TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Joonas Nikunen

Object-based Modeling of Audio for Coding and Source Separation

Julkaisu 1276 • Publication 1276

Tampere 2015

Tampereen teknillinen yliopisto. Julkaisu 1276
Tampere University of Technology. Publication 1276

Joonas Nikunen

Object-based Modeling of Audio for Coding and Source Separation

Thesis for the degree of Doctor of Science in Technology to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB223, at Tampere University of Technology, on the 16th of January 2015, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2015

Supervisor:

Tuomas Virtanen, Adjunct Professor
Department of Signal Processing
Tampere University of Technology
Tampere, Finland

Pre-examiner:

Alexey Ozerov, Ph.D.
Technicolor
Cesson-Sévigné, France

Pre-examiner and opponent:

Ville Pulkki, Associate Professor
Department of Signal Processing and Acoustics
Aalto University
Helsinki, Finland

Opponent:
Derry FitzGerald, Ph.D.
Cork Institute of Technology
Cork, Ireland

ISBN 978-952-15-3438-6 (printed)
ISBN 978-952-15-3452-2 (PDF)
ISSN 1459-2045

Abstract

THIS thesis studies several data decomposition algorithms for obtaining an object-based representation of an audio signal. The estimation of the representation parameters are coupled with audio-specific criteria, such as the spectral redundancy, sparsity, perceptual relevance and spatial position of sounds. The objective is to obtain an audio signal representation that is composed of meaningful entities called audio objects that reflect the properties of real-world sound objects and events. The estimation of the object-based model is based on magnitude spectrogram redundancy using non-negative matrix factorization with extensions to multichannel and complex-valued data. The benefits of working with object-based audio representations over the conventional time-frequency bin-wise processing are studied. The two main applications of the object-based audio representations proposed in this thesis are spatial audio coding and sound source separation from multichannel microphone array recordings.

In the proposed spatial audio coding algorithm, the audio objects are estimated from the multichannel magnitude spectrogram. The audio objects are used for recovering the content of each original channel from a single downmixed signal, using time-frequency filtering. The perceptual relevance of modeling the audio signal is considered in the estimation of the parameters of the object-based model, and the sparsity of the model is utilized in encoding its parameters. Additionally, a quantization of the model parameters is proposed that reflects the perceptual relevance of each quantized element.

The proposed object-based spatial audio coding algorithm is evaluated via listening tests and comparing the overall perceptual quality to conventional time-frequency block-wise methods at the same bitrates. The proposed approach is found to produce comparable coding efficiency while providing additional functionality via the object-based coding domain representation, such as the blind separation of the mixture of sound sources in the encoded channels.

For the sound source separation from multichannel audio recorded by a microphone array, a method combining an object-based magnitude model and spatial covariance matrix estimation is considered. A direction of arrival-based model for the spatial covariance matrices of the sound sources is proposed. Unlike the conventional approaches, the estimation of the parameters of the proposed spatial covariance matrix model ensures a spatially coherent solution for the spatial parameterization of the sound sources. The separation quality is measured with objective criteria and the proposed method is shown to improve over the state-of-the-art sound source separation methods, with recordings done using a small microphone array.

Preface

THIS work has been carried out at the Department of Signal Processing, Tampere University of Technology, during 2010-2014. First and foremost I want to thank my supervisor Tuomas Virtanen for all the guidance and support during my doctoral studies. His help and advice in the process of accomplishing the research discoveries leading to the included publications of this thesis has been irreplaceable. I also want to extend my gratitude to Anssi Klapuri who was leading the audio research group when I originally joined the group as a Master's thesis worker, which has eventually led to the completion of this doctoral thesis.

I thank the pre-examiners of this thesis, Dr. Alexey Ozerov from Technicolor and Associate Professor Ville Pulkki from Aalto University, for providing valuable feedback for improving and finalizing the thesis. The opponents of the public defense of my thesis, Associate Professor Ville Pulkki and Dr. Derry FitzGerald from Cork Institute of Technology, also receive my highest gratitude.

The research work included in this thesis was made possible by funding from Nokia Research Center (NRC). My gratitude goes to the people at NRC, Mikko Tammi, Miikka Vilermo, Adriana Vasilache, Matti Hämäläinen and Mauri Väänänen. Their scientific input in evaluating the work and discussion of ideas and directions to look into have had a major role in developing the methods presented in the included publications of this thesis.

I want to thank all the past and present colleagues at the audio research group, this includes but is not limited to, Pasi Pertilä, Antti Hurmalainen, Katariina Mahkonen, Hanna Silén, Mikko Parviainen, Toni Heittola, Annamaria Mesaros, Aleksandr Diment, Tom Barker and Toni Mäkinen. Over the years they have made the audio research group a welcoming and relaxed, yet the most professional place to conduct research in the field of audio signal processing.

Finally, my most sincere thanks and love goes to Susanna, without her support over the years this thesis would have not been possible. I also want to thank my parents Erkki and Satu for their endless support and my sister Emmi for all the shared joys in life.

Joonas Nikunen
Tampere, December 2014

Contents

Abstract i

Preface iii

Contents v

List of Abbreviations ix

List of Included Publications x

1 Introduction 1
- 1.1 Object-based Audio Coding . . . 2
- 1.2 Sound Source Separation . . . 3
- 1.3 Objectives of the Thesis . . . 5
- 1.4 Main Results of the Thesis . . . 6
- 1.5 Organization of the Thesis . . . 9

2 Background 11
- 2.1 Definition of an Audio Object . . . 11
- 2.2 Hearing . . . 12
  - 2.2.1 Ear Anatomy . . . 13
  - 2.2.2 Critical Bands and Masking . . . 13
  - 2.2.3 Spatial Hearing . . . 14
- 2.3 Sound and Recording . . . 15
  - 2.3.1 Sound Sources and Propagation . . . 15
  - 2.3.2 Microphones and Microphone Arrays . . . 16
  - 2.3.3 Sound Capture . . . 18
  - 2.3.4 Mixing Model . . . 19
  - 2.3.5 Sound Source Separation . . . 19

2.4 Frequency-domain Representations 20
2.4.1 Mixing in Frequency Domain 21
2.4.2 Spatial Covariance Domain 22

3 Object-based Models of Audio Signals 25
3.1 Spectral Redundancy and Spectrogram Decompositions 25
3.2 Independent Component Analysis 27
3.3 Non-negative Matrix and Tensor Factorization 28
3.3.1 Model for Single Channel Audio 29
3.3.2 Model for Multichannel Audio 30
3.3.3 Optimization Criteria and Parameter Estimation . . . 31
3.3.4 Algorithm Description 32
3.4 Complex-valued Non-negative Matrix Factorization 32
3.4.1 Component-wise Spatial Covariance Estimation 33
3.4.2 Source-wise Spatial Covariance Estimation 34
3.4.3 Interpretations of Complex-valued NMF 34

4 Object-based Audio Coding 37
4.1 Audio Coding Background 37
4.1.1 Measuring Coding Distortions with Perceptual Criteria 38
4.1.2 Spatial Audio Coding 39
4.1.3 Spatial Audio Object Coding 40
4.1.4 Blind Upmixing 41
4.1.5 Entropy Coding and Redundancy 42
4.2 Channel-wise Object-Based Audio Coding Using NMF 42
4.2.1 Coding Framework Overview 43
4.2.2 Perceptual Optimization Criterion for NMF 45
4.2.3 Quantization and Rate of the Model Parameters 45
4.2.4 Sparsity of the Model Parameters 47
4.2.5 Results and Findings 48
4.2.6 Conclusion on Channel-wise Approach 49
4.3 Multichannel Audio Upmixing by NTF 50
4.3.1 Spatial Cues and NTF 50
4.3.2 Upmixing by Time-Frequency Filtering 51
4.3.3 Cost Function for Upmixed Signal Perceptual Quality . 53
4.3.4 Evaluation of Coding Efficiency 55
4.3.5 Audio Source Separation with NTF-based SAC 56

5 Source Separation from a Multichannel Audio Recording 59
5.1 Problem Definition and Earlier Work 60
5.2 Array Signal Processing 61

5.2.1 Time Difference of Arrival 61
5.2.2 Beamforming 63
5.2.3 Direction of Arrival Estimation 66
5.3 Source Separation Using Complex-valued NMF 67
5.3.1 Spatial Covariance Matrix Estimation 68
5.4 Direction of Arrival-based Model for Spatial Covariance 69
5.4.1 SCM Model by Superposition of DOA Kernels 69
5.4.2 Complex-valued NMF with the DOA-based SCM Model 71
5.5 Direction of Arrival Estimation Performance 73
5.6 Source Separation Results 75

6 Conclusions and Future Work **79**
6.1 Conclusions 79
6.2 Future Work 81

Bibliography **83**

P1 Publication 1 **99**

P2 Publication 2 **105**

P3 Publication 3 **117**

P4 Publication 4 **123**

P5 Publication 5 **131**

P6 Publication 6 **147**

P7 Publication 7 **163**

List of Abbreviations

ASR	Automatic speech recognition
BSS	Blind source separation
CNMF	Complex-valued non-negative matrix factorization
DFT	Discrete Fourier transform
DOA	Direction of arrival
DSB	Delay and sum beamformer
EM	Expectation maximization
GCC	Generalized cross correlation
HRTF	Head-related transfer function
IBM	Ideal binary mask
ICA	Independent component analysis
ICC	Inter-channel coherence
ICLD	Inter-channel level difference
ICTD	Inter-channel time difference
ILD	Interaural level difference
ISR	Image-to-spatial distortion ratio
ISS	Informed source separation
ITD	Interaural time difference
IVA	Independent vector analysis
KLD	Kullback-Leibler divergence
MVDR	Minimum variance distortionless beamformer
NMF	Non-negative matrix factorization
NMR	Noise-to-mask ratio
NTF	Non-negative tensor factorization
RIR	Room impulse response
SAC	Spatial audio coding
SAOC	Spatial audio object coding
SAR	Signal-to-artefact ratio

SCM	Spatial covariance matrix
SDR	Signal-to-distortion ratio
SED	Squared Euclidean distance
SIR	Signal-to-interference ratio
STFT	Short-time Fourier transform
TDOA	Time-difference of arrival
WSED	Weighted squared Euclidean distance

List of Included Publications

This thesis is a compound thesis consisting of the following publications, preceded by an introduction to the research field and summary of the main findings of the publications. Parts of this thesis have been previously published and the original publications are reprinted with a permission from the respective copyright holders. The publications are referred in the text by the notation [P1],[P2] and so forth.

P1 **J. Nikunen and T. Virtanen**, "Noise-to-Mask Ratio Minimization by Weighted Non-negative Matrix factorization," in *Proceedings of the 35th International Conference on Acoustics, Speech, and Signal Processing*, Dallas, Texas, USA, pp. 25–28, 2010

P2 **J. Nikunen and T. Virtanen**, "Object-based Audio Coding Using Non-negative Matrix Factorization for the Spectrogram Representation," in *Proceedings of the 128th Audio Engineering Society Convention* , London, U.K., 2010

P3 **J. Nikunen, T. Virtanen and M. Vilermo**, "Multichannel Audio Upmixing Based on Non-negative Tensor Factorization Representation," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, pp. 33–36, 2011

P4 **J. Nikunen, T. Virtanen, P. Pertilä and M. Vilermo**, "Permutation Alignment in Frequency-domain ICA by the Maximization of Intra-source Envelope Correlations," in *Proceedings of the 20th European Signal Processing Conference*, Bucharest, Romania, pp. 1489–1493, 2012

P5 **J. Nikunen, T. Virtanen and M. Vilermo**, "Multichannel Audio Upmixing by Time-Frequency Filtering Using Non-Negative Tensor

Factorization," in *Journal of the Audio Engineering Society*, vol. 60, no. 10, pp. 794–806, 2012

P6 **J. Nikunen and T. Virtanen**, "Direction of Arrival Based Spatial Covariance Model for Blind Sound Source Separation," in *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 3, pp. 727–739, 2014

P7 **J. Nikunen and T. Virtanen**, "Multichannel audio separation by Direction of Arrival Based Spatial Covariance Model and Non-negative Matrix Factorization," in *Proceedings of the 39th International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, pp. 6727–6731, 2014

Author's Contributions to the Publications

The thesis author is the main author in all the included publications. The included works have been done in close collaboration with, and under the guidance of T. Virtanen who has provided feedback in preparation of all the publications, while the author has done most of the writing. The motivation and initial idea for Publication [P1] was proposed by T. Virtanen who helped in formulating the problem to its final form. The idea for the algorithm proposed in Publication [P4] was provided by T. Virtanen, while the implementation and formulation of the algorithm are by the author. The multichannel audio coding framework in Publications [P3] and [P5] was initially discussed among the thesis author, T. Virtanen and M. Vilermo, while the implementation and refinement of the idea was done by the author. The evaluations of audio coding quality in [P3] and [P5] are done by M. Vilermo while the separation quality evaluation in [P5] is done by the author. The original ideas of Publications [P2] and [P6] were proposed by the author, while formulation of the final form includes feedback from T. Virtanen. Evaluation material for publication [P4] was provided by P. Pertilä. All the other evaluations and implementations were done by the author.

Chapter 1

Introduction

RECORDINGS of natural auditory scenes are composed of mixtures of different sounds. For us humans the separation of different sound sources and events by their spatial position or spectral features comes naturally. Our hearing functions subconsciously by enhancing, separating or focusing on different sound sources of interest. For example, at a concert we can easily shift the focus of our hearing to different instruments or, in the case of multiple simultaneous speech sources, we can hear and recognize the speech of our conversation partner. It is evident that our hearing processes the auditory information using a higher level representation which is assumed to be object-based [93, 94, 152], meaning that our consciousness of sounds is based on entities referred to as audio objects.

The adoption of object-based processing into computer audio is an active and ongoing field of research [13, 103, 131, 145]. It has been enabled by the recent advances in machine learning algorithms that are applicable for structural analysis of audio signals [19, 80]. Machine learning in this case can be divided into supervised and unsupervised methods. In supervised learning the system is presented with annotated data of which structure and properties it tries to learn. In operation the system is applied to process unforeseen data which is assumed to reflect the statistics and structure of the data used in learning. In unsupervised methods the learning occurs during the operation, and assumptions on the structure of the data can be included in a definition of the model it uses, for example, assuming certain statistics of the input data [78].

A fair amount of the analysis abilities of human hearing can be attributed to the memory and combining information from other sensory systems such as vision and touch. Our hearing is presented with annotated stimuli over the whole course of our life and the brain makes associations and interpreta-

tions of it. This learning allows us to recognize and classify different sounds and recognize speech and the speaker's identity even in difficult and unfamiliar environments [14]. The functioning of the hearing extends to concepts that seem trivial in our everyday life, which include, for example, associating different notes as originating from the same instrument even though their fundamental frequency and spacing of harmonics vary. Integrating such processing and decision making into an algorithm analyzing an audio signal is nearly impossible, at least with the level of precision we humans are capable of. Specializing the algorithm for a single task, for example speech recognition or instrument classification, and using adequate training material can result in satisfying performance [7, 21, 44]. However, there exists a need for more generic audio models for cases where training is not viable, the most notable being separation of sound sources from an audio recording done in unforeseen conditions with no knowledge of the sources involved in the recording. In such cases the focus of the research shifts towards blind source separation approaches [20, 127] that operate on information derived solely from the data they observe.

In this thesis the aspects of unsupervised machine learning are studied with the intent of deriving an object-based models and representations of audio signals. The audio objects within the studied models reflect the various properties of sound objects and entities, such as phonemes, words and notes, all the way up to entire sound sources. The two applications focused upon in the thesis are object-based audio coding and sound source separation. These fields are briefly introduced in the following sections in order to provide the problem statement and preliminaries of conventional approaches contrasted with object-based methods.

1.1 Object-based Audio Coding

The consumption of audiovisual content has been revolutionized by the increase in both the bandwidth and processing power of portable devices. The streaming of video with several audio tracks may be viewed with devices ranging from those comprised of only one channel of audio playback to the full scale multichannel home theater setups. The scalability of audio coding in terms of the number of channels addresses this issue and is known as spatial audio coding (SAC) [33, 125]. The SAC methods share a similar framework, where a perceptually encoded downmix, along with auxiliary information consisting of the spatial position of each time-frequency (TF) block, is used to recover the original multichannel audio signal. The parameters for encoding the spatial position are based on the magnitude and time difference

between the input channels [4, 34]. Additionally, coherence and diffuseness of the time-frequency blocks between the channels are used [55, 110].

The conventional SAC methods [54, 55] do not utilize any object-based model and estimate the spatial position of each time-frequency block individually. In the context of this thesis an audio object is a spectral pattern with a time-dependent activation which represents a part of or an entire sound source. These audio objects occur redundantly during the audio signal that is under analysis and finally encoded. Obtaining an object-based model of the multichannel audio signal allows representation of the spatial position of each audio object, spanning the whole frequency range, using a single parameter. This greatly decreases the amount of auxiliary information as compared to the block-wise approaches. Additionally, the long term redundancy in audio signals become utilized with the object-based representation. In return, the representation and transmission of the object-based model requires extra bitrate.

A class of related methods known as spatial audio object coding (SAOC) [13, 56] assumes that the individual audio tracks of each audio object are available during the encoding. In such a setting the object-wise spatial parameterization of the multichannel signal can be done without first blindly decomposing the mixture into audio objects.

The additional benefit of using blindly estimated object-based representation for SAC is its source separation capability. The sound source separation ability is based on learning recurrent and repetitive spectral patterns using audio spectrogram decompositions, such as the non-negative matrix factorization (NMF) and non-negative tensor factorization (NTF) [19, 37, 80]. The NMF and NTF applied on decomposing the audio signal spectrogram are known to produce audio objects that can be used for sound source separation [20, 147]. Such object-based models of audio signals allow manipulation of its content based on meaningful entities, i.e., the audio objects. Using the NMF and the NTF for SAC not only allows channel-wise scalability, but also audio object-based control for the user over the reconstruction. This enables the possibility of the removal or amplification of sound sources present in the encoded content, without separate encoding of each source track. Additionally, it is not necessary for the separate source tracks to be present in the encoding stage as is the case in most music content to be encoded.

1.2 Sound Source Separation

The field of sound source separation studies the process of obtaining separate source audio tracks from the mixture audio signal that it is composed

of. The audio signal is recorded by one or several microphones in a situation where multiple sources are emitting sound simultaneously. The use of sound source separation has been conventionally associated with telephony, teleconferencing and automatic speech recognition (ASR), where the enhancement and separation of multiple speakers is sought after [149]. Applications related to entertainment, such as gaming and the sharing of videos in social media, have generated an increased demand for generic sound source separation. The recording of audio and video has never been more accessible than during the current age of mobile phones with capabilities to record audio using miniature microphones that not only handle high dynamics [74] but are also capable for capturing high-quality audio in any practical scenario. The recordings are usually done in public places with lots of interfering noise, which prevents understanding of the speech or other essential content present in the recording. The need for sound source separation in such cases is evident. Another example is the voice interface and communication used in gaming consoles [75] which employ ASR in a home environment containing numerous sources of interfering noise.

The field of sound source separation can be roughly divided into supervised [128, 132, 147], informed [82, 91, 106] and blind approaches [65, 108, 118, 150]. The supervised and semi-supervised methods either require training or other annotated information to operate. In the field of informed source separation (ISS), a set of known sources are mixed together and need to be separated from the mixture after transmission. The completely blind approaches assume no information being available for the sources, their parameters or the environment used in recording. For the ASR, the reconstruction of separated signals is not a mandatory step. The feature extraction through means of source separation can be used directly as the input of the recognition module (recognition back-end), for example, as in [45, 60, 95].

Furthermore, the sound source separation methods can be categorized based on operation in either a single [132, 147] or multichannel context [65, 108, 118, 150]. This thesis concentrates on blind sound source separation from a multichannel audio signal assuming that no information regarding the sources or the recording environment is available. The use of portable devices for audio recordings and the growing number of microphones embedded in one device encourage research on algorithms that are suitable for arrays of microphones of a very compact size. For example, there already exist systems that run independently on a mobile phone [100] and are able to separate two sources from a two-channel audio recording.

The conventional multichannel blind sound source separation methods include independent component and vector analysis (ICA and IVA) [62, 81] applied in frequency domain and spatial filtering by beamforming and spa-

tial post-filters [139]. ICA and IVA are usually restricted to determined cases where the number of microphones is higher or equal to the number of sources to be separated. In beamforming the good spatial selectivity, and thus separation, requires a high number of microphones, which is a contradictory requirement with respect to the size and cost of a portable device.

The benefits of object-based models of audio signals for sound source separation were mentioned in the previous section regarding object-based audio coding. In the single channel case, where no spatial information exists to be utilized, the NMF-based models of audio signals have been considered for supervised source separation [128, 147]. The NMF is known for being able to learn repetitive spectral patterns and events from the audio signal spectrogram, and such audio objects can be used as a basis for building a model for sound sources and their separation. More recently the NMF model has been extended to multichannel audio recordings and estimation of spatial properties of the audio objects [101, 121]. These methods utilize the spatial information in combining several audio objects to model complete sound sources that are spatially discriminant.

Several ISS methods utilize an object-based model which is obtained using NMF or NTF [83, 103]. In the coding-based ISS [103], the goal is to represent and encode the sources for transmission. The sources are parameterized by an object-based model obtained by applying NTF to the magnitude spectrogram. The encoded object-based model is used for recovering the source signals from a mixture signal in the decoding stage. Coding-based ISS is alternatively known in the literature as spatial audio object coding (SAOC) [13], which extends the SAC framework for separate audio source tracks. The ISS methods based on NMF and NTF are thus closely related to both topics of the thesis by combining object-based separation and coding.

1.3 Objectives of the Thesis

The main objective of this thesis is to study and utilize conventional blind data decomposition methods, such as the NMF and the NTF, in developing object-based audio signal models for multichannel audio coding and sound source separation. Models inherently incorporating structures suitable for modeling spectral content and spatial position of audio objects are proposed and estimations of the parameters of the models are presented.

In this thesis the object-based models for audio signals are obtained by means of unsupervised machine learning that relies on the spectral redundancy, perceptual relevance, sparsity, independence and analysis of the spatial position of the audio object. This thesis introduces two main applications

for these blindly estimated object-based models of audio, multichannel audio coding and sound source separation, both in single and multichannel cases.

The developed models and audio signal representations for the purpose of audio coding are compared to conventional methods that operate on time-frequency blocks with fixed size. Additionally, the performance of object-based models in sound source separation are compared to generic data decomposition algorithms applied to audio signals.

1.4 Main Results of the Thesis

The development and proposed use of object-based models of audio signals for the tasks of audio coding and sound source separation is the main contribution of this thesis. The main results include introduction of new functionality to spatial audio coding and improving separation performance over conventional time-frequency block-wise approaches. Additionally, estimating object-based representation of audio signals by combining spectral redundancy, perceptual relevance and spatial position of the audio objects can be regarded as the main novelty of the thesis.

The main results regarding audio coding include the proposed object-based SAC approach in [P3] which uses non-negative tensor factorization (NTF) [37, 80] for estimation of object-based representation of the multichannel magnitude spectrogram. The SAC algorithm relies on the perceptually motivated optimization criteria for the object-based approximation of the audio magnitude spectrogram proposed in [P1] and utilizes the quantization and perceptual encoding framework considered in [P2]. The coding efficiency of the proposed SAC is found to be comparable to conventional methods in [P3], while a completely new functionality, the blind separation of audio sources mixed in the encoded channels, is studied in [P5].

The sound source separation from a multichannel audio recording introduced in this thesis consists of both conventional approach of ICA applied in frequency domain [P4] and estimation of spatial properties of audio objects in [P6] and [P7]. In the case of ICA-based separation, a source activation envelope estimation for improving the permutation alignment of frequency-wise source estimates is proposed in [P4] and is found to improve the separation quality over conventional approaches. Sound source separation based on estimation of spatially coherent audio objects derived by the NMF is proposed in [P6]. The spatially coherent audio object estimation is extended to cover estimation of entire sound sources in [P7] and is found to exceed the separation performance of other spatial NMF algorithms and conventional methods.

The results and contributions of the individual publications included in the thesis are summarized in the following listing.

Publication 1 : Noise-to-Mask Ratio Minimization by Weighted Non-negative Matrix Factorization

The publication introduces a perceptually motivated cost function for modeling the magnitude spectrogram of audio using the NMF. The cost function proposed minimizes the noise-to-mask ratio [140] between the magnitude observations and its NMF approximation. The proposed cost function is found to improve perceptual quality of the NMF model over conventional cost functions such as squared Euclidean distance and Kullback-Leibler divergence.

Publication 2 : Object-based Audio Coding Using Non-negative Matrix Factorization for the Spectrogram Representation

In this publication a single channel object-based audio coding algorithm is proposed which utilizes the perceptually motivated NMF from [P1]. The publication concentrates on developing a quantization and entropy coding scheme for the NMF model parameters, i.e., the object-based model. Additionally the quantization of the phase spectrogram required for time-domain signal reconstruction is proposed. The NMF model is found to be very efficient for representing magnitude information due to its sparsity properties, but overall coding efficiency remains low due to the required coding of the phase information.

Publication 3 : Multichannel Audio Upmixing Based on Non-negative Tensor Factorization Representation

The publication proposes a spatial audio coding algorithm based on multichannel upmixing utilizing an object-based model obtained by using the NTF. The coding framework consists of the transmission of a perceptually encoded downmix and quantized NTF model, which is reconstructed on the decoder side and used for recovering multiple channels from the downmix via upmix filtering. The perceptually motivated cost function from [P1] is extended for multichannel observations, and perceptual quality is optimized for the upmix filtering operation. The proposed SAC algorithm is evaluated using listening tests and is concluded to achieve perceptual quality similar to conventional SAC methods [54].

Publication 4 : Permutation Alignment in Frequency-domain ICA by the Maximization of Intra-source Envelope Correlations

The publication investigates blind source separation using frequency domain ICA and proposes a permutation alignment based on correlation of source activation envelopes. The estimation of the source envelopes is based on the singular value decomposition applied to the source spectrum after an initial frequency-wise alignment by clustering the time difference of arrival (TDOA) of ICs at each frequency. The method is shown to improve the separation quality of multiple speakers recorded in a real environment where the capture consists of a high amount of spatial aliasing. This poses difficulties for the TDOA-based permutation alignment due to the aliasing of the phase differences in frequency domain processing.

Publication 5 : Multichannel Audio Upmixing by Time-Frequency Filtering Using Non-Negative Tensor Factorization

This publication further examines the source separation properties and encoding benefits of the object-based SAC proposed in [P3]. The upmixing process using the NTF is formulated in a more general manner, allowing use of either mono or stereo downmix. The evaluation of the source separation quality with user-created clustering of the NTF components to entire sources is presented and found to be comparable to ideal binary mask separation. The manipulation of the selected individual sources of the upmixed mixture by attenuating or boosting is concluded to be plausible, allowing similar abilities as the SAOC [13] without the need of individual source tracks being available during the encoding.

Publication 6 : Direction of Arrival Based Spatial Covariance Model for Blind Sound Source Separation

The publication proposes a direction of arrival-based spatial covariance model to be used in conjunction with NMF and utilized for blind source separation of multichannel audio. The proposed spatial model and estimation of its parameters independent of frequency ensures NMF components to be spatially coherent, i.e., originating from one direction. Furthermore, a k-means clustering applied to direction parameters of NMF components is proposed for combining them to entire acoustical sources. The separation quality of the

proposed method, measured with energy-based and perceptual metrics, is shown to exceed the conventional approach of frequency-wise spatial covariance matrix estimation of NMF components and the ICA based separation from [P4].

Publication 7 : Multichannel audio separation by Direction of Arrival Based Spatial Covariance Model and Non-negative Matrix Factorization

This publication proposes an improved version of the NMF based sound source separation algorithm proposed in [P6] by introducing the estimation of the NMF component to the source clustering parameter as a part of the entire optimization problem. Additionally, the direction of arrival-based spatial covariance model is defined as being estimated source-wise, instead of NMF component-wise as in [P6]. Utilizing the concurrent optimization, with all the parameters affecting the separation and enforcing spatially coherent sources, the algorithm is able to improve the separation over all previous attempts and other state-of-the-art sound source separation methods.

1.5 Organization of the Thesis

The thesis is organized as follows. First, an introduction to the field of audio signal processing consisting of topics such as sound, capture, hearing and frequency domain signal processing is presented in Chapter 2. The decomposition models of audio signals as a basis for object-based representations are introduced in Chapter 3.

The first main part of this thesis, consisting of the object-based model for spatial audio coding, is presented in Chapter 4. The section includes a short review of audio coding and is followed by an investigation of the main findings of Publications [P1],[P2],[P3] and [P5]. The topics include the perceptually motivated cost function for estimation of the object-based model in Section 4.2.2, quantization in Section 4.2.3 and entropy coding of its parameters in Section 4.2.4. Optimizing the model for the recovery of multiple channels for SAC is introduced in Sections 4.3.2 and 4.3.3. Additionally, the source separation possibilities of the proposed SAC framework are discussed in Section 4.3.5 based on the evaluations presented in [P5].

The second main application of blind sound source separation from a multichannel audio signal is introduced in Chapter 5. The chapter starts by presenting preliminaries of array signal processing in Section 5.2. The direction of arrival-based spatial covariance model, as proposed in Publication

[P6] and utilized in [P7], is presented in Section 5.4 along with an introduction to source separation using the NMF with spatial covariance estimation. The separation performance analyses from [P4], [P6] and [P7] are summarized in Section 5.6.

The research included in this thesis is concluded in Chapter 6, followed by a discussion of future work regarding improving the algorithms and possible other uses of the object-based audio signal models.

Chapter 2

Background

The auditory scenes we perceive are composed of multiple simultaneous sound sources. Their combination is called a mixture. The auditory event the mixture invokes is characterized by attributes such as the source radiation pattern, their location and the interacting environment affecting the sound propagation in the form of reflections and diffraction. Our hearing processes the auditory event and is able to resolve certain quantities for the sources, such as their direction of arrival (DOA). Additionally, our hearing is able to distinguish sound sources as separate entities based on their spectral content and timbre [94].

This chapter introduces the background of sound and audio as related to topics discussed in this thesis. The introduced concepts include sound propagation, capture using microphone arrays and the processing of the captured multichannel signal in sampled time-domain and frequency-domain. Additionally, the basics of hearing related to resolving the spatial location of sound sources and the perceptual relevance of sounds is presented.

2.1 Definition of an Audio Object

The term sound source refers to the physical entity causing the sound, whereas the sound object refers to the physical sound wave (movement of air molecules) the source causes. The sound sources we perceive and conceptualize vary, depending on the context. A concert might be considered as a single sound source and object when heard from a distance, whereas each instrument (and the crowd) can be considered separate sound sources as well. The relation of a sound object to an audio signal-domain object is defined in order to standardize the terminology used in the context of this thesis.

The term audio object has been conventionally used in the field of audio coding to denote different audio source tracks available for separate encoding. The spatial audio object coding (SAOC) methods [13, 53, 56], have been developed to cope and efficiently encode sources such as speech and different instruments. However, within the scope of this thesis considering the object-based representations blindly estimated from the mixture requires a broader definition of an audio object.

The recent popular matrix and tensor decomposition models (NMF and NTF) [19, 80] applied on the audio spectrogram [39, 131, 147] produce an object-based approximation of the audio signal. The representation is based on spectral templates (single or multiple time frames) and their time-dependent activation. These matrix and tensor factorization tools enable the extraction of meaningful repetitive content from an audio spectrogram, such as words, phonemes, notes and chords. It is justifiable to refer to these representations as being audio object-based and to consider the pair of spectral template and its time dependent activation as a fundamental audio object. The representation of the audio object thus spans the whole frequency range making it broad-band and it is estimated over the whole duration of the analyzed signal segment. The spectral content of the obtained audio object and the redundancy reduction ability of such a model in representing audio signal magnitude spectrogram is discussed in more detail in Section 3.1.

Including the NMF-based signal models in the category of object-based audio representations is justified by their powerful analysis properties proved by numerous applications based on them. Applications in the field of audio signal processing include single-channel source separation [128, 147], multi-pitch detection in score transcription [6, 130] and multichannel source separation [101, 121], [P6],[P7].

2.2 Hearing

The understanding of the operation of human hearing in resolving sound objects and their spatial properties is essential in building object-based models of audio signals. The field studying the perception of sounds is known as psychoacoustics and it tries to explain the relation of the physical stimulus and the auditory event it produces. The essential theories related to the topics of perceptual relevance, and spatial position of sound sources are introduced in the following sections.

The functioning of hearing as a physical process is well known through the studying of the anatomy of the ear and measurement of the neural activity caused by different stimuli. However, the way in which cognitive processing

of the auditory event occurs in the brain is less agreed upon. The auditory models [84, 152] are based on a knowledge of the physical process of hearing and can be used in predicting the perceptual relevance of individual sound components causing an auditory event [140]. This compact presentation regarding operation of the ear and perception of sound is based on literature from hearing research [93, 94], psychoacoustics [152] and neuroscience [5].

2.2.1 Ear Anatomy

The ear is divided into the outer, middle and inner ear and all have different characteristics regarding the perception of sound. The outer ear consists of the pinna and auditory canal which, ends at the tympanic membrane. The pinna causes a direction- and frequency-dependent filtering for sounds arriving from different directions; this is discussed in more details in Section 2.2.3 regarding spatial hearing. The middle ear is composed of the tympanic membrane, small bones called the ossicles and another membrane known as the oval window, which is located at the beginning of the cochlea. The air pressure in the auditory canal causes the tympanic membrane to vibrate, and its mechanical movement is further transmitted by the ossicles to the oval window. The inner ear consists of the cochlea, which is filled with fluid. The movement of the oval window causes the basilar membrane inside the cochlea to vibrate and causes a resonance to occur at a location proportional to the frequency of the sound. The hair cells attached to the basilar membrane bend due the resonance and cause neural activity which is then sent through the auditory nerve, ultimately ending up in the auditory cortex of the brain. The brain then resolves the auditory event based on the neural stimulus.

2.2.2 Critical Bands and Masking

Human hearing is known to operate on frequency bands known as the critical bands. Additionally, the perception of pitch is logarithmic, meaning that doubling the frequency of a tone is perceived as an equal increment, regardless of the reference frequency. In audio signal processing the concept of critical bands refers to the mapping of the linear frequency scale to frequency bands in a non-linear (logarithmic) relationship with bandwidth increasing towards the higher frequencies. The critical band decompositions include, for example, the Bark bands [152], and equivalent rectangular bandwidth [92]. Frequency bands with constant bandwidth in the Mel-frequency scale [137] are also widely used for critical band decomposition in audio signal processing. The logarithmic relation in specifying the bandwidth causes each obtained band to have equal informational relevancy.

Within the body of hearing research, the concept of critical bands is known as auditory filter [46]. It refers to a phenomenon where a complex sound composed of several neighboring tones (frequency components) creates an auditory filter around the most energetic tone and the neighboring tones are not resolved and thus become inaudible. This complex sound causes a joint neural stimulus that is not proportional to the linear combination of the amplitude or energy of the individual tones, but is rather determined by the most energetic tone.

The auditory filter can be explained by the physical properties of the ear [5]. The frequency-resolving ability of human hearing can be attributed to the different locations in the basilar membrane invoking neural activity, which is known as the mapping of the location of hair cells in the basilar membrane to the perceived frequency of a tone. Additionally, the basilar membrane is an elastic organ and its floppiness and width increase towards the end of the membrane where low frequencies resonate. In the case of a single tone, the resonance peak caused by the physical stimulus slowly attenuates towards neighboring locations and the hair cells of corresponding frequencies are also activated. Such a sound is perceived as a pure tone even though the hair cells corresponding to neighboring frequencies are bent by the resonance and this indicates that the neural stimulus from them is being suppressed.

The auditory filters are the reason for the masking phenomenon, i.e., faint tones can become inaudible if a stronger tone is present on the neighboring frequencies [152]. The masking and auditory filters behave similarly to the pitch perception, i.e., in the linear frequency scale the masking extends further towards higher frequencies and thus the auditory filters are wider at higher frequencies. The masking effect also continues for a short time after the sound causing the mask has ended and is then called post-masking. For an even shorter time period the masking occurs before the sound event and is called pre-masking. The utilization of the masking effect can be used in measuring the perceptual quality of processed audio [140] and thus is utilized in the perceptual coding of audio for specifying what information can be disregarded without altering the perceived sound [134].

2.2.3 Spatial Hearing

Spatial hearing, or alternatively binaural hearing, refers to the ability to sense the direction and spaciousness of different sound sources and is based on perceiving slightly different sound with different ears. The differences between ears can be characterized by following binaural cues, interaural time-difference (ITD), interaural level-difference (ILD) and interaural coherence

(IC). In determining the location of the sound source the spatial cues work in different frequency regions.

The ITD is utilized in the frequency range of 200 - 1500 Hz [152]. The time-difference is determined by locking on the phase of the signal [5]. The fact that the ITD does not work at frequencies above 1500 Hz is caused by the wavelength being less than the average distance between human ears (20 cm) and several cycles of such signal being able to fit within that distance. In such a situation the phase difference becomes ambiguous. It has been reported that ITD is also utilized in higher frequencies, through the process of detecting ITD from amplitude envelope changes [8].

At higher frequencies the ILD is used for localization of the sound source and is caused by head blocking the wavefront and attenuating high frequency components of the signal. Regarding low frequencies, the physical size of the head is not proportional to the wavelength, and causes no acoustical shade. The IC is considered to be the measure of how ambient and diffuse the perceived sound is.

Additionally, the pinna and its directional properties have a large role in spatial hearing. The localization by direction-dependent attenuation and amplification of certain frequencies is a learned ability, and every person's pinna causes a unique frequency response. The direction-dependent filtering property is especially used in locating sources in the region called the cone of confusion, where ITD and ILD are identical for sources located at the surface of a cone. All the cues working together and associated to one direction of arrival form the head-related transfer function (HRTF) for both ears for that given direction of arrival. By filtering anechoic signals with the HRTF and then listening them through headphones produces an illusion of binaural hearing, where the sound is perceived to originate from the direction associated to the HRTF that was used to filter the anechoic signal.

Binaural hearing is considered to be one of the most important factors in speech intelligibility within a multi-speech scenario [26, 51]. This ability to concentrate on a specific sound source within a noisy environment is known as the cocktail-party effect.

2.3 Sound and Recording

2.3.1 Sound Sources and Propagation

Sound is a pressure wave caused by a sound source. The sound wave propagates through air from its originating location to the observer which senses small changes in the atmospheric pressure. In free field propagation the sound

pressure is attenuated in proportion to the inverse of the distance between the sound source and the observer. Also absorption of sound energy by air causes attenuation [31].

An auditory scene is composed of multiple sound sources. Sound sources can have varying radiation patterns, from omni-directional point sources (start pistol) to line sources (highway noise). Additionally the auditory scene can include diffuse sounds that are considered as originating from all directions and thus do not have an interpretable origin.

With non-diffuse sounds the propagation is characterized by the direct path and interaction with the surrounding environment which consists of reflections, diffraction and scattering of the original source signal. The reflection is caused by the sound wave hitting a boundary with which it is not fully absorbed. In practice some amount of energy is always absorbed in the reflection. The phase of the reflected sound pressure wave is reversed. Diffraction is caused by the wavefront interacting with an obstacle or a slit which creates a new sound source through interference. The scattering of sound refers to a sound wave being reflected to a direction other than the opposite direction of arrival when interacting with an obstacle. All the above are generally referred to as reverberation in the case of indoor sound propagation. The amount of reverberation is characterized by reverberation time T_{60}, which indicates the time when the sound has attenuated 60 dB from the original sound pressure level [139].

The sound propagation and observing a sound event in outdoor and indoor environments differs in terms of reverberation and the amount of reflections occurring. Outdoors the reverberation consists mainly of first-order reflections from the ground whereas in indoors higher order reflections are observed. Typical floor, ceiling and wall materials do not absorb a significant amount of the energy of the sound wave and several interactions (reflections) with such low absorbing materials are required for 60 dB attenuation of the sound pressure level. A special indoor space is the anechoic room, where all surfaces are treated so as to absorb all sound energy and no reverberation occurs.

2.3.2 Microphones and Microphone Arrays

Sound is recorded using microphones which have various directivity patterns. The most typical types of microphones are omni-directional, dipole, cardioid and hyper-cardioid. Spatial audio signal processing with estimation and utilization of channel-wise properties requires using two or more microphones in the capture. Such a recording setup is called a microphone array.

Microphone arrays can be divided into spaced arrays and coincident arrays. Spaced arrays are composed of omni-directional microphones and no assumption on direction-dependent level difference between microphones can be made. In practice the body of the array (where the microphones are attached) causes an acoustical shadow, particularly at higher frequencies, and differences in sound level. In coincident arrays the microphones are directive (cardioid) and the level difference with respect to sound source direction of arrival can be utilized. The coincident arrays are used in field recording of X-Y stereo, B-format and other directional recordings. The coincident arrays can be used to determine the intensity vector which point in the opposite direction of the DOA of the sound wave recorded.

The signal processing with spaced arrays is based on observing time delays in terms of phase difference between the array elements. The microphone array signal processing is analogous to the processing developed for antenna arrays, but it involves few notable fundamental differences mainly due to the size of the aperture with respect to the wavelength of the electromagnetic versus acoustical waves.

Far Field Assumption

The propagation and observation in considered to happen in far field when the curvature of the wavefront is no longer a significant factor with respect to time delay between array elements and thus the sound wave can be considered to be a plane wave. An established rule of thumb for distinguishing near and far field propagation is defined as follows: the capturing scenario can be treated as far field when the receiving array is at distance $r > \frac{2D^2}{\lambda}$ from the emitting source, λ being the wavelength and D is the maximum distance between the array elements. The above definition originates from antenna array signal processing literature [136].

Contrary to electromagnetic waves at radio frequencies, in audio signal processing the recorded and observed frequency range (20Hz - 20 kHz) consists of several decades and corresponds to a wavelength of from several meters to a few centimeters. This makes use of the above mentioned rule difficult, i.e., the distance that fulfills the far field condition is much larger for high frequencies than for low frequencies of the audible spectrum. Without further scientific reasoning, in the context of this thesis far field wave propagation is always assumed. However, it is worth noting that there exist numerous spatial audio applications where near-field propagation and curvature of the wavefront needs to be considered, for example [18, 22].

Spatial Aliasing

The observed phase difference is unambiguous only up to the spatial aliasing frequency $f = \frac{v}{2D}$, where v is the speed of sound. It means that frequencies corresponding to a wavelength greater than half of the microphone spacing have multiple cycles between the microphones of the array and thus the phase difference is ambiguous.

The spatial aliasing causes several problems in array signal processing algorithms. Beamforming is one of the most prevalent applications of microphone arrays where the phase differences are used for aligning the microphone signals in time to enhance sounds originating from certain direction. The spatial aliasing causes amplification of undesired directions, as will be explained in more detail in Section 5.2.2. These problems are not usually faced in antenna array signal processing, since the relative bandwidth of the signal observed is much narrower, and antenna elements can be placed optimally at every half wavelength [139].

Microphone arrays in practical applications are usually small, at most several tens of centimeters in diameter, and the number of microphones is limited. These properties make them highly non-optimal at both ends of the usable audio frequency range. The ratio of the wavelength of the lowest frequencies compared to the overall aperture (size) is poor and spatial aliasing increases if the size is increased.

Practical Considerations

A comprehensive introduction of specialties in microphone array signal processing compared to antennas can be found in Chapter 5 in [139]. It covers topics such as the noise properties of the capturing environment, self noise of the array elements and computing power needed for a high number of microphones. Based on the aforementioned restrictions and the fact that the microphone arrays need to be of a size that allows for embedment in devices such as cellphones, laptops or placed on the desktop in a teleconferencing scenario, makes the spatial audio signal processing a challenging research area.

2.3.3 Sound Capture

Recording an auditory scene with one or several microphones captures the source signals $s_q(t)$ convolved with their spatial impulse response $h_{mq}(\tau)$. The recording is considered to happen by sampling the continuous change of air pressure at discrete time instances and thus the variables involved are

indexed by the time-domain sample index t and discrete sample delay index τ. The spatial impulse response is the transfer function for a sound source q at location $\mathbf{s}_q \in \mathcal{R}^3$ to each microphone m at location $\mathbf{m} \in \mathcal{R}^3$. The spatial impulse response $h_{mq}(\tau)$ incorporates all propagation attributes of a specific source, which are the radiation pattern of the sound source and its position with respect to the microphone array and the surrounding environment. In room acoustics the spatial response is referred as the room impulse response (RIR) and this thesis will use this notation hereafter when referring to any form of sound source capture and the associated spatial impulse response.

In measurement of RIR a known input signal is produced by a loudspeaker, and the output after interaction with the environment is measured using a microphone. Assuming that the system under analysis is linear time-invariant, the impulse response can be calculated based on the measured output and knowing the input. Over the years numerous methods for measurement of the RIR have been proposed [135] with the most prevalent being exponentially the swept sine [35] and the maximum length sequence (MLS) technique [124]. Regarding the topics covered in this thesis, the most important observation regarding measurement of the RIR is that it includes the directivity of the loudspeaker and the microphone used in capturing it.

