CHALMERS
UNIVERSITY OF TECHNOLOGY

Headphone Auralization of Acoustic Spaces Recorded with Spherical Microphone Arrays

Carl Andersson

Department of Civil Engineering
Chalmers University of Technology
Gothenburg, 2017

Master's Thesis BOMX60-16-03

Headphone Auralization of Acoustic Spaces Recorded with Spherical Microphone Arrays

Carl Andersson

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Civil Engineering
Division of Applied Acoustics
Chalmers University of Technology
Gothenburg, 2017

Headphone Auralization of Acoustic Spaces Recorded with Spherical Microphone Arrays

Carl Andersson

Supervisor: Jens Ahrens, Division of Applied Acoustics
Examiner: Jens Ahrens, Division of Applied Acoustics

Master's Thesis BOMX60-16-03
Department of Civil Engineering
Division of Applied Acoustics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Chalmers Reproservice
Gothenburg, 2017

Abstract

Binaural auralizations using headphones have become increasingly popular in virtual reality applications to create a realistic three dimensional sound scape. Conventional methods use dummy heads to capture the signals that arise at the ears of a listener. Such recordings cannot be played back with head-tracking, which is a significant limitation. Spherical microphone arrays allow capture of the spatial structure of a sound field. It is possible to impose the acoustic properties of the human body onto such a sound field and to calculate the signals that would arise at the ears of a listener who is coincident with the microphone array. Rotations of the listener, and therefore head-tracking, are straightforward to achieve.

This thesis investigates signal processing of sound fields captured with spherical microphone arrays in combination with spherical measurements of head related impulse responses. When implemented, such signal processing enables true three dimensional audio with possible head tracking for all rotations, without limitations on the number of simultaneous users.

The thesis has three major parts; the theory behind sound field expansions to the spherical harmonics domain and the related signal processing, controlled simulations to quantify the effects different signal processing chains have on the resulting binaural signals, and a user study to see the subjective effects of the proposed signal processing chains.

A magnitude-phase separated expansion method is proposed to reduce time domain errors due to discrete spatial sampling, which is evaluated both with simulations and in the user study. The proposed expansion method performs worse than previously used expansion methods for non-anechoic spaces, resulting in artificial echoes in the binaural sound signals.

Contents

1 Introduction

This report is in the field of binaural audio with focus on spatial audio signal processing using spherical harmonics expansion. Applications to spatial audio processing is in virtual or augmented reality systems, where the audio signals necessarily need to be localized both in time and space in order to present a good user experience.

The overall aim in this report is to study the effects of spherical harmonics expansion used to create binaural signals. One part of this will study the impact of spherical harmonics expansion applied on head related impulse responses (HRIR) and reverberant sound fields. The other part is the effects of using the expanded HRIR with an expanded room impulse response (RIR) in the spherical harmonics domain to create a set of binaural room impulse responses (BRIR) that also include the influence of the room.

1.1 Background

The oldest method to achieve binaural audio recording is using a dummy head with two microphones placed where the ear channels should be. This give a very good representation of the effects our hearing system uses to localize sound sources, but has issues in larger scale applicability. This is because the dummy head impinges a direction that only can be changed at the time of the recording, and only by mechanical means. This method does therefore not allow head tracking of multiple recipients or at playback at later times.

A digital approach called Dynamic Binaural Synthesis uses digital simulations of scenes or rooms to create binaural signals. This enables good use of head tracking to multiple recipients and playback of saved scenes, but can only be used with simulated environments. This system can therefore not be used to record live fields.

The method of choice in this project will instead be expansion of the incoming field and head related impulse response in spherical harmonics. This enables live recordings and playback of a sound field, with the possibility of head tracking to multiple recipients. The disadvantage of this method lies in the recording of the sound field, which requires a high number of spatial sampling points on a sphere. Since no system has yet been able to record a live sound field at a sufficient high spatial resolution, aliasing artifacts are introduced in the signal chain.

1.2 Related works

Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording is the title of the doctoral dissertation of Benjamin Bernschütz [1]. Said thesis

is the foundation on which this work is built, and contains most of the used theoretical work and related practical considerations. Many of the mathematical relations described in the above, which are needed for sound field analysis using spherical harmonics, are implemented in the MATLAB toolbox *Sound Field Analysis Toolbox* (SOFiA), which is one of the major tools used in this work [2]. The SOFiA toolbox primarily includes algorithms for calculating the spherical harmonics coefficients of a sound field measured at certain sampling points using a spherical microphone array.

SoundScape Renderer (SSR) is a software that can be used for binaural playback [3]. The software uses HRIR:s or BRIR:s for angles in the horizontal plane with a 1° resolution, stored in wave files. These HRIR are convolved with input audio signals in real-time to produce binaural audio signals, with the possibility of using a head-tracker to render a sound scene where the sound sources do not move when the listener rotates their head.

MCRoomSim is a MATLAB toolbox that can make basic simulations of room impulse responses [4]. This will be used to have easy access to a variety of noise free RIR and to enable simulated measurements with a variety of microphone arrays. MCRoomSim also has the option to produce spherical harmonics coefficients, which will be integrated with the SOFiA toolbox to enable comparisons between the fields descriptions using microphone array expansions and a reference set of coefficients.

2 Theory

2.1 Spherical Harmonics

The underlying physical equation for propagating sound is the wave equation

$$\nabla^2 p - \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = 0$$

where p is the pressure at some point $\vec{x}$, and c is the wave speed. If we assume a separable solution in spherical coordinates $p(\vec{x},t) = R(r)\Theta(\theta)\Phi(\varphi)T(t)$ and take the Fourier transform from $T(t) \to \hat{T}(\omega)$, this equation transforms to

$$\begin{aligned} \frac{d}{dr}\left(r^2\frac{dR}{dr}\right) + \left(k^2r^2 - n(n+1)\right)R = 0 \\ \frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) + \left(n(n+1) - \frac{m^2}{\sin^2\theta}\right)\Theta = 0 \\ \frac{d^2\Phi}{d\varphi^2} + m^2\Phi = 0. \end{aligned}$$

The solutions to these differential equations are known, and a general solution can be written as a linear combination of certain basis functions [5]. These basis functions consist of one set for the radial behavior, and one set for the angular behavior. The radial equation is solved by spherical Bessel functions

$$j_n(kr) = \sqrt{\frac{\pi}{2kr}} J_{n+1/2}(kr)$$
$$y_n(kr) = \sqrt{\frac{\pi}{2kr}} Y_{n+1/2}(kr)$$

where J and Y are the Bessel functions of the first and second kind. The angular equation is solved by functions called spherical harmonics $Y_n^m(\theta,\varphi)$. There are different conventions for how to define the spherical harmonics, that give different symmetries and normalizations. In this work multiple versions are used in parallel and conversions are applied when necessary. The choice of definition does not affect the overall theory but will influence some details, which will be brought to attention in the appropriate sections. The three most important versions are

$$Y_n^m = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta) e^{im\varphi} \tag{1}$$

$$Y_n^m = (-1)^m \sqrt{\frac{2n+1}{4\pi}\frac{(n-|m|)!}{(n+|m|)!}} P_n^{|m|}(\cos\theta) e^{im\varphi} \tag{2}$$

$$Y_n^m = (-1)^m \sqrt{\frac{2n+1}{4\pi}\frac{(n-|m|)!}{(n+|m|)!}} P_n^{|m|}(\cos\theta) \cdot \begin{cases} \sqrt{2}\cos(m\phi), & m > 0 \\ 1, & m = 0 \\ \sqrt{2}\sin(m\phi), & m < 0 \end{cases} \tag{3}$$

where P_n^m is the associated Legendre polynomial of order n and mode m. Convention (1) will in this work be referred to as complex asymmetric, (2) as complex symmetric, and (3) as the real convention.

For all three definitions the spherical harmonics are a normalized set of basis functions, that is

$$< Y_n^m, Y_{n'}^{m'} >= \int_S Y_n^m (Y_{n'}^{m'})^* d\Omega = \delta_{n,n'}\delta_{m,m'}$$

where S is a unit sphere and $d\Omega$ is the solid angle element on said sphere. The conjugate of a spherical harmonic is an important operation, and one that will differ in the three definitions. For the complex asymmetric (1) we have

$$(Y_n^m)^* = (-1)^m Y_n^{-m},$$

for the complex symmetric (2) we have

$$(Y_n^m)^* = Y_n^{-m},$$

and for the real convention the conjugate have no effect.

2.2 Spherical Harmonics Expansion

Since the spherical harmonics Y_n^m is a complete set for the solutions of the wave equation on a sphere, any function f that fulfills the wave equation can be written as a linear combination of the spherical harmonics [5]. That is

$$f = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} P_{nm} Y_n^m \tag{4}$$

where P_{nm} are the so called expansion coefficients for f. If f also depend on the frequency, so will the expanion coefficients. The coefficients can be calculated with the so called "expansion integral"

$$P_{nm} = \int_S f(\Omega,\omega) Y_n^m(\Omega)^* d\Omega \tag{5}$$

where the integral is over a given sphere S. The radius of the sphere used in the expansion will define a set of expansion coefficients, and the integral over another sphere will define another set of coefficients.

The relationship between the coefficient sets from spheres of different radii is given by the radial solutions of the wave equation. It can be shown that for internal problems, i.e. where the sources are outside the region of interest, one must choose the spherical Bessel function of the first order as the solutions for the radial part of the wave equation. Now, the relation between one set of coefficients $P_{nm}(r_0)$ from a sphere of radius r_0 and the set $P_{nm}(r)$ from a sphere of radius r can be shown to be

$$P_{nm}(r) = \frac{j_n(kr)}{j_n(kr_0)} P_{nm}(r_0)$$

when both surfaces are in free field [1]. If the surface is not an open surface in the free field or the expanded function f does not represent the sound pressure, this expression must change. The factor that removes the dependency on the expansion surface is called "radial filter", and can be used to remove the dependency of the expansion coefficients on the expansion surface. For the case of an open surface the radial filter is

$$d_n = \frac{1}{4\pi i^n j_n(kr_0)} \tag{6}$$

where the $4\pi i^n$ is included for convenience in a later stage, and r_0 is the radius of the expansion sphere. Another important example of this is when the expanded function is a combination of sound pressure and sound pressure gradient, e.g. sound measured with cardioid microphones. In this case the radial filter will be

$$d_n = \frac{1}{2\pi i^n (j_n(kr_0) - i j_n'(kr_0))}.$$

A third common scenario is the pressure expanded at a rigid surface. Now, the radial filters should also handle the scattering effects of the surface, as can be shown to be

$$d_n = \frac{1}{4\pi i^n \left(j_n(kr_0) - \frac{j_n'(kr_0)}{h_n'^{(2)}(kr_0)} h_n^{(2)}(kr_0) \right)}$$

where $h_n^{(2)}$ is the second spherical Hankel function.

