS rendering methods were conducted exclusively under static conditions, considering a fixed target extent range for the entire signal duration. Furthermore, no direct visual representation of the SESS was available to the listeners. As these restrictions may impact the evaluation results, it would be of interest to assess the performance of the proposed rendering methods inside an interactive VR or AR environment. Moreover, a combination and joint evaluation with further simulated acoustic propagation effects, such as source directivity, distance attenuation, and wall reflections, may be insightful.

For the heterogeneous SESS rendering method, only simulated two-channel input signals have been considered for the evaluation. Consequently, assessing the plausibility of the proposed method using real-world recordings would be of interest. Additionally, for sources whose direction-dependent energy is not approximated well by linear interpolation based on a two-channel input signal, it would be beneficial to consider a greater number of input channels to improve the degree of agreement between the rendering model and the source's characteristics. An extension of the proposed method to more than two input channels would be straightforward.

Appendices

Appendix A

Supplemental Material

A.1 Decorrelation Tree Structure

Given are a time-domain input signal $x(n)$, with discrete time index n , as well as I mutually independent single-channel decorrelator instances. The process to determine the output of the i -th decorrelator instance is denoted by $\text{dec}_i\{\cdot\}$, $i \in \{1, 2, \dots, I\}$. Under the assumption of ideal decorrelation operations, the following conditions hold:

$$\mathcal{E}\{x(n) \text{dec}_i\{x(n)\}\} = 0, \forall i, \quad (\text{A.1})$$

$$\mathcal{E}\{\text{dec}_i\{x(n)\} \text{dec}_j\{x(n)\}\} = \begin{cases} 0, & i \neq j \\ \mathcal{E}\{x^2(n)\}, & i = j, \end{cases} \quad (\text{A.2})$$

where $\mathcal{E}\{\cdot\}$ denotes the statistical expectation.

The output signals of the individual single-channel decorrelator instances can be used directly for multi-channel playback or further processing. Alternatively, an $(I + 1)$ -channel decorrelated output signal can be obtained by mixing the output signals of the individual single-channel decorrelator instances with the original input signal. In the following sections, this mixing procedure is outlined and the properties of the resulting multi-channel decorrelated output signal are discussed. First, the special case of a single decorrelator instance is considered (i.e., $I = 1$). Subsequently, a generalization to multiple decorrelator instances follows (i.e., $I \geq 2$). The described mixing procedure is equivalent to the upmixing procedure employed in MPEG Surround [26, 92], provided that the upmixing parameters are selected to yield uncorrelated output signals with equal power spectral density (PSD).

A.1.1 Single Decorrelator Instance

Considering a single decorrelator instance, the individual channels of the two-channel decorrelated output signal $\mathbf{y}(n) = [y_{11}(n), y_{12}(n)]^T$ are obtained as follows:

$$\begin{aligned} y_{11}(n) &= \frac{1}{\sqrt{2}} (x(n) + \text{dec}_1\{x(n)\}), \\ y_{12}(n) &= \frac{1}{\sqrt{2}} (x(n) - \text{dec}_1\{x(n)\}). \end{aligned} \quad (\text{A.3})$$

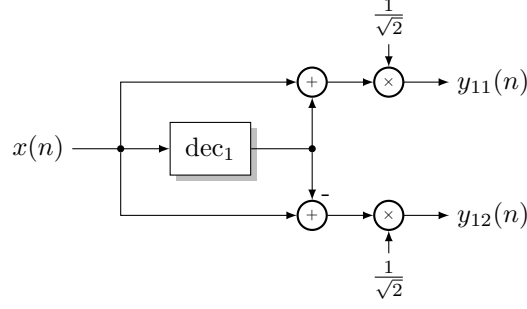


Figure A.1: Block diagram to obtain two-channel decorrelated output signal according to (A.3).

This is equivalent to passing the input signal $x(n)$ through an MPEG Surround one-to-two (OTT) decoding block, wherein both the inter-channel coherence and the inter-channel level difference parameters are set to 0. A block diagram corresponding to (A.3) is depicted in Figure A.1.

Using (A.1) and (A.2), it can be shown that $y_{11}(n)$ and $y_{12}(n)$ are uncorrelated and do have the same PSD as $x(n)$:

$$\begin{aligned}
 \mathcal{E}\{y_{11}(n)y_{12}(n)\} &= \frac{1}{2}\mathcal{E}\{(x(n) + \text{dec}_1\{x(n)\})(x(n) - \text{dec}_1\{x(n)\})\} \\
 &= \frac{1}{2}\mathcal{E}\{x^2(n) - (\text{dec}_1\{x(n)\})^2\} \\
 &= \frac{1}{2}\mathcal{E}\{x^2(n)\} - \frac{1}{2}\mathcal{E}\{(\text{dec}_1\{x(n)\})^2\} \\
 &= 0,
 \end{aligned} \tag{A.4}$$

$$\begin{aligned}
 \mathcal{E}\{y_{11}^2(n)\} &= \frac{1}{2}\mathcal{E}\{(x(n) + \text{dec}_1\{x(n)\})^2\} \\
 &= \frac{1}{2}\mathcal{E}\{x^2(n) + (\text{dec}_1\{x(n)\})^2 + 2x(n)\text{dec}_1\{x(n)\}\} \\
 &= \frac{1}{2}\mathcal{E}\{x^2(n)\} + \frac{1}{2}\mathcal{E}\{(\text{dec}_1\{x(n)\})^2\} + \mathcal{E}\{x(n)\text{dec}_1\{x(n)\}\} \\
 &= \mathcal{E}\{x^2(n)\},
 \end{aligned} \tag{A.5}$$

$$\begin{aligned}
 \mathcal{E}\{y_{12}^2(n)\} &= \frac{1}{2}\mathcal{E}\{(x(n) - \text{dec}_1\{x(n)\})^2\} \\
 &= \frac{1}{2}\mathcal{E}\{x^2(n) + (\text{dec}_1\{x(n)\})^2 - 2x(n)\text{dec}_1\{x(n)\}\} \\
 &= \frac{1}{2}\mathcal{E}\{x^2(n)\} + \frac{1}{2}\mathcal{E}\{(\text{dec}_1\{x(n)\})^2\} - \mathcal{E}\{x(n)\text{dec}_1\{x(n)\}\} \\
 &= \mathcal{E}\{x^2(n)\}.
 \end{aligned} \tag{A.6}$$