2.3.4 Mixing Model

Audio sources in the discretely sampled time-domain behave similarly to their physical counterpart, i.e., source signals are additive, but instead of the movement of air molecules, the superposition principle is carried out by summing the source signals in each time instance in which they were sampled. The audio signal capture by $m = 1 \dots M$ microphones in a sampled time-domain can be given as

$$x_m(t) = \sum_{q=1}^{Q} \sum_{\tau} h_{mq}(\tau) s_q(t - \tau), \tag{2.1}$$

where the $x_m(t)$ is the mixture signal consisting of Q sources. The single channel sources $s_q(t)$ are convolved with their associated spatial impulse responses $h_{mq}(\tau)$.

2.3.5 Sound Source Separation

The field of sound source separation studies the estimation of the original source signals $s_q(t)$ from the observed mixture of $x_m(t)$. The separation methods can be based on single or multichannel captures, and conventional

methods were referred to in Section 1.2. In the case of single channel source separation, the algorithm usually relies upon supervised learning of the parameters of the sources in order to apply the learned model to unforeseen data. The multichannel source separation utilizes spatial array signal processing, i.e., observation of level and time-differences between the recorded channels [65, 118, 150].

In the development and comparison of the different sound source separation methods a set of objective metrics are used. The energy-based separation metrics proposed in [144, 146] are some of the most widely used and consists of the following: signal-to-distortion ratio (SDR), image-to-spatial distortion Ratio (ISR), signal-to-interference ratio (SIR), and signal-to-artefact ratio (SAR). The SDR measures how much the separated signal resembles the original signal from signal energy perspective, ISR measures the correct spatial positioning of the separated source, SIR measures the interference between separated sources and SAR denotes how many artefacts are added in the process of encoding and separation. Perceptually motivated mapping of the above metrics have been proposed in [30]. Perceptual speech quality metrics are mainly used in measuring enhancement of noisy speech and include perceptual evaluation of speech quality (PESQ) [114] and short-time objective intelligibility measure (STOI) [138]. These can also be used for measuring sound source separation quality.

2.4 Frequency-domain Representations

The human hearing operating on frequency bands for resolving a mixture of sounds is one of the most utilized concepts in audio signal processing. Thus spectral representations, the digital domain equivalent of frequency bands in hearing, are widely used as an input for audio processing algorithms. One of these is the discrete time short-time Fourier transform (STFT), which is obtained by dividing the time-domain signal into overlapping frames, applying an (analysis) window function and then calculating the discrete Fourier transform (DFT) of each windowed frame. In practice fast Fourier transform (FFT) is used in calculating the DFT.

The result of the STFT, calculated from a multichannel mixture signal $x_m(t)$, is time-frequency representation $\mathbf{x}_{il} = [x_{il1}, \ldots, x_{ilM}]^T$ defined for each frequency bin $i = 1...I$ in each analyzed frame $l = 1...L$ of length $N = 2I - 1$. The complex-valued result of the STFT is hereafter referred to as spectrogram and the absolute value of each of its elements is referred to as magnitude spectrogram. The argument of the STFT results in phase spectrogram.

In audio signal processing a time-frequency transform is said to be invertible if the original signal can be perfectly reconstructed by applying the inverse operations. The STFT produces a spectrogram which can be perfectly transformed back to the original signal. In case of the STFT the operations are the inverse DFT/FFT, application of a suitable synthesis window function [134] and the combining of the frames using the overlap-add method.

The perceptually motivated time-frequency representation can be obtained by grouping the magnitude spectrogram frequency indices $i = 1...I$ into critical bands. The most common divisions of frequencies into critical bands are Bark bands [152] or the use of triangular bands equally spaced on Mel-frequency scale [137]. It is worth noting that grouping the frequency bins causes information loss, and the original signal cannot be reconstructed from the critical band decomposed spectrogram. However, the result of an audio signal processing algorithm operating on critical bands can usually be applied to the linear frequency scale STFT and the output signal can be reconstructed without loss of information, as in, for example, time-frequency mask-based separation [50].

The STFT is a redundant representation due to the fact that the DFT is applied to overlapping data. Especially in audio coding this is not desired property and instead of the STFT the modified discrete cosine transform (MDCT) is used [134] as a spectral representation. The invertibility of the MDCT can be interpreted through the analysis-synthesis filter bank which it realizes and through its perfect reconstruction requirements [134]. More recently in musical signal analysis the method of constant-Q transform [15] and its invertible realizations [123] have been considered instead of the STFT.

2.4.1 Mixing in Frequency Domain

The convolution of sources and their spatial impulse responses in the time domain can be approximated in the STFT domain by a simple multiplication of DFTs of the source signal and its associated spatial impulse response. It equals out to instant mixing at each frequency bin individually and can be written as

$$\mathbf{x}_{il} \approx \sum_{q=1}^{Q} \mathbf{h}_{iq} s_{ilq} = \sum_{q=1}^{Q} \mathbf{y}_{ilq}, \tag{2.2}$$

where $\mathbf{x}_{il}$ is the STFT of the array capture. The mixing of each source q is denoted at each frequency bin i by spatial frequency response $\mathbf{h}_{iq} = [h_{ik1}, \ldots, h_{ikM}]^T$ and the source signals are denoted by their STFT s_{ilq}. The source signals as seen by the array are given as $\mathbf{y}_{ilq} = \mathbf{h}_{iq} s_{ilq}$. The approximation in Equation (2.2) is due to the fact that the effective length (before the

impulse tail is attenuated to negligible energy) of spatial impulse response h_{mq} in time-domain can be longer than its corresponding DFT $\mathbf{h}_{iq}$ of the length $N = 2I - 1$. Thus the spatial impulse response is truncated to the window length used in calculating STFT.

The frequency-domain processing of audio signals can be reasoned as originating from the analogue to human hearing and its ability to discriminate sounds by their spectrum. Additionally, certain properties of mixtures of sounds can be described and formulated in frequency domain, for example, the approximate W-disjoint orthogonality of speech [113], which indicates that the time-frequency information of multiple simultaneous speakers does not overlap.

Another benefit of processing audio signals in frequency domain is the approximation of the convolutive mixing process in Equation (2.1) by instantanous mixing in Equation (2.2). The instantaneous mixing indicates that each time-frequency point is linear combination of the source signals. This is utilized in, for example, using of ICA for sound source separation [61, 128].

2.4.2 Spatial Covariance Domain

In audio signal processing applications operating on channel-wise properties it is beneficial to operate with spatial covariance matrices calculated from the input signal STFT. The spatial covariance matrix in the context of this thesis is calculated from the square rooted STFT

$$\hat{\mathbf{x}}_{il} = [|x_{il1}|^{1/2} \operatorname{sign}(x_{il1}), \ldots, |x_{ilM}|^{1/2} \operatorname{sign}(x_{ilM})]^T, \tag{2.3}$$

where $\operatorname{sign}(\cdot)$ is the signum function. The covariance matrix of a single time-frequency point is obtained as the outer product

$$\mathbf{X}_{il} = \hat{\mathbf{x}}_{il} \hat{\mathbf{x}}_{il}^H, \tag{2.4}$$

where ^{H} stands for Hermitian transpose. The result of Equation (2.4) for each time-frequency point is the spatial covariance matrix of the observed STFT. The elements of matrix $\mathbf{X}_{il}$ for each time-frequency point encode the spatial behavior of the captured signal in the form of amplitude and phase difference with respect to each microphone pair. The use of square rooted STFT $\hat{\mathbf{x}}_{il}$ as proposed in [120] means that the diagonal of each matrix $\mathbf{X}_{il} \in \mathbb{C}^{M \times M}$ contains the STFT magnitudes $\mathbf{x}_{il} = [|x_{il1}|, \ldots, |x_{ilM}|]^T$. The off-diagonal values $[\mathbf{X}_{il}]_{nm}, n \neq m$ represent the magnitude correlation and phase difference $|x_{iln} x_{ilm}|^{1/2} \operatorname{sign}(x_{iln} x_{ilm}^*)$ between each microphone pair (n, m).

The mixing defined in Equation (2.2) can be expressed in the spatial covariance domain as

$$\mathbf{X}_{il} \approx \sum_{q=1}^{Q} \mathbf{H}_{iq} \hat{s}_{ilq}, \tag{2.5}$$

where $\mathbf{H}_{iq}$ is the spatial covariance matrix (SCM) for each source q at each frequency i and $\hat{s}_{ilq} = (s_{ilq}\overline{s_{ilq}})^{(1/2)} = |s_{ilq}|$ is the magnitude spectrogram of each source q. SCMs $\mathbf{H}_{iq}$ for all frequencies $i = 1, \ldots, I$ defines the mixing of qth source in spatial covariance domain.

Equivalence between Equations (2.5) and (2.2) is achieved by defining

$$\mathbf{H}_{iq} = \frac{\mathbf{h}_{iq}\mathbf{h}_{iq}^H}{||\mathbf{h}_{iq}\mathbf{h}_{iq}^H||_F} \tag{2.6}$$

and assuming that $||\mathbf{h}_{iq}||_1 = 1$. In practice the connection between $\mathbf{h}_{iq}$ and $\mathbf{H}_{iq}$ is not utilized and thus it is stated only for the completeness of the derivation.

It can be assumed that the source covariances are uncorrelated and thus the source magnitudes can be considered to be approximately additive, i.e., the diagonal of $\mathbf{X}_{il}$ is $|\mathbf{x}_{il}| \approx \sum_q |\mathbf{y}_{ilq}|$. In the context of this thesis the covariance mixing defined in Equation (2.5) is referred to as spatial covariance domain.

Expressing the instantaneous mixing defined in Equation (2.2) in spatial covariance domain by model in Equation (2.5) has several benefits. The source spectrum $\hat{s}_{ilq}$ is real valued and estimation of its absolute phase is not required in separation applications. Additionally the source mixing is expressed by the magnitude correlation and phase difference encoded in the covariance matrices $\mathbf{H}_{iq}$, which is known to be easily estimated by aid of time-difference of arrival (TDOA) estimation with a generalized cross-correlation method [139]. Many separation methods also rely on the estimation of the TDOA [65, 117], which is equivalent to the argument of the covariance matrix entries (phase difference). The benefits of defining the SCMs and source magnitude spectrum separately are made more evident in the context of complex-valued NMF for spatial sound source separation presented in Chapter 5.

Chapter 3

Object-based Models of Audio Signals

In digital audio signal processing the sampled time-domain signal is often transformed to a mid-level representation [29] in order to analyze it and to apply other signal processing operations. In typical applications a frequency domain representation such as the STFT introduced in Section 2.4 or critical band decompositions of it are used. The desired property of a mid-level representation for an object-based audio model is that they allow utilization of similar processing as known to be applied by human hearing for resolving different audio objects. Additionally, the invertibility of the representation is desired, which allows reconstruction of the time-domain signal by minimal added artefacts caused by loss of information.

The derivation of an object-based model in the context of this thesis consists of using two mid-level representations. The first is the STFT of the mixture, which is further approximated using the NMF, producing yet another mid-level representation composed of spectral bases and their activations. This chapter introduces the NMF and NTF models for the magnitude spectrogram and the complex-valued NMF for spatial covariance matrices given in Section 2.4.2. Additionally, frequency domain ICA for sound source separation is considered for contrasting the object analysis in the later discussion of source separation results in Section 5.6.

3.1 Spectral Redundancy and Spectrogram Decompositions

Natural auditory scenes are constructed of events that repeat over time, such as, individual phonemes in speech and identical notes of musical instruments. The magnitude spectrum, i.e., the absolute value of the STFT, of such an au-

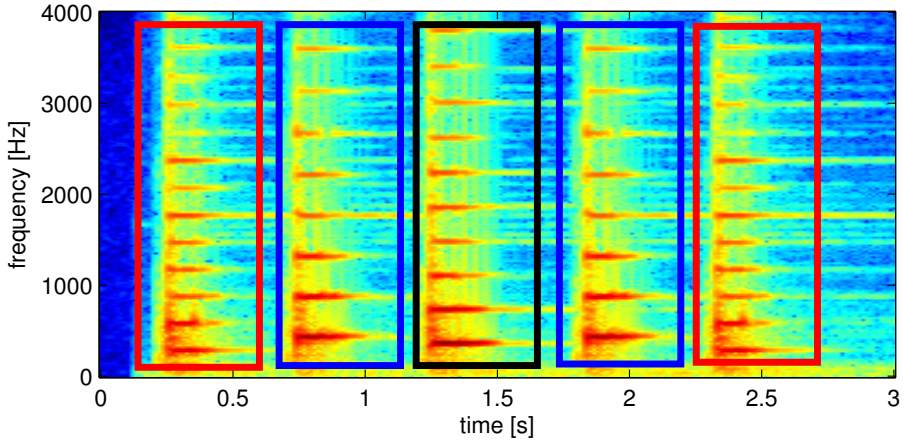


Figure 3.1: Example of magnitude spectrogram redundancy of piano notes. Repeating events are denoted by rectangle of same color, the note in the middle occurs only once.

dio signal is constructed of spectral patches that also have a repetitive manner and resemblance. This can be visually confirmed from Figure 3.1 depicting a few different notes played on a piano in repeating sequence. Additionally, the STFT is already redundant due to its definition having analysis frames overlapping, and thus DFT is being applied to partially redundant content. The object-based representation based on magnitude spectrum templates derived in the frequency domain also benefits from the shift-invariance property of the spectral transform, meaning that the absolute phase of the audio object or event can be disregarded.

The redundancy of the sound events can be utilized by a model that is composed of a few spectral templates and their time-dependent activation. The degree of the model, i.e., the number of spectral templates, restricts the modeling accuracy. Given that a certain degree of error is allowed in representing the spectrogram, similar yet slightly different sound events can be modeled using the same spectral template. Such a model reduces the redundancy in representing the overall spectral content. Additionally, the single spectral template representing redundant content can be interpreted to perform a separation at a very basic level.

The purpose of decomposing an audio spectrogram into objects is for the learning of underlying structures from the original data, which are beneficial for further signal processing stages. These stages include, for example, speech recognition [111], audio coding [145] and source separation [121, 147]. In the course of this thesis the decomposing techniques of ICA and the family of algorithms based on non-negative matrix and tensor factorization are utilized

and introduced.

Algorithms suited for monaural spectrogram decomposition rely on the machine learning approaches that learn redundant parts occurring over time in the spectrogram. The difference in decomposing monaural and multichannel spectrogram is that the latter includes inter-channel information regarding spatial position of the audio objects and events. The inter-channel information can be utilized in the decomposition algorithm instead of processing each channel individually. In spatial audio coding, the spatial properties of audio objects are referred to as spatial cues, namely inter-channel level and time difference, ICLD and ICTD, respectively. The decomposition methods can utilize either one [37, 39], [P3], [P5] or both [121],[P6],[P7], which can improve learning of audio objects by combining spectrogram evidence over multiple channels.

3.2 Independent Component Analysis

The most well known method for separating sources from multichannel audio is independent component analysis (ICA) [61] applied individually for each frequency of the STFT spectrogram [127]. The frequency-wise estimation of source spectrum and its spatial mixing causes frequency permutation problems, i.e., independent components belonging to different sources at different frequencies. After solving the permutation problem ICA can be considered to produce a representation that is interpretable as being object-based.

Different realizations of ICA aim at maximizing the statistical independence of extracted components by estimating a separation matrix $\mathbf{W}_i$ of size $Q \times M$ and $Q \leq M$ for each frequency i. The ICA decomposition can be given as

$$\mathbf{y}_{il} = \mathbf{W}_i \mathbf{x}_{il}, \tag{3.1}$$

where $\mathbf{y}_{il} = [y_{il1}, \ldots, y_{ilQ}]^T$ are the (unmixed) separated signals or in other words the independent components. The ordering of components $q = 1, \ldots, Q$ at each frequency of $\mathbf{y}_{il}$ is permutated and for time-domain signal reconstruction the inverse STFT cannot be applied directly. The ordering of components is aligned by a permutation matrix $\mathbf{P}_i$ of size $Q \times Q$. When multiplying a column vector or a matrix by $\mathbf{P}_i$ from the left, it changes the ordering of the rows. Thus it is applied to each frequency bin as $\mathbf{y}_{il} \leftarrow \mathbf{P}_i \mathbf{y}_{il}$ or $\mathbf{W}_i \leftarrow \mathbf{P}_i \mathbf{W}_i$ since $\mathbf{P}_i = \mathbf{P}_i^{-1}$.

Various methods have been proposed over the years to solve the permutation ambiguity. One of the first methods was based on maximizing the smoothness of the frequency response of the mixing filter [127]. Later methods considered temporal gain structure of the source signals [2, 122]

and TDOA interpretation of ICA mixing parameters [118], or both [122]. Solving the permutation problem by combining TDOA-based clustering of independent components aided with intra-source envelope analysis, and correlation maximization is presented in Publication [P4] of this thesis. Results of the permutation alignment from [P4] are presented in conjunction with the NMF-based sound source separation in Section 5.6.

More recently, methods inherently avoiding permutation ambiguity have been proposed and have been reported to have better performance over the conventional frequency-wise ICA in sound source separation. These methods include independent vector analysis (IVA) [81, 99] and ICA regularized over frequency [96]. However, some restrictions exist; the algorithm proposed in [96] allows no spatial aliasing and in [99] the algorithm is derived only for $M = Q = 2$. Methods estimating convolutive mixing such as TRINICON (Triple-N ICA for convolutive mixtures) [16] have gained success especially in blind deconvolution and multichannel source separation of time-domain audio signals [151]. Estimation of the convolutive mixing in time-domain causes no permutation problem.

3.3 Non-negative Matrix and Tensor Factorization

The family of algorithms known as non-negative matrix factorization (NMF), are based on a simple model that consists of a linear combination of basis functions and their activations. The non-negative constraint on the parameters makes the model purely additive and efficient algorithms for estimating optimal parameters have been developed [80]. The NMF decomposition of an audio signal magnitude spectrogram produces a dictionary of spectral templates that model redundant parts of the audio signal (in spectral domain) and thus is able to utilize long term redundancy in representing the signal. The NMF model of audio spectrograms have been utilized in a wide range of applications due to their ability to represent recurrent spectral templates using a single pair of basis function and its activation. The representation of the magnitude spectrogram of an audio signal using only several of such components allows, for example, easy labeling and classification of the spectral templates for separation, recognition and transcription types of applications.

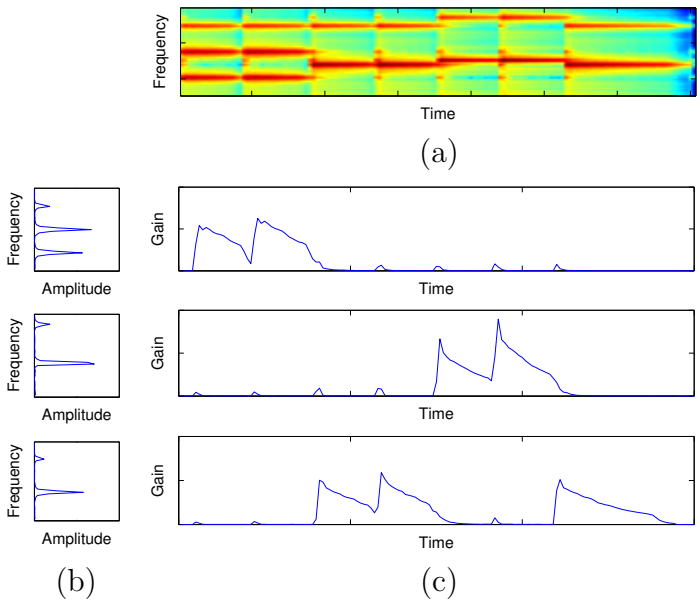


Figure 3.2: The NMF decomposition parameters estimated on spectrogram $\hat{x}_{il}$ illustrated in (a), NMF basis b_{ik} illustrated in left row (b) and component activations illustrated in right row g_{kl} (c)

3.3.1 Model for Single Channel Audio

Considering the magnitude of single channel STFT $\hat{x}_{il} = |\mathbf{x}_{il}| = [|x_{il1}|]$, the NMF model of it can be given as

$$\hat{x}_{il} \approx v_{il} = \sum_{k=1}^{K} b_{ik} g_{kl}, \quad b_{ik}, g_{kl} \geq 0. \tag{3.2}$$

The parameters of the model arranged into matrices $[\mathbf{B}]_{i,k} = b_{ik}$ and $[\mathbf{G}]_{k,l} = g_{kl}$ allows for interpretation as follows; the kth column of $[\mathbf{B}]_{i,k}$ contains the fixed spectrum of kth NMF component and corresponding row of $[\mathbf{G}]_{k,l}$ represents its gain in each STFT frame. The number of components used for approximating $\hat{x}_{il}$ is denoted by K, later referred to as the degree of the NMF model. The estimation of the optimal parameters b_{ik} and g_{kl} with respect to different optimization criteria is introduced in Section 3.3.3.

One NMF component with a fixed spectrum can only model part of the actual sound sources with varying spectral features. However, in most cases the NMF components represent meaningful entities that are interpretable as sound objects. Thus the entire NMF model consisting of several components serves as a mid-level representation for an audio signal based on audio objects. Such a structured model is utilized in applications such as the object-based

audio coding and source separation discussed later in this thesis. An example of a three-component decomposition of piano notes is illustrated in Figure 3.2 where it can be seen that each NMF component represents the harmonic structure of one note and that the components are active during the presence of associated spectral features observed in the mixture.

The required degree of the approximation denoted by K depends highly on the application and data to be processed. In the case of the NMF utilized in audio signal processing in a blind manner (without any training material), the appropriate value varies from several to tens of components. In supervised NMF, where spectral templates of known source types are learned from a large dataset of training material, the number of components can be several thousands up to tens of thousands. Concerning the representation of the audio magnitude spectrogram for audio coding purposes, the study of the required degree is discussed later in Section 4.2.

3.3.2 Model for Multichannel Audio

When dealing with multichannel data, there are two alternatives for defining and deriving an object-based signal model with the aid of matrix and tensor factorization, the first one being the non-negative tensor factorization (NTF), which is applied to the magnitude spectrogram of each channel stacked along the third dimension, producing a $I \times L \times M$ tensor with non-negative entries. Such a NTF model has been utilized in, for example, source separation [37, 40] and estimation of spatial position of spectral components [104]. The second approach in decomposing multichannel audio is factorization of the entire complex-valued STFT, incorporating both magnitude and phase. These types of matrix factorization algorithms differ greatly from the magnitude models and thus are discussed separately in Section 3.4.

The magnitude spectrogram of a multichannel signal is denoted as $\hat{x}_{ilm} = |\mathbf{x}_{il}| = [|x_{il1}|, \ldots, |x_{ilm}|]$. The NTF model consists of the same basic linear model introduced in Equation (3.2), composed of several fixed spectral bases and their corresponding gain. The added third dimension is modeled using a channel dependent gain for each component, resulting in the NTF model of

$$\hat{x}_{ilm} \approx v_{ilm} = \sum_{k=1}^{K} b_{ik} g_{kl} a_{km}, \quad b_{ik}, g_{kl}, a_{km} \geq 0. \tag{3.3}$$

The parameter a_{km} denotes the gain of NMF component k in each input channel $m = 1, \ldots, M$. In the above formulation the mixing of NMF components i.e., sound objects, is only considered to occur via level difference between channels, corresponding to the ICLD. The phase relation determined by ICTD is not considered.

3.3.3 Optimization Criteria and Parameter Estimation

Estimating parameters for the NMF and NTF models introduced in Equations (3.2) and (3.3) can be done with respect to different optimization criteria. In this section the most commonly used and general ones are introduced.

The optimization of the model parameters is done by iteratively updating them. The update equations are required to monotonically decrease the value of the cost function measuring the fit of the model to the observed data. The decrease of the cost function is referred to as convergence. In the vicinity of the local or global minimum the decrease in the cost function becomes small between iterations and the algorithm is considered to have converged. No globally optimal solution can be guaranteed to be obtained, since multiple local minima exist even in the simplest NMF model formulation and cost functions combinations.

Original work by Lee and Seung [80] includes multiplicative updates for two different cost functions, squared Euclidean distance (SED)

$$d_{SED} = \sum_{i=1}^{I} \sum_{l=1}^{L} (\hat{x}_{il} - v_{il})^2, \tag{3.4}$$

and Kullback-Leibler divergence (KLD)

$$d_{KLD} = \sum_{i=1}^{I} \sum_{l=1}^{L} \hat{x}_{il} \log \frac{\hat{x}_{il}}{v_{il}} - \hat{x}_{il} + v_{il}. \tag{3.5}$$

The convergence towards local minimum by the multiplicative updates proposed in [80] are proven using an auxiliary function technique familiarized by an expectation maximization (EM) algorithm [24]. The multiplicative updates in general tend to converge faster than additive ones based on gradient descent. Additionally, additive updates require determination of the step size for the update. The above-mentioned cost functions, with the addition of the Itakura-Saito divergence [64], are generalized in the work by Kompass [73]. The update rules for different non-negative models and modified cost functions can be obtained utilizing the framework proposed in [80, 116] for deriving the updates.

When linear operations are used to modify the cost function, such as weighting of each time frequency point of the model, new update rules can be derived from the unweighted ones by simple multiplicative operations. One of the most commonly used modified cost functions is weighted squared Euclidean distance (WSED), which is defined as

$$d_{WSED} = \sum_{i=1}^{I} \sum_{l=1}^{L} w_{il} (\hat{x}_{il} - v_{il})^2. \tag{3.6}$$

The term w_{il} contains linear weights corresponding to each time-frequency point (i, l) in the spectrogram. The WSED cost function is useful when different parts of the input signal spectrogram have unequal importance. This includes, for example, the perceptual relevance of the NMF approximation based on the masking phenomenon [152] discussed in more detail in Section 4.1.1 and in Publication [P1]. Several weighted NMF approaches exist for image classification [49] and audio separation [98].

3.3.4 Algorithm Description

The algorithm for estimating optimal parameters for the NMF and NTF models consists of a few common steps that can be used to describe all cost function variants and different formulations of the model. Given that the input signal STFT $\mathbf{x}_{il}$ and the cost function with associated update equations is specified, the process can be described in the following steps:

1. Initialization of the model parameters $(b_{ik}, g_{kl}, \dots)$ with random positive values uniformly distributed between]0,1].

2. Applying the update equation for each optimized parameter, for example with multiplicative form as proposed in [80].

3. Fixing the scaling of one of the parameters for numerical stability, which is then compensated by the rescaling of the other, for example l^2 -norm for the basis defined as $a_k = \left(\sum_{i=1}^{I} b_{ik}^2\right)^{1/2}, b_{ik} \leftarrow b_{ik}/a_k, g_{kl} \leftarrow g_{kl}a_k$.

4. Repeating steps 2-3 for a fixed number of iterations or until the chosen cost function does not change significantly between updates.

3.4 Complex-valued Non-negative Matrix Factorization

The multichannel microphone array audio signal encoding the spatial position of the sound sources by time-difference of arrival between the channels is not considered in the magnitude-based NMF models discussed in Section 3.3. This is due to the fact that the phase of the STFT $\mathbf{x}_{il}$ is not retained in the magnitude-only decomposition models defined in Equations (3.2) and (3.3). Considering the phase of the STFT in the object-based decomposition requires processing of complex-valued input data, as well as consideration of the additivity of the NMF components with different phases.

This section consists of an introduction to the complex-valued NMF model suitable for processing multichannel audio including the phase difference between the channels. The complex-valued NMF model and algorithm used in Publications [P6] and [P7] for spatial sound source separation are more closely investigated in Section 5.4 where the proposed direction of arrival-based model for spatial covariance matrices is introduced.

The first complex-valued NMF algorithms were introduced in [67, 101, 105]. In [67, 105] only a single channel complex-valued spectrogram is considered with an attempt to cover the exact additivity of NMF components by estimation of absolute phase of components for each time-frequency point. The problem in single channel estimation is that the absolute phase of an audio signal is not recurrently structured in the STFT domain and it is dependent of the exact time of the frame used for calculating the STFT. The multichannel complex valued NMF proposed in [101] considers the actual spatial covariance estimation and directly uses the EM algorithm framework for estimation of its parameters. In this thesis the concept and complex-valued NMF framework proposed in [119–121] is adopted for spatial sound source separation in a blind setting.

3.4.1 Component-wise Spatial Covariance Estimation

The complex-valued NMF model is formulated in the spatial covariance domain introduced in Section 2.4.2. The NMF framework can be used to model the spatial covariance observations and the spatial covariance domain mixing defined in Equation 2.5 as follows. The NMF magnitude model defined in Equation (3.2) is used for representing the real-valued mixture spectrogram $\hat{x}_{il} \approx \sum_q \hat{s}_{ilq} \approx \sum_k b_{ik} g_{kl}$, and the spatial properties, the SCMs $\mathbf{H}_{ik}$, are estimated separately for each NMF component k. This strategy is hereafter referred to as a component-wise SCM model. Such a model cannot be directly used for sound source separation, since several NMF components are needed for representing one actual acoustic sound source $\hat{s}_{ilq}$, and a separate NMF component linking strategy would be needed, as proposed in Publication [P6]. However, the component-wise SCM model serves as the most general complex-valued NMF decomposition for multidimensional signals.

The complex-valued NMF model with component-wise SCMs is defined as

$$\mathbf{X}_{il} \approx \hat{\mathbf{X}}_{il} = \sum_{k=1}^{K} \mathbf{H}_{ik} b_{ik} g_{kl}, \tag{3.7}$$

where $\mathbf{H}_{ik}$ are the SCMs for each NMF component k at each frequency index i. The above model and estimation of its parameters using auxiliary

functions as in the EM algorithm is introduced in [120]. The optimization criterion used is the squared Frobenius norm between the observed covariance matrices and the model given as

$$d_{FRO} = \sum_{i=1}^{I} \sum_{l=1}^{L} ||\mathbf{X}_{il} - \hat{\mathbf{X}}_{il}||_F^2. \tag{3.8}$$

Additionally, the model parameter estimation using KLD and Itakura-Saito divergence as a cost function is given in [121].

3.4.2 Source-wise Spatial Covariance Estimation

Regarding the complex-valued NMF model in Equation (3.7), two or more components modeling the same sound source will ideally end up having equal SCM properties (up to the estimation accuracy) determined by the spatial position of the source. An explicit parameterization of this underlying spatial property has been proposed in [36, 102] and utilized in [120] by introducing a component to source linking parameter to the model and only estimating a single set of SCMs for a group of NMF components. The component linking can be estimated in the same way as any other non-negative parameter of the model, and thus no separate clustering of NMF components for separation is needed. Additionally, the number of SCMs which need to be estimated at each frequency index decreases to the number of sources present in the mixture.

The complex-valued NMF model with source-wise SCMs is defined as

$$\mathbf{X}_{il} \approx \hat{\mathbf{X}}_{il} = \sum_{q=1}^{Q} \sum_{k=1}^{K} \mathbf{H}_{iq} c_{qk} b_{ik} g_{kl}, \tag{3.9}$$

where c_{qk} denotes the association of component k to source q. The association parameter c_{qk} is a non-negative scalar and the association is soft instead of binary, which means that a one NMF component can be associated to multiple sources with different weights. Also note that the SCMs denoted by $\mathbf{H}_{iq}$ are defined for the actual number of sources Q.

3.4.3 Interpretations of Complex-valued NMF

The interpretation of the complex-valued NMF model in Equation (3.9) based on the covariance mixing defined in Equation (2.5) can be described as follows. The diagonal entries of $\mathbf{H}_{iq}$ model the ICLD of each source q with

respect to the input channels, and are not time-dependent due to the assumption of stationary sources and mixing. Furthermore, the NMF magnitude model $\hat{s}_{ilq} \approx \sum_k c_{qk} b_{ik} g_{kl}$ for a single source denotes its overall spectrogram in all channels. The ICLD at the diagonal of $\mathbf{H}_{iq}$ represents frequency-dependent acoustical amplification or attenuation caused by the source positioning with respect to the capturing device. The off-diagonal values of $\mathbf{H}_{iq}$ model the cross-channel magnitude and phase difference properties of which the latter represents the ICTD between the input channels.

Calculating the input values of the algorithm, i.e., the spatial covariance matrices defined in Equation (2.4), from a single frame causes the observed covariance matrices to be rank-1 and positive definite Hermitian. Further, the definition of the SCMs $\mathbf{H}_{iq}$ in Equation (2.6) produces rank-1 and positive definite Hermitian matrices. Thus within a single observed STFT frame the covariance mixing is always rank-1, whereas in reality the actual spatial covariance matrices of a source can be full-rank. Given the definitions the magnitude differences at the diagonal would also determine the cross-channel magnitude differences, but this property is not taken into consideration in the covariance estimation of algorithms proposed in [120,121] or in the works of this thesis in Publications [P6],[P7].

The difference between the magnitude-based NTF model in Equation (3.3) and the magnitude part of the complex-valued NMF in Equation (3.7) is that the ICLD in the former is defined component-wise, whereas the complex-valued NMF model also allows magnitude differences frequency-wise. Frequency-wise modeling is more accurate when considering real recordings done using microphone arrays, where frequency dependent attenuation may exist, especially at high frequencies which are easily absorbed by physical obstacles, as discussed in Section 2.3.

Chapter 4

Object-based Audio Coding

OBJECT-BASED audio coding is a relatively new area of research within the field of audio coding. It aims at coding and representing separate audio objects that can be efficiently encoded and allow modification of the mixture content based on the objects at the decoding stage. In the coding of spatial audio especially, the benefits of such models include scalability in the number of channels and reconstructed content. The demand for an object-based spatial audio coding arises from the fact that the devices used for consuming the same audiovisual content vary from mobile devices to full-scale home theaters. The audio playback capabilities of these devices range from mono and stereo to multichannel loudspeaker arrays, thus requiring scalability of channel quantity at the coding perspective. Additionally, control over the reconstruction with respect to audio objects and sound sources allows novel entertainment aspects of mixed audio content.

4.1 Audio Coding Background

The field of audio coding studies the concepts of representing single or multichannel audio using a minimum amount of storage requirements and often aiming at minimal algorithmic delay required for real time streaming of the audio content. In lossy audio compression the aim is to reduce the amount of data needed for representing the signal without noticeably degrading its perceptual quality. The lossy compression methods are also known in the literature as perceptual coders [134]. The process of lossy compression can be seen as an optimization problem of minimizing the perceived distortions while reducing the amount of data needed for representing the signal. The basis of modern-day audio coding is enabled by a large combination of source

coding and signal processing techniques, such as perfect reconstruction filter banks, entropy coding, nonuniform quantization with dynamic bit allocation and psychoacoustic analysis.

4.1.1 Measuring Coding Distortions with Perceptual Criteria

Minimizing perceptual error in the encoding of audio requires modeling the human hearing in order to analyze which degradations in signal are disregarded by our hearing and which cause audible artefacts. Experiments for discovering the limits of our hearing in discriminating minor changes in physical stimulus has created a basis upon which to build models of our hearing [11, 12, 68, 84, 140].

The audibility of distortions in the processed signal is determined by estimating the masking threshold (see section 2.2.2) that the signal to be encoded creates and then comparing that to the level of the artefacts caused by the encoding. The psychoacoustic concepts used for building a perceptual distortion measurement criteria include critical bands [92, 137, 152], equal loudness contours [41] (Fletcher-Munson curves) and masking with temporal integration in the role of pre- and post-masking [152].

Noise-to-mask Ratio

One of the perceptual criteria used for measuring the audibility of distortions is the noise-to-mask ratio (NMR) as defined in ITU-R recommendation BS.1387 for perceptual evaluation of audio quality (PEAQ) [140]. The PEAQ is aimed at evaluating overall perceptual quality of different audio codecs by combining several audio quality criteria, including the NMR. The main part of the recommendation consist of defining and standardizing the masking level estimation, but it also includes definitions for audio quality criteria which are independent of it, such as overall distortion loudness, harmonic structure of the error and modulation measures. All the derived individual criteria are combined to obtain an overall perceptual score.

The NMR is defined as a measure between the reference and processed audio and in [140] it is defined as the ratio between error energy and the masked threshold in each analysis band corresponding to critical bands. The error is calculated as a difference in frequency domain magnitudes weighted by the outer- and middle ear model responsible for compensating for the different frequency sensitivity of hearing. The resulting error of each audio frame is grouped in the critical bands. The resulting error is considered as noise. The masking threshold is estimated from the reference signal.

The masking threshold estimations consist of the following steps. Before the actual mask estimation, the time-domain input signal level is scaled to match the average listening level and its STFT is calculated. The first step applied is the outer- and middle-ear modeling, which consist of frequency dependent weight similar to equal loudness contour. The weighted magnitude spectrum of each frame is grouped into critical bands with bandwidth of one quarter of a Bark band. An offset representing internal noise of the ear is added to each band. The spreading of the masking energy in each band is then applied, which equals to adding a proportion of the energy of each band to its neighboring bands. This process is referred to as simultaneous masking. In the last step the masking energy is spread in time, which corresponds to the pre- and post- masking phenomenon. The detailed operations of each step can be found from [66].

The perceptual audio coding methods such as MP3, AAC and OGG Vorbis [10,134] use the mask estimation in allocating bits for each frequency band in such a way that the audible distortions are minimized. A long and systematic research guided by listening tests has perfected the psychoacoustical models of the prevalent lossy compression algorithms.

4.1.2 Spatial Audio Coding

The multichannel audio codecs encoding each channel separately, such as AC3 [141], DTS [133] and DTS-HD, are bound to the fixed number of channels in reconstruction and bitrate increasing in proportion to the number of channels to be transmitted. These restrictions have been surpassed by parametric approaches developed for spatial audio coding (SAC) such as MPEG surround [55] and MP3 Surround [54]. These parametric approaches are capable of transmitting 6-channel audio (for 5.1 speaker configuration) with bitrates comparable to conventional perceptual encoding of stereo audio [56]. In this section the principles of parameterizing spatial audio are introduced which largely build the foundation for object-based spatial audio coding.

The parametric coding of spatial audio is based on a downmix-upmix framework, where the reconstruction of multiple channels is obtained from one or several downmixed channels using spatial cue parameters transmitted as auxiliary information. The multichannel SAC can be seen as a generalization of the binaural cue coding principles proposed in [4, 34]. The generalized block diagram of the SAC is illustrated in Figure 4.1. The synthesis of multiple channels is achieved from the downmixed signal by adjusting the level, time delay and decorrelation of each frequency band. The frequency bands usually follow the resolution of the critical bands of hearing [92, 152]. The above-mentioned process is based on an estimation of the

following spatial cues at each frequency band: inter-channel level difference (ICLD), inter-channel time difference (ICTD) and inter-channel coherence (ICC). The existing SAC codecs utilizing the above-mentioned spatial cues range from parametric stereo [125] to full multichannel codecs [33, 54, 55].

The field of SAC also covers approaches that try to cope with spatial analysis, coding and reproduction as a whole. The method known as directional audio coding (DirAC) [110] analyses the direction and diffuseness of frequency blocks using B-format recordings, which are usually used in research on ambisonics [23]. Combined with spatial impulse response rendering [88] the DirAC can not only encode the spatial sound field but can also separate sources from a mixture, based on their direction. The DirAC can thus be viewed as a method that is situated somewhere in the middle of fields like SAC, 3D sound synthesis and sound source separation.

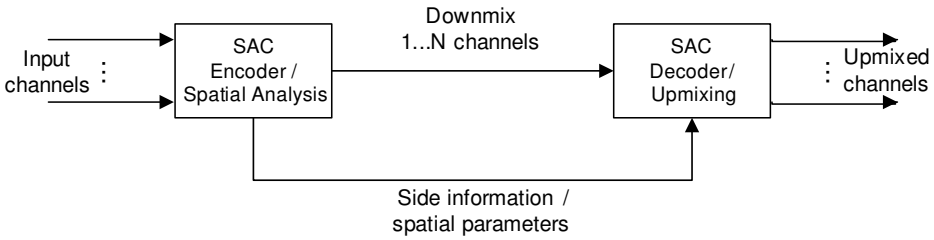


Figure 4.1: Block diagram of general spatial audio coding algorithm utilizing upmixing based on auxiliary information regarding spatial properties of each channel.

4.1.3 Spatial Audio Object Coding

Problem definition in spatial audio object coding (SAOC) starts from the assumption that the audio objects to be encoded are available as separate tracks in the encoding. Additionally, mixing information regarding the objects is available, which is used to construct one or more mixture channels. The encoding of a multichannel mixture of spatially positioned audio objects is an extension of the SAC approach, with the addition that the individual objects in each channel can be separately synthesized. The coding principles follow the SAC framework, where a downmix signal with additional auxiliary information is transmitted in order to recover the individual objects from the downmix.

The cues extracted and used for separating individual audio objects from the dowmixed mixture signal in SAOC [13] are object level difference, inter-object cross coherence, downmix channel level difference, downmix gains and object energies. A more detailed explanation of the object cues and their

quantization can be found from [13, 32]. These cues are estimated for 68 frequency bands in the case of a sampling frequency of 44.1kHz, but compared to the SAC the bands are more equally spaced regarding bandwidth, and a greater number of bands is used in total. This is due to the fact that separation of multiple audio objects from the mixture signal needs a higher resolution in determining which time-frequency block is assigned to which audio object. This may be validated by the approximate W-disjoint orthogonality of speech [113] sources for linear time-frequency domain and separation algorithms utilizing this property [65]. However, the degree of W-disjoint orthogonality of the audio objects decreases as the width of the frequency bands is increased, i.e., two sources may occupy the same frequency band with equal energy levels.

The quality of the object separation when completely removing a certain sound source such as vocals, is limited in terms of perceptual quality. For improving the separation a perceptually encoded residual can be transmitted and used as error correction of the standard object separation [13, 32].

One of the conceptual problems in SAOC is the lack of applications where the separate audio object tracks are available in the encoding stage. The usual case of a recorded or produced mixture of sound sources would require first solving the blind source separation problem. One of such methods is proposed in [53] where DirAC is used as a source separation tool and a preprocessing step for MPEG SAOC. The method proposed in Publication [P5] can moreover be viewed as solution both blind source separation and coding of spatial audio objects.

More recently a related field of informed source separation (ISS) [83, 106, 107] has been extended to the problem of spatial audio object-coding [47, 70, 103]. Many of these methods utilize NMF and NTF for parameterizing the audio objects [70, 82, 83, 103]. In addition to using NMF model for representing the sources, a joint framework for the encoding of the source parameters and the residual is proposed in [103]. The process of an NMF- and NTF-based ISS can be seen as identical to the method proposed in Publications [P3] and [P5] for multichannel audio coding via a dowmix-upmix framework.

4.1.4 Blind Upmixing

A related field to SAC is blind upmixing, which aims at creating additional channels containing ambient audio content [25, 63]. The ambience is usually placed in the back channels in the 5.1 speaker configuration. The problem of blind upmixing has also been addressed using the NMF [112, 142, 143]. In NMF-based ambience extraction the residual of the approximation is consid-

ered to contain the non-directive and ambient sound content. The operation of these methods can be interpreted as the NMF or NTF model to capture the content of the most prominent directive sound sources. Additionally, source separation by NTF as part of a mono-to-stereo upmixing approach has been proposed in [38].

4.1.5 Entropy Coding and Redundancy

The goal of general data compression is to reduce the redundancy in representing the signal to be compressed [115]. In lossless compression the amount of achievable compression is bounded on the lower end by the Shannon's source coding theorem, i.e., the entropy of the signal in question. The compression algorithms function by mapping symbols (patterns) from the data to the variable length codewords in such a way that the most frequent symbols are given the shortest codeword. The most well-known design of a predefined codebook (prefix code) based on known statistics of the data is Huffman coding [59], which guarantees an optimal codebook for the given symbols with given statistics.

The practical problem faced in data compression prior to the entropy coding stage is the analysis and recognition of the recurrent bit patterns that are considered as symbols. Audio in a purely binary form such as, for example, 44.1kHz / 16bit pulse code modulated (PCM), does not contain redundant parts, even though in terms of understanding music and speech they are constructed of repeating events. PCM audio is known to be rather incompressible by conventional all-purpose compression algorithms, but can be compressed by up to 50% using specially designed lossless audio compression algorithms such as free lossless audio coding (FLAC) (`xiph.org/flac/`).

In conventional lossy audio compression, perceptually motivated quantization combined with entropy coding is the most effective tool in reducing the size needed to represent the audio content without a significant perceptual loss. However, the algorithms do not utilize the semantic redundant structures of music and speech, as is used in other audio signal processing tasks such as automatic speech recognition (phonemes), automatic score transcription (notes and chords) and many more.

4.2 Channel-wise Object-Based Audio Coding Using NMF

This section introduces object-based audio coding utilizing NMF as proposed in Publication [P2] for the encoding of mono and stereo content. The benefits

of NMF-based signal representation for audio coding are evident and can be summarized as follows. The NMF representation is able to utilize long-term redundancy by modeling the signal over time using recurrent audio objects. This reduces the redundancy in representing the audio signal which is the foundation in all data compression. Additionally, the well known sparsity property of the NMF model, i.e., only a small number of components are active simultaneously, is beneficial for coding applications. The sparsity also holds for the NMF component spectrum, as is shown and investigated in Publication [P2].

The object-based audio coding of mono or stereo audio signals is not widely covered in the literature. One of such methods proposed in [145] constructs audio objects based on the harmonic structure of musical instruments. The SAOC concept and coding framework introduced in Section 4.1.3 also differs from the object-based coding covered in this section in that it assumes no separated sources present in the encoding stage.