2.3 Plane wave decomposition

Since the basic solution of the wave equation in Cartesian coordinates are plane waves, these also constitute a complete set for the solutions of the wave equation. This means that any function that fulfills the wave equation can be written as a combination of plane waves. Since the set of plane waves is a continuous set instead of a discrete set, the function f can be written as

$$f(\omega,\Omega) = \int_S D(\omega,\Omega')P(\Omega',\Omega))dS$$

where P is a unity plane wave arriving from the direction Ω' evaluated at the position Ω at a sphere, and D is a frequency dependent coefficient for that direction. The integral is to be taken over all possible direction from where plane waves can arrive. It can be shown that a unity plane wave has the spherical harmonics expansion

$$P(\Omega',\Omega) = 4\pi \sum_{n=0}^{\infty} \sum_{m=-n}^{n} i^n j_n(kr) Y_n^m(\Omega) Y_n^{m*}(\Omega').$$

if expanded at an open sphere. This forces the weighting coefficient D to be

$$D(\omega,\Omega') = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \frac{1}{4\pi i^n j_n(kr)} P_{nm}(\omega) Y_n^m(\Omega')$$

where P_{nm} are the spherical harmonics expansion coefficients of the function f. This gives us

$$f(\omega,\Omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} P_{nm}(\omega) Y_n^m(\Omega)$$

by inserting the expressions for D and P into the integral above. The weighting coefficient D is the plane wave components for f and the whole process is termed plane wave decomposition. If the original function f does not represent pressure at an open surface, this can be compensated by using the correct radial filters as

$$D(\omega,\Omega') = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} d_n(kr) P_{nm}(\omega) Y_n^m(\Omega') \tag{7}$$

2.4 Head Related Impulse Responses

Head related impulse responses (HRIR) are a way to describe the effects our outer ears, head and torso have on sound waves. They contain the aural cues that our hearing system uses to locate sound sources in space, such as level and time differences between the right and left ears. The HRIR for a specific direction is the response to a plane wave from that direction, at the left and right ear channels. The frequency domain counterpart to the HRIR is the head related transfer function (HRTF). This can be seen as a filter that our ears, head, and torso apply to all sound that we hear.

A useful approximation to the plane wave response is the response to a sound source at a sufficient large distance from the head. In practice, this is measured with small microphones placed in the ear channels of a test subject. The test subject in question can be either a human person, or an acoustic test dummy for more repeatable results. If this is measured for all directions, one at a time, the measurement set describe the information necessary for our hearing system to locate a sound source in space.

Since a HRTF can be seen as a filter, measurements of HRTF:s can be applied to audio signals that lack the aural locating cues. This will create an audio signal with a virtual position relative to the listener, the same position as the source position used in the recording of the HRTF. If the original audio signal contains reverberation or echoes, this method will place all reverberation and echoes at the same virtual position as the direct sound source. In other words, the spatial information about the room is lost in the process. One possible solution to this problem is to have all the sounds separated, and apply different HRTF:s to place the sounds at different virtual positions. This would require that the sounds have recorded in a manner that isolates the directions from each other. If the direct sounds are recorded in an anechoic environment, the reverberation can be added later using digitaly synthesized reverberation. This is, however, not possible for sounds recorded in a reverberant room.

There are two different kinds of fields relevant for the uses in binaural reproduction. The first field describe a physical sound field, such as sound originating from a source in a reverberant room. The second type of field is the head related impulse response field, which describe the influence of our head and outer ears on the sounds we hear. Since the HRTF is defined as the response at the ears to a plane wave, if we have a complex sound field described as a set of plane waves we can apply the HRTF to all of them. This will result in a set of sound signals $S^{l,r}$ that is the same sound as a person would hear if present in the sound field. For a field expressed as a set of plane waves, this can be written as

$$S^{l,r} = \int_S H^{l,r}(\omega,\Omega) D(\omega,\Omega) d\Omega \tag{8}$$

where D are the plane wave components for the sound field, as described in section 2.3. Since D itself is a function defined on a sphere, we can expand its conjugate in spherical harmonics, as

$$\begin{aligned}(D^*)_{nm} &= \int_S D^*(\Omega) Y_n^m(\Omega)^* d\Omega \\ &= \int_S \sum_{n'=0}^{\infty} \sum_{m'=-n'}^{n'} d_n^*(kr) P_{n'm'}^* Y_{n'}^{m'}(\Omega)^* Y_n^m(\Omega)^* d\Omega \\ &= \int_S \sum_{n'=0}^{\infty} \sum_{m'=-n'}^{n'} d_n^*(kr) P_{n'm'}^* a_{m'} Y_{n'}^{-m'}(\Omega) Y_n^m(\Omega)^* d\Omega\end{aligned}$$

where $a_{m'}$ is a different factor depending on the spherical harmonics convention used. For the complex asymmetrical convention $a_{m'} = (-1)^{m'}$, while for the complex symmetrical convention, $a_{m'} = 1$. For the real convention the conjugate expansion is not needed. Using the orthonormality of the spherical harmonics, we get

$$(D^*)_{nm} = d_n^*(kr) P_{n(-m)}^* a_m.$$

Using the expansion for the plane wave decomposition above and a spherical harmonics expansion of the HRTF, we can write

$$\begin{aligned}S^{l,r} &= \int_S \left(\sum_{n=0}^{\infty} \sum_{m=-n}^{n} d_n^*(kr) P_{n(-m)}^* Y_n^m(\Omega) a_m \right)^* \sum_{n'=0}^{\infty} \sum_{m'=-n'}^{n'} H_{n'm'} Y_{n'}^{m'}(\Omega) d\Omega \\ &= \sum_{n=0}^{\infty} \sum_{m=-n}^{n} d_n a_m P_{n(-m)} H_{nm}\end{aligned} \tag{9}$$

where the orthonormality for the spherical harmonics have been used again. Since the orthonormality for the real convention of the spherical harmonics does not require a conjugate on one of the terms (both are real), there is no need to expand the conjugate expression. The corresponding integral will then be given by

$$S^{l,r} = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} d_n P_{nm} H_{nm}.$$

It is possible to calculate the binaural signals for a rotated sound field by considering either a rotated set of HRTF coefficients or a rotated set of sound field coefficients. This involves so called Wigner-D functions or matrices and the theory is based on group algebra and tensors. In the simplified case of rotations in the horizontal plane it reduces to a multiplication of complex exponentials, as

$$S^{l,r} = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} d_n a_m P_{n(-m)} H_{nm} e^{-jm\alpha}$$

where α is defined as the counter-clockwise rotation of the HRTF (head) for a fixed sound field (room). This can be seen by rotating the azimuth angle in the definitions of the spherical harmonics.

It is also possible to calculate binaural room impulse responses (BRIR), simply by using room impulse responses (RIR) as the sound field input for the spherical convolution. This BRIR can then be convolved with an anechoic sound signal in the time domain to create a binaural audio signal that corresponds to the sound signal playing in the room described by the RIR.

3 Implementation

This section gives an overview of the signal processing needed to implement the theory of sound field expansions to the spherical harmonics domain, and calculations of binaural signals from a spherical harmonics description. Figure 1 shows a schematic over the signal processing from a measured (simulated) sound field to a final pair of binaural signals. Subsequent sections will describe what specific parts of this processing chain do, but the focus will not be on the actual implementations in code.

3.1 Expansion of fields

The aim of expansion methods is to calculate the spherical harmonics coefficients in (4) from the sampled frequency domain response at some sampling points on a sampling sphere. In this work, three major methods have been used for this purpose; a discrete approximation to the continuous integral (5), a standard least-squares fit of the coefficients to the measurement data, and a regularized least-squares fit to the measured data.

The discrete approximation to the continuous integral is the standard method used in the SOFiA toolbox. This is the method that closest represents an analytic calculation of the coefficients. As with discrete approximations to integrals in one dimension, in order to evaluate the expansion integral it is necessary to choose a suitable calculation grid, also known as quadrature grid. Different grids have different properties, mainly in how many grid points are necessary to resolve spherical harmonics up to a certain order [1]. If this criterion is not met the expansion will not be stable and spatial aliasing will be introduced. This can be compared with the Nyquist sampling criterion in one dimension. In general, the number of sampling points required to resolve the spherical harmonics of order N can be shown to be $\gamma(N+1)^2$, where γ is a constant depending on the type of grid. Typical values for γ are in the range [1,2]. For a given grid, each grid point will have a corresponding grid weight w_i related to the local density of the grid. The

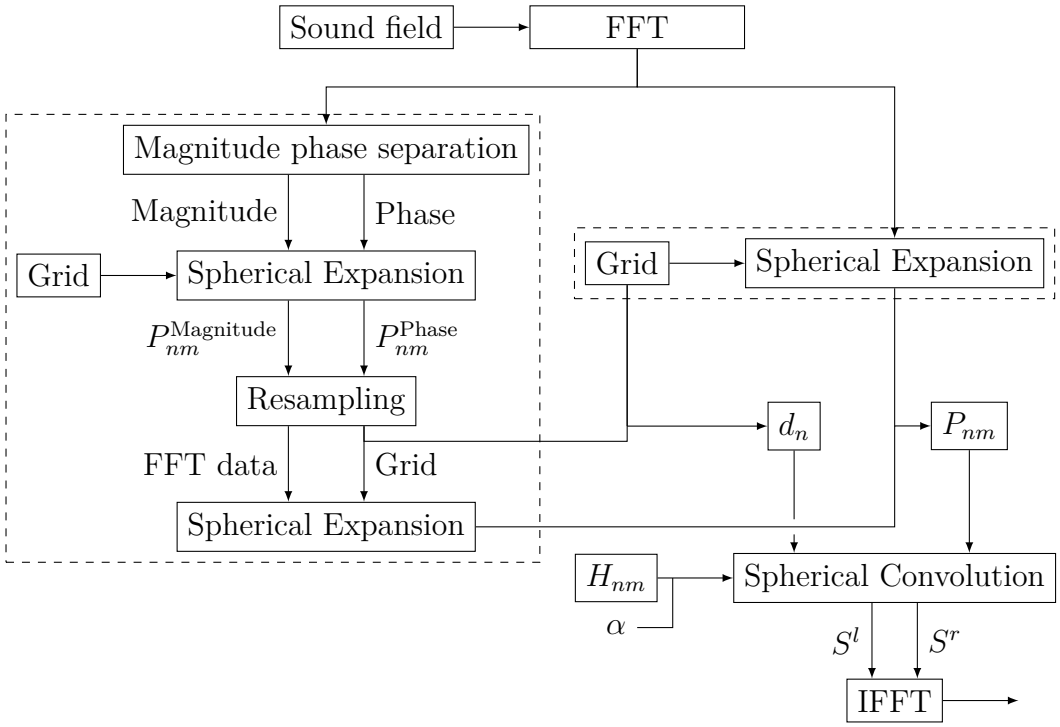


Figure 1: Schematic overview of the signal processing. The two dashed blocks are exclusive, so only one of the blocks is used to expand a given field. The left dashed block is the magnitude-phase separated expansion, white the right dashed block is a normal expansion. Not included in the schematic is the processing of the HRIR measurements. This processing is the same as for a sound field, except that no radial filters d_n will be generated.

theory of quadrature grids is well documented, and will not be covered further in this thesis. For a quadrature grid consisting of sampling points Ω_i the numerical approximation to the expansion integral is

$$P_{nm} = \sum_i f(\Omega_i) Y_n^m(\Omega_i)^* w_i$$

where $f(\Omega_i)$ is the field at the corresponding sampling point i. This expansion needs to be applied for each frequency bin.