Since $y_{11}(n)$ and $y_{12}(n)$ are obtained by mixing $x(n)$ with the decorrelator's output, $x(n)$ is partly correlated with both $y_{11}(n)$ and $y_{12}(n)$. More specifically, again using (A.1) and (A.2), the following

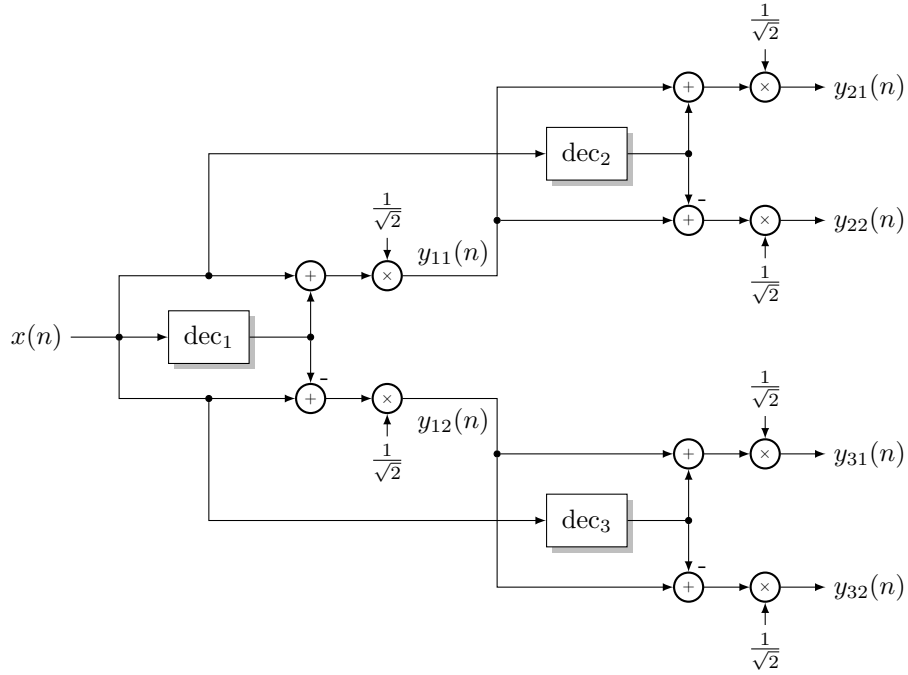


Figure A.2: Block diagram of tree structure used to generated multi-channel decorrelated output signal, for $I = 3$ individual single-channel decorrelator instances.

expressions for $\mathcal{E}\{x(n)y_{11}(n)\}$ and $\mathcal{E}\{x(n)y_{12}(n)\}$ are obtained:

$$\begin{aligned}
 \mathcal{E}\{x(n)y_{11}(n)\} &= \frac{1}{\sqrt{2}}\mathcal{E}\{x(n)(x(n) + \text{dec}_1\{x(n)\})\} \\
 &= \frac{1}{\sqrt{2}}\mathcal{E}\{x^2(n)\} + \frac{1}{\sqrt{2}}\mathcal{E}\{x(n)\text{dec}_1\{x(n)\}\} \\
 &= \frac{1}{\sqrt{2}}\mathcal{E}\{x^2(n)\},
 \end{aligned} \tag{A.7}$$

$$\begin{aligned}
 \mathcal{E}\{x(n)y_{12}(n)\} &= \frac{1}{\sqrt{2}}\mathcal{E}\{x(n)(x(n) - \text{dec}_1\{x(n)\})\} \\
 &= \frac{1}{\sqrt{2}}\mathcal{E}\{x^2(n)\} - \frac{1}{\sqrt{2}}\mathcal{E}\{x(n)\text{dec}_1\{x(n)\}\} \\
 &= \frac{1}{\sqrt{2}}\mathcal{E}\{x^2(n)\}.
 \end{aligned} \tag{A.8}$$

A.1.2 Multiple Decorrelator Instances

In the case of multiple decorrelator instances (i.e., $I \geq 2$), I of the basic mixing blocks depicted in Figure A.1 are concatenated in a tree structure to obtain an $(I + 1)$ -channel decorrelated output signal. This is equivalent to concatenating I MPEG Surround OTT decoding blocks, setting both the inter-channel coherence and the inter-channel level difference parameters to 0.

Using $I = 3$ as an example, the resulting block diagram is depicted in Figure A.2, for which the four-channel decorrelated output signal $\mathbf{y}(n) = [y_{21}(n), y_{22}(n), y_{31}(n), y_{32}(n)]^T$ is obtained. Each of

the I basic mixing blocks employs an independent decorrelator instance. To minimize decorrelation artifacts, the individual decorrelator instances all process the original input signal $x(n)$. The resulting multi-channel output signal is constructed using all basic mixing block output signals that correspond to leaf nodes of the tree structure.

For the two output signals $y_{i1}(n)$ and $y_{i2}(n)$ of the i -th basic mixing block, $i \geq 2$, the following mathematical expressions are obtained:

$$\begin{aligned} y_{i1}(n) &= \frac{1}{\sqrt{2}} (y_{ja}(n) + \text{dec}_i\{x(n)\}), \\ y_{i2}(n) &= \frac{1}{\sqrt{2}} (y_{ja}(n) - \text{dec}_i\{x(n)\}), \end{aligned} \quad (\text{A.9})$$

with

$$j = \lfloor i/2 \rfloor, \quad (\text{A.10})$$

$$a = (i \bmod 2) + 1. \quad (\text{A.11})$$

Given (A.1) and (A.2), it can be shown similarly to (A.5) and (A.6) that

$$\mathcal{E}\{y_{ia}^2(n)\} = \mathcal{E}\{x^2(n)\}, \forall i, a. \quad (\text{A.12})$$

Furthermore, the following condition holds for all pairs of basic mixing block output signals that do not possess a direct hierarchical relationship:

$$\mathcal{E}\{y_{ia}(n)y_{jb}(n)\} = 0, i \neq j \vee a \neq b. \quad (\text{A.13})$$

The individual channels of the resulting multi-channel output signal are thus mutually uncorrelated and do have the same PSD as $x(n)$, provided that (A.1) and (A.2) hold. As shown for $i = 1$ in (A.7) and (A.8), $y_{ia}(n)$ is partly correlated with $x(n)$, $\forall i, a$. The degree of correlation depends on i ; the more decorrelated signal energy is mixed into $y_{ia}(n)$, the lower the correlation.

