Multi-Perspective Scene Analysis from Tetrahedral Microphone Recordings

BSc Matthias Blochberger
Matr.Nr.:01273011

Supervisor: Ass.Prof. DI Dr.rer.nat. Franz Zotter
Assessor: O.Univ.Prof. Mag.art. DI Dr.techn. Robert Höldrich

Graz, April 21, 2020

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Place & Date	Signature

Eidesstaatliche Erklärung [1]

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die benutzten quellenwörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Ort & Zeit	Unterschrift

1. Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008. Genehmigung des Senates am 01.12.2008

Kurzfassung

Eine überzeugende Einschließung von Nutzer_innen in eine virtuelle Realität setzt die Ermöglichung echtzeitfähigen interaktiven Hörens in dreidimensionalen Klangszenen voraus. Für ein realitätsnahes Hörerlebnis muss sich die akustische Perspektive und Orientierung in Echtzeit variabel durch die körpereigenen Bewegungen steuern lassen. Diese Arbeit befasst sich damit, virtuellen Hörer_innen eine Klangszene aus einer interpolierten variablen Perspektive zu präsentieren, während die ursprüngliche Szene lediglich an wenigen gleichzeitig aufgenommenen statischen Einzelperspektiven vorliegt. Für die variabelperspektivische Interpolation wird die Klangszene in mehrperspektivisch lokalisierbare Klangobjekte und einen Rest zerlegt. Die Information über lokalisierbare Objekte entspringt einer Wahrscheinlichkeitslandkarte, die aus der Zusammenfassung der Richtungsdetektionen aller einzelnen Perspektiven hervorgeht. Diese Arbeit schlägt einen Partikelfilter-Ansatz zur laufenden Lokalisierung von Klangobjekten in der Klangszene vor. Dieser findet in der Wahrscheinlichkeitslandkarte einen geeigneten, zeitlich zusammenhängenden Positionspfad pro Klangobjekt. Zur Wiedergabe wird für jedes Klangobjekt seinem Positionspfad gemäß ein Klang aus der Aufnahme extrahiert und relativ zum virtuellen Subjekt in das Restsignal eingebettet.

Abstract

Convincing immersion in virtual reality requires to enable the user to engage in interactive listening within three-dimensional audio scenes. To achieve a realistic listening experience, the acoustic perspective and orientation has to be real-time controlled with the own body movements. This thesis addresses the task of presenting an interpolated variable perspective to an interacting listener, while the original audio scene is recorded simultaneously at only a few static perspectives. The scene is decomposed into localizable sound objects and a residual signal for the variable-perspective interpolation. Information regarding localizable objects is extracted from a probability map that is composed from the directions detected by the collective of the available single perspectives. This work proposes a particle-filter-based approach for a continuous position estimation of sound objects. The particle filter uses the probability map to estimate a continuous trajectory for each sound object in the scene. The rendering approach extracts signals from the recording for each localized sound object according to its estimated trajectory and embeds it relative to the virtual listener into the residual signal.

Contents

Chapter 1

Introduction

The number of possibilities in the digital virtual world is growing thanks to recent developments that boost availability and affordability of virtual reality hardware. The entertainment industry, in particular the growing gaming market, is responsible for a big part of the recent push to bring virtual reality to the masses. Due to the demand of high quality content for VR including high quality and immersive audio playback, rendering algorithms for audio in VR is an emerging area of research.

An application of this technology is approached in this work, namely the recording and playback of entire sound scenes. The information of spatiality is vital to facilitate playback featuring naturalness that holds true for 6DoF rendering. Therefore, a method to record, analyse and re-synthesise is necessary. This thesis focuses on the development and evaluation of such a method.

1.1 Overview

The basic motivation of this work is to formulate an approach which is able to analyse a sound scene from *multiple perspectives* and recreate it for an interacting virtual listener. This requires a method which allows the listener to move freely between these static perspectives by facilitating some way of interpolation.

The use of interpolation on *Binaural Room Impulse Responses* (BRIRs) for auralisation in interactive virtual spaces has notable research such as [NK17, NKKK17, NR18] as well as [Mül20] or a simpler approach using other auditory cues such as [Dep17].

This work however will not use BRIRs. Instead, the perspectives are represented by surround recordings that would allow to switch playback between multiple perspectives. Research on how to interpolate these recordings for a variable-perspective playback has been approached by interpolation of surround signals in the Ambisonics domain in [TDTH13, PP15, TC16, AK17, TC19, Tyl19]. Further methods are proposed and analysed in [GZS+18,

CJ20, ZFSH20, RZF17].

The aforementioned recording perspectives are captured by microphone arrays that are located at multiple static positions. As adopted in this work, each of the recording perspectives could be captured by a tetrahedral microphone array, which is compact and has four cardioid microphones directed outwards. The 4 microphone look directions are gathered in the array direction matrix $\mathbf{V}$ as the unit length Cartesian vectors $\boldsymbol{\theta}_{c_i}$ denoted as

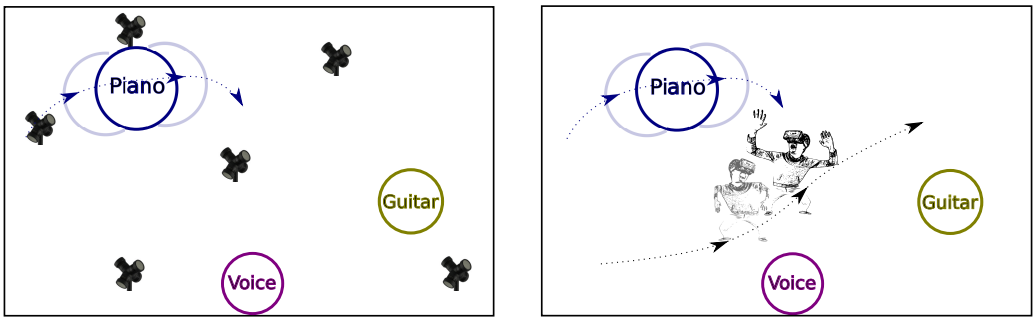
$$\mathbf{V} = \begin{bmatrix} | & & | \\ \boldsymbol{\theta}_{c_1} & \cdots & \boldsymbol{\theta}_{c_i} \\ | & & | \end{bmatrix}. \tag{1.1}$$

Each of those perspectives contributes some information about the spatial distribution of sound objects seen from the respective perspective position, contained in its array signals. Since for general signals, a single-perspective recording by tetrahedral arrays mainly provides directional information, it requires additional distance data for object localisation, as obtained by projection on a convex hull, for instance, in [PST+18]. However, a multi-perspective approach can provide object distance by intersecting single-perspective directional information. This idea is topic of works such as [BOS08, BOS10, Hac15]

Based on this idea, this work establishes a system consisting of particle filters is employed to track sound objects in the sound scene. Particle filters are a proven method of tracking peaks in unknown probability functions and generate a continuous spatial trajectory in time when following moving objects. Particle filters are used in acoustic source tracking in [WLW03b, KG18, VMR07] and a non-exhaustive list of fundamental literature on particle filters is [Efr93, LC95, Fea98, Kit96, LC98, Kit96, CCF99, DdFG01, Sär13].

The information retrieved by the particle filtering on time-varying object locations enables signal extraction of the object sound and variable-perspective rendering thereof. For each sound object, a direct signal is approximated using beam-forming and perspective merging to minimize signal cross contamination. Encoding these signals with direction and amplitude attenuation according to spatial constellation of the objects and with regards to the listener perspective yields a spatially accurate reconstruction of the major sound

Figure 1.1: Example of a tetrahedral microphone array. (Oktava MK-4012. Image Credit: [Okt19])



(a) Recording situation (b) Virtual playback situation

Figure 1.2: The goal of this work is to record spatially diverse scenes and facilitate playback for virtual listeners with full freedom of movement and accurate spatial reproduction.

objects in the scene. In addition the enveloping room response and those sounds whose location features are not salient enough to be tracked by the particle filters remain as residual signals. To preserve there residuals as important background cues, they are rendered with less precise reproduction methods, which are nevertheless suitable to render consistently with the listener perspective. The residual signal reproduction method, as introduced in this thesis, is based on the approach described in [GZS+18, ZFSH20].

1.2 Thesis Outline

Chapter 2 describes the concept of direction-of-arrival estimation and the computation of directional distributions. Further, perspective combination is explained yielding the acoustic activity map.

Chapter 3 delves into details about the particle filter system, the peak picking procedure and a probabilistic birth-death algorithm for object detection.

Chapter 4 explains the concepts behind the proposed rendering algorithm using the analysis data acquired by the detection and tracking procedure.

Chapter 5 is the extensive summary of (1) a numerical evaluation of the proposed detection and tracking algorithm and (2) a listening evaluation assessing scene playback by the rendering algorithm.

Chapter 6 concludes on the research tasks accomplished in this thesis, answers the major questions, and states suggestions for future research on the topics at hand.

Chapter 2

Acoustic Activity Map

The *acoustic activity* map characterizes the measured sound object activity in a three-dimensional space. Starting from perspective recordings from P microphone arrays located at positions $\boldsymbol{p}_p$ in an acoustic scene, the procedure consists of computing directional distributions for all and combining these into one three-dimensional map. This activity map is furthermore used to locate sound objects in the scene as peaks.

The underlying concept of this map, which is an expansion of the ones in [BOS08, BOS10, Hac15], is explained in this chapter.

2.1 Perspective Directional Distribution

Direction-of-arrival (DOA) refers to the direction from which a propagating wave arrives at a measurement setup, which is usually a sensor array and, as in this case, a tetrahedral microphone array. The estimation of the DOA is not a novel concept and has been topic of research and applied in a multitude of ways, e.g. MUSIC [Sch86], ESPRIT [RK89], beamformer-based approaches as found in [KV96, TH13] or instantaneous estimates such as in [WVHK06, PDPM15, PDP15a, Pol16, Wil16]. Related methods are used in acoustic algorithms like [PF06, Pul06, PTP18, BB10, WEJ12].

In this section, a practice-proof algorithm for DOA estimation based on [PDP15b, Wil16] is explained and subsequently applied to compute direction estimates and further directional distributions using the transformation into the spherical harmonic domain.

2.1.1 Direction of Arrival Estimation

The smoothed magnitude sensor response as proposed in [Wil16] is a method which superposes the directions of the array microphones according to the magnitudes of the microphone signals in the frequency domain. The enhancement compared to the original

magnitude sensor response [PDP15b] is a separation into signal and noise subspace for better estimation results.

Assuming microphones with identical directivity patterns, the resulting vector is an estimation of the DOA. However, the condition for this method to work is a balanced directional layout and microphone directivity pattern that monotonically decreases with the angle enclosed with the direction of each microphone. This condition is fulfilled by tetrahedral sensor arrays with a first order cardioid pattern.

The time-domain array microphone signals

$$\boldsymbol{s}(t) = \begin{bmatrix} s_1(t) & s_2(t) & s_3(t) & s_4(t) \end{bmatrix}^{\mathrm{T}}$$

are transformed into frequency-time signals

$$\boldsymbol{S}^{(k)}[m] = \begin{bmatrix} S_1^{(k)}[m] & S_2^{(k)}[m] & S_3^{(k)}[m] & S_4^{(k)}[m] \end{bmatrix}^{\mathrm{T}}$$

and used to calculate a running estimate of the covariance matrix $\boldsymbol{\Sigma}^{(k)}(m)$ over M time frames. The frequency bin index and time frame index are denoted k and m respectively. The estimation for a time instant m is done as an average over M frames

$$\boldsymbol{\Sigma}^{(k)}(m) = \frac{1}{M} \sum_{m'=m-M/2}^{m+M/2-1} \boldsymbol{S}^{(k)}[m'] \cdot \boldsymbol{S}^{(k)}[m']^{\mathrm{H}}. \tag{2.1}$$

Note that the time frame index m is dropped from the explanations below for readability, so keep in mind that the subsequent steps are applied at all time steps m.

The first step is the decomposition of the estimated covariance matrix for all k into the 4×4 column matrix of eigenvectors $\boldsymbol{U}^{(k)}$ and the diagonal eigenvalue matrix $\boldsymbol{\Lambda}^{(k)}$:

$$\boldsymbol{\Sigma}^{(k)} = \boldsymbol{U}^{(k)} \boldsymbol{\Lambda}^{(k)} \boldsymbol{U}^{\mathrm{H}(k)}. \tag{2.2}$$

The eigenvalues on the diagonal are ordered from largest to smallest ($\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_i$). Now, separating the signal and noise subspace requires the selection of relevant eigenvalues and corresponding eigenvectors. Assuming one incoming signal, the largest eigenvalue and its eigenvector is the best candidate. However, this estimation can also be applied to arrays with more microphones and the introduction of methods, facilitating eigenvalue selection, such as SORTE [HCXC10] can be useful. Depending on the method, different numbers of selected eigenvalues per frequency bin k can appear which will be denoted as L.

For all frequency bins k, regardless of the selection method, each four-dimensional eigenvector $\boldsymbol{u}_{*l}^{(k)}$ is recombined into a three-dimensional direction estimate by matrix multiplication with the array directions $\mathbf{V}$, see Eq. (1.1), and normalized to unit length. This

becomes

$$\hat{\boldsymbol{\theta}}_l^{(k)} = \frac{\mathbf{V}|\boldsymbol{u}_{*l}^{(k)}|}{\|\mathbf{V}|\boldsymbol{u}_{*l}^{(k)}|\|} \quad \text{for } l = 1...L \tag{2.3}$$

where $|\boldsymbol{u}_{*l}^{(k)}|$ denotes the element-wise complex magnitude of the vector $\boldsymbol{u}_{*l}^{(k)}$.

2.1.2 Spherical Harmonics Directional Distribution

Following the selection of the L relevant eigenvalues $\lambda_l^{(k)}$ and the calculation of the direction estimates $\hat{\boldsymbol{\theta}}_l^{(k)}$ (cf. Eq. (2.3)), they are used to compute a directional distribution based on directional energy histograms. For a strictly defined, smooth resolution of this histogram, it is composed in the spherical harmonics domain. A further advantage of SHs are the straightforward procedures for the discrete spherical harmonic transform (DSHT) and its inverse (IDSHT).

The fully normalized spherical harmonic functions for any order N and their derivation are available in literature, e.g. [ZF19]. The definitions used there is

$$Y_n^m(\varphi, \vartheta) = \underbrace{N_n^{|m|}}_{\substack{\text{norm-}\\\text{alization}}} \underbrace{P_n^{|m|}(\cos\vartheta)}_{\substack{\text{assoc.}\\\text{Legendre}\\\text{functions}}} \underbrace{\Phi_m(\varphi)}_{\substack{\text{azimuth}\\\text{harmonics}}}.$$

Numerous implementations for computations are available [1] [2]. For readability they will be denoted as $Y_n^m(\boldsymbol{\theta})$. The $(N+1)^2$ transformation coefficients for $n = 0...N$ and $m = -n...n$ for any direction vector $\boldsymbol{\theta}$ are defined as

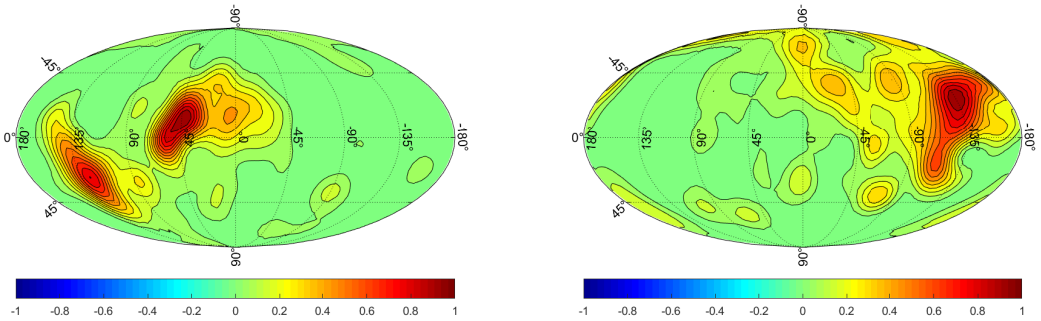
$$\boldsymbol{y}(\boldsymbol{\theta}) = \begin{bmatrix} Y_0^0(\boldsymbol{\theta}) & Y_1^{-1}(\boldsymbol{\theta}) & \dots & Y_N^N(\boldsymbol{\theta}) \end{bmatrix}^{\mathrm{T}}.$$

With this, the L direction estimates $\boldsymbol{\theta}_l^{(k)}$ are transformed into the SH domain using the appropriate DSHT coefficients $\boldsymbol{y}(\boldsymbol{\theta}_l^{(k)})$. These transformed direction-spread functions are linearly combined using a function $\mathcal{L}(k, \lambda_l^{(k)})$ to define a directional distribution

$$\boldsymbol{\chi} = \sum_{k=1}^{K} \sum_{l=1}^{L} \boldsymbol{y}(\boldsymbol{\theta}_l^{(k)})\, \mathcal{L}(k, \lambda_l^{(k)}). \tag{2.4}$$

The eigenvalue-dependent and frequency-dependent function $\mathcal{L}(k, \lambda_l^{(k)})$ can be defined arbitrarily, but in this case a useful definition was found in a combined frequency-domain

1. MATLAB® implementation *Spherical-Harmonic-Transform* by Archontis Politis available at https://github.com/polarch/Spherical-Harmonic-Transform
2. Pure Data externals *iem_ambi, iemmatrix*



(a) Example 1

(b) Example 2

Figure 2.1: Two examples of directional distributions of perspectives in a scene with 2 active sound objects. Depending on distance and loudness more ore less distinct patterns are visible. For visualization, the SH distribution has been decoded to 2000 equally distributed directions and normalized to a maximum value of 1.

attenuation and compression function:

$$\mathcal{L}(k,\lambda) = \begin{cases} k\sqrt{\lambda} & \text{if } \frac{k}{K} f_s > 200Hz \\ 0 & \text{else} \end{cases} . \tag{2.5}$$

Here, f_s is the sampling rate. Such a distribution $\boldsymbol{\chi}$ is generated to map the observed directional energy distribution of each perspective microphone array $p = 1...P$.

2.2 Perspective Combination

Similar to the proposed approaches in [BOS08, BOS10, Hac15], the directional distributions captured at multiple, single-perspective recording positions are combined into the three-dimensional acoustic activity map. In order to explain the procedure involved, we want to define a set of arbitrary positions $\boldsymbol{s}_i$ in addition to the known perspective positions. The new set represents points on the acoustic activity map. To calculate the map values for these positions, an algorithm of perspective merging is introduced. The procedure involves the mapping of the aforementioned positions to the single-perspective SH directional distributions values and a subsequent additional weighting thereof. It is explained in the following sections.

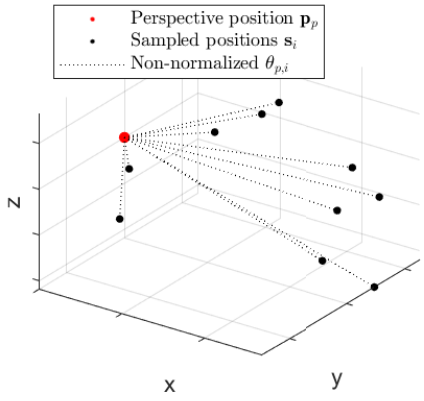


Figure 2.2: The three-dimensional space is sampled at an arbitrary set of points $\boldsymbol{s}_i$ whose positions can be evaluated for acoustic activity by the perspective merging procedure.