Another possibility is to use a least-squares fit of the coefficients to the measured data. The field at a point Ω_i is approximated by

$$f(\Omega) = \sum_{n=0}^{N} \sum_{m=-n}^{n} P_{nm} Y_n^m(\Omega)$$

i.e. the truncated variant of (4), which for many grid points can be written as a matrix equation

$$\boldsymbol{Y}\boldsymbol{P} = \boldsymbol{F}$$

where $\boldsymbol{Y}$ is a $l \times (N+1)^2$ matrix with the values of the spherical harmonics at the grid points, $\boldsymbol{P}$ is a $(N+1)^2 \times 1$ vector with the corresponding expansion coefficients, and $\boldsymbol{F}$ is a $l \times 1$ vector with the values of the field at the grid points. Assuming that the number of grid points is larger than the number of spherical harmonics coefficients this will be an overdetermined equation system, with the expansion coefficients as the unknowns. A least-square solution $\hat{\boldsymbol{P}}$ to this equation is the one that minimizes the squared error $|\boldsymbol{F} - \boldsymbol{Y}\hat{\boldsymbol{P}}|^2$. This can be calculated in more than one way, but the approach used in this work is based on the pseudo-inverse of $\boldsymbol{Y}$, i.e.

$$\boldsymbol{Y}^\dagger = (\boldsymbol{Y}^*\boldsymbol{Y})^{-1}\boldsymbol{Y}^*$$

which give the solution as

$$\hat{\boldsymbol{P}} = \boldsymbol{Y}^\dagger \boldsymbol{F}.$$

As can be seen this does not have any formal requirements on the sampling grid, except that the number of grid points must be more than $(N+1)^2$. We do however expect that different kinds of grid will yield different results in the same way as the normal SOFiA approximation to the expansion integral. One extreme example would be a sampling grid that only has sampling points along the equator. This would obviously not be able to tell the difference between waves from above or below, but the least-squares method would still give a result. It it however a good method to use if the sampling grid was not chosen to any specific known quadrature so that the grid weights w_i are unknown, but the grid is still sufficiently dense and regular.

One version of the least-squares expansion method is the regularized least-squares expansion. This has been suggested as a good approach to expand fields where the sampling grid does not cover the entire sphere, in order to reduce the errors in the unsampled region [6]. The idea is very similar to the normal least-squares method, but the target is to minimize $|\boldsymbol{F} - \boldsymbol{Y}\hat{\boldsymbol{P}}|^2 + \lambda|\boldsymbol{D}\hat{\boldsymbol{P}}|$, where λ is a parameter that specifies the strength of the regularization, and $\boldsymbol{D}$ is a diagonal matrix that specifies which components of $\hat{\boldsymbol{P}}$ to regularize. The effects of regularization are that the specified components of $\hat{\boldsymbol{P}}$ are suppressed in the solution. A proposed choice of $\boldsymbol{D}$ has the diagonal elements $d_{l,l} = 1 + n(n+1)$, where the n is the order of the corresponding spherical harmonic in $\hat{\boldsymbol{P}}$. Choosing $\boldsymbol{D}$ in this manner will suppress higher order harmonics, which will have a smoothing effect on the field described by the resulting coefficients. The explicit solution to this minimization problem is

$$\hat{\boldsymbol{P}} = (\boldsymbol{Y}^*\boldsymbol{Y} + \lambda^2\boldsymbol{D})^{-1}\boldsymbol{Y}^*\boldsymbol{F}.$$

Note that by setting $\lambda = 0$ the normal unregularized least-squares solution is obtained.

3.2 Magnitude-phase separated expansions

The impulse response to a single plane wave is a pulse at a specific time. The impulse response to a single plane wave arriving some time later is a pulse at some time later. Intuitively the average impulse response to these two plane waves is a pulse at a time between them. However, the actual calculated mean of these two plane waves will be two pulses with half the magnitude of the original ones, occurring at the same times as the original pulses. This means that the average response to a group of delayed pulses will not be one single pulse at the average time of the pulses, but a series of pulses at the same times as the original pulses with a decreases magnitude. In this section this will be called *complex averaging*, since it is the result of averaging the real and imaginary part of pulses.

For the purposes in this work it would be preferable to find the average of a set of pulses as one pulse with the average magnitude and the average time delay, or phase. In order to do this, the pulse must be represented in the frequency domain using the magnitude and the phase of the pulse. The problem can be explained as follows: Given two pulses $r_1 e^{-i\varphi_1}$ and $r_2 e^{-i\varphi_2}$ we wish to construct the average pulse as $(r_1 + r_2)/2 \cdot e^{-i(\varphi_1+\varphi_2)/2}$. For simple cases like this it is possible to calculate the magnitude and phase of the individual pulses and average them separately, but for more complex cases, e.g. a diffuse sound field, where individual pulses cannot be differentiated this is not possible using such easy means.

This is relevant due to the fact that the response at the expansion surface between receiving grid points should be the average in magnitude and phase of the

surrounding points. When using expansion coefficients from the normal expansion methods, the SOFiA approximation and the two least-squares methods, to calculate the field at a location where there originally was no receiving grid point, the response at this grid point will be the *complex average* of the surrounding points. This section describes an attempt to give a better representation of the field at locations in between the sampling points by separating the magnitude and the phase of the field.

Since the magnitude and the phase of a complex description of a field are themselves fields it is possible to expand the magnitude and the phase in the spherical harmonics domain. This description of the magnitude and the phase will not suffer from the averaging issues described earlier, since the value calculated in between sampling points will have the average magnitude and the average phase of the surrounding points. This enables a more physical calculation of what the field should be in between the sampling points, so the magnitude and phase fields can be resampled at a higher grid density. Using these resampled magnitude and phase fields, a new complex field can be calculated. The idea is that the new complex dense field is what should have been measured in between the original grid points, with the modification that the first spherical harmonics expansion of the magnitude and the phase will smooth the fields so that the resolving power is limited. This new complex dense field can in turn be expanded in spherical harmonics to the same order as the original expansions, using the more dense grid. Since the distances between the grid points are smaller there are more points accurately represented by the expansion coefficients, which should decrease the overall averaging problems.

This magnitude-phase separated expansion step can be used with any combination of the expansion methods described above, but in this work all variants use the same expansion method for the initial magnitude and phase expansions and the final dense field expansion.

3.3 Convolution in the Spherical Harmonics Domain

In order to produce a usable binaural audio signal the spherical HRTF-field convolution integral (8) could be used. This would require methods not unlike the discrete solving of the expansion integral, using a quadrature grid and the values of the HRTF set and the plane wave decomposition of the field at said quadrature grid. The disadvantages of this arise when the HRTF and the sound field are not sampled at the same grid. In these cases it is necessary to expand at least one to spherical harmonics and recalculate the field at the sampling grid points of the other field. The subjective response to the binaural signals generated using this method has been shown to depend very much on the the quadrature grid used in the calculation of the convolution integral [1].

The method used in this work instead is based on the multiplication of the spherical harmonics coefficients shown in (9) where the spherical harmonics coefficients are used directly. This has the advantage of independence of matching of quadrature grids which give the possibility to use the most apt grid for each of the two fields. It will also reduce the computational load, both since the number of spherical harmonics coefficients is lower than the number of grid points so that the actual calculation of the integral is quicker, but also since the fields do not have to be recalculated at the quadrature grid points from the spherical harmonics description. This is an important consideration if the system is to be used for real-time applications using a microphone array in a live sound field. It is important to note that the theoretically infinite sum must be truncated at the lowest order of expansion of the sound field and the HRTF field. If one of the fields is expanded to a higher order than the other field, the surplus coefficients will simply be disregarded.

The radial filters described in section 2.2, e.g. (6), might have singularities or very large gain at certain frequencies or frequency ranges. This is problematic when working with fields from actual measurements since there will always be some noise and errors in the data, which would be amplified by the radal filters. Because of this it can be advantageous to limit the radial filters to some maximum gain, called the maximum modal amplification. There are several possible implementations of this but since the SOFiA toolbox already has an built in smooth limiter for the radial filters, that implementation is used. The limited radial filters is calculated as

$$\hat{d}_n = \frac{2a_{max}}{\pi} \frac{d_n}{|d_n|} \arctan\left(\frac{\pi}{2a_{max}}|d_n|\right)$$

where a_{max} is the maximum modal amplification, and d_n is the corresponding unlimited filter.

4 Simulations

To develop a general understanding of how much the different stages of the processing changes the final result, simulations were performed to compare different kinds of fields before and after processing. This section contains descriptions of the calculations and the results of said comparisons. Since all fields in this context are sampled at discrete grid points the overall error were used, instead of studying specific grid points. The chosen method for this overall error was the root mean square (RMS) error in magnitude and in phase. The RMS error in the magnitude is in this context defined as

$$E_{\text{mag}} = \sqrt{\left\langle \left(|H| - |\hat{H}| \right)^2 \right\rangle} \tag{10}$$

where H is the original signal, $\hat{H}$ is the reconstructed signal after processing, and the brackets signify averaging over the sampling positions in the array. The RMS error in the phase is defined in the same way, taking care to unwrap the phases first,

$$E_{\text{phase}} = \sqrt{\left\langle \left(\arg(H) - \arg(\hat{H}) \right)^2 \right\rangle}. \tag{11}$$

The magnitude errors are shown in decibels to encompass the entire range, but are not normalized by any means. One option would be to normalize the errors by the magnitude of the original signal, either using the mean value for all sampling points or individually for each sampling point. This would however not be a better representation of the actual error, since a small error at a frequency or sampling point that originally had a low signal would influence as much as a larger error at a frequency or sampling point with a higher original signal. This would not be a good measure of the perceived difference since a small error in a part of the signal that is not audible will still not be audible, but a larger error in a dominant part of the signal would be heard clearly. It should be noted that the error is not averaged over frequency bands, which would decrease the level of the high frequency errors.

The phase errors are shown as the difference in unwrapped phase calculated as above, in units of π. Since the slope in the unwrapped phase of an impulse corresponds to the time delay it is expected that the error in unwrapped phase is related to temporal differences in some way. It should be noted that since the unwrapped phase for a delayed signal is larger (in absolute terms) at higher frequencies, the difference in unwrapped phase will have the same behavior. This should not be interpreted as a larger difference, since the corresponding time delay is the same.

4.1 Expansions of Impulses

The response of a microphone array to an impulse plane wave was simulated using the built in tools in the SOFiA toolbox. These responses were expanded to the spherical harmonics domain using the different methods described in section 3.1. In order to study the behavior of the expansions, the sound field was reconstructed using the spherical harmonics representation and compared with the original field, as described above.

Figure 2 shows the RMS error in magnitude and phase for a field expanded using the standard SOFiA algorithm, for expansion orders between 3 and 60. It clearly shows that lower orders of expansion are less capable to resolve the higher frequencies, as expected from the theory. The change from the noise floor in magnitude error at $-300\,\text{dB}$ to the maximum error at $0\,\text{dB}$ is quite steep at the higher expansion orders but more gradual at the lower orders. When comparing

the errors in magnitude and the errors in phase, it indicates that the behavior is similar. The first rise in the phase error corresponds to the edge of the maximum error plateau in the magnitude. The phase error does however start its slope at this point, while the magnitude error ends its slope. This behavior is present for all the different expansion methods, and the differences lie in how large the errors are and how they change over frequency.