A.2 Objective Evaluation Plots Complementing Section 7.3.3

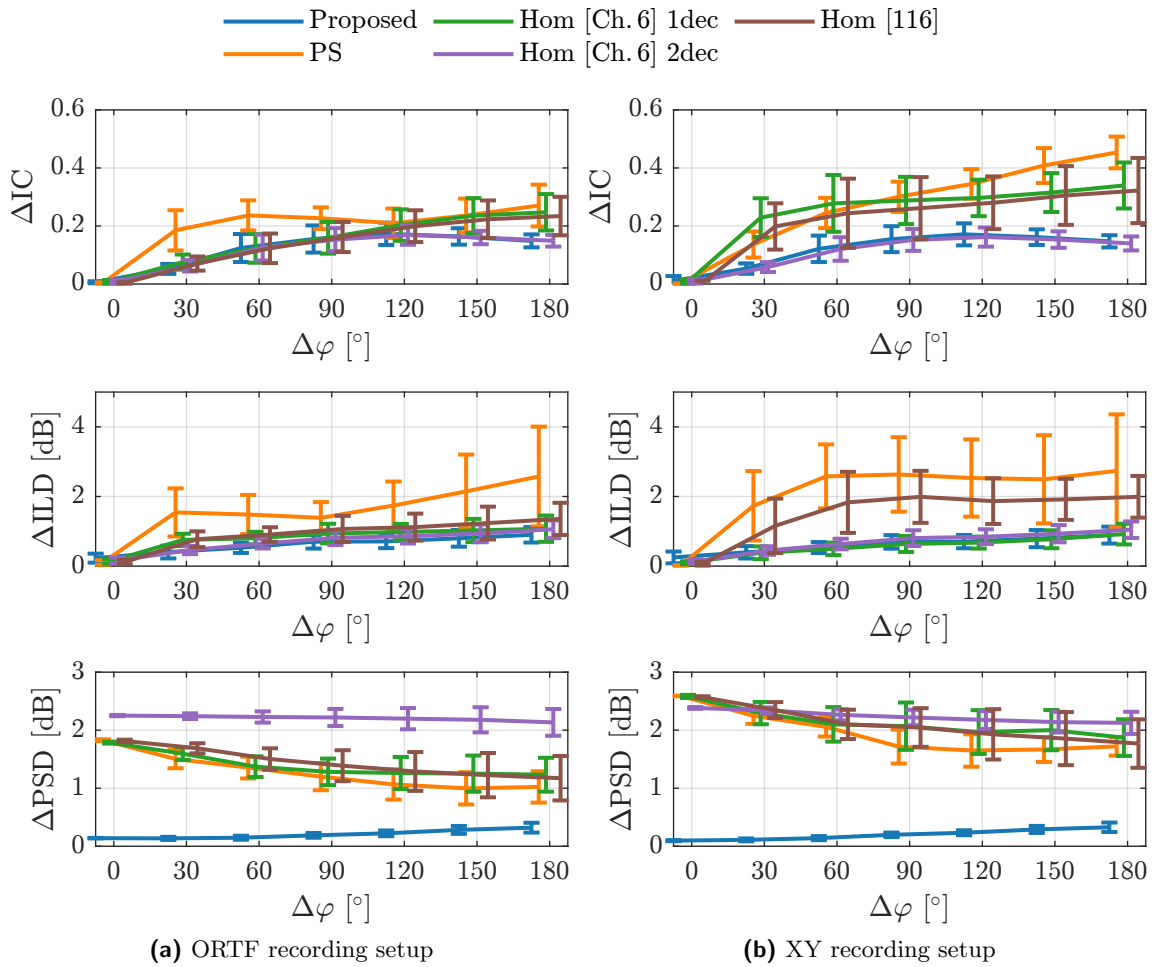


Figure A.3: Mean and standard deviation of RMSE between objective metrics of considered processing methods and binaural reference depending on the extent $\Delta\varphi$, averaged over all considered values for the center $\bar{\varphi}$ of the SESS, for the sparse applause signal.