2.2.1 Direction-Position Mapping

The set of positions intended to be evaluated for their acoustic activity has to be mapped to SH directional distribution values. This is done by the calculation of perspective-position directions $\boldsymbol{\theta}_{p,i}$ as visualized in Fig. 2.2 and using SH sampling decoding matrices. Such a matrix consists of the coefficients of the inverse discrete SH transformation for the directions pointing from perspective to the sample points:

$$\boldsymbol{\theta}_{p,i} = \frac{\boldsymbol{s}_i - \boldsymbol{p}_p}{\|\boldsymbol{s}_i - \boldsymbol{p}_p\|}. \tag{2.6}$$

The directions for p and all $i = 1...G$ are computed and gathered in the direction matrix

$$\boldsymbol{\Theta}_p = \begin{bmatrix} | & | & & | \\ \boldsymbol{\theta}_{p,1} & \boldsymbol{\theta}_{p,2} & \cdots & \boldsymbol{\theta}_{p,G} \\ | & | & & | \end{bmatrix}.$$

Next, a DSHT sampling decoder $\tilde{\mathbf{Y}}_p$ of order N is calculated from the direction matrix $\boldsymbol{\Theta}_p$:

$$\tilde{\boldsymbol{Y}}_p(\boldsymbol{\Theta}_p) = \begin{bmatrix} | & | & & | \\ \mathbf{y}(\boldsymbol{\theta}_{p,1}) & \mathbf{y}(\boldsymbol{\theta}_{p,2}) & \ldots & \mathbf{y}(\boldsymbol{\theta}_{p,G}) \\ | & | & & | \end{bmatrix}.$$

The directional distribution values of the sample positions $\boldsymbol{s}_i$ mapped from a single perspective are denoted as $\tilde{\boldsymbol{w}}_p$ of size $G \times 1$. This vector is yielded by the mapping step

$$\tilde{\boldsymbol{w}}_p = \tilde{\boldsymbol{Y}}_p \boldsymbol{\chi}_p. \tag{2.7}$$

Gathering the decoding matrices into the columns of a larger matrix

$$\tilde{\boldsymbol{\Upsilon}} = \begin{bmatrix} \tilde{\boldsymbol{Y}}_1(\boldsymbol{\Theta}_1) & \tilde{\boldsymbol{Y}}_2(\boldsymbol{\Theta}_2) & \dots & \tilde{\boldsymbol{Y}}_P(\boldsymbol{\Theta}_P) \end{bmatrix} \tag{2.8}$$

and arranging the single-perspective SH distributions $\boldsymbol{\chi}_p$ as a block diagonal matrix

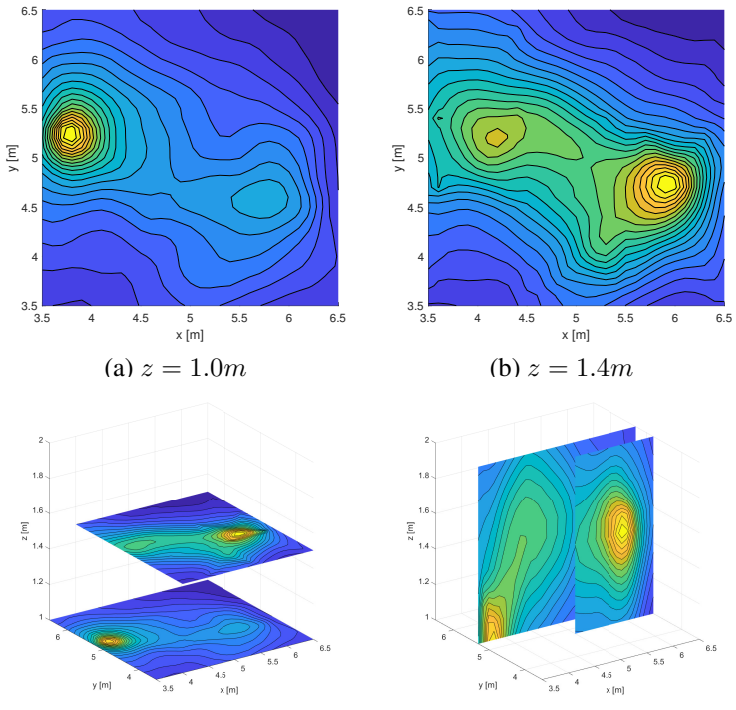
$$\mathbf{X} = \begin{bmatrix} \boldsymbol{\chi}_1 & & \\ & \ddots & \\ & & \boldsymbol{\chi}_P \end{bmatrix} \tag{2.9}$$

permits the joint mapping step mapping of all perspectives yielding a $G \times P$ perspective activity map

$$\tilde{\mathbf{W}} = \begin{bmatrix} | & & | \\ \tilde{\boldsymbol{w}}_1 & \dots & \tilde{\boldsymbol{w}}_P \\ | & & | \end{bmatrix} = \tilde{\boldsymbol{\Upsilon}}\mathbf{X} \tag{2.10}$$

holding the acoustic activity values for each sample point and perspective.

The unweighted linear combination $\tilde{\mathbf{W}}\,\mathbf{1}_{\mathrm{P}\times 1}$ of the columns yields a set of activity values for the sample positions. Figure 2.3 shows cross sections of such an activity map.



(a) $z = 1.0m$

(b) $z = 1.4m$

(c) View of evaluated horizontal planes in spatial context.

(d) View of evaluated vertical planes in spatial context.

Figure 2.3: The acoustic activity map evaluated at equidistant grids on different planes in the scene at a time frame. Here 2 sound object peaks are visible.

2.2.2 Perspective Weighting

The perspective activity values are combined into merged activity values γ_i for all points $i = 1...G$. The single perspective values in $\tilde{\mathbf{W}}$ are weighted according to distance as proposed in [Hac15] with small modifications.

Distance Weighting

Perspectives are assigned higher or lower importance depending on the distance between perspective position and the point to be evaluated, $\|\boldsymbol{s}_i - \boldsymbol{p}_p\|$. The distance function is decreasing exponentially limited by a factor δ_d starting from a minimum distance d_0. It is defined as

$$D(p,i) = \exp\left(-\frac{f\left(\|\boldsymbol{s}_i - \boldsymbol{p}_p\|\right)^2}{\delta_d}\right), \tag{2.11}$$

$$f(d) = \max\left\{0, d - d_0\right\}.$$

The distance map $\mathbf{D}$ is the arrangement of $D(p,i)$ evaluated for all sample points $i = 1...G$ and perspectives $p = 1...P$.

$$\mathbf{D} = \begin{bmatrix} D(1,1) & \dots & D(P,1) \\ \vdots & \ddots & \vdots \\ D(1,G) & \dots & D(P,G) \end{bmatrix}. \tag{2.12}$$

The distance map $\mathbf{D}$ is applied to $\tilde{\mathbf{W}}$

$$\mathbf{W} = \tilde{\mathbf{W}} \circ \mathbf{D}. \tag{2.13}$$

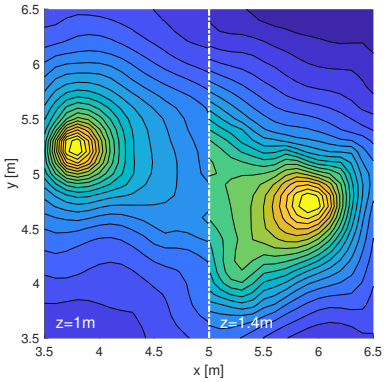
The denotation "$\circ$" describes the *Hadamard* product which is the element-wise multiplication of two $M \times N$ matrices, e.g. $\mathbf{C} = \mathbf{A} \circ \mathbf{B}$ where $C_{ij} = A_{ij}B_{ij}$ for all $i = 1...M, j = 1...N$.

Acoustic Activity Value Computation

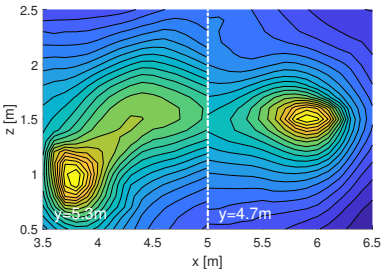
Row-wise computation of l-norms of the distance weighted activity map $\mathbf{W}$

$$\gamma_i = \left[\sum_{j=1}^{P} (W_{ij})^l\right]^{\frac{1}{l}} \qquad \text{for } i = 1...G \tag{2.14}$$

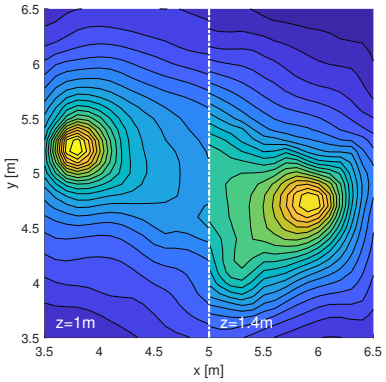
yields the acoustic activity values for all sample positions as the $G \times 1$ vector $\boldsymbol{\gamma}$. In practice, norms close to one, $0 \ll l \leq 1$, proved to be best suitable.



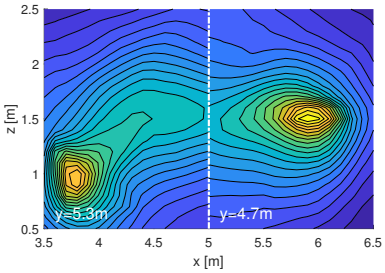
(a) Raw.



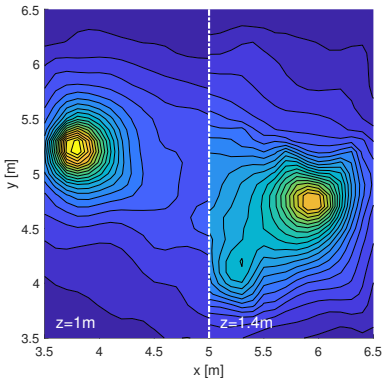
(b) Raw.



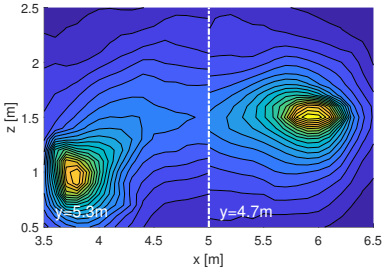
(c) Effect of the distance map with $\delta_d = 12$ and $d_0 = 0m$ shown on horizontal planes.



(d) Effect of the Distance map with $\delta_d = 12$ and $d_0 = 0m$ shown on vertical planes.



(e) Effect of the distance map with $\delta_d = 2$ and $d_0 = 0m$ shown on horizontal planes.



(f) Effect of the Distance map with $\delta_d = 2$ and $d_0 = 0m$ shown on vertical planes.

Figure 2.4: The distance map is applied and results in more focused peaks for sound objects. The distance factor δ_d should be chosen with regards to the perspective distribution in use. All data visualized here is normalized to a maximum of 1.

Chapter 3

Sound Object Tracking

Objects in the scene are tracked by *particle filters*, which are a proven method of temporal object tracking from noisy observations of various types of measurement such as distance measurements in robotics or in this case directional information of the perspectives.

Derived from [VMR07] and [KG18], the observations are evaluated by applying transitional probabilities to already tracked source positions as well as the probabilities of new sources and false positive observations. The particle filters for tracking are managed by a probabilistic *birth-death* algorithm.

The algorithm and particle-filter tacking of sound objects are explained in detail in this chapter.

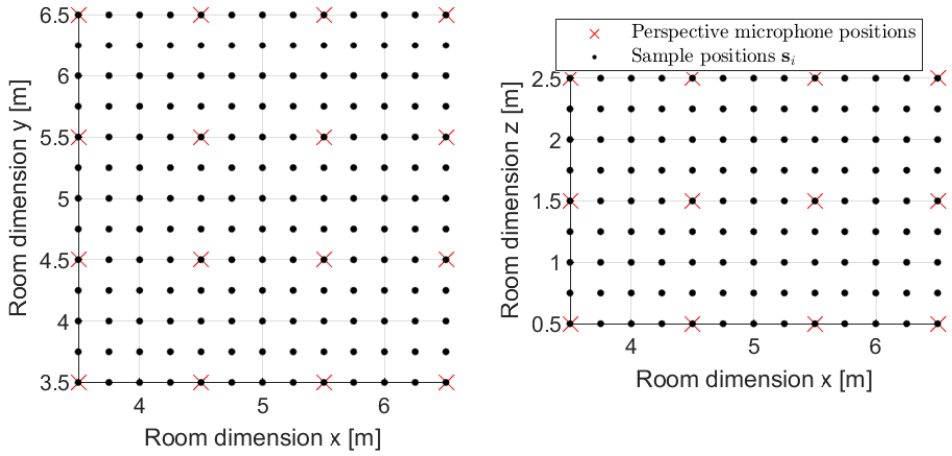
3.1 Observations of Acoustic Activity

Multiple particle filters are used to track sound objects via peaks detected in the acoustic activity map. However, it is necessary to introduce a method to initialize and remove particle filters depending on observations of detected sound objects. Objects can appear or disappear in a dynamic sound scene over time, thus an algorithm is introduced to detect such events.

The acoustic activity map introduced in Sec. 2.2 represents the instantaneous sound object activity in the scene and it is used by a probabilistic *birth-death* algorithm that detects the number of potential sound objects and determines the number of required particle filters.

The concept of particle filter application in this work will be explained in Sec. 3.3.

The target space is sampled by an equidistant grid in three dimensions, yielding the grid positions $\mathbf{g}_i = \begin{bmatrix} g_{\mathrm{x},i} & g_{\mathrm{y},i} & g_{\mathrm{z},i} \end{bmatrix}^{\mathrm{T}}$ where $i = 1...G$ is the number of grid points. Along this grid, the activity map is evaluated and used subsequently. Since grid and perspective positions are static, decoding matrices of Eq. (2.7) that are required for the activity map



(a) X-Y (b) X-Z

Figure 3.1: The relevant parts of the sound scene is sampled by a equidistantly spaced three-dimensional grid.

can be pre-computed for all perspectives. Furthermore, the grid-direction mapping and distance is static as well, so also the distance map of Sec. 2.2.2 can be pre-computed.

3.1.1 Activity Peaks

The acoustic activity values γ_i after Eq. (2.14) have to be analysed for peaks. Due to the changing map whenever a detected peak is to be removed, there is no analytical solution to the problem of finding global and local maxima. Therefore a different strategy has to be employed. The procedure introduced here is a greedy sequential algorithm involving a grid search for the global maximum and a successive deletion of components associated with each peak detection.

The index j of the maximum of all γ_i is used to get the grid position $\boldsymbol{g}_i$. j is the argument of the maximizer

$$j = \arg\max_i \gamma_i, \quad \text{for } i = 1...G. \tag{3.1}$$

The grid position and the activity values will be denoted as $\boldsymbol{o}_q = \boldsymbol{g}_j$ and $\Gamma_q = \gamma_j$ respectively. Due to this, a sequential method of de-emphasizing and peak picking is applied, resulting in a set of Q observations per time instant m, denoted as $O = \begin{bmatrix}\boldsymbol{o}_1 & \dots & \boldsymbol{o}_Q\end{bmatrix}$.

For each observed time frame m, the procedure is repeated unless the successive peaks are unable to stay within a certain threshold of $L_{\text{o,dB}}$ compared to the first observation $\Gamma_1^{(m)}$.

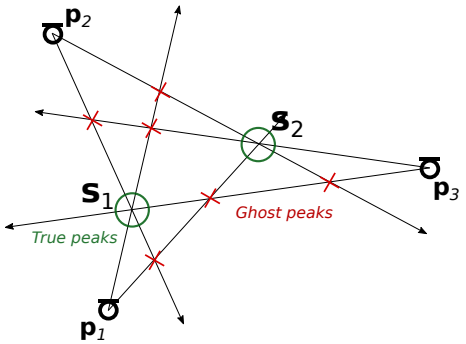


Figure 3.2: Combining perspective directional information, here simplified as one line per sound object and perspective, can yield incorrect peaks in addition to the true peaks. This is alleviated with (1) the distance map from Sec. 2.2.2 and (2) the peak deletion procedure from Sec. 3.1.2

.

3.1.2 Peak Deletion

Ghost peaks appear when multiple sound objects and perspectives are in such positions that the combination of directional information gives more peaks than objects, best explained by Fig. 3.2. To avoid such peaks, in addition to the distance map (cf. Sec. 2.2.2), a peak deletion algorithm is introduced.

Once a maximum has been picked, the directional distributions $\boldsymbol{\chi}_p$ for all perspectives are modified to exclude directional values corresponding to this maximum. A similar concept is described in [BOS08, BOS10] and especially [Hac15] which uses discrete directional histograms and Gaussian filters for the removal of peaks. Here, the method relies on peak deletion in the SH domain.

To achieve the removal of SH components corresponding to a direction pointing to a selected grid position $\boldsymbol{g}_i$, a subtraction from the corresponding distribution $\boldsymbol{\chi}$ has to happen for each perspective. To fully remove the directional information belonging to the peak, the distribution $\boldsymbol{\chi}$ has to be zeroed at the direction associated with the peak. Doing this in the SH domain requires an evaluation of the initial value regarding that direction and subtracting correspondingly. The idea is outlined in the subsequent derivation.

Directional Peak Deletion

In a continuous-direction pattern, here denoted as $x(\boldsymbol{\theta})$, we want to zero the value for a certain direction $\boldsymbol{\theta}_0$ by subtracting another pattern $g(\boldsymbol{\theta})$ yielding a modified directional pattern $\tilde{x}(\boldsymbol{\theta})$,

$$\tilde{x}(\boldsymbol{\theta}) = x(\boldsymbol{\theta}) - x(\boldsymbol{\theta}_0)g(\boldsymbol{\theta}).$$

The modified pattern $\tilde{x}(\boldsymbol{\theta})$ has to vanish for $\boldsymbol{\theta}_0$

$$\tilde{x}(\boldsymbol{\theta}_0) = x(\boldsymbol{\theta}_0) - x(\boldsymbol{\theta}_0)g(\boldsymbol{\theta}_0) = 0. \tag{3.2}$$

This is satisfied whenever the pattern $g(\boldsymbol{\theta})$ is 1 at $\boldsymbol{\theta}_0$,

$$g(\boldsymbol{\theta}_0) = 1.$$

To further look at this, the arbitrary directivity pattern $\boldsymbol{G}$ is introduced and the following transformations into the SH domain are performed:

$$\tilde{x}(\boldsymbol{\theta}) = \boldsymbol{y}(\boldsymbol{\theta})^{\mathrm{T}}\tilde{\boldsymbol{\chi}}, \tag{3.3}$$
$$x(\boldsymbol{\theta}) = \boldsymbol{y}(\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{\chi}, \tag{3.4}$$
$$g(\boldsymbol{\theta}) = \boldsymbol{y}(\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{G}. \tag{3.5}$$

The orthogonality of the SHs gives the identity,

$$\int_{\boldsymbol{\theta}\in S^2} \boldsymbol{y}(\boldsymbol{\theta})\boldsymbol{y}(\boldsymbol{\theta})^{\mathrm{T}}\mathrm{d}\boldsymbol{\theta} = \mathbf{I}.$$

The subtraction of the pattern $\boldsymbol{G}$ from the original $\boldsymbol{\chi}$ can be denoted as

$$\tilde{\boldsymbol{\chi}} = \boldsymbol{\chi} - x(\boldsymbol{\theta}_0)\boldsymbol{G},$$

and using Eq. (3.4), $x(\boldsymbol{\theta}_0)$ can be substituted and results in the modification step of the directional distribution $\boldsymbol{\chi}$:

$$\tilde{\boldsymbol{\chi}} = \left(\mathbf{I} - \boldsymbol{G}\boldsymbol{y}(\boldsymbol{\theta}_0)^{\mathrm{T}}\right)\boldsymbol{\chi}. \tag{3.6}$$

Since the pattern $\boldsymbol{G}$ to suppress the peak is arbitrary, it can be chosen to be an order-weighted version of $\boldsymbol{y}(\boldsymbol{\theta}_0)$, $\boldsymbol{G} = \mathrm{diag}\left\{\tilde{\mathbf{a}}\right\}\boldsymbol{y}(\boldsymbol{\theta}_0)$ to adjust its width. With this weighting, Eq. (3.6) turns into the application of a peak deletion matrix

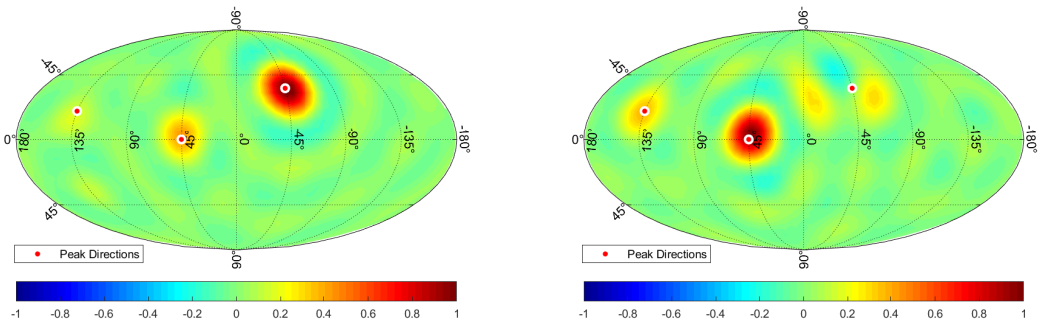
$$\tilde{\boldsymbol{\chi}} = \left(\mathbf{I} - \mathrm{diag}\left\{\tilde{\mathbf{a}}\right\}\boldsymbol{y}(\boldsymbol{\theta}_0)\boldsymbol{y}(\boldsymbol{\theta}_0)^{\mathrm{T}}\right)\boldsymbol{\chi}. \tag{3.7}$$

Normalization: Due to the condition Eq. (3.2) for zeroing at $\boldsymbol{\theta}_0$, now written as

$$\boldsymbol{y}(\boldsymbol{\theta}_0)^{\mathrm{T}}\left(\left(\mathbf{I} - \mathrm{diag}\left\{\tilde{\mathbf{a}}\right\}\boldsymbol{y}(\boldsymbol{\theta}_0)\boldsymbol{y}(\boldsymbol{\theta}_0)^{\mathrm{T}}\right)\boldsymbol{\chi}\right) = 0$$

we can see that there is a restriction on $\tilde{\mathbf{a}}$. Simplifying this yields

$$\boldsymbol{y}(\boldsymbol{\theta}_0)^{\mathrm{T}}\mathrm{diag}\left\{\tilde{\mathbf{a}}\right\}\boldsymbol{y}(\boldsymbol{\theta}_0) \overset{!}{=} 1,$$



(a) Original (peak at -45,-40)

(b) After peak deletion (peak at 45,0)

Figure 3.3: The SH deletion function removes parts of the SH directional distribution. On the left it shows a 8th order SH directional distribution. 3 distinct peaks from sound sources are visible. The values are normalized to a maximum value of 1. The right depicts two remaining peaks after deletion with $\boldsymbol{\theta}_0$ towards a peak direction and a constant order weight of $a_i = \frac{4\pi}{N}$, $i = 1...(N+1)^2$ and re-normalization. Both are *Mollweide* projections [Sny87] decoded to 1000 equally distributed sample directions for visualization.

which prompts that any $\tilde{\mathbf{a}}$ can be normalized to be satisfactory by normalization as follows:

$$\mathbf{a} = \frac{\tilde{\mathbf{a}}}{\boldsymbol{y}(\boldsymbol{\theta}_0)^{\mathrm{T}} \mathrm{diag}\left\{\tilde{\mathbf{a}}\right\} \boldsymbol{y}(\boldsymbol{\theta}_0)}. \tag{3.8}$$

Whilst this deletion function assures a value of zero at the direction $\boldsymbol{\theta}_0$, negative values can result from this operation at neighbouring locations $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, depending on $\mathbf{a}$, i.e. the width of the deletion pattern. These negative values generally do not pose an issue when the $\mathbf{a}$-order-weighted SH pattern is kept narrow enough. Furthermore, negative values do not affect the result in any impeding way, as the combination of perspectives stays monotonically increasing towards peak values.