Figure 3 shows the RMS error in magnitude and phase when expanding an impulse plane wave field using different expansion methods, all for order $n = 12$. The three different methods described in section 3.1 were used both with and without separation of the magnitude and the phase in the expansion stage, see section 3.2. It is clear that for the SOFiA algorithm and the unregularized least squares fit, the separated magnitude-phase step is of little importance for the error in magnitude. It also shown that for a simple impulse field the least-squares fit performs equally good as the SOFiA algorithm. The regularized least-squares fit does however perform much worse in the lower frequencies, where the lowest error in magnitude is at $-80\,\mathrm{dB}$, compared to the $-300\,\mathrm{dB}$ level for the other two methods. In the considered frequency range there is no plateau for the lowest magnitude error, and the increase is more gradual than for the other methods.

It is clear that the phase is resolved very well up to 7 kHz, where the errors increase drastically. For the phase it does not seem to matter as much which method is used to expand the field, but more so if the separated magnitude and phase step was used. For all the three methods the magnitude-phase separation introduced a much larger error in the phase at the higher frequencies, contrary to what was expected. A more detailed comparison showed that the phase errors are overall increased using separated magnitude and phase, for all orders of expansion tested.

Figure 4 shows comparisons of the impulse responses of the array, using the SOFIA expansion algorithm with and without the magnitude-phase separation step. Comparing the original IR and the reconstructed IR after the normal expansion, it can be seen that the expansion will mix the responses from the receivers. This can be seen as the time extension of the responses along the equator all have the same start and the same end, but the primary part of the impulse matches the original. This mixing of the responses appears to be symmetric along the wave, i.e. the line from 0° to 180°. The magnitude-phase separated expansion also show the same behavior, but the whole response appears to be smoothed. The frontal direction (0°) and the rear receiver (180°) have a good representation of the direct sound, but with smooth versions of the other responses. The other directions lacks a sharp direct pulse, and instead only show smoothed versions.

The corresponding effect but in the frequency domain can be seen in Figure 5. The additional pulses from the mixing of the responses show up as a decreased

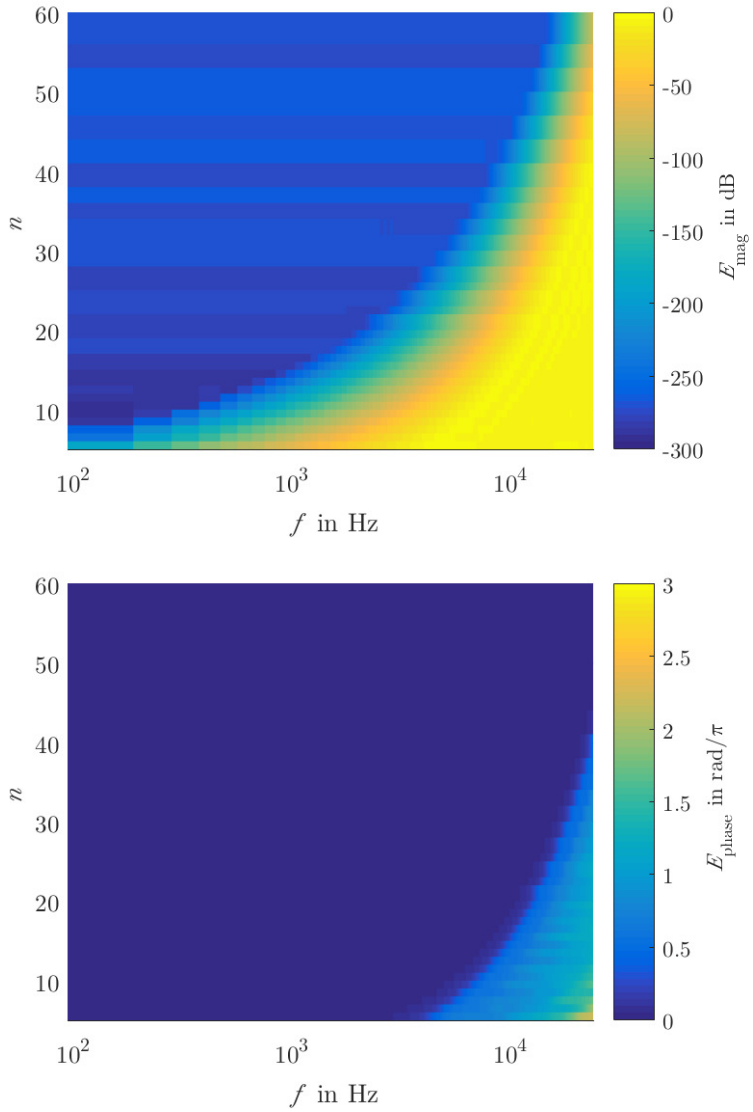


Figure 2: RMS error in magnitude (top) and phase (bottom), for simulated impulse plane waves expanded with the SOFiA algorithm for different orders.

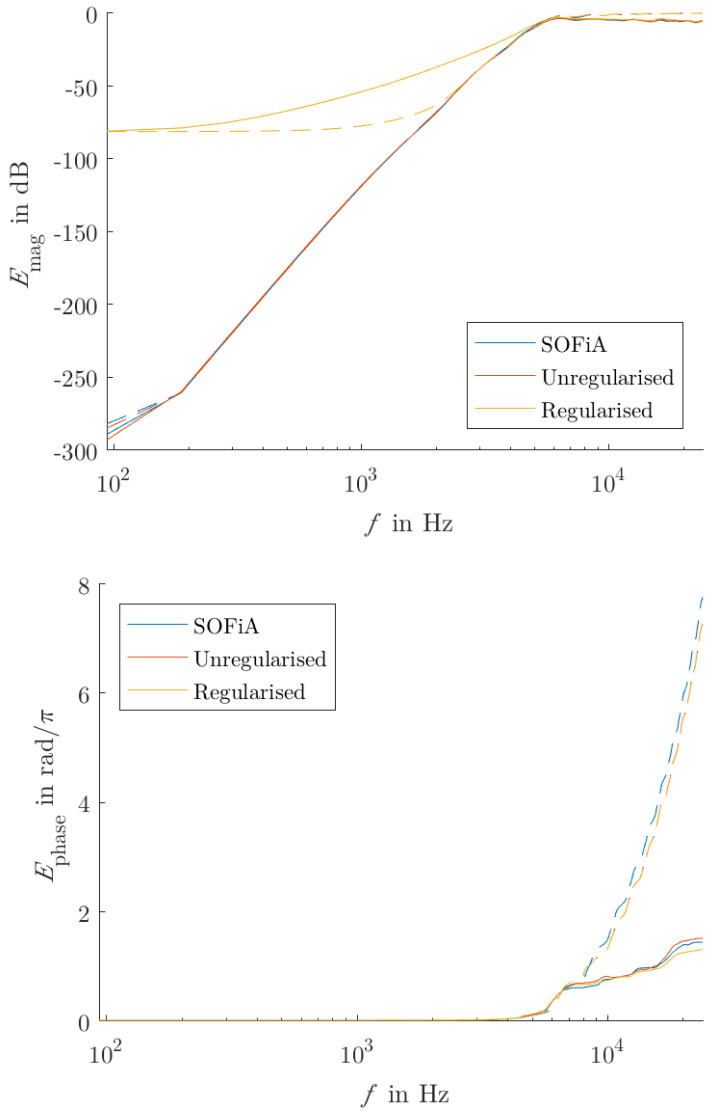


Figure 3: RMS error in the magnitude (top) and phase (bottom), for simulated impulse plane waves expanded using different methods for order $n = 12$. The dashed lines are for a separated variant of the corresponding solid line, see section 3.2.

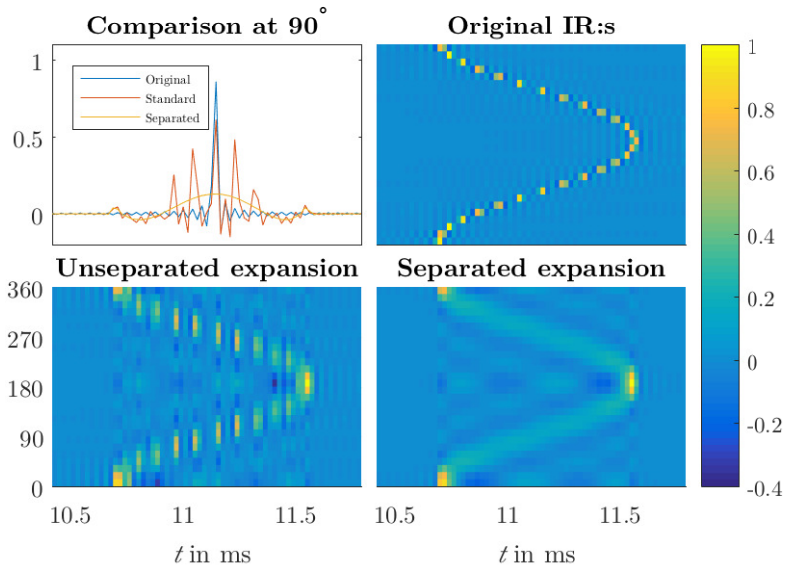


Figure 4: Comparisons between impulse responses from different expansion methods. The original measured plane wave impulse is shown, along with reconstructed responses after expansion using the normal SOFiA algorithm and the magnitude-phase separated version. The field was expanded at expansion order $n = 5$.

high frequency response at the sides of the array. The slope of the unwrapped phase is reduced for many directions. A possible explanation for this is that the multiple pulses introduce a symmetry that reduces the effect of any single pulse of the phase slope.

In short, the results from the expansions of impulse plane waves show that regularization increases the error in magnitude and the magnitude-phase separated expansion increases the error in the phase.