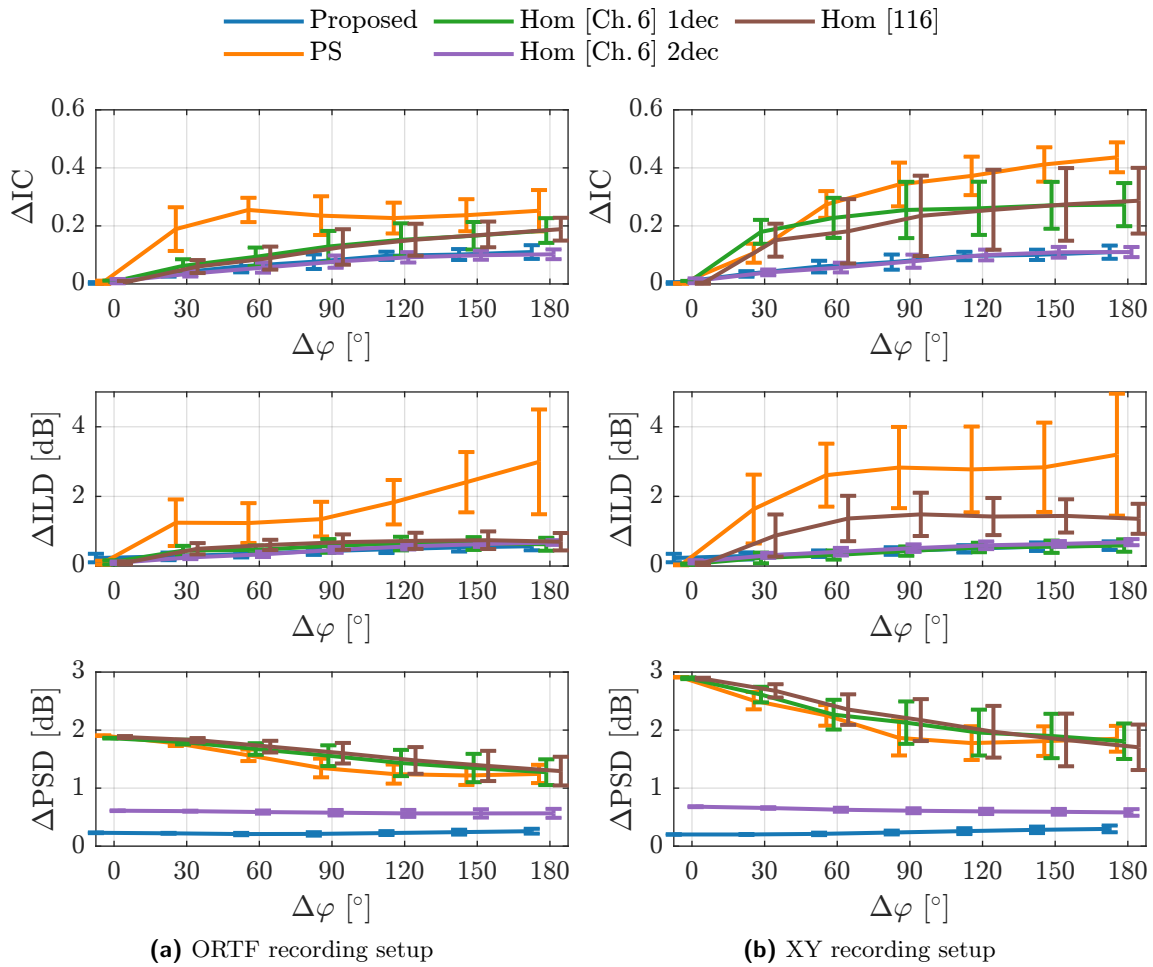


Figure A.4: Mean and standard deviation of RMSE between objective metrics of considered processing methods and binaural reference depending on the extent $\Delta\varphi$, averaged over all considered values for the center $\bar{\varphi}$ of the SESS, for the speech signal.

Bibliography

- [1] J. Berg, “The contrasting and conflicting definitions of envelopment,” in *Proc. AES 126th Conv.*, May 2009.
- [2] J. Blauert and W. Lindemann, “Spatial mapping of intracranial auditory events for various degrees of interaural coherence,” *J. Acoust. Soc. Am.*, vol. 79, no. 3, pp. 806–813, Mar. 1986.
- [3] J. S. Bradley and G. A. Soulodre, “Objective measures of listener envelopment,” *J. Acoust. Soc. Am.*, vol. 98, no. 5, pp. 2590–2597, Oct. 1998.
- [4] G. Kendall, “The decorrelation of audio signals and its impact on spatial imagery,” *Comput. Music J.*, vol. 19, no. 4, pp. 71–87, 1995.
- [5] M. Boueri and C. Kyriakakis, “Audio signal decorrelation based on a critical band approach,” in *Proc. AES 117th Conv.*, Oct. 2004.
- [6] C. Faller, “Parametric multichannel audio coding: synthesis of coherence cues,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 1, pp. 299–310, Jan. 2006.
- [7] V. Pulkki, “Spatial sound reproduction with directional audio coding,” *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, Jun. 2007.
- [8] Y. G. Kim, C. J. Chun, H. K. Kim, Y. J. Lee, D. Y. Jang, and K. Kang, “An integrated approach of 3D sound rendering techniques for sound externalization,” in *Proc. PCM*, Sep. 2010, pp. 682–693.
- [9] N. Xiang, U. Trivedi, and B. Xie, “Artificial enveloping reverberation for binaural auralization using reciprocal maximum-length sequences,” *J. Acoust. Soc. Am.*, vol. 145, no. 4, pp. 2691–2702, Apr. 2019.
- [10] G. Potard and I. Burnett, “Decorrelation techniques for the rendering of apparent sound source width in 3D audio displays,” in *Proc. DAFX-04*, Jan. 2004, pp. 280–284.
- [11] M.-V. Laitinen, S. Disch, and V. Pulkki, “Reproducing applause-type signals with directional audio coding,” *J. Audio Eng. Soc.*, vol. 59, no. 1/2, pp. 29–43, Jan. 2011.
- [12] V. Pulkki, M.-V. Laitinen, and C. Erkut, “Efficient spatial sound synthesis for virtual worlds,” in *Proc. AES 35th Int. Conf.: Audio for Games*, Feb. 2009.
- [13] M. Weger, G. Marentakis, and R. Höldrich, “Auditory perception of spatial extent in the horizontal and vertical plane,” in *Proc. DAFX-16*, Sep. 2016, pp. 301–308.
- [14] G. Potard and J. Spille, “Study of sound source shape and wideness in virtual and real auditory displays,” in *Proc. AES 114th Conv.*, Mar. 2003.
- [15] J. Herre and S. Disch, “MPEG-I Immersive Audio – Reference model for the virtual/augmented reality audio standard,” *J. Audio Eng. Soc.*, vol. 71, no. 5, pp. 229–240, May 2023.
- [16] C. Anemüller, O. Thiergart, and E. A. P. Habets, “A data-driven approach to audio decorrelation,” *IEEE Signal Process. Lett.*, vol. 29, pp. 2477–2481, Nov. 2022.
- [17] —, “Neural audio decorrelation using generative adversarial networks,” in *Proc. WASPAA*, Oct. 2023.