4.2.1 Coding Framework Overview

In order to utilize recurrent structures in lossy audio compression, an object-based approximation of the original audio signal needs to be estimated. This can be achieved by estimating NMF approximation, Equation (3.2), of the signal to be encoded. The learning of the audio object structures and the NMF model require batch processing of an audio signal with a segment length varying from several seconds to tens of seconds. In the context of this thesis the proposed audio coding algorithms operate offline and a fixed 10 second segment is used for estimating the object-based model, and utilization of segment-wise information is not studied.

The entire compression algorithm based on an NMF spectrogram representation is illustrated in Figure 4.2. It consists of the following steps.

1. Calculating the STFT spectrogram of the input signal to be encoded and estimating the masking threshold it creates.

2. Magnitudes of the STFT are approximated using the NMF model given in Equation (3.2), with WSED as cost function as defined in Equation (3.6), and weighting corresponds to NMR as proposed in Publication [P1].

3. The NMF model parameters are quantized and entropy coded.

4. The phase of the STFT is quantized and entropy coded in a separate processing branch.

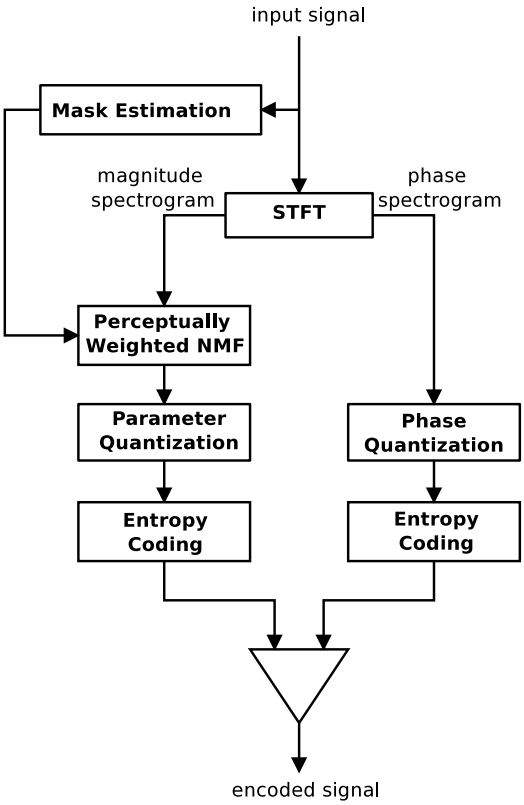


Figure 4.2: Block diagram of the NMF-based discrete channel audio coding algorithm.

5. The object-based magnitude model and quantized phases are multiplexed to a single bitstream.

The above process needs to be applied to each channel in the signal to be encoded, i.e., it does not utilize any cross-channel redundancies. The decoding is achieved by reconstructing the magnitude spectrum using the NMF parameters and combining it with a transmitted and quantized phase spectrogram. Inverse STFT is applied in order to obtain the time-domain signals.

The concept of perceptually motivated NMF model criterion and quantization and entropy coding of its parameters are more closely examined in Sections 4.2.2 – 4.2.4. The proposed framework is utilized in Publications [P3] and [P5], and is comprised of the SAC algorithm based on the NTF. Further studies in NMF-based coding developed by Ozerov, et. al. in [103] utilize a similar concept for combination of audio coding and informed source separation.

4.2.2 Perceptual Optimization Criterion for NMF

The development of object-based audio coding algorithms began with research on perceptually motivated cost function for the NMF proposed in Publication [P1]. In order to model perceptually relevant parts of the magnitude spectrogram with high precision, a cost function approximating human hearing was required, i.e., weighting the modeling error by the masking level. The perceptual relevance measured by NMR was introduced in Section 4.1.1. In [P1] the NMR criterion was formulated in the form of the WSED cost function in Equation (3.6), where a weight for each time-frequency point is applied. The NMR cost criterion for modeling the audio spectrogram using the NMF was studied in the M.Sc. thesis [97] of the author and the contributions relevant to the object-based audio coding using the NMF are cited in the following sections.

Implementing the NMR cost criteria as a WSED cost function and using the updates proposed in [9] for optimization of the model parameters requires inversing the operations specified in the NMR error signal calculation introduced in Section 4.1.1. The operations consists of weighting by the middle and outer ear transfer function and the grouping of the error into critical bands. Additionally, the masking threshold calculated in critical bands needs to be brought back to linear frequency scale. In Publication [P1] it is shown that these operations can be simply converted to a weighting matrix w_{il} for linear frequency scale error, as in WSED cost function in Equation (3.6). By using associated update formulas for the WSED, the NMR error of the overall approximation of v_{il} is minimized. The NMR cost function for NMF audio spectrogram modeling has been proposed in [98] but it evaluates the cost function in Bark bands, leading to effective resolution of the NMF model to be equal to the number of Bark bands used in masking estimation. More recently the coding-based ISS approach proposed in [70] has investigated the use of perceptual criteria, both in estimation of the NTF model and in the encoding of the residual. The findings favor the use of perceptual criterion over mean squared error by improving the separation quality measured with perceptual similarity metric.

4.2.3 Quantization and Rate of the Model Parameters

Quantizing the NMF Model Parameters

The quantization of the NMF representation of audio magnitude spectrogram was studied in [P2] and the following quantization scheme was proposed. The quantization of both NMF parameters b_{ik} and g_{kl} is based on non-uniform quantization, achieved by logarithmic compression before uniform quantiza-

tion. The function used to map the values of both parameters is the μ-law companding. The quantization scheme developed equals to using more quantization levels for the small values of b_{ik} and g_{kl}, whereas larger values are represented more coarsely. The justification of the non-uniform quantization is based on the perception of loudness being approximately logarithmic [152], i.e., doubling the perceived loudness requires amplification of the signal by a factor of ten. That leads to the fact that high magnitudes of the NMF model can be represented with less precision, since perceived differences are larger compared to at a lower magnitude level.

Regarding the proposed logarithmic mapping of the NTF parameters values, a related work in coding-based ISS [103] provides a more detailed investigation of optimal NTF model quantization based on techniques introduced in [71]. The analysis in [103] ends up with logarithmic compression with a scalar quantizer being optimal for encoding the NTF model, with an assumption of independence of the quantization error caused by quantizing individual parameters.

Parameter Rate and Effect on Perceptual Quality

The quantization of the NMF parameters decreases the overall perceptual quality of the NMF approximation. Assuming that the quantization of all elements of b_{ik} and g_{kl} is independent, meaning that the overall bitrate of the NMF model is determined by the total number of elements in b_{ik} and g_{kl}, thus minimizing the parameter rate while optimizing the perceptual quality before quantization is subject of great interest. This occurrence was studied in [97] and Publication [P1].

The perceptual quality of the non-quantized model is determined by the number of components used for approximating the spectrogram, given a fixed window length used for calculating the STFT. The results from [97] indicated that the NMR quality of the approximation increases almost linearly with respect to the number of components used for the NMF model. The window size affects the perceptual quality by determining the time resolution of the gains of the NMF model. Additionally, the window size of STFT determines the time resolution of the optimization criterion, meaning that the masking threshold is estimated for each STFT frame with a fixed duration. Transient sounds (for example, drums) may have a significantly shorter duration than the typical window size for STFT, varying from approximately 10 ms to 100 ms in audio signal processing applications. This leads to a situation where the masking threshold for a transient sound event is estimated from a signal content averaged from its neighborhood (the entire long STFT frame), and which may not reflect the masking properties at the time of the transient

sound. Therefore the STFT window size determines the time resolution for the perceptual optimization criterion and the NMF model. The shortening of the window size from 40 ms to 20 ms was found to improve the NMR quality by 2 dB [97], while substantially increasing the rate of NMF model gains g_{kl} due to the increased number of STFT frames per second.

The parameter rate and the perceptual quality of the NMF model can be concluded as follows. Increasing the degree of approximation increases the number of both parameters, i.e., the gains and the basis functions, whereas shortening of the window size increases the number of gain parameters g_{kl} per second and leads to better perceptual quality. In contrast, increasing the window size increases the number of frequency bins used to represent each NMF component basis b_{ik}, which means a greater number of parameters for the spectral basis part which in turn acts as an overhead.

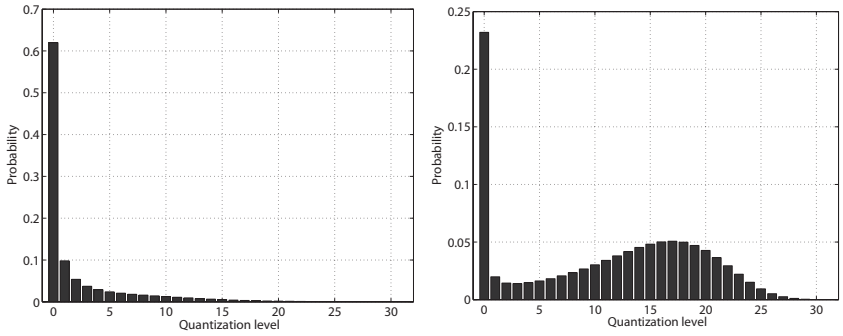
4.2.4 Sparsity of the Model Parameters

A signal representation is said to be sparse when only a few atoms of which it is composed of are activated at a time. The field studying sparse signal representations divides the process for obtaining the approximation into two stages, dictionary learning and sparse coding [28]. The methods used for these are, for example, K-SVD dictionary learning [1], and matching pursuit (MP) and orthogonal matching pursuit (OMP). NMF does not make an explicit difference in these stages, since alternating updates for all its parameters are derived using the same process. However, the parameterization is clearly divisible into NMF basis (dictionary) and their activations (component gains).

The similarity of NMF to non-negative sparse coding [58] applied for audio source separation [148] is evident. Even without any sparsity constraints applied to the NMF optimization criteria, the updates originally proposed in [80] produce rather sparse solutions when approximating an audio magnitude spectrogram.

Model Parameter Statistics

The probability distribution of the NMF parameters when compressed and quantized as described in Section 4.2.3 is studied in Publication [P2]. The collected probability distributions are reproduced in Figure 4.3, which clearly indicates the sparsity when parameters are quantized. Not only the activations but also the component basis (dictionary elements) are dominated by zero values. A large portion of the parameters without quantization would be non-zero, but their numerical value would be very small and negligible re-



(a) b_{ik} (b) g_{kl}

Figure 4.3: Probability distributions of the compressed and quantized values of the NMF parameters when 5 bits per parameter are used.

garding the approximation accuracy. This is proved by the small perceptual degradation of the proposed quantization as reported in Publication [P2].

The probabilities in Figure 4.3 were collected by processing a large dataset of speech and music from various genres which are all possible material for the object-based audio coding algorithm. The collected probabilities were used in building the prefix codes for entropy coding of the model parameters. The Huffman codes for each quantized value were generated leading to the shortest codeword associated for the zero value in both sets of parameters. The infrequent nonzero values end up having long prefix codes.

4.2.5 Results and Findings

The overall results achieved in channel-wise audio coding using the NMF as an object-based model for the audio magnitude spectrogram are provided in Publication [P2]. In general, the evaluation concentrated on the perceptual quality loss of the proposed quantization of the NMF model, and no direct comparison to conventional perceptual codecs, such as MP3, was reported. The test set consisted of 10-second signals of speech and music from various genres with the total number of test signals being 100. A detailed description of the encoding parameters and setting can be found from [P2].

Effect of Quantization and Entropy Coding

The decrease in NMR quality caused by the proposed quantization scheme with 5 bits used for representing each entry of b_{ik} and g_{kl} was less than 1 dB averaged over a large test set. The average NMR quality being approximately -14 dB indicates that the distortions are on average 14 dB below the masking

level. However, the instantaneous NMR for individual frames can be closer to the masking threshold and indicate the presence of audible distortions. With the reported NMR quality the overall bitrate using 5 bits for each parameter of the NMF model of one channel equals out roughly to bitrate of 64 kbps.

After the entropy coding the average bitrate was reduced to 50 kbps. For the encoding of a stereo signal, the representation of the magnitude spectrogram by the proposed object-based audio coding algorithm already requires a bitrate of 100 kbps. It should be noted that no mutual information and redundancy between the channels would be utilized in this case.

Encoding the Phase Spectrogram

The proposed object-based model is comprised of only the magnitude spectrogram and requires representing and encoding the phase of each time-frequency point individually. In publication [P2] the quantization and encoding of the phase spectrogram was based on a combination of using random phases for the highest frequencies due to their irrelevance in overall perceptual quality and different bit allocation schemes for the phases at lower frequencies.

The evaluation using random phases indicated that random phase content can be used at frequencies above 15 kHz with practically no perceptual quality loss when the magnitude spectrogram is represented using the NMF and quantized using the proposed scheme. Three different bit allocation methods were experimented with for quantizing the phase for the rest of the frequencies. The baseline consisted of uniform quantization with an equal number of bits allocated for each time-frequency point; the second approach used dynamic bit allocation based on the magnitude of the time-frequency point, which assumes that time-frequency points with high magnitudes are also perceptually more relevant from the phase representation point of view. The third approach used frame-wise differential encoding and dynamic bit allocation based on magnitude for encoding the residual. The dynamic bit allocation schemes were able to reduce the bitrate while maintaining the same perceptual quality as the uniform quantization, but none of the methods were found to be efficient enough for actual object-based audio coding comparable to conventional methods. The bitrates for the phase quantization for single channel were in the range of 70-80 kbps.

4.2.6 Conclusion on Channel-wise Approach

The overall channel-wise object-based audio coding algorithm based on the NMF model was not comparable to conventional methods in coding efficiency.

The shortcoming of the proposed audio coding framework is the utilization of an object-based model solely for representing the magnitude spectrogram of the audio signal to be encoded and not including object-based modeling and coding of the phase spectrogram. The time-frequency point-wise encoding of the phase spectrogram was found inefficient. Additionally, no cross-channel information in representing stereo signals was utilized, resulting in a linear increase in bitrate with respect to the number of audio channels to be encoded.

4.3 Multichannel Audio Upmixing by NTF

The multichannel audio downmix-upmix procedure, where multiple channels are recovered from a perceptually encoded time-domain downmix signal by filtering or other synthesis procedure, was introduced in Section 4.1.2. The term spatial audio coding (SAC) is often used to denote such methods. In the context of this thesis the term upmixing refers to the reconstruction stage of the process of first parameterizing the multichannel audio, downmixing and reconstructing the original multichannel audio by means of assigning the downmix signal content to multiple channels.

4.3.1 Spatial Cues and NTF

In the SAC framework the spatial parameterization was determined based on ICLD, ICTD and ICC. In the simplest case the spatial positioning of the audio object is determined only by the level difference between channels. If the audio object occurs on a single channel only, the effect of ICTD with respect to other channels becomes inherently encoded by the downmix signal phase. This assumes that the spectral content of audio objects does not overlap. Additionally, in the case of a 5.1 speaker configuration [86] and stereo downmix, the ICTD of the audio object present in the left and right front or rear channels becomes encoded by the stereo downmix. Taking into account the aforementioned special cases, it can be assumed that fairly accurate spatial upmixing can be carried out by only estimating the level differences of audio objects in each respective channel.

The NTF model as given in Equation (3.3) can be used to approximate the magnitude spectrogram of multichannel audio. The parameter a_{km} denotes the audio object occurrence with respect to each channel, making it a very natural object-based representation for the SAC. The quantization and encoding of the NMF model parameters introduced in Section 4.2 can be di-

rectly utilized in encoding the NTF model used for multichannel upmixing, as proposed in Publications [P3] and [P5].

The NTF model allows representing spatial audio objects that can overlap in frequency. Comparing the NTF model of multichannel audio to the conventional SAC approaches, which estimate only a single set of spatial cues for each time-frequency block, the NTF model estimates the cues over frequency for each audio object and allows accurate upmixing of simultaneous sound events overlapping in frequency.

4.3.2 Upmixing by Time-Frequency Filtering

The general approach in audio upmixing consists of estimating a compact representation of multiple audio channels and using it to synthesize the original content by filtering a downmixed signal containing a mixture of all the channels. The block diagram of the object-based SAC proposed in Publication [P3] is illustrated in Figure 4.4. The overview of the algorithm for obtaining the object-based model for the multichannel magnitude spectrogram and utilizing it for recovering the multichannel data by time-frequency filtering consists of the following steps.

1. The magnitude spectrogram $\hat{x}_{ilm}$ consisting of M channels is calculated from the input signal and corresponding perceptual weighting w_{ilm} is estimated from it.

2. The time-domain multichannel signal is downmixed to $\hat{M} < M$ channels and perceptually encoded using MP3, AAC or similar. For simplicity we assume a mono downmix ($\hat{M} = 1$).

3. The perceptually encoded downmix is decoded and its STFT is computed, denoted by d_{il}. It is used in optimizing the NTF model estimation for recovering the original channels from the downmix, not just for optimizing the perceptual quality in approximating the multichannel input $\hat{x}_{ilm}$. See section 4.3.3 for more details.

4. The NTF model defined in Equation (3.3) of the input signal $\hat{x}_{ilm}$ is estimated using WSED optimization criterion defined in Equation (3.6) with weighting w_{ilm} corresponding to the NMR [P1].

5. The parameters of the NTF model are quantized and encoded based on the framework introduced in Publication [P2] and extended to account for the channel-wise gain parameters a_{km}.

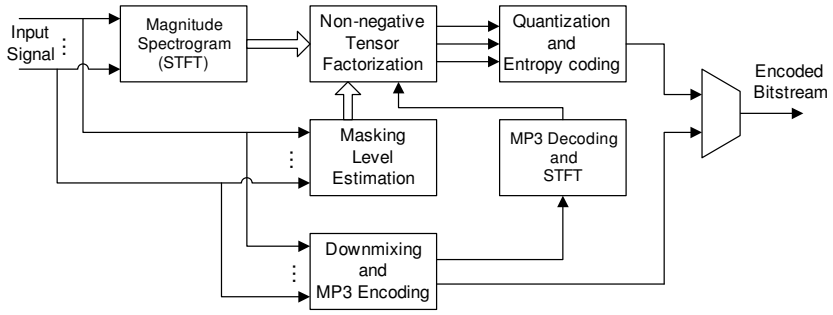


Figure 4.4: Block diagram of the SAC algorithm based on upmixing using the NTF model for the magnitude spectrogram.

In practice the downmixed signal is mono or stereo, but the algorithm presented in Publications [P3] and [P5] is extendable to any combination of input channels M downmixed to $\hat{M}$ channels.

In the decoding stage the NTF model based on the quantized parameters is reconstructed, the downmix is decoded and the STFT of it is computed. The upmixing by time-frequency domain Wiener filtering is implemented as

$$y_{ilm} = \frac{\sum_{k=1}^{K} b_{ik} g_{kl} a_{km}}{\sum_{m'=1}^{M} \sum_{k=1}^{K} b_{ik} g_{kl} a_{km'}} d_{il} = A_{ilm} d_{il}, \tag{4.1}$$

where y_{ilm} are the recovered channels and d_{il} is the STFT of the downmix signal. The term in the numerator corresponds to the NTF model for channel m. In the denominator one finds the sum of the NTF approximation over channels which corresponds to the STFT of the downmix. The above type of Wiener minimum mean squared error (MMSE) estimate is widely used in NMF- and NTF-based source separation [101]. The time-domain signals can be directly reconstructed from the upmixed STFT defined in Equation (4.1) by inverse DFT and overlap-add synthesis.

The process of upmixing using the NTF model and mono downmix is illustrated in Figure 4.5. In this example the first two channels contain a structured spectral content (piano), whereas the third channel contains a wideband noise-like signal (water fountain). The downmix magnitude spectrogram contains all of this spectral content overlapped. The upmixing masks A_{ilm} denote gains in the range of [0, 1] and the overall mask shape resembles the original multichannel data, but with the exception of also being able to handle rejection of spectrally overlapping content. This is especially evident in the upmixing mask of the third channel, which contains the spectral shape of the piano with reversed gain, i.e., rejecting it. In the last column the upmixed magnitude spectrograms before time-domain reconstruction can

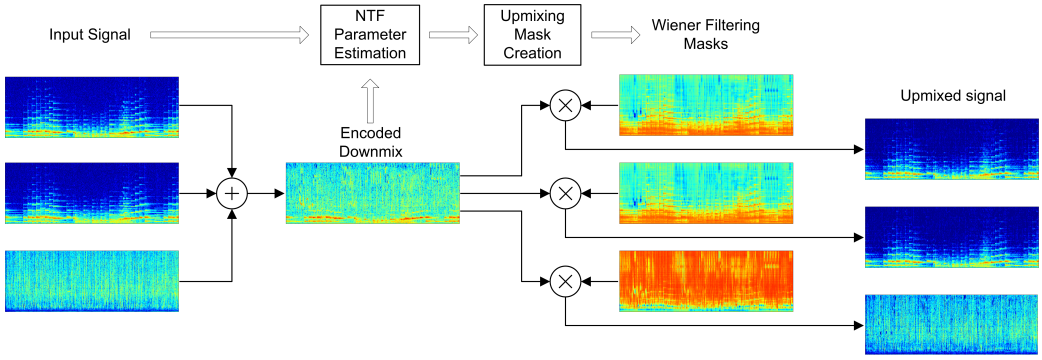


Figure 4.5: Process of upmixing based on the object-based NTF model of multichannel spectrogram and perceptually encoded mono downmix. The first column contains the magnitude spectrograms of each individual channel. In the second column is the magnitude spectrogram of the perceptually encoded downmix. The third column contains the Wiener filtering masks denoting the gain [0,1] for each time-frequency point. The combination of the second and the third column by element-wise multiplication realizes the Equation (4.1) and the upmixed channels can be seen in the last column.

be seen. There are only faint remarks from the upmix processing, namely some noise of the third channel added to the first two and vice versa. This equates to some crosstalk between the channels, which in general have little or no perceptually degrading effect due to the masking effect caused by the correctly upmixed content.

4.3.3 Cost Function for Upmixed Signal Perceptual Quality

In the above section describing the overall algorithm of the NTF-based SAC, it was mentioned that the STFT of the downmix is fed into the NTF parameter estimation stage. The downmix is used in optimizing the NTF for the upmix operation and not just for modeling the magnitudes of the input channels. This equates to the designing of the cost function and the model parameter update formulas so as to minimize the upmixed signal NMR.

The initial cost function for modeling the magnitude spectrogram of multichannel audio is the WSED introduced in Section 3.3.3, which is extended for multiple channels as

$$c = \sum_{i=1}^{I} \sum_{l=1}^{L} \sum_{m=1}^{M} w_{ilm}(\hat{x}_{ilm} - v_{ilm})^2. \tag{4.2}$$

The weights w_{ilm} correspond to NMR criterion [140] utilizing the formulation proposed in Publication [P1]. The masking level used in defining w_{ilm} is estimated separately for each channel, based on specifications in [66, 140] and thus no spatial unmasking effects are considered. Given that the actual upmixed output is achieved using Equation (4.1) the above cost does not minimize the upmixed NMR, which can be defined as

$$c = \sum_{i=1}^{I} \sum_{l=1}^{L} \sum_{m=1}^{M} w_{ilm} (\hat{x}_{ilm} - |y_{ilm}|)^2. \tag{4.3}$$

The difficulty of the upmixed NMR as a cost function is that the updates minimizing it cannot be derived directly, due to the fact that the parameters of the model are defined as nested and twice in the Wiener time-frequency filtering masks.

Optimizing the NTF model parameters using multiplicative updates obtained by partial derivation of Equation (4.2) causes the following errors when the estimated model is used for upmixing. If the spectrum of the audio objects from different channels overlap or are only closely separated in time and frequency the regular NMR cost function may allow too coarse of an NTF-magnitude model for them, which leads to crosstalk when the model is used for upmixing. All the audio objects being downmixed to one or two channels causes the magnitude spectrum to become less disjointed. The approximation accuracy of the NTF model corresponds to time and frequency selectivity, which needs to be increased at the time-frequency points where the downmix contains overlapping content.

The method for approximating the upmixed NMR cost function in Equation (4.3) was proposed in Publication [P3]. It consists of emphasizing the error in time-frequency points where the downmix STFT has a high magnitude with respect to the NTF channel sum in the denominator of the Wiener mask in Equation (4.1). Such time-frequency points are assumed to contain overlapping spectral content from audio objects in different channels. Using the proposed weighting, the NTF model is restricted so as to have smaller margin of error for spectrally overlapping audio objects. The proposed approximation can be implemented by modifying the weights according to

$$w_{ilm} \leftarrow w_{ilm} \frac{|d_{il}|}{\sum_{m'=1}^{M} \sum_{k=1}^{K} b_{ik} g_{kl} a_{km'}}. \tag{4.4}$$

There is no exact mathematical interpretation of the modified weights, and it cannot be guaranteed to minimize cost function defined in Equation (4.3), but as investigated in Publication [P3] the weighting was found to improve the upmixed NMR over the unmodified weights and has no negative effects.

The proposed weighting was applied only after the initial model estimation was obtained using the regular cost function in Equation (4.2) and its associated multiplicative updates for the first 500 iterations. This ensures that the initial values of the model parameters are close to optimal when starting to minimize the upmixed NMR, due to the fact that the developed approximation is not guaranteed to converge to the sought minimum. In Publication [P3] it was experimentally shown that the proposed approximation is able to converge to a smaller error measured using the cost function in Equation (4.3) when evaluated over the whole test set used in the listening test mentioned in that paper. The behavior of both cost functions in Equations (4.2), and (4.3) averaged over the test set, is shown in Figure 3 of Publication [P3]. After 500 iterations the weights are modified according to Equation (4.4). The analysis of the results can be summarized by the following observations

1. Both cost functions decrease and converge well up to the 500 iterations using the updates derived for the multichannel NMR in Equation (4.2) even though the rapid decrease of the upmixed NMR occurs later regarding the number of iterations.

2. After changing to weighting defined in Equation (4.4) the upmixed NMR again rapidly decreases but converges in a very few iterations. Due to the cost function behavior being averaged over all test signals, those containing less overlap of audio object spectrum do not benefit as much, whereas those with more overlapping spectral content as in, for example, the one illustrated in Figure 4.5, benefit from the proposed approximation significantly.

3. Lastly it is clear that the optima of the two cost functions are not equal; once the weights have been changed to approximate the upmixed NMR, the general magnitude NMR criterion start to increase in an almost exactly reversed manner.

4.3.4 Evaluation of Coding Efficiency

The overall encoding results of the NMF-based SAC coding method are reported in Publications [P3] and [P5]. The perceptual quality produced by the algorithms was evaluated using a methodology of multiple stimuli with hidden reference and anchor (MUSHRA) [85], which allows simultaneous evaluation of multiple methods using experienced listeners. The dataset used was the MPEG multichannel test samples intended for playback with a 5.1 speaker

configuration [86]. The complete description of the dataset and content of the individual samples can be found from Table 1 in Publication [P5].

The proposed object-based SAC method was compared to MP3 surround [54] based on conventional spatial cues estimated for fixed time-frequency blocks and stereo downmix. The target bitrate was set at 128 kbps, which was exactly produced by the MP3 surround. The bitrate of the proposed method consisted of an MP3-encoded stereo downmix at 96 kbps and a quantized and entropy-coded magnitude model producing 26 kbps using $K = 64$ audio objects with window length of 810 samples (20 milliseconds) for the STFT. The total bitrate of the proposed method thus equals to 122 kbps and more detailed parameter and bitrate definition are given in [P5].

Evaluation using the MUSHRA methodology and encoding parameters specified above indicated that neither of the tested methods were transparent compared to the reference, and the proposed object-based SAC algorithm produced a slightly lower perceptual quality than the MP3 surround baseline. However, the quality impairment being very small and the fact that the proposed method produced slightly lower overall bitrate indicated that the proposed method is a suitable alternative in encoding 5.1 multichannel audio at 128 kbps bitrate.

The benefits of the proposed model over the conventional SAC approaches are that the upmixed content can overlap in time and frequency without either a severe decrease in the perceptual quality of the upmixed content or increased crosstalk between the channels. The most important aspect of the object-based SAC, where audio objects are estimated blindly is that it also allows a separation of mixture of sound sources in one channel. The aspect of source separation with the proposed SAC algorithm is discussed in more details in Section 4.3.5.

One disadvantage of the proposed algorithm is the lack of estimating the diffuseness (ICC) or time-delay (ICTD) of the objects with respect to channels. This was identified as being one of the causes of lower perceptual quality in samples requiring time-domain decorrelation of upmixed channels to create perception of ambient and diffuse sound sources. In contrast, the test samples consisting of many point like sources were evaluated as being better than the ones encoded using the baseline method.

4.3.5 Audio Source Separation with NTF-based SAC

The studies on NMF and NTF representations as a basis for audio and sound source separation [37, 40, 147] have shown the potential of these spectrogram factorization methods, especially if a suitable supervised or semi-supervised clustering of the objects to entire audio sources are used [52, 128]. Given

that the clustering of NTF audio objects for entire audio sources can be determined, the separation of the original sources present in the multichannel mixture can be achieved. The performance of such source separation method using user created NTF component to source clustering is examined in Publication [P5], and the findings are reviewed briefly in the next paragraphs.

The proposed object-based SAC allows for manipulation of the reconstructed content by audio objects directly in the coding domain. Given that the clustering of the audio objects for the audio sources is determined on the decoder's side, the separation can be achieved without any additional bitrate in representing the source tracks individually. Assuming that the association of NTF audio object k belonging to an audio source q is known, a binary decisions (a scalar value of either zero or one) denoted by c_{qk} can be constructed to denote classification of NTF audio objects to entire sources. The reconstruction of individual source q constructed of subset NTF audio objects denoted by the clustering parameter c_{qk} is defined as

$$y_{ilmq} = \frac{\sum_{k=1}^{K} c_{qk} b_{ik} g_{kl} a_{km}}{\sum_{m'=1}^{M} \sum_{k=1}^{K} b_{ik} g_{kl} a_{km'}} d_{il}, \tag{4.5}$$

where the scaling of the clustering fulfills $\sum_{q,k} c_{qk} = 1$.

In Publication [P5] multichannel music samples consisting of mixture of instruments and vocals were generated and the sources, i.e., the different instrument and vocal tracks, were separated using user-defined NTF component clustering and the above formulation for reconstructing the source signals. The source separation occurs after the decoding process of the proposed SAC algorithm and thus the separated source signals also contain all the coding artefacts.

Separation Performance

The separation performance of the proposed method was compared to the ideal binary mask (IBM) separation, which determines binary clustering of a time-frequency point by dominance of energy in the original separated source signals. The separation performance measured by quantities introduced in Section 2.3.5 was reported in Tables 4 and 5 of Publication [P5].

The SDR score for IBM varied from a good 7-9 dB separation of more prominent sources such as the vocals, to a poor 1-2 dB separation of less energetic ambient sources. In the case of vocals the proposed method produced SDR scores 1.3 dB and 1.2 dB lower than the IBM separation, which can be regarded as an excellent result. Sources with intermediate separation with IBM, SDR being approximately 5 dB, the proposed method achieved

relatively good separation performance. For the most difficult sources the proposed method failed to produce meaningful separation. The other separation metrics followed the trend of the SDR differences between the proposed and the IBM separation. Most notably, the proposed method achieved a better SAR score in a few cases, indicating that the Wiener upmixing in Equations (4.1) and (4.5) produces less artifacts than the binary clustering of time-frequency points, as in IBM separation.

The proposed SAC method with user-guided clustering of the blindly estimated NTF audio objects for entire sources was concluded to be plausible for audio source separation, both in terms of separation quality and the relatively low amount of user input needed for annotating $K = 64$ audio objects for several audio sources. In the case of the proposed method the separated audio signals went through downmixing and perceptual encoding, whereas the IBM separation was applied for unprocessed multichannel audio signals and uses the original source signals for creating separation masks. The perceptual encoding of the downmix may lose some of the time-frequency details of the low energy sources, which may be attributed as one of the reasons for the proposed method performing worse in separation of such sources. Additionally, the separation and reconstruction of multichannel audio signals are both done by filtering a downmix, where the sparsity of the TF-points is decreased in comparison to the original multichannel audio signal, i.e., the original audio sources may occupy disjoint TF points in multiple channels whereas they can overlap when downmixed to stereo.

Connection to Informed Source Separation

The SAC application of the NTF algorithm presented in Publications [P3] and [P5] is highly related to the ISS problem definition. Let us consider the scenario where the audio objects to be encoded are all presented in their respective channel. In this situation the downmix contains the mixture of the audio objects and the NTF model-based upmixing is used in a sound source separation manner. The difference from the SAC approach is that each channel is not necessarily composed of a single source but rather a mixture of several sources. Such an approach is considered in, for example [83], where the encoding of the NTF parameters is achieved by embedding them in the mixture spectrogram to be sent to the receiver.

Chapter 5

Source Separation from a Multichannel Audio Recording

The separation of sound sources from the multichannel mixture is often required in many audio processing applications. One of the most well known is the preprocessing done in automatic speech recognition (ASR), which is set to enhance or separate speech from noise before the actual recognition task [87, 129]. When the result of the separation is interpreted by a human, as in source separation in teleconferencing [53], the perceptual quality of the separation is essential. Human listeners are known to be very critical regarding perceptual aspects of speech, i.e., artifacts in the separation.

The task discussed in this section is widely known as the cocktail party problem, i.e., given a multichannel recording of the party at a given location the goal is to separate individual sound sources. The source separation using multichannel array recordings and spatial cues is related to spatial parameterization of auditory scenes [88]. The separation methods proposed in Publications [P6] and [P7] combine the strengths of array signal processing for estimation of the direction of arrival of the source, and machine learning for the estimation of the redundant spectrum of sources.

The section is organized as follows. First, a problem definition and short review on related blind sound source separation approaches are given in Section 5.1. The preliminaries related to array geometry and array signal processing in general are discussed in Section 5.2. The more detailed explanation of the complex-valued NMF model as a multichannel separation method is given in Section 5.3. The direction of arrival-based spatial covariance matrix model proposed in [P6] and to be used in conjunction with the complex-valued NMF is presented in Section 5.4. The DOA estimation properties of the proposed spatial covariance model is presented in Section 5.5, and fi-

nally the separation performance of the proposed methods is presented and discussed in Section 5.6.

5.1 Problem Definition and Earlier Work

Considering multichannel capture and the time-domain mixing defined in Equation (2.1) the blind source separation (BSS) problem consists of estimating the source signals $s_q(t)$ and their convolutive spatial mixing $h_{mq}(\tau)$. The mixing can be estimated by observing the time difference of arrival between the input channel which is interpreted as the phase difference in frequency domain. In the context of this thesis we consider the separation of the sources in the frequency domain and the mixing model in Equation (2.2). In the frequency domain the BSS problem can be formulated as estimation of the source spectrogram s_{ilq} and the instantaneous mixing at each frequency $\mathbf{h}_{iq}$.

One of the most-used approaches for BSS is the ICA, as presented in Section 3.2. The arbitrary frequency-wise ordering of the ICs, referred to as the permutation problem, can be solved based on the phase difference of the independent components. However, the phase difference becomes ambiguous when the frequency exceeds the spatial aliasing limit.The permutation is generally problematic to be solved at a post-processing step together with the fact that the actual estimates of the source parameters are derived with no regularization over frequency. It means that the algorithm cannot benefit from the mutual information over frequency even though sources originating from the same spatial position are known to have the same direct path propagation properties. The ICA- and IVA-based approaches which avoid the permutation problem were discussed in Section 3.2.

The observed phase differences between microphones at each frequency can also be directly clustered to sources. The time-frequency point belonging to a cluster (source) can be used to create binary or soft separation masks, which can then be applied to the mixture to filter out the sources. Such methods include, for example, DUET [65] and bin-wise clustering [117].

More recently, methods based on NMF [19, 79, 80] have been used in the separation of monaural and multichannel mixtures. The methods for single channel separation are usually based on a separate clustering algorithm for solving the NMF component assignment to sources and mostly use supervised approaches [52]. The ability of NMF to find and to represent the mixture spectrogram using spectrally redundant components was discussed in Section 4.2 regarding single channel audio coding. The decomposition of mixture signal into redundant audio objects is as useful for the source separation and only requires finding the clustering for real acoustical sources.

The use of NMF in sound source separation from multichannel audio mixtures has been more recently studied in [3,37,40,101,119–121] where the NMF framework is extended to complex-valued inputs and paired with estimation of spatial parameters of sources. The methods share similar parameterization of the spatial properties of the sources by spatial covariance matrix (SCM) at each STFT frequency bin.

5.2 Array Signal Processing

The concept of microphone array was introduced in Section 2.3.2 and in this section the principles of microphone array signal processing are discussed. The topics include the estimation of time-difference of arrival (TDOA) and direction of arrival (DOA). Additionally the beamforming as a basis for enhancing signals originating from certain spatial direction is introduced. These topics lead to the proposed direction of arrival-based spatial covariance model introduced in Section 5.4.

5.2.1 Time Difference of Arrival

A source at a certain spatial location is distinguished by its DOA with respect to the microphone array. Assuming that the geometry of the array is known, the DOA can be translated to TDOA, which is further interpretable as a phase difference in the frequency domain. The purpose of the calculation of the ideal TDOAs is that the array can be steered towards the source of interest, which is done in its simplest form by a delay and sum beamformer (DSB). DSB time aligns and sums the microphone signals to enhance a certain direction. The process of estimating TDOA by assuming known DOA can be also seen in a reversed manner, i.e., observing TDOAs and estimating the most likely DOA which is causing the observations.

The work regarding microphone array signal processing included in this thesis assumes far field propagation. The wavefront-arrival direction corresponds to a set of TDOA values between each microphone pair and the TDOAs depend on the geometry of the array. The geometry of an array consisting of two microphones n and m located on the xy-plane at locations $\mathbf{n} \in \mathcal{R}^3$ and $\mathbf{m} \in \mathcal{R}^3$ is illustrated in Figure 5.1. In the illustration a look direction vector $\mathbf{k}_o$ is pointing towards the source at location $\mathbf{s} \in \mathcal{R}^3$ from the geometrical center $\mathbf{p} \in \mathcal{R}^3$ of the array. The geometrical center of the array is in the origin of the Cartesian coordinate system, i.e., $\mathbf{p} = [0, 0, 0]^T$, and the norm of the look direction vector is $||\mathbf{k}_o|| = 1$. Any given array geometry can be translated and rotated in such a way that its geometrical center is

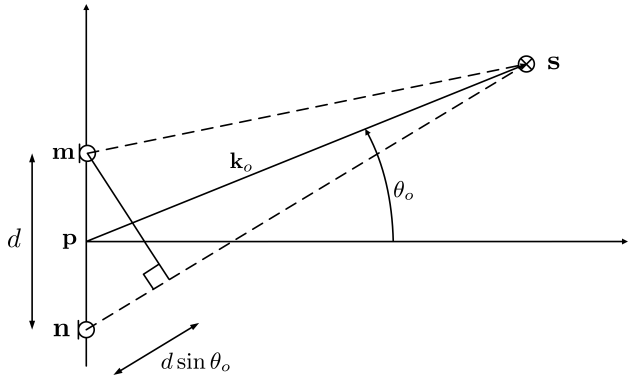


Figure 5.1: Geometry of an array consisting of microphones at locations **n** and **m** and look direction vector $\mathbf{k}_o$ pointing towards a source located at **s**.

located at the origin of the coordinate system, which is done to simplify the TDOA calculations. The different look directions are indexed by o and their direction is given in a spherical coordinate system using elevation $\theta_o \in [0, \pi]$, azimuth $\varphi_o \in [0, 2\pi]$ and fixed radius of $r = 1$. The reference axis $\theta, \varphi = 0$ of the array can be chosen arbitrarily.

With the above definitions and use of basic Euclidean geometry the TDOA of the microphone n with respect to array center point **p** in seconds is defined for direction $\mathbf{k}_o$ as

$$\tau_n(\mathbf{k}_o) = \frac{-\mathbf{k}_o^T(\mathbf{n} - \mathbf{p})}{v} = \frac{-\mathbf{k}_o^T\mathbf{n}}{v}, \tag{5.1}$$

where v is the speed of sound. The TDOA corresponds to a phase difference of $-j2\pi f_i \tau_n(\mathbf{k}_o)$ in the frequency domain, where $f_i = (i-1)F_s/N$ is frequency in Hz of ith STFT bin. F_s is the sampling frequency and N is the STFT window length.

The Equation (5.1) determined the time difference between microphone n located at **n** and array center **p** for a source at a direction $\mathbf{k}_o$. However, in spatial processing the interest is in the TDOA between microphone pairs, which can be given for a pair (n,m) as

$$\tau_{nm}(\mathbf{k}_o) = \tau_n(\mathbf{k}_o) - \tau_m(\mathbf{k}_o) = \frac{-\mathbf{k}_o^T(\mathbf{n} - \mathbf{m})}{v}. \tag{5.2}$$

Its corresponding phase difference can be given similarly as above and we denote such phase differences of all microphone pairs $n = 1 \dots M$ and $m = 1 \dots M$ using matrices of size $M \times M$ defined for each frequency index i and each direction o as

$$[\mathbf{W}_{io}]_{n,m} = \exp\big(-j2\pi f_i \tau_{nm}(\mathbf{k}_o)\big). \tag{5.3}$$

Hereafter in this thesis we denote these matrices $\mathbf{W}_{io} \in \mathbb{C}^{M \times M}$ as DOA kernels, since they denote the ideal phase difference for DOA hypothesis towards look directions $o = 1 \ldots O$.

5.2.2 Beamforming

The basis of beamforming is time aligning the microphone signals in such a way that they sum up coherently and thus emphasize the direction of interest. Additionally, different weights for each time-frequency point in each array element can be applied to further suppress unwanted noise sources in other directions. The field of beamforming is reviewed in this thesis due to its close relation to the spatial covariance model and separation method proposed in [P6]. The separation method can be seen as blindly estimating the direction of interest, aligning the microphone signals to enhance that direction and estimate a spatial post-filter to further enhance signals originating from that direction. The similarities of the complex-valued NMF for sound source separation and adaptive beamforming are discussed in more detail in Section 5.3.

The process of beamforming can be expressed in the time-frequency domain as obtaining a single channel enhanced signal from direction o as

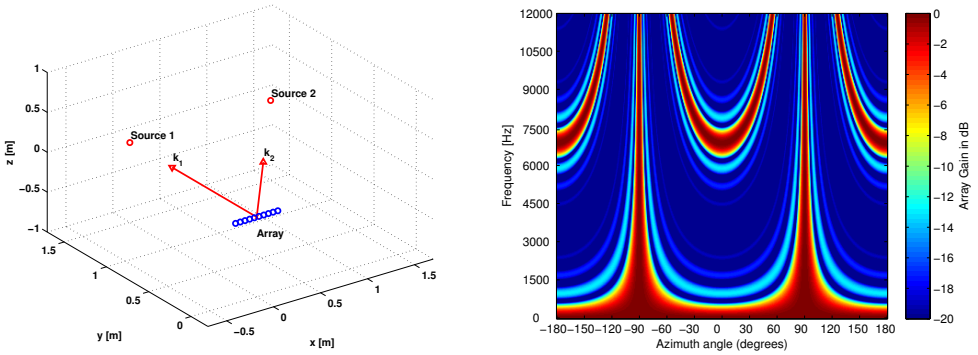
$$y_{il} = \mathbf{w}_{io}^H \mathbf{x}_{il}, \tag{5.4}$$

where $\mathbf{w}_{io}^H = [w_{io1}, \ldots, w_{ioM}]^T$ are the beamforming weights for each frequency index i and each sensor m towards look direction o.

Static Design

The simplest beamformer design is the DSB, which consists of time aligning and summing the microphone signals. The DSB beamforming weights can be obtained by taking the first row of matrix $\mathbf{W}_{io}$ specified in Equation (5.3), which results in $\mathbf{w}_{io}^H = [e^{-j\omega_i \tau_{11}(\mathbf{k}_o)}, \ldots e^{-j\omega_i \tau_{1m}(\mathbf{k}_o)}]$. DSB alignment correspond to a time delay caused by the target source DOA, i.e., beamformer look direction indexed with o, and any sound sources originating from this direction become enhanced. The DSB weights are also known as steering vector since they act as the basis of time-alignment for the adaptive beamforming designs.

The spatial selectivity of beamformers is measured using a directivity index, which is defined as the ratio between the array sensitivity and the average sensitivity in all other directions (surface of a sphere). In practice the directivity index can be regarded as a measure of the main lobe width with respect to suppression of all other directions.



(a) Scenario layout

(b) DSB beampattern

Figure 5.2: Example case of eight-microphone linear array with 5 cm spacing and DSB steered towards source 1 in broadside of the array. Zero attenuation of other directions occurs around 6.9kHz where the phase difference of all the microphone pairs become aliased.

Limitations of DSB include a poor directivity index [87]. The strength of DSB is its simplicity; only the array geometry and direction of the source need to be known and thus it is easy to implement as a preprocessing step to enhance the target source in almost any audio signal processing task.

All beamformers suffer from the spatial aliasing which causes amplification of unwanted directions. This is due to the fact that above the spatial aliasing frequency the phase difference corresponding to the given look direction has wrapped around the limit of $\pm\pi$. The delay according to Equation (5.2) corresponds to multiple directions at aliased frequencies, which all become enhanced when summing the time-aligned signal. The effect of spatial aliasing is illustrated in Figure 5.2 for an arbitrary example case, which consists of two sources and an eight-microphone linear array with spacing of 5 cm between microphones. The array is steered using DSB towards source 1 located in the broadside of the array. The array location, source locations and look direction vector are illustrated in Figure 5.2 (a) and the beampattern in Figure 5.2 (b). The beampattern depicts the beamformer gain and attenuation towards all direction and at all frequencies. The aliased sidelobe with 0 dB gain occurs around 6.9 kHz when the the wavelength of the frequency is equal to microphone spacing and has exactly the same phase difference as the steered direction of $-90°$. Additionally, smaller side lobes occur between the extremes, which is due to the fact that aliasing starts to occur in a subset of microphone pairs with a distance of larger than the 5 cm, but are fainter in terms of overall beampattern gain.