4.2 Expansions of Diffuse Room Impulse Responses

Since a diffuse sound field is a superposition of impulse plane waves, the general result from section 4.1 still hold. The difference is that for a diffuse field it is relevant how the different waves are related to each other. A simple measure of how diffuse a sampled field is how correlated the individual impulse responses are. Instead of studying the correlation the interesting measure is the maximum of the cross-correlation, which will be the same as the correlation of signals time shifted so that the direct sound occurs at the same time. This can be illustrated by two extreme cases; a pure plane wave and a pure noise signal. For the single plane wave, all the shifted signals will be perfectly correlated since they will be identical after the time shift. For the noise signal the signals will be fully uncorrelated

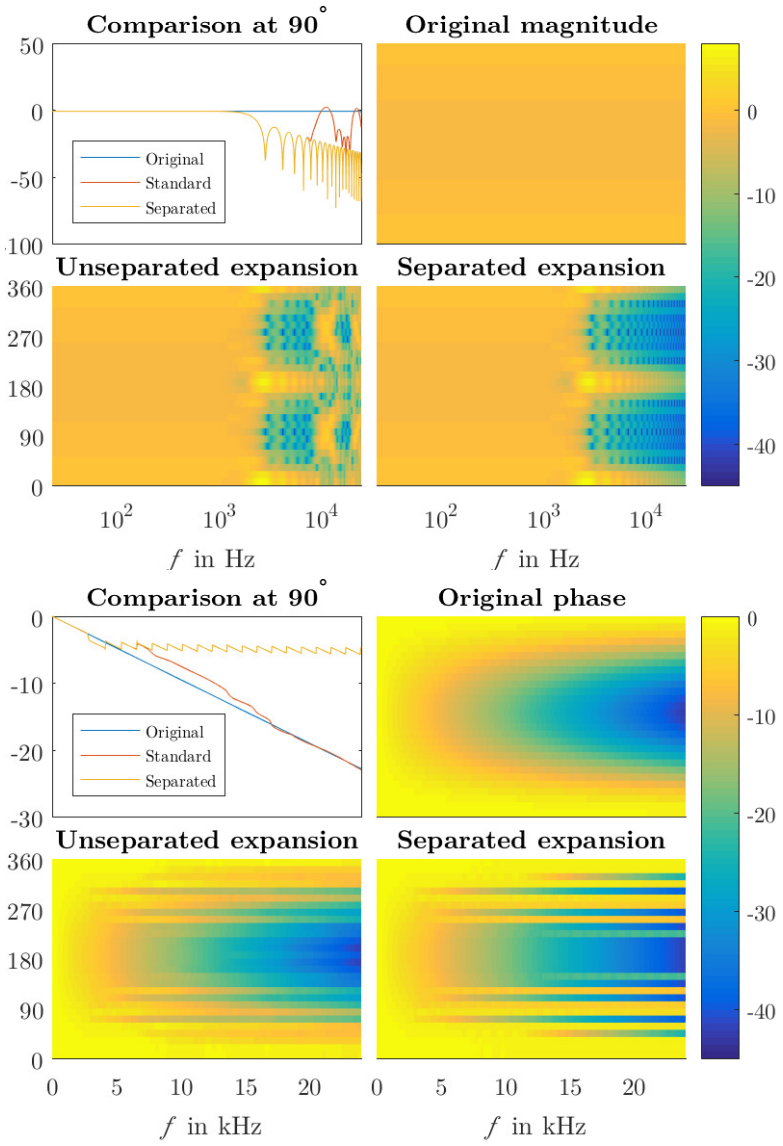


Figure 5: Comparisons between magnitude (top) and phase (bottom) from different expansion methods applied to a impulse plane wave field. The field was expanded at expansion order $n = 5$.

in the ideal case that the noise is uncorrelated. For a sound field consisting of a direct sound and a diffuse tail a measure of the diffuseness is the correlation of the shifted signals when the direct sound is disregarded, i.e. the maximum cross-correlation of the tail part only. In order to study the effects of spherical harmonics expansion on the diffuseness of a diffuse sound field, a field with an artificial reverberation was considered. The reverberation was implemented as a white noise with a exponentially decaying envelope. Since the reverberant tail was generated using new noise for each of the sampling positions, the resulting reverberation is as uncorrelated as the noise generator.

Figure 6 shows the correlation of the sound field before and after expansion followed by reconstruction at different orders and different expansion methods. For the study of different orders, the SOFiA method of expansion was used. The top left axis shows the correlations between the sampling positions at the equator of the sampling sphere. It is clear that the correlation between different channels is very low, as expected. The other three axes in the top half show the correlation of the same channels, but after expansion and reconstruction at the orders $n =$ 3,7,12. Interestingly the process introduces a symmetry in the signals, where the correlation increases for channels which are at opposite angles from the incident wave. There is also an increase in correlation for sampling point close to the the incident wave direction, or directly opposite to the incident direction. Both these two effects decrease at higher orders of expansion, and particularly the symmetry is negligible at order $n = 12$. It should be noted that the mean correlation between the other channels is not influenced enough to be considered a clear change. The bottom half of Figure 6 shows the correlation between the impulse responses at the equator for different expansion methods, all at expansion order $n = 3$. This shows that there is little difference between the SOFiA discrete evaluation of the expansion integral and the least-squares fit of the coefficients, but there is a larger difference between the unseparated expansions and the magnitude-phase separated expansions. Both the SOFiA method and the least-squares method have increased correlations between many of the responses when the magnitude-phase separation step is used.

To further understand the increase in correlation between the channels using the magnitude-phase separated expansion process, the impulse responses at different stages of this expansion process can be seen in Figure 7. The direct sound used in this study was a bell curve 20 samples long, since a 1 sample impulse was indistinguishable from the diffuse tail in the response. The diffuse tails were created in the same way as previously. The decaying tail can clearly be seen in the original responses, and the time difference between the front and the back portion of the sphere is seen as the curvature of the first wave front. For the responses reconstructed after expansion using the standard SOFiA algorithm the diffuse tail

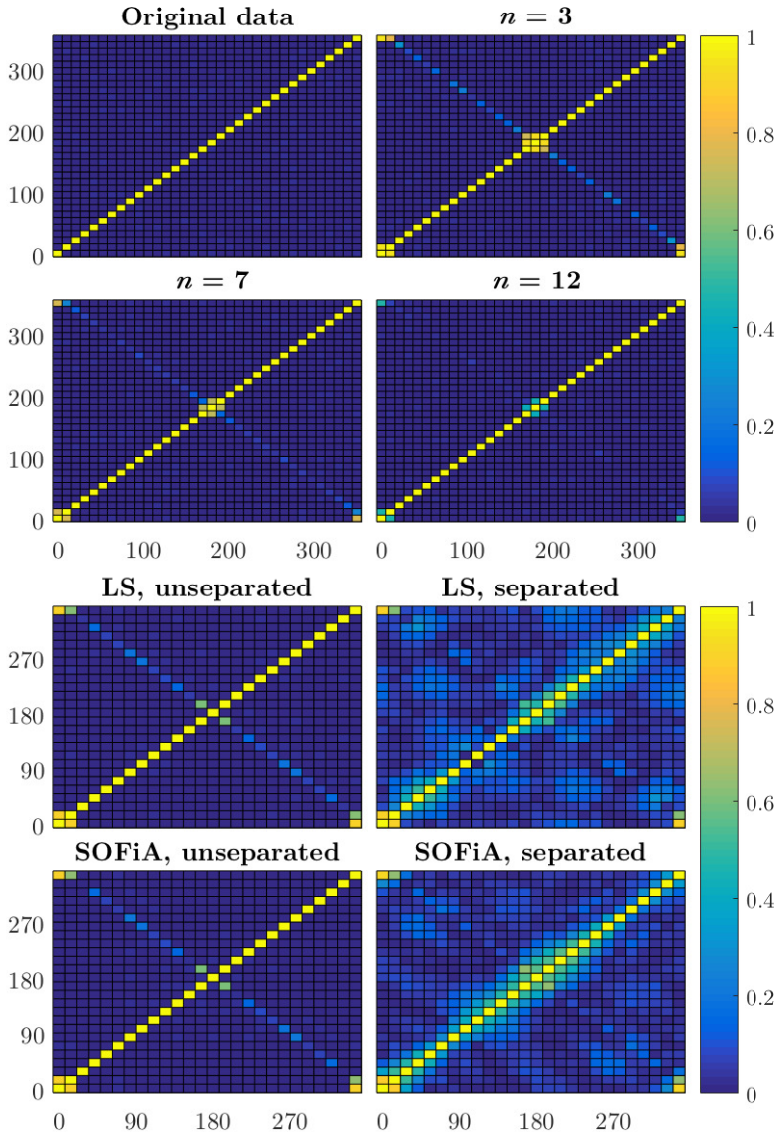


Figure 6: Correlation between sampling positions along the equator of an artificial diffuse sound field, using the standard SOFiA method of expansion at different orders (top), and different expansion methods at order $n = 3$ (bottom). Each colored square represents the correlation between the sampling point at the angles at one axis and the sampling point at the angle on the other axis. Since the correlation is symmetrical in nature, these plots are symmetrical around the main diagonal (lower left to upper right). It is clear that there is a mirror symmetry introduced, which is seen along the off-diagonal (upper left to lower right). This symmetry is reduced at higher orders of expansion. Note that the main diagonals have unity amplitude since the correlation of a signal with itself always is one.

is decreased in magnitude compared to the direct sound, but the decay is well-behaved. For these responses a non-causal portion can be seen in the channels at the back portion of the sphere. This is because the channels closest to the source will cause the expansion coefficients to have an equally early portion. This earliest part will spread to all the channels due to incomplete cancellation as a result of the non-perfect representation in the spherical harmonics domain. This is a clear example of why a magnitude-phase separated expansion process would be preferable.

The responses reconstructed from the separated magnitude and phase after resampling but before additional expansions, see section 3.2 and Figure 1, show a second wave front roughly 5 ms after the direct sound. This wave front is not a perfect echo, but a spread out and more diffuse version. The cause to this second wave front is not understood at this point, since the magnitude and phase responses to the corresponding separated impulse responses seem well-behaved. This "echo" is probably what causes the correlation between the channels to increase, but it is not clear if this actually reduces the subjective diffuseness of the sound field. The reconstructed impulse responses after the complete magnitude-phase separated expansion also show the same diffuse echo, but not as clear. The major difference seems to be a large reduction of the length of the diffuse tail in the responses. When comparing to the reconstructed responses after the standard unseparated expansion the separated-reconstructed responses have a stronger response in the first 5 ms after the direct sound, after which the tail decays much faster. The shown data is from a simulation using an impulse response 4096 samples long with a simulated reverberation time of 0.1 s. It was noted that a longer impulse response with longer reverberation time delayed the second wave front ever further. A 2 s simulated reverberation time increased the delay to approximately 75 ms and the diffuse spread out to roughly 30 ms. The longer reverberation time also influenced the strength of the second wave, to the point where the original direct sound was not apparent.

Figure 8 show the corresponding magnitude and phase of the same time signals seen in Figure 7. When comparing the four phase responses it is clear that the phase in the separated stage is much smoother than the original phase. The end result from the separated expansion have a smoother phase than the normal expansion method, which was the original intent. It is however apparent from the time responses that a smooth phase response for a diffuse field does not guarantee a good representation of the sound field.

4.3 Expansions of Head Related Impulse Responses

Since the head related impulse responses are very important in a binaural reproduction system, the effects of spherical harmonics expansion on a measured set of

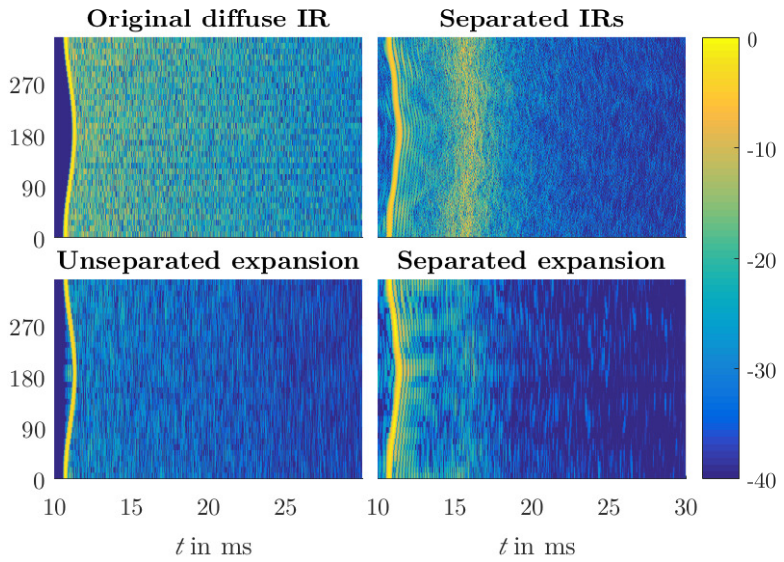


Figure 7: The impulse responses at the equator during the separated expansion of a diffuse sound field. The colored squares represent the magnitude of the impulse response at the angle on the y-axis and the time on the x-axis, measured in dB relative to the maximum of the corresponding impulse response. The top left axis shows the original diffuse sound field. The direct sound used here is a bell curve 20 samples long, since an impulse of 1 sample was too short to differentiate from the diffuse field. The bottom left axis shows the responses at the same grid points after expansion and reconstruction using the standard SOFiA algorithm. The top right axis shows the impulse responses reconstructed at the intermediate magnitude-phase separated stage, see section 3.2. The lower right axis shows the impulse responses after the complete expansion-reconstruction process using the separated expansion.