-
- [18] —, “Multi-channel neural audio decorrelation using generative adversarial networks,” *EURASIP J. Audio, Speech Music. Process.*, vol. 2024, no. 58, Nov. 2024.
- [19] C. Anemüller, A. Adami, and J. Herre, “Efficient binaural rendering of spatially extended sound sources,” *J. Audio Eng. Soc.*, vol. 71, no. 5, pp. 281–292, May 2023.
- [20] C. Anemüller, O. Thiergart, and E. A. P. Habets, “Binaural rendering of heterogeneous sound sources with extent,” in *Proc. ICASSP*, Apr. 2024, pp. 471–475.
- [21] —, “Sector-based parametric sound field reproduction in the circular harmonic domain using covariance based rendering,” in *Proc. IWAENC*, Sep. 2022.
- [22] C. Anemüller and J. Herre, “Calculation of directivity patterns from spherical microphone array recordings,” in *Proc. AES 147th Conv.*, Oct. 2019.
- [23] J. Blauert, *Spatial Hearing: Psychophysics of Human Sound Localization*, 2nd ed. Cambridge: MIT Press Ltd, 1996.
- [24] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th ed. BRILL, 2013.
- [25] S. A. Gelfand, *Hearing: An Introduction to Psychological and Physiological Acoustics*, 5th ed. New York: CRC Press, Dec. 2009.
- [26] J. Breebaart and C. Faller, *Spatial Audio Processing: MPEG Surround and Other Applications*. Chichester, UK: Wiley, 2007.
- [27] L. Rayleigh, “XII. On our perception of sound direction,” *Phil. Mag.*, vol. 13, no. 74, pp. 214–232, 1907.
- [28] J. Blauert, “Sound localization in the median plane,” *Acta Acust. united Ac.*, vol. 22, no. 4, pp. 205–213, Jan. 1969.
- [29] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, “Auditory distance perception in humans: A summary of past and present research,” *Acta Acust. united Ac.*, vol. 91, no. 3, pp. 409–420, May 2005.
- [30] D. S. Brungart and W. M. Rabinowitz, “Auditory localization of nearby sources. Head-related transfer functions,” *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1465–1479, Aug. 1999.
- [31] M. B. Gardner, “Distance estimation of 0° or apparent 0° -oriented speech signals in anechoic space,” *J. Acoust. Soc. Am.*, vol. 45, no. 1, pp. 47–53, Jan. 1969.
- [32] P. D. Coleman, “Failure to localize the source distance of an unfamiliar sound,” *J. Acoust. Soc. Am.*, vol. 34, no. 3, pp. 345–346, Mar. 1962.
- [33] A. D. Little, D. H. Mershon, and P. H. Cox, “Spectral content as a cue to perceived auditory distance,” *Perception*, vol. 21, no. 3, pp. 405–416, Jun. 1992.
- [34] P. D. Coleman, “Dual role of frequency spectrum in determination of auditory distance,” *J. Acoust. Soc. Am.*, vol. 44, no. 2, pp. 631–632, Aug. 1968.
- [35] D. S. Brungart, N. I. Durlach, and W. M. Rabinowitz, “Auditory localization of nearby sources. II. Localization of a broadband source,” *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 1956–1968, Oct. 1999.
- [36] D. H. Mershon and L. E. King, “Intensity and reverberation as factors in the auditory perception of egocentric distance,” *Percept. Psychophys.*, vol. 18, no. 6, pp. 409–415, Nov. 1975.
- [37] S. H. Nielsen, “Auditory distance perception in different rooms,” *J. Audio Eng. Soc.*, vol. 41, no. 10, pp. 755–770, Oct. 1993.
- [38] C. Borß, “An improved parametric model for the design of virtual acoustics and its applications,” Ph.D. Thesis, Ruhr-Universität Bochum, Jan. 2011.
- [39] M. Tohyama and A. Suzuki, “Interaural cross-correlation coefficients in stereo-reproduced sound fields,” *J. Acoust. Soc. Am.*, vol. 85, no. 2, pp. 780–786, Feb. 1989.
-

-
- [40] I. M. Lindevald and A. H. Benade, "Two-ear correlation in the statistical sound fields of rooms," *J. Acoust. Soc. Am.*, vol. 80, no. 2, pp. 661–664, Aug. 1986.
- [41] T. Hidaka, L. L. Beranek, and T. Okano, "Interaural cross-correlation, lateral fraction, and low- and high-frequency sound levels as measures of acoustical quality in concert halls," *J. Acoust. Soc. Am.*, vol. 98, no. 2, pp. 988–1007, Aug. 1995.
- [42] S. Klockgether, "The dependence of the spatial impression of sound sources in rooms on interaural cross-correlation and the level of early reflections," in *Proc. DAGA*, Mar. 2015, pp. 1099–1102.
- [43] L. L. Beranek, "Concert Hall Acoustics—2008," *J. Audio Eng. Soc.*, vol. 56, no. 7/8, pp. 532–544, Aug. 2008.
- [44] H. Kuttruff, *Room acoustics*, 4th ed. Spon Press, 2000.
- [45] T. Okano, L. L. Beranek, and T. Hidaka, "Relations among interaural cross-correlation coefficient (IACC(E)), lateral fraction (LFE), and apparent source width (ASW) in concert halls," *J. Acoust. Soc. Am.*, vol. 104, no. 1, pp. 255–265, Jul. 1998.
- [46] M. Barron and A. Marshall, "Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure," *J. Sound Vib.*, vol. 77, no. 2, pp. 211–232, Jul. 1981.
- [47] J. S. Bradley, "Comparison of concert hall measurements of spatial impression," *J. Acoust. Soc. Am.*, vol. 96, no. 6, pp. 3525–3535, May 1994.
- [48] J. Ahrens and S. Spors, "Two physical models for spatially extended virtual sound sources," in *Proc. AES 131st Conv.*, Oct. 2011.
- [49] K. Kurozumi and K. Ohgushi, "The relationship between the cross-correlation coefficient of two-channel acoustic signals and sound image quality," *J. Acoust. Soc. Am.*, vol. 74, no. 6, pp. 1726–1733, Dec. 1983.
- [50] R. Mason, T. Brookes, and F. Rumsey, "Frequency dependency of the relationship between perceived auditory source width and the interaural cross-correlation coefficient for time-invariant stimuli," *J. Acoust. Soc. Am.*, vol. 117, no. 3, pp. 1337–50, Apr. 2005.
- [51] C. Gribben and H. Lee, "The frequency and loudspeaker-azimuth dependencies of vertical interchannel decorrelation on the vertical spread of an auditory image," *J. Audio Eng. Soc.*, vol. 66, no. 7/8, pp. 537–555, Aug. 2018.
- [52] D. R. Perrott and T. N. Buell, "Judgments of sound volume: Effects of signal duration, level, and interaural characteristics on the perceived extensity of broadband noise," *J. Acoust. Soc. Am.*, vol. 72, no. 5, pp. 1413–1417, Nov. 1982.
- [53] D. Cabrera and S. Tilley, "Parameters for auditory display of height and size," in *Proc. ICAD*, Jul. 2003.
- [54] M. P. Cousins, S. Bleeck, and F. M. Fazi, "The effect of inter-channel cross-correlation coefficient on perceived diffuseness," in *Proc. ICSA*, Sep. 2017, pp. 123–130.
- [55] M. P. Cousins, F. M. Fazi, S. Bleeck, and F. Melchior, "Subjective diffuseness in layer-based loudspeaker systems with height," in *Proc. AES 139th Conv.*, Oct. 2015.
- [56] K. Hiyama, S. Komiyama, and K. Hamasaki, "The minimum number of loudspeakers and its arrangement for reproducing the spatial impression of diffuse sound field," in *Proc. AES 113th Conv.*, Oct. 2002.
- [57] A. W. Mills, "Lateralization of high-frequency tones," *J. Acoust. Soc. Am.*, vol. 32, no. 1, pp. 132–134, Jan. 1960.
- [58] W. A. Yost and R. H. Dye, "Discrimination of interaural differences of level as a function of frequency," *J. Acoust. Soc. Am.*, vol. 83, no. 5, pp. 1846–1851, May 1988.
- [59] C. Shuixian, H. Ruimin, L. Yutian, and C. Wenqin, "Frequency dependence of spatial cues and its implication in spatial stereo coding," in *Proc. ICCSSE*, vol. 4, Dec. 2008, pp. 1066–1069.
-