Algorithm Summary

A complete run of the activity peak picking procedure in an observed time frame m is summarized in Algorithm 3.1. The sequential peak picking and deletion yields global and local maxima.

Algorithm 3.1: Peak picking and deletion

Data: Perspective directional distributions $\boldsymbol{\chi}_p^{(m)}$, grid points $\boldsymbol{g}_g$, level threshold $L_{\text{o,dB}}$, maximum number of observations Q

Result: Set of observation coordinates $O^{(m)} = \begin{bmatrix}\boldsymbol{o}_1 & \dots & \boldsymbol{o}_Q\end{bmatrix}$ and observation values Γ_q

(3) **while** $\Gamma_q \geq \Gamma_1 \cdot 10^{\frac{L_{o,dB}}{20}}$ *AND* $q \leq Q$ **do** `/* as long as above level threshold and max. number of observations not reached */`

(1) Combination of directional information into map according to Eq. (2.14):

$$\gamma_i = \left[\sum_{j=1}^{P} (W_{ij})^l\right]^{\frac{1}{l}} \qquad \text{for } i = 1...G.$$

(2) Pick the maximum and get grid position (Sec. 3.1.1):

$$\Gamma_q = \max_i \gamma_i^{(m)}, \quad \text{for } i = 1...G.$$

$$\boldsymbol{o}_q = \boldsymbol{g}_j, \quad \text{where } j = \arg\max_i \gamma_i^{(m)}, \quad \text{for } i = 1...G.$$

foreach $p = 1...P$ **do** `/* Peak deletion for all perspectives */`

(a) Calculate direction vector θ_p from perspective $\boldsymbol{p}_p$ to the observation $\boldsymbol{o}_q$.

(b) Apply peak deletion to $\boldsymbol{\chi}_p$ in direction θ_p (Eq. (3.7)).

end

end

3.2 Validation and Detection Algorithm

The instantaneous observations from the sequential peak picking procedure have to be evaluated for their viability. This is achieved by assigning *transitional* probabilities between observations and probabilistic hypotheses for each possible state including current tracked sound objects, new objects as well as false detections. This probabilistic procedure is modified from [VMR07, KG18], while the basic concept is related to the field of *Signal Detection Theory* (cf. [MC04]). Fig. 3.4 gives a simplistic overview.

3.2.1 Transitional Probabilities

All instantaneous observations that are potential sound object locations for time instant m are denoted as $O^{(m)} = \begin{bmatrix}\boldsymbol{o}_1 & \dots & \boldsymbol{o}_Q\end{bmatrix}$. The set of all observations up until the current time is denoted as $\boldsymbol{O}^{(m)}$. To express how likely an observation is a true sound object and not a false detection, the probability P_q is introduced.

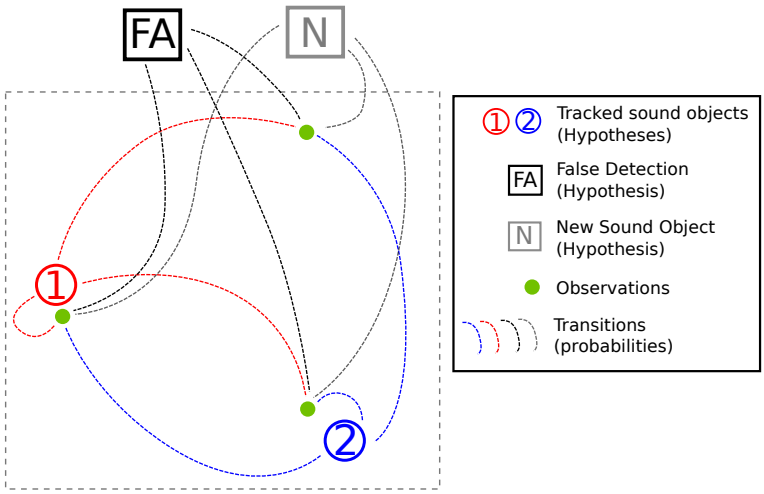


Figure 3.4: Illustration of transitional probabilities. The instantaneous observation magnitudes express probability values controlling their potential of getting instantiated or continued as sound object or on the other hand getting identified as a false detection. The sum over all classification probabilities has to fulfil $\sum P = 1$.

This value is dependent on the observed peak magnitude relative to the largest, first peak Γ_1 (cf. Algorithm 3.1) via the following definition:

$$P_q = \frac{\Gamma_q}{\Gamma_1} \tag{3.9}$$

The probability of observing at the position $\boldsymbol{o}_q^{(m)}$ at the time instant m is modelled by a multivariate Gaussian distribution around a known sound object position

$$p(\boldsymbol{o}_q|\hat{\boldsymbol{x}}_s) = \mathcal{N}(\boldsymbol{o}_q, \hat{\boldsymbol{x}}_s, \boldsymbol{\Sigma}_s). \tag{3.10}$$

This known position is the estimate $\hat{\boldsymbol{x}}_s$ of a particle filter and the distribution is dependent on the covariance matrix $\boldsymbol{\Sigma}_s$, which is estimated from a sequence of particle positions (cf. Sec. 3.3)

$$\boldsymbol{\Sigma}_s = \mu \,\mathrm{Cov}\left\{\boldsymbol{x}_i\right\}. \tag{3.11}$$

The factor μ allows the covariance to be scaled. This gives us the possibility to change the detection radius of a tracked sound object on the observation algorithm. A larger detection radius means that observations in proximity of a tracked source are more likely to be regarded as associated with it and vice versa. Also, for further description of the transitional probabilities, we define the following hypotheses:

$\mathcal{H}_{\text{fa}}$: The observation is a false detection.
$\mathcal{H}_{\text{new}}$: The observation is an untracked/new sound object.
$\mathcal{H}_s$: The observation belongs to an already tracked sound object s.

These hypotheses, represent a mapping between observations and detected sound objects, false detections, and new objects. The mapping is expressed as association functions which are defined as

$$f_r(q) = \begin{cases} -2, & \mathcal{H}_{\text{fa}} \\ -1, & \mathcal{H}_{\text{new}} \\ s, & \mathcal{H}_s \end{cases}. \tag{3.12}$$

This mapping is done in a probabilistic sense, so that the conditional probability of a consistent mapping from the given observations $O^{(m)}$ is denoted as $P(f_r|O^{(m)})$. These association probabilities are given in greater detail below. All possible observation-hypothesis mapping combinations have to be evaluated for the marginal probabilities of the hypotheses. With S currently tracked sound objects and Q observations for the time instant m, this results in $r = (S+2)^Q$ possible combinations to be evaluated. The marginals, as transitional probabilities, are defined as the sum over all possible combinations

$$P_q(\mathcal{H}_{\text{fa}}) = \sum_r \delta_{-2,f_r(q)}\, P(f_r|O^{(m)}), \tag{3.13}$$

$$P_q(\mathcal{H}_{\text{new}}) = \sum_r \delta_{-1,f_r(q)}\, P(f_r|O^{(m)}), \tag{3.14}$$

$$P_q(\mathcal{H}_s) = \sum_r \delta_{s,f_r(q)}\, P(f_r|O^{(m)}). \tag{3.15}$$

Their sum is one, $P_q(\mathcal{H}_{\text{fa}}) + P_q(\mathcal{H}_{\text{new}}) + \sum_S P_q(\mathcal{H}_s) = 1$. The Kronecker delta denotes the selection from $P(f_r|O^{(m)})$ for the specific result of the association function $f_r(q)$ that numerically represents the hypothesis. To illustrate this, Fig.3.5 shows a minimal example.

For all mapping combinations r, all the existing association probabilities $P(f_r|O^{(m)})$ are defined by Bayes' rule

$$P(f_r|O^{(m)}) \propto p(O^{(m)}|f_r)P(f_r) \tag{3.16}$$

with omitted denominator as normalization can be done later on.

Assuming conditional independence of the observations for all q, the probability densities $p(O^{(m)}|f_r)$ for a specific hypothesis f_r given the particular probabilities of the individual Q observations is

$$p(O^{(m)}|f_r) = \prod_{q=1}^{Q} p(\boldsymbol{o}_q^{(m)}|f_r(q)), \tag{3.17}$$

$$\text{with } p(\boldsymbol{o}_q^{(m)}|f_r(q)) = \begin{cases} P_{\text{false}}(\boldsymbol{o}_q^{(m)}) & \text{if } f_r(q) = -2 \\ P_{\text{new}}(\boldsymbol{o}_q^{(m)}) & \text{if } f_r(q) = -1 \\ \text{use Eq. (3.10)} & \text{else} \end{cases}. \tag{3.18}$$

$P_{\text{false}}(\boldsymbol{o}_q^{(m)})$ and $P_{\text{new}}(\boldsymbol{o}_q^{(m)})$ represent knowledge of probable sound object locations prior to this whole process, for example the exclusion of or emphasis on certain volumes, be it physical obstacles or a more frequented stage area. In general, if there is no knowledge, the probability is assumed to be an equal distribution over the entire room volume, they are consequently $P_{\text{false}}(\boldsymbol{o}_q^{(m)}) = P_{\text{new}}(\boldsymbol{o}_q^{(m)}) = \frac{1}{G}$. G is the number of grid locations being used to select observations.

Further, the prior probabilities for the association functions are calculated by using the ratio P_q (cf. Eq. (3.9))

$$P(f_r) = \prod_{q=1}^{Q} p(f_r(q)), \tag{3.19}$$

$$\text{with } p(f_r(q)) = \begin{cases} P_{\text{assoc,false}}(1 - P_q) & \text{if } f_r(q) = -2 \\ P_{\text{assoc,new}} P_q & \text{if } f_r(q) = -1 \\ P_q P_{\text{obs}}^{(m)}(f_r(q)|\boldsymbol{O}^{(m-1)}) & \text{else} \end{cases}. \tag{3.20}$$

This introduces a new set of probabilities:

$P_{\text{assoc,new}}$ and $P_{\text{assoc,false}}$ are empirically set to 0.2 and 0.8 respectively and are basically parameters that fine-tune the impact of the observation viability P_q, as defined by Eq. (3.9), which in turn is dependent on acoustic activity map peak values.

The expression $P_{\text{obs}}(f_r(q)|\boldsymbol{O}^{(m-1)}) = P_{\text{obs}}(s|\boldsymbol{O}^{(m-1)})$ is the probability of a sound object s being observable at the time instant m. This is defined as the product of the *existence probability* and the *activity probability*, which represents the probability of the sound object being active regardless of being observed or not.

This existence probability is defined recursively as proposed in [VMR07] as

$$P_{\text{exist}}^{(m)}(s|\boldsymbol{O}^{(m-1)}) = P_q^{(m-1)} + (1 - P_q^{(m-1)}) \frac{P_0 P_{\text{exist}}^{(m-1)}(s|\boldsymbol{O}^{(m-2)})}{1 - (1 - P_0) P_{\text{exist}}^{(m-1)}(s|\boldsymbol{O}^{(m-2)})}. \tag{3.21}$$

Here, $P_0 = 0.5$ is the prior probability that a sound object is not observed even if it exists. This recursion evaluates the probability of existence at time instant m using the value $P_{\text{exist}}^{(m-1)}(s|\boldsymbol{O}^{(m-2)})$ of the previous time instant as well as the observation viability P_q.

In practice, whenever the existence probability reaches a certain threshold of 0.98, it is set to 1 and will not be updated further, since the existence of the sound object is considered sufficiently certain.

The activity probability is calculated involving past and current sets of observations. First, the instantaneous activity probability $P_{\text{active}}^{(m)}$ is defined as the sound object probability of the last time frame,

$$P_{\text{active}}^{(m)}(s|O^{(m-1)}) = P_s^{(m-1)}. \tag{3.22}$$

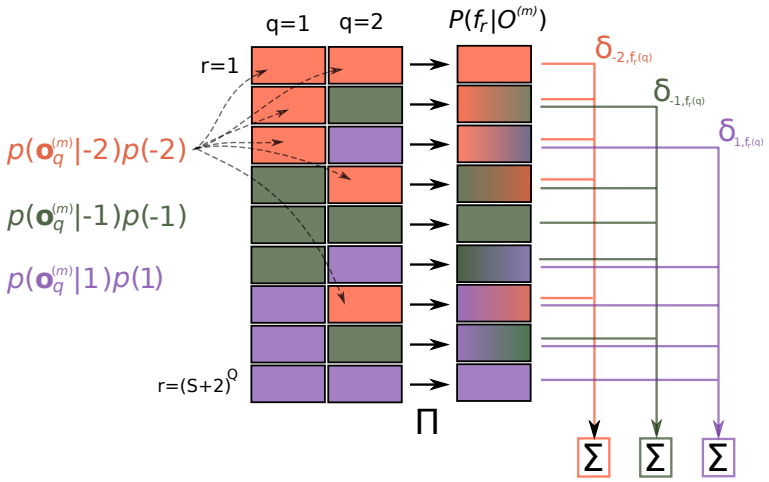


Figure 3.5: The set of mapping functions f_r is applied to all observations in all combinations of hypotheses. The $r = (S+2)^Q$ possible combinations are evaluated by computing the probabilities $p(\boldsymbol{o}_q^{(m)}|f_r(q))$, see Eq. (3.18) and the priors $p(f_r(q))$, see Eq. (3.20) for all $q = 1...Q$ and multiplication per combination. This yields probability values $P(f_r|O^{(m)})$ in Eq. (3.16) for the r combinations. At this point, the previously omitted normalization happens in discrete form through division by the sum. Then, the Kronecker deltas are used to compute a selective sum.

Secondly, an activity state model is defined as a first-order Markov process using probabilities for transitioning from (1) active state to active state and (2) from inactive state to active state again (cf. [VMR07]):

$$P_{\text{active}}^{(m)}(s|\boldsymbol{O}^{(m-1)}) = P(\text{active}|\text{active})P_{\text{active}}^{(m-1)}(s|\boldsymbol{O}^{(m-1)}) \tag{3.23}$$

$$+ P(\text{active}|\text{inactive})\left[1 - P_{\text{active}}^{(m-1)}(s|\boldsymbol{O}^{(m-1)})\right]. \tag{3.24}$$

These state transition probabilities are set to $P(\text{active}|\text{active}) = 0.95$ and $P(\text{active}|\text{inactive}) = 0.05$. For equiprobable active and inactive states, the probability for the current set of observations is provable to be

$$P_{\text{active}}^{(m)}(s|\boldsymbol{O}^{(m)}) = \left(1 + \frac{\left[1 - P_{\text{active}}^{(m)}(s|\boldsymbol{O}^{(m-1)})\right]\left[1 - P_{\text{active}}^{(m)}(s|O^{(m)})\right]}{P_{\text{active}}^{(m)}(s|\boldsymbol{O}^{(m-1)})P_{\text{active}}^{(m)}(s|O^{(m)})}\right)^{-1}. \tag{3.25}$$

Finally, the algorithm described in this section yields the following values:

$P_q^{(m)}(\mathcal{H}_{\text{fa}})$ is the probability that the observation $\boldsymbol{o}_q^{(m)}$ is a false detection.

$P_q^{(m)}(\mathcal{H}_{\text{new}})$ is the probability that the observation $\boldsymbol{o}_q^{(m)}$ is a new sound object.

$P_s^{(m)} = \sum_{q=1}^{Q} P_q^{(m)}(\mathcal{H}_s)$ is the probability that the sound object s has been observed at time instant m.

3.3 Object Tracking

The procedure introduced in Sec. 3.1 yields a set of observations O and the algorithm from Sec. 3.2 maps these to existing sound objects, detects new and false object observations and yields transitional probabilities for all these hypotheses. The probabilities will be used to manage a set of particle filters, by instantiating and removal, as explained later in Sec. 3.3.2.

First however, since the observations are instantaneous for m, most certainly noisy and do not necessarily yield time-continuous trajectories for sound objects, particle filters are introduced to track objects. Moreover, these particle filters do not use the set of observations for estimation but instead directly use the acoustic activity map, as introduced in Sec. 2.2.

3.3.1 Particle Filters

Particle filtering is a method of numerical integration to track objects from noisy observations. It is part of the Monte-Carlo type analysis methods which replace closed form computation of statistical quantities with drawing samples from distributions and estimating by sample averages. Important literature on the basic concepts is [Efr93, LC95, Fea98, Kit96, LC98, Kit96, CCF99, DdFG01, Sär13].

When a probability density function is not known directly then calculation of statistical quantities is not possible. In the context of this application, the density function is the three-dimensional probability density of sound objects $p_{obj}(\boldsymbol{x})$ and unknown. However, the acoustic activity map can be used to compute approximations of such statistical quantities, e.g. the mean. For now, assume exactly one active sound object and therefore one maximum in the acoustic activity map.

We represent the sound object as a set of N particles, each represented by a state vector holding position $\boldsymbol{x}_i$ and velocity $\dot{\boldsymbol{x}}_i$

$$\boldsymbol{s}_i = \begin{bmatrix} x_i & y_i & z_i & \dot{x}_i & \dot{y}_i & \dot{z}_i \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{x}_i \\ \dot{\boldsymbol{x}}_i \end{bmatrix}. \tag{3.26}$$

To approximate the density function, the particle set $\boldsymbol{S}^{(m)} = \begin{bmatrix} \boldsymbol{s}_1 & \dots & \boldsymbol{s}_N \end{bmatrix}$ is introduced as well as and the weights $\boldsymbol{q}^{(m)} = \begin{bmatrix} q_1 & \dots & q_N \end{bmatrix}^{\mathrm{T}}$, which express the *importance* of each particle. The particles are sampled from a known distribution, e.g. a uniform distribution. Each particle $\boldsymbol{s}_i$ represents a hypothesis of sound object. The likelihood of this hypothesis is expressed by its weight q_i, which is calculated in the *importance sampling* (IS) step. For a complete derivation of the probabilistic framework behind this step, refer to [Sär13]. For this application, the computation is done by evaluating the acoustic activity, as defined by

Eq. (2.14), at the N particle positions $\boldsymbol{s}_i$ followed by the normalization of the sampled activity values

$$q_i = \frac{\gamma_i}{\sum_{j=1}^{N} \gamma_j} \qquad \text{for } i = 1...N. \tag{3.27}$$

The weighted sum over the set of N particles $\boldsymbol{S}$ is

$$\hat{\boldsymbol{s}} = \begin{bmatrix} \hat{\boldsymbol{x}} \\ \hat{\dot{\boldsymbol{x}}} \end{bmatrix} = \boldsymbol{S}\boldsymbol{q}. \tag{3.28}$$

The position $\hat{\boldsymbol{x}}$ is the estimate of the mean of the probability density of sound objects $p_{obj}(\boldsymbol{x})$. Since we assumed one active sound object for now, this holds true, however if more object peaks are present in the acoustic activity map then this does not represent a valuable estimation. This problem is addressed by limiting particle sampling to local maximums. This concept is visualized in Figure 3.6b.