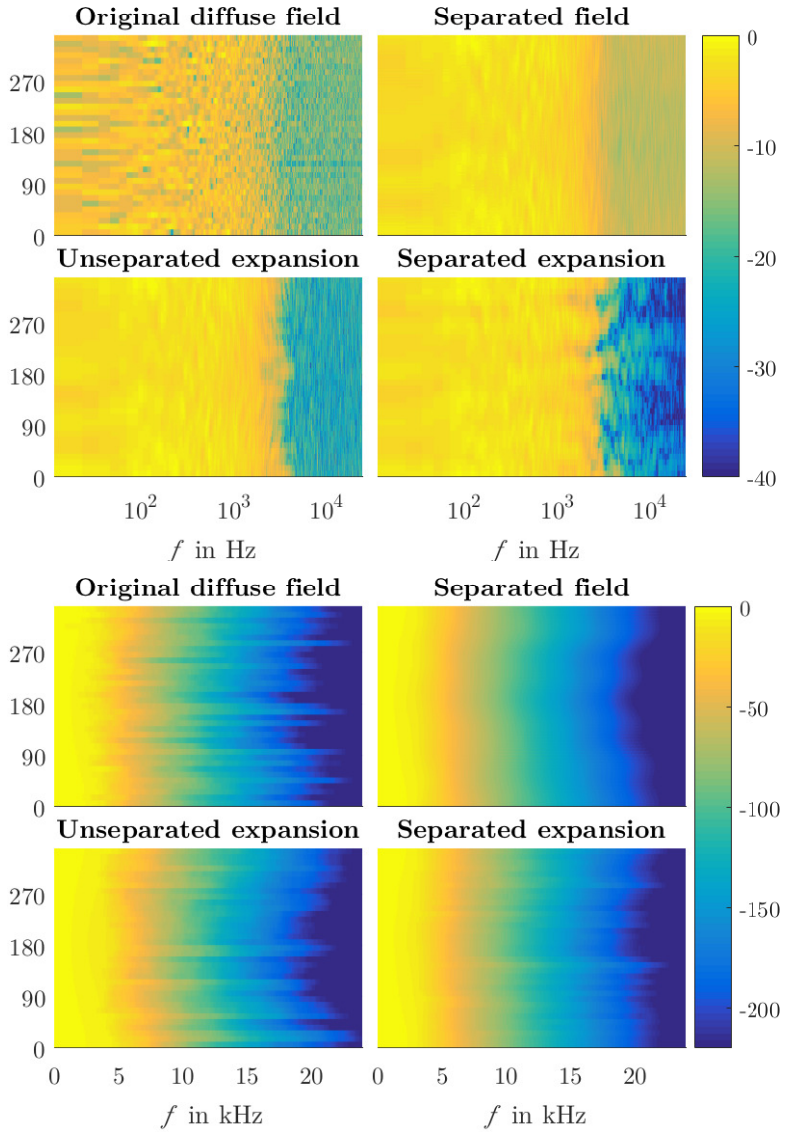


Figure 8: Comparisons between the magnitude (top) and the phase (bottom) of a diffuse field in different stages of the expansion, See Figure 7.

HRIR needs to be understood. Therefore a very dense set of HRIR was expanded to the spherical harmonics domain [7]. This set was chosen since it theoretically supports a very high order of expansion, which removes the possibility of spatial undersampling and ensures that the measured errors are from the actual expansion. In order to quantify the error introduced, the spherical harmonics expansion was used to reconstruct the original set of HRTFs, and the RMS difference in magnitude and phase was calculated according to (10) and (11) respectively.

Figure 9 shows the error in the magnitude and phase using the standard SOFiA algorithm for expansions of orders 3 to 60. This shows that the error in the spherical harmonics representation of the HRTF set decreases at higher orders of expansion, but not to the same degree as the expansions of impulse plane waves. Of special importance is the decrease in the error when increasing the order of expansion to $n = 30$, which significantly decreases the mean error in the high frequency range. The error in phase converges in the high frequency range around order 30, and does not continue to decrease. Together this indicates that there are diminishing returns from this point.

Figure 10 shows a comparison between a measured set of HRIR and the same set after expansion and reconstruction using the normal SOFiA method and its magnitude-phase separated variant. The effects of head shadowing can clearly be seen in the original HRIR:s, both as a delay in the initial response but also a decrease in the response magnitude. Comparing with the original responses, the reconstructed responses both show a mixing of the responses. This can clearly be seen for the responses around 270° as a small response before the actual wavefront. The reconstructed responses also show reduced amplitudes for the frontal directions. The frequency domain counterparts to this can be seen in Figure 11. The most notable difference in magnitude is at high frequencies for the frontal direction, where the magnitude have been reduced with between 15 dB and 30 dB. It is apparent that both expansion methods remove the smaller details from the original response. The phase responses have the largest difference around the 270° portion, where the phase slopes for the original response are much steeper. This is probably due to the mixing of the channels causing a response before the actual wavefront for the reconstructed HRIR:s. Both expansion methods show regular patterns at the high frequencies that are not present in the original response.

When comparing different methods of expansion for the HRTF set it was found that the regularized least-squares expansion did not have the same error decrease around the orders 20-30, see Figure 12. It can also be seen that the magnitude-phase separated methods perform worse in regards to magnitude errors. At higher orders of expansion, e.g. 40, the peak at 20 kHz for the SOFiA and least-squares methods had decreased to around −30 dB, compared to the −12 dB level at order 25. Both the regularized least-squares expansion and all three separated expansion

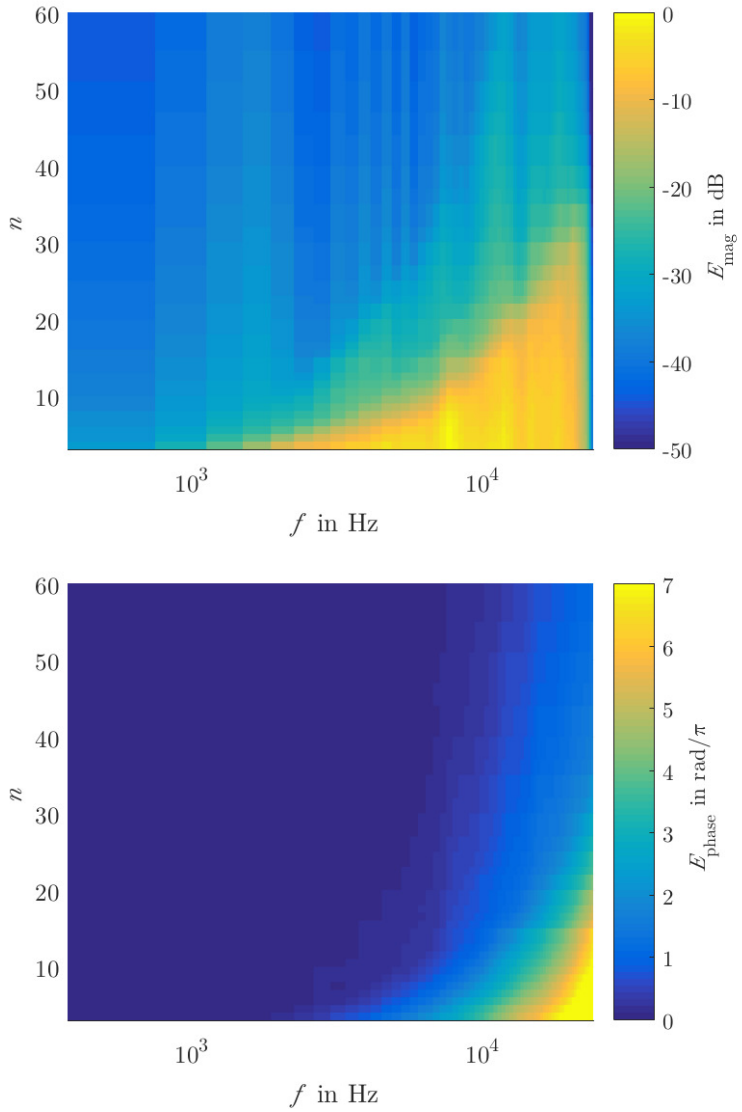


Figure 9: The RMS error in magnitude (top) and phase (bottom) for an expansion and reconstruction of head related impulse responses using the standard SOFiA algorithm, for expansion orders between 3 and 60. Note in particular the decrease in the magnitude error around orders 20 to 30.

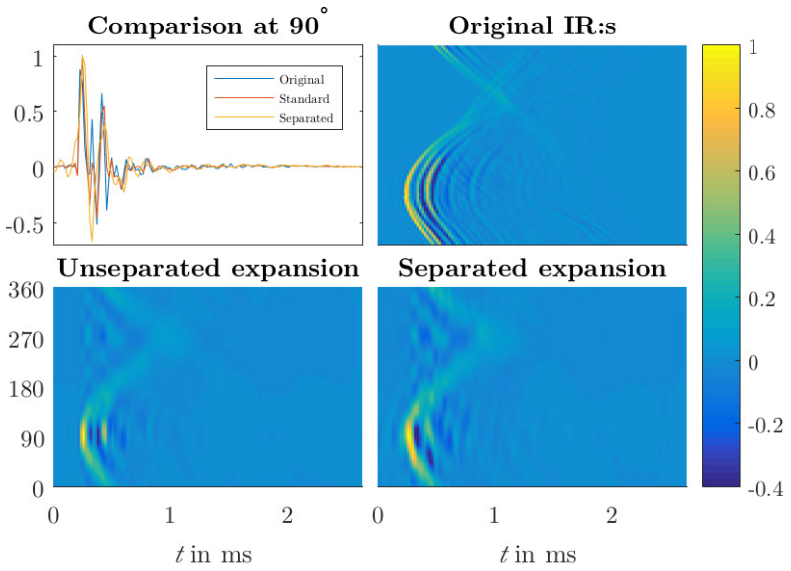


Figure 10: Comparison between different expansion methods used to expand a set of HRIR:s. Shown are the original HRIR set and the recombined HRIR:s using the normal SOFiA expansion method and the magnitude-phase separated variant.

have roughly the same RMS magnitude error at order 40 as at order 25. The phase errors in the different methods were more similar. At the lower orders of expansion, e.g. order 5, the normal regularized method was slightly better than the other methods at the highest octave band. At intermediate orders, e.g. order 12, the normal regularized method was slightly worse than the other methods from 4 kHz and higher, and the other methods were very close. At high orders of expansion, e.g. 25, the standard SOFiA method and the normal least-squares methods were slightly better than the other methods. Since the differences in the RMS error in phase are small among the different methods, no figures are included to show this.