Another phenomenon related to beamforming and array signal processing, visible from Figure 5.2 (b), is that the main lobe widens towards low frequencies and in practice the directivity index of the DSB decreases to zero. It means that the DSB cannot enhance low frequency content from the steered direction. This is due to the fact that the aperture of the array in terms of distance between microphones is small with respect to the wavelength of the frequency it is observing. In the case where the frequency term in Equation (5.3) is small with respect to the TDOA term, which is the case with reasonably sized microphone arrays and low frequencies, the phase difference will be very small and not even large changes in the look direction and TDOA will produce significant change in terms of phase. This means that the low frequency content from the array perspective are in almost equal phase regarding the incident angle and aligning and summing will not cancel signal elements from other directions.

Adaptive Beamformers

The DSB introduced in the above section is only dependent on the source direction and is thus considered as a static design. However, the choice of the beamformer weights $\mathbf{w}_{io}^{H}$ allows frequency dependent suppression of the noise or unwanted source direction. The name adaptive refers to change in the beamforming weights over time and adaptation to time-varying statistics of the noise or interfering source.

One of these is the minimum variance distortionless response (MVDR) beamformer, which aims at minimizing the variance of the beamformer output while maintaining unit gain towards the given steered direction. It equates to placing null towards the direction of the interfering sources. This is done by estimating the second order statistics of the noise, i.e., its covariance matrix and the exact formulation can be found from, for example, [87, 139]. However, estimating the noise covariance matrix can be problematic, especially if it is changing over time. In practice it is done by updating the statistics of the noise while the source of interest is inactive which has to be estimated using a voice activity detector. The detection of target source activity is not necessarily perfectly accurate, especially in very noisy conditions. One alternative is to use a minimum power distortionless response beamformer, which uses the whole input signal covariance as such in place of the noise covariance matrix, but may easily cancel out the desired source content in the output of the beamformer.

The MVDR beamformer can also be implemented using a general sidelobe canceller structure [42,43,48,57], which uses a separate blocking matrix which tries to cancel out the interfering sources, but in practice also requires use of

voice activity detection in audio signal processing applications. One emerging approach in beamforming-based source separation includes spherical array beamforming [89, 90] with the advantage of spherical arrays having uniform directivity properties towards every look direction. Additionally, higher order statistics have been proposed for use in constraining the beamformer output, namely the maximum kurtosis beamformer [78] and the maximum negentropy beamformer [76, 77]. They are reasoned out through assumption of human speech characteristics following the super Gaussian distribution and thus the output of the beamformer focused on speech should output such statistics.

5.2.3 Direction of Arrival Estimation

In previous sections regarding beamforming it was assumed that the direction of the target source is known and thus the steering of the array can be applied to the said direction. Beamforming can also be used to estimate the target source DOA by following processing. The array is steered to a set of directions $o = 1 \dots O$ that sample the spatial space around it and the average output power from each direction over all frequencies and frames is calculated. The direction o producing the largest output is chosen as the estimate of the source DOA. The estimate is based on observing where the largest energy is emitted from the beamformers perspective.

A more common way of determining source DOA is by time delay estimation through searching out the maximum of steered response power (SRP) [139] function, which is defined for look direction $\mathbf{k}_o$ as

$$R(\tau_{nm}(\mathbf{k}_o)) = \int_{-F_s/2}^{F_s/2} \Psi(f) X_n(f) X_m^*(f) \exp(j2\pi f \tau_{nm}(\mathbf{k}_o)) df, \tag{5.5}$$

where $X_n(f)$ is the STFT of the signal of microphone n at frequency f and $\Psi(f)$ is the weighting for magnitudes. For simplifying the equations, the STFT of the microphone signal is given without frame index and with continuous frequency index, which differs from the earlier used discrete time STFT $\mathbf{x}_{il}$. The magnitude weighting used in most cases is the phase transform (PHAT) [17] which gives equal importance of phase difference at each frequency and is defined as

$$\Psi_{PHAT}(f) = \frac{1}{|X_n(f) X_m^*(f)|}. \tag{5.6}$$

The SRP-PHAT functions are combined over all microphone pairs to get an estimate of energy originating from direction $\mathbf{k}_o$ as

$$E(\mathbf{k}_o) = \sum_{n=1}^{M-1} \sum_{m=n+1}^{M} |R(\tau_{nm}(\mathbf{k}_o)|^2. \tag{5.7}$$

The look direction vectors $o = 1 \dots O$ define the search space for the source direction.

Alternatively, for calculating the directional energy in Equation (5.7) the generalized cross correlation (GCC) [72] function can used instead of SRP. The GCC refers to indexing only fixed sample delays (multiples of the sampling period) whereas the SRP function given in Equation (5.5) is defined for a specific delay $\tau_{nm}(\mathbf{k}_o)$.

5.3 Source Separation Using Complex-valued NMF

The complex-valued NMF model used for approximating multichannel STFT was introduced in Section 3.4, but its use in sound source separation was not discussed. In Section 4.3.2 NMF-based upmixing and a Wiener filtering of a mixture signal using the NMF parameters was introduced. The use of complex valued NMF for sound source separation follows the same framework. The source spectrum estimates obtained by the complex-valued NMF are used to construct a Wiener mask to filter out the sources from the mixture.

Considering the mixture model in covariance domain given in Equation (2.5) it is evident that being able to estimate the source magnitudes $\hat{s}_{ilq}$ and their corresponding spatial mixing in form of SCMs $\mathbf{H}_{iq}$ allows designing a Wiener filter to separate the sources from the observed mixture $\mathbf{x}_{il}$. Assuming that a complex-valued NMF model in Equation (3.9) of the observed mixture covariance is estimated, then the Wiener MMSE estimates of the sources are obtained as

$$\mathbf{y}_{ilq} = \frac{\sum_{k=1}^{K} c_{qk} b_{ik} g_{kl}}{\sum_{q'=1}^{Q} \sum_{k=1}^{K} c_{q'k} b_{ik} g_{kl}} \mathbf{x}_{il}. \tag{5.8}$$

Alternatively, a multichannel Wiener filter can be used, which also utilizes the estimated spatial covariance information. The multichannel Wiener filter based on the complex-valued NMF model is given as

$$\mathbf{y}_{ilq} = \sum_{k=1}^{K} c_{qk} b_{ik} g_{kl} \mathbf{H}_{iq} \hat{\mathbf{X}}_{il}^{-1} \mathbf{x}_{il}, \tag{5.9}$$

where $\hat{\mathbf{X}}_{il}$ is the model according to Equation (3.9). The structure of filter in Equation (5.9) corresponds to an MVDR followed by a single channel post-filter [126]. The estimated source SCMs $\mathbf{H}_{iq}$ are in the role of the MVDR weights and the magnitude model constructs the single-channel Wiener post-filter. The Wiener estimates can be defined similarly to the complex-valued

NMF model with component-wise SCM estimates given in Equation (3.7) corresponding to reconstruction of the individual NMF components in an MMSE sense.

Estimating $\mathbf{H}_{iq}$ in such a way that it corresponds to a single source at all frequencies requires an algorithm that operates jointly over frequencies and thus ties together possible aliased phase differences. In source separation with complex-valued NMF [120, 121] the estimation of source SCM assumes that the NMF magnitude model enforces $\hat{s}_{ilq}$ to correspond to a single source, and thus estimating $\mathbf{H}_{iq}$ yields an estimate that corresponds to a single source over frequency. However, it is not guaranteed that each NMF component models spatially coherent audio objects. For example, two audio objects having similar spectral characteristics may become modeled by the same NMF component even though they reside at different spatial locations.

A related method proposed in [27] introduces a direct estimation of full rank SCM of the source and its magnitudes at each frequency, but requires solving the frequency-wise permutation. It is further developed by combining the full rank SCM estimation with the NMF magnitude model for the sources [3], which similar to as in [120] avoids the permutation ambiguity by assuming the NMF components to be spatially coherent. A direct investigation on whether the assumption is violated is difficult. However, the evaluation of the method proposed in [120] against the one proposed in Publication [P6] indicates better separation performance by unifying the SCM properties over frequency by the TDOA of the direct path of the source. The separation performance is further discussed and investigated in Section 5.6.

5.3.1 Spatial Covariance Matrix Estimation

Given the complex-valued NMF model in Equation (3.9) the source-wise estimation of the covariance properties as proposed in [120] is achieved using update

$$\mathbf{H}_{iq} \leftarrow \mathbf{H}_{iq}\Big[\sum_{l,k} c_{qk} b_{ik} g_{kl} v_{il} + \sum_{l,k} c_{qk} b_{ik} g_{kl} \mathbf{E}_{il}\Big], \tag{5.10}$$

where $\mathbf{E}_{il} = \mathbf{X}_{il} - \sum_k \sum_q \mathbf{H}_{iq} c_{qk} b_{ik} g_{kl}$ is the error of the model and $v_{il} = \sum_k \sum_q c_{qk} b_{ik} g_{kl}$ is the approximation of the mixture source magnitude spectrum $\hat{x}_{il} \approx \sum_q \hat{s}_{ilq}$.

By investigating the update in Equation (5.10) and the model in Equation (3.9) it is clear that the updating of the SCM $\mathbf{H}_{iq}$ of each source q is done frequency-wise. Contrasting this to the spatial covariance domain mixing in Equation (2.5), it can be stated that the only thing tying the estimated

source SCMs over the frequency is the NMF-based source magnitude model, defined as $\hat{s}_{ilq} \approx \sum_k c_{qk} b_{ik} g_{kl}$.

5.4 Direction of Arrival-based Model for Spatial Covariance

The problem in estimation of the spatial mixing of a source concentrates on the fact that the observed evidence, phase difference, is dependent of the frequency and starts to alias at rather low frequencies with microphone arrays of practical size. In section 5.2.1 the connection between source DOA and phase difference was drawn by the aid of TDOA, meaning that the direct path TDOA explains the observed phase difference even with the aliasing. Additionally, as seen in Section 5.2.2 regarding beamforming, this property was utilized allowing the beamformers implemented in a frequency domain to be able to integrate the phase difference over the whole frequency range. A similar concept can be adopted to the SCM estimation in complex-valued NMF framework by finding which direct path TDOA defined in the frequency domain explains the observed phase difference evidence the best. This is analogous to directly estimating the TDOA by observing phase differences over the whole frequency range.

The above-described concept has not been utilized in the sound source separation earlier, due to the fact that it is difficult to include in the parameter estimation. In Publication [P6] the frequency independent estimation of spatial properties of sources was achieved by estimating non-negative weights for DOA kernels defined in Equation (5.3) containing phase difference for look direction vectors sampling the spatial space around the array. The parameter estimation framework was based on techniques introduced in [120] and the proposed SCM model was further utilized in Publication [P7].

5.4.1 SCM Model by Superposition of DOA Kernels

The direction of arrival-based SCM model proposed in Publication [P6] is based on superposition, i.e., weighted linear combination, of the DOA kernels defined in Equation (5.3). The look directions of the kernels are set to approximately uniformly sample the spatial space around the array. The SCM of a point source in anechoic capturing conditions could be described by a single DOA kernel, which is analogous to the direct path propagation in reverberant conditions. Due to the echoes and diffractions from surfaces and objects in regular capturing conditions, a combination of several direct paths is proposed and for each direction a non-negative weight is estimated.

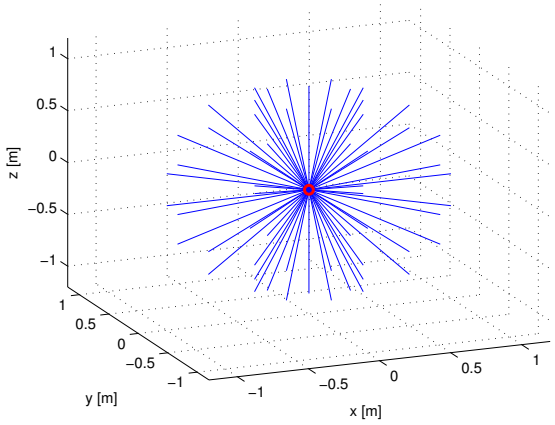


Figure 5.3: Look direction vectors approximating uniform sampling of the unit sphere around the geometric center of the array illustrated by the red circle; discretization of look directions is sparse for illustrative purposes.

The weight of each DOA kernel describe the signal power originating from said direction and is comparable to DSB output power to the same direction.

The look directions vectors $\mathbf{k}_o$, defined in Section 5.2.1, are now set to spatially sample the surface of a unit sphere set around the geometrical center $\mathbf{p}$ of the array. A sparsely sampled grid is illustrated in Figure 5.3 and in the real implementation a more densely spaced grid is needed to sample the continuous DOA space around the array. Given the DOA kernels $\mathbf{W}_{io}$ for each look direction $o = 1 \dots O$ according to Equation (5.3), the proposed SCM model based on their superposition is given as

$$\mathbf{H}_{iq} = \sum_{o=1}^{O} \mathbf{W}_{io} z_{qo}, \tag{5.11}$$

where z_{qo} are the direction weights corresponding to each look direction.

The proposed model can be placed back into the spatial covariance domain Equation (2.5) to obtain

$$\mathbf{X}_{il} \approx \sum_{q=1}^{Q} \mathbf{H}_{iq} \hat{s}_{ilq} = \sum_{q=1}^{Q} \Big[\sum_{o=1}^{O} \mathbf{W}_{io} z_{qo} \Big] \hat{s}_{ilq}. \tag{5.12}$$

As described in Section 5.3, the aim is to estimate $\mathbf{H}_{iq}$ in such a way that it is spatially coherent and thus corresponds to a single source over all the STFT frequency bins. The direction weights z_{qo} in the proposed model are independent of frequency which makes the estimation of the entire SCM

$\mathbf{H}_{iq}$ being optimized over all frequencies. The frequency dependencies of the phase differences are taken into account in the definition of the DOA kernels and the estimation can be seen as finding the most probable DOA in terms of how well TDOA of each look direction explains the observations. The direction weights z_{qo} are restricted to being non-negative and can be estimated using multiplicative updates based on techniques proposed in [120, 121]. The magnitude model is based on the NMF components and will be presented in the next section.

5.4.2 Complex-valued NMF with the DOA-based SCM Model

The complex-valued NMF model with the DOA-based SCM model is simply obtained from the source-wise SCM model defined in Equation (3.9) by replacing the source SCMs $\mathbf{H}_{iq}$ with the proposed model $\sum_{o=1}^{O} \mathbf{W}_{io} z_{qo}$. The entire model can be then written as

$$\mathbf{X}_{il} \approx \hat{\mathbf{X}}_{il} = \sum_{q=1}^{Q} \sum_{k=1}^{K} [\sum_{o=1}^{O} \mathbf{W}_{io} z_{qo}] c_{qk} b_{ik} g_{kl}. \tag{5.13}$$

Comparing the above model to the covariance mixing in Equation (2.5) it can be identified that the NMF magnitude model represents source spectra $\hat{s}_{ilq} \approx \sum_k c_{qk} b_{ik} g_{kl}$, and all the spatial properties are encoded by the kernels $\mathbf{W}_{io}$ and the direction weights z_{qo}.

In [P6] the DOA-based SCM model is defined with NMF component-wise SCMs and it omits the estimation of the clustering c_{qk} within the algorithm. Thus a separate clustering strategy for linking the NMF components to entire acoustical sources is required. The model defined in Equation (5.13) follows the formulation proposed in [P7] with source-wise SCM estimation. It is more generic due to producing direct estimates of the source spectra $\hat{s}_{ilq}$ which can be used to obtain Wiener estimates of the sources as defined in Equation (5.8). The source reconstruction produces estimates of the sources as seen by the array, i.e., convolved with their spatial impulse responses. The time-domain signals are obtained by inverse STFT of $\mathbf{y}_{ilq}$.

The proposed method is not free from the spatial aliasing and the phase difference in DOA kernels are ambiguous above the spatial aliasing frequency. This causes side lobes in the spatial gain structure if the DOA kernels are used as DSB steering of the array signal. However, the NMF magnitude model acts as a spatial post filter which already attenuates time-frequency bins that do not fit the spectral structure of the source.

Estimating the DOA-based SCM model and the NMF model for magnitudes in turn enforces the NMF components to be spatially coherent due to the fact that the SCMs are constrained to being a linear combination of direct path phase differences. Additionally, the NMF model enforces the spatial post filter being structured by recurrent spectral events, which is not taken into account in beamforming and TDOA-based spatial filtering methods. This can improve the selectivity of spatial enhancement by Wiener filtering when the derivation of the filter coefficients is not based solely on the observed spatial information (phase difference), but also employs an object-based magnitude model.

Parameter Estimation

Finding optimal parameters for the model defined in Equation (5.13) with the cost function being the squared Frobenius norm in Equation (3.8) can be achieved using the EM algorithm [24] and interpretation of its latent components in the case of NMF with spatial covariance matrices as proposed in [120]. The update equations and their derivation for both introduced complex-valued NMF models (Equations (3.7) and (3.9)) are given in [P6] and [P7]. In this section only the update of the SCM parameters, the spatial weights z_{qo} and the DOA kernels $\mathbf{W}_{io}$ with source-wise SCM in Equation (3.9) from Publication [P7] is presented. The optimization of the magnitude model parameters c_{qk}, b_{ik} and g_{kl} is similar to the multiplicative updates proposed in [120] and can be also found from the above mentioned publications [P6] and [P7].

The update of the DOA kernels $\mathbf{W}_{io}$ requires a special process due to the fact that the argument of each complex-valued entry of $\mathbf{W}_{io}$ needs to be kept fixed. This is done to maintain the original phase difference, i.e., the original delay caused by a certain look direction, while the magnitudes are subject to updating. The magnitudes in the diagonal of $\mathbf{W}_{io}$ combined with the non-negative weights z_{qo} determine the magnitude difference between the channels. The off-diagonal magnitudes determine the cross-channel magnitude correlation.

The following update scheme for the magnitudes of $\mathbf{W}_{io}$ was proposed in [P6]. First a preliminary update for both the magnitudes and the phases of the DOA kernels is derived based on the framework proposed in [120] resulting to

$$\hat{\mathbf{W}}_{io} \leftarrow \mathbf{W}_{io}\Big[\sum_{l,k,o} z_{qo}c_{qk}b_{ik}g_{kl}v_{il} + \sum_{l,k,o} z_{qo}c_{qk}b_{ik}g_{kl}\mathbf{E}_{il}\Big]. \tag{5.14}$$

To prevent a subtractive model and negative values in diagonal, each matrix $\hat{\mathbf{W}}_{io}$ is forced to be positive semidefinite. This can be done as proposed in [120] by calculating its eigenvalue decomposition, setting all negative eigenvalues to zero and reconstructing each matrix with the modified eigenvalues. The final update for $\mathbf{W}_{io}$ is achieved by combining the magnitudes of the modified eigenvalue decomposition and the unmodified phase difference. The phase difference of the original look direction is maintained and the magnitudes are updated. Due to the updated DOA kernels forced being semidefinite and retaining the original phases while modifying the magnitudes, it is no longer guaranteed that the proposed update scheme is optimal with respect to the actual cost function in Equation (3.8). However, experiments with the algorithm have shown no such occasion where the cost function has not decreased after the proposed update of the magnitudes.

The optimization of the spatial weights is achieved with the update

$$z_{qo} \leftarrow z_{qo}\Big[1 + \frac{\sum_{i,l,k} c_{qk} b_{ik} g_{kl} \mathrm{tr}(\mathbf{E}_{il}\mathbf{W}_{io})}{\sum_{i,l,k} c_{qk} b_{ik} g_{kl} v_{il}}\Big]. \tag{5.15}$$

The spatial weights and their above defined update behave similarly to the magnitude model parameters, i.e., it is non-negative and a multiplicative update rule is obtained based on partial derivation of the cost function.

Comparing the conventional source-wise SCM update given in Equation (5.10) and updates in Equations (5.14) - (5.15) of the proposed SCM model $\mathbf{H}_{iq} = \sum_{o=1}^{O} \mathbf{W}_{io} z_{qo}$, the following differences can be observed. The update of the DOA kernels in Equation (5.14) is still frequency-wise, but it only updates the spatial magnitude cues associated in each direction o. It can be interpreted as learning the magnitude properties of each look direction. These properties can be, for example, the frequency-dependent attenuation caused by the acoustical shade of the array and its surrounding environment. The actual SCM of the sources is estimated via update of parameter z_{qo} in Equation (5.15) which is truly independent of frequency.

The actual algorithm is a straightforward extension of the generic NMF flow chart in Section 3.3.4. Each parameter is updated in turns and scaling is applied between updates. The updates are iterated for a fixed amount of iterations.

5.5 Direction of Arrival Estimation Performance

The algorithm proposed in Publication [P6] requires a separate clustering algorithm to associate the NMF component with entire sound sources. The

assumption on NMF components originating from the same or neighboring look direction models the spectral content of the same spatially discriminated sound source. In Figure 5 of Publication [P6] the estimated directions of several NMF components are illustrated. From the figure it is evident that the sources at the left, right and bottom with exactly 90-degree spacing are present and the NMF component belonging to each source is visually easy to determine.

For the automatic assigning of a component to source, the following k-means clustering based on the direction weights z_{ko} was proposed. Values $z_{ko}, o = 1 \ldots O$ were interpreted as a feature vector for NMF component k and k-means clustering was applied to the features. The number of clusters equals to the number of acoustic sources to be separated, which is assumed to be known. As a result of the clustering, the c_{qk} with binary decisions were generated, i.e., one NMF component belongs to one source and not to others, whereas soft decisions are allowed in [P7].

The simple clustering approach was chosen to demonstrate the DOA estimation accuracy of the proposed SCM model. The performance of the chosen clustering was evaluated using known locations of the sources and retrieving the oracle clustering based on it. The decrease in separation performance compared to the oracle clustering in terms of SDR was measured. A small SDR decrease indicated the effectiveness of the k-means clustering and the performance of the proposed method in estimating the spatial direction of each NMF component.

The source-wise SCM based on the DOA kernels in [P7] allowed initialization of the spatial search space of each source to prominent locations. In Section 5.2.3 the SRP-PHAT for source DOA estimation was introduced, and the result of it can be used to initialize the direction weights z_{qo} prior to the estimation of the other model parameters. The exact procedure for the preliminary DOA estimation is given in [P7]. The estimated DOA for each source q was used to set the weights of the look direction indices o in z_{qo} within ± 25 degrees from it to one and all the other direction weights of the source were set to zero. The window of 50 degrees was chosen to account any errors in the preliminary DOA estimation. This was found to be an important procedure in order to avoid the direction weights of each source pointing in direction of the most energetic source. The cost function of the Frobenius norm of the overall modeling error does not distinguish the underlying sources, i.e., the modeling of less energetic sources may become undermined and only details of the most energetic source will be accounted for by the model.

Both of the above concepts, estimating either the DOA of an NMF component or an entire source, allows positioning of the source in playback by

binaural synthesis or with a 3D loudspeaker array. The binaural reconstruction can be achieved by retrieving the DOA of the source by finding the mean of the direction weights z_{qo} and mapping its index o back to the look direction azimuth φ_o and elevation θ_o pair. The associated HRTF from the nearest direction is used for convolving the estimated source signal. When a loudspeaker array is used for reconstruction and playback of the 3D sound, the source positioning, for example, with vector base amplitude panning [109] can be used.

5.6 Source Separation Results

The separation quality evaluation from Publications [P6] and [P7] is summarized next. The separation quality was evaluated using the energy-based objective metrics introduced in Section 2.3.5 with the addition of perceptually motivated scores proposed in [30] were reported in the Publication [P6].

The test material was generated by convolving anechoic material (male and female speech, pop music and various everyday noise sources) with RIRs from different angles captured using an array consisting of four omnidirectional microphones enclosed in a metal casing of the size 30 mm x 60 mm x 1150 mm. The geometry of the array used in recording is illustrated in Figure 9 in [P6] and the locations of the microphones are given in Table 1 in [P6]. The spatial aliasing frequency of the array used is 1563 Hz. The room used for recording the RIRs is a normal meeting room with moderate reverberation time of $T_{60} = 350$ ms. The room layout and directions of the sources are shown in Figure 5.4. The description of the exact capturing conditions, anechoic material used and the overall test set description with used angle combinations for simultaneous sound sources can be found from [P6]. In short, the evaluation material included two datasets consisting of two and three simultaneous sources originating from different angles. The total number of test signals was 48 and 42, respectively, and each test signal was 10 seconds in duration.

The results of the separation quality evaluation from the Publications [P4], [P6] and [P7] are considered all at once.The algorithms considered are

Alg. 1: Complex-valued NMF with component-wise SCM based on DOA kernels [P6].

Alg. 2: Complex-valued NMF with source-wise SCM based on DOA kernels [P7].

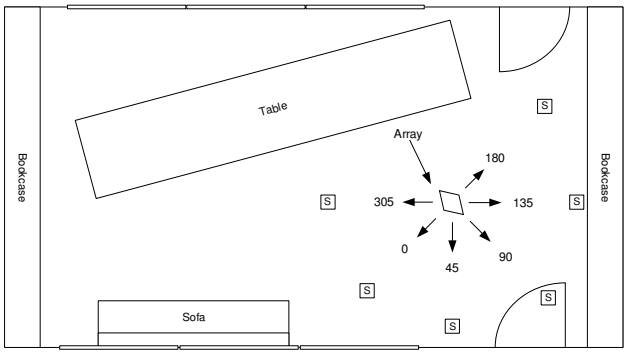


Figure 5.4: Room layout used for capturing RIRs for creating the source separation evaluation material.

Alg. 3: Complex-valued NMF with unconstrained source-wise SCM estimation [120].

Alg. 4: Frequency domain ICA with TDOA and intra-source envelope-based permutation alignment [P4].

Alg. 5: Frequency domain ICA with TDOA-based permutation alignment [118].

The combined results can be summarized as follows. Both proposed methods Alg. 1 and Alg. 2 with the DOA-based SCM model exceed the separation performance of the unconstrained SCM estimation in Alg. 3 when measured using the SDR criterion. In general, the NMF-based separation Alg. 1-3 exceeds the SDR performance of the frequency domain ICA approaches Alg. 4 and Alg. 5. The evaluation in [P6] and [P7] replicates the result from [P4] where the proposed source envelope-based permutation alignment in Alg. 4 improves separation over permutation alignment-based only on the TDOA estimation in Alg. 5.

The perceptually motivated scores reported in Publication [P6] indicate that the methods based on ICA produce an almost perceptually equivalent separation quality when compared to NMF-based separation and even exceeds the perceptual performance with three simultaneous sources. It is worth noting that none of the tested algorithms managed to separate the three simultaneous sources with moderate quality, with a few exceptions of selected angle and source type combinations. Methods based on ICA do not produce artefacts to the reconstructed signal even if no accurate estimation of the unmixing matrix is obtained, whereas in similar situations the NMF-based methods may produce unwanted rapid changes in the spectrum resulting in separation artefacts. Particularly, the binary clustering used to

link NMF component to sources in [P6] is affected from the said problem and the lower SAR score can be partially attributed to the errors in the clustering.

Regarding the most general and best performing method proposed in this thesis, the source-wise DOA-based SCM with NMF magnitude model from [P7], achieves an SDR increase of 1.9 dB and 1.0 dB over the baseline from [120] with two and three simultaneous sources, respectively.

In Publication [P6] the effect of the increasing angle between sources with a fully simulated scenario was investigated. The angle between two simultaneous sources starting from 15 degrees, with 15 degree increments up to 90 degrees, were evaluated in terms of SDR between the method proposed in [P6] and the baseline [120]. With 15- and 30-degree spacing the proposed method already has a minor advantage, and starting with the 45-degree spacing the proposed method gives a significant increase in SDR, up to 3 dB in the case of 90-degree spacing between sources.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

THIS thesis investigated object-based models for representing audio signals and presented applications of them for audio coding and source separation. The suitability of the object-based audio models for the covered applications was analyzed and the benefits found either by better performance or added features, were analyzed and critically assessed.

The focus of the thesis in obtaining an object-based approximation of a time-frequency representation of an audio signal was upon machine learning techniques. The methods included independent component analysis (ICA) [61] along with non-negative matrix factorization (NMF) [80] and its multiway extensions [37, 120]. The method of decomposing an audio spectrogram to a finite set of spectral templates and their activations combined in a linear manner is equivalent to reducing the redundancy in representing the audio signal. The machine learning approach for finding a small amount of spectral templates with associated temporal activity that best explains the observed spectrogram leads to structures that model coherent spectral content and have meaningful interpretation through audio objects. In the thesis one NMF component, with the possible addition of spatial parameters, was regarded as a fundamental audio object.

Approximating natural and rich audio signals with such an object-based model was shown to have only a minor loss of significant signal details in Publications [P1] and [P2], given that the optimization criteria used in deriving the model parameters reflect the perceptual relevance of time-frequency points of the audio spectrogram in question. The perceptual relevance was accounted for by the ratio of the modeling error compared to the masking threshold created by our hearing [140] formulated as NMF model cost func-

tion in Publication [P1].

The natural sparseness of the NMF decomposition, in terms of a few simultaneously active audio objects, was proposed to be harnessed for audio signal compression in Publication [P2]. The NMF model consisting of only several tens of spectral templates and their associated time-dependent gains for representing the magnitude spectrogram of audio was found to have favorable data and redundancy reduction properties. A single channel object-based audio coding algorithm was realized by developing a quantization and entropy coding scheme for the NMF model parameters. The difficulties in representing the phase of each time-frequency point caused inefficiency in audio coding performance.

In Publications [P3] and [P5] a spatial audio coding (SAC) method based on non-negative tensor factorization (NTF) was proposed. Recovery of multiple channels was based on filtering a downmixed signal using an object-based model encoded and transmitted as a auxiliary information. The object-based model was obtained by applying the NTF to the multichannel magnitude spectrogram, producing a channel-wise level difference for each audio object to denote its spatial position. The conventional SAC methods [33, 125] estimate spatial cues, such as the time and level differences between channels, for fixed time-frequency blocks, whereas the proposed method operates with an object-based model requiring spatial position estimated for a fewer number of elements. The quantization and encoding of the NTF model was based on findings from [P2]. An evaluation done using listening tests indicated that the proposed algorithm was comparable to the conventional upmixing-based SAC method with block-wise spatial parameter estimation in terms of perceptual quality at similar bitrates.

The benefits of the NTF-based model for SAC include its ability to learn and represent audio objects with overlapping frequency content. Additionally, the learned audio objects are suitable for blind source separation, given that a suitable clustering of components to entire acoustical sources can be obtained. Publication [P5] investigated the aspect of user-created clustering of the components to the original sources present in the mixture. The evaluation of the separation quality by objective criteria [146] resulted in a similar separation performance as ideal binary mask separation with selected sources, including vocals. Blind separation allows, for example, removal of vocals or other instruments without having the source tracks separately at the encoding stage. The separation does not increase the bitrate of the encoding due to the fact that the manipulation of the upmixed content is based on the gain of the objects directly in the coding domain.

The source separation part of the thesis in Chapter 5 concentrated on the use of same object-based model obtained by NMF with the addition of spatial

extension that can model both level and time differences between multiple channels of a microphone array recording. The strengths of the object-based model for sound source separation are similar as in audio coding, reducing the amount of elements for which the spatial parameters are estimated and the components being defined over frequency alleviating the spatial aliasing issues. In [P6] a direction of arrival-based spatial covariance matrix model was proposed for use in representing the spatial properties of NMF components as a function of their direction. The proposed model unifies the estimation of spatial parameter over frequency, thus ensuring that the audio objects estimated are spatially coherent, i.e., the object spectrogram modeled by one NMF component is originating from the same spatial location over the whole duration of the signal under analysis. In [P7] the direction of arrival-based covariance matrix estimation was formulated for entire sources instead of NMF components, removing the need for clustering the NMF components to sources based on their estimated direction of arrival. The independent component analysis (ICA) based separation with the permutation alignment proposed in Publication [P4] was in the role of baseline for evaluation of the object-based separation framework in [P6], [P7]. In both publications, the proposed alternatives were able to improve the objective measures of separation quality over the ICA baseline and earlier covariance matrix estimation methods [27, 120].

This thesis has introduced different methods for obtaining an object-based model of audio and has demonstrated the benefits of such models in application fields of audio coding and source separation. The use of the decomposition models and machine learning-based audio models is certainly not limited to the applications covered in this thesis, and the constant ongoing research on improving the object analysis performance and utilization in new audio signal processing problems and applications is evident.

6.2 Future Work

The concepts of NMF-based spatial audio coding and the direction of arrival-based covariance matrix estimation for NMF-based source separation introduced in this thesis are relatively novel and undiscovered fields. They allow several new aspects for future work in improving and extending the methods.

The perceptually motivated NMF criterion considered in this thesis for audio coding have been additionally considered for sound source separation [69]. However, the spatial extensions of NMF with covariance matrix estimation have not been yet developed to account for such perceptual relevance. The results in [121] with Itakura-Saito divergence suggest that the squared Eu-

clidean distance is not optimal in terms of source separation, and it is yet to be discovered whether improvements with perceptual criteria can be achieved.

A possible aspect for future work on improving the concept of object-based SAC by NTF would be to include estimation of inter-channel time delay for each audio object. The complex-valued NMF models with covariance matrix estimation allow for modeling the time delay between the channels but with respect to the coding and compression efficiency the number of parameters used for representing each covariance matrix is high. A modification of the direction of arrival-based SCM model in [P6] to a general kernel-based covariance estimation without assumption on array capture could be used to reduce the number of SCM parameters to be encoded. However, the arbitrary mixing in professionally produced material does not obey the assumptions made on the time-delay behavior of a known array geometry and thus would require a generalization to any arbitrary time-delay in each channel pair. A related field of NMF-based informed source separation [103] can be considered as a future work for improving and extending the object-based coding of multichannel audio, and the techniques developed can be utilized in the coding of audio containing a mixture of sources as in the proposed SAC framework.

The general direction-based estimation of the source spatial covariance properties proposed in Publications [P6] and [P7] pave the way for numerous minor modifications and alterations, possibly improving the separation quality. Several topics for future work include the following: incorporating the perceptually motivated criteria, further studies on 3D sound reconstruction based on the estimated source directions, investigation of rank-1 versus full-rank SCM estimates while considering the fact that the input of algorithm is strictly rank-1. Additionally, investigating the possibility of re-sampling and allowing dynamic look directions for more accurate estimation of the source direction and possibility to reduce the amount of DOA kernels used.

Bibliography

[1] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[2] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proceedigs of International Conference on Independent Component Analysis and Blind Signal Separation*, 2000, pp. 215–220.

[3] S. Arberet, A. Ozerov, N. Q. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Proceedings of International Conference on Information Sciences Signal Processing and their Applications (ISSPA)*, 2010, pp. 1–4.

[4] F. Baumgarte and C. Faller, "Binaural Cue Coding-Part I: Psychoacoustic Fundamentals and Design Principles," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 509–519, 2003.

[5] M. Bear, B. Connors, and M. Paradiso, *Neuroscience: Exploring the brain*. Lippincott Williams & Wilkins, 2006.

[6] E. Benetos, A. Klapuri, and S. Dixon, "Score-informed transcription for automatic piano tutoring," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, 2012, pp. 2153–2157.

[7] E. Benetos, M. Kotti, and C. Kotropoulos, "Musical instrument classification using non-negative matrix factorization algorithms," in *Proceedings of IEEE International Symposium on Circuits and Systems*, 2006, pp. 1844–1847.

[8] L. R. Bernstein and C. Trahiotis, "Detection of interaural delay in high-frequency sinusoidally amplitude-modulated tones, two-tone complexes, and bands of noise," *The Journal of the Acoustical Society of America*, vol. 95, no. 6, pp. 3561–3567, 1994.

[9] V. Blondel, N.-D. Ho, and P. van Dooren, "Weighted nonnegative matrix factorization and face feature extraction," *Image and Vision Computing*, 2007.

[10] K. Brandenburg, "MP3 and AAC explained," in *Proceedings of the 17th Audio Engineering Society International Conference on High-Quality Audio Coding*, Florence, Italy, 1999.

[11] K. Brandenburg and T. Sporer, "NMR and masking flag: Evaluation of quality using perceptual criteria," in *Proceedings of the 19th Audio Engineering Society Conference: Test and Measurement*, Portland, USA, 1992, pp. 169–179.

[12] K. Brandenburg, "Evaluation of quality for audio encoding at low bit rates," in *Proceedings of the 82nd Audio Engineering Society Convention*, London, UK, 1987.

[13] J. Breebaart, J. Engdegård, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, J. Koppens, W. Oomen, B. Resch, E. Schuijers, and L. Terentiev, "Spatial audio object coding (SAOC) - the upcoming MPEG standard on parametric object based audio coding," in *Proceedings of the 124th Audio Engineering Society Convention*, Amsterdam, The Netherlands.

[14] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica United with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[15] J. C. Brown and M. S. Puckette, "An efficient algorithm for the calculation of a constant Q transform," *The Journal of the Acoustical Society of America*, vol. 92, no. 5, pp. 2698–2701, 1992.

[16] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A versatile framework for multichannel blind signal processing," in *Proceedings of the 29th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, Quebec, Canada, 2004, pp. 889–892.

[17] G. C. Carter, "Time delay estimation for passive sonar signal processing," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 3, pp. 463–470, 1981.

[18] J. C. Chen, R. E. Hudson, and K. Yao, "Maximum-likelihood source localization and unknown sensor location estimation for wideband sig-

nals in the near-field," *IEEE Transactions on Signal Processing*, vol. 50, no. 8, pp. 1843–1854, 2002.

[19] A. Cichocki, R. Zdunek, A. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. New York, NY, USA: Wiley, 2009.

[20] A. Cichocki, R. Zdunek, and S. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *Proceedings of the 31th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, 2006, pp. 621–624.

[21] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[22] J. Daniel, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format," in *Proceedings of the 23rd International Audio Engineering Society Conference: Signal Processing in Audio Recording and Reproduction*, 2003.

[23] J. Daniel, S. Moreau, and R. Nicol, "Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging," in *Proceedings of the 114th Audio Engineering Society Convention*, Amsterdam, The Netherlands.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[25] P. F. Driessen and Y. Li, "An unsupervised adaptive filtering approach of 2-to-5 channel upmix," in *Proceedings of the 119th Audio Engineering Society Convention*, New York, NY, USA, 2005.

[26] R. Drullman and A. W. Bronkhorst, "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation," *The Journal of the Acoustical Society of America*, vol. 107, no. 4, pp. 2224–2235, 2000.

[27] N. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[28] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing.* Berlin, Germany: Springer, 2010.

[29] D. Ellis and D. Rosenthal, "Mid-level representations for computational auditory scene analysis," in *International Joint Conference on Artificial Intelligence - Workshop on Computational Auditory Scene Analysis*, 1995.

[30] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[31] F. A. Everest and K. C. Pohlmann, *The Master Handbook of Acoustics.* New York, NY, USA: McGraw-Hill, 2001.

[32] C. Falch, L. Terentiev, and J. Herre, "Spatial audio object coding with enhanced audio object separation," in *Proceedings of 13th International Conference on Digital Audio Effects (DAFx)*, Graz, Austria, 2010.

[33] C. Faller, "Parametric coding of spatial audio," in *Proceedigs of 7th International Conference on Audio Effects (DAFx)*, Naples, Italy, 2004.

[34] C. Faller and F. Baumgarte, "Binaural Cue Coding-Part II: Schemes and Applications," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 520–531, 2003.

[35] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proceedings of the 108th Audio Engineering Society Convention*, Paris, France, 2000.

[36] C. Févotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation: statistical insights and towards self-clustering of the spatial cues," in *Exploring Music Contents*, 2011, pp. 102–115.

[37] D. FitzGerald, M. Cranitch, and E. Coyle, "Sound source separation using shifted non-negative tensor factorisation," in *Proceedings of the 31th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, 2006.

[38] D. FitzGerald, "Upmixing from mono-a source separation approach," in *International Conference on Digital Signal Processing*, Corfu, Greece, 2011.

[39] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative tensor factorisation for sound source separation," in *Proceedings of the Irish Signals and Systems Conference*, Dublin, Ireland, 2005.

[40] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience*, 2008.

[41] H. Fletcher and W. A. Munson, "Loudness, its definition, measurement and calculation," *Journal of the Acoustical Society of America*, vol. 5, pp. 82–108, 1933.

[42] O. L. Frost III, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[43] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[44] J. T. Geiger, F. Weninger, A. Hurmalainen, J. F. Gemmeke, M. Wöllmer, B. Schuller, G. Rigoll, and T. Virtanen, "The TUM + TUT + KUL approach to the 2nd CHiME challenge: Multi-stream ASR exploiting BLSTM networks and sparse NMF," *Proceedings of the 2nd International Workshop on Machine Listening in Multisource Environments (CHiME)*.

[45] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[46] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1, pp. 103–138, 1990.

[47] S. Gorlow and S. Marchand, "Informed audio source separation using linearly constrained spatial filters," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 3–13, 2013.

[48] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[49] D. Guillamet, J. Vitria, and B. Schiele, "Introducing a weighted non-negative matrix factorization for image classification," *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2447–2454, 2003.

[50] S. Harding, J. Barker, and G. J. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 58–67, 2006.

[51] M. L. Hawley, R. Y. Litovsky, and J. F. Culling, "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," *The Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 833–843, 2004.

[52] M. Helén and T. Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," in *Proceed of the 13th European Signal Processing Conference (EUSIPCO)*, 2005.

[53] J. Herre, C. Falch, D. Mahne, G. Del Galdo, M. Kallinger, and O. Thiergart, "Interactive teleconferencing combining spatial audio object coding and DirAC technology," *Journal of the Audio Engineering Society*, vol. 59, no. 12, pp. 924–935, 2010.

[54] J. Herre, C. Faller, C. Ertel, J. Hilpert, A. Hoelzer, and C. Spenger, "MP3 surround: Efficient and compatible coding of multi-channel audio," in *Proceedings of the 116th Audio Engineering Society Convention*, Berlin, Germany, 2004.

[55] J. Herre, K. Kjürling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Röden, W. Oomen, K. Linzmeier, and C. K. C., "MPEG surround: The ISO/MPEG standard for efficient and compatible multichannel audio coding," *Journal of the Audio Engineering Society*, vol. 56, no. 11, pp. 932–955, 2008.

[56] J. Herre and S. Disch, "New concepts in parametric coding of spatial audio: From SAC to SAOC," in *Proceedigs of IEEE International Conference on Multimedia and Expo*, 2007, pp. 1894–1897.

[57] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, pp. 2677–2684, 1999.

[58] P. O. Hoyer, "Non-negative sparse coding," in *Proceedings of 12th IEEE Workshop on Neural Networks for Signal Processing*, 2002, pp. 557–565.

[59] D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proceedings of the Institute of Radio Engineers (IRE)*, vol. 40, no. 9, pp. 1098–1101, 1952.

[60] A. Hurmalainen, J. Gemmeke, and T. Virtanen, "Non-negative matrix deconvolution in noise robust speech recognition," in *Proceedings of the 36th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 4588–4591.

[61] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley-interscience, 2001.

[62] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[63] R. Irwan and R. M. Aarts, "Two-to-five channel sound processing," *Journal of the Audio Engineering Society*, vol. 50, no. 11, pp. 914–926, 2002.

[64] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Proceedings of 6th International Congress on Acoustics*, 1968.

[65] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *Proceedings of the 25th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, 2000, pp. 2985–2988.

[66] P. Kabal, "An examination and interpretation of ITU-R BS. 1387: Perceptual evaluation of audio quality," *TSP Lab Technical Report, Department of Electrical & Computer Engineering, McGill University*, pp. 1–89, 2002.

[67] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proceedings of IEEE the 34th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 3437–3440.

[68] M. Karjalainen, "A new auditory model for the evaluation of sound quality of audio systems," in *Proceedings of the 10th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Tampa, FL, USA, 1985, pp. 608–611.

[69] S. Kırbız and B. Günsel, "Perceptually enhanced blind single-channel music source separation by non-negative matrix factorization," *Digital Signal Processing*, vol. 23, no. 2, pp. 646–658, 2013.

[70] S. Kırbız, A. Ozerov, A. Liutkus, and L. Girin, "Perceptual coding-based informed source separation," in *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, Portugal, 2014.

[71] W. B. Kleijn and A. Ozerov, "Rate distribution between model and signal," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2007, pp. 243–246.

[72] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[73] R. Kompass, "A generalized divergence measure for nonnegative matrix factorization," *Neural Computation*, vol. 19, no. 3, pp. 780–791, 2007.

[74] A. Koski, "Rich recording technology technical overall description," in *(http://i.nokia.com/blob/view/-/1696152/data/2/-/Download2.pdf)*.

[75] K. Kumatani, T. Arakawa, K. Yamamoto, J. McDonough, B. Raj, R. Singh, and I. Tashev, "Microphone array processing for distant speech recognition: Towards real-world deployment," *Proceedings of the Asia Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference*, 2012.

[76] K. Kumatani, L. Lu, J. McDonough, A. Ghoshal, and D. Klakow, "Maximum negentropy beamforming with superdirectivity," in *Proceedings of the 18th European Signal Processing Conference (EUSIPCO)*, Aalborg, Denmark, 2010, pp. 2067–2071.