In case prior knowledge of sound object distribution is available, the sampling of particle positions can be more intricate. This is the case when a new object is detected by the algorithm introduced in Sec. 3.2. Whenever this happens, a multivariate Gaussian distribution with the mean at the observation position $\boldsymbol{o}_q$ is used to sample the initial particle state vectors

$$\boldsymbol{s}_i \sim \mathcal{N}(\boldsymbol{o}_q, \boldsymbol{\Sigma}_{\text{inst}}). \tag{3.29}$$

The covariance matrix of this distribution is defined depending on the grid spacing d_{grid} of the observation picking algorithm

$$\text{with } \boldsymbol{\Sigma}_{\text{inst}} = \text{diag}\left\{\begin{bmatrix} d_{\text{grid}}^2 & d_{\text{grid}}^2 & d_{\text{grid}}^2 & 0 & 0 & 0 \end{bmatrix}\right\}$$

Since this initial distribution of particles is now centred around this new presumed sound object, the approximation is not representing the entire object density function but just sampling the local maximum. The mean $\hat{\boldsymbol{x}}$ yields now a good estimation of this local maximum position and further allows us to estimate multiple of those.

If a physical sound object moves then the estimation loses accuracy as the activity peak moves away from the high but static sampling density. Therefore a state-space model is introduced. We can predict the state of a particle from its previous state

$$\boldsymbol{s}_i^{(m)} = \mathbf{M}\boldsymbol{s}_i^{(m-1)}$$

by applying the system matrix

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 & \Delta t & 0 & 0 \\ 0 & 1 & 0 & 0 & \Delta t & 0 \\ 0 & 0 & 1 & 0 & 0 & \Delta t \\ 0 & 0 & 0 & a_{\text{dyn}} & 0 & 0 \\ 0 & 0 & 0 & 0 & a_{\text{dyn}} & 0 \\ 0 & 0 & 0 & 0 & 0 & a_{\text{dyn}} \end{bmatrix}.$$

However, the unpredictability of real dynamic systems requires the addition of process noise which is modelled as a multivariate Gaussian distribution

$$\mathbf{Q}^{(m)} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\text{process}}) \tag{3.30}$$

to the model which becomes

$$\boldsymbol{s}_i^{(m)} = \mathbf{M}\boldsymbol{s}_i^{(m-1)} + \mathbf{Q}^{(m)}. \tag{3.31}$$

This process noise only affects the velocities of particles, as seen in the definition of the covariance matrix

$$\text{with } \mathbf{\Sigma}_{\text{process}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & b_{\text{dyn}} & 0 & 0 \\ 0 & 0 & 0 & 0 & b_{\text{dyn}} & 0 \\ 0 & 0 & 0 & 0 & 0 & b_{\text{dyn}} \end{bmatrix}.$$

This model is called the excitation-damping dynamical model [WLW03a] and uses two factors defined as

$$a_{\text{dyn}} = e^{-\alpha_{\text{dyn}}\Delta t}, \tag{3.32}$$

$$b_{\text{dyn}} = \beta_{\text{dyn}}\sqrt{1 - a_{\text{dyn}}^2}. \tag{3.33}$$

They factors for α_{dyn} and β_{dyn} can be defined for different dynamic scenarios. The first factor influences the damping of the velocity of existing particles, the second one the variance of the process noise. For the given problem, the chosen values were $\alpha_{\text{dyn}} = 2$ and $\beta_{\text{dyn}} = 0.04$. The time step Δt is the time between prediction steps which is the length of a time frame in seconds. Each time instant m, the prediction is done first before weight computation. This procedure is called *sequential importance sampling* (SIS) in literature such as [Sär13].

Finally, *importance resampling*, the replacement of low-weight particles with high-weight

particles, is introduced. This is necessary to solve the *degeneracy problem* which can lead into situations where almost all particles have zero or close to zero weights. The resampling can be done by methods such as multinomial (cf. [Efr93, CCF99]), residual (cf. [Whi98, LC98]), stratified (cf. [Kit96, Fea98]) and systematic resampling (cf. [Whi98, CCF99]) which is the approach which proved effective for this application.

A simplified one-dimensional example of the entire procedure being, in literature referred to as *sequential importance resampling* (SIR), is shown in Figure 3.6a and applied to this application listed in Algorithm 3.2.

The end result and purpose of applying particle filters in this work is a set of S position estimates which are the Cartesian coordinates $\hat{\boldsymbol{x}}_s$ gathered in the weighted particle mean computed using Eq. (3.28) for each detected sound object.

Algorithm 3.2: Sequential importance resampling

On particle filter initialization, draw N samples $\boldsymbol{s}_i$, Eq. (3.29):

$$\boldsymbol{s}_i \sim \mathcal{N}(\boldsymbol{o}_q, \boldsymbol{\Sigma}_{\text{inst}}), \quad i = 1, ..., N;$$

set $q_i^{(0)} = \frac{1}{N}$.

for $m = 1$ **to** M **do** `/* for all time steps` M `*/`

(2) Process noise, Eq. (3.30):

$$\mathbf{Q}^{(m)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\text{process}})$$

(2) Prediction step, Eq. (3.31):

$$\boldsymbol{s}_i^{(m)} = \mathbf{M}\boldsymbol{s}_i^{(m-1)} + \mathbf{Q}^{(m)}, \quad i = 1, ..., N.$$

(3) Acoustic map evaluation, Eq. (2.14) / weight calculation:

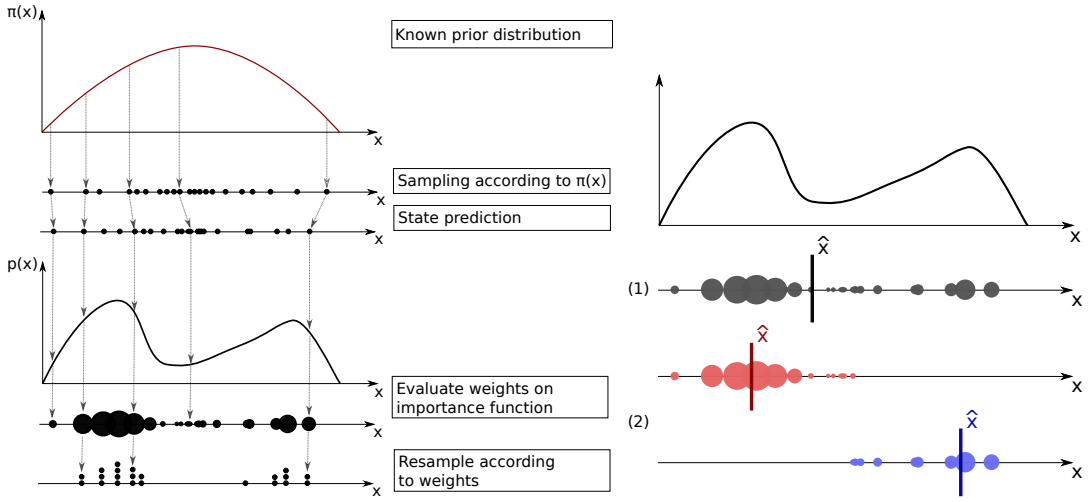
$$q_i = \frac{\gamma_i}{\sum_{j=1}^{N} \gamma_j} \qquad \text{for } i = 1...N.$$

(4) Particle mean / position estimation, Eq. (3.28):

$$\hat{\boldsymbol{s}} = \boldsymbol{S}\boldsymbol{q}$$

(5) Systematic particle resampling, [Whi98, CCF99].

end



(a) Sequential Importance Resampling (b) (1) global mean, (2) local means

Figure 3.6: (a) The process of SIR as explained in Sec. 3.3. The steps pictured are repeated for each time frame m. (b) The weighted particle means of locally limited sample groups (2) are good estimates of local peaks in a density function compared to global sampling (1).

3.3.2 Particle Filter Management

As described in the previous section, the particle filter tracks single sound objects using the acoustic activity map. Therefore, increasing the number of sound objects to track requires the same increase in the number of particle filters which in turn necessitates a management system for particle filters. For each time instant m, a set of observations and transition probabilities is computed as described in Sec. 3.2.1. These probabilities describe the association between observations and tracked sound objects. The probability of an observation being a new sound object $P_q^{(m)}(\mathcal{H}_{\text{new}})$ is used to instantiate a new particle filter when certain conditions are met. Similarly, particle filters that are tracking sources with a low probability for continuation $P_s^{(m)}$ are removed from calculations.

Instance Pool

To manage the resources of the systems, a pool of particle filters is used. This entails a set number of pre-allocated particle filters to be used when necessary. When a new instance is required, the first free space in the pool is selected. In case no pre-allocated instances are available, the request is ignored. The size of the pool should therefore be chosen so it roughly reflects the maximum number of sound objects expected to be active concurrently.

Initialization

Whenever the transitional probability $P_q^{(m)}(\mathcal{H}_{\text{new}})$ of an instantaneous observation reaches a certain threshold, e.g. 0.8, a new particle filter is requested from the instance pool. The particles of the filter are initially sampled from a three-dimensional Gaussian distribution with its mean at the observation $\boldsymbol{o}_q$ and a variance in relation to the grid spacing as was explained in Sec. 3.3.1. The sound object probability $P_s^{(m)}$ is set to $P_q^{(m)}(\mathcal{H}_{\text{new}})$. The sound object is marked as active when the sound object probability $P_s^{(m)}$ is above a threshold for a certain duration t_{active}, e.g. 0.1 seconds, to avoid spurious detections.

Deletion

Inversely to the activation, when $P_s^{(m)}$ stays or falls below the threshold for a certain duration t_{inactive}, e.g. 0.5, it is regarded as inactive and the particle filter is deactivated and its resources in the instance pool are freed.

3.4 Anti-causal Processing of Onsets

Since the algorithm needs to ensure certainty of object existence before beginning tracking, it can happen that objects are detected a few frames later that they actually appear. This could mean that important transients or onsets would be lost. Therefore, parts of the process are repeated as an anti-causal process. This restores the timely tracking of sound objects right from the moment on they physically appear in the scene and therefore improves the audio quality in playback. The acoustic activity map values of the causal processing can be stored and reused to save computation time.

For each detected sound object in the scene being analysed, the anti-causal processing starts in frames in which a new particle filter becomes active. For the anti-causal processing, the sequence of time instants m is reversed from $(m) \rightarrow (m+1)$ to $(m) \rightarrow (m-1)$ and the detection algorithm as well as the dynamic system require some changes outlined below.

3.4.1 Anti-causal Detection Algorithm

The algorithm is executed in the anti-causal time direction of acoustic map values in the same manner as on the causal direction, with the notable change of switching the durations t_{active} and t_{inactive}. Since this is used for the continuation of sound object trajectories which have been detected already, the decision operation for the detection of new sound objects is deactivated and $P_q^{(m)}(\mathcal{H}_{\text{new}})$ remains unused.

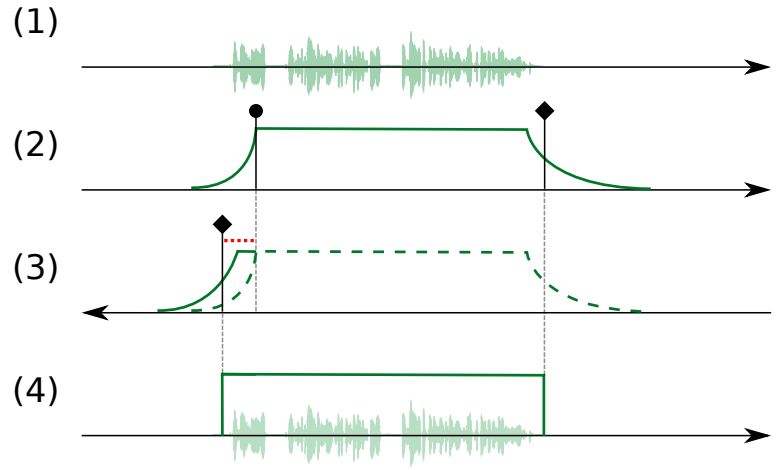


Figure 3.7: Causal and anti-causal processing. (1) A sound object is active. (2) Causal processing which yields one object detection event (○) and one object inactivation event (◊). (3) The detection event determines the starting frame of the anti-causal processing, which is active until the objects is considered inactive. The anti-causal processing interval is marked with a dotted red line. (4) This greatly improves the detection of object signals in regards to onsets.

This leads to the possibility that anti-causal computation only occurs whenever necessary. Hence, only time instants are considered at which detection events are located by the causal processing, and the anti-causal processing indicates that the sound objects is still active. This is visualized in Figure 3.7 as a dotted red line.

3.4.2 Anti-causal Particle Filter Prediction

The prediction step of the particles dynamic model, as defined by Eq. 3.31, can simply be reversed in time by changing the sign of the frame time Δt in the system matrix $\mathbf{M}$. The prediction, computation of particle weights and position estimation are only done for times instants m at which the causal processing has not instantiated the object as active yet.

Chapter 4

Six-Degrees-of-Freedom Rendering

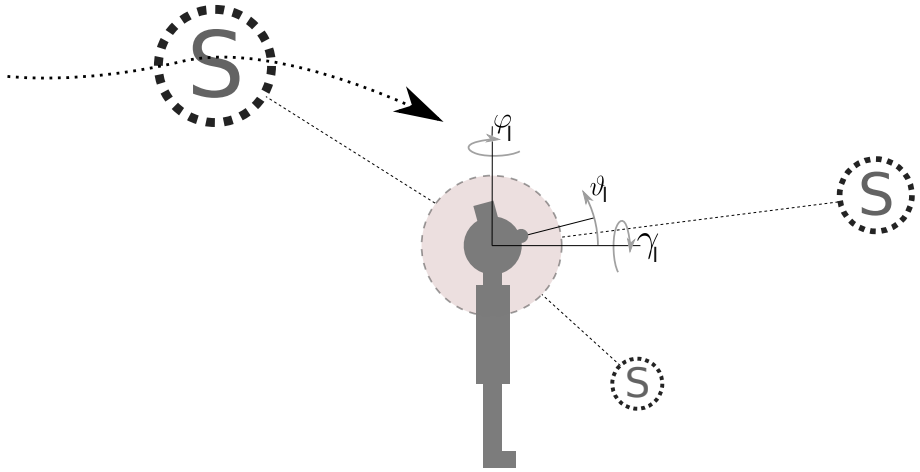


Figure 4.1: The sound objects have to be encoded regarding their relative position to the listener perspective. This happens through the concept of Ambisonics which enables easy methods of rotation and decoding to different playback devices. The listeners head position and rotation is tracked and used to compose the presented perspective.

After analysis of the sound scene using the perspective recordings, the information retrieved is used to recreate the sound scene for a virtual listener as closely to the original as possible. A listener should be able to move freely in all three dimensions including a free rotation of their perspective. This is facilitated by composing a listener perspective in the Ambisonics domain. The sound objects are encoded and attenuated depending on the relative position to the virtual listener.

To facilitate audio rendering, the virtual listener head position and head rotation is required. First one, further referred to as *listener position* and denoted as the three dimension column vector $\boldsymbol{p}_\mathrm{l}$, as well as the 3×3 head rotation matrix $\boldsymbol{R}_\mathrm{l}$ can be made available through a data stream of various kinds. Numerous types of devices such as dedicated head trackers or head-mounted displays of commercial VR hardware are capable of delivering

this data.

The target auditory perspective is represented as a fifth-order Ambisonics signal rendered for the listener position. This perspective is rotated to fit the listeners head rotation and decoded to a binaural listening device, i.e. headphones.

4.1 Sound Object Encoding

The main non-audio information characterizing sound objects in the sound scene is its trajectory. Its analysis and tracking was described above. Since the audio information is available only at the perspective positions in form of the microphone recordings of the tetrahedral arrays, signals to be embedded as sound objects have to be extracted.

This is done via a *minimum-variance distortionless response* (MVDR) beamformer using available positional information to separate the sound object signals.

The following sections explain the method to extract perspective-wise *sound object signals*

$$s_{\text{obj},s}(t)$$

from the *array microphone signals*

$$s_{p,i}(t)$$

and combine them into the *approximated direct signals*

$$\bar{s}_s(t).$$

4.1.1 Object Signal Extraction with Mixed MVDR

The mixed MVDR approach is a beamforming technique that is based on a combination of minimizing the overall energy of the beam towards a steering vector and minimizing

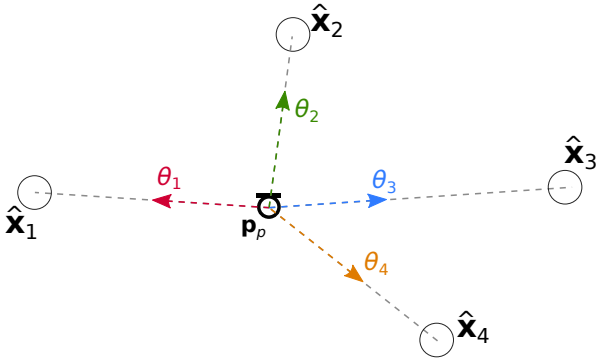


Figure 4.2: One perspective and four sound objects.

the influence of certain disturbing signals whose directions are known.

Taking into account only one of the perspectives indexed as p with its position $\boldsymbol{p}_p$, we can calculate unit direction vectors for the estimated sound object positions $\hat{\boldsymbol{x}}_i$ (cf. Eq. (3.28)), as shown in Fig. 4.2, as

$$\boldsymbol{\theta}_i(t) = \frac{\hat{\boldsymbol{x}}_i(t) - \boldsymbol{p}_p}{\|\hat{\boldsymbol{x}}_i(t) - \boldsymbol{p}_p\|}. \tag{4.1}$$

We assume four distinct $s_{\text{obj},i}(t)$ arriving at the microphone array from the directions $\boldsymbol{\theta}_i$, both with $i = 1...4$. The SH domain lets us describe this as the following:

$$\boldsymbol{\psi}(t) = \sum_{i=1}^{4} \boldsymbol{y}(\boldsymbol{\theta}_i) s_{\text{obj},i}(t) = \mathbf{Y}\boldsymbol{s}_{\text{obj}}(t). \tag{4.2}$$

However, since the object signals are unknown, they have to be extracted from the SH signals which are first-order Ambisonics in our case, and which are characterized by the four array microphone signals $s_{p,j}(t)$, $j = 1...4$ for perspective p, encoded in their array directions $\boldsymbol{\theta}_{\text{c}_i}$ (cf. Eq. (1.1)), which can be denoted as

$$\boldsymbol{\psi}(t) = \sum_{j=1}^{4} \boldsymbol{y}(\boldsymbol{\theta}_{\text{c}_i}) s_{p,j}(t). \tag{4.3}$$

Now, trying to extract the object signal arriving from the direction $\boldsymbol{\theta}_1$, we define a signal estimation by left multiplication of a combination vector $\boldsymbol{g}_1$ with Eq. (4.2):

$$s_{\text{obj},1}(t) \approx \boldsymbol{g}_1^{\text{T}} \mathbf{Y} \boldsymbol{s}_{\text{obj}}(t). \tag{4.4}$$

To find the weights in this vector, we formulate a minimization problem:

$$\text{minimize } a\, \boldsymbol{g}_1^{\text{T}} \boldsymbol{g}_1 + b\, \boldsymbol{g}_1^{\text{T}} \mathbf{Y}\mathbf{Y}^{\text{T}} \boldsymbol{g}_1, \qquad \text{subject to } \boldsymbol{g}_1 \boldsymbol{y}(\boldsymbol{\theta}_1) = 1. \tag{4.5}$$

For the subsequent steps the SH weights $\boldsymbol{y}(\boldsymbol{\theta}_1)$ are simply denoted as $\boldsymbol{y}_1$.
Eq. (4.5) denotes a mixed problem which is best explained per the following two cases:

$a = 1, b = 0$: The minimization of

$$\boldsymbol{g}_1^{\text{T}} \boldsymbol{g}_1$$

results in a minimum-energy directivity pattern of the entire beamformer.