4.4 Convolutions of Impulse Fields with Head Related Transfer Functions

The primary focus in this section will be the errors introduced in the spherical convolution between a HRTF field and a sound field, described in section 3.3. The numerical approximation of the convolution integral will not be studied, in favor of the multiplication of the expansion coefficients. Since the result from the convolution of the two fields depends on the two fields in question, the errors from the expansions will propagate to the convolution. The major new component in this part is the radial filters, see (7), which in turn depends on the type of expansion surface where the sound field has been measured. In order to investigate

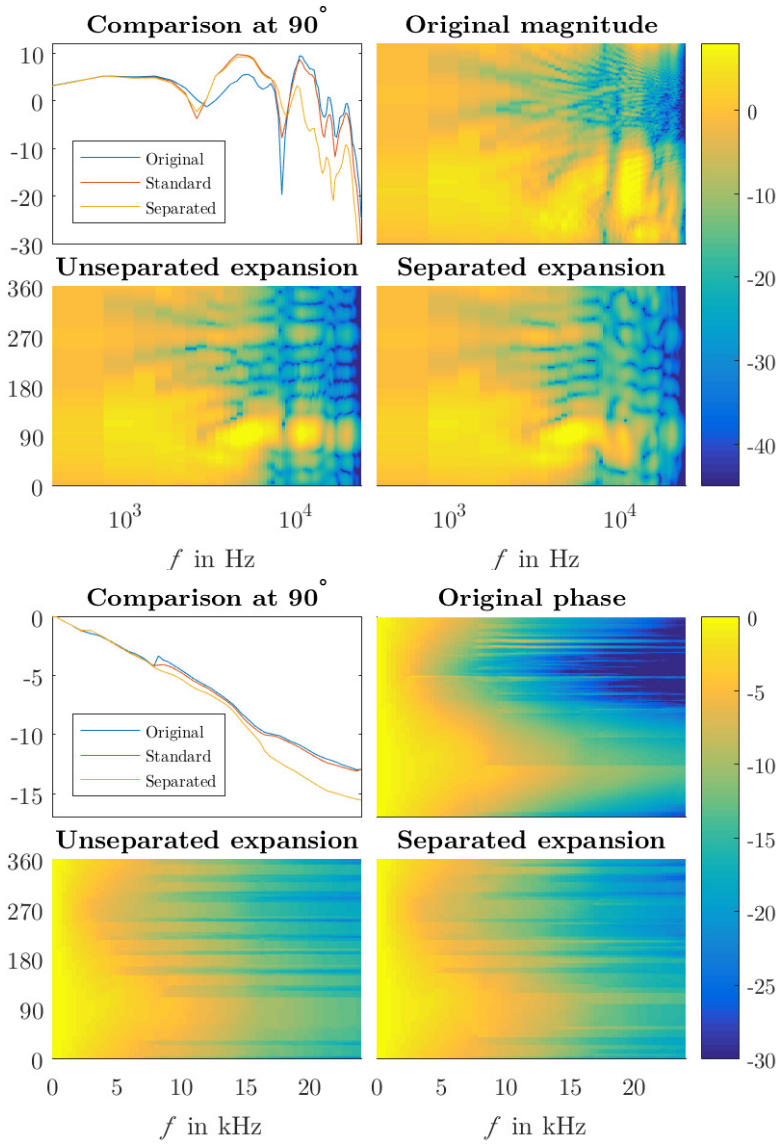


Figure 11: Magnitude (top) and phase (bottom) comparisons of different expansion methods used on a measured set of HRIR:s.

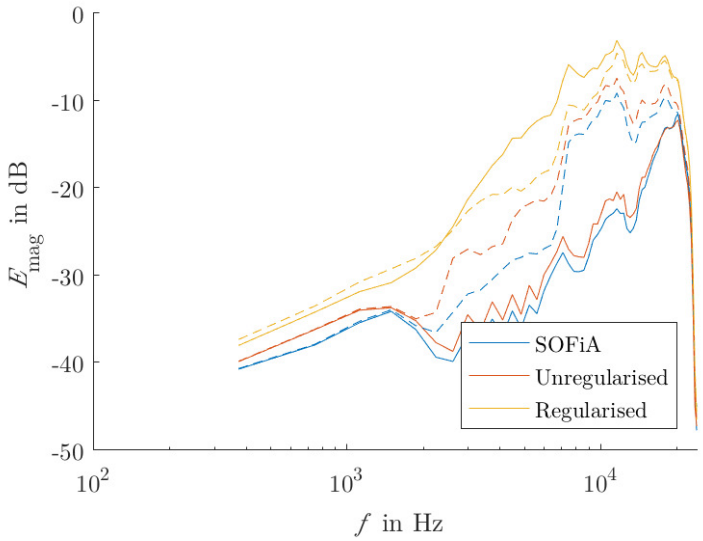


Figure 12: A comparison of the RMS error in magnitude for different methods of expansion of a HRTF set, for order $N = 25$. The solid lines correspond to normal unseparated expansions, while the dashed lines correspond to magnitude-phase separated expansions using the same method as the solid line of the same color.

these errors a HRTF set was expanded to the spherical harmonics domain. Instead of reconstructing the field directly as in section 4.3, the HRTF coefficients were used in a convolution with simulated coefficients for a plane wave. This was done for plane waves arriving from directions corresponding to the sampling grid points in the original HRIR set, which enables a comparison between the new calculated HRTF set and the measured HRTF set. The plane waves were simulated as the response of an array, so the simulation includes the effects of spatial sampling of the sound field.

As a first study, different array configurations were tested, all using order $n = 5$. Since the SOFiA toolbox supports both open and rigid arrays using either omni-directional or cardioid microphones, these were the configurations used in the test. However, the two rigid arrays were so similar that it was impossible to see any difference in the results. The following discussion will only consider the rigid array using omni-directional microphones, but the same results and comparisons is valid for both configurations. The initial test also includes two variants of each configuration, one using unlimited radial filters and one using radial filters limited to 0 dB.

Figure 13 shows the RMS error in the magnitude and phase for the three array configurations. When comparing the three array configurations with unlimited radial filters, it is clear that the open array with omni-directional microphones

has large errors in magnitude at the higher frequencies. This is likely due to that the radial filters for this array have singularities at certain frequencies, which amplify the errors from the expansions significantly [1]. The other two array configurations behave very similar in the higher frequency range, and all three configurations are indistinguishable at the lower frequencies. The three variants with a maximum modal amplification of 0 dB all have the same trend. The errors at the lower frequencies are increased by the limited modal amplification, while the errors at the higher frequencies are reduced. At the lower frequencies the rigid array has the lowest error, while in the higher frequency range the open array with cardioid microphones is slightly better. The open array with the omni-directional microphones also has the largest error in the phase, regardless of the maximum modal amplification. The other array configurations are very similar in average phase performance, and the maximum modal amplification does not seem to matter for the phase of the more stable arrays. It should be noted that while the open array with omni-directional transducers is the configuration that benefits the most from the limited modal amplifications, that also means that it is the configuration needs the most modifications to produce a stable result.

In order to more closely investigate the effects of maximum modal amplification, Figure 14 shows the RMS error for a rigid array configuration using different maximum modal amplifications. It shows that a high maximum modal amplification (18 dB) has roughly the same RMS error as using unlimited radial filters, as expected for a well behaved array configuration. A stronger limitation on the radial filters seems to increase the errors at the lower frequencies while decreasing the errors at the higher frequencies. However, there is diminishing returns in the higher frequency range but increasing errors at the lower frequencies. Of the cases shown here a maximum modal amplification of 0 dB seems to be a good balance between the frequency ranges.

Figure 15 show BRIR:s measured in two acoustic spaces before and after the whole signal processing chain, using different expansion methods. The three expansion methods shown are variants of the SOFiA expansion; one normal expansion, one with the separation stage applied on the field, and one with the separation stage applied on the HRIR set. For the anechoic space, the same general effects seem in section 4.3 can also be seen here. Of particular interest is that the separation step applied on the plane wave sound field seem to cause more smoothing than the separation step applied to the HRIR set. This can be compared with the BRIR from the small control room also used in the user study (see section 5). The original BRIR show a clear set of reflections as decaying repetitions of the original first wave front. Both the normal expansion and the HRTF separated version show the same reflections, considering the same smoothing effect mentioned above. The BRIR created with separation of the field does however have very strange behav-

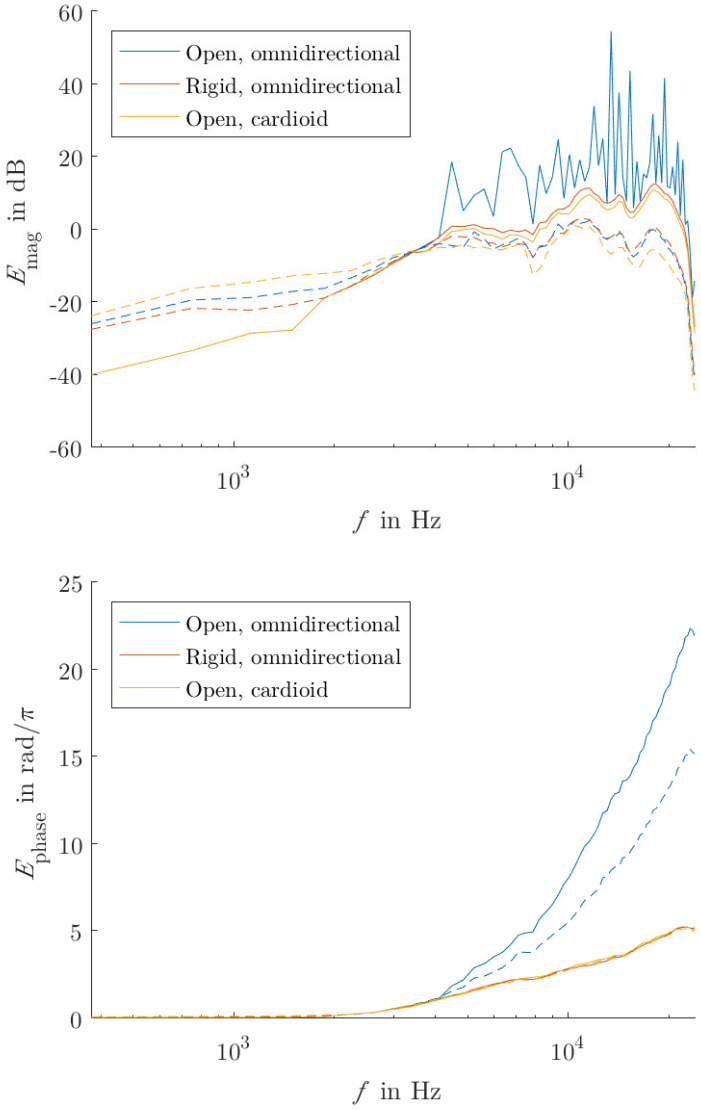


Figure 13: The RMS error in magnitude (top) and phase (bottom) after the spherical convolution of order $n = 5$ for different array configurations. The solid lines correspond to unlimited modal amplification while the dashed lines correspond to a maximum modal amplification of 0 dB. The rigid sphere array with cardioid microphones was indistinguishable from the rigid array with omni-directional microphones, and is therefore not included in the figure. The general shape is comparable to the error shapes in section 4.3, and is a consequence of error propagation.