[77] K. Kumatani, J. McDonough, D. Klakow, P. N. Garner, and W. Li, "Adaptive beamforming with a maximum negentropy criterion," in *Proceedings of Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, Trento, Italy, 2008, pp. 180–183.

[78] K. Kumatani, J. McDonough, B. Rauch, P. N. Garner, W. Li, and J. Dines, "Maximum kurtosis beamforming with the generalized sidelobe canceller," in *Proceedings of the 9th Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2008, pp. 423–426.

[79] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[80] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, pp. 556–562, 2002.

[81] I. Lee, T. Kim, and T. Lee, "Independent vector analysis for convolutive blind speech separation," *Blind speech separation*, pp. 169–192, 2007.

[82] A. Liutkus, R. Badeau, and G. Richard, "Informed source separation using latent components," in *Proceedings of 9th International Conference on Latent Variable Analysis and Signal Separation*, 2010, pp. 498–505.

[83] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, vol. 92, no. 8, pp. 1937–1949, 2012.

[84] R. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Proceedings of the 7th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Paris, France, 1982, pp. 1282–1285.

[85] ITU-R BS.1534-1, "Method for the subjective assessment of intermediate quality level of coding systems," *International Telecommunications Union Radiocommunication Assembly*, 2003.

[86] ITU-R BS.775, " Multichannel Stereophonic Sound System With and Without Accompanying Picture," *International Telecommunications Union Radiocommunication Assembly*, 1994.

[87] J. McDonough and K. Kumatani, "Microphone arrays," in *Techniques for Noise Robustness in Automatic Speech Recognition*, T. Virtanen, R. Singh, and B. Raj, Eds. John Wiley & Sons, 2012.

[88] J. Merimaa and V. Pulkki, "Spatial impulse response rendering I: Analysis and synthesis," *Journal of the Audio Engineering Society*, vol. 53, no. 12, pp. 1115–1127, 2012.

[89] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proceedings of the 27th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, FL, USA, 2002, pp. 1781–1784.

[90] J. Meyer and G. Elko, "Spherical microphone arrays for 3d sound recording," in *Audio Signal Processing: For Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds. Springer, 2004.

[91] Y. Mitsufuji and A. Roebel, "On the use of a spatial cue as prior information for stereo sound source separation based on spatially weighted non-negative tensor factorization," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, pp. 1–9, 2014.

[92] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *The Journal of the Acoustical Society of America*, vol. 74, pp. 750–753, 1983.

[93] B. C. J. Moore, *Hearing*. London, UK: Elsevier Academic Press, 1995.

[94] B. C. J. Moore, *An Introduction to Psychology of Hearing*. London, UK: Elsevier Academic Press, 1997.

[95] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proceedings of the 38th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 7092–7096.

[96] F. Nesta, M. Omologo, and P. Svaizer, "Multiple TDOA estimation by using a state coherence transform for solving the permutation problem in frequency-domain BSS," in *Proceedings of IEEE Workshop on Machine Learning for Signal Processing*, 2008, pp. 43–48.

[97] J. Nikunen, "Perceptual audio quality criterion in non-negative matrix factorization," Master's thesis, Tampere University of Technology, Department of Information Technology, 2010.

[98] P. D. O'Grady, "Sparse separation of under-determined speech mixtures," Ph.D. dissertation, 2007.

[99] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2011, pp. 189–192.

[100] N. Ono, "Fast stereo independent vector analysis and its implementation on mobile phone," in *Proceedings of International Workshop on Acoustic Signal Enhancement*, 2012, pp. 1–4.

[101] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[102] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *Proceedings of the 36th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 257–260.

[103] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Coding-based informed source separation: Nonnegative tensor factorization approach," *IEEE Transactions on Audio, Speech and Language Processing*, 2012.

[104] R. M. Parry and I. A. Essa, "Estimating the spatial position of spectral components in audio," in *Proceedings of International Conference on*

Independent Component Analysis and Blind Signal Separation, 2006, pp. 666–673.

[105] R. M. Parry and I. Essa, "Incorporating phase information for source separation via spectrogram factorization," in *Proceedings of the 32nd International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, HI, USA, 2007, pp. 661–664.

[106] M. Parvaix and L. Girin, "Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1721–1733, 2011.

[107] M. Parvaix, L. Girin, and J. Brossier, "A watermarking-based method for informed source separation of audio signals with a single sensor," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1464–1475, 2010.

[108] P. Pertilä, "Online blind speech separation using multiple acoustic speaker tracking and time-frequency masking," *Computer Speech & Language*, vol. 27, no. 3, pp. 683–702, 2012.

[109] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997.

[110] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, 2007.

[111] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition." in *Proceedings of the 11th Annual Conference of International Speech Communication Association (INTERSPEECH)*, Makuhari, Japan, 2010, pp. 717–720.

[112] M. Ramteke, "Method and apparatus for stereo to five channel upmix," 2012, US Patent Application 13/579,561. [Online]. Available: http://www.google.com/patents/US20120308015

[113] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *Proceedings of the 27th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, FL, USA, 2002, pp. 529–532.

[114] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality

assessment of telephone networks and codecs," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, UT, USA, 2001, pp. 749–752.

[115] D. Salomon, *Data Compression: the Complete Reference*. Springer, 2004.

[116] L. K. Saul, F. Sha, and D. D. Lee, "Statistical signal processing with nonnegativity constraints," in *Proceedings of the Eighth European Conference on Speech Communication and Technology*, 2003, pp. 1001–1004.

[117] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[118] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592–1604, 2007.

[119] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Formulations and algorithms for multichannel complex NMF," in *Proceedings of the 36th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 229–232.

[120] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "New formulations and efficient algorithms for multichannel NMF," in *Proceedigns of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2011, pp. 153–156.

[121] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[122] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, 2004.

[123] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *7th Sound and Music Computing Conference*, Barcelona, Spain, 2010.

[124] M. R. Schroeder, "Integrated-impulse method measuring sound decay without using impulses," *The Journal of the Acoustical Society of America*, vol. 66, no. 2, pp. 497–500, 1979.

[125] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegård, "Low complexity parametric stereo coding," in *Proceedings of the 116th Audio Engineering Society Convention*, Berlin, Germany, 2004.

[126] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*. Springer, 2001, pp. 39–60.

[127] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1, pp. 21–34, 1998.

[128] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[129] P. Smaragdis, "Extraction of speech from mixture signals," in *Techniques for Noise Robustness in Automatic Speech Recognition*, T. Virtanen, R. Singh, and B. Raj, Eds. John Wiley & Sons, 2012.

[130] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003, pp. 177–180.

[131] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Independent Component Analysis and Blind Signal Separation*. Springer, 2004, pp. 494–499.

[132] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Independent Component Analysis and Signal Separation*, 2007, pp. 414–421.

[133] S. Smyth, W. Smith, M. Smyth, M. Yan, and T. Jung, "DTS coherent acoustics. delivering high quality multichannel sound to the consumer," in *Proceedings of the 100th Audio Engineering Society Convention*, Copenhagen, Denmark, 1996.

[134] A. Spanias, T. Painter, and V. Atti, *Audio Signal Processing and Coding*. John Wiley & Sons, 2007.

[135] G.-B. Stan, J.-J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 249–262, 2002.

[136] B. D. Steinberg, *Principles of Aperture and Array System Design: Including Random and Adaptive Arrays*, 1976.

[137] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[138] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[139] I. Tashev, *Sound Capture and Processing: Practical Approaches*. John Wiley & Sons Inc, 2009.

[140] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Kheyl, G. Stoll, K. Brandenburg, and B. Feiten, "PEAQ - the ITU standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, pp. 3–29, 2000.

[141] C. Todd, G. Davidson, M. Davis, L. Fielder, B. Link, and S. Vernon, "AC-3: Flexible perceptual coding for audio transmission and storage," in *Proceedings of the 96th Audio Engineering Society Convention*, Amsterdam, The Netherlands, 1994.

[142] C. Uhle, J. Herre, A. Walther, O. Hellmuth, and C. Janssen, "Apparatus and method for generating an ambient signal from an audio signal, apparatus and method for deriving a multi-channel audio signal from an audio signal and computer program," 2013, US Patent 8,346,565. [Online]. Available: http://www.google.com/patents/US8346565

[143] C. Uhle, A. Walther, O. Hellmuth, and J. Herre, "Ambience separation from mono recordings using non-negative matrix factorization," in *Proceedings of the 30th Audio Engineering Society Conference: Intelligent Audio Environments*, 2007.

[144] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[145] E. Vincent and M. D. Plumbley, "Low bit-rate object coding of musical audio using bayesian harmonic models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1273–1282, 2007.

[146] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and

results," *Independent Component Analysis and Signal Separation*, pp. 552–559, 2007.

[147] T. Virtanen, "Monoaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[148] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in *Proceedings of International Computer Music Conference (ICMC)*, 2003, pp. 231–234.

[149] T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*. John Wiley & Sons, 2012.

[150] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[151] Y. Zheng, K. Reindl, and W. Kellermann, "BSS for improved interference estimation for blind speech signal extraction with two microphones," in *3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, 2009, pp. 253–256.

[152] E. Zwicker and H. Fastl, *Psychoacoustics Facts and Models*. Springer-Verlag, 1990.

Publication P1

J. Nikunen and T. Virtanen, "Noise-to-Mask Ratio Minimization by Weighted Non-negative Matrix factorization," in *Proceedings of 35th International Conference on Acoustics, Speech, and Signal Processing*, Dallas, Texas, USA, 2010, pp. 25–28.

NOISE-TO-MASK RATIO MINIMIZATION BY WEIGHTED NON-NEGATIVE MATRIX FACTORIZATION

Joonas Nikunen, Tuomas Virtanen

Department of Signal Processing, Tampere University of Technology
Korkeakoulunkatu 1, FI-33720 Tampere, Finland

ABSTRACT

This paper proposes a novel algorithm for minimizing the perceptual distortion in non-negative matrix factorization (NMF) based audio representation. We formulate the noise-to-mask ratio audio quality criterion in a form where it can be used in NMF and propose an algorithm for optimizing the criterion. We also propose a method for compensating the spreading of the representation error in the synthesis filterbank. The objective perceptual quality produced by the proposed method is found to outperform all the reference methods. We also study the trade-off between the window length and the rank of factorization with a fixed data rate, and find that the best performance is obtained with window lengths between 10 and 30 ms.

Index Terms— Non-negative matrix factorization, Noise-to-mask ratio, Audio coding, Signal representations

1. INTRODUCTION

In audio signal processing, an acoustic time-domain signal is often represented using a mid-level representation [1], which allows more efficient analysis or manipulation of the signal. Commonly used mid-level representations include, for example, the time-frequency representations such as the short-time Fourier transform (STFT), and parametric representations such as the sinusoidal model. More advanced models can take into account the structure of the sounds in more detail, for example by using a harmonic model [2]. The latter two can be also viewed as lossy compression, since they reduce the amount of information needed to approximate the original signal. The parameters of a representation can be estimated by using a statistical criterion such as the mean-square error, but also the properties of the human audio perception can be taken into account.

All present-day perceptual audio coders are essentially based on a sub-band bit allocation upon a psychoacoustical masking model. They quantize a time-frequency representation of audio signal in such way that the quantization noise stays below the masking threshold and thus remains inaudible [3]. An objective measure of the perceptual quality of a compressed signal is the noise-to-mask ratio (NMR) [4], which measures the relative level of the quantization noise in comparison with the masking threshold. An alternative approach to audio compression is object-based audio coding, where individual sound sources or objects (e.g. musical instruments, speakers, notes) in an audio recording are represented separately [5]. Object-based coding allows using the most efficient codec for each object, but as well interactive synthesis of the signal.

Recently, non-negative matrix factorization (NMF) has been applied in many audio signal processing tasks, such as sound source separation [6]. Its main advantage is the ability to automatically decompose a mixture signal into a representation where each sound source is represented as an individual object [6]. The NMF decomposition also effectively finds repetitive structures in the signal, thus being able to reduce redundancy and being attractive from signal compression point of view.

This paper proposes a novel algorithm for NMF which minimizes the noise-to-mask ratio of the signal decomposition. The NMR objective is formulated as a cost function for NMF and it is minimized using a weighted NMF algorithm. We also propose to filter the estimated masking patterns in time, which effectively reduces the pre-echo caused by the spreading of errors in the synthesis filterbank. Potential applications of the proposed method include object-based audio coding and analysis of audio signals.

The block diagram of the proposed system is shown in Figure 1. First, the magnitude spectrogram of an input signal is calculated for the NMF algorithm. Masking thresholds are estimated from an input signal, which are then used for NMF weighting. Approximation of original spectrogram is obtained from the weighted NMF algorithm and the signal is reconstructed by assigning the original phases to it and taking the inverse FFT. Frames are finally combined in the synthesis filterbank by overlap-add.

The structure of the paper is as follows: Section 2 gives short review of the noise-to-mask ratio which is the objective of the proposed method. In Section 3 we derive a weighted cost function for NMF corresponding to the NMR. Section 4 presents a synthesis procedure and proposes a technique to reduce the pre-echo effect. The proposed method is compared to conventional NMF algorithms in Section 5. Section 5 also presents results from an experiment on finding the best combination of coding parameters in case of constant data rate.

2. NOISE-TO-MASK RATIO

Human hearing includes a masking phenomenon, which causes low-intensity frequency components to become masked by more intense ones, that occur spatially and temporally close to each other. It means that a loud frequency component can make a fainter component become completely inaudible to our hearing [7, p. 56]. The masking concept can be utilized in audio coding, where it is used to decide, which parts of the audio can be disregarded without perceptual difference.

A quality metric to measure the audibility of distortions is the noise-to-mask ratio, which was introduced by Brandenburg [4]. The metric consists of the following processing steps: 1) The error between a distorted signal and a reference signal is calculated. 2) The masking threshold is estimated from the reference. 3) The noise-to-mask ratio in each time frame is calculated in Bark scale. 4) The final measure is average over all the time-frequency points. Distortions having a NMR value of -10 dB or below can be assumed to be

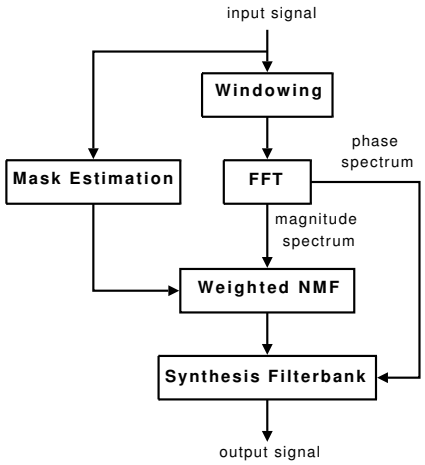


Fig. 1. Block diagram of proposed method

inaudible.

NMR has been included into recommendation BS.1387 [8] for perceptual evaluation of audio quality (PEAQ). The recommendation includes specifications for the auditory model to be used for estimating the masking threshold required for NMR evaluation. PEAQ auditory model (with clarifications from [9]) is used here for masking threshold estimation. The model includes parameter L_p for scaling the mask estimation to correspond to desired listening sound pressure level (SPL). This is due to the fact that spatial and temporal spreading functions are dependent on the energy of the masker component.

The NMR in PEAQ can be described using the equation

$$\mathrm{NMR_B} = 10\log_{10}\Big(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{B}\sum_{b=1}^{B}[\mathbf{M}]_{b,t}[\mathbf{CH}(\mathbf{X}-\hat{\mathbf{X}})^{\cdot 2}]_{b,t}\Big), \quad (1)$$

and it consists of the operations below: 1) Squared difference between the magnitude spectrograms of the original signal $\mathbf{X}$ and the estimated signal $\hat{\mathbf{X}}$ is calculated. $\mathbf{X}^{\cdot 2}$ denotes element-wise power of two. The spectrograms are calculated using a 42.7 ms Hanning window and discrete Fourier transform (DFT). 2) The error is weighted by middle- and outer ear transfer function, which is implemented by multiplying the squared error spectrogram by a diagonal matrix $\mathbf{H}$ having the values of the transfer function on the diagonal. 3) The error is decimated to a bark scale representation, which is implemented by multiplication by matrix $\mathbf{C} \in \mathbb{R}^{\geq 0, B\times K}$, where each row contains the power response of a bark band for all the DFT indices. 4) The error in bark scale is weighted by $\mathbf{M} \in \mathbb{R}^{\geq 0, B\times T}$, which is the element-wise inverse of the masking threshold in each frame t and bark band b. Both the error and masking patterns are having a quarter bark band frequency resolution, which results to 109 bands with 48 kHz sampling frequency. 5) The results are averaged over frequency and time and converted to the dB scale. T is the total number of frames, and the total number of bark bands is B.

3. PROPOSED PERCEPTUALLY WEIGHTED NMF

NMF approximates the observation matrix $\mathbf{X} \in \mathbb{R}^{\geq 0, K\times T}$ as a product of basis matrix $\mathbf{B} \in \mathbb{R}^{\geq 0, K\times R}$ and gain matrix $\mathbf{G} \in \mathbb{R}^{\geq 0, R\times T}$ as $\mathbf{X} \approx \mathbf{BG}$. Matrix $\mathbf{X}$ consists of magnitudes of frame-wise DFTs of the observed audio signal, calculated in frames $t = 1, \ldots, T$. Only positive frequencies $k = 1, \ldots, K$ of the DFT are used. The rank of the decomposition is denoted by R, which is a free parameter chosen by the user.

Matrices $\mathbf{B}$ and $\mathbf{G}$ are estimated by minimizing the error of the approximation. Measures for the error include, for example, the squared Euclidean distance (EUC), generalized Kullback-Leibler divergence (KLD), and the Itakura-Saito divergence (ISD) [10].

3.1. NMR as cost function for NMF

The masking thresholds in $\mathbf{M}$ for certain observations $\mathbf{X}$ are calculated before the NMF algorithm. The mask estimation and NMR evaluation in PEAQ is defined in bark scale, but due to its lower resolution, we wish to perform the NMF decomposition in a linear frequency scale provided by the DFT. In the following we formulate the NMR objective into a weighted squared error, calculated in a linear frequency scale. Let use denote the squared error in Equation (1) as $\mathbf{E} = (\mathbf{X} - \hat{\mathbf{X}})^{\cdot 2}$. The measure (1) is a monotonic function ($\log_{10}$ and scalar multipliers) of term $\sum_{t=1}^{T}\sum_{b=1}^{B}[\mathbf{M}]_{b,t}[\mathbf{CHE}]_{b,t}$. Thus, minimizing the NMR is equivalent to minimizing the above term. In each frame t, the term can be formulated as

$$\sum_{b=1}^{B}[\mathbf{M}]_{b,t}[\mathbf{CHE}]_{b,t} = \sum_{k=1}^{K}\sum_{b=1}^{B}[\mathbf{M}]_{b,t}[\mathbf{CH}]_{b,k}[\mathbf{E}]_{k,t}$$
$$= \sum_{k=1}^{K}[\mathbf{W}]_{k,t}[\mathbf{E}]_{k,t}\ , \quad \text{where} \quad \mathbf{W} = (\mathbf{CH})^{\mathrm{T}}\mathbf{M}$$

The above formulation can be placed back to Equation (1) and the result is an NMR metric defined for linear frequency scale error

$$\mathrm{NMR_L} = 10\log_{10}\Big(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{B}\sum_{k=1}^{K}[\mathbf{W}]_{k,t}[\mathbf{X}-\hat{\mathbf{X}}]^{.2}_{k,t}\Big). \quad (2)$$

When applying the above equation as NMF cost function, we model $\hat{\mathbf{X}}$ using $\mathbf{BG}$. The resulting NMF criterion is the weighted squared Euclidean distance:

$$\mathrm{D_{WEUC}}(\mathbf{X}, \mathbf{BG}, \mathbf{W}) = \sum_{kt}[\mathbf{W}]_{k,t}([\mathbf{X}]_{k,t} - [\mathbf{BG}]_{k,t})^2, \quad (3)$$

The NMR quality criterion has been also implemented as cost function for NMF by O'Grady in [11]. His method calculated the error between the observed magnitude spectrogram and the model in bark bands, which does not allow modeling the fine spectral structure, that the linear frequency scale models.

3.2. Algorithm for minimizing the NMR

The weighted squared Euclidean distance and thus the proposed cost function can be minimized by the update rules proposed in [12] and applied in [11]. First, the entries of matrices $\mathbf{B}$ and $\mathbf{G}$ are initialized with random values normally distributed between zero and one. The matrices are updated iteratively using the update rules

$$\mathbf{B} \leftarrow \mathbf{B}. \times \frac{(\mathbf{W}. \times \mathbf{X})\mathbf{G}^{\mathrm{T}}}{(\mathbf{W}. \times (\mathbf{BG}))\mathbf{G}^{\mathrm{T}}}$$
$$\mathbf{G} \leftarrow \mathbf{G}. \times \frac{\mathbf{B}^{\mathrm{T}}(\mathbf{W}. \times \mathbf{X})}{\mathbf{B}^{\mathrm{T}}(\mathbf{W}. \times (\mathbf{BG}))}, \quad (4)$$

where operators $.\times$ and $\frac{\mathbf{X}}{\mathbf{Y}}$ denote element-wise multiplication and division, respectively. The update rules are repeated until the algorithm converges.

4. SIGNAL RECONSTRUCTION AND WEIGHT SMOOTHING

The above section described the model parameter estimation stage of the algorithm. In signal analysis the estimated parameters can be used as such, but for example in audio coding applications a signal needs to be reconstructed from the parameters.The synthesis procedure requires generating the phases for the reconstructed magnitude spectrogram $\mathbf{BG}$, applying inverse DFT in each frame, and combining the frames by overlap-add.

An example of an algorithm that can be used to generate the phases has been proposed in [13]. Our main focus in this study is in the magnitude spectrogram modeling and in order to prevent the artefacts caused by the phase reconstruction from affecting the evaluation, we use the phase spectrogram estimated from the original signal, as illustrated in Figure 1.

The NMF cost function derived in the previous section does not take into account the synthesis procedure, i.e., it assumes that the magnitude spectrogram of the synthesized signal equals $\hat{\mathbf{X}}$ in (1). In practice, the overlap-add synthesis procedure affects the quality in the sense that an error produced in a frame is spread to the neighboring frames where it may become audible. Specifically, the phenomenon becomes prominent if a quiet frame is followed by an intense one where fair amount of error is produced. In audio coding the phenomenon is called pre-echo.

We approximate the effect of the synthesis procedure by assuming that the modeling error $[\mathbf{E}]_{k,t}$ of the magnitude spectrograms in frame t is divided into frames $t-1$, t, and $t+1$ by weights h_{-1}, h_0, and h_1, respectively. We use values α, $1-2\alpha$, and α for the weights, where the amount of spreading defined by the parameter α is dependent on the shape of the window function. We also assume that the errors produced in adjacent frames are independent from each other, so that the errors (represented by energies) are additive. In practice the spreading depends on the lengths and relative positions of the windows of the synthesis filter bank and the analysis filter bank in NMR, but for simplicity we restrict ourselves to the above approximation. The spread error is given as

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{B}\sum_{\tau=-1}^{1}\sum_{k=1}^{K}[\mathbf{W}]_{k,t}[\mathbf{E}]_{k,t-\tau}h_\tau,$$

which can be formulated as

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{B}\sum_{k=1}^{K}[\mathbf{W}']_{k,t}[\mathbf{E}]_{k,t},$$

where $[\mathbf{W}']_{k,t}=\sum_{\tau=-1}^{1}[\mathbf{W}']_{k,t+\tau}h_\tau$. Thus the effect of the synthesis filterbank can be taken into account by filtering the weights $\mathbf{W}$ in time. The simulation results show that the overall quality is slightly improved by the spreading.

5. SIMULATION AND RESULTS

The proposed NMF algorithm was tested by applying it to various styles of audio signals and measuring the NMR of the synthesized signals. The test set consisted of 10-second monaural excerpts from following categories (number of entries in brackets): classical music (16), drum patterns (20), western pop music (24), solo instruments (20), solo singing (10) and speech (10), equaling to total of 100 samples. The speech samples have a 16 kHz sampling frequency, whereas the rest of them have a 44.1 kHz sampling frequency.

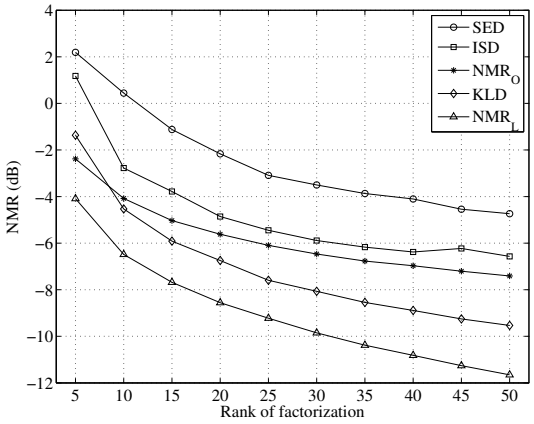


Fig. 2. NMR of the tested NMF algorithms as the function of the rank of factorization

Each test sample was processed using the method illustrated in Figure 1. We used Kaiser-Bessel derived window function [3, p 171] in analysis and synthesis, since it was found to produce the best performance among various tested window functions. We used 50% overlap between adjacent windows. The synthesized signals were evaluated with NMR criterion described in PEAQ and the average NMR over the whole test set was calculated. The scaling parameter L_p was set to 40 dB.

The tested NMF algorithms were EUC, KLD, ISD and proposed NMR_L. The weighting method from [11] is denoted as NMR_O. The masking estimation for NMF was done using a 42.7ms window, but the hop size was set equal to the NMF hop size. The number of iterations was chosen by calculating NMR after each iteration to determine the rate of convergence for a subset of the test signals. The experiments showed that EUC and NMR_L needed more iterations to converge. The number of iterations was set to 200 for KLD and ISD and 400 for EUC and NMR_L.

The results of different ranks of factorization with a 20 ms window are shown in Figure 2. Results indicate that the proposed method enables on average 1.9 dB better NMR than the best reference method. The test was also repeated for 40ms window and the results were very similar, the advantage of the proposed method being again approximately 1.6 dB. Few demonstrative test signals are available at http://www.cs.tut.fi/sgn/arg/nikunen/demo/icassp2010/.

Increasing the hop size will reduce the amount frames per second. From audio coding point of view this decreases the amount of gains to be represented. The number of frequency indices for each source in $\mathbf{B}$ is half of the window length, since the DFT length equals the window length and only positive frequencies are retained. We restrict the hop size to be 50% of the window size, and therefore longer windows will result to longer DFTs, which need to be encoded as well. We consider each parameter to be represented as a particle, and study the effect of the frame length and the rank of factorization when constraining a fixed amount of particles per second. The total amount of particles per second in a decomposition is $P=(Z+K/S)R$, where Z denotes the number of frames per second, K is the number of positive DFT coefficients, S is the signal length in seconds and R is the rank of factorization.

We fixed the amount particles per second to 3000, and determined the parameters by selecting a certain rank of factorization and searching for the shortest possible window that did not exceed the particle rate. The results with different ranks of factorization are

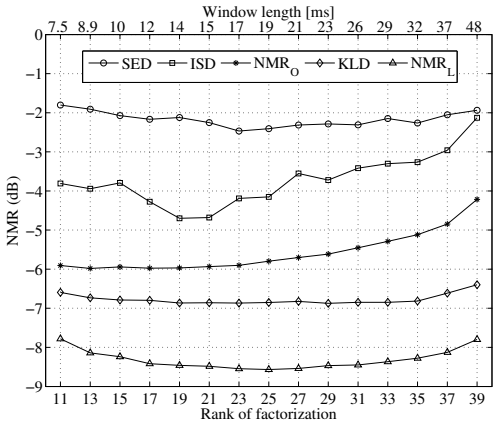


Fig. 3. NMR as the function of the window length and the rank of factorization when 3000 particles per second are used

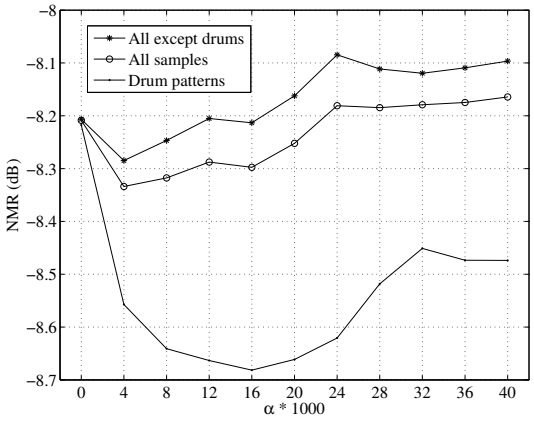


Fig. 4. Weights $\mathbf{W}$ filtered with different time averaging filters, NMR evaluated over whole test set, without drum patterns and only drum patterns

shown in Figure 3. For this test we used 30-second excerpts where the total number of samples was 50. The window lengths depend on the sampling frequency. In the figure they are denoted for the signals with sampling frequency of 44100 Hz. Considering the average quality, the range of equally good parameter combinations seems to be wide for all the NMF algorithms. The quality decreases only when a too short or a too long window is used. By examining the results of individual samples it seems that a good combination depends greatly on the signal to be composed.

Figure 4 illustrates the average NMR as the function of the spreading parameter α. The average is calculated separately for drum signals, which contain lot of transients, and thus the pre-echo phenomenon is assumed to be the largest. It can be seen that with a suitable value of α, the filtering improves the average quality NMR of drums by 0.4 dB. For other signals the filtering does not improve the quality.

6. CONCLUSION

We have proposed a method for minimizing the noise-to-mask ratio using non-negative matrix factorization. We have formulated the noise-to-mask ratio calculated on bark-band signal representation as a cost function for linear-frequency NMF. Simulation experiments show that the proposed method allows better quantitative perceptual quality than the reference methods. The proposed method for spreading the masking patterns in time enables a better quality for signals with plenty of transient sounds. The overall results show improvement of audio quality in benefit for proposed method and it could be plausible for future object-based audio coding applications.

7. REFERENCES

[1] D. Ellis and D.F. Rosenthal, "Mid-level representations for computational auditory scene analysis," in *Proceedings of International Joint Conference on Artificial Intelligence – Workshop on Computational Auditory Scene Analysis*, Montreal, Canada, 1995.

[2] E. Vincent and M.D. Plumbley, "Low bit-rate object coding of musical audio using Bayesian harmonic models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1273–1282, 2007.

[3] Andreas Spanias, Ted Painter, and Venkatraman Atti, *Audio Signal Processing and Coding*, John Wiley & Sons, 2007.

[4] K. Brandenburg and T. Sporer, "NMR and Masking Flag: Evaluation of Quality Using Perceptual Criteria," in *Proceedings of the AES 11th International Conference on Test and Measurement*, Portland, USA, May 1992, pp. 169–179.

[5] E. D. Scheirer, "Structured audio and effects processing in the MPEG-4 multimedia standard," *Multimedia Systems*, vol. 7, no. 1, pp. 11–22, 1999.

[6] Tuomas Virtanen, "Monoaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE Transactions on Audio, Speech and Language processing*, vol. 15, pp. 1066–1074, 2007.

[7] E. Zwicker and H. Fastl, *Psychoacoustics Facts and Models*, Springer-Verlag, 1990.

[8] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, "PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality," *Journal of the Audio Engineering Society*, vol. 48, pp. 3–29, 2000.

[9] P. Kabal, "An Examination and Interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality," Tech. Rep., Department of Electrical & Computer Engineering, McGill University, 2002.

[10] C. Févotte and Cemgil A. T., "Nonnegative matrix factorizations as probabilistic inference incomposite models," in *17th European Signal Processing Conference*, Scotland, 2009.

[11] Paul D. O'Grady, *Sparse Separation of Under-Determined Speech Mixtures*, Ph.D. thesis, Nui Maynooth, 2007.

[12] Tuomas Virtanen, "Separation of Sound Sources by Convolutive Sparse Coding," in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, Jeju, Korea, 2004.

[13] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, Brisbane, Australia, 2008.

Publication P2

J. Nikunen and T. Virtanen, "Object-based Audio Coding Using Non-negative Matrix Factorization for the Spectrogram Representation," in *Proceedings of the 128th Audio Engineering Society Convention*, London, U.K., 2010

Publication P3

J. Nikunen, T. Virtanen and M. Vilermo, "Multichannel Audio Upmixing Based on Non-negative Tensor Factorization Representation," in *Proceedings of Workshop on Applications of Signal Processin to Audio and Acoustics*, New Paltz, NY, USA, 2011, pp. 33-36.

MULTICHANNEL AUDIO UPMIXING BASED ON NON-NEGATIVE TENSOR FACTORIZATION REPRESENTATION

J. Nikunen, T. Virtanen

Tampere University of Technology
Korkeakoulunkatu 1, 33720 Tampere, Finland
joonas.nikunen@tut.fi, tuomas.virtanen@tut.fi

M. Vilermo

Nokia Research Center
Visiokatu 1, 33720 Tampere, Finland
miikka.vilermo@nokia.com

ABSTRACT

This paper proposes a new spatial audio coding (SAC) method that is based on parametrization of multichannel audio by sound objects using non-negative tensor factorization (NTF). The spatial parameters are estimated using perceptually motivated NTF model and are used for upmixing a downmixed and encoded mixture signal. The performance of the proposed coding is evaluated using listening tests, which prove the coding performance being on a par with conventional SAC methods. The novelty of the proposed coding is that it enables controlling the upmix content by meaningful objects.

Index Terms— Spatial audio coding, Object-based audio coding, Non-negative tensor factorization

1. INTRODUCTION

The audio coding research have recently focused on spatial audio coding (SAC), where one or few discretely encoded signal channels are transmitted with additional spatial cues for synthesis of multiple channels. The existing SAC algorithms are mostly based on binaural cue coding principles [1]. Algorithms include for example parametric stereo coding [2] and coding of multichannel audio [3].

The spatial synthesis from a downmixed signal is based on adjusting the level, time delay and decorrelation of the time-frequency blocks used for spatial parameter estimation. The parametrization relies on assumption of non-overlapping sound sources in the frequency domain or momentary dominance of certain sound source in perception of direction. Directional audio coding (DirAC) [4] shares the above assumptions but improves the parameter estimation and spatial sound field reproduction by diffuseness measure and spatial impulse response rendering. Another degree of parametrization in audio coding is using objects having interpretable structure, for example audio objects based on harmonic components of instruments [5]. In this paper we focus on incorporating SAC with an object-based model for spatial parametrization.

We propose a new object-based SAC algorithm utilizing audio spectrogram parametrization by non-negative tensor factorization (NTF) algorithm, which estimates object spectra and their spatial parameters simultaneously. The NTF spatial parametrization is used for recovering the multichannel signal from a downmixed and perceptually encoded stereo signal by filtering the downmix short-time Fourier transform (STFT) in Wiener filtering manner using the NTF model as a time-frequency filter kernel.

The proposed approach relies on object parametrization from the mixture signal in a blind sound separation manner using non-negative matrix factorization (NMF) and its extension to multidimensional data by NTF [6]. The NMF algorithm with various extensions has been intensively studied for blind sound source separation [7, 8]. The separation is based on ability of NMF algorithm to find and model repetitive structures from audio signals using a single object. The NMF objects usually represent sound structures such as individual notes of an instrument, chords or drum hits.

In addition to object separation, the advantage of NTF representation for SAC is that it utilizes long-term redundancy present in an audio signal by using a single object to describe repetitive sound events. Additionally, the NTF signal model estimates the spatial parameters and the object spectrum simultaneously allowing utilization of inter-channel redundancy in representation of the spatial information. Such non-redundant spatial representation is efficient with respect to coding and bitrate performance. In comparison to most existing SAC methods, NTF allows representing overlapping frequency content of the objects, which enables better spatial synthesis and separation of such simultaneous sound events.

The rest of the paper is organized as follows. In Section 2 the novel method for object-based spatial audio encoding and decoding utilizing NTF for signal parametrization is proposed. In Section 2.1 a perceptually motivated NMF cost function [9] is extended for NTF and multichannel observations. The upmix filtering framework for spatial synthesis with NTF is proposed in Section 2.2. The estimation of NTF parameters optimized for the upmix operation is proposed in Section 2.3 and the quantization end encoding of the parameters is shortly revised in Section 2.4. The results from a listening test are provided in Section 3.

2. PROPOSED METHOD FOR SPATIAL AUDIO CODING

The encoding and decoding of the proposed SAC algorithm are illustrated in Figures 1 and 2, respectively. The encoding starts by calculating the STFT of each input signal channel to obtain a magnitude spectrogram tensor. The masking level is estimated from the input signal and the NTF algorithm with perceptual weighting is applied to the spectrogram tensor to obtain an object-based spatial parametrization of it. The input signal is downmixed to stereo and perceptually encoded. The encoded downmix is STFT analyzed and used as an additional time-frequency weighting to optimize the NTF spatial parametrization for the upmix filtering operation. The estimated spatial parameters are quantized and entropy encoded.

The decoding to recover the multichannel signal starts by decoding of the downmix and calculating its STFT. The spatial parameters are decoded and dequantized. The upmixing is done by filtering the downmix STFT in a Wiener filtering manner where the channel dependent filter kernels are obtained from the NTF model. The time-domain signals are synthesized using the phases obtained from the downmix STFT analysis for every upmixed channel.

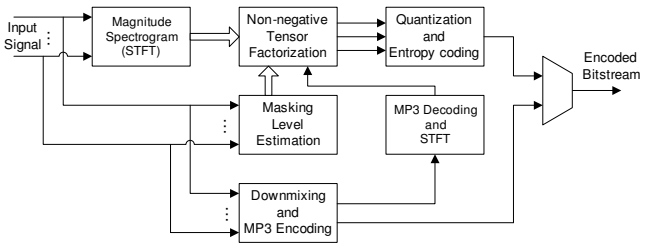


Figure 1: Block diagram of the encoding part of the proposed coding algorithm.

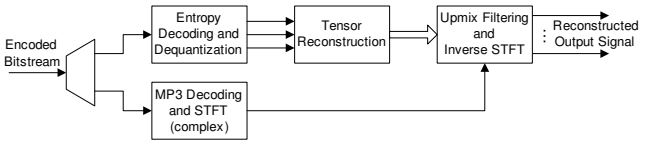


Figure 2: Block diagram of the decoding part of the proposed coding algorithm.

2.1. Non-negative Tensor Factorization

In this section the NTF representation of multichannel magnitude spectrogram tensor is introduced and we extend the perceptual weighting proposed in [9] for multichannel observations. We will use the following notation. Tensors are denoted by capital bold letters and a single entry of rank-j tensor $\mathbf{X}$ is denoted as $\mathbf{X}_{i_1,i_2,\ldots i_j}$.

For a multichannel time-domain audio signal $x(n,c)$, of sample index $n = 1,\ldots N$ and in channels $c = 1,\ldots,C$, absolute values of its STFT are denoted by $\mathbf{X}_{k,t,c}$, where $k = 1,\ldots,K$ is the positive DFT frequency bin index and $t = 1,\ldots,T$ is the STFT frame index. STFT is calculated using frame length of $N = 2(K-1)$ samples and consecutive frames are overlapping by $N/2$ samples, Hanning window function is used.

The NTF signal model for approximating spectrogram tensor $\mathbf{X}$ of rank three can be written as a product of three matrix entries summed over the decomposition objects r as

$$\mathbf{X}_{k,t,c} \approx \hat{\mathbf{X}}_{k,t,c} = \sum_{r=1}^{R} \mathbf{B}_{k,r}\mathbf{G}_{r,t}\mathbf{A}_{r,c}, \tag{1}$$

where R is the number of NTF objects used for the approximation. Each column of $\mathbf{B}$ contains the DFT spectrum of an object. The corresponding row of $\mathbf{G}$ represents its gain in each STFT frame, and the corresponding row of $\mathbf{A}$ represents the channel-dependent gain of the object. The NTF model has been found to produce good results in sound source separation in [6]. The NTF model (1) constitutes from a set of fixed object spectra that have time-varying gain and a channel-dependent gain.

2.1.1. Perceptually Motivated Weighting for NTF

The cost function to be minimized in finding the NTF approximation is the noise-to-mask ratio (NMR) [10], which evaluates perceptual quality of encoded audio by determining audibility of encoding artefacts based on masking phenomenon invoked by the desired signal content. The perceptually motivated NMF algorithm, minimizing the NMR of the approximation by multiplicative updates of the model parameters was proposed in [9].

We propose to extend the NMR cost function for NMF [9] to be used with the NTF signal model (1). The masking level for each frame is estimated in Bark band domain and the masking level conversion to any desired DFT frequency resolution is given in [9]. We will denote the masking level for the each time-frequency point in each input channel c by a tensor $\mathbf{W}_{k,t,c}$. The NMR measure equals to squared Euclidean distance of the original and NTF spectrogram weighted by the masking level and can be defined as

$$c_{\mathrm{NMR}} = \sum_{c=1}^{C}\sum_{t=1}^{T}\sum_{k=1}^{K} \mathbf{W}_{k,t,c}(\mathbf{X}_{k,t,c} - \sum_{r=1}^{R}\mathbf{B}_{k,r}\mathbf{G}_{r,t}\mathbf{A}_{r,c})^2. \tag{2}$$

2.1.2. Estimation of the Perceptually Motivated NTF

The estimation of the NTF model minimizing NMR (2) is achieved by iterative multiplicative updates, which can be derived using same principles as in [7]. The update rules are given as

$$\begin{aligned}
\mathbf{B}_{k,r} &\leftarrow \mathbf{B}_{k,r}\frac{\sum_t\sum_c \mathbf{W}_{k,t,c}\mathbf{X}_{k,t,c}\mathbf{G}_{r,t}\mathbf{A}_{r,c}}{\sum_t\sum_c \mathbf{W}_{k,t,c}\hat{\mathbf{X}}_{k,t,c}\mathbf{G}_{r,t}\mathbf{A}_{r,c}},\\
\mathbf{G}_{r,t} &\leftarrow \mathbf{G}_{r,t}\frac{\sum_k\sum_c \mathbf{B}_{k,r}\mathbf{W}_{k,t,c}\mathbf{X}_{k,t,c}\mathbf{A}_{r,c}}{\sum_k\sum_c \mathbf{B}_{k,r}\mathbf{W}_{k,t,c}\hat{\mathbf{X}}_{k,t,c}\mathbf{A}_{r,c}},\\
\mathbf{A}_{r,c} &\leftarrow \mathbf{A}_{r,c}\frac{\sum_k\sum_t \mathbf{B}_{k,r}\mathbf{G}_{r,t}\mathbf{W}_{k,t,c}\mathbf{X}_{k,t,c}}{\sum_k\sum_t \mathbf{B}_{k,r}\mathbf{G}_{r,t}\mathbf{W}_{k,t,c}\hat{\mathbf{X}}_{k,t,c}},
\end{aligned} \tag{3}$$

where $\hat{\mathbf{X}}_{k,t,c}$ is the reconstructed NTF model evaluated according to (1) before each update.

The complete NTF algorithm is as follows. First the entries of matrices $\mathbf{B}, \mathbf{G}$ and $\mathbf{A}$ are initialized with random values uniformly distributed between zero and one. The decomposition matrices are then iteratively updated by applying the updates (3) for each of the matrices at a time. A fixed number of iterations is used.

2.2. Object-based Spatial Upmixing Using NTF Model

In this section we will propose an object-based spatial upmixing method for recovering the multichannel signal. For streamlined representation we will derive the upmixing model only for stereo downmix, but it can be defined for any other desired input signal and dowmix channel configuration.

The original multichannel time-domain signal $x(n,c)$ is downmixed to stereo by

$$l(n) = \sum_{c\in\mathcal{L}} x(n,c), \qquad r(n) = \sum_{c\in\mathcal{R}} x(n,c), \tag{4}$$

where $l(n)$ and $r(n)$ are the left and right channel respectively. For a 5.1 speaker configuration $\mathcal{L}$ contains front and rear left channel with center and low-frequency extension, $\mathcal{R}$ contains respectively the right side counterparts. The downmixed time-domain signal is perceptually encoded and is available at the decoder. The decoded downmix signal is STFT analyzed using same analysis parameters (window length, etc.) to obtain complex downmix STFT spectrogram $\mathbf{L}_{k,t}$ and $\mathbf{R}_{k,t}$ for left and right stereo channels respectively.

The object-based multichannel signal model is used for upmixing the downmixed STFT as follows. The STFT of the upmixed

signal is obtained as

$$\mathbf{Y}_{k,t,c} = \begin{cases} \mathbf{L}_{k,t} \dfrac{\sum_{r=1}^{R} \mathbf{B}_{k,r}\mathbf{G}_{r,t}\mathbf{A}_{r,c}}{\sum_{c\in\mathcal{L}}\sum_{r=1}^{R} \mathbf{B}_{k,r}\mathbf{G}_{r,t}\mathbf{A}_{r,c}}, & \text{(5a)} \quad c \in \mathcal{L}, c \notin \mathcal{R} \\ \mathbf{R}_{k,t} \dfrac{\sum_{r=1}^{R} \mathbf{B}_{k,r}\mathbf{G}_{r,t}\mathbf{A}_{r,c}}{\sum_{c\in\mathcal{R}}\sum_{r=1}^{R} \mathbf{B}_{k,r}\mathbf{G}_{r,t}\mathbf{A}_{r,c}}, & \text{(5b)} \quad c \in \mathcal{R}, c \notin \mathcal{L} \\ \dfrac{1}{2}\left[\text{(5a)} + \text{(5b)}\right] & c \in \mathcal{L}, \mathcal{R} \end{cases} \tag{5}$$

Note that $\mathbf{L}_{k,t}, \mathbf{R}_{k,t}$ and $\mathbf{Y}_{k,t,c}$ are complex-valued. Time-domain signals are obtained by inverse STFT and overlap-add. The above can be viewed as filtering the downmix to multiple channels using a time-varying Wiener filter.