$a = 0, b = 1$: Minimizing

$$\boldsymbol{g}_1^{\text{T}} \mathbf{Y}\mathbf{Y}^{\text{T}} \boldsymbol{g}_1$$

or equivalently

$$\left\|\boldsymbol{g}_1^{\text{T}} \boldsymbol{y}(\boldsymbol{\theta}_2)\right\|^2 + \left\|\boldsymbol{g}_1^{\text{T}} \boldsymbol{y}(\boldsymbol{\theta}_3)\right\|^2 + \left\|\boldsymbol{g}_1^{\text{T}} \boldsymbol{y}(\boldsymbol{\theta}_4)\right\|^2$$

reduces the influence of the signals $s_{\text{obj},2}(t)$,$s_{\text{obj},3}(t)$ and $s_{\text{obj},4}(t)$ on the approximation of $s_{\text{obj},1}(t)$ (cf. Eq. (4.4)) and thus the cross talk. A composite cost function for this mixed minimization problem becomes

$$\hat{J}_{\boldsymbol{g}_1} = \boldsymbol{g}_1^{\mathrm{T}} \left(a\mathbf{I} + b\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right) \boldsymbol{g}_1, \qquad \text{subject to } \boldsymbol{g}_1\boldsymbol{y}_1 = 1. \tag{4.6}$$

Using Lagrange multipliers to implicate the constraint leads to the new cost function

$$J_{\boldsymbol{g}_1} = \boldsymbol{g}_1^{\mathrm{T}} \left(a\mathbf{I} + b\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right) \boldsymbol{g}_1 + \lambda \left(\boldsymbol{g}_1\boldsymbol{y}_1 - 1\right), \tag{4.7}$$

$$\tag{4.8}$$

which can be differentiated regarding $\boldsymbol{g}_1^{\mathrm{T}}$ as well as λ and zeroed,

$$\frac{\partial}{\partial \boldsymbol{g}_1^{\mathrm{T}}} J_{\boldsymbol{g}_1} = 2\left(a\mathbf{I} + b\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right) \boldsymbol{g}_1 + \lambda\boldsymbol{y}_1 = 0, \tag{4.9}$$

$$\frac{\partial}{\partial \lambda} J_{\boldsymbol{g}_1} = \boldsymbol{g}_1\boldsymbol{y}_1 - 1 = 0. \tag{4.10}$$

Solving Eq. (4.9) for $\boldsymbol{g}_1$ results in

$$\boldsymbol{g}_1 = -\frac{\lambda}{2}\left(a\mathbf{I} + b\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right)^{-1} \boldsymbol{y}_1, \tag{4.11}$$

and substituting this in Eq. (4.10) lets us solve for the Lagrange multiplier yielding

$$\lambda = -\frac{2}{\boldsymbol{y}_1^{\mathrm{T}}\left(a\mathbf{I} + b\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right)^{-1} \boldsymbol{y}_1}. \tag{4.12}$$

Further, re-substituting Eq. (4.12) into Eq. (4.11) gives us $\boldsymbol{g}_1$, being the solution to the minimization problem for the approximation of signal $s_{\text{obj},1}(t)$. Expanding this to all signals yields a matrix $\mathbf{G}$

$$\boldsymbol{g}_1 = \left(a\mathbf{I} + b\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right)^{-1} \boldsymbol{y}_1 \left[\boldsymbol{y}_1^{\mathrm{T}}\left(a\mathbf{I} + b\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right)^{-1} \boldsymbol{y}_1\right]^{-1}, \tag{4.13}$$

$$\mathbf{G} = \left(a\mathbf{I} + b\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right)^{-1} \mathbf{Y}\,\mathrm{diag}\,\mathrm{diag}\left\{\mathbf{Y}^{\mathrm{T}}\left(a\mathbf{I} + b\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right)^{-1} \mathbf{Y}\right\}^{-1}. \tag{4.14}$$

The approximation of the desired object signal can now be calculated by Eq. (4.4).

In case that less than 4 sound object positions are concurring at a time, the remaining directions have to be selected systematically for good conditioning so that the inversions in Eq. (4.13) and Eq. (4.14) are possible.

In practice, the case where $0 \ll a \leq 1$ and $0 \leq b \ll 1$ is better suited for signal extraction in this application, since it has a stable directional pattern not dependent on the 4 signal directions. Other cases can lead to instabilities in gains when non-optimal conditioning of the directions appears. e.g. when all four sound objects lie close together seen from

a perspective. This can lead to strong interference of unwanted signals in environments such as reverberant rooms or more generally diffuseness in perspective recordings.

4.1.2 Triplet-Based Signal Extraction

Extraction of direct signals, as explained in the previous section, is done for three perspectives spanning the smallest triangle around the sound object location to get the best SNR. It is to note that since the perspective positions are not necessarily distributed on a plane, a projection of these positions onto the horizontal plane by omitting the z or more generally the upward coordinate is necessary for this. All subsequent explanations are expandable to three dimensions introducing tetrahedrons instead of triangles.

The triplets in question can be determined by algorithms pre-computing all smallest triangle combinations for a set of perspective positions or can be done manually for small sets. Selecting three perspectives obviously implies that three extracted signals are available and have to be combined reasonably.

First, the areal coordinates are introduced, cf. [Hil82]. This type of coordinates is best explained by partitioning the area of the triangle into three sub triangles, as Fig. 4.3 shows for vertex $p = 1$. The relation of areas

$$\frac{\text{Area sub triangle}}{\text{Area whole triangle}}$$

defines the weight or in this case gain of the vertex. The concept is explained for the two dimensional case only as expanding to three is straightforward.

Given the sound object position $\hat{\boldsymbol{x}}_s$ and the triplet perspectives p with $i = 1...3$, the areal coordinates superimposing the signals from the perspectives are denoted as $\boldsymbol{g}_{\text{tri},s,i}$ here. If the object position is within the triangle then $g_{\text{tri},i} \in [0, 1]$. The calculation is as follows: By inverse multiplication we get from

$$\begin{bmatrix}\boldsymbol{p}_2 - \boldsymbol{p}_1 & \boldsymbol{p}_3 - \boldsymbol{p}_1\end{bmatrix} \boldsymbol{g}_{s,23} = (\hat{\boldsymbol{x}}_s - \boldsymbol{p}_1)\,,$$

to

$$\boldsymbol{g}_{s,23} = \begin{bmatrix}\boldsymbol{p}_2 - \boldsymbol{p}_1 & \boldsymbol{p}_3 - \boldsymbol{p}_1\end{bmatrix}^{-1} (\hat{\boldsymbol{x}}_s - \boldsymbol{p}_1)\,.$$

Then the remaining gain for the first perspective is simply

$$g_{\text{tri},s,1} = 1 - \boldsymbol{g}_{s,23}^T \mathbf{1}_{2\times 1},$$

$$\boldsymbol{g}_{\text{tri},s} = \begin{bmatrix} g_{\text{tri},s,1} \\ \boldsymbol{g}_{s,23} \end{bmatrix} = \begin{bmatrix} g_{\text{tri},s,1} \\ g_{\text{tri},s,2} \\ g_{\text{tri},s,3} \end{bmatrix}.$$

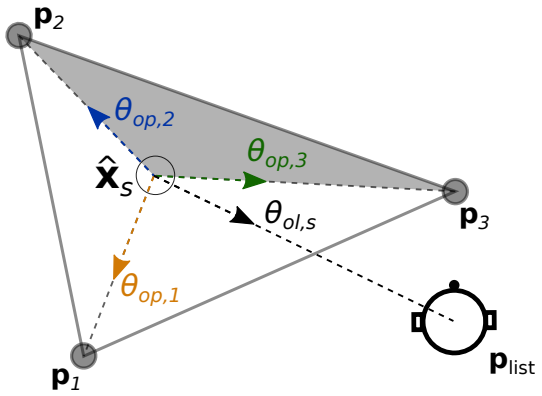


Figure 4.3: The three perspectives forming the smallest triangle around the sound object position are used for signal extraction. These signals are combined by a mixed weighting procedure. First the aerial coordinates provide preliminary gains for the perspectives by relating the sub triangles opposite of the perspective, as e.g. indicated by the grey area for $p = 1$, to the entire area spanned by $\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3$ (cf. Eq. (4.15)). Secondly, these gains are multiplied with a directional gain as introduced by Eq. (4.19) using the unit length vectors $\boldsymbol{\theta}_{\text{ol},s}$ and $\boldsymbol{\theta}_{\text{op},i}$.

These gains are continuous for all cross-triplet transitions. However, to consider the directional radiation of the sound object, the areal coordinates approach is expanded. Fig. 4.3 shows the object-listener direction and distance

$$\boldsymbol{\theta}_{\text{ol},s} = \frac{\boldsymbol{p}_{\text{list}} - \hat{\boldsymbol{x}}_s}{\|\boldsymbol{p}_{\text{list}} - \hat{\boldsymbol{x}}_s\|} \tag{4.15}$$

$$d_{\text{ol},s} = \|\boldsymbol{p}_{\text{list}} - \hat{\boldsymbol{x}}_s\| \tag{4.16}$$

and object-perspective directions and distances

$$\boldsymbol{\theta}_{\text{op},i} = \frac{\boldsymbol{p}_i - \hat{\boldsymbol{x}}_s}{\|\boldsymbol{p}_i - \hat{\boldsymbol{x}}_s\|} \tag{4.17}$$

$$d_{\text{op},i} = \|\boldsymbol{p}_i - \hat{\boldsymbol{x}}_s\| \tag{4.18}$$

for the detected sound object s. The expansion considers the alignment between the directionally radiated sound that should arrive at the listener and the recording perspective that is best aligned with this direction. This is done a cardioid function in the form of

$$g_{\text{dir},s,i} = \left[\frac{1}{2}\left(\boldsymbol{\theta}_{\text{op},i}^{\text{T}}\boldsymbol{\theta}_{\text{ol},s} + 1\right)\right]^{\alpha} \tag{4.19}$$

which is a straightforward way to acquire a measure for this alignment. The exponent α facilitates control over the cardioid order and influence of the directional alignment of the recording perspectives with the radiation direction to the listener. Finally, perspective-wise multiplication of the corresponding values obtained by combining both methods and

normalizing the result to a sum of 1 yields the squared single signal gains for sound object s and triplet perspective i

$$g_{s,i} = \sqrt{\frac{g_{\text{tri},s,i}\, g_{\text{dir},s,i}}{\sum_{j=1}^{3} g_{\text{tri},s,j}\, g_{\text{dir},s,j}}}. \tag{4.20}$$

The approximated direct signal is now a combination of the extracted object signals $s_{\text{obj},i}(t)$ after Eq. 4.4. Before combination, the signals are shifted to compensate the delayed arrival of sound at the triplet perspectives using the differences of the object-perspective distances $\Delta t_i \propto \Delta d_{\text{op},i}$

$$\bar{s}_s(t) = \sum_{i=1}^{3} g_{s,i}\, s_{\text{obj},i}(t - \Delta t_i) \tag{4.21}$$

The sound object gain g_s for encoding is now calculated according to the distance law $\propto \frac{1}{d}$. The individual gains $g_{s,i}$ are taken into consideration to ensure consistency with Eq. (4.21). The values are limited to a maximum of 4 to avoid complications at the edge case $d_{\text{ol},s} \to 0$:

$$g_s = \min\left\{4,\ \frac{1}{d_{\text{ol},s}} \sum_{i=1}^{3} g_{s,i}\, d_{\text{op},i}\right\}. \tag{4.22}$$

4.1.3 Encoding

The direction and amplitude are dependent on the relative position of the sound object to the listener perspective, previously introduced as the object-listener direction $\boldsymbol{\theta}_{\text{ol},s}$ in Eq. 4.15. The Ambisonics signals are computed by multiplication of the approximated direct signals $\bar{s}_s(t)$ with the appropriate encoder $\boldsymbol{y}_{\text{N}}(-\boldsymbol{\theta}_{\text{ol},s})$. The sum of all encoded sound object signals yields the direct signal part of the virtual listener perspective

$$\boldsymbol{\chi}_{\text{direct}}(t) = \sum_{s=1}^{S} \boldsymbol{y}_{\text{N}}(-\boldsymbol{\theta}_{\text{ol},s})\, \bar{s}_s(t)\, g_s. \tag{4.23}$$

Additional signal delay proportional to the object-listener distance $d_{\text{ol},i}$ could be introduced in Eq. (4.23) to ensure time-exact reconstruction, however this showed no audible improvement compared to the non-compensation approach. Rather, it could yield a time-varying delay and thus Doppler shifts while the virtual listener is moving fast, which makes its omission the more robust variant.

4.2 Residual Signals

The extracted signals are an approximation of the direct signals of the sound objects at the estimated positions. Under ideal conditions, this implies the absence of room information in those object signals. As however, room acoustic information is an important background information, it has to be considered in the the playback algorithm separately. The high-order encoding of the direct signals ensures a high definition in auditory localization while the residual signal is necessary to convey the room impression and support externalization. The computation of the residual signal is explained subsequently.

4.2.1 Perspective Residual Signals

In [GZS+18, ZFSH20] an approach for rendering multi-perspective recordings of scenes by using *Virtual Loudspeaker Objects* (VLO) was proposed. In practice, this approach proved well in terms of object localisation, as shown in [RZF17] but more importantly yields an appropriate room impression.

Assuming the virtual listener is located at a microphone array center position and only considering this one recording perspective for now, the approach is best described as a virtual surround setup with the array microphone signals $s_{p,i}(t)$ as virtual loudspeakers, see Fig.4.4a. These virtual loudspeakers are located at a finite distance R in the array microphone directions $\boldsymbol{\theta}_{\mathrm{c}_i}$ (from Eq. (1.1)). The distance R permits a shift of the listener perspective off the array center position and the encoding of the VLOs must therefore happen according to the shifted distance and directions

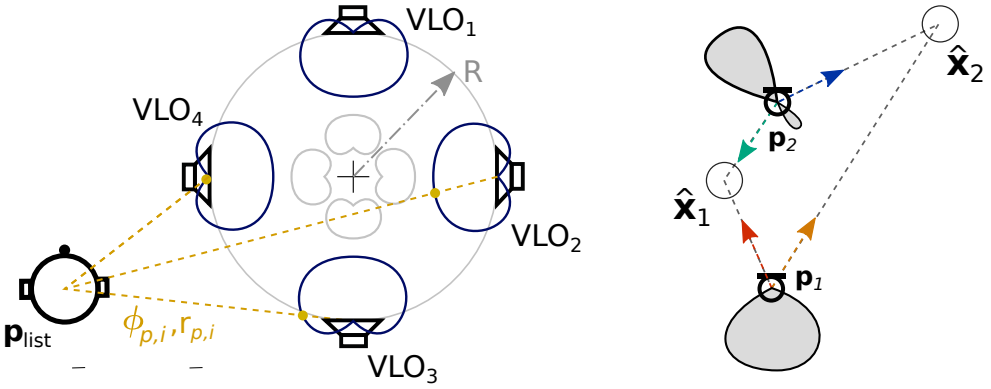
$$r_{p,i} = \| \left(\boldsymbol{p}_p + R\boldsymbol{\theta}_{\mathrm{c}_i}\right) - \boldsymbol{p}_{\mathrm{list}} \| \tag{4.24}$$

$$\boldsymbol{\phi}_{p,i} = \frac{\left(\boldsymbol{p}_p + R\boldsymbol{\theta}_{\mathrm{c}_i}\right) - \boldsymbol{p}_{\mathrm{list}}}{r_{p,i}} \tag{4.25}$$

Before encoding, each signal of the array microphones $i = 1...4$ is attenuated according to a gain function incorporating the distance to the VLO and directivity therefore of. Using the relative distance $\frac{r_{p,i}}{R}$ ensures unit gain for all array microphones for a centered listener but needs an extension to avoid gains above unity which is realised as the inverse function. Therefore the distance gain becomes

$$g_{\mathrm{dis},p,i}(r, R) = \begin{cases} \frac{R}{r_{p,i}}, & \text{for } r_{p,i} > R \\ \frac{r_{p,i}}{R}, & \text{for } r_{p,i} \leq R \end{cases}. \tag{4.26}$$

Assuming the location is shifted outside of the radius R, additional attenuation is necessary to avoid incorrect directional mapping when being positioned behind a VLO. Therefore, the VLOs are modelled with a cardioid radiation pattern resulting in a directional



(a) Virtual Loudspeaker Objects of one perspective p

(b) Direct signal de-emphasis of two perspectives and two sound objects

Figure 4.4: The VLO approach is used as a basis for the residual signals. (a) Virtual loudspeaker objects are placed around the perspective position and attenuated according to cardioid radiation patters (blue) and relative listener direction (yellow). (b) The gain patterns after direct signal de-emphasis are shown in grey.

gain,

$$g_{\text{dir},p,i}(r) = \left(1 - \frac{\alpha}{2}\right) + \frac{\alpha}{2}\boldsymbol{\theta}_{\text{c}_i}^{\text{T}}\boldsymbol{\phi}_{p,i}, \tag{4.27}$$

$$\alpha = \frac{r_{p,i}}{r_{p,i} + R_{\text{dir}}}. \tag{4.28}$$

While the object direct signals should not contain room information, the residuals should in turn not contain object direct signals. When object signals cross-talk to the residual signals, their VLOs would yield multiple instances of the object and hereby deteriorate their spatial definition. Therefore the approach above is expanded. We apply de-emphasis to the object signals contained in each perspective to approximate the residuals. Each extracted sound object signal is used to suppress the perspective signals by attenuation proportional to the signal strength coming from the object towards the perspective. This is achieved by computing a moving RMS measure over a time period of Δt_{RMS}, e.g. 5ms, of the object signal $s_s(t)$ determined as

$$s_{\text{RMS},s}(t) = \sqrt{\frac{1}{\Delta t_{\text{RMS}}} \int_{t-\Delta t_{\text{RMS}}}^{t} \bar{s}_s(\tau)^2 \mathrm{d}\tau}. \tag{4.29}$$

Now, a de-emphasis directivity pattern is constructed. This entails the aforementioned RMS signal values of the detected sound objects and their estimated positions $\hat{\boldsymbol{x}}_s$ for the corresponding time frame. The directional attenuation pattern is computed as the product of single cardioid patterns directed towards the estimated object positions.

In order to scale these single directional attenuation patterns, the moving RMS signal value is compared to a threshold value RMS_s. This value is determined manually, empirically, or as the maximum moving RMS value of the extracted signal beforehand. The attenuation factor $a_{p,s}(t)$ is introduced dependent on the RMS ratio and a distance factor analogue to Eq. (4.26) but using the object-perspective distance $d_{\text{op},p}$ (cf. Eq. (4.18)) and the reference distance R_{de},

$$a_{p,s}(t) = \frac{s_{\text{RMS},s}(t)}{\text{RMS}_s}\, g_{\text{dis}}(d_{\text{op},p}, R_{\text{de}}). \tag{4.30}$$

The unit vector indicating the direction from perspective to sound object is the inverted directed object-perspective direction $-\boldsymbol{\theta}_{\text{op},p}$ (cf. Eq. (4.17)). The combination of patterns now becomes

$$G_p(t, \boldsymbol{\theta}) = 1 - \prod_{s=1}^{S} a_{p,s}(t) \left[\frac{1}{2}\left(1 - \boldsymbol{\theta}_{\text{op},p}(t)\boldsymbol{\theta}\right)\right]^{\beta} \tag{4.31}$$

with an exponent β for higher order directivity. Now, this allows us to calculate attenuation factors for arbitrary directions. In our application, this is done with the array microphone directions for the perspective p, $\boldsymbol{\theta}_{\text{c}_i}$ (cf. Eq.(1.1)) yielding the final attenuation factors for the array microphone signals. Combined with the factors from Eq. (4.27) and Eq. (4.26) this becomes

$$\tilde{s}_{p,i}(t) = s_{p,i}(t)\, G_p(t, \boldsymbol{\theta}_{\text{c}_i})\, g_{\text{dis},p,i}\, g_{\text{dir},p,i}. \tag{4.32}$$

These signals are the attenuated VLO signals which have to be encoded accordingly. Figure 4.4b visualises the idea of the resulting directivity patterns.

4.2.2 Residual Signal Encoding

The now appropriately attenuated 4 array microphone signals $\tilde{s}_{p,i}(t)$ are then multiplied with the fifth-order Ambisonics encoder [ZF19] for the VLO directions $\boldsymbol{y}(\boldsymbol{\phi}_{p,i})$ (cf. Eq. (4.25)). This yields the residual Ambisonics signals of one microphone perspective from the perspective of the listener

$$\boldsymbol{\chi}_p(t) = \sum_{i=1}^{4} \boldsymbol{y}(\boldsymbol{\phi}_{p,i})\, \tilde{s}_{p,i}(t). \tag{4.33}$$

Applying this single-perspective procedure to the recorded perspectives and summation of the resulting Ambisonics signals delivers the total residual signals for the listener perspective

$$\boldsymbol{\chi}_{\text{residual}}(t) = \sum_{p=1}^{P} \boldsymbol{\chi}_p(t). \tag{4.34}$$

Encoding of time-delays to the listener can be omitted to avoid Doppler shifts or interference varying with the listener position.