-
- [60] R. C. Rowland and J. V. Tobias, "Interaural intensity difference limen," *J. Speech Hear. Res.*, vol. 10, no. 4, pp. 745–756, Dec. 1967.
- [61] R. M. Hershkowitz and N. I. Durlach, "Interaural time and amplitude jnds for a 500-Hz tone," *J. Acoust. Soc. Am.*, vol. 46, no. 6B, pp. 1464–1467, Dec. 1969.
- [62] W. A. Yost, "Discriminations of interaural phase differences," *J. Acoust. Soc. Am.*, vol. 55, no. 6, pp. 1299–1303, Jun. 1974.
- [63] R. G. Klumpp and H. R. Eady, "Some measurements of interaural time difference thresholds," *J. Acoust. Soc. Am.*, vol. 28, no. 5, pp. 859–860, Sep. 1956.
- [64] J. Zwislocki and R. S. Feldman, "Just noticeable differences in dichotic phase," *J. Acoust. Soc. Am.*, vol. 28, no. 5, pp. 860–864, Sep. 1956.
- [65] D. McFadden and E. G. Pasanen, "Lateralization of high frequencies based on interaural time differences," *J. Acoust. Soc. Am.*, vol. 59, no. 3, pp. 634–639, Mar. 1976.
- [66] S. E. Boehnke, S. E. Hall, and T. Marquardt, "Detection of static and dynamic changes in interaural correlation," *J. Acoust. Soc. Am.*, vol. 112, no. 4, pp. 1617–1626, Oct. 2002.
- [67] I. Pollack and W. J. Trittipoe, "Binaural listening and interaural noise cross correlation," *J. Acoust. Soc. Am.*, vol. 31, no. 9, pp. 1250–1252, Sep. 1959.
- [68] I. Pollack and W. Trittipoe, "Interaural noise correlations: Examination of variables," *J. Acoust. Soc. Am.*, vol. 31, no. 12, pp. 1616–1618, Dec. 1959.
- [69] K. J. Gabriel and H. S. Colburn, "Interaural correlation discrimination: I. Bandwidth and level dependence," *J. Acoust. Soc. Am.*, vol. 69, no. 5, pp. 1394–1401, May 1981.
- [70] J. Culling, H. S. Colburn, and M. Spurchise, "Interaural correlation sensitivity," *J. Acoust. Soc. Am.*, vol. 110, no. 2, pp. 1020–1029, Sep. 2001.
- [71] L. R. Bernstein and C. Trahiotis, "Discrimination of interaural envelope correlation and its relation to binaural unmasking at high frequencies," *J. Acoust. Soc. Am.*, vol. 91, no. 1, pp. 306–316, Jan. 1992.
- [72] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, May 2007.
- [73] J. Middlebrooks, "Virtual localization improved by scaling external-ear transfer functions in frequency," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1493–1510, Oct. 1999.
- [74] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?" *J. Audio Eng. Soc.*, vol. 44, no. 6, pp. 451–469, Jun. 1996.
- [75] M. D. Burkhard and R. M. Sachs, "Anthropometric manikin for acoustic research," *J. Acoust. Soc. Am.*, vol. 58, no. 1, pp. 214–222, Jul. 1975.
- [76] B. Gardner and K. Martin, "HRTF measurements of a KEMAR dummy-head microphone," MIT Media Lab, Tech. Rep. 280, 1994.
- [77] H. S. Braren and J. Fels, "A high-resolution head-related transfer function data set and 3D-scan of KEMAR," Lehrstuhl für Hörtechnik und Akustik, Tech. Rep. RWTH-2020-11307, 2020.
- [78] F. Brinkmann, A. Lindau, S. Weinzierl, G. Geissler, S. van de Par, M. Müller-Trapet, R. Opdam, and M. Vorländer, "The FABIAN head-related transfer function data base," 2020. [Online]. Available: <https://depositonce.tu-berlin.de/handle/11303/6153.5>
- [79] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. WASPAA*, Oct. 2001, pp. 99–102.
-

-
- [80] R. Bomhardt, M. de la Fuente Klein, and J. Fels, “A high-resolution head-related transfer function and three-dimensional ear model database,” *Proc. Meet. Acoust.*, vol. 29, no. 1, Jun. 2017.
- [81] Y.-W. Liu and J. Smith, “Perceptually similar orthogonal sounds and applications to multichannel acoustic echo canceling,” in *Proc. AES 22nd Intl. Conf. on Virtual, Synthetic and Entertainment Audio*, Jun. 2002.
- [82] M. L. Valero and E. A. P. Habets, “Insight into a phase modulation technique for signal decorrelation in multi-channel acoustic echo cancellation,” in *Proc. ICASSP*, Mar. 2016, pp. 519–523.
- [83] H. Lauridsen, “Experiments concerning different kinds of room-acoustics recording,” *Ingeniøren*, vol. 47, pp. 906–910, 1954.
- [84] E. K. Canfield-Dafilou and J. S. Abel, “A group delay-based method for signal decorrelation,” in *Proc. AES 144th Conv.*, May 2018.
- [85] M. Hawksford and N. Harris, “Diffuse signal processing and acoustic source characterization for applications in synthetic loudspeaker arrays,” in *Proc. AES 112th Conv.*, May 2002.
- [86] B. Alary, A. Politis, and V. Välimäki, “Velvet-noise decorrelator,” in *Proc. DAFX-17*, Sep. 2017, pp. 405–411.
- [87] S. Schlecht, B. Alary, V. Välimäki, and E. A. P. Habets, “Optimized velvet-noise decorrelator,” in *Proc. DAFX-18*, Sep. 2018, pp. 87–94.
- [88] B.-s. Xie, B. Shi, and N. Xiang, “Audio signal decorrelation based on reciprocal-maximal length sequence filters and its applications to spatial sound,” in *Proc. AES 133rd Conv.*, Oct. 2012.
- [89] E. Kermit-Canfield and J. Abel, “Signal decorrelation using perceptually informed allpass filters,” in *Proc. DAFX-16*, Sep. 2016, pp. 225–231.
- [90] A. Politis, J. Vilkkamo, and V. Pulkki, “Sector-based parametric sound field reproduction in the spherical harmonic domain,” *IEEE J. Sel. Top. Signal Process.*, vol. 9, no. 5, pp. 852–866, Aug. 2015.
- [91] R. Penniman, “A general-purpose decorrelation algorithm with transient fidelity,” in *Proc. AES 137th Conv.*, Oct. 2014, pp. 99–107.
- [92] J. Herre, H. Purnhagen, J. Breebaart, C. Faller, S. Disch, K. Kjörning, E. Schuijers, J. Hilpert, and F. Myburg, “The reference model architecture for MPEG spatial audio coding,” in *Proc. AES 118th Conv.*, May 2005.
- [93] S. Disch, “Decorrelation for immersive audio applications and sound effects,” in *Proc. DAFX-23*, Sep. 2023.
- [94] M. Schroeder and B. Logan, “‘Colorless’ artificial reverberation,” *IRE Trans. Audio*, vol. AU-9, no. 6, pp. 209–214, Nov. 1961.
- [95] P. Kechichian, A. Ravi, and E. Schuijers, “A cross-domain approach to temporal envelope shaping in parametric stereo coding using deep learning,” in *Proc. IWAENC*, Sep. 2024, pp. 354–358.
- [96] A. Kuntz, S. Disch, T. Backstrom, J. Robilliard, and C. Uhle, “The transient steering decorrelator tool in the upcoming MPEG unified speech and audio coding standard,” in *Proc. AES 131st Conv.*, Oct. 2011.
- [97] S. Disch and A. Kuntz, “A dedicated decorrelator for parametric spatial coding of applause-like audio signals,” in *Microelectronic systems. Circuits, systems and applications*. Berlin: Springer, Dec. 2011, pp. 363–371.
- [98] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, “Creating interactive virtual acoustic environments,” *J. Audio Eng. Soc.*, vol. 47, no. 9, pp. 675–705, Sep. 1999.
-