Similar filtering methods are widely used in reconstruction of source signals when NMF or NTF is used for sound source separation, for example in [8]. The proposed method allows synthesizing only selected objects. In this case the summation in numerator of (5) is evaluated over desired group of objects $r \in \mathcal{O}$. This is refererred as control of the upmix content based on meaningful objects.

2.3. NTF Parameter Optimization for the Upmix Filtering

Taking into account the filtering operation (5) the NMR cost of NTF model becomes

$$c = \sum_{c=1}^{C}\sum_{t=1}^{T}\sum_{k=1}^{K} \mathbf{W}_{k,t,c}(\mathbf{X}_{k,t,c} - |\mathbf{Y}_{k,t,c}|)^2 \tag{6}$$

where $\mathbf{Y}$ is evaluated according to (5). The difference between NTF cost functions (2) and (6) is that the latter takes into account the downmixing process and particularly the case where sound events from different spatial positions are overlapping or are closely separated in time and frequency. Such case will introduce cross-talk to the upmixed channels if cost function (2) and updates (3) are used. This equals to filtering undesired downmix STFT details to the upmixed channels.

We propose to approximate the upmixing cost function (6) by giving bigger weighting for time-frequency bins in which the downmix STFT has high magnitude with respect to the NTF channel sum in equation (8). These time-frequency bins are assumed to contain overlapping but spatially separated signal content and the proposed weighting assigns more NTF modeling accuracy for regions of such sound events. This is achieved by replacing $\mathbf{W}_{k,t,c}$ in (3) with

$$\hat{\mathbf{W}}_{k,t,c} = \begin{cases} \mathbf{W}_{k,t,c}\dfrac{\mathbf{L}_{k,t}}{\hat{\mathbf{L}}_{k,t}}, & c \in \mathcal{L}, c \notin \mathcal{R} \\ \mathbf{W}_{k,t,c}\dfrac{\mathbf{R}_{k,t}}{\hat{\mathbf{R}}_{k,t}}, & c \in \mathcal{R}, c \notin \mathcal{L} \\ \mathbf{W}_{k,t,c}\dfrac{1}{2}\left(\dfrac{\mathbf{L}_{k,t}}{\hat{\mathbf{L}}_{k,t}} + \dfrac{\mathbf{R}_{k,t}}{\hat{\mathbf{R}}_{k,t}}\right), & c \in \mathcal{L}, \mathcal{R} \end{cases} \tag{7}$$

where

$$\hat{\mathbf{L}}_{k,t} = \sum_{c\in\mathcal{L}}\sum_{r=1}^{R} \mathbf{B}_{k,r}\mathbf{G}_{r,t}\mathbf{A}_{r,c}, \quad \hat{\mathbf{R}}_{k,t} = \sum_{c\in\mathcal{R}}\sum_{r=1}^{R} \mathbf{B}_{k,r}\mathbf{G}_{r,t}\mathbf{A}_{r,c} \tag{8}$$

The weights $\hat{\mathbf{W}}$ need to be updated after every NTF iteration due the change of $\hat{\mathbf{L}}_{k,t}$ and $\hat{\mathbf{R}}_{k,t}$.

The practical implementation of the proposed weighting is done by first using cost function (2) and parameter updates (3) for several hundreds of iterations to estimate an initial NTF model and then change to (7) to optimize the NTF model to the given downmix. Even though we cannot proof that algorithm described above would minimize (6), the experiments with the implementation have shown to produce desired result.

The behavior of the cost function (6) measuring the upmixed signal NMR was investigated to prove its decrease with the proposed weighting. The evaluation of (6) was done first with the update rules (3) and then changing to the proposed updates (7). The cost function was averaged over the whole test set described in Section 3 and the resulting cost is illustrated in Figure 3. Detailed encoding settings are given in Section 3. The decrease of the upmix filtering NMR cost is evident when switching to the proposed updates at iteration 500.

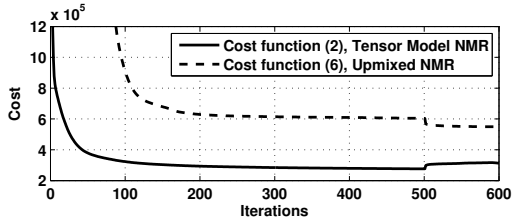


Figure 3: Behaviour of cost function (6) with the NTF algorithm updates (3) for the first 500 iterations and with updates (7) for successive 100 iterations.

2.4. Quantization and Encoding of the NTF Parameters

The NTF model derived in Section 2.3 is quantized and entropy coded and sent as a side information for spatial synthesis by the proposed upmix filtering. We use the quantization framework proposed in [11], which applies a non-uniform quantization for object spectrum $\mathbf{B}_{k,r}$ and gains $\mathbf{G}_{r,t}$ in such way that more quantization levels are assigned for smaller parameter values. For quantization of the channel gain parameter $\mathbf{A}_{r,c}$ we use uniform quantization.

In [11] the frequency of occurrence of quantized values of $\mathbf{B}_{k,r}$ and $\mathbf{G}_{r,t}$ was gathered from a large test set resulting to distributions having high probability of zeros and rest of the quantization levels had relatively small probability. Such distributions of quantized values can be effectively utilized for reducing the output bitrate by entropy coding. In the case of proposed SAC algorithm we calculated the entropy of each individual model parameter $\mathbf{B}_{k,r}, \mathbf{G}_{r,t}$ and $\mathbf{A}_{r,c}$ to estimate the final bitrate after entropy coding.

3. EVALUATION

In this section we will present listening test results of the proposed SAC algorithm when evaluated using multiple stimuli with hidden reference and anchor (MUSHRA) [12] methodology and comparing the coding quality to MP3 surround [3] at similar bitrates. Test samples used were the MPEG multichannel evaluation samples.

The listening test was run in Nokia Research Center listening room, which is fully conformant with ITU-R BS.1116-1 [13]. Speakers were set up according to ITU-R BS.775. 3.5kHz low-pass filtered original was used as a lower anchor. A lower anchor with spatially reduced quality was deemed unnecessary since all listeners

were experts. 10 listeners participated in the test. Listeners were instructed to grade the samples taken into account all coding artefacts including spatial sound image.

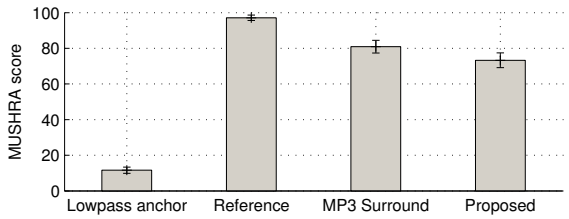


Figure 4: Listening test mean score and 95% confidence intervals.

Encoding parameters were chosen as follows, the window length was set to $N = 880$ samples, which equals to 20ms with the sampling frequency $F_s = 44.1$ kHz. The length of the NTF segment was chosen to 15 seconds. The number of NTF algorithm iterations were set to 500 with updates (3), and 100 with updates (7). The number of bits for representing each NTF parameter with quantization described in Section 2.4 was defined by preliminary listening tests. Evaluation resulted to using $n_b = 4, n_g = 4$ and $n_a = 6$ bits per parameter for $\mathbf{B}_{k,r}$, $\mathbf{G}_{r,t}$ and $\mathbf{A}_{r,c}$ respectively.

The stereo core encoding algorithm was MP3 at 96 kbps and the target bitrate with the NTF upmixing side information was 128 kbps. The number of NTF objects was determined by allocating the remaining bitrate after downmix encoding to the NTF model. The bitrate reduction by entropy coding was estimated as described in Section 2.4. The resulting number of objects $R = 64$ corresponds to a NTF bitrate of 26.0 kbps after quantization and averaging the estimated entropy coding bitrate over the whole test set.

The listening test results are given in Figure 4. The results indicate that with similar bitrates the proposed coding attains slightly lower mean score than the compared SAC method, MP3 Surround. However, the score difference is small which indicates that coding performance achieved by the proposed algorithm is comparable to the existing SAC methods. Both tested SAC methods do not achieve transparency compared to the hidden reference, but the overall quality level can be considered to be moderate and suitable for coding of multichannel audio for consumer applications. The listening results prove that the proposed SAC method can be used for coding of multichannel audio with at bitrates equivalent to 128 kbps or similar.

The proposed algorithm achieved coding performance comparable to conventional SAC approaches and additionally the proposed upmix filtering allows manipulation of the upmix content by NTF objects, corresponding to meaningful sound events. The NMF and NTF signal decomposition models have been shown to achieve promising results in blind sound source separation [6, 7, 8]. Combining the proposed coding with separation would produce SAC with possibility to control the content of the upmix for example by instruments. The sound separation performance of the proposed SAC method was informally evaluated by k-means clustering of the NTF objects with spectral and time-gain based features. The separation performance was determined to be promising and comparable to separation results achieved in [6, 7].

4. CONCLUSION

We proposed a novel method for spatial audio coding (SAC) by using non-negative tensor factorization (NTF) for deriving an object-based spatial upmixing model. The spatial synthesis was done in Wiener filtering manner using NTF representation as a time-frequency filtering kernel and we proposed an experimental algorithm for estimating the NTF parameters found to minimize the filtering cost. The listening test showed that the proposed SAC algorithm achieved the performance of conventional spatial audio coding methods, but additionally enabling the control of the upmix by objects in blind sound separation manner. The future work will include more extensive evaluation of sound source separation performance of the proposed method.

5. REFERENCES

[1] F. Baumgarte and C. Faller, "Binaural Cue Coding-Part I: Psychoacoustic Fundamentals and Design Principles," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 509–519, 2003.

[2] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegård, "Low complexity parametric stereo coding," in *Proc. of 116th AES Convention, Berlin, Germany*, 2004.

[3] J. Herre, C. Faller, C. Ertel, J. Hilpert, A. Hoelzer, and C. Spenger, "MP3 Surround: Efficient and compatible coding of multi-channel audio," in *Proc. of 116th AES Conv.*, 2004.

[4] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of AES*, vol. 55, no. 6, p. 503, 2007.

[5] E. Vincent and M. Plumbley, "Low bit-rate object coding of musical audio using Bayesian harmonic models," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1273–1282, 2007.

[6] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative Tensor Factorisation for Sound Source Separation," in *Proc. of the Irish Signals and Systems Conference, Dublin, Ireland*, 2005.

[7] T. Virtanen, "Monoaural Sound Source Separation by Non-negative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, pp. 1066–1074, 2007.

[8] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. on Audio, Speech and Language Process.*, vol. 18, no. 3, pp. 550–563, 2010.

[9] J. Nikunen and T. Virtanen, "Noise-to-Mask Ratio Minimization by Weighted Non-negative Matrix factorization," in *Proc. of ICASSP '10*, Dallas, USA, 2010.

[10] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Kheyl, G. Stoll, K. Brandenburg, and B. Feiten, "PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality," *Journal of AES*, vol. 48, pp. 3–29, 2000.

[11] J. Nikunen and T. Virtanen, "Object-based Audio Coding Using Non-negative Matrix Factorization for the Spectrogram Representation," in *Proc. of 128th AES Conv.*, London, UK, 2010.

[12] ITU-R BS. 1534-1, "Method for the subjective assessment of intermediate quality level of coding systems," *ITU Radiocomm. Assembly*, 2003.

[13] ITU-R BS. 1116-1, "Methods for the Subjective Assessment of Small Impairments in Audio Systems including Multichannel Sound Systems," *ITU Radiocomm. Assembly*, 1994.

Publication P4

J. Nikunen, T. Virtanen, P. Pertilä and M. Vilermo, "Permutation Alignment of Frequency-Domain ICA by the Maximization of Intra-Source Envelope Correlations," in *Proceedings of the European Signal Processing Conference*, Bukarest, Romania, 2012, pp. 1489–1493.

PERMUTATION ALIGNMENT OF FREQUENCY-DOMAIN ICA BY THE MAXIMIZATION OF INTRA-SOURCE ENVELOPE CORRELATIONS

J. Nikunen, T. Virtanen, P. Pertilä

Tampere University of Technology
Korkeakoulunkatu 1, 33720 Tampere, Finland

M. Vilermo

Nokia Research Center
Visiokatu 1, 33720 Tampere, Finland

ABSTRACT

This paper presents a novel method for solving the permutation ambiguity of frequency-domain independent component analysis based on source signal envelope correlation maximization. The proposed method is developed for blind source separation with high sampling frequency and significant spatial aliasing. We propose a method that analyzes the source envelope using a rank-one singular value decomposition (SVD) applied to an initial source magnitude spectrogram obtained by a time difference of arrival (TDoA) based permutation alignment method. The permutation for frequencies with incoherent TDoA are corrected by maximizing the cross-correlation of the SVD analyzed source activation vector and each independent component magnitude envelope. We evaluate the separation quality using real high sampling frequency speech captures and the proposed method is found to improve the separation over the baseline algorithm.

Index Terms— Blind Source Separation, Independent Component Analysis

1. INTRODUCTION

The blind source separation (BSS) of simultaneously emitting sound sources, generally known as the cocktail party problem, has been intensively studied over the years, but is however still categorized as an unsolved problem. In the course of this paper we pursue blind separation of high sampling frequency speech using independent component analysis (ICA) applied in frequency domain leading into frequency-wise permutation ambiguity. The permutation alignment have been previously solved for example based on mixing filter frequency response smoothness [1], temporal structure of the source signals [2], and time-difference of arrival (TDoA) and direction of arrival (DoA) [3, 4] interpretation of ICA mixing parameters. The latter can be considered as generally robust with no assumptions on the source characteristics, however their performance starts to degrade in reverberant conditions and with captures involving lot of spatial aliasing frequencies.

In this paper we propose a novel method for ICA permutation alignment that resolves the component ordering via maximization of intra-source envelope correlations. TDoA based algorithm [4] is used for obtaining an initial solution for the ICA parameter alignment which is to be improved by the proposed method. The proposed algorithm is applied for frequencies where the source TDoA is incoherent due spatial aliasing and reverberation making the source magnitude envelopes more accurate method for permutation alignment. The separation quality of the proposed method is evaluated using high sampling frequency speech captures and the results show an increase in separation quality measured using quantities proposed in [5].

The rest of the paper is organized as follows, in Section 2 we review the frequency domain ICA and the permutation alignment algorithms used in prior art. The proposed method is presented in Section 3. In Section 3.1 we shortly present the TDoA based permutation algorithm [4] used for obtaining an initial permutation solution. The proposed singular value decomposition (SVD) based source envelope analysis and the permutation alignment by maximization of intra-source envelopes is presented in Section 3.2. The source separation quality of speech samples is presented in Section 4.

2. BLIND SOURCE SEPARATION AND INDEPENDENT COMPONENT ANALYSIS

The array capture can be considered by the following convolutive mixture model in the time-domain

$$x_m(t) = \sum_{j=1}^{J} \sum_{\tau} h_{mj}(\tau) s_j(t-\tau) \quad (1)$$

where $x_m(t)$ is the mixture of $j = 1...J$ source signals capture by sensor $m = 1...M$ and sampled in time instances t. The spatial response from the source j to the sensor m is denoted by $h_{mj}(\tau)$ and the source signals are given as $s_j(t)$. Convolutive model (1) is usually approximated by instantaneous mixing in frequency domain as

$$\mathbf{x}(f,n) = \sum_{j=1}^{J} \mathbf{h}_j(f) s_j(f,n) \quad (2)$$

where $\mathbf{x}(f,n) = [x_1, \ldots, x_M]^T$ is the short-time Fourier transform (STFT) of the array capture $x_m(t)$, $f = 1...F$ is

the frequency index and $n = 1...N$ is the frame index. The impulse response $h_{mj}(\tau)$ is replaced with the frequency response denoted by $\mathbf{h}_j(f) = [h_{1j}, \ldots, h_{Mj}]^T$ and the STFTs of source signals are denoted by $s_j(f, n)$.

The ICA applied to the frequency domain model (2) has been successfully used for determined BSS [1, 2, 3, 4] where $M \geq J$. ICA is applied separately for each frequency bin f to obtain $J \times M$ unmixing matrix $\mathbf{W}$ as in

$$\mathbf{y}(f, n) = \mathbf{W}(f)\mathbf{x}(f, n). \tag{3}$$

where $\mathbf{y}(f, n) = [y_1, \ldots, y_J]^T$ corresponds to the sources $s_j(f, n)$ with an arbitrary permutation of sources indices at each frequency f. Further we assume that the unmixing matrix is invertible and define $\mathbf{A}(f) = \mathbf{W}(f)^{-1}$, thus we can write the ICA model as,

$$\mathbf{x}(f, n) = \mathbf{A}(f)\mathbf{y}(f, n) \tag{4}$$

If $J < M$, the mixing matrix is obtained via Moore–Penrose pseudoinverse $\mathbf{A}(f) = \mathbf{W}(f)^+$. $\mathbf{A}$ is constructed of column vectors $[\mathbf{a}_1, ..., \mathbf{a}_J]$ and each vector denotes the response of single source j to the each capturing sensor $m = 1, \ldots, M$.

In the earliest frequency-domain ICA based BSS methods [1] the permutation alignment was solved by assuming a smooth frequency response of the mixing filters $\mathbf{h}_j(f)$. Later in [2] the temporal structure of the separated signals $\mathbf{y}(f, n)$ was considered and the permutation was solved by maximizing cross-correlation of magnitudes of neighboring frequencies. TDoA and DoA interpretation of component bases $\mathbf{a}_j(f)$ has been proposed in [3] and in [6] the TDoA approach was combined with the magnitude envelope correlation maximization. More recently a method only relying on anechoic source signal propagation model estimation was proposed in [4], which will be used as a baseline in this paper.

There also exists ICA-based methods that unify the source dependencies across frequencies, independent vector analysis [7] and recursively regularized ICA across frequencies [8]. In this paper we will concentrate only to the frequency bin-wise ICA model (3) and improving the permutation alignment in case of high sampling frequency captures and severe spatial aliasing over the baseline [4]. Other related work combining TDoA with envelope correlation maximization include for example [6, 9].

3. PROPOSED METHOD

The proposed method for ICA permutation alignment combines a TDoA based algorithm [4] with a novel source envelope analysis by rank-one SVD and source temporal activity cross-correlation maximization across frequencies. With the proposed algorithm we aim for improving performance of TDoA based algorithms with high sampling frequency captures by using source magnitude envelope information in the permutation alignment.

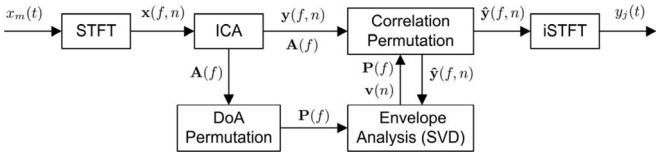


Fig. 1. Block diagram of the frequency domain ICA and the proposed permutation alignment by source envelope cross-correlation.

The block diagram of the proposed method is illustrated in Figure 1. First the input signal x_m is STFT analyzed to get $\mathbf{x}(f, n)$. The ICA is applied for each frequency f separately to obtain mixing matrix $\mathbf{A}(f)$ and the source signals $\mathbf{y}(f, n)$. The mixing matrix entries $\mathbf{a}_j(f)$ are clustered using a TDoA permutation alignment algorithm to get an initial permutation matrices $\mathbf{P}(f)$. The source signals $\mathbf{y}(f, n)$ are aligned using the initial permutations to obtain $\hat{\mathbf{y}}(f, n)$ and the source envelopes are analyzed using rank-one SVD. The obtained source envelope $\mathbf{v}(n) = [v_1, \ldots, v_J]^T$ is used for finding the permutation that maximizes the cross-correlation with $|\hat{\mathbf{y}}(f, n)|$ at each frequency f. The SVD envelope analysis and cross-correlation matching is repeated until no changes are made for $\hat{\mathbf{y}}(f, n)$ The time domain source signals are obtained via inverse STFT.

3.1. Permutation Alignment by Signal Propagation Model

The initial alignment of separated components is obtained by algorithm presented in [4], which is shortly reviewed in this section. The algorithm provides initial magnitude spectrogram matrices $|\hat{\mathbf{y}}(f, n)|$ in order to be SVD analyzed and corrected by the proposed algorithm presented in Section 3.2.

The parameters $\mathbf{a}_j(f)$ are phase and amplitude normalized with respect to a chosen reference sensor by subtracting the reference sensor phase and dividing by its norm, result is denoted by $\tilde{\mathbf{a}}_j(f)$. Normalization gives the relative TDoA of the mixing parameters in terms of phase difference with respect to the reference sensor. The source propagation model is defined as

$$\hat{h}_{mj}(f) = \lambda_{mj} \exp(-i2\pi f \tau_{mj}) \tag{5}$$

which approximates the mixing filter frequency response $h_{mj}(f)$ by having a fixed time delay τ_{mj} and attenuation λ_{mj} from source j to each capturing sensor m over all frequencies. The propagation model (5) translates into a fixed spatial position in means of TDoA in anechoic conditions, which further can be viewed as DoA estimate of the source.

The permutations are solved by minimizing the cost function

$$D = \sum_{j=1}^{J} \sum_{f=1}^{F} ||\tilde{\mathbf{a}}_{P_f(j)}(f) - \hat{\mathbf{h}}_j(f)||^2 \tag{6}$$

where the permutation of $\tilde{\mathbf{a}}_j(f)$ for each frequency f is given by $P_f(j)$ and the propagation model (5) is given in vector

form $\hat{\mathbf{h}}_j(f)$. The permutation alignment and the propagation model estimation is solved simultaneously and the correct permutations depend on the accuracy of the estimated propagation model. With no further algorithm details we assume to obtain the permutation matrix $\mathbf{P}(f)$ for changing the rows of $\mathbf{y}(f,n)$ and estimated propagation model $\hat{\mathbf{h}}_j(f)$ that minimizes the cost function (6). The details of the algorithm can be found from [4].

3.2. Source Envelope Analysis and Cross-correlation Maximization of Magnitude Envelopes

We start the derivation of the proposed algorithm by considering which of the frequency indices after the permutation alignment given in Section 3.1 have high confidence of being correct. These frequency indices are used as a reference for analyzing the source envelopes using a rank-one SVD. The proposed algorithm is applied for correcting the permutation of the rest of the frequency indices.

The confidence of correct permutation at each frequency after TDoA permutation can be extracted by evaluating the following distance measure,

$$D(f) = \sum_{j=1}^{J} ||\tilde{\mathbf{a}}_{P_f(j)}(f) - \hat{\mathbf{h}}_j(f)||^2 \tag{7}$$

and sorting $D(f)$ in ascending order. Choosing the k_R first frequencies, denoted by set $\mathcal{F}_R$, will serve as a reference having the lowest distance to the estimated propagation model $\hat{\mathbf{h}}$. The frequency indices to be corrected by the proposed method are chosen by taking indices $k_Q, \ldots, F$ from the sorted $D(f)$, denoted by set $\mathcal{F}_Q$. These have the most incoherent TDoA and amplitude difference with respect to the estimated propagation model. Note that $\mathcal{F}_R$ and $\mathcal{F}_Q$ can have overlapping frequencies if $k_Q < k_R$.

The confidence measure (7) assumes that the estimation of the propagation model $\hat{\mathbf{h}}_j(f)$ has converged close to the actual spatial position in terms of TDoA and that the anechoic source propagation assumption holds for the observed data. It is shown by an example in Section 4 that the lowest frequencies fit to the model (5) more accurately whereas the ICs at higher frequencies suffer from the spatial aliasing and phase modification by reverberation making the permutation uncertain according to (7).

The permutation matrix $\mathbf{P}(f)$ obtained from the TDoA based alignment is used to change the ordering of rows of vector $\mathbf{y}(f,n)$ to correspond to a single source signal defined as $\hat{\mathbf{y}}(f,n) = \mathbf{P}(f)\mathbf{y}(f,n)$. The magnitude spectrogram matrix of the sources after initial permutation is denoted as $[\hat{\mathbf{Y}}_{(j)}]_{f,n} = |\hat{y}_j(f,n)|$.

The permutation correction algorithm is described as follows. For each source $j = 1...J$ we apply the SVD to the magnitude spectrogram of the sources given as,

$$\hat{\mathbf{Y}}_{(j)} = \mathbf{U}_{(j)}\mathbf{\Sigma}_{(j)}\mathbf{V}^*_{(j)}, \quad f \in \mathcal{F}_R \tag{8}$$

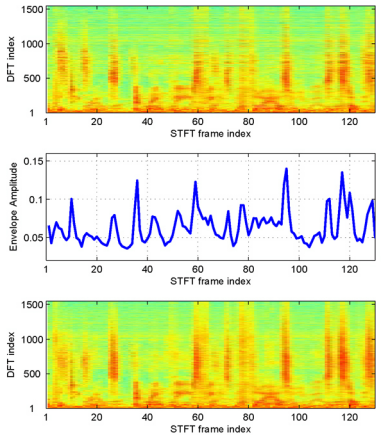


Fig. 2. An example of the SVD analyzed source envelope in the middle, the source magnitude spectrogram after baseline [4] on the top and the source magnitude spectrogram after proposed permutation alignment in the bottom.

where subindices (j) denote the matrix indexing corresponding to each source and each $\hat{\mathbf{Y}}_{(j)}$ is of size $k_R \times N$. To obtain a rank-one approximation of the source magnitude spectrogram we take the singular vectors $\mathbf{U}_{(j)i,:}$ and $\mathbf{V}_{(j)i,:}$ corresponding to the largest singular value $\mathbf{\Sigma}_{(j)i,i}$. The singular vector $\mathbf{U}_{(j)i,:}$ contains the average source spectrum and the corresponding temporal activity is given by $\mathbf{V}_{(j)i,:}$ which we propose to use as the reference source envelope.

The analyzed source envelope for each STFT frame n is hereafter denoted by $\mathbf{v}(n) = [v_1, \ldots, v_J]^T = \mathbf{V}_{(j)i,n}$. The SVD analyzed envelope is assumed to capture quintessential temporal activity features of the source and thus can be used as a reference for aligning permutation for frequencies $f \in \mathcal{F}_Q$ by maximizing the cross-correlation of source magnitudes and $\mathbf{v}(n)$. An example of the SVD analyzed envelope and the source magnitude spectrogram before and after the proposed permutation alignment is illustrated in Figure 2.

The permutation optimization with the obtained source envelopes $\mathbf{v}(n)$ can be defined as

$$\mathbf{P}(f) \leftarrow \text{argmax}_{\mathbf{P}(f)} \sum_{n=1}^{N} \mathbf{v}(n)^T \mathbf{P}(f)\hat{\mathbf{y}}(f,n), \quad \forall f \in \mathcal{F}_Q \tag{9}$$

which equals finding a new permutation matrix $\mathbf{P}(f)$ which maximizes the cross-correlation of $\mathbf{v}(n)$ and source magnitude envelopes $\mathbf{P}(f)\hat{\mathbf{y}}(f,n)$ within the frequency set $f \in \mathcal{F}_Q$. In practice the maximization is implemented by searching through all combinations of $\mathbf{P}(f) : \{1, \ldots J\} \rightarrow \{1, \ldots J\}$ and choosing the one producing largest cross-correlation, this is computationally feasible for low number of sources. As a result we obtain a new permutation matrix $\mathbf{P}(f)$ which is used for aligning the permutations as

$$\hat{\mathbf{y}}(f,n) \leftarrow \mathbf{P}(f)\hat{\mathbf{y}}(f,n) \tag{10}$$

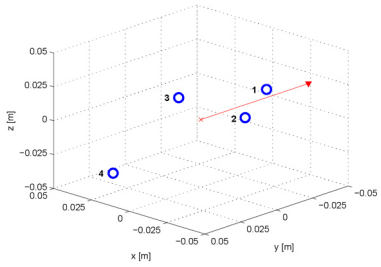


Fig. 3. The capturing array used in the simulations. Microphones are denoted by circles and the zero angle references axis by an arrow.

The experiments with the algorithm has shown that choosing $k_Q < k_R$ produces generally best results regarding the separation quality. In this case the evaluation of permutation optimization given in Equation (9) may also change permutation for frequency indices $f \in F_R$ which further affects on SVD analysis (8). The proposed algorithm is implemented by iteratively evaluating Equations (8) - (10) until no permutation changes are made in (10). With a suitable choice of k_R and k_Q the algorithm usually converges in less than 10 iterations. The choice of k_R and k_Q is discussed in more details in Section 4.

4. EVALUATION OF SEPARATION QUALITY

In this section we evaluate the separation performance of the proposed algorithm against the TDoA based algorithm proposed in [4]. The evaluation consist of real audio captures recorded in following conditions: sampling frequency was 48kHz, the room dimensions were 4.53 × 3.96 × 2.59 m and the reverberation time (T60) was approximately 0.26s.

The capturing array consists of four DPA 4060-BM prepolarized omnidirectional miniature condensator microphones.The array dimensions are given in Table 1 and the array geometry with reference axis is illustrated in Figure 3. The spatial aliasing frequency for the given array is 1563 Hz which corresponds to STFT frequency bin $f = 133$.

The test samples used included three male and one female speakers from Librivox audiobook database which were played with Genelec 1029A speakers. The utterance length is 10 seconds. Each speaker was captured separately and signals were combined into mixtures of three simultaneous speakers. The angle of the speakers with respect the reference axis of the microphone array are given in Table 2.

4.1. Implementation Considerations

For the ICA parameter estimation we used the complex-valued version of JADE algorithm [10]. Other parameters were chosen as follows: STFT window length = 4096 with 50% window overlap, number of target sources = 3. Two

Mic	x (mm)	y (mm)	z (mm)
1	0	-46	6
2	-22	-8	6
3	22	-8	6
4	0	61	-18

Table 1. Geometry of the array used for evaluation. Illustrated in Figure 3.

Identification	Angle
Speaker 1	180°
Speaker 2	90°
Speaker 3	45°
Speaker 4	0°

Table 2. Speaker positions with respect to array zero angle axis.

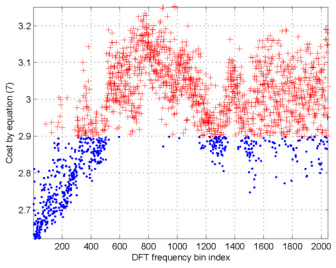


Fig. 4. Cost function (7) for an individual test sample. The reference frequencies $f \in F_R$ are denoted by dots and rest of the cost function entries are denoted by plus-marks.

datasets were used, dataset one consisting of speakers 1, 2 and 4 and dataset two consisting of speakers 2, 3 and 4. The total number of 10-second utterances in both datasets is five. It is shown in Section 4.2 that no separate training stage or development set is needed for the choice of k_R and k_Q due the separation quality not being affected by a wide range of choosing k_R and k_Q. The values for the separation evaluation were chosen as $k_R = 600$ and $k_Q = 300$ producing a good average performance.

An example of the TDoA coherence cost defined by Equation (7) is illustrated in Figure 4. The reference bins chosen are denoted by dots and the rest of the frequencies are denoted by plus-marks. It is clear from the shape of the cost function that the lowest frequencies have the most coherent TDoA regarding the estimated propagation model and are chosen mostly for the reference frequency group $f \in F_R$. Also some higher frequencies fit the model well and serve as a reference.

4.2. Separation Results

The results from the separation quality evaluation using metrics signal-to-distortion ratio (SDR), image-to-spatial distortion Ratio (ISR), signal-to-interference ratio (SIR) and signal-to-artefact ratio (SAR) proposed in [5] are given in Table 3. The measures are averaged over all sources and all utterances. With the proposed method the SDR separation quality increases by 0.72 dB and 0.48 dB in the datasets one and two, respectively. The source interference (SIR) is improved noticeably, the separated source spatial image accuracy (ISR) improves as well and the separation artifacts are decreased

Dataset	Baseline [4]		Proposed	
	1	2	1	2
SDR (dB)	3.88	0.92	4.60	1.40
ISR (dB)	8.92	5.04	10.19	5.63
SIR (dB)	8.37	1.42	9.64	2.93
SAR (dB)	6.24	3.85	6.82	4.43

Table 3. Separation results for datasets one and two.

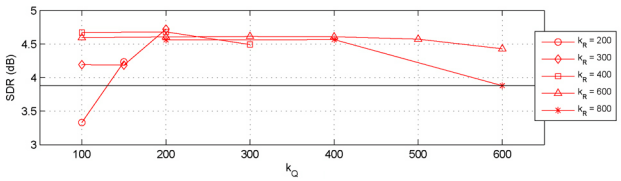


Fig. 5. Separation quality in terms of SDR with different combinations of k_R and k_Q with the dataset one. Solid line denotes the baseline performance.

(SAR).

Each source in dataset one are spatially separated at least by 90° whereas in the dataset two the spatial separation is 45°, which significantly decreases the separation performance. In case of dataset two where the initial separability of the sources is poor the proposed algorithm is still able to improve the average separation of sources, considering the fact that the derivation of the algorithm assumes obtaining a fair initial separation for the envelope analysis.

The effect of the algorithm parameters k_R and k_Q is illustrated in Figure 5 where the SDR separation performance is given with different combinations of k_R and k_Q. The performance of the proposed algorithm is almost equivalent regardless of the choice of the parameters. Only too few reference frequencies $k_R = 200$ degrades the SDR quality below the baseline performance. The results in Figure 5 indicate high robustness towards the choice of the parameters and eliminates the need of a separate training stage.

Temporal activity based permutation alignment algorithms are known to be less efficient with short signals and thus the proposed method was additionally tested with the signals from the test set one split to duration of 2.5 seconds. The average SDR was 2.82 dB and 3.15 dB for the baseline and the proposed algorithm, respectively, indicating improved separation with the proposed method also in such cases.

5. CONCLUSION

In this paper we proposed an algorithm for independent component analysis (ICA) permutation alignment when used for blind source separation (BSS) of simultaneous speakers. The proposed method is based on analysis of source envelopes by rank-one SVD and maximizing the cross-correlations of the analyzed envelope and source magnitude envelopes at each individual frequency. The proposed method is aimed for improving the time difference of arrival (TDoA) based alignment algorithms suffering from spatial aliasing in case of high sampling frequency speech and it was found to improve the separation quality in such conditions.

6. REFERENCES

[1] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1, pp. 21–34, 1998.

[2] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proc. of ICA*. Helsinki, Finland, 2000, pp. 215–220.

[3] M.Z. Ikram and D.R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *Proc. of ICASSP*, 2002, pp. 881–884.

[4] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Trans. on ASLP*, vol. 15, no. 5, pp. 1592–1604, 2007.

[5] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[6] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 5, pp. 530–538, 2004.

[7] I. Lee, T. Kim, and T.W. Lee, "Independent vector analysis for convolutive blind speech separation," *Blind speech separation*, pp. 169–192, 2007.

[8] F. Nesta, P. Svaizer, and M. Omologo, "Convolutive bss of short mixtures by ica recursively regularized across frequencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 624–639, 2011.

[9] F. Nesta, T.S. Wada, and B.H. Juang, "Coherent spectral estimation for a robust solution of the permutation problem," in *in Proc. of WASPAA*, 2009, pp. 105–108.

[10] J.F. Cardoso and A. Souloumiac, "Blind beamforming for non-gaussian signals," *IEE Proc-F. Radar and Signal Processing*, vol. 140, no. 6, pp. 362–370, 1993.

Publication P5

J. Nikunen, T. Virtanen and M. Vilermo, "Multichannel Audio Upmixing by Time-Frequency Filtering Using Non-Negative Tensor Factorization," in *Journal of the Audio Engineering Society*, Vol. 60, No. 10, pp. 794–806, 2012.

Publication P6

J. Nikunen and T. Virtanen, "Direction of Arrival Based Spatial Covariance Model for Blind Sound Source Separation," in *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 3, pp. 727–739, 2014.

Direction of Arrival Based Spatial Covariance Model for Blind Sound Source Separation

Joonas Nikunen, *non-Member* and Tuomas Virtanen, *Member, IEEE*

***Abstract*—This paper addresses the problem of sound source separation from a multichannel microphone array capture via estimation of source spatial covariance matrix (SCM) of a short-time Fourier transformed mixture signal. In many conventional audio separation algorithms the source mixing parameter estimation is done separately for each frequency thus making them prone to errors and leading to suboptimal source estimates. In this paper we propose a SCM model which consists of a weighted sum of direction of arrival (DoA) kernels and estimate only the weights dependent on the source directions. In the proposed algorithm, the spatial properties of the sources become jointly optimized over all frequencies, leading to more coherent source estimates and mitigating the effect of spatial aliasing at high frequencies. The proposed SCM model is combined with a linear model for magnitudes and the parameter estimation is formulated in a complex-valued non-negative matrix factorization (CNMF) framework. Simulations consist of recordings done with a hand-held device sized array having multiple microphones embedded inside the device casing. Separation quality of the proposed algorithm is shown to exceed the performance of existing state of the art separation methods with two sources when evaluated by objective separation quality metrics.**

***Index Terms*—multichannel source separation, spatial covariance models, non-negative matrix factorization, direction of arrival estimation, array signal processing**

I. Introduction

WHEN recording an auditory scene using one or multiple microphones, it is preferred that the sound source dependent information can be separated for the uses of great variety of subsequent audio processing tasks. The examples of such applications include spatial audio coding (SAC) [1], [2], 3D sound analysis and synthesis [3], and signal enhancement for various purposes, such as automatic speech recognition (ASR) [4], [5]. When no prior information of the sources involved in the capture is available, the process is called blind source separation (BSS). The BSS problem in the case of spatial audio captures consist of decomposing the multichannel mixture signal into source signals and representing information about their spatial position or response from their originating location to each receiving microphone.

A well known BSS approach is the independent component analysis (ICA) [6] applied separately at each frequency of a short-time Fourier transformed (STFT) array input. It leads to an arbitrary frequency-wise source ordering referred to as the permutation problem. Source permutation is usually solved based on time difference of arrival (TDoA) interpretation of ICA mixing parameters [7]–[9]. The TDoAs calculated from phase differences become ambiguous when the frequency exceeds the spatial aliasing limit, which corresponds to a wavelength greater than half of the microphone spacing. As a result, the TDoAs cannot be directly utilized in solving the permutation problem for high frequencies. Additionally, the ICA parameters for a single source concatenated over frequency do not explicitly have a connection to the spatial position of a source but only to the phase difference caused by it. Separation methods directly utilizing TDoAs between microphones and creating time-frequency separation masks by clustering the measured TDoAs at each frequency include for example DUET [10] and binwise clustering [11].

More recently, methods based on finding spectrally redundant parts by non-negative matrix factorization (NMF) [12]–[14] have been proposed for separation of sound sources both with single [15], [16] and multichannel mixtures [17]–[20]. NMF is applied in the magnitude spectrogram domain and it finds an approximation of the mixture spectrogram using a linear combination of components that have a fixed spectrum and time-dependent gain. In the NMF separation framework the spatial properties of the sources can be modeled using a spatial covariance matrix (SCM) for each source at each STFT frequency bin [18]–[22]. Such extensions are hereafter referred to as complex-valued NMF (CNMF). The SCM denotes the mixing of the sources by magnitude and phase differences between the recorded channels, and is not dependent on the absolute phase of the source signal. Additionally, non-negative tensor factorization with spatial cues based on the magnitude panning of sources have been proposed in [23].

The CNMF algorithms [19]–[21] estimate unconstrained SCMs at each frequency, thus relying on the ability of the NMF magnitude model to separate repetitive parts that correspond to sources at a single spatial location. In the case of spectrally similar sources, for example two speakers, a single NMF component and the corresponding SCM can end up representing both the sources at different spatial locations. In such case, the estimated parameters cannot provide separation of the two sources. Spatial aliasing makes the algorithm prone to SCM estimation errors and separation is dependent on the magnitude model separation abilities. The CNMF method proposed in [18] assigns a fixed number of NMF components per source but it is reported to have a poor separation quality without an oracle initialization of the source parameters.

A spatial signal processing field of beamforming [24] can also be considered as a separation technique. The simplest design is the delay and sum beamformer (DSBF), which consists of time aligning and summing the microphone signals.

J. Nikunen and T. Virtanen are with the Department of Signal Processing, Tampere University of Technology, Tampere, Finland, email: firstname.lastname@tut.fi

The alignment correspond to a time delay caused by the target DoA, i.e. beamformer look direction, and the sources originating from this direction become enhanced. Other types of beamformers, such as the minimum variance distortionless (MVDR) beamformer, aim at suppressing and canceling interfering signals originating from other than the beamformer look direction. More recent advances in beamforming are based on adaptively estimating the noise characteristics and designing blocking matrix for the general sidelobe canceller structure [25], [26].

Beamforming methods assume that the geometry of the array is known and require a high number of microphones to work efficiently, to form a narrow beam that is useful for source separation. This is conventionally only achieved with physically large arrays. Beamformers also suffer from spatial aliasing which causes signal amplification from undesired source directions, i.e. peaks in the sidelobe structure. An emerging approach for beamforming based source separation is spherical array beamforming [27], [28] where a high number of densely spaced sensors in a physically small sphere is used to obtain uniform directivity properties towards every look direction.

In this paper we propose a novel BSS method that combines SCM estimation by beamforming-inspired DoA kernels and object-based signal analysis using NMF. We propose to model the source SCMs as the weighted combination of DoA kernels and the magnitudes of sources by the NMF. The DoA kernels represent the phase difference between array channels caused by a single source TDoA at a certain spatial position. The main benefit of the method comes from making the connection between the SCMs at each frequency by representing SCMs over frequency as a weighted sum of DoA kernels thus avoiding source and frequency ambiguity issues. By only estimating direction-dependent weights for the DoA kernels of each source, the proposed SCM model is jointly optimized over all frequencies, and produces better estimates of the source SCMs. The direction weights estimation also mitigates the effect of spatial aliasing in high frequencies due the estimation algorithm taking into account phase difference evidence across frequency by single time delays of individual DoA kernels.

The NMF components represent parts of the sources by estimating repetitive magnitude structures from the mixture signal. Components sharing the same spatial position are assumed to originate from the same source and can be thus linked together by simple clustering applied on the estimated direction weights. In addition to doing separation of sources, the method produces a parameterization of their spatial properties, and can therefore be used in 3D sound synthesis of the recorded mixture.

We evaluate the separation quality of the proposed method against the reference CNMF approach [19] and frequency-domain ICA [8], [29]. The simulations are done using a small microphone array consisting of four microphones enclosed in a casing similar in size to a hand-held mobile device. The evaluation is based on objective separation quality metrics proposed in [30], [31] and perceptually motivated metrics [32]. The proposed method is shown to produce considerably better separation quality over the conventional methods.

The rest of the article is organized as follows. In Section II we present the background of spatial audio processing for sound source separation. The general principle of the proposed SCM model as a superposition of DoA kernels is presented in Section III. Formulation of the proposed SCM model into the CNMF framework and update rules for the model parameter optimization are presented in Section IV. The source reconstruction based on clustering of DoA kernel weights is presented in Section V. The simulations and separation evaluation are given in Section VI. In Section VII we discuss future work for improving the proposed SCM model and 3D sound synthesis using the proposed spatial audio parameterization.

II. Background

In this section we define the problem of the sound source separation with spatial audio captures and present the spatial processing background for the proposed SCM model and CNMF algorithm it is used with. The section consist of stating the source mixing model in Section II-A, definition of the signal representation and the spatial covariance matrices, in Section II-B, and interpretation of the convolutive mixing model in the spatial covariance domain in Section II-C.

A. Source Mixing Model

In time domain an array capture consists of a mixture of sound sources convolved with their spatial responses. The mixing model can be described as

$$\tilde{x}_m(t) = \sum_{k=1}^{K} \sum_{\tau} h_{mk}(\tau) s_k(t-\tau) \tag{1}$$

where the mixture $\tilde{x}_m(t)$ consists of $k = 1...K$ sources captured by microphones $m = 1...M$, and the time-domain sample index is denoted by t. The spatial response from source k to microphone m is represented by a mixing filter $h_{mk}(\tau)$ and the single-channel source signals are denoted by $s_k(t)$.

The convolutive mixing model (1) can be approximated in the STFT domain by instantaneous mixing at each frequency bin as

$$\mathbf{x}_{il} \approx \sum_{k=1}^{K} \mathbf{h}_{ik} s_{ilk} = \sum_{k=1}^{K} \mathbf{y}_{ilk} \tag{2}$$

where $\mathbf{x}_{il} = [x_{il1}, \ldots, x_{ilM}]^T$ is the STFT of the capture $\tilde{x}_m(t)$ with analysis window length of $N = 2I-1$, the positive DFT bin frequencies are denoted by $i = 1...I$ and the STFT frame index by $l = 1...L$. The frequency-domain mixing filter is denoted at each frequency bin by $\mathbf{h}_{ik} = [h_{ik1}, \ldots, h_{ikM}]^T$ and the STFTs of the sources are denoted by s_{ilk}. The spatial images of the sources are denoted as $\mathbf{y}_{ilk} = \mathbf{h}_{ik} s_{ilk}$, which are the source signals as seen by the array, i.e. convolved with their spatial impulse responses. The effective length of the mixing filter $h_{mk}(\tau)$ can be several hundreds of milliseconds but its approximation in frequency domain with an analysis window length of tens of milliseconds works well in practice due to the negligible energy after the main reverberant part of the source spatial response.