4.2.3 Dynamic Binaural Rendering

As the final step, the direct signal perspective $\boldsymbol{\chi}_{\text{direct}}(t)$ (Eq. (4.23)) and residual signal perspective $\boldsymbol{\chi}_{\text{residual}}(t)$ (Eq. (4.34)) are merged into the complete virtual listener perspective $\boldsymbol{\chi}$. Using gain values $0 \leq a_{\text{direct}} \leq 1$ and $0 \leq a_{\text{residual}} \leq 1$ to balance the two partial perspectives gives control over the effects of both direct and residual part on the listening experience. The merging is simply done as

$$\boldsymbol{\chi}(t) = a_{\text{direct}}\, \boldsymbol{\chi}_{\text{direct}}(t) + a_{\text{residual}}\, \boldsymbol{\chi}_{\text{residual}}(t). \tag{4.35}$$

These Ambisonics signals can now be used for any decoding operation, such as binaural rendering. At this point, the head rotation of the listener has to be taken into account. Ambisonics signals can be rotated with a $(N+1)^2 \times (N+1)^2$ listener head rotation matrix $\boldsymbol{R}_{\text{l}}(\varphi_{\text{l}}, \vartheta_{\text{l}}, \gamma_{\text{l}})$ calculated from the Euler angles $\varphi_{\text{l}}, \vartheta_{\text{l}}, \gamma_{\text{l}}$ being azimuth, elevation and roll respectively (cf. Fig. 4.1) as

$$\boldsymbol{\chi}_R(t) = \boldsymbol{\chi}(t)\boldsymbol{R}_{\text{l}}(\varphi_{\text{l}}, \vartheta_{\text{l}}, \gamma_{\text{l}}). \tag{4.36}$$

This can be done in the time or frequency domain. Since the methods to acquire Ambisonics rotation matrices from given angles or quaternions are well known, they will not be further explained, if necessary refer to literature such as [IR96, ZF19]. The rendering of the Ambisonics signals to a binaural signal is done by an existing MagLS binaural decoder [SZH18, ZF19] implemented as the *IEM BinauralDecoder* [1] [Rud19].

4.3 Implementation

The rendering is realized as an application in Pure data [Puc16], which allows users to create signal processing algorithms using a graphical object oriented programming language. The application is the real-time implementation of the signal processing procedures introduced and explained in the previous sections and uses the stored time frame-based analysis data from the object tracking algorithm introduced in earlier chapters.

1. https://plugins.iem.at/

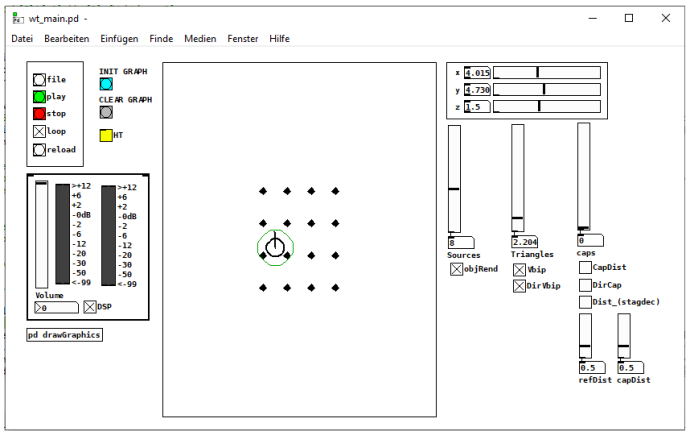
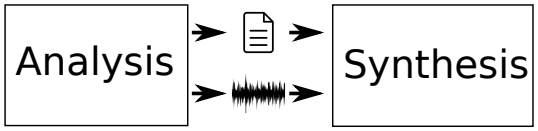


Figure 4.5: The user interface of the Pure data application exposes the main features and controls.

The implementation of a real-time analysis system is not in the scope of this work. However, a real-time rendering system is necessary for the interactive perceptual evaluation of the proposed method of rendering. Therefore the analysis procedure introduced in Chapter 3 as well as the direct signal extraction is done beforehand and the information is stored for repeated use in the playback algorithm. The gathered trajectory information is stored as the frame-wise estimated positions of the detected sound objects. Direct signal extraction is not required to be real-time computable either, as it is only dependent on object position. The storage of the extracted audio data is done as *.wav* files. In a playback situation, the necessary object and audio data is loaded on-demand.



This separated implementation still satisfies all requirements for a working prototype of a rendering application used for evaluation purposes.

Chapter 5

Evaluation

To assess the performance and usability of the proposed method, evaluation steps have been taken. First, a numerical one performs error measurements in regards to tracking and detection accuracy. Further, a listening evaluation consisting of 2 experiments on aspects of the quality of the audio playback algorithm has been conducted. The following sections introduce the methods and errors measurements used and discuss the results.

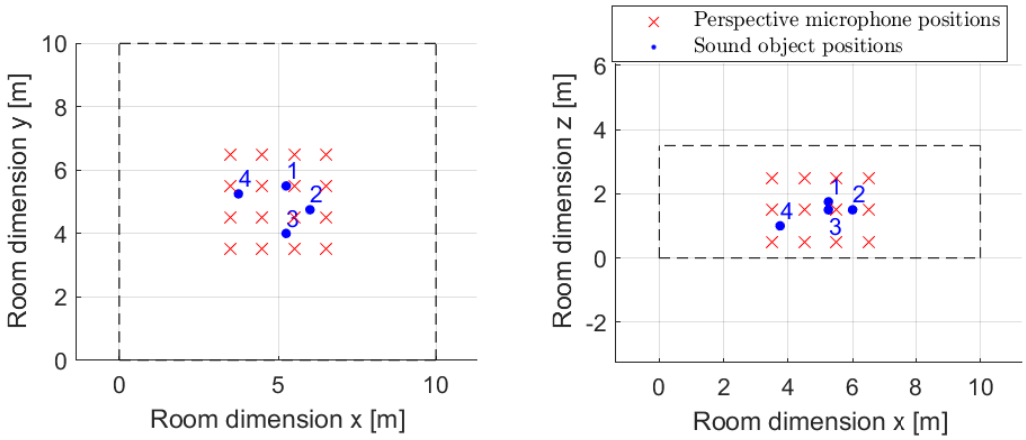
5.1 Numerical Evaluation of Object Tracking

5.1.1 Method

The effectiveness of the object tracking algorithm is evaluated by analysing simulated scenarios with changing variables, more specifically the noise contamination of the recordings and the SH order of the directional distributions in the analysis. A 10m by 10m by 3.5m room is simulated with the MATLAB® library *MCRoomSim* [1] [WEJvS10] using the standard parameters for reflection and attenuation.

An exemplary arrangement of perspectives is used, this being a 4 by 4 by 3 grid at the center of the room with 1 meter grid spacing as pictured in Fig. 5.1. Illustrated as well are the four sound object positions used in the static scenario whose positions are listed in Tab. 5.1. The sound objects are speaker signals from the *EBU SQAM* collection (Track Nr. 49, 50, 51, 52) [EBU08]. The simulation computes error measures to compare different noise situations and the influence of the SH order used for the directional distributions.

1. available at https://github.com/Andronicus1000/MCRoomSim



(a) X-Y view

(b) X-Z view

Figure 5.1: The perspective microphone and sound object positions for the analysis. Microphone arrays are located at 1m intervals ranging from 3.5m to 6.5m on the x and y coordinate. On the z coordinate axis, the are spaced at 1m as well from 0.5m to 2.5m this results in 48 virtual microphone arrays. The room dimensions are marked dashed black.

5.1.2 Error Measures

The analysis yields estimated position trajectories for the sound objects which are then compared to the ground truth employing the following measures.

Mean Distance Error: This measure is defined as the mean distance between the ground truth to the nearest sound object. Only 1-to-1 mappings are allowed, therefore if an object is already in use for calculation then the next-nearest will be used if existent. The measure is computed for every sound object $j = 1...4$. The distance to the ground truth is averaged over N_{trial} trials and M time frames, if active detected sound objects are present:

$$\mathrm{MDE}_j = \frac{1}{N_{trial}} \sum_{n=1}^{N_{trial}} \frac{1}{M} \sum_{m=1}^{M} \min_i \left\| \boldsymbol{s}_{j,\mathrm{true}}^{(m,n)} - \hat{\boldsymbol{s}}_i^{(m,n)} \right\|. \tag{5.1}$$

Combined Mean Distance Error: This defines the sum of the mean distance errors over all $J = 4$ sound objects:

$$\mathrm{CMDE} = \sum_{j=1}^{J} \mathrm{MDE}_j. \tag{5.2}$$

Object number	X	Y	Z
1	5.25 m	5.50 m	1.75 m
2	6.00 m	4.75 m	1.50 m
3	5.25 m	4.00 m	1.50 m
4	3.75 m	5.25 m	1.00 m

Table 5.1: The true sound object positions.

Activation Error Time: The detected sound object activity is the time where false positives or false negatives in sound object activity appear. This is done by summation of time frame lengths where there is no nearest sound object existent or the sound object probability is not 1. This measure is potentially harsh as the object detection algorithm for the rendering has more relaxed thresholds but should nevertheless help assess the responsiveness of the sound object detection depending on variables such as SNR and SH order:

$$\mathrm{AET}_j = \sum_{m=1}^{M} \phi_{j,\mathrm{active}}(m)\Delta t. \tag{5.3}$$

$$\phi_{j,\mathrm{active}}(m) = \begin{cases} A^{(m)}_{j,\mathrm{truth}} & \text{if no nearest unassociated sound object existent,} \\ |P^{(m)}_s - A^{(m)}_{j,\mathrm{truth}}| & \text{else, s is nearest sound object.} \end{cases} \tag{5.4}$$

The truth $A^{(m)}_{j,\mathrm{truth}}$ is defined by manual analysis of the sound object signals.

5.1.3 Parameters

The simulation uses a set of parameters listed in Table 5.2 and additionally references their definitions.

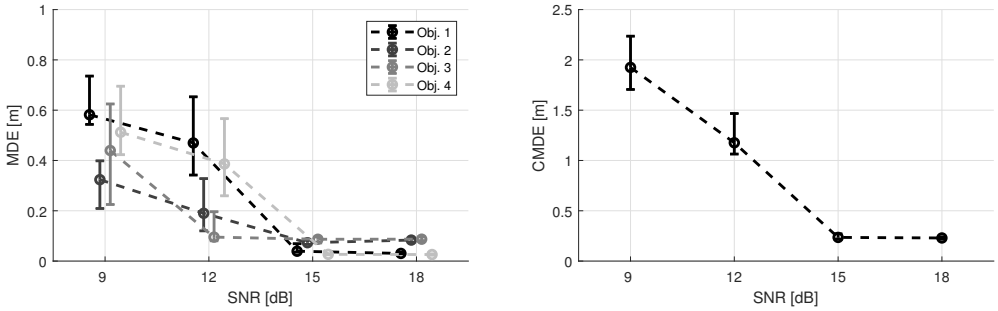
Parameter	Reference	Value
Nr. grid points G	Sec. 3.1	1521
Grid spacing	Sec. 3.1	0.25 m
SH order N	Sec. 2.1.2	15
Distance factor δ_d	Sec. 2.2.2	5
Reference distance d_0	Sec. 2.2.2	0
SH subtraction weights $\boldsymbol{a}$	Sec. 3.1.2	maxR_E
Observation threshold	Sec. 3.1.2	-20 dB
SFTF length	Sec. 2.1.2	2048 samples
Hopsize	Sec. 2.1.2	1024 samples
Smoothing size M	Sec. 2.1.2	16
Activation time	Sec. 3.3.2	0.1 s
Disable time	Sec. 3.3.2	0.7 s
$P_{\text{assoc,false}}$	Sec. 3.2.1 Eq. 3.19	0.6
$P_{\text{assoc,new}}$	Sec. 3.2.1 Eq. 3.19	0.1
P_0	Sec. 3.2.1 Eq. 3.21	0.5
$P(\text{active}\|\text{inactive})$	Sec. 3.2.1 Eq. 3.23	0.05
$P(\text{active}\|\text{active})$	Sec. 3.2.1 Eq. 3.23	0.95
μ	Sec. 3.2.1 Eq. 3.10	4
Particles N	Sec. 3.3.1	100
Damping α_{dyn}	Sec. 3.3.1 Eq. 3.32	2
Process noise factor β_{dyn}	Sec. 3.3.1 Eq. 3.33	0.04

Table 5.2: The simulation parameters.

5.1.4 Results

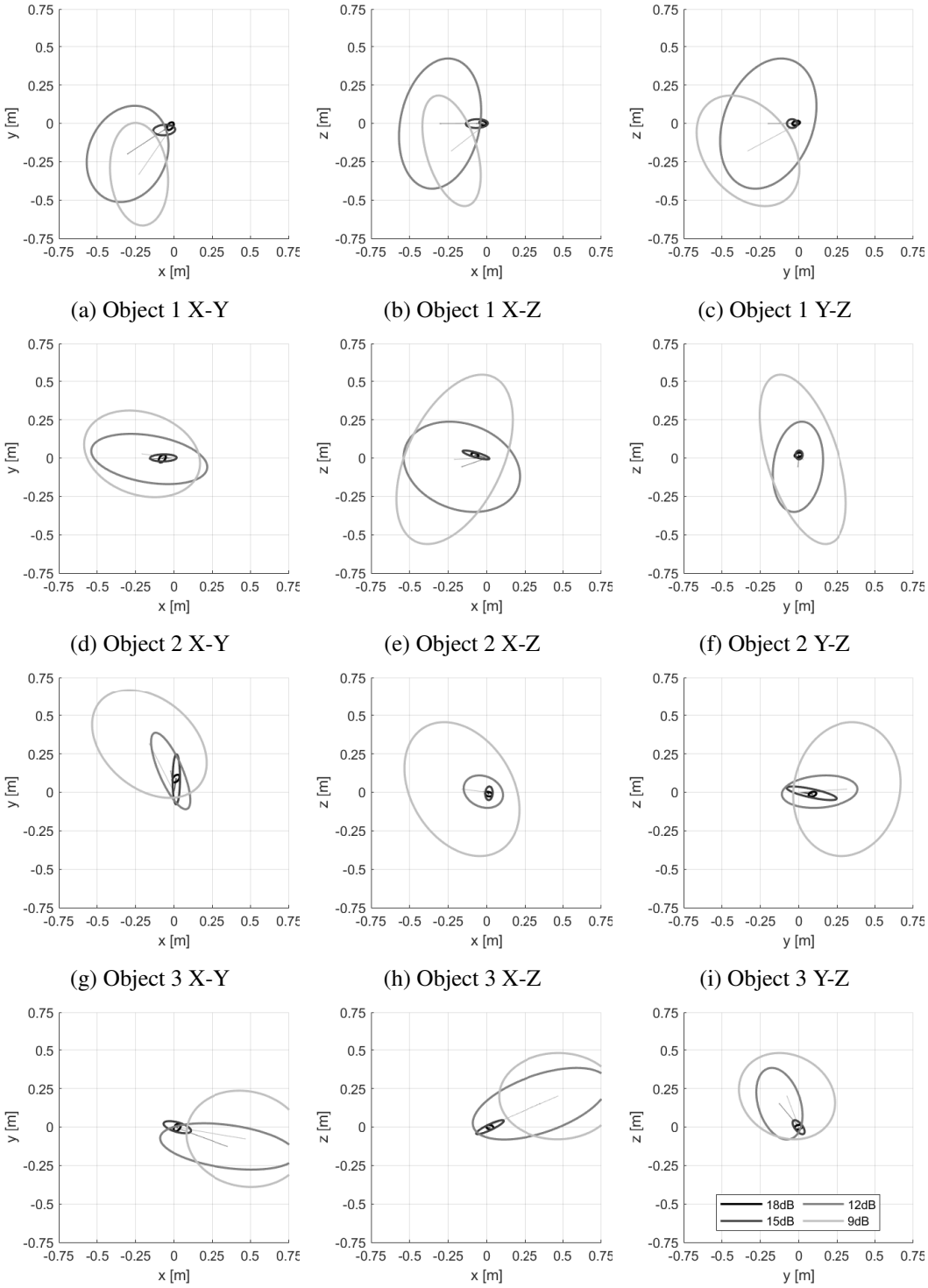
Robustness to Noise

To evaluate the robustness of the sound object tracking to uncorrelated noise, trials are run for variable noise levels. $N_{trial} = 20$ are run for each SNR in the set of $\text{SNR}_n \in \{9, 12, 15, 18\}$ dB which has been identified as representative since higher and lower values yield no significant increase or decrease in results. White noise with normal distribution is generated independently for each array microphone. The noise has equal energy on all microphone as the SNR taken into account for noise level generation is measured between noise and the loudest signal only. The effectiveness of the algorithm is affected by increased noise level. The results in terms of positional accuracy are relatively constant up until a SNR between 15dB and 12dB.



(a) MDE (b) CMDE

Figure 5.2: MDE of the sound objects 1 to 4 and the CMDE. The decrease of error with increasing SNR is clearly visible. SNRs larger than the picture maximum do not increase accuracy further, although a decrease in trial error spread can be seen comparing 15dB and 18dB. SNRs of 12dB and below are not feasible, as the detection becomes unreliable, and the position accuracy suffers greatly since the estimates start jumping heavily which the large 95% confidence intervals suggest.



(a) Object 1 X-Y (b) Object 1 X-Z (c) Object 1 Y-Z

(d) Object 2 X-Y (e) Object 2 X-Z (f) Object 2 Y-Z

(g) Object 3 X-Y (h) Object 3 X-Z (i) Object 3 Y-Z

(j) Object 4 X-Y (k) Object 4 X-Z (l) Object 4 Y-Z

Figure 5.3: The 95% confidence ellipses of the misalignment between estimated and true sound object positions after 20 numerical simulation trials for different noise levels. The increase of spread and therefore decrease in accuracy is observable with decreasing SNR.

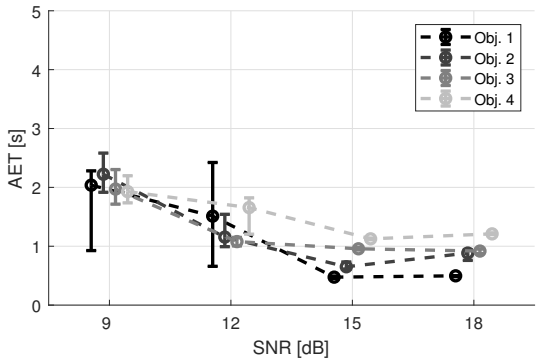
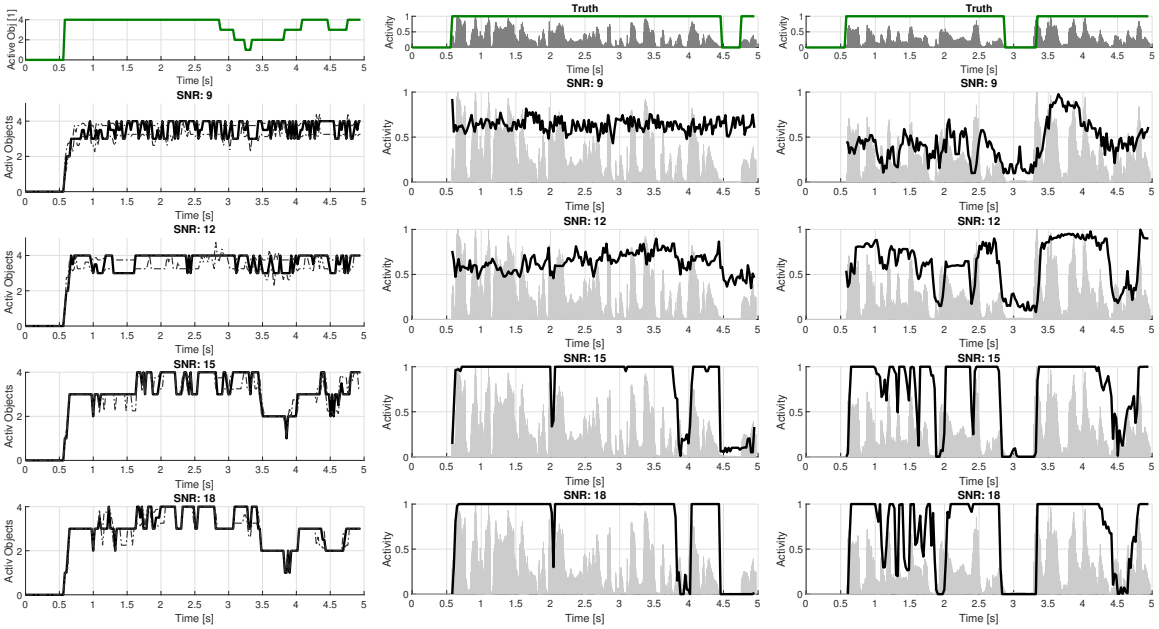


Figure 5.4: As expected, the AET of the sound objects decreases with higher SNR. The confidence intervals suggest again a strong variation in results indicative of jumping measurements at lower SNRs. Pictured are median, and 95% confidence intervals.