-
- [99] S. Schlecht, A. Adami, E. Habets, and J. Herre, “Apparatus and method for reproducing a spatially extended sound source or apparatus and method for generating a bitstream from a spatially extended sound source,” US Patent 11 937 068, Mar., 2024.
- [100] C. Schissler, A. Nicholls, and R. Mehra, “Efficient HRTF-based spatial audio for area and volumetric sources,” *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 4, pp. 1356–1366, Apr. 2016.
- [101] V. Pulkki, “Uniform spreading of amplitude panned virtual sources,” in *Proc. WASPAA*, Oct. 1999, pp. 187–190.
- [102] M.-V. Laitinen, T. Pihlajamäki, C. Erkut, and V. Pulkki, “Parametric time-frequency representation of spatial sound in virtual worlds,” *ACM Trans. Appl. Percept.*, vol. 9, no. 2, pp. 1–20, Jun. 2012.
- [103] T. Pihlajamäki, O. Santala, and V. Pulkki, “Synthesis of spatially extended virtual source with time-frequency decomposition of mono signals,” *J. Audio Eng. Soc.*, vol. 62, no. 7/8, pp. 467–484, Aug. 2014.
- [104] H. Su, A. Marui, and T. Kamekawa, “The auditory source widening effect in binaural synthesis with spatial distribution of frequency bands,” *J. Audio Eng. Soc.*, vol. 67, no. 9, pp. 691–704, Sep. 2019.
- [105] T. Hirvonen and V. Pulkki, “Center and spatial extent of auditory events as caused by multiple sound sources in frequency-dependent directions,” *Acta Acust. united Ac.*, vol. 92, no. 2, pp. 320–330, Mar. 2006.
- [106] —, “Perception and analysis of selected auditory events with frequency-dependent directions,” *J. Audio Eng. Soc.*, vol. 54, no. 9, pp. 803–814, Oct. 2006.
- [107] F. Zotter, M. Frank, M. Kronlachner, and J.-W. Choi, “Efficient phantom source widening and diffuseness in ambisonics,” in *Proc. EAA Joint Symposium on Auralization and Ambisonics*, Apr. 2014, pp. 69–74.
- [108] A. Franck, F. M. Fazi, and F. Melchior, “Optimization-based reproduction of diffuse audio objects,” in *Proc. WASPAA*, Oct. 2015.
- [109] S. Khan, “Investigations on modeling spatial extent of sound sources in virtual reality,” Master’s thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg, Jan. 2020.
- [110] J. Schmidt and E. F. Schroeder, “New and advanced features for audio presentation in the MPEG-4 standard,” in *Proc. AES 116th Conv.*, May 2004.
- [111] J.-M. Jot, A. Philp, and M. Walsh, “Binaural simulation of complex acoustic scenes for interactive audio,” in *Proc. AES 121st Conv.*, Oct. 2006.
- [112] C. Verron, M. Aramaki, R. Kronland-Martinet, and G. Pallone, “A 3-D immersive synthesizer for environmental sounds,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 6, pp. 1550–1561, Sep. 2010.
- [113] A. Sibbald, “ZoomFX for 3D-sound,” Sensaura, White Paper DEVPC/012, 2000.
- [114] T. Schmele and U. Sayin, “Controlling the apparent source size in ambisonics using decorrelation filters,” in *Proc. AES Intl. Conf. on Spatial Reproduction - Aesthetics and Science*, Jul. 2018.
- [115] F. Zotter and M. Frank, “Phantom source widening by filtered sound objects,” in *Proc. AES 142nd Conv.*, May 2017.
- [116] L. McCormack, A. Politis, and V. Pulkki, “Rendering of source spread for arbitrary playback setups based on spatial covariance matching,” in *Proc. WASPAA*, Oct. 2021, pp. 371–375.
- [117] J. Vilkamo, T. Bäckström, and A. Kuntz, “Optimized covariance domain framework for time-frequency processing of spatial audio,” *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 403–411, Jul. 2013.
- [118] ISO/IEC JTC1/SC29/WG6, “WD1 of ISO/IEC 23090-4, MPEG-I Immersive Audio,” 140th MPEG Meeting, Document N0168, Oct. 2022.
-