B. Signal Representation

The proposed method uses SCMs calculated at each time-frequency point as the signal representation. Spatial covariance calculation translates the absolute phase of the mixture to a phase difference between each microphone pair. In the CNMF work by Sawada et. al [19] it was proposed that for the calculation of the SCMs a magnitude square-rooted version of the array capture is used. This ensures that the nonnegative part in the diagonal of the SCM, modeled by the NMF, contains the magnitude spectrum of the mixture and the individual source spectra are approximately additive.

The magnitude square-rooted version $\hat{\mathbf{x}}_{il}$ of the capture $\mathbf{x}_{il} = [x_{il1}, \dots x_{ilM}]^T$ for a time-frequency point (i, l) is obtained as

$$\hat{\mathbf{x}}_{il} = [|x_{il1}|^{1/2} \operatorname{sign}(x_{il1}), \dots, |x_{ilM}|^{1/2} \operatorname{sign}(x_{ilM})]^T \quad (3)$$

where $\operatorname{sign}(z) = z/|z|$ is the signum function for complex numbers. The SCM for a single time-frequency point is obtained from the array capture vector $\hat{\mathbf{x}}_{il} = [\hat{x}_{il1}, \dots, \hat{x}_{ilM}]^T$ as outer product

$$\mathbf{X}_{il} = \hat{\mathbf{x}}_{il} \hat{\mathbf{x}}_{il}^H, \quad (4)$$

where ^{H} stands for Hermitian transpose. Matrices $\mathbf{X}_{il} \in \mathbb{C}^{M \times M}$ for each time frequency point (i, l) point consist of observation magnitude $|\mathbf{x}_{il}| = [|x_{il1}|, \dots, |x_{ilM}|]^T$ in its diagonal $[\mathbf{X}_{il}]_{nn}$, and off-diagonal values $[\mathbf{X}_{il}]_{nm}, n \neq m$ represent the magnitude correlation and phase difference $|x_{iln} x_{ilm}|^{1/2} \operatorname{sign}(x_{iln} x_{ilm}^*)$ between each microphone pair (n, m).

C. Convolutive Mixing Model in Spatial Covariance Domain

The convolutive mixing model defined in Equation (2) can be expressed in the SCM domain by replacing each term by its covariance counterpart. The SCM domain mixing is expressed as

$$\mathbf{X}_{il} \approx \sum_{k=1}^{K} \mathbf{H}_{ik} \hat{s}_{ilk} = \sum_{k=1}^{K} \mathbf{S}_{ilk}, \quad (5)$$

where $\mathbf{H}_{ik}$ is the spatial covariance matrix for each source at each frequency and $\hat{s}_{ilk}$ is the corresponding source magnitude spectrum.

The matrix $\mathbf{H}_{ik} \in \mathbb{C}^{M \times M}$ denotes the source spatial response $\mathbf{h}_{ik}$ expressed in the form of covariance matrix $\mathbf{h}_{ik} \mathbf{h}_{ik}^H$. The complex-valued monoaural source spectrogram s_{ilk} in the SCM domain results to a real-valued power spectrum $s_{ilk} \overline{s_{ilk}}$. Due to the square-rooted STFT used to calculate the observed SCMs, we denote the sources using their magnitude spectra $\hat{s}_{ilk} = (s_{ilk} \overline{s_{ilk}})^{1/2}$. We can approximate the SCMs being additive since the sources are approximately uncorrelated but also sparse, meaning that only a single source is to be active within each time-frequency point [33]. When using the SCM domain representation defined by Equations (3) - (5), the absolute phase of the sources is not significant from the parameter estimation point of view, and we only model the phase differences between all microphone pairs.

Estimating the source magnitudes $\hat{s}_{ilk}$ and the corresponding SCMs denoted by $\mathbf{H}_{ik}$ by turn would provide the desired BSS properties. However estimating $\mathbf{H}_{ik}$ jointly over all frequencies requires a model that ties together the phase difference over frequencies, which is a difficult constraint to be included in the parameter estimation. For the CNMF-based source separation the SCM estimation proposed in [19] relies on the NMF model to enforce magnitudes $\hat{s}_{ilk}$ to correspond to a single source, which is assumed to yield an estimate of $\mathbf{H}_{ik}$ associated to a single source. A direct estimation of the source SCM and variances at each frequency is done for example in [34], but it again requires solving the frequency-wise permutation. The covariance estimation strategy from [34] with NMF as a source magnitude model has been proposed in [22], thus avoiding permutation ambiguity. However, in both cases the spatial properties estimated separately for each STFT frequency bin i do not utilize the fact that the SCM properties are connected by the TDoA of the direct path and early reflections.

III. Proposed Spatial Covariance Matrix Model by Superposition of DoA Kernels

In the case of direct path propagation or anechoic conditions, the source direction with respect to a receiving array corresponds to a specific time delay between the microphones. In beamforming, the TDoA defined by the look direction of a beamformer is used to align the received microphone signals in order to enhance sources originating from the look direction. A single TDoA determines the desired phase difference over frequencies, making the beamformer implemented in the frequency domain able to integrate source evidence over the whole frequency spectrum. Such a concept has not been widely utilized in BSS since it is difficult to include it to the parameter estimation, and the spatial aliasing makes the delays unambiguous at high frequencies.

The proposed DoA-based SCM model can be used to unify the STFT bin dependencies when estimating the source spatial responses, and to avoid optimizing the model parameters individually for each frequency. The difference of the proposed source separation algorithm to beamforming is that the CNMF optimization algorithm is set to fit a collection of predefined DoA kernels (beamforming kernels) to the observed data and that way to find the most likely DoA of the source in question. By defining the SCM model using only direction-dependent parameters, we can utilize the time delay dependency of the spatial covariance values across frequencies in a CNMF algorithm framework for estimating the source magnitude and spatial properties.

A. Time-difference of Arrival

A specific wavefront-arrival direction corresponds to a set of TDoA values between each microphone pair. The TDoA values depend on the geometry of the array and the relationship is shortly explained in the following. We first consider the array illustrated in Figure 1, where one pair of microphones n and m lie on the xy-plane at locations $\mathbf{n}$ and $\mathbf{m}$, respectively. A unit vector $\mathbf{k}_o$ is pointing towards the look direction from the geometrical center $\mathbf{p}$ of the array. For simplicity, we define that the geometrical center of the array is in the origin of the Cartesian coordinate system, i.e. $\mathbf{p} = [0, 0, 0]^T$. The look

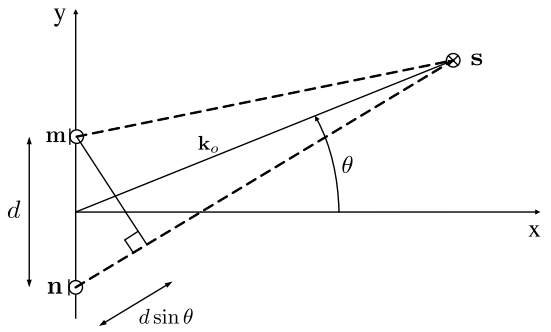


Fig. 1. Example array geometry consisting of two microphones m and n as seen from above, source $\mathbf{s}$ azimuth angle given as θ.

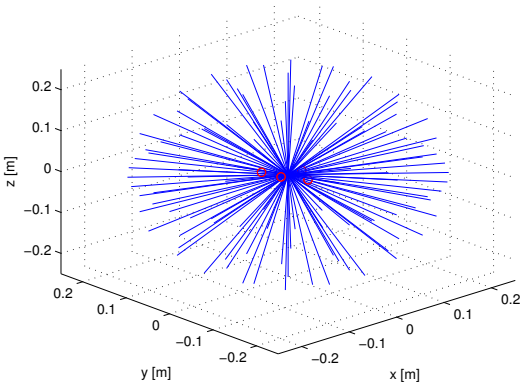


Fig. 2. Look direction vectors approximating uniform sampling of the unit sphere around the array.

directions can be denoted in the spherical coordinate system using elevation $\theta \in [0, \pi]$, azimuth $\varphi \in [0, 2\pi]$ and fixed radius of $r = 1$. We define ranges of $-90° \leq \theta \leq 90°$ and $0° \leq \varphi \leq 360°$ for elevation and azimuth, respectively.

Assuming the far field model, i.e. the wavefront being planar when arriving to the array, we can write the TDoA of the microphone n with respect to array center point $\mathbf{p}$ in seconds as

$$\tau_n(\mathbf{k}_o) = \frac{-\mathbf{k}_o^T(\mathbf{n} - \mathbf{p})}{v} = \frac{-\mathbf{k}_o^T\mathbf{n}}{v} \quad (6)$$

where v is the speed of sound. Each look direction $o = 1 \dots O$ translates to a TDoA for each microphone, which further translates into a phase difference linearly proportional to the frequency in the STFT domain. The TDoA in Equation (6) equals to frequency-domain phase difference of $-j2\pi f \tau_n(\mathbf{k}_o)$, where f is the frequency in Hertz. The phase difference is unambiguous only up to the spatial aliasing frequency $f = \frac{v}{2d}$, where d is the smallest distance between any two microphones in the array.

We define TDoA between a microphone pair (n, m) as $\tau_{nm}(\mathbf{k}_o) = \tau_n(\mathbf{k}_o) - \tau_m(\mathbf{k}_o)$. The phase differences corresponding to the TDoA $\tau_{nm}(\mathbf{k}_o)$ between every microphone pair $n = 1 \dots M$ and $m = 1 \dots M$ are represented as a matrix $\mathbf{W}_{io} \in \mathbb{C}^{M \times M}$ for each each STFT frequency index $i = 1 \dots I$ and each look direction $o = 1 \dots O$. We define these to be DoA kernel matrices which are obtained as

$$[\mathbf{W}_{io}]_{nm} = \exp\left(j2\pi f_i \tau_{nm}(\mathbf{k}_o)\right), \quad f_i = (i-1)F_s/N \quad (7)$$

where the F_s denotes sampling frequency and N is the STFT length.

B. Superposition of DoA Kernels

Assuming a point source and an anechoic capturing condition, a single DoA kernel would be enough to describe the SCM of a source. However, because of echoes and diffractions from surfaces and objects, a more complex model is needed. For SCM modeling, we propose to use a weighted linear combination of DoA kernels that uniformly sample a surface of the unit sphere around the receiving array. The gain of each DoA kernel describe the signal power emanating from each sampled look direction around the array.

We define a set of fixed look directions vectors $\mathbf{k}_o$ that spatially sample the surface of a unit sphere set around the geometrical center $\mathbf{p}$ of the array. An example set of the look direction vectors is illustrated in Figure 2. DoA kernels for each look direction $o = 1 \dots O$ at each frequency $i = 1 \dots I$ are denoted using $\mathbf{W}_{io} \in \mathbb{C}^{M \times M}$ and are calculated according to Equation (7). Entries of kernel matrices $[\mathbf{W}_{io}]_{nm}$ denote a TDoA in terms of phase difference expressed as a complex number for a microphone pair (n, m).

In Section II-C the source spatial image was defined as $\mathbf{S}_{ilk} = \mathbf{H}_{ik}\hat{s}_{ilk}$, consisting of the magnitudes $\hat{s}_{ilk}$ and the mixing defined by the source SCM $\mathbf{H}_{ik}$. The proposed SCM model equals the weighted superposition of multiple DoA kernels and is given as

$$\mathbf{H}_{ik} = \sum_{o=1}^{O} \mathbf{W}_{io} z_{ko}, \quad (8)$$

where z_{ko} are the direction weights corresponding to the DoA kernels into each look direction.

We want to estimate $\mathbf{H}_{ik}$ in such a way that it corresponds to a single acoustical source over all the STFT frequencies. This is directly achieved in the proposed SCM model by estimating the spatial weights z_{ko} which are independent of frequency. The definition of the DoA kernels in Equation (7) directly takes into account the frequency dependencies that a certain source DoA causes through a single TDoA. The spatial weights z_{ko} are restricted to be non-negative and they can be estimated in the CNMF framework with a corresponding magnitude model for $\hat{s}_{ilk}$ as will be shown in Section IV. An example of the estimated SCM model weights z_{ko} for three sources are illustrated in Figure 3. The illustration depicts the weighted look direction vectors denoted by $z_{ko}\mathbf{k}_o$ and the result is projected on to the xy-plane. The experimental conditions for obtaining Figure 3 is described in Section VI-A.

IV. Complex-valued Non-negative Matrix Factorization with the Proposed SCM Model

In this section we present a BSS algorithm that combines an NMF-based source magnitude model and the DoA kernel based SCM model which together produce a complex valued NMF (CNMF) model. The proposed BSS algorithm is able to

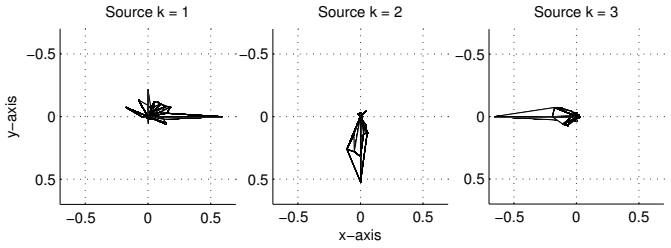


Fig. 3. Illustration of the weighted look direction vectors $z_{ko}\mathbf{k}_o$ of the estimated SCM model projected on to the xy-plane. Sources are at 0, 90 and 180 degrees in azimuth, pictured above the array, and azimuth increasing counterclockwise.

jointly estimate source SCMs across frequencies by using the proposed SCM model and its parameterization of source spatial properties by the direction weights, which are independent of frequency.

The block diagram of the proposed algorithm is given in Figure 4. First the STFT is calculated from the time domain microphone array input and the SCM of each time-frequency point is calculated as defined in Section II-B. The SCM representation serves as an input for the CNMF algorithm. Prior to model parameter estimation a set of DoA kernels with fixed look directions are constructed as defined in Section III-A. The DoA kernels are set to sample the spatial space approximately uniformly around the array. The CNMF algorithm with the proposed DoA-based SCM model is applied to estimate the source parameters, i.e. magnitude spectra and DoA kernel direction weights. In the separation stage the sources are reconstructed from the mixture signal by clustering the components obtained by the CNMF to construct a magnitude mask which is used for obtaining Wiener filter estimates of the source spatial images.

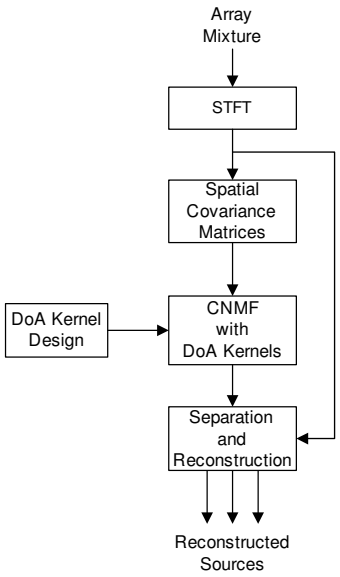


Fig. 4. Block diagram of the proposed BSS system.

A. CNMF Model for SCM Observations

The proposed spatial model consist of a NMF magnitude model [13], [14] for the source magnitude spectra denoted by $\hat{s}_{ilk}$ and the DoA kernel based SCM for denoting the spatial position of the source. In practice, several NMF components are used for representing one actual acoustic sound source, but for the algorithm derivation we define that one NMF component represents one sound source. Later in source signal reconstruction we will cluster the NMF components based on their estimated direction weights z_{ko}.

The model for SCM observations is obtained by replacing $\mathbf{H}_{ik}$ in Equation (5) by the proposed SCM model defined in Equation (8) to get

$$\mathbf{X}_{il} \approx \hat{\mathbf{X}}_{il} = \sum_{k=1}^{K} \mathbf{H}_{ik}\hat{s}_{ilk} = \sum_{k=1}^{K}\sum_{o=1}^{O} \mathbf{W}_{io} z_{ko}\hat{s}_{ilk}. \quad (9)$$

Source magnitudes $\hat{s}_{ilk}$ are to be obtained by the NMF estimation framework. A rank-1 NMF model for the magnitude spectrogram of a single source k is defined as

$$\hat{s}_{ilk} = t_{ik}v_{kl}, \quad t_{ik}, v_{kl} \geq 0, \quad (10)$$

where column vector $t_{:k}$ contains the spectrum of the source, and the corresponding row $v_{k:}$ represents its gain in each STFT frame. The NMF magnitude model with a fixed source spectrum is extremely simplified and can only model parts of real acoustic sources, but serves as an intermediate representation for the spatial parameter estimation.

Substituting the NMF model (10) into the SCM model (9) and rearranging the parameters gives us the whole CNMF model

$$\mathbf{X}_{il} \approx \hat{\mathbf{X}}_{il} = \sum_{k=1}^{K}\sum_{o=1}^{O} \mathbf{W}_{io} z_{ko} t_{ik} v_{kl}. \quad (11)$$

Additionally, the CNMF model can be given using the source SCMs $\mathbf{H}_{ik}$ which equals to the model

$$\hat{\mathbf{X}}_{il} = \sum_{k=1}^{K} \mathbf{H}_{ik} t_{ik} v_{kl}. \quad (12)$$

Comparing the models defined in Equations (11) and (12), we observe that the real-valued entries in the diagonal of $\mathbf{H}_{ik}$ are responsible for modeling the absolute source magnitude level with respect to each channel, and the off-diagonal values model the cross-channel magnitude and phase difference properties. This further means that the magnitudes $|\mathbf{W}_{io}|$ combined with the non-negative weights z_{ko} determine the magnitude difference between the channels.

The DoA kernels generated using Equation (7) have unit magnitudes and for modeling the magnitude differences between each channel, the algorithm needs to estimate and update magnitudes of $\mathbf{W}_{io}$ accordingly. This is due to the fact that the sources have gain differences with respect to each microphone. The gain differences are caused by the microphones being at different distance from the source and the possible acoustical shade of the array casing which produces direction-dependent gain even if omnidirectional microphones are used. While the SCM magnitudes are subject to updating, we keep the original DoA kernel phase difference the same, i.e. the

original time delay caused by certain direction of the source. In this way we retain the frequency dependency when modeling the phase difference by estimating the frequency independent spatial weights z_{ko}.

B. The CNMF Algorithm

NMF algorithms typically use multiplicative updates iteratively in order to minimize a given cost function, for example the squared Euclidean distance or the Kullback-Leibler divergence [12]. In this paper we present a method for obtaining the algorithm updates via auxiliary functions and EM-algorithm structure similarly as presented in [19].

1) CNMF Cost Function: We aim to minimize the squared Frobenius norm between the observed $\mathbf{X}_{il}$ and the model $\hat{\mathbf{X}}_{il}$ summed over frequency and time indices, which is defined as

$$\sum_{i=1}^{I}\sum_{l=1}^{L}||\mathbf{X}_{il}-\hat{\mathbf{X}}_{il}||_F^2. \tag{13}$$

In [19] the statistical interpretation of the CNMF model error (13) is shown to be equivalent to the negative log-likelihood (up to terms independent of the model parameters)

$$\mathcal{L}(\mathbf{W},\mathbf{Z},\mathbf{T},\mathbf{V})=\sum_{i=1}^{I}\sum_{l=1}^{L}||\mathbf{X}_{il}-\sum_{k=1}^{K}\sum_{o=1}^{O}\mathbf{W}_{io}z_{ko}t_{ik}v_{kl}||_F^2. \tag{14}$$

We use this result in deriving the algorithm update rules for optimization of the model parameters $\theta=\{\mathbf{W},\mathbf{Z},\mathbf{T},\mathbf{V}\}$. We introduce latent components $\mathbf{C}_{ilko}$ given as

$$\mathbf{C}_{ilko}=\mathbf{W}_{io}z_{ko}t_{ik}v_{kl}+r_{ilko}(\mathbf{X}_{il}-\sum_{k,o}\mathbf{W}_{io}z_{ko}t_{ik}v_{kl}), \tag{15}$$

where

$$r_{ilko}=\frac{z_{ko}t_{ik}v_{kl}}{\hat{x}_{il}},\quad \hat{x}_{il}=\sum_{k,o}z_{ko}t_{ik}v_{kl}. \tag{16}$$

The parameters satisfy $\sum_{k,o}r_{ilko}=1$ and $r_{ilko}>0$. The latent components obey

$$\sum_{k=1}^{K}\sum_{o=1}^{O}\mathbf{C}_{ilko}=\mathbf{X}_{il}. \tag{17}$$

Based on techniques introduced in [19] the negative log-likelihood (14) can be minimized using an auxiliary function incorporating the latent components. The auxiliary function is defined as

$$\mathcal{L}^{+}(\mathbf{W},\mathbf{Z},\mathbf{T},\mathbf{V},\mathbf{C})= \sum_{i=1}^{I}\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{o=1}^{O}\frac{1}{r_{ilko}}||\mathbf{C}_{ilko}-\mathbf{W}_{io}z_{ko}t_{ik}v_{kl}||_F^2. \tag{18}$$

According to [19], the likelihood function (18) can be used for an indirect optimization of (14). This is due to the auxiliary function having the properties

$$\mathcal{L}(\mathbf{W},\mathbf{Z},\mathbf{T},\mathbf{V})\leq\mathcal{L}^{+}(\mathbf{W},\mathbf{Z},\mathbf{T},\mathbf{V},\mathbf{C}) \tag{19}$$

$$\mathcal{L}(\mathbf{W},\mathbf{Z},\mathbf{T},\mathbf{V})=\min_{\mathbf{C}}\mathcal{L}^{+}(\mathbf{W},\mathbf{Z},\mathbf{T},\mathbf{V},\mathbf{C}), \tag{20}$$

which indicate that minimizing $\mathcal{L}^{+}$ with respect to $\mathbf{W},\mathbf{Z},\mathbf{T}$ and $\mathbf{V}$ corresponds to the minimization of $\mathcal{L}$ which yields optimization of the model parameters with respect to (13). Substituting the definition of $\mathbf{C}_{ilko}$ in Equation (15) to Equation (18) makes it equal to original likelihood (14) and allows indirect optimization of the whole model using the auxiliary variables.

2) Algorithm Updates for the Non-negative Parameters: The derivation of the algorithm updates is achieved via partial derivation of (18) with respect to each model parameter and setting the derivative to zero. The derivations are given in Appendix A. For non-negative model parameters z_{ko},t_{ik} and v_{kl}, the following update rules are obtained:

$$z_{ko}\leftarrow z_{ko}\Big[1+\frac{\sum_{i,l}t_{ik}v_{kl}\mathrm{tr}(\mathbf{E}_{il}\mathbf{W}_{io})}{\sum_{i,l}t_{ik}v_{kl}\hat{x}_{il}}\Big] \tag{21}$$

$$t_{ik}\leftarrow t_{ik}\Big[1+\frac{\sum_{l,o}z_{ko}v_{kl}\mathrm{tr}(\mathbf{E}_{il}\mathbf{W}_{io})}{\sum_{l,o}z_{ko}v_{kl}\hat{x}_{il}}\Big] \tag{22}$$

$$v_{kl}\leftarrow v_{kl}\Big[1+\frac{\sum_{i,o}z_{ko}t_{ik}\mathrm{tr}(\mathbf{E}_{il}\mathbf{W}_{io})}{\sum_{i,o}z_{ko}t_{ik}\hat{x}_{il}}\Big]. \tag{23}$$

where $\mathbf{E}_{il}=\mathbf{X}_{il}-\sum_{k,o}\mathbf{W}_{io}z_{ko}t_{ik}v_{kl}$ is the error of the model.

3) Algorithm Updates for the SCM Model Parameters: The optimization of the DoA kernels needs a different update scheme, since we desire to retain the phase differences of the predefined kernels but update the relative magnitude differences. For estimation of the DoA kernel magnitudes we first derive the update for complex $\mathbf{W}_{io}$, but restrict the update to its magnitude.

The update rule for $\mathbf{W}_{io}$ via partial derivation is given in Appendix A and results to multiplicative update

$$\hat{\mathbf{W}}_{io}\leftarrow\mathbf{W}_{io}\Big[\sum_{l,k}z_{ko}t_{ik}v_{kl}\hat{x}_{il}+\sum_{l,k}z_{ko}t_{ik}v_{kl}\mathbf{E}_{il}\Big], \tag{24}$$

where $\hat{\mathbf{W}}_{io}$ is a preliminary update with a modified phase difference compared to the actual desired update of magnitudes of $\mathbf{W}_{io}$.

In particular at the highest frequencies the update (24) may produce matrices that are not positive semidefinite. For example, negative values at the diagonal equal to a subtractive magnitude model even though the model assumes purely additive sources. Based on [19] to enforce positive semidefinite matrices an eigenvalue decomposition $\hat{\mathbf{W}}_{io}=\mathbf{V}\mathbf{D}\mathbf{V}^H$ is applied and eigencomponents with negative eigenvalues are set to zero, denoted as $\hat{\mathbf{D}}$. The positive semidefinite matrices are obtained as

$$\hat{\mathbf{W}}_{io}\leftarrow\mathbf{V}\hat{\mathbf{D}}\mathbf{V}^H. \tag{25}$$

For the final update of actual DoA kernels $\mathbf{W}_{io}$ we apply

$$\mathbf{W}_{io}\leftarrow|\hat{\mathbf{W}}_{io}|\exp(i\arg(\mathbf{W}_{io})), \tag{26}$$

which only updates the magnitude part of the DoA kernels and thus the magnitudes of the SCMs.

4) Parameter Scaling: We constrain the scale of DoA kernels as

$$||\mathbf{W}_{io}||_F = 1, \tag{27}$$

which is achieved by applying

$$\mathbf{W}_{io} \leftarrow \frac{\mathbf{W}_{io}}{||\mathbf{W}_{io}||_F} \tag{28}$$

after evaluating the final stage of the update given in Equation (26). The scaling ensures that the SCM part is only responsible of modeling phase differences and relative magnitude differences between the input channels (diagonal values).

Additionally we introduce following constraints for numerical stability:

$$\sum_{o=1}^{O} z_{ko}^2 = 1, \quad \sum_{l=1}^{L} v_{kl}^2 = 1. \tag{29}$$

The scaling of z_{ko} to unity l^2-norm along DoA kernel direction dimension is compensated by multiplying t_{ik} by the same norm. Similarly, enforcing unity l^2-norm to v_{kl} is compensated by scaling of t_{ik}. The scaling of the model parameters is achieved by applying

$$\hat{a}_k = \big(\sum_{l=1}^{L} v_{kl}^2\big)^{1/2}, \quad v_{kl} \leftarrow \frac{v_{kl}}{\hat{a}_k}, \quad t_{ik} \leftarrow t_{ik}\hat{a}_k \tag{30}$$

$$\hat{b}_k = \big(\sum_{o=1}^{O} z_{ko}^2\big)^{1/2}, \quad z_{ko} \leftarrow \frac{z_{ko}}{\hat{b}_k}, \quad t_{ik} \leftarrow t_{ik}\hat{b}_k, \tag{31}$$

after updates of v_{kl} and z_{ko}, respectively.

5) Algorithm Implementation: The proposed CNMF algorithm consists of the following steps.

1) Initialize z_{ko}, t_{ik} and v_{kl} with random values uniformly distributed between zero and one.
2) Initialize $\mathbf{W}_{io}$ according to (7) and apply scaling (28).
3) Recalculate magnitude model $\hat{x}_{il}$ according to (16).
4) Update t_{ik} according to (22).
5) Recalculate magnitude model $\hat{x}_{il}$ according to (16).
6) Update v_{kl} according to (23).
7) Scale v_{kl} to unity l^2-norm and compensate by rescaling t_{ik} as specified in (30).
8) Recalculate magnitude model $\hat{x}_{il}$ according to (16).
9) Update z_{ko} according to (21).
10) Scale z_{ko} to l^2-norm and compensate by rescaling t_{ik} as specified in (31).
11) Recalculate magnitude model $\hat{x}_{il}$ according to (16).
12) Calculate $\hat{\mathbf{W}}_{io}$ according to (24) and enforce it to be positive semidefinite by (25).
13) Update $\mathbf{W}_{io}$ according to (26) and apply scaling (28).

The algorithm is implemented by repeating steps 3-13 for a fixed amount of iterations or until the parameter updates converge.

V. Source Reconstruction and Direction Weights Clustering

The separation of sources corresponding to whole physical entities requires clustering the CNMF components that were earlier interpreted as individual sources. The components span over frequency, but due to their fixed spectrum over time they can only model simple audio objects which need to be clustered based on their spatial orientation. This can be compared to the clustering of ICA components consisting of estimates for single frequency bin, whereas the CNMF components are audio objects that are sematically at a higher level.

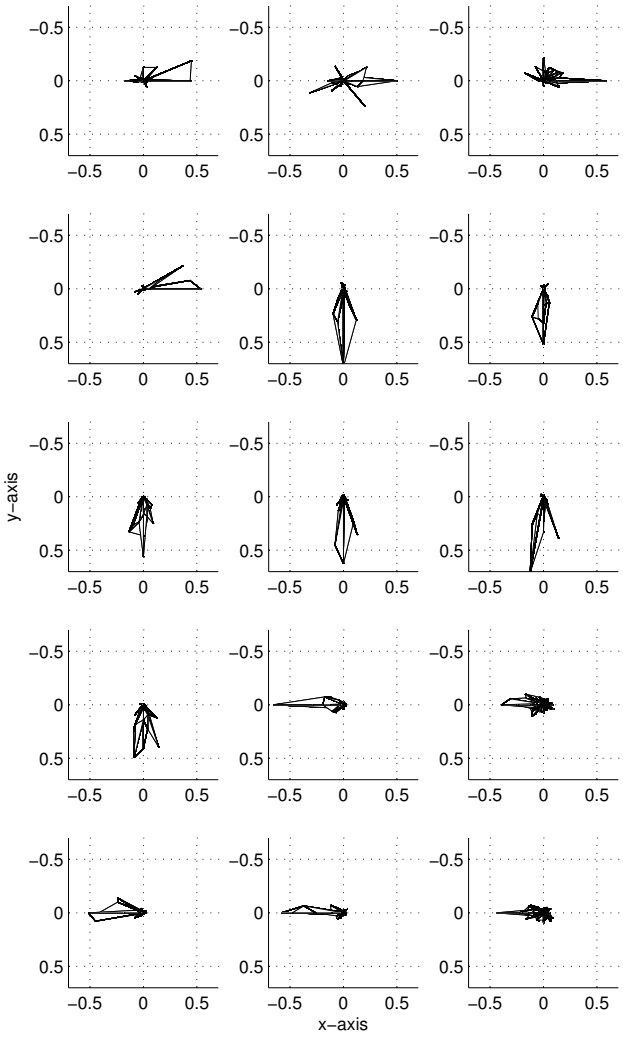


Fig. 5. Look direction vectors weighted by the estimated SCM model parameters for several CNMF components. Vectors denoted by $z_{ko}\mathbf{k}_o$ are projected on to the xy-plane and illustrated as seen above the array. Experimental conditions are described in Section VI-A.

CNMF components originating from the same acoustic source share similar spatial covariance properties determined by their spatial weights z_{ko}. This is illustrated in Figure 5 which depicts SCM model direction weights for several CNMF components showing distinct segmentation to sources at three separate directions. Based on the spatial weight similarity, a separate clustering algorithm can be used to associate CNMF components to the acoustic sources.

We propose to use k-means clustering on the spatial weights z_{ko}. Each $z_{k,:}$ acts as a feature vector and we apply k-means clustering with the cluster count being equal to the number of acoustic sound sources which is defined by the user of the algorithm. We now define the acoustic source index as

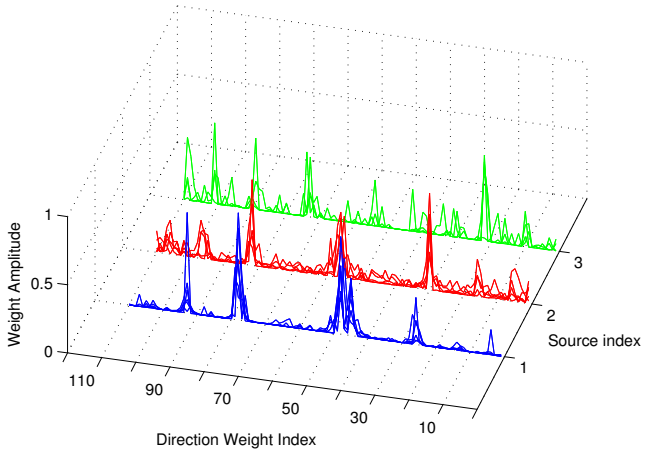


Fig. 6. Direction weights z_{ko} for sources 1, 2 and 3 determined by proposed clustering are illustrated in separated rows. The weights z_{ko} for multiple components k overlaid in each row have a similar peak structure determined by the spatial position of the source.

$q = 1 \ldots Q$. As a result from clustering we get binary cluster decision b_{qk} denoting component k belonging to a source q. An example of direction weights associated to three different sources are illustrated in Figure 6 and direction weights for multiple components k associated to each source are plotted on separate rows. The weights in Figure 6 corresponds to the ones illustrated in Figure 5 projected to the xy-plane.

The CNMF magnitude model for the magnitude spectrogram of an acoustic source q is defined as

$$s_{ilq} = \sum_{ko} b_{qk} z_{ko} t_{ik} v_{kl}. \tag{32}$$

The reconstruction of sources $\mathbf{y}_{ilq}$ as seen by the array, i.e. convolved with their spatial impulse response, based on (32) is given as

$$\mathbf{y}_{ilq} = \mathbf{x}_{il} \frac{\sum_{ko} b_{qk} z_{ko} t_{ik} v_{kl}}{\sum_{qko} b_{qk} z_{ko} t_{ik} v_{kl}}, \tag{33}$$

The time-domain signals are obtained by inverse FFT and frames are combined by weighted overlap-add.

Any other clustering algorithm or CNMF component to source linking strategy can be used to estimate either a binary or a soft decision b_{qk}. We have chosen to use the k-means clustering using spatial weights as features to demonstrate the DoA analysis performance of the proposed SCM model. Other features extracted from the CNMF component parameters such as spectral similarity and gain behavior over time can be used in parallel for associating the CNMF components to the sources to improve the clustering decision [35], [36]. In Section VI-D we study the performance of the chosen k-means clustering strategy against oracle clustering based on known source locations.

VI. Simulations and Source Separation Evaluation

In this section we evaluate the source separation performance of the proposed algorithm. The evaluation consist of a comparison against conventional BSS methods suitable for the case of small microphone array captures with a reasonable amount of reverberation. The separation performance is determined by objective measures, the signal-to-distortion ratio (SDR), image-to-spatial distortion ratio (ISR), signal-to-interference ratio (SIR), and signal-to-artefact ratio (SAR). Additionally, perceptually motivated scores proposed in [32] are reported.

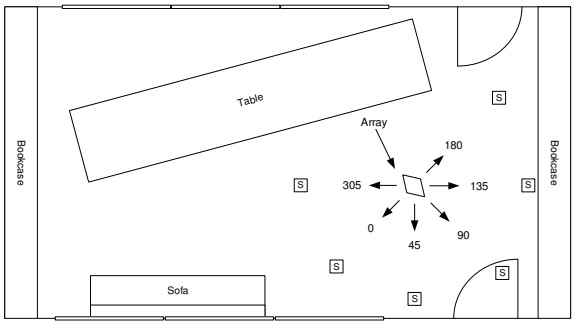


Fig. 7. Capturing room layout, and array and source positions used for datasets one and two.

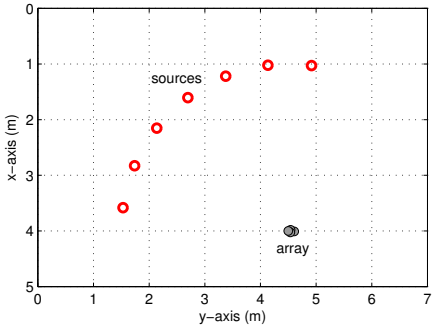


Fig. 8. Simulated room layout for dataset three, microphones illustrated by gray circles and sources by red circles.

A. Evaluation Datasets

For evaluation purposes, a set of room impulse responses (RIR) were measured in a regular meeting room by using an array consisting of four Sennheiser MKE2 omnidirectional condensator microphones inside a metal casing similar to a regular hand-held device. The room dimensions were 7.95 m x 4.90 m x 3.25 m and the reverberation time averaged over all the impulse responses from all locations was $T_{60} = 350$ ms. For obtaining the RIRs, an MLS sequence of order 18 was played using a Genelec 1029 monitor loudspeaker and captured using the array. The room layout and angles of the speaker with respect to the array are given in Figure 7. The angles of the speakers were 0, 45, 90, 135, 180 or 305 degrees, the height of the speaker was set to 1.40 m and the array was placed on a tripod with elevation of 1.08 m. The distance of the loudspeaker to the array was approximately 1.50 m. The microphone locations are given in Table I and the array geometry with a reference axis is illustrated in Figure 9. The spatial aliasing frequency for the given array is 1563 Hz.

Mic	x (mm)	y (mm)	z (mm)
1	0	-46	6
2	-22	-8	6
3	22	-8	6
4	0	61	-18

TABLE I
GEOMETRY OF THE ARRAY USED FOR EVALUATION.

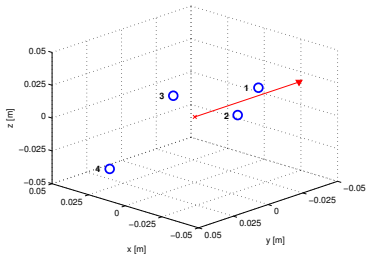


Fig. 9. Array geometry and reference axis.

The test material was generated by convolving the source-to-array RIRs with varying anechoic source material. The material consist of samples of male and female speech, pop music and various everyday noise sources as listed in Tables II and III. The length of the signals is 10 seconds and original sampling frequency of 48 kHz was downsampled to $F_s = 24$ kHz to reduce the computational complexity. The produced spatial images of sources were summed to obtain an array mixture containing specific sources at specific angles.

Datasets with two and three simultaneous sources were generated, referred as dataset one and two respectively. The datasets consist of cases which refer to combination of different types of sources which are described in Tables II and III. For all cases a set of angles were used which are given in Table IV. The case and angle combinations result to 48 different mixture signals for two simultaneous sources and 42 different mixture signals for three simultaneous sources. It results to eight and seven minutes of evaluation material for datasets one and two, respectively.

Additionally, a simulated room with dimensions of 7 m x 4 m x 3 m was generated using a room simulator based on the source-image method [37]. The simulated room was used to study separation of two sources as the function of the angular spacing between the sources, starting from $15°$ spacing to $90°$ with $15°$ increments. The array was rotated $5°$ to prevent any special geometry of it being parallel to the walls, and the first source is located at $8°$ with respect to the reference axis of the array. Distance of sources to the array center is 3 m. The source types used were the same as given in Table II and the corresponding angles are reported in Table IV. The simulated room is illustrated in Figure 8. The target reverberation time was set to 300 ms and the default reflection coefficients for surfaces were used. This test material is referred to as dataset three.

B. Evaluated Methods and Algorithm Parameters

The evaluated algorithms are the proposed CNMF with the DoA kernel based SCM model, the baseline CNMF with unconstrained SCM estimation [19], fullrank SCM estimation [34], ICA variant 1 with magnitude envelope permutation alignment [29] and ICA variant 2 with TDoA permuation alignment [8]. The baseline CNMF was included in the evaluated algorithms to prove the advantages of the proposed DoA kernel based SCM model over the unconstrained SCM model. The ICA methods were chosen to provide comparison to a well known and established BSS techniques.

The results for fullrank SCM estimation [34] are only reported for two sources case. The reference implementation requiring inverting the estimated source covariance matrices which in case of low energy at higher frequencies may become close to singluar and eventually becoming not invertible and preventing the algorithm to proceed from such a state. In the case of three sources dataset, majority of the test samples could not be processed due to the above issue. Additionally, for the same reason one test signal is omitted from the average scores of the fullrank SCM estimation method in two sources dataset.

The proposed separation method only requires three parameters set by the user: the window length, the number of NMF components and the number of iterations for the algorithm updates. The parameters were set to similar values as used in related works [18], [19], and are as follows. The window length of the short-time Fourier transform was set to

Case	Source 1	Source 2
1	music 1	music 2
2	male speech 1	power drill
3	male speech 2	hairdryer
4	male speech 3	music 1
5	male speech 1	male speech 2
6	male speech 3	female speech 1
7	female speech 1	movie trailer
8	male speech 2	vacuum cleaner

TABLE II
DESCRIPTION OF SOURCES IN EACH CASE IN DATASET WITH TWO SIMULTANEOUS SOURCES.

Case	Source 1	Source 2	Source 3
1	music 1	music 2	music 3
2	male speech 1	music 1	movie trailer
3	male speech 1	male speech 3	power drill
4	male speech 2	female speech 1	hairdryer
5	male speech 2	male speech 3	music 1
6	male speech 1	male speech 2	female speech 1
7	male speech 2	male speech 3	vacuum cleaner

TABLE III
DESCRIPTION OF SOURCES IN EACH CASE IN DATASET WITH THREE SIMULTANEOUS SOURCES.

	Dataset 1		Dataset 2			Dataset 3	
Angles	1	2	1	2	3	1	2
1	45°	90°	0°	45°	90°	8°	23°
2	135°	180°	45°	90°	135°	8°	38°
3	0°	90°	0°	45°	305°	8°	53°
4	45°	135°	0°	90°	180°	8°	68°
5	0°	135°	0°	135°	180°	8°	83°
6	45°	180°	45°	135°	305°	8°	98°

TABLE IV
ANGLE COMBINATIONS FOR BOTH DATASETS GIVEN IN DEGREES.

$N = 2048$ with 50% window overlap, the window function used was the square root of Hanning window. The number of NMF components for all the CNMF based algorithms was set to $K = 60$ and the algorithms were run for 300 iterations. The true number of sources was given to the methods. ICA and fullrank SCM estimation methods use same STFT analysis parameters given above. For description of the ICA based separation and its method-specific parameters, please refer to [29]. Fullrank SCM estimation was run for 20 iterations.

The look direction vectors for the proposed CNMF are illustrated in Figure 2. It consists of 110 beam directions which sample the unit sphere surface around the array approximately uniformly. The lateral resolution at zero elevation, i.e. at the xy-plane of the array, is 10 degrees, and the different elevations are at 22.5 degrees spacing. The azimuth resolution is decreased close to the poles of the unit sphere.

C. Separation Metrics

The evaluation is done by comparing each separated signal to the spatial images of the original sources and using objective measures by BSS Eval toolbox [30], [31]. Additionally, perceptually motivated scores proposed in [32] are reported.

The discussion of the separation performance is mainly based on signal-to-distortion ratio (SDR) and signal to interference ratio (SIR). The SDR determines how much of the original signal can be explained by the reconstructed source estimates. It is known to emphasize frequency bins with high energy and thus is somewhat dominated by low frequency content especially in case of music samples. The SDR is, however, an established evaluation technique for separation quality comparison. The interference metric SIR determines the amount of cross-talk, and is therefore a good measure of how well each algorithm can separate sources.

Other metrics, SAR and ISR, measure the amount of additional artefacts produced by the separation, and the accuracy of the spatial image of the reconstructed signals, i.e. how well the spatial position of the reconstructed sources is preserved after reconstruction. The used perceptual metrics are overall perceptual score (OPS), target-related perceptual score (TPS), Interference-related Perceptual Score (IPS) and artifact-related perceptual score (APS) [32].

D. Overall Results

The separation scores averaged over all test samples are given in the Tables V and VI for datasets one and two, respectively. The last row labeled as "mixture" contains the separation metrics evaluated without processing, i.e. calculated for the mixture signal as input for the evaluation toolbox. The results show that the proposed method achieves better average SDR and SIR over the all the compared methods. In the three source dataset the overall separation scores of all the tested methods are fairly low which makes the separation improvement of the proposed method less evident as compared to two sources separation. In the case of three simultaneous sources the SIR performance of the baseline CNMF goes below the performance of ICA separation while the proposed method maintains a better separation with respect to the SIR score. The proposed method also performs best in reconstructing the spatial image of the sources in both test sets.

The score regarding added artefacts to the separated signals measured by the SAR score are lower with the proposed method when compared to the baseline CNMF. This may be attributed to the binary clustering of the proposed method. Faults in the clustering decisions may introduce unwanted rapid changes in the spectrum of the sources, which produces artefacts when reconstructed using the phase of the mixture signal. The baseline CNMF allows soft component-to-source decisions which prevents the added artefacts by smoother spectral discrimination between sources but adds unwanted crosstalk between them. Examples of separated signals from all the evaluated methods are provided at http://www.cs.tut.fi/sgn/arg/nikunen/demo/TASLP2013/.

Method	**SDR** [dB]	**SIR** [dB]	**SAR** [dB]	**ISR** [dB]
CNMF proposed	**4.59**	**7.71**	10.25	**10.29**
CNMF baseline	3.57	4.46	**11.97**	8.31
ICA variant 1	2.86	5.93	9.20	7.87
ICA variant 2	2.03	4.47	8.20	6.95
Fullrank SCM[1]	3.24	6.06	9.19	8.5
Mixture	0.00	0.05	256.76	23.79
Method	**OPS**	**TPS**	**IPS**	**APS**
CNMF proposed	**26.44**	45.70	31.15	54.69
CNMF baseline	16.36	37.13	17.36	**64.01**
ICA variant 1	22.81	46.63	29.99	53.84
ICA variant 2	23.81	46.22	27.83	52.66
Fullrank SCM[1]	25.17	**48.66**	**33.67**	56.86

TABLE V
SEPARATION METRICS FOR DATASET WITH TWO SOURCES.