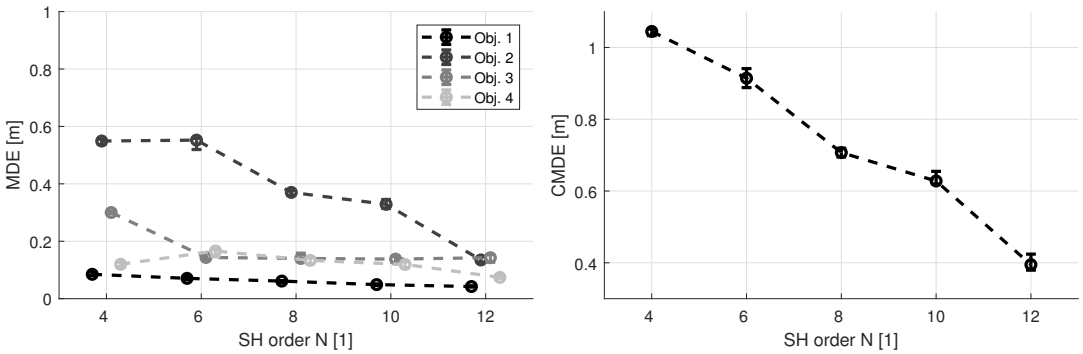


(a) Active Objects (b) Mean activity 1 (c) Mean activity 2

Figure 5.5: The number of objects being considered active is determined by counting the tracked objects with $P_s > 0.8$. Comparing that to the ground truth shows the degrading effect of lower SNRs (a). Mean activity over all trials of the sound objects 1 and 2 shows the activity detection working well at higher SNRs. The mean activities jumping between zero and one as seen at lower SNRs indicates that activity detection stops working properly and falls to the randomness introduced by the noise.

Dependency on Order of Spherical Harmonics

The error measures are computed for different orders of spherical harmonics used in the directional distribution calculation. Here, $N_{trial} = 20$ trials are run to compute the measurements. While higher orders bring more positional accuracy, the effect is not as strong as seen in case of SNR.



(a) MDE (b) CMDE

Figure 5.6: The medians of the MDE and 95% confidence intervals depending on the SH order shows a decrease towards higher orders. The low spread is indicative of consistent location detection with more or less constant error distance.

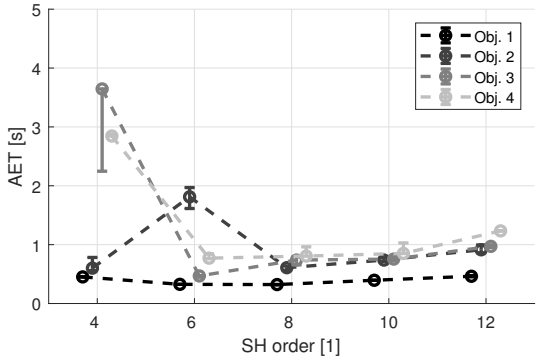
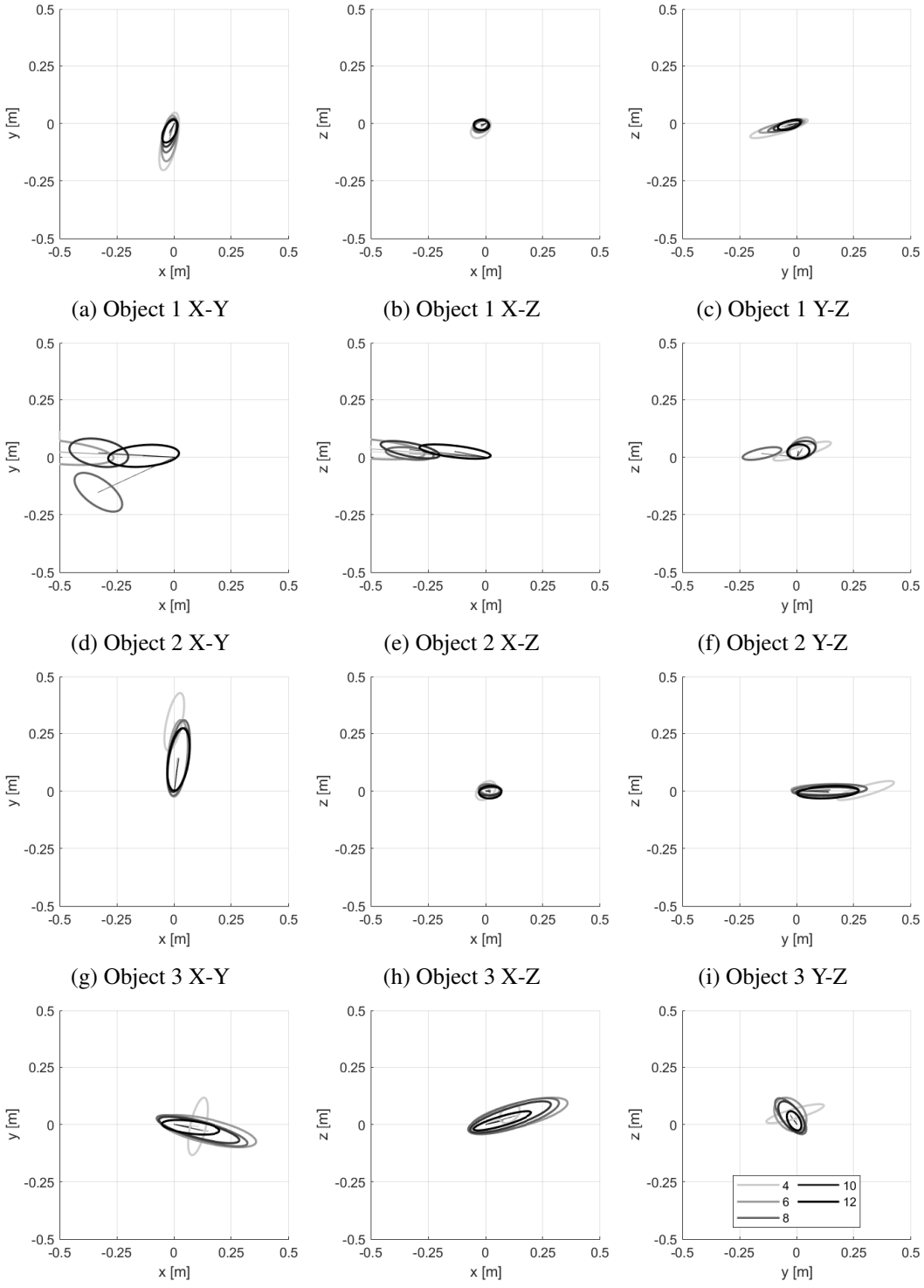


Figure 5.7: The AET of the sound objects does seem to be affected by the order of SHs, but is in conjunction with the object location, as e.g. a comparison of object 1 and 3 suggests.



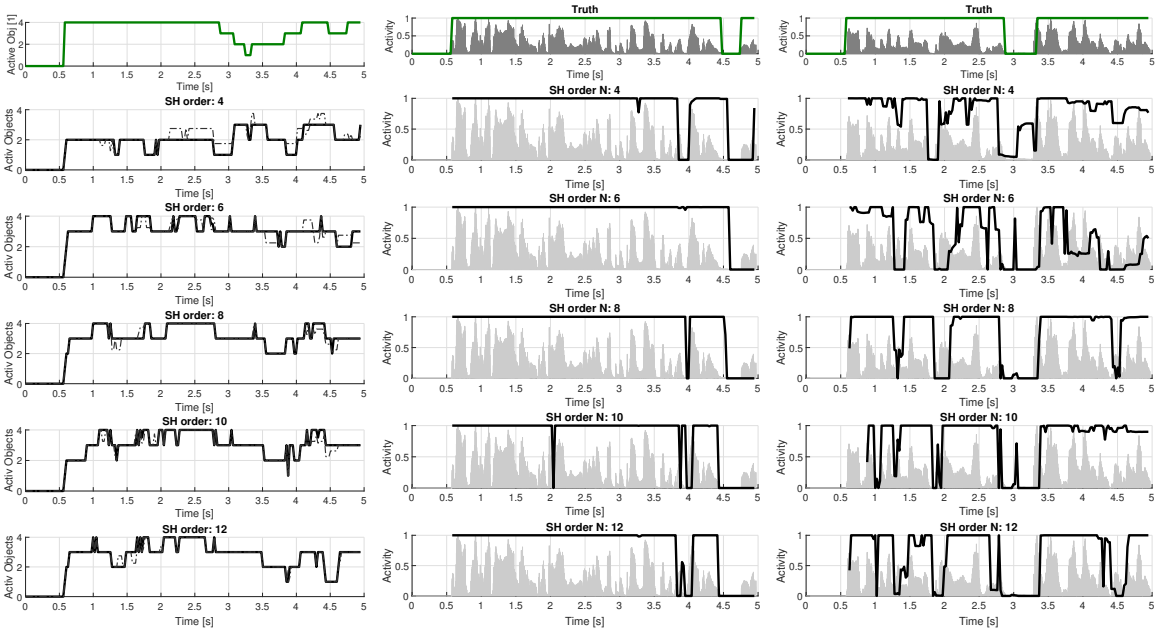
(a) Object 1 X-Y (b) Object 1 X-Z (c) Object 1 Y-Z

(d) Object 2 X-Y (e) Object 2 X-Z (f) Object 2 Y-Z

(g) Object 3 X-Y (h) Object 3 X-Z (i) Object 3 Y-Z

(j) Object 4 X-Y (k) Object 4 X-Z (l) Object 4 Y-Z

Figure 5.8: The 95% confidence ellipses of the misalignment between estimated and true sound object positions after 20 numerical simulation trials for different orders of spherical harmonics. The spread does not increase drastically, however the mean distance appears to be jointly dependent on order and object-microphone constellation.



(a) Active sound objects (b) Mean activity of object 1 (c) Mean activity of object 2

Figure 5.9: The number of objects being considered active is determined by counting the tracked objects with $P_s > 0.8$. Comparing that to the ground truth shows the degrading effect of lower SH orders (a). Mean activity over all trials of the sound objects 1 (b) and 2 (c) shows the activity on one hand fitting very well over all noise levels. On the other, looking at sound object 2, shows stronger influence of SH order.

5.1.5 Discussion

The numerical analysis provides insight into the algorithms dependency on noise in the scene and the order of SHs used for the computation of directional distributions. The results prove good performance up to 15 dB SNR and showed the necessity of SH order higher than 8.

The effectiveness of the proposed method seems not to have a linear dependency on SNR but rather a sharp drop off between 15dB and 12dB with further decrease at 18dB. This leads to the assumption that a quadratic or exponential dependency on the overall SNR lies at hand. This could be due to the similar degradation of the directional accuracy of all microphone perspectives concurrently and the combination thereof through perspective merging. Further, the non-linearity of the detection algorithm in regards to observation validation could be a possible cause of the jump in error rates between 12dB and 15dB SNR.

The order of SH directional distributions is equivalent to the directional resolution with which the DOA detection of the individual first-order perspectives is considered and intersected in space. It only affects the accuracy of objects in certain positions. Assuming DOA estimation of the microphone perspective is accurate and without disturbance, then

large errors due to low-order directional distributions could be assumed to occur whenever the intersecting angles of the different directions enclose angles close to $0°$ or $180°$. Further, multiple object in close proximity to each other influence single object detection accuracy. Possible quieter objects get pulled towards louder ones especially when directional resolution is not good enough and smaller peaks are masked by a larger ones. It can be assumed that the detections of object 2 are pulled towards the center by objects 1 and 3 as observed in Figure 5.8.

5.2 Listening Evaluation

In order to understand how the proposed method compares to selected existing ones, a MUSHRA-like [ITU15] comparative evaluation was conducted. This method of evaluation presents conditions to a listener while allowing to switch between them and rate them on a scale of 0 to 100 comparing them to a reference and with each other. For this particular evaluation, two major parts containing a number of trials were used. The first part presented static, while the second one dynamic listener perspectives.

Because of the COVID-19 pandemic at the time of authoring this thesis, experimentation was done using static binaural rendering. Also an initial design of the evaluation using visualization and interfacing with a head-mounted display was skipped to allow the participants to do the test at home. Nevertheless, there is the strong believe that despite this small inconvenience, all the major characteristics that stem from user-interactive real-time manipulation were enclosed and could be covered representatively in the evaluation.

5.2.1 Experiment Setup

The evaluation uses a virtual recorded sound scene simulated by the *MCRoomSim* [WEJvS10] library. A 6m by 6m room with a microphone array grid of a 1m spacing, as shown in Fig. 5.11a, is set up using the default room settings. The array microphone signals are simulated by using *cardioid*-type receivers at positions equivalent to Oktava MK-4012 arrays. A mono recording of a male speaker of the *EBU SQAM* collection (Track-Nr. 50) [EBU08] was convoluted with the simulated RIRs for all array microphones and ARIRs of the reference perspectives. A MUSHRA application (cf. Fig. 5.10) was used to present the conditions to the participant while Reaper [2] provided the audio playback. The communication between programs was facilitated by OSC [3]. The setup of the evaluation was designed to allow distribution over the internet and participation at home. A requirement for participation was a high-quality set of headphones and the

2. https://www.reaper.fm/
3. http://opensoundcontrol.org

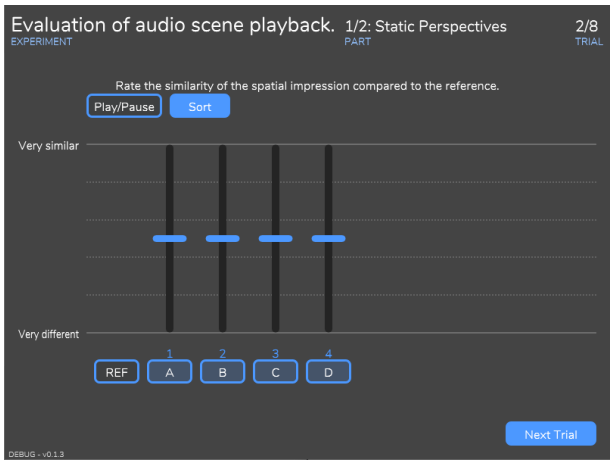


Figure 5.10: The user interface of the listening experiment.

experiment was pre-rendered for the most common high-end *AKG* and *Beyerdynamic* headphone models.

Conditions

The four selected conditions were presented randomly to avoid learning effects or fatigue. The reference was always visible and not doubled by a hidden reference. All conditions were pre-rendered binaural signals using the following rendering methods:

Reference: Using a *sphharm*-type receiver of fifth order in the *MCRoomSim* library, a perspective using ARIRs was simulated and rendered to a binaural signal using the *IEM BinauralDecoder*.

Proposed: The condition was rendered following the method of analysis and synthesis as described in this work. The Pure Data application implemented for interactive listening with head tracking was sent simulated head position and rotation data to ensure consistency with the reference.

An audio scene involving two sound objects was analysed by the detection algorithm. However, for rendering, only one sound object was used, while still using the object tracking data. This should the effect multiple sound objects have on the object tracking and therefore scene playback assessable, while keeping the listening experiment at minimal audible complexity for the participants.

VLO: This robust scene rendering approach was proposed in [GZS+18, ZFSH20] and introduced in Sec. 4.2.1.

VBIP: The condition is a binaural rendering of a linear combination of the perspective FOA signals using areal coordinates to linearly mix the signals of the triplet perspectives that surround the listener position. The microphone signals of the perspectives are mixed corresponding with their respective microphone directions. The 4 resulting combined signals are encoded as third order Ambisonics signals using the microphone array directions projected onto the horizontal plane and rendered to a binaural signal.

Anchor: This condition is the VBIP binaural signal merged into a mono signal played on both headphone channels. The purpose of this anchor is to provide a low-rating reference to listeners.

Experiment 1 - Static Perspectives

To assess the playback algorithm in terms of quality of reconstruction of the sound scene, a set of static perspectives was selected. There are 4 positions in this set, 2 exactly at a microphone array position and 2 at triplet centers, each being a pair of small and large distance to the speaker. These perspectives, as shown in Fig. 5.11a, are all facing the speaker. Each position is rated twice, giving $4 \times 2 = 8$ trials for this part of the experiment.

Before the start of the rating experiment, the participants were given the following task description:

You are tasked to rate the presented stimuli regarding the similarity of their spatial impression when compared to a reference. You are presented a virtual listening perspective facing a male speaker. When rating the stimuli, please take into consideration the following aspects: Is the speaker clearly locatable? Does the distance perception fit? Does the room impression fit?

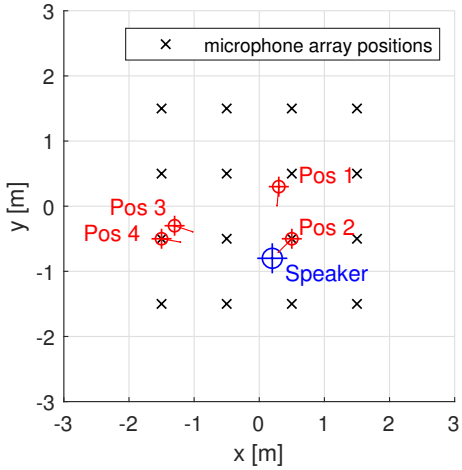
The rating scale was labelled with *very different* as the lower scale end and *very similar* as the upper scale end.

Experiment 2 - Dynamic Perspectives

To assess the playback algorithm in terms of quality of reconstruction of the sound scene, the virtual listener perspective was moved along a trajectory in the scene with varying look directions $\{A,\ B,\ C,\ D\}$ (cf. Fig. 5.12). Each look direction is rated twice providing $4 \times 2 = 8$ trials for this part of the experiment.

Before the start of the rating experiment, the participants were given the following task description:

You are tasked to rate the presented stimuli regarding the similarity of their spatial impression when compared to a reference. You are presented a virtual listening perspective



(a) Selected positions

Position	x [m]	y [m]	z [m]
Speaker	0.2	-0.8	1.5
Pos 1	0.3	0.3	1.5
Pos 2	0.5	-0.5	1.5
Pos 3	-1.3	-0.3	1.5
Pos 4	-1.5	-0.5	1.5

(b) Position coordinates

Figure 5.11: The static perspectives used in the first experiment.

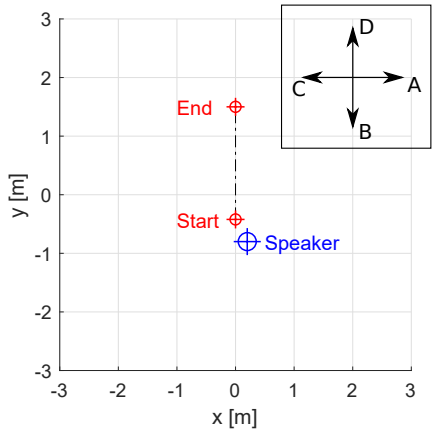


Figure 5.12: The trajectory and defined look directions A,B,C,D of the virtual listener perspective in the evaluation scene as presented by the experiment.

moving on the trajectory seen on the handout. Several look directions are present in trials and given in the trial headline. When rating the stimuli, please take into consideration the following aspects: Is the distance and the direction as well as the movement of the speaker fitting? Is the room impression fitting? Also take into account the stability/smoothness/linearity of these aspects.

The rating scale was labelled with *very different* as the lower scale end and *very similar* as the upper scale end.

5.2.2 Results

6 expert listeners aged between 26 and 39 (average age: 29) took part in the listening experiment taking 34 minutes on average to complete it. A Wilcoxon signed rank test [Wil45] with Bonferroni-Holm correction [Hol79] was used to assess the statistical significance of pairs-wise comparisons of conditions. Each condition has 12 data points by itself, so merging data over positions in experiment 1 and look directions in experiment 2 yields $4 \times 12 = 48$ data points for each experiment overall. The data proved to be consistent enough to be merged.

Experiment 1 - Static Perspectives

The single position ratings are visualized in Figure 5.13a (Median and 95% confidence intervals). It shows that participants rated the *proposed* approach higher than the other conditions, with statistical significance ($p < 0.001$) at positions 1 to 3.