-
- [119] R. J. Beaton, J. G. Beerends, M. Keyhl, and W. C. Treurniet, "Objective perceptual measurement of audio quality," in *Collected Papers on Digital Audio Bit-Rate Reduction*. Audio Eng. Soc., May 1996, pp. 126 – 152.
- [120] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv:1609.03499*, 2016.
- [121] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans. Audio Speech Signal Process.*, vol. 37, no. 3, pp. 328–339, Mar. 1989.
- [122] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or well done?" in *Proc. ICASSP*, May 2019, pp. 626–630.
- [123] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimitakis, and R. Bittner, "MUSDB18-HQ - An uncompressed version of MUSDB18," Aug. 2019. [Online]. Available: <https://doi.org/10.5281/zenodo.3338373>
- [124] A. Adami, S. Disch, G. Steba, and J. Herre, "Assessing applause density perception using synthesized layered applause signals," in *Proc. DAFx-16*, Sep. 2016, pp. 183–189.
- [125] A. Prodeus and I. Kotvytskyi, "On reliability of log-spectral distortion measure in speech quality estimation," in *Proc. APUAVD*, Oct. 2017, pp. 121–124.
- [126] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Trans. Audio Speech Signal Process.*, vol. 24, no. 5, pp. 380–391, Oct. 1976.
- [127] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 829–852, 2022.
- [128] EBU, "Sound quality assessment material recordings for subjective tests - Users' handbook for the EBU-SQAM compact disk," European Broadcasting Union, Tech. Rep. 3253-E, Apr. 1988.
- [129] ITU-T, "Methods for subjective determination of transmission quality," International Telecommunication Union, Recommendation P.800, Aug. 1996.
- [130] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webMUSHRA — A comprehensive framework for web-based listening tests," *J. Open Res. Softw.*, vol. 6, no. 8, Feb. 2018.
- [131] M. J. Lew, "Principles: When there should be no difference – How to fail to reject the null hypothesis," *Trends Pharmacol. Sci.*, vol. 27, no. 5, pp. 274–278, May 2006.
- [132] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NeurIPS*, Dec. 2014.
- [133] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, Dec. 2020, pp. 17 022–17 033.
- [134] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brebisson, Y. Bengio, and A. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Proc. NeurIPS*, Dec. 2019, pp. 14 910–14 921.
- [135] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, May 2020, pp. 6199–6203.
- [136] S. Kim and V. Sathe, "Bandwidth extension on raw audio via generative adversarial networks," *arXiv:1903.09027*, Mar. 2019.
- [137] R. Kumar, K. Kumar, V. Anand, Y. Bengio, and A. Courville, "NU-GAN: High resolution neural up-sampling with GAN," *arXiv:2010.11362*, Oct. 2020.
-

-
- [138] J. Chen, X. Tan, J. Luan, T. Qin, and T.-Y. Liu, “HiFiSinger: Towards high-fidelity neural singing voice synthesis,” *arXiv:2009.01776*, Sep. 2020.
- [139] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, “Least squares generative adversarial networks,” in *Proc. IEEE Intl. Conf. Comput. Vis.*, Oct. 2017, pp. 2794–2802.
- [140] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Proc. ICLR*, May 2019.
- [141] ITU-R, “Method for the subjective assessment of intermediate quality level of audio systems,” International Telecommunication Union, Recommendation BS.1534-3, Oct. 2015.
- [142] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, “Enabling factorized piano music modeling and generation with the MAESTRO dataset,” in *Proc. ICLR*, May 2019.
- [143] ITU-R, “Method for objective measurements of perceived audio quality,” International Telecommunication Union, Recommendation BS.1387-2, May 2023.
- [144] P. Kabal, “An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality,” Department of Electrical & Computer Engineering, McGill University, Tech. Rep., Dec. 2003.
- [145] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, and N. Harte, “ViSQOLAudio: An objective audio quality metric for low bitrate codecs,” *J. Acoust. Soc. Am.*, vol. 137, no. 6, pp. 449–455, Jun. 2015.
- [146] A. Silzle, S. Geysersberger, G. Brohasga, D. Weninger, and M. Leistner, “Vision and technique behind the new studios and listening rooms of the Fraunhofer IIS audio laboratory,” in *Proc. AES 126th Conv.*, May 2009.
- [147] E. A. P. Habets, I. Cohen, and S. Gannot, “Generating nonstationary multisensor signals under a spatial coherence constraint,” *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.
- [148] D. Mirabilli, S. Schlecht, and E. Habets, “Generating coherence-constrained multisensor signals using balanced mixing and spectrally smooth filters,” *J. Acoust. Soc. Am.*, vol. 149, no. 3, pp. 1425–1433, Mar. 2021.
- [149] B. Gold and K. Jordan, “A direct search procedure for designing finite duration impulse response filters,” *IEEE Trans. Audio and Electroacoust.*, vol. 17, no. 1, pp. 33–36, Mar. 1969.
- [150] B. Rakerd and W. M. Hartmann, “Localization of sound in rooms. V. Binaural coherence and human sensitivity to interaural time differences in noise,” *J. Acoust. Soc. Am.*, vol. 128, no. 5, pp. 3052–3063, Nov. 2010.
- [151] C. Trahiotis, L. R. Bernstein, and M. A. Akeroyd, “Manipulating the “straightness” and “curvature” of patterns of interaural cross correlation affects listeners’ sensitivity to changes in interaural delay,” *J. Acoust. Soc. Am.*, vol. 109, no. 1, pp. 321–330, Jan. 2001.
- [152] F. Brinkmann, A. Lindau, S. Weinzierl, S. v. d. Par, M. Müller-Trapet, R. Opdam, and M. Vorländer, “A high resolution and full-spherical head-related transfer function database for different head-above-torso orientations,” *J. Audio Eng. Soc.*, vol. 65, no. 10, pp. 841–848, Oct. 2017.
- [153] ITU-R, “Methods for the subjective assessment of small impairments in audio systems,” International Telecommunication Union, Recommendation BS.1116-3, Feb. 2015.
- [154] E. Pfanzagl-Cardone, “Surround Microphone Techniques,” in *The Art and Science of Surround and Stereo Recording: Including 3D Audio Techniques*. Vienna: Springer, 2020, pp. 97–170.
- [155] V. Pulkki, “Virtual sound source positioning using vector base amplitude panning,” *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, Jun. 1997.
- [156] J. N. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *Proc. ICML*, Jul. 2015, pp. 2256–2265.
-

-
- [157] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Proc. NeurIPS*, Dec. 2020, pp. 6840–6851.
 - [158] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *Proc. ICLR*, Apr. 2014.
 - [159] E. G. Tabak and E. Vanden-Eijnden, “Density estimation by dual ascent of the log-likelihood,” *Commun. Math. Sci.*, vol. 8, no. 1, pp. 217–233, Mar. 2010.
 - [160] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts, “DDSP: Differentiable digital signal processing,” in *Proc. ICLR*, Apr. 2020.
-