Method	**SDR** [dB]	**SIR** [dB]	**SAR** [dB]	**ISR** [dB]
CNMF proposed	**2.06**	**4.59**	7.92	**6.37**
CNMF baseline	1.65	-0.10	**9.69**	4.44
ICA variant 1	1.38	3.14	6.39	5.88
ICA variant 2	0.51	1.33	5.59	4.99
Mixture	-3.50	-3.43	251.75	19.62
Method	**OPS**	**TPS**	**IPS**	**APS**
CNMF proposed	21.00	28.51	29.52	40.04
CNMF baseline	17.05	19.64	9.63	**45.10**
ICA variant 1	**26.78**	**49.03**	**37.49**	43.79
ICA variant 2	23.03	44.25	33.95	42.68

TABLE VI
SEPARATION METRICS FOR DATASET WITH THREE SOURCES.

The separation quality measured by the SDR for different cases are reported in Figures 10 and 11 for dataset one and two, respectively. With two simultaneous sources, the proposed method exceeds the baseline CNMF and ICA based separation in most cases. Only in one case the proposed method performs worse than the baseline in terms of the SDR. For very difficult broad-band noise produced by vacuum cleaner, all the tested methods fail to produce adequate separation of the sources. For the three source case, the performance of the evaluated methods have more deviation. The proposed method with simple k-means clustering for determining the NMF component to sources mapping does not produce as constant separation performance as in the case of only two simultaneous

[1]One test signal omitted from the averaging, see Section VI-B.

sources. All the methods fail to provide meaningful separation except in cases 3 and 5.

The overall perceptual score of the proposed method in the case of dataset one indicate the best performance among the tested methods. However, with three simultaneous sources the ICA variant 1 produces better perceptual scores. This can be due to the fact that for most of the cases all the tested methods fail to separate the sources but the reconstruction of sources with ICA based separation are pleasant sounding despite of high amount of crosstalk between the sources. The proposed method in case of faulty component-to-source clustering decision might produce rapid changes in the spectrum which may decrease the perceptual quality and associated scores.

The SDRs for dataset three for the proposed CNMF and baseline CNMF are illustrated in Figure 12 for different angles for the sources. The results clearly indicate that the proposed method benefits from the increased angle between the sources. With the two closest spatial separation of 15 and 30 degrees both the methods produce similar separation with minor advantage for the proposed method. The separation score difference increases when the angle between the sources is increased and starting from 45 degrees the proposed method produces significant improvement for the SDR score over the baseline method.

For performance analysis of the k-means clustering of estimated source spatial weights z_{ko}, a comparison to oracle clustering is provided. The oracle clustering is implemented by searching for the largest value of z_{ko} for each component k and comparing the azimuth of the found index o to the known source angular positions, determined by their azimuth. The component is assigned to the closest known source azimuth. Using the described oracle clustering we only rely on the spatial information of the NMF components but eliminate the effect of possibly faults made by the k-means clustering. The average SDR with oracle clustering are 5.25 dB and 2.94 dB for datasets one and two, respectively. The increase from the k-means clustering are 0.66 dB and 0.88 dB which indicate that the robustness of the k-means clustering is acceptable. Additionally, it can be stated that the spatial weights of the proposed contain the information of the real source spatial positions.

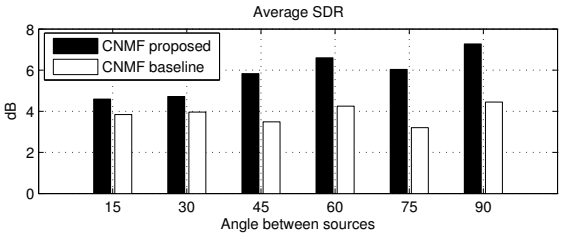


Fig. 12. Averaged SDR for different angle between sources in dataset three with simulated room.

VII. Discussion and Future Work

For the evaluation of the proposed algorithm, a method proposed in [18] that use SCM estimation was also considered. However, it did not prove to be suitable for the given test case according to the separation results and thus no exact separation measures are reported. The lack of any oracle information of the sources and their mixing to give the algorithm a good initial starting point is arguably the reason for low separation performance. Regarding the fullrank SCM estimation [34] of which only partial separation results were reported, issues in solving the permutation ambiguity originating from frequency-wise processing, and the algorithm producing singular SCM estimates for high frequencies were identified as the reason for low separation scores. This indicates the difficulty of the tested case and the efficiency and robustness of the proposed algorithm in analyzing the source spatial covariance.

The future work related to improving the proposed separation framework includes investigating better clustering strategies based on the estimated SCM model direction weights. The k-means clustering does not take into account the geometrical interpretation of the direction weights but solely treats them as feature vectors. Also the clustering decision could be included in the CNMF parameter estimation framework as in done in [19], [21].

In the development of the proposed SCM model, the computational cost of the model was not considered as a design parameter and only good separation performance was sought after. Regarding the computational cost of the model, the number of DoA kernels used for the SCM model increases the computational complexity compared to for example [19]. The average time for performing one iteration with the proposed method takes approximately 9.2 times longer when compared to [19]. The result is obtained with a desktop computer equipped with Intel core2duo E8400 3GHz processor and no special optimization of codes regarding computational complexity for either of the algorithms was made. The computational complexity of the proposed method is approximately linearly proportional to the number of DoA kernels. For example, halving the number of DoA kernels from the used 110 directions to 55 directions decreases the factor to 4.8, when compared to [19]. In general all the CNMF algorithms based on SCM estimation are computationally heavy as they require a high number of iterations in order to converge to a feasible solution.

Reconstruction of the 3D spatial sound field recorded by an irregular array such as the one used in the evaluation requires not only the separation of the sources but information regarding the spatial orientation of the sources. Conventional sound source separation methods do not provide such information, whereas the proposed algorithm can be directly utilized in reconstruction of the spatial sound field the array records. Other soundfield reconstruction methods [38], [39] do not require separation of the sources but utilize more specialized array constellations such as B-format arrays or large linear arrays. For 3D sound field synthesis with the proposed method, the individual CNMF components of the model can be reconstructed without the clustering introduced in Section V, and their direction and spatial spread is determined by the direction weights of the SCM model. Each CNMF component could be synthesized individually and panned and positioned to their analyzed spatial location for example by VBAP [40] or binaural synthesis by HRTF filtering.

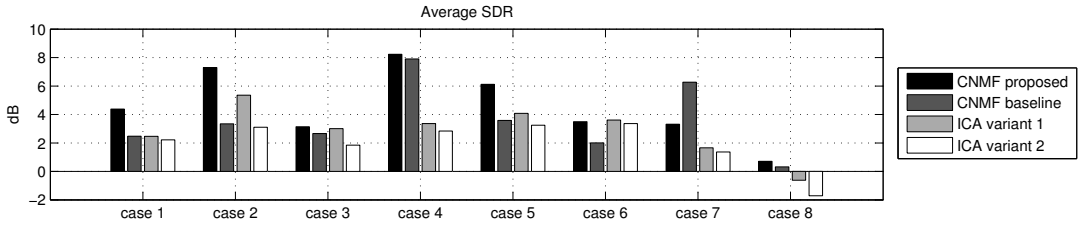


Fig. 10. Averaged SDR for each case for the dataset with two sources.

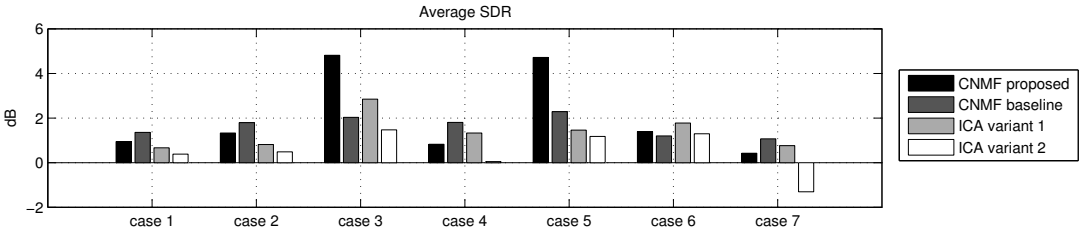


Fig. 11. Averaged SDR for each case for the dataset with three sources.

VIII. Conclusion

In this paper we have proposed a direction of arrival (DoA) based spatial covariance matrix (SCM) model for the purpose of spatial sound source separation using complex-valued non-negative matrix factorization (CNMF). The proposed parameterization of the source SCM by direction-dependent weights allows deriving parameters for the SCM model simultaneously over all frequencies. This improves the overall converge to a spatially coherent solution and mitigates the effect of spatial aliasing which causes problems to many conventional audio separation algorithms. We have shown the separation performance of the proposed algorithm to exceed best performing conventional methods with various types of audio recorded by a small microphone array. The proposed method is a novel approach for spatial parameter estimation in frequency-domain blind source separation, which makes it interesting concept to be utilized in different separation model structures.

Appendix A
Derivation of the CNMF update rules

For the estimation of z_{ko}, t_{ik} and v_{kl}, we redefine the likelihood function (18) by expanding the Frobenius form by using the equality $||\mathbf{A}||^2 = \mathrm{tr}(\mathbf{A}^H\mathbf{A})$ into the form

$$\mathcal{L}^+(\theta) = \sum_{i,l,k,o} \frac{1}{r_{ilko}}[||\mathbf{C}_{ilko}||_F^2 + ||\mathbf{W}_{io}||_F^2 z_{ko}^2 t_{ik}^2 v_{kl}^2 - 2z_{ko}t_{ik}v_{kl}\mathrm{tr}(\mathbf{C}_{ilko}\mathbf{W}_{io})]. \quad (34)$$

Based on the scaling introduced in (28) the second term simplifies to $z_{ko}^2 t_{ik}^2 v_{kl}^2$.

The partial derivatives of (34) with respect to parameters z_{ko}, t_{ik} and v_{kl} are given as

$$\frac{\partial\mathcal{L}^+(\theta)}{\partial z_{ko}} = \sum_{i,l} \frac{2}{r_{ilko}}[z_{ko}t_{ik}^2v_{kl}^2 - t_{ik}v_{kl}\mathrm{tr}(\mathbf{C}_{ilko}\mathbf{W}_{io})] \quad (35)$$

$$\frac{\partial\mathcal{L}^+(\theta)}{\partial t_{ik}} = \sum_{l,o} \frac{2}{r_{ilko}}[z_{ko}^2t_{ik}v_{kl}^2 - z_{ko}v_{kl}\mathrm{tr}(\mathbf{C}_{ilko}\mathbf{W}_{io})] \quad (36)$$

$$\frac{\partial\mathcal{L}^+(\theta)}{\partial v_{kl}} = \sum_{i,o} \frac{2}{r_{ilko}}[z_{ko}^2t_{ik}^2v_{kl} - z_{ko}t_{ik}\mathrm{tr}(\mathbf{C}_{ilko}\mathbf{W}_{io})], \quad (37)$$

where $\theta = \{\mathbf{W}, \mathbf{Z}, \mathbf{T}, \mathbf{V}, \mathbf{C}\}$. Setting the derivatives to zero, substituting r_{ilko} with its definition in Equation (16), and solving the equations with respect to the parameter to be updated, results to update rules

$$z_{ko} \leftarrow \frac{\sum_{i,l}\hat{x}_{il}\mathrm{tr}(\mathbf{C}_{ilko}\mathbf{W}_{io})}{\sum_{i,l}t_{ik}v_{kl}\hat{x}_{il}} \quad (38)$$

$$t_{ik} \leftarrow \frac{\sum_{l,o}\hat{x}_{il}\mathrm{tr}(\mathbf{C}_{ilko}\mathbf{W}_{io})}{\sum_{l,o}z_{ko}v_{kl}\hat{x}_{il}} \quad (39)$$

$$v_{kl} \leftarrow \frac{\sum_{i,o}\hat{x}_{il}\mathrm{tr}(\mathbf{C}_{ilko}\mathbf{W}_{io})}{\sum_{i,o}z_{ko}t_{ik}\hat{x}_{il}}. \quad (40)$$

Above updates can be brought into a multiplicative form (Equations (21) - (23)) by substituting the term in the numerators of the above equations as

$$\hat{x}_{il}\mathrm{tr}(\mathbf{C}_{ilko}\mathbf{W}_{io}) = z_{ko}t_{ik}v_{kl}(\hat{x}_{il} + \mathrm{tr}(\mathbf{E}_{il}\mathbf{W}_{io})) \quad (41)$$

and applying some trivial manipulations of the equations.

The update rule for the spatial covariance matrices $\mathbf{W}_{io}$ is obtained via partial derivation of the negative log-likelihood (18) with respect to $\mathbf{W}_{io}$ which is

$$\frac{\partial\mathcal{L}^+(\theta)}{\partial\mathbf{W}_{io}} = \sum_{l,k} \frac{2}{r_{ilko}}(\mathbf{C}_{ilko} - \mathbf{W}_{io}z_{ko}t_{ik}v_{kl})(-z_{ko}t_{ik}v_{kl}). \quad (42)$$

Setting the above derivative to zero and substituting r_{ilko} with its definition in Equation (16) results in the update

$$\hat{\mathbf{W}}_{io} \leftarrow \frac{\sum_{l,k} \hat{x}_{il} \mathbf{C}_{ilko}}{\sum_{l,k} \hat{x}_{il} z_{ko} t_{ik} v_{kl}} \tag{43}$$

Due to the scaling defined in (29) the divisor in the above update can be disregarded. Substituting $\mathbf{C}_{ilko}$ with its definition (15) the above update can be modified into the multiplicative update given in (24).

REFERENCES

[1] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of Audio Engineering Society*, vol. 55, no. 6, p. 503, 2007.

[2] J. Herre, C. Falch, D. Mahne, G. Del Galdo, M. Kallinger, and O. Thiergart, "Interactive teleconferencing combining spatial audio object coding and dirac technology," *Journal of Audio Engineering Society*, vol. 59, no. 12, pp. 924–935, 2010.

[3] J. Merimaa and V. Pulkki, "Spatial impulse response rendering I: Analysis and synthesis," *Journal of the Audio Engineering Society*, vol. 53, no. 12, pp. 1115–1127, 2005.

[4] P. Smaragdis, "Extraction of speech from mixture signals," in *Techniques for Noise Robustness in Automatic Speech Recognition*, T. Virtanen, R. Singh, and B. Raj, Eds. John Wiley & Sons, 2012.

[5] J. McDonough and K. Kumatani, "Microphone arrays," in *Techniques for Noise Robustness in Automatic Speech Recognition*, T. Virtanen, R. Singh, and B. Raj, Eds. John Wiley & Sons, 2012.

[6] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. Wiley-interscience, 2001.

[7] M. Ikram and D. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, pp. 881–884.

[8] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592–1604, 2007.

[9] F. Nesta, M. Omologo, and P. Svaizer, "Multiple tdoa estimation by using a state coherence transform for solving the permutation problem in frequency-domain bss," in *IEEE Workshop on Machine Learning for Signal Processing*, 2008, pp. 43–48.

[10] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, 2000, pp. 2985–2988.

[11] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[12] D. Seung and L. Lee, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, pp. 556–562, 2001.

[13] D. Lee, H. Seung *et al.*, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[14] A. Cichocki, R. Zdunek, A. Phan, and S. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. Wiley, 2009.

[15] T. Virtanen, "Monoaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 1066–1074, 2007.

[16] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[17] D. FitzGerald, M. Cranitch, and E. Coyle, "Sound Source Separation Using Shifted Non-negative Tensor Factorisation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, 2006.

[18] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[19] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "New formulations and efficient algorithms for multichannel nmf," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 153–156.

[20] ——, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[21] ——, "Formulations and algorithms for multichannel complex nmf," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 229–232.

[22] S. Arberet, A. Ozerov, N. Q. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *International Conference on Information Sciences Signal Processing and their Applications (ISSPA)*, 2010, pp. 1–4.

[23] Y. Mitsufuji and A. Roebel, "Sound source separation based on non-negative tensor factorization incorporating spatial cue as prior knowledge," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 71–75.

[24] I. Tashev, *Sound capture and processing: practical approaches*. John Wiley & Sons Inc, 2009.

[25] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, pp. 2677–2684, 1999.

[26] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[27] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2002.

[28] ——, "Spherical microphone arrays for 3d sound recording," in *Audio Signal Processing: For Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds. Springer, 2004.

[29] J. Nikunen, T. Virtanen, P. Pertila, and M. Vilermo, "Permutation alignment of frequency-domain ica by the maximization of intra-source envelope correlations," in *20th European Signal Processing Conference (EUSIPCO)*, 2012, pp. 1489–1493.

[30] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[31] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," *Independent Component Analysis and Signal Separation*, pp. 552–559, 2007.

[32] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[33] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[34] N. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[35] M. Helén and T. Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, 2005.

[36] T. Barker and T. Virtanen, "Non-negative tensor factorisation of modulation spectrograms for monaural sound source separation," in *Proceedings of INTERSPEECH*, 2013.

[37] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *The Journal of the Acoustical Society of America*, vol. 124, p. 269, 2008.

[38] J. Daniel, S. Moreau, and R. Nicol, "Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging," in *Proceedings of 114th Audio Engineering Society Convention*, 2003.

[39] S. Koyama, K. Furuya, Y. Hiwasaki, and Y. Haneda, "Design of transform filter for reproducing arbitrarily shifted sound field using phase-shift of spatio-temporal frequency," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 381–384.

[40] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997.

Joonas Nikunen (joonas.nikunen@tut.fi) received the M.Sc Degree in signal processing and communications engineering in 2010 from Tampere University of Technology (TUT), Finland. He is currently working as a researcher and post-graduate student at the Department of Signal Processing in TUT and pursuing towards his Ph.D. degree. His research interests include spatial audio representations and coding, sound source separation and object-based representations for audio.

Tuomas Virtanen (tuomas.virtanen@tut.fi) is an Academy Research Fellow and an adjunct professor at Department of Signal Processing, Tampere University of Technology (TUT), Finland. He received the M.Sc. and Doctor of Science degrees in information technology from TUT in 2001 and 2006, respectively. He has also been working as a research associate at Cambridge University Engineering Department, UK. He is known for his pioneering work on single-channel sound source separation using non-negative matrix factorization based techniques, and their application to noise-robust speech recognition, music content analysis and audio event detection. In addition to the above topics, his research interests include content analysis of audio signals in general and machine learning. He has authored about 100 scientific publications on the above topics. He has received the IEEE Signal Processing Society 2012 best paper award.

Publication P7

J. Nikunen and T. Virtanen, "Multichannel audio separation by Direction of Arrival Based Spatial Covariance Model and Non-negative Matrix Factorization," in *Proceedings of the 39th International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, pp. 6727–6731, 2014.

MULTICHANNEL AUDIO SEPARATION BY DIRECTION OF ARRIVAL BASED SPATIAL COVARIANCE MODEL AND NON-NEGATIVE MATRIX FACTORIZATION

Joonas Nikunen, and Tuomas Virtanen

Tampere University of Technology
Korkeakoulunkatu 1, 33720 Tampere, Finland
joonas.nikunen@tut.fi, tuomas.virtanen@tut.fi

ABSTRACT

This paper studies multichannel audio separation using non-negative matrix factorization (NMF) combined with a new model for spatial covariance matrices (SCM). The proposed model for SCMs is parameterized by source direction of arrival (DoA) and its parameters can be optimized to yield a spatially coherent solution over frequencies thus avoiding permutation ambiguity and spatial aliasing. The model constrains the estimation of SCMs to a set of geometrically possible solutions. Additionally we present a method for using a priori DoA information of the sources extracted blindly from the mixture for the initialization of the parameters of the proposed model. The simulations show that the proposed algorithm exceeds the separation quality of existing spatial separation methods.

Index Terms— Spatial sound separation, non-negative matrix factorization, spatial covariance models

1. INTRODUCTION

Sound source separation has many applications which include for example signal enhancement for speech recognition [1] and object-based audio coding [2]. The separation of multichannel audio is usually based on the estimation of the mixing filter in time or frequency domain. Along with the underlying mixing, there exists spectral structure of the sources that can be analyzed from the mixture for example by non-negative matrix factorization (NMF). The utilization of NMF in separation of spatial audio captures in combination with spatial covariance matrix (SCM) estimation has been studied in [3, 4, 5]. Their benefits over conventional methods such as frequency-domain independent component analysis (ICA) [6, 7, 8] are the absence of permutation problem, and the utilization of audio spectrogram redundancy in estimating audio objects, i.e. NMF components, that span over frequency and time. The previous approaches estimate SCMs separately for each frequency of each source, without placing any constraints on the SCMs.

The unconstrained estimation of source SCMs causes several problems. Estimating SCMs separately for each frequency leads to not only the permutation problem [9], but may also produce solutions that are not spatially coherent. Using the NMF as a source magnitude model introduces frequency dependency, but sources at different spatial locations with similar spectral characteristics may become modeled using a single NMF component. Therefore estimating SCMs for NMF components or a group of them still not guarantee a spatially coherent solution.

In this paper, we introduce a direction of arrival (DoA) based SCM model for spatial audio separation and use NMF as the source magnitude model. We propose to model the source SCMs as a weighted combination of DoA kernels which are derived similarly to array manifold vectors towards a certain look direction as in the field of beamforming [10]. A benefit of the model over ones used in [4, 5, 11] is that the proposed structure of the SCMs is constrained by geometrically possible source directions by knowing the array geometry and phase difference caused by each DoA. Additionally, parameterizing the source SCMs by a set of DoA kernels with fixed look directions and their weights results in a model that unifies the phase difference over frequency and thus its parameters are independent of frequency. The proposed model ensures that the SCM for a source is spatially coherent. Furthermore, conventional DoA analysis tools can be used to initialize its parameters.

This paper proposes an improved version of the system [12] and differs from it by estimating the SCM model for entire sources instead of individual NMF components. Additionally, we propose a blind DoA analysis front-end to initialize the SCM model direction weights. We evaluate the performance of the proposed method compared to the method proposed in [4] and to conventional ICA separation [6].

The rest of the paper is organized as follows. In Section 2 we give the problem definition of spatial source separation and source mixing in the spatial covariance domain. The proposed DoA kernel based SCM model is given in Section 3 and a source DoA estimation front-end for initialization of its parameters is explained in Section 4. A complex-valued NMF model incorporating the direction of arrival based SCM model and the optimization of its parameters is presented in Section 5. Simulations for separation quality evaluation with the proposed method are presented in Section 6.

2. PROBLEM DEFINITION

We assume convolutive mixing of sources in time domain, which is approximated by instantaneous mixing in frequency domain. The mixing model is defined as

$$\mathbf{x}_{il} \approx \sum_{p=1}^{P} \mathbf{h}_{ip} s_{ilp} = \sum_{p=1}^{P} \mathbf{y}_{ilp} \tag{1}$$

where $\mathbf{x}_{il} = [x_{il1}, \ldots, x_{ilM}]^T$ is the short-time Fourier transformed (STFT) mixture signal consisting of M channels, and $i = 1...I$ and $l = 1...L$ are the frequency and frame index, respectively. The source index is denoted by $p = 1...P$ and mixing filters by $\mathbf{h}_{ip} = [h_{ip1}, \ldots, h_{ipM}]^T$. The STFTs of the sources are denoted by s_{ilp}. Sources convolved with their impulse responses are denoted by $\mathbf{y}_{ilp} = \mathbf{h}_{ip} s_{ilp}$.

As proposed in [4] we use magnitude square rooted STFT

$$\hat{\mathbf{x}}_{il} \quad = [|x_{il1}|^{1/2} \operatorname{sign}(x_{il1}), \ldots, |x_{ilM}|^{1/2} \operatorname{sign}(x_{ilM})]^T \tag{2}$$

for the calculation of the spatial covariance matrices $\mathbf{X}_{il} = \hat{\mathbf{x}}_{il}\hat{\mathbf{x}}_{il}^H \in \mathbb{C}^{M\times M}$. With the above definitions the magnitude spectrum of each channel is at the diagonal of $\mathbf{X}_{il}$, and the spatial properties of the mixture are represented by its off-diagonal values, which encode the magnitude cross correlation and the phase difference between each microphone pair. The spatial covariances are invariant of the absolute phase, which allows estimation of their spatial properties by phase difference only.

The mixing model (1) in SCM domain equals to

$$\mathbf{X}_{il} \approx \sum_{p=1}^{P} \mathbf{H}_{ip}\hat{s}_{ilp}, \tag{3}$$

where $\mathbf{H}_{ip}$ is the covariance matrix for each source at each frequency and $\hat{s}_{ilp} = (s_{ilp}\overline{s_{ilp}})^{1/2}$ is the corresponding source magnitude spectrum. The problem now becomes estimating the source spectrum and its covariance matrices in such a way that they correspond to spatially coherent sources.

3. SPATIAL COVARIANCE MATRIX MODEL

The proposed SCM model for a single source consists of a weighted sum of DoA kernels that each correspond to a fixed look direction. Each DoA kernel represent the phase difference of a source at a specific spatial location and is obtained by knowing the array geometry. The DoA kernels sample the spatial space around the array approximately uniformly. By estimating the weights corresponding to each direction, the estimation of SCMs is constrained to a search space of geometrically feasible solutions. Additionally, the direction weights are independent of frequency which further unifies the estimation of phase difference over frequency.

Assuming direct path propagation, a point source at a specific spatial location causes a set of TDoAs between all the microphone pairs, which translates into a phase difference in the frequency domain. We introduce a look direction vector $\mathbf{k}_o$ pointing from the geometric center of the array to the source determined by azimuth φ and elevation θ. By knowing the array geometry, we can calculate the time delays between every microphone pair $n = 1 \dots M$ and $m = 1 \dots M$ a source at this direction causes. This is analogous to finding array steering vectors for a sum-and-delay beamformer.

We denote the time delay between microphone pair (n, m) corresponding to look direction $\mathbf{k}_o$ as

$$\tau_{n.}(\mathbf{k}_o) = (\mathbf{k}_o^T(\mathbf{n} - \mathbf{m}))/v, \tag{4}$$

where v is the speed of sound and $\mathbf{n}$ and $\mathbf{m}$ are vectors representing the locations of microphones n and m, respectively. The time delay translates into a phase difference that is linearly proportional to frequency f_i in Hertz. The spatial covariance matrix of a specific look direction $\mathbf{k}_o$, termed here as the DoA kernel, is given as

$$[\mathbf{W}_{io}]_{nm} = \exp\left(j2\pi f_i \tau_{nm}(\mathbf{k}_o)\right), \quad f_i = (i-1)F_s/N, \tag{5}$$

for each STFT frequency index i. The sampling frequency is denoted by F_s and N is the FFT length.

Each DoA kernel $\mathbf{W}_{io} \in \mathbb{C}^{M\times M}$ has a fixed look direction index by $o = 1 \dots O$ which sample the spatial space around the array approximately uniformly. In case of a point source in anechoic capturing conditions, a single look direction would be enough to describe the SCM of the source using Equation (5). However, due to reverberation and diffraction, a more complex model is needed. We propose to use a weighted superposition of DoA kernels, i.e. point sources, resulting in the proposed SCM model

$$\mathbf{H}_{ip} = \sum_{o=1}^{O} \mathbf{W}_{io} z_{po}, \tag{6}$$

where z_{po} are the direction weights corresponding to DoA kernels into each look direction $\mathbf{k}_o$.

By estimating the direction weights that are independent of frequency, the proposed model produces an estimate of $\mathbf{H}_{ip}$ that is spatially coherent over frequency. We restrict the direction weights z_{po} to be non-negative and in Section 5 we introduce an estimation algorithm for them. An example of the direction weights estimated as described later in Section 5 is illustrated in Figure 1.

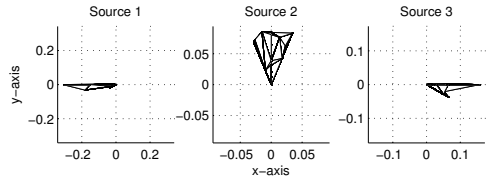


Fig. 1: Illustration of the weighted look direction vectors $z_{po}\mathbf{k}_o$ of the SCM model projected on to the xy-plane. Sources are at 0, 90 and 180 degrees in azimuth.

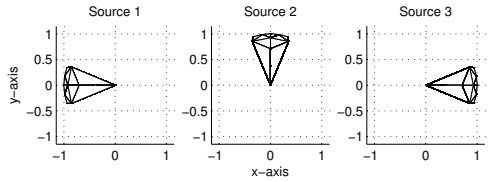


Fig. 2: Illustration of the initialization of the spatial search space for three sources corresponding to Figure 1

4. INITIALIZATION OF DIRECTION WEIGHTS

The parameterization of the source SCM by direction weights z_{po} allows initializing the spatial search space for each source based on DoA analysis of the mixture prior to the model parameter estimation. Based on estimated DoAs defined by azimuth φ_p for each source $p = 1 \dots P$, the direction weights z_{po} are initialized as follows. For each source p the direction weights z_{po} corresponding to look direction indices o within ± 25 degrees from the estimated azimuths φ_p are set to one and all other direction weights of the source are set to zero. The spatial window of 50 degrees accounts for possible errors in the estimation of the source direction in this preprocessing step. An example of the search space used for obtaining the source direction weights in Figure 1 is illustrated in Figure 2.

In the simulations we use the following process to obtain the initial DoA estimates of the sources. Steered response power (SRP) with phase transform (PHAT) [10] is calculated from the STFT of the array signal. The SRP is evaluated for $\varphi = [-180, 180]$ degrees in azimuth with one degree increments and at zero elevation ($\theta = 0$). The maximum of the SRP function at each time frame is scaled to one. The resulting SRP function at each time frame represents the likelihood of a source in each direction. The separation model assumes stationary sources we can therefore average the SRP functions over time. Before averaging, the 15 largest values of the

SRP function are taken from each time frame and the rest of the values are set to zero. Taking only the largest values is equivalent to considering only likelihoods with high confidence. Local maxima that are at least 20 degrees apart from each other are searched from the averaged SRP function. Found locations are set as the initial source DoA estimates. If the number of the found maxima is higher than the number of target sources, the largest maxima are chosen.

5. SEPARATION MODEL

In this section we present the model for the NMF-based spatial sound source separation utilizing the proposed SCM model. Estimation of the parameters of the model follows the framework proposed originally in [4].

The separation model consist of a NMF magnitude model for source spectra $\hat{s}_{ilp} = \sum_{q=1}^{Q} b_{pq} t_{iq} v_{ql}$, where $b_{pq}, t_{iq}, v_{ql} \geq 0$. Each $t_{:q}$ represents the magnitude spectrum of an NMF component, and $v_{q:}$ is its gain in each frame. One NMF component models a single spectrally repetitive event from the mixture and one source is modeled as a sum of multiple components. Parameter b_{pq} represents a soft decision of NMF component q belonging to source p. The second part of the separation model comprises the spatial properties of the sources denoted by $\mathbf{H}_{ip}$, which are represented using the DoA kernel based SCM model $\sum_{o=1}^{O} \mathbf{W}_{io} z_{po}$ as defined in Equation (6). Parameters b_{pq}, t_{iq} and v_{ql} are constrained to non-negative values.

Placing the above definitions into the SCM mixing model defined in Equation (3) results in

$$\mathbf{X}_{il} \approx \hat{\mathbf{X}}_{il} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{o=1}^{O} \mathbf{W}_{io} z_{po} b_{pq} t_{iq} v_{ql}. \tag{7}$$

The cost function for the parameter optimization is the squared Frobenius norm summed over frequency and time as $\sum_{i,l} ||\mathbf{X}_{il} - \hat{\mathbf{X}}_{il}||_F^2$. As proposed in [4], finding the optimal parameters $\theta = \{\mathbf{W}, \mathbf{Z}, \mathbf{B}, \mathbf{T}, \mathbf{V}\}$ for model (7) is shown to be equivalent to minimizing the following negative log-likelihood

$$\mathcal{L}^{+}(\theta, \mathbf{C}) = \sum_{i,l,p,q,o} \frac{1}{r_{ilpqo}} ||\mathbf{C}_{ilpqo} - \mathbf{W}_{io} z_{po} b_{pq} t_{iq} v_{ql}||_F^2, \tag{8}$$

with latent components obeying $\sum_{p,q,o} \mathbf{C}_{ilpqo} = \mathbf{X}_{il}$ and being defined as

$$\mathbf{C}_{ilpqo} = \mathbf{W}_{io} z_{po} b_{pq} t_{iq} v_{ql} + r_{ilpqo} (\mathbf{X}_{il} - \sum_{q,o} \mathbf{W}_{io} z_{po} b_{pq} t_{iq} v_{ql}). \tag{9}$$

The parameters $r_{ilpqo} > 0$ are defined as

$$r_{ilpqo} = \frac{z_{po} b_{pq} t_{iq} v_{ql}}{\hat{x}_{il}}, \quad \hat{x}_{il} = \sum_{q,o} z_{po} b_{pq} t_{iq} v_{ql} \tag{10}$$

For optimizing the model parameters multiplicative update equations are derived. The procedure for solving the update rules is based on setting the partial derivatives of (8) with respect to each updated parameter $b_{pq}, z_{po}, t_{iq}, v_{ql}$ and $\mathbf{W}_{io}$ to zero. Substituting $\mathbf{C}_{ilpqo}$ by its definition (9) and applying simple manipulations, this leads to the multiplicative updates

$$b_{pq} \leftarrow b_{pq} \Big[1 + \frac{\sum_{i,l,o} z_{po} t_{iq} v_{ql} \mathrm{tr}(\mathbf{E}_{il} \mathbf{W}_{io})}{\sum_{i,l,o} z_{po} t_{iq} v_{ql} \hat{x}_{il}} \Big] \tag{11}$$

$$z_{po} \leftarrow z_{po} \Big[1 + \frac{\sum_{i,l,q} b_{pq} t_{iq} v_{ql} \mathrm{tr}(\mathbf{E}_{il} \mathbf{W}_{io})}{\sum_{i,l,q} b_{pq} t_{iq} v_{ql} \hat{x}_{il}} \Big] \tag{12}$$

$$t_{iq} \leftarrow t_{iq} \Big[1 + \frac{\sum_{l,p,o} z_{po} b_{pq} v_{ql} \mathrm{tr}(\mathbf{E}_{il} \mathbf{W}_{io})}{\sum_{l,p,o} z_{po} b_{pq} v_{ql} \hat{x}_{il}} \Big] \tag{13}$$

$$v_{ql} \leftarrow v_{ql} \Big[1 + \frac{\sum_{i,p,o} z_{po} b_{pq} t_{iq} \mathrm{tr}(\mathbf{E}_{il} \mathbf{W}_{io})}{\sum_{i,p,o} z_{po} b_{pq} t_{iq} \hat{x}_{il}} \Big], \tag{14}$$

where $\mathbf{E}_{il} = \mathbf{X}_{il} - \sum_{p,q,o} \mathbf{W}_{io} z_{po} b_{pq} t_{iq} v_{ql}$ is the error of the model in each time-frequency point. To ensure numerical stability, the scale of the parameters are normalized as

$$\hat{a}_q = (\sum_{o=1}^{O} z_{po}^2)^{1/2}, \quad z_{po} \leftarrow z_{po} / \hat{a}_q, \quad b_{pq} \leftarrow b_{pq} \hat{a}_q \tag{15}$$

$$\hat{c}_q = (\sum_{l=1}^{L} v_{ql}^2)^{1/2}, \quad v_{ql} \leftarrow v_{ql} / \hat{c}_q, \quad t_{iq} \leftarrow t_{iq} \hat{c}_q. \tag{16}$$

The diagonal entries of $\mathbf{W}_{io}$ model the relative source magnitude level in each channel, and its off-diagonal values model the cross-channel magnitude and phase difference. This means that their unit magnitude as defined by (6) has to be updated in order to model the magnitude level differences in each channel. The update has to maintain the original phase difference, i.e. the original delay caused by a certain look direction.

For updating the magnitudes of $\mathbf{W}_{io}$, we apply the following scheme, also used in [12]. An initial update with a modified phase is calculated as given by the partial derivation of (8)

$$\hat{\mathbf{W}}_{io} \leftarrow \mathbf{W}_{io} \Big[\sum_{l,p,q} b_{pq} z_{po} t_{iq} v_{ql} (\hat{x}_{il} + \mathbf{E}_{il}) \Big]. \tag{17}$$

In order to avoid a subtractive model, matrices $\hat{\mathbf{W}}_{io}$ are forced to be positive semidefinite, which is achieved as proposed in [4] by calculating an eigenvalue decomposition $\hat{\mathbf{W}}_{io} = \mathbf{V}\mathbf{D}\mathbf{V}^H$ and setting negative eigenvalues to zero. Using the modified eigenvalue matrix $\hat{\mathbf{D}}$ the update is reconstructed as $\hat{\mathbf{W}}_{io} \leftarrow \mathbf{V}\hat{\mathbf{D}}\mathbf{V}^H$. The final update preserving the original DoA kernel phase difference is obtained by

$$\mathbf{W}_{io} \leftarrow |\hat{\mathbf{W}}_{io}| \exp(i \arg(\mathbf{W}_{io})). \tag{18}$$

The overall estimation algorithm is implemented as follows. Values of z_{po} are initialized as explained in Section 4 and other parameters are initialized with positive random numbers. The DoA kernels are initialized according to Equation (5). The updates (11) - (14) and (17) - (18) are repeated for a fixed amount of iterations and the parameter scaling as defined by Equations (15) - (16) are applied between iterations. The procedure results in optimizing the model parameters with respect to the squared Frobenius norm between the observations and the model.

The sources are reconstructed as

$$\mathbf{y}_{ilp} = \mathbf{x}_{il} \frac{\sum_{qo} b_{pq} z_{po} t_{iq} v_{ql}}{\sum_{pq'o'} b_{pq'} z_{po'} t_{iq'} v_{q'l}}, \tag{19}$$

which represents Wiener estimates of the sources as seen by the array, i.e. convolved with their spatial impulse responses. The time-domain signals are obtained by inverse STFT and frames are combined by weighted overlap-add.

6. SIMULATIONS

We evaluate the separation quality of the proposed method using separation metrics proposed in [13] and compare its performance against the following methods: NMF with component-wise DoA

Mic	x (mm)	y (mm)	z (mm)
1	0	-46	6
2	-22	-8	6
3	22	-8	6
4	0	61	-18

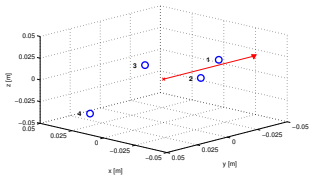


Table 1: Geometry of the array.

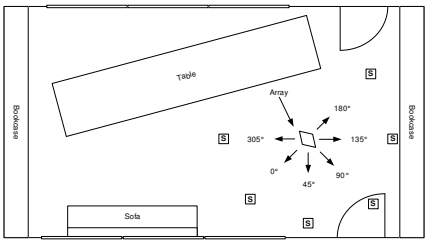


Fig. 3: Room layout, array (tetragon) and source positions (S).

kernel SCM, where the NMF components are grouped to sources by clustering [12], NMF with unconstrained SCM estimation [4] and frequency domain ICA with TDoA based permutation alignment [6].

The test material was generated from anechoic samples that were convolved with room impulse responses (RIR) captured using an array consisting of four omnidirectional microphones enclosed in a metal casing of size 30 mm x 60 mm x 1150 mm. Locations of the microphones are given in Table 1. A Genelec 1029 loudspeaker was used to capture the RIRs from different directions around the array. The height of the loudspeaker was set to 1.40 m and the array was placed on a tripod with elevation of 1.08 m. The distance of the loudspeaker to the array was approximately 1.50 m. The recording location was a meeting room with dimensions of 7.95 m x 4.90 m x 3.25 m and the reverberation time averaged over all the impulse responses from all directions was $T_{60} = 350$ ms. The room layout and directions are shown in Figure 3.

The anechoic samples consisted of male and female speech, pop music and various everyday noise sources. The speech samples were obtained from Librivox audiobooks database, the music samples are from RWC Music Genre Database [14] and the noise sources were recorded at an anechoic chamber. Each sample was 10 seconds in duration and they were downsampled from sampling frequency of 48 kHz to $F_s = 24$ kHz. Different datasets for two and three simultaneous sources were generated by convolving the anechoic material with the measured RIRs and summing separate sources from different angles. The used angles are given in Table 2. Using eight combinations of source types for dataset one and seven combinations for dataset two resulted in 48 different mixture signals for two simultaneous sources and 42 different mixture signals for three simultaneous sources.

Dataset 1		**Dataset 2**		
source 1	source 2	source 1	source 2	source 3
45°	90°	0°	45°	90°
135°	180°	45°	90°	135°
0°	90°	0°	45°	305°
45°	135°	0°	90°	180°
0°	135°	0°	135°	180°
45°	180°	45°	135°	305°

Table 2: Angle combinations for both datasets given in degrees.

Method	**SDR**	**SIR**	**SAR**	**ISR**
Proposed	**5.6**	6.8	**13.1**	9.9
NMF clustering [12]	4.8	**8.1**	10.3	**10.5**
NMF Unconstrained [4]	3.7	4.5	12.7	8.4
ICA [6]	2.0	4.5	8.2	6.9

Table 3: Separation metrics for dataset with two sources. All figures in decibels.

Method	**SDR**	**SIR**	**SAR**	**ISR**
Proposed	**3.0**	2.6	**10.7**	6.0
NMF clustering [12]	1.9	**3.8**	7.6	**6.2**
NMF Unconstrained [4]	2.0	0.4	9.9	4.7
ICA [6]	0.5	1.3	5.6	5.0

Table 4: Separation metrics for dataset with three sources. All figures in decibels.

The parameters of the algorithms were set to values similar to the ones used in related studies and are as follows. The window length of the STFT was set to $N = 2048$ with 50% overlap, the window function was square root of Hanning window. The number of NMF components was set to $Q = 60$ and the algorithms were run for 500 iterations. The true number of sources was given to the methods. The DoA kernels for the proposed SCM model consists of 110 directions which sample the unit sphere surface around the array approximately uniformly. The lateral resolution at zero elevation is 10 degrees, and the different elevations are at 22.5 degrees spacing. The azimuth resolution is decreased close to the poles of the unit sphere.

The separation performance is determined by objective measures, the signal-to-distortion ratio (SDR), image-to-spatial distortion ratio (ISR), signal-to-interference ratio (SIR) and signal-to-artefact ratio (SAR). The results averaged over all test samples and all separated sources are given in Tables 3 and 4. The method in [12] is denoted in the tables by "NMF clustering".

The results show that the separation performance of the proposed method exceeds the unconstrained SCM estimation method and frequency-domain ICA across all the measured quantities. The separation measured by SDR when comparing to [4] is increased by 1.9 dB and 1.0 dB in the dataset with two and three sources, respectively. The SIR score denoting source interference is slightly decreased from the NMF component-wise SCM estimation, but it is mostly due to the method in [12] using binary NMF component to source clustering.

7. CONCLUSION

We have presented a spatial audio separation method based on the NMF magnitude model combined with a source SCM model consisting of direction of arrival (DoA) kernels. The strength of the method is the parameterization of the spatial properties of sources by their direction instead of unconstrained estimates which also allows the initialization of the model parameters by a DoA analysis preprocessing step. The separation based on the NMF magnitude model was shown to exceed the quality of the most recent spatial separation method which use unconstrained SCM estimation. An additional benefit of the proposed spatial parameterization is the possibility of the reconstruction of the 3D spatial sound field by positioning the separated sources by their analyzed direction.

8. REFERENCES

[1] A. Hurmalainen, J.F. Gemmeke, and T. Virtanen, "Modelling non-stationary noise with spectral factorisation in automatic speech recognition," *Computer Speech & Language*, 2012.

[2] J. Herre, C. Falch, D. Mahne, G. Del Galdo, M. Kallinger, and O. Thiergart, "Interactive teleconferencing combining spatial audio object coding and DirAC technology," *Journal of Audio Engineering Society*, vol. 59, no. 12, pp. 924–935, 2010.

[3] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[4] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "New formulations and efficient algorithms for multichannel nmf," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2011, pp. 153–156.

[5] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[6] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592–1604, 2007.

[7] F. Nesta, M. Omologo, and P. Svaizer, "Multiple TDoA estimation by using a state coherence transform for solving the permutation problem in frequency-domain BSS," in *IEEE Workshop on Machine Learning for Signal Processing*. IEEE, 2008, pp. 43–48.

[8] F. Nesta and M. Omologo, "Convolutive underdetermined source separation through weighted interleaved ICA and spatio-temporal source correlation," *Latent Variable Analysis and Signal Separation*, pp. 222–230, 2012.

[9] N.Q.K. Duong, E. Vincent, and R. Gribonval, "Underdetermined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[10] I.J. Tashev, *Sound capture and processing: practical approaches*, John Wiley & Sons Inc, 2009.

[11] S. Arberet, A. Ozerov, N.Q.K. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for underdetermined reverberant audio source separation," in *International Conference on Information Sciences Signal Processing and their Applications (ISSPA)*. IEEE, 2010, pp. 1–4.

[12] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 727–739, 2014.

[13] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," *Independent Component Analysis and Signal Separation*, pp. 552–559, 2007.

[14] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Database: Music Genre Database and Musical Instrument Sound Database," in *Proceedings of the 4th International Conference on Music Information Retrieval*, 2002, pp. 229–230.

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-3438-6
ISSN 1459-2045