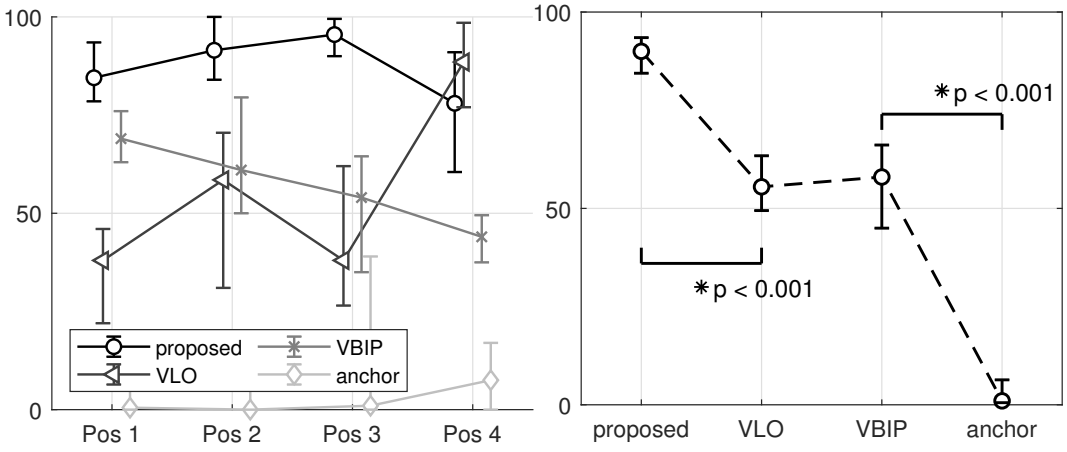
One exception to this is position 4: The *VLO* rendering shows significant higher rating ($p < 0.001$) than the same approach at the other positions and shows no significant difference to the *proposed* one ($p = 0.1963$). The comparison of merged *VLO* data from Position 1&3 on the one hand and 2&4 on the other shows a weak advantage ($p = 0.0049$) of the direct perspectives over interpolated.

The ratings of the *VBIP* show a decrease with distance, as seen in Fig. 5.11. The difference is significant when comparing the farthest and closest position ($p < 0.001$).

To assess the overall ratings, the datasets of positions 1 to 4 were merged and are pictured in Fig. 5.13b (Median and 95% confidence intervals). The *proposed* approach is rated higher ($p < 0.001$) than all other conditions. No notable difference between *VLO* and *VBIP* ($p = 0.4764$) in overall rating although unsurprisingly both show strong difference to the *anchor* ($p < 0.001$).

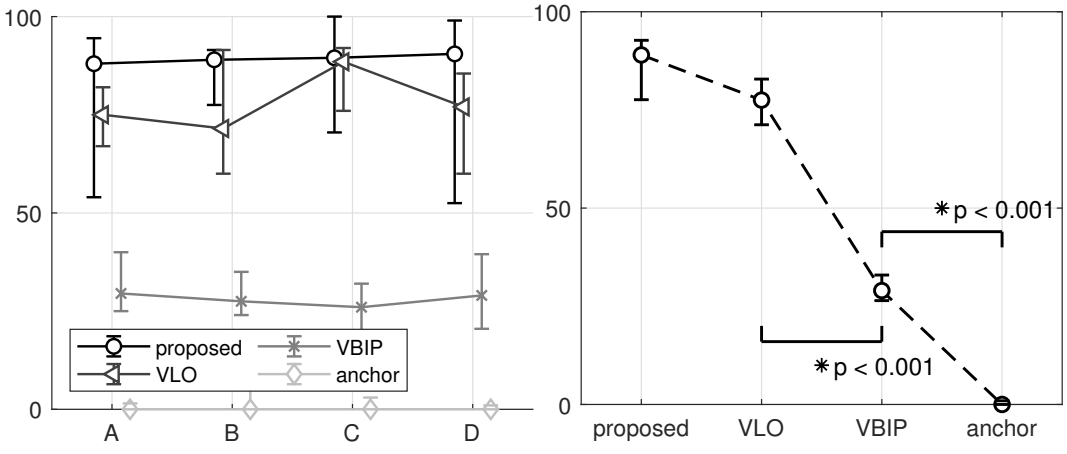
Experiment 2 - Dynamic Perspectives

The rating for all conditions is very similar over the look directions, as Fig. 5.14a shows. The *proposed* and the *VLO* methods are consistently rated higher ($p < 0.001$) than the *VBIP* condition and unsurprisingly the *anchor* ($p < 0.001$). Between *proposed* and *VLO* conditions, the difference is statistically not provable or weak at all look direction A ($p = 0.5332$), B ($p = 0.9766$), C ($p = 0.0596$) or D ($p = 0.3320$).
The merged data conveys a similar picture, as Fig. 5.14b shows. The results from the static-perspective experiment would imply a pronounced mean difference between overall rating of *proposed* and *VLO*, but ultimately, only a weak advantage ($p = 0.0624$) could be determined in the dynamic scenario.



(a) Single Conditions/Positions (b) Merged position data

Figure 5.13: The results of the first experiment comparing static perspectives for (a) single position data and conditions as well as the (b) merged-position data. Pictured are median, 95% confidence intervals and notable statistical significance is marked.



(a) Single Conditions/Look directions (b) Merged look direction data

Figure 5.14: The results of the second experiment comparing dynamic perspectives for (a) single look direction data and conditions as well as the (b) merged-look-direction data. Pictured are median, 95% confidence intervals and notable statistical significance is marked.

5.2.3 Discussion

Despite the relatively small number of participants, some of the statistical results were significant. The listening evaluation confirms the effectiveness of the proposed method by providing comparative ratings to references and established methods. The evaluated aspect is the overall similarity to a reference. The first experiment, presenting static perspectives facing a speaker in a room shows significant improvement over existing methods. The second experiment, evaluating the same aspect but with moving perspectives, did not turn out as conclusive but supports the results of the first.

The proposed method is rated consistently very close to the reference and consistently better than the alternatives under test under various settings. We can observe a slight drop off at the farthest position but this is most likely due to the very high rating of the *VLO* approach as it is a comparative evaluation with a limited scale.

VBIP shows some effectiveness at locations close to the sound object, and its rating deteriorates at larger distance. This is mostly due to the fact that this approach delivers a recording of the sound object that subjectively appears to be too dry and auditory cues for the room impression such as reverberation and early reflections are not prominent enough. Moreover, the directional diversity of the *VBIP* approach is not very high, as only 4 virtual playback directions are used.

The *VLO* approach shows ratings seemingly depending on the listener position. The significant increase in rating at position 4 is due to the fact that this position is a direct microphone array position and far from a source. The direct signal is accurately encoded by the surround perspective of the microphone position, just as in *VBIP* with additional cues from the other perspectives providing a better room impression. The interpolated positions have lower rating as the spatial reproduction suffers. The excellent rating for listening positions that are far away from the source might owe to the rich diversity of VLOs and their directions, which appear to be capable of rendering a convincing envelopment from the low-resolution single-perspective recordings.

The dynamic experiment shows a stark increase in ratings for the *VLO* approach as the interpolation is perceived smoother and is rated almost as high as the proposed method. The residual signals of the *proposed* approach are based in the concept *VLO* therefore the smoothness and general impression appears similar. However the direct signal extraction provides better object localisation for the listener and hereby achieves the improvement intended.

Chapter 6

Conclusion

In this work, an approach to analyse acoustic scenes and extract data for interactive playback for virtual listeners was developed. The goal was to detect and track sound objects in-scene so that their positional information can be used for signal extraction and composition of a virtual listener perspective. The strength of the proposed approach is that it was shown to already operate well with a frequency-independent signal conditioning for the real-time rendering.

The first step of analysis, using in-scene perspective surround (FOA) recordings to calculate directional distributions of acoustic activity in the spherical harmonics domain, was established by expanding on existing research. This was achieved using direction-of-arrival estimation and weighted cumulation of directional signal energy. Following this, a method of merging perspective-directional information into a three-dimensional acoustic map was developed. The method is based on the inverse spherical harmonics transformation and perspective weighting to compute acoustic map values at positions between perspective positions.

The acoustic map leads to the introduction of the concept of particle filters to track acoustic activity. The well-known method uses acoustic map values for the computation of particle weights and proved effective in delivering trajectories for detected sound objects.

Additionally, a probabilistic birth-death algorithm was introduced to detect the emergence and disappearance of sound objects in the acoustic scene. The algorithm performs the inverse SH transform on the directional distributions as well as perspective merging to evaluate an equidistant grid of positions and a sequential peak picking algorithm to extract frame-based observations of acoustic activity. These observations feed probabilistic estimations by calculating transitional probabilities to hypotheses of object existence and activity. This lays ground for the management of particle filter instances, such as creation and removal, which was described in detail.

To counteract the unreliably detected onsets of emerging sound objects, the concept of

anti-causal tracking was introduced. Since the prediction step of a particle filter system is easily adapted to operate along the anti-causal time direction and the birth-death algorithm is applicable anti-causal as well, this was adopted as an effective solution.

The effectiveness of the object detection algorithm was assessed by conducting a numerical evaluation, focusing on positional accuracy and object activity detection.

The first component of the 6DoF rendering algorithm, the object direct signal extraction, was developed based on a composite minimum-variance-distortionless-response (MVDR) beamformer and perspective merging. This allows to renderer direct sound with high definition in terms localization. Additionally, a method to approximate the residual signal was developed, which is necessary to preserve and convey the room information, e.g. for distance perception , immersion, and envelopment.

Both components in tandem proved to be working well in providing a convincing experience to listeners when they are moving in the virtual reconstruction of the sound scene. This was evaluated by listening experiments assessing the perception of static and moving listener perspectives in terms of naturalness and using a comparison to a reference.

In combination, both experiments suggest that the proposed method is a viable alternative to the known methods to interpolate multi-perspective recordings, supporting improved spatial definition and therefore a superior listening experience.

Outlook

The analysis of sound scenes is not an efficient task as of now, therefore a possible continuation of this research could be the optimization of procedures and a faster implementation. Moving from MATLAB to more effective ways of achieving computation could speed up the process. Real-time analysis seems to be a unrealistic goal but suggests itself anyway.

The method, as described in this work, is effective for broadband signals. As another possible continuation, research on band-wise operations suggests itself. This could lead to improved object detection in dense acoustic scenes.

The algorithm currently makes no attempt in estimating the room geometry, which could prove useful for future work for better estimation of the object positions or the residual signal by including wall-reflections into estimation techniques.

Lastly, due to the ongoing COVID-19 pandemic at the time of authoring this thesis and the resulting problems in availability of university facilities and equipment, an experiment part of the listening evaluation had to be scrapped. A setup using the commercial VR hardware HTC VIVE was developed to evaluate the listening experience in a fully interactive virtual environment. A strong suggestion for future work would be to pursue and conduct of the evaluation in question.

Bibliography

[AK17] A. Allen and B. Kleijn, "Ambisonic soundfield navigation using directional decomposition and path distance estimation," in *4th International Conference on Spatial Audio (ICSA)*, Graz, Austria, September 2017.

[BB10] S. Berge and N. Barret, "High angular resolution planewave expansion," in *Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, Paris, France, 2010.

[BOS08] A. Brutti, M. Omologo, and P. Svaizer, "Localization of multiple speakers based on a two step acoustic map analysis," 05 2008, pp. 4349 – 4352.

[BOS10] ——, "Multiple source localization based on acoustic map de-emphasis," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, 01 2010.

[CCF99] J. Carpenter, P. Clifford, and P. Fearnhead, "An improved particle filter for non-linear problems," *IEE Proceedings Radar Sonar and Navigation*, vol. 146, no. 1, pp. 2–7, 2 1999.

[CJ20] E. Y. Choueiri and T. Joseph, "System and method for virtual navigation of sound fields through interpolation of signals from an array of microphone assemblies," US Patent US 2020/0021940 A1, 2020.

[DdFG01] A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*, 1st ed., ser. Information Science and Statistics. Springer-Verlag New York, 2001.

[Dep17] T. Deppisch, "Erstellung einer walkthrough-fähigen webapplikation unter verwendung von ambisonics und binauraler auralisation," Bachelor Thesis, Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Graz, Austria, 2017.

[EBU08] EBU, *Sound Quality Assessment Material recordings for subjective tests*, 2008, https://tech.ebu.ch/publications/sqamcd.

[Efr93] B. Efron, *An introduction to the Bootstrap*, 1993.

[Fea98] P. Fearnhead, "Sequential monte carlo methods in filter theory," Ph.D. dissertation, University of Oxford, 1998.

[GZS+18] P. Grosche, F. Zotter, C. Schörkhuber, M. Frank, and R. Höldrich, "Method and apparatus for acoustic scene playback," Patent WO2018077379A1, 2018.

[Hac15] P. Hack, "Multiple source localization with distributed tetrahedral microphone arrays," Master's Thesis, Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Graz, Austria, 2015.

[HCXC10] Z. He, A. Cichocki, S. Xie, and K. Choi, "Detecting the number of clusters in n-way probabilistic clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 2006–2021, Nov 2010.

[Hil82] E. Hille, *Analytic Function Theory*, 2nd ed. Chelsea Publishing Company New York, 1982, vol. 1.

[Hol79] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979. [Online]. Available: http://www.jstor.org/stable/4615733

[IR96] J. Ivanic and K. Ruedenberg, "Rotation matrices for real spherical harmonics. direct determination by recursion," *The Journal of Physical Chemistry*, vol. 100, no. 15, pp. 6342–6347, 1996. [Online]. Available: https://doi.org/10.1021/jp953350u

[ITU15] ITU, "ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems," 2015.

[KG18] S. Kitić and A. Guerin, "Tramp: Tracking by a real-time ambisonic-based particle filter," *LOCATA Challenge Workshop*, 2018.

[Kit96] G. Kitagawa, "Monte carlo filter and smoother for non-gaussian nonlinear state space models," *Journal of Computational and Graphical Statistics*, vol. 5, no. 1, pp. 1–25, 1996. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/10618600.1996.10474692

[KV96] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *Signal Processing Magazine, IEEE*, vol. 13, pp. 67 – 94, 08 1996.

[LC95] J. S. Liu and R. Chen, "Blind deconvolution via sequential imputations," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 567–576, 1995. [Online]. Available: https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.10476549

[LC98] J. Liu and R. Chen, "Sequential monte carlo methods for dynamic systems," *Journal of the American Statistical Association*, vol. 93, 04 1998.

[MC04] N. Macmillan and D. Creelman, *Detection Theory: A User's Guide*, 01 2004, vol. xix.

[Mül20] K. Müller, "Variable-perspective rendering of virtual acoustic environments based on distributed first-order room impulse responses," Master's thesis, Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Graz, Austria, 2020.

[NK17] A. Neidhardt and N. Knoop, "Binaural walk-through scenarios with actual self-walking using an htc vive," in *43. Jahrestagung für Akustik*, Kiel, Germany, 03 2017.

[NKKK17] A. Neidhardt, F. Klein, N. Knoop, and T. Köllmer, "Flexible python tool for dynamic binaural synthesis applications," in *142nd AES Convention*, Berlin, Germany, 05 2017.

[NR18] A. Neidhardt and B. Reif, "Minimum brir grid resolution for interactive position changes in dynamic binaural synthesis," in *30th Tonmeistertagung, Int. VDT Convention*, Cologne, Germany, 11 2018.

[Okt19] Oktava GmbH. (2019) Oktava mk-4012. [Online]. Available: http://www.oktava-shop.com/images/product_images/popup_images/4012.jpg

[PDP15a] A. Politis, S. Delikaris-Manias, and V. Pulkki, "Direction-of-arrival and diffuseness estimation above spatial aliasing for symmetrical directional microphone arrays," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 6–10.

[PDP15b] ——, "Direction-of-arrival and diffuseness estimation above spatial aliasing for symmetrical directional microphone arrays," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 6–10.

[PDPM15] D. Pavlidi, S. Delikaris-Manias, V. Pulkki, and A. Mouchtaris, "3d localization of multiple sound sources with intensity vector estimates in single source zones," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 1556–1560.

[PF06] V. Pulkki and C. Faller, "Directional audio coding: Filterbank and stft-based design," *Audio Engineering Society - 120th Convention Spring Preprints 2006*, vol. 4, 01 2006.

[Pol16] A. Politis, "Microphone array processing for parametric spatial audio techniques," Ph.D. dissertation, Aalto University, 2016.

[PP15] T. Pihlajamaki and V. Pulkki, "Synthesis of complex sound scenes with transformation of recorded spatial sound in virtual reality," *J. Audio Eng. Soc*, vol. 63, no. 7/8, pp. 542–551, 2015. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=17840

[PST+18] A. Plinge, S. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, and E. Habets, "Six-degrees-of-freedom binaural audio reproduction of first-

order ambisonics with distance information," in *AES International Conference on Audio for Virtual and Augmented Reality (AVAR)*, Redmond, WA, USA, 08 2018.

[PTP18] A. Politis, S. Tervo, and V. Pulkki, "Compass. coding and multidirectional parameterization of ambisonic sound scenes," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Calgary, Canada, 2018.

[Puc16] M. Puckette, "Pd documentation chapter 2: theory of operation," 2016.

[Pul06] V. Pulkki, "Directional audio coding in spatial sound reproduction and stereo upmixing," 01 2006.

[RK89] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

[Rud19] D. Rudrich, *IEM Plugin Suite*. IEM, 2019, https://plugins.iem.at/.

[RZF17] D. Rudrich, F. Zotter, and M. Frank, "Evaluation of interactive localization in virtual acoustic scenes," in *DAGA 2017 Kiel*, Kiel, Germany, 2017.

[Sär13] S. Särkkä, *Bayesian Filtering and Smoothing*, ser. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2013.

[Sch86] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION*, vol. AP-34, no. 3, 1986.

[Sny87] J. P. Snyder, "Map projections: A working manual," U.S. Government Printing Office, Tech. Rep., 1987.

[SZH18] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural rendering of ambisonic signals via magnitude least squares," in *Fortschritte der der Akustik (DAGA)*, Munich, Germany, 03 2018.

[TC16] J. G. Tylka and E. Choueiri, "Soundfield navigation using an array of higher-order ambisonics microphones," in *Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality*, Sep 2016. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=18502

[TC19] J. G. Tylka and E. Y. Choueiri, "Domains of practical applicability for parametric interpolation methods for virtual sound field navigation," *J. Audio Eng. Soc*, vol. 67, no. 11, pp. 882–893, 2019. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=20702

[TDTH13] O. Thiergart, G. Del Galdo, M. Taseska, and E. A. P. Habets, "Geometry-based spatial sound acquisition using distributed microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2583–2594, 2013.

[TH13] O. Thiergart and E. Habets, "An informed lcmv filter based on multiple instantaneous direction-of-arrival estimates," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 05 2013.

[Tyl19] J. G. Tylka, "Virtual navigation of ambisonics-encoded sound fields containing near-field sources," Ph.D. dissertation, Princeton University, 2019.

[VMR07] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Elsevier Science*, vol. 55, no. 3, pp. 216–228, 2007.

[WEJ12] A. Wabnitz, N. Epain, and C. Jin, "A frequency-domain algorithm to upscale ambisonic sound scenes," 03 2012, pp. 385–388.

[WEJvS10] A. Wabnitz, N. Epain, C. Jin, and A. van Schaik, "Room acoustics simulation for multichannel microphone arrays," in *Proceedings of the International Symposium on Room Acoustics (ISRA)*, Melbourne, Australia, August 2010.

[Whi98] D. Whitley, "A genetic algorithm tutorial," *Statistics and Computing*, vol. 4, 10 1998.

[Wil45] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945. [Online]. Available: http://www.jstor.org/stable/3001968

[Wil16] T. Wildling, "System parameter estimation of acoustic scenes using first order microphones," Master Thesis, Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Graz, Austria, 2016.

[WLW03a] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 826–836, Nov 2003.

[WLW03b] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, vol. 11, no. 6, 2003.

[WVHK06] E. G. Williams, N. Valdivia, P. C. Herdic, and J. Klos, "Volumetric acoustic vector intensity imager," *The Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 1887–1897, 2006. [Online]. Available: https://doi.org/10.1121/1.2336762

[ZF19] F. Zotter and M. Frank, *Ambisonics*, 1st ed., ser. Springer Topics in Signal Processing. Springer International Publishing, 2019, vol. 19.

[ZFSH20] F. Zotter, M. Frank, C. Schörkhuber, and R. Höldrich, "Signal-independent approach to variable-perspective (6dof) audio rendering from simultaneous surround recordings taken at multiple perspectives," in *DAGA Hannover*, 04 2020.