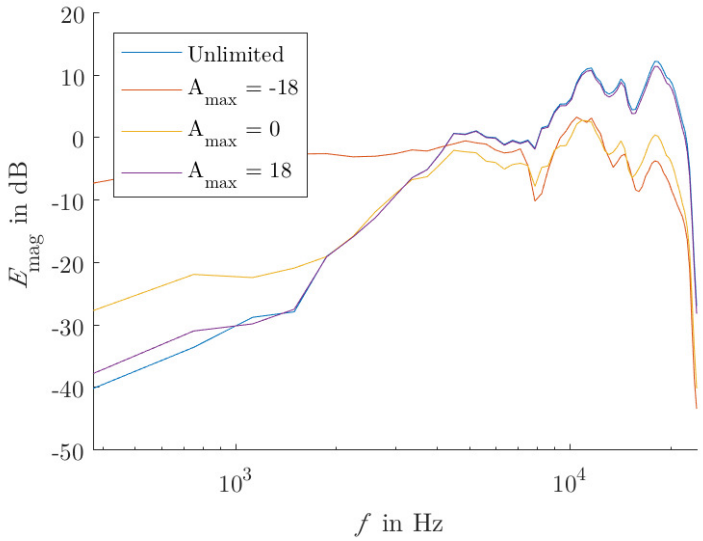


Figure 14: The RMS error in the magnitude for a convolved field for different maximum modal amplifications, at expansion order $n = 5$. The field response was simulated using a rigid array with omni-directional microphones.

ior. The original wavefront and the reflected waves are almost completely missing, and replaced by a diffuse response. This response starts at the very first measured time sample and lingers longer than the other responses.

5 User Study

5.1 Test Procedure

In order to test the subjective influence of spherical harmonics expansion of diffuse fields, a user study was conducted. The major focus in this user study was on the effects of the magnitude-phase separated expansion methods. For a more general study, primarily on the influence of the order of expansion, refer to [1]. The user study was a semantic differential listening test. The participants were presented with a pair of stimuli and asked to rate eight parameters on a scale between -100 and 100, see Table 1. The two stimuli were the same audio signal but processed using different BRIR filters, of which one was always a dummy head reference recording. The convolution of the audio signal with the filters was done using SoundScapeRenderer, with head tracking applied. Three different audio signals were used but in order to keep the tests reasonably short for the participants, each subject rated the BRIR filters using one audio signal, but with two hidden repetitions. The audio signals were one drum kit and percussion piece,

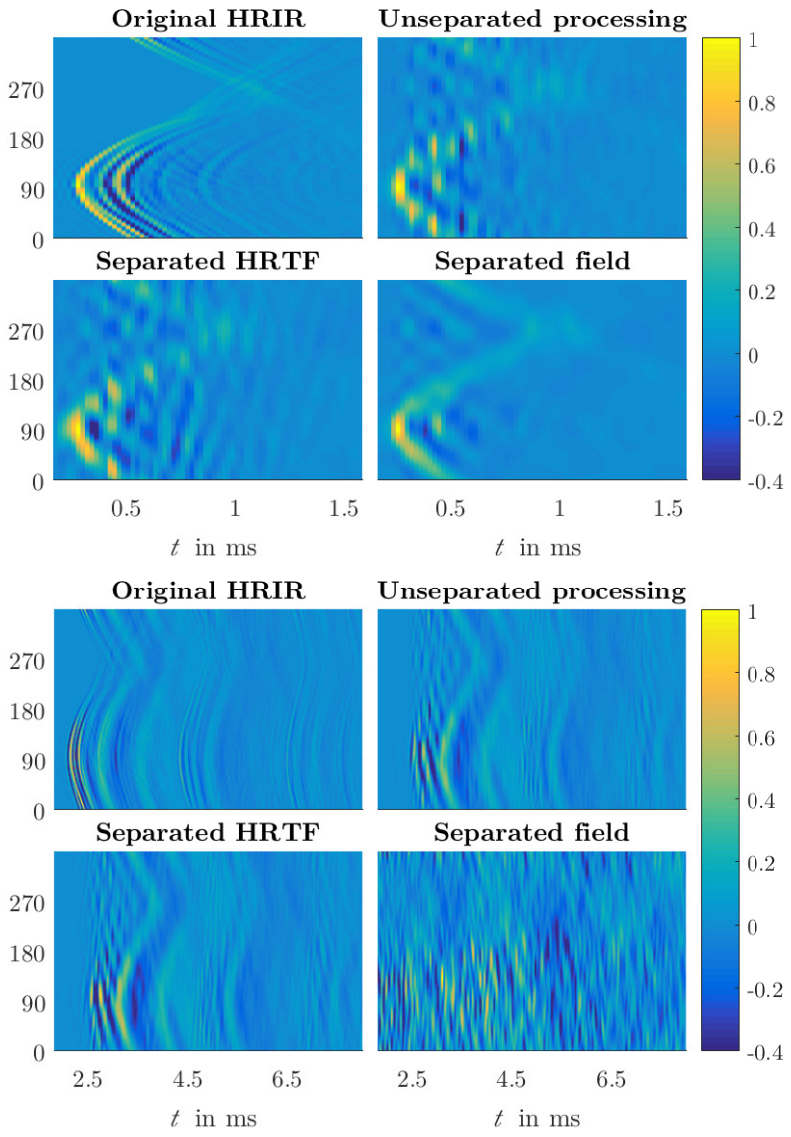


Figure 15: BRIR comparisons of different expansion methods in the complete signal processing chain. Shown are two acoustic spaces, an anechoic space (top) and a small control room (bottom).

one saxophone quartet piece, and one speech signal. All three audio signals were recorded in an anechoic environment. There were 15 participants in the test, so each audio signal was used for five participants. The participants were mostly MSc students at the division of applied acoustics at Chalmers University of Technology, along with some members of staff at the same division. The participants were between 23 and 50 years old.

Measurements of HRIR sets of a Neumann KU100 dummy head were used, both as the HRIR fields for the signal processing and the reference sets. The HRIR set used as the HRTF field in the signal processing was measured using a Lebedev grid of order 43 [7]. There were four different acoustic environments in the tests; an anechoic environment, a small control room, a medium performance room, and a large performance room. The RIR fields and dummy head BRIR references required for the latter three were measured at WDR broadcast studios [8]. These fields were expanded and used to generate BRIR filters usable with the SoundScapeRenderer software, as described in 3.3. For all spatial convolutions the maximum modal amplification was set to 0 dB, and all RIR fields were measured using a rigid array supporting expansion order $n = 5$. In order to test the effects of magnitude-phase separated expansions, a total of three BRIR filters were calculated for each acoustic environment. In addition to this, the anechoic space allows for use of analytic expansion coefficients of a plane wave, which removes the influence of the recording array of the RIR. This was included to have a baseline of what the recording array contributes with, i.e. the effects of the measured HRIR field. It was noted before the study that there were very small differences between the different expansion methods for the anechoic space. These were omitted from the test, in order to not have too many paired stimuli without audible differences. In total there were eleven processed BRIR filters, and four dummy head filters in the test.

5.2 Test Results

This section contains a summary of the results from the user study. The full set of measurements can be found in Appendix A. The figures below use box plot to show how the ratings was distributed. The box plots show seven values; the maximum and the minimum rating, the median rating, the first and the third quartiles, and the 10:th and 90:th percentiles. The minimum and maximum values are shown as the end points of the extruding whiskers. The median rating is shown as a thick line. The first and third quartiles are represented by the edges of the box, within which 50% of the rating lie. The 10:th and 90:th percentiles are shown as small circular marks on the whiskers, within which 80% of the ratings lie. It was noted that some participants have answered very differently in the two repetitions of the same comparison. The most plausible explanation to this is confusion between

Table 1: Subjective parameters in the user study. The first column correspond to the labels in figures that show the results from the user study. The second and third column correspond to the label for the -100 and +100 side of the scale that were presented to the participants.

Parameter	**Lower end**	**Upper end**
High frequencies	Weak high frequencies	Strong high frequencies
Low frequencies	Weak low frequencies	Strong low frequencies
Source size	Small source	Large source
Source distance	Close source	Far away source
Room size	Small room	Large room
Echoes	Fewer echoes	More echoes
Reverberation	Dry	Reverberant
Realistic	Synthetic	Realistic

which way the comparison is, so the participant rated A compared to B instead of B compared to A. Because of this the spread in the results is larger than the spread in subjective appreciation. Note that the maximum and the minimum values are very susceptible to this effect, since a single mis-rating might change the value by a lot. The labels below the axes correspond to the subjective parameters in Table 1, where the corresponding positive and negative scale can be found.

Figures 16 and 17 show the ratings of the parameters for the anechoic condition without respective with a sampling array for the sound field. This show a large maximum spread in the results, but the majority of the results are in a much smaller range. Both these conditions show a clear trend of decreased high frequencies. It is very likely that this comes from the maximum modal amplification, which is also indicated in [1]. Most participants considered the source and room size to be very similar to the reference, but with a tendency towards slightly smaller sources and rooms. The large spread in the source distance indicate difficulties in determining the distance to the source. A more detailed study of this shows no special tendencies in either audio signal. Interestingly there is a tendency towards less reverberation than the reference in the tests without the array. The largest difference between the test with and without an array for the sound field, is that in the tests without the array is rated less realistic than the tests with the array. This might be a coincidence, either due to the relatively small sample size, or that some effect from the array sampling cancels out some other effect in the signal processing. Otherwise it seems as if the spatial sampling had benefits in terms of producing a result that is believable.

Figure 18 shows the ratings for the three rooms with measured RIR, expanded using the normal SOFiA expansion method for both the RIR field and the HRTF

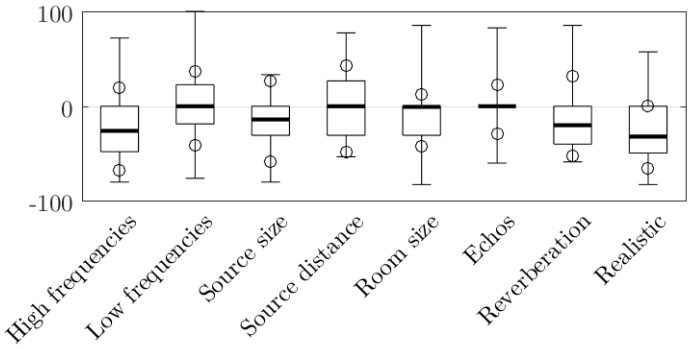


Figure 16: The results from the user study, anechoic conditions without array sampling for the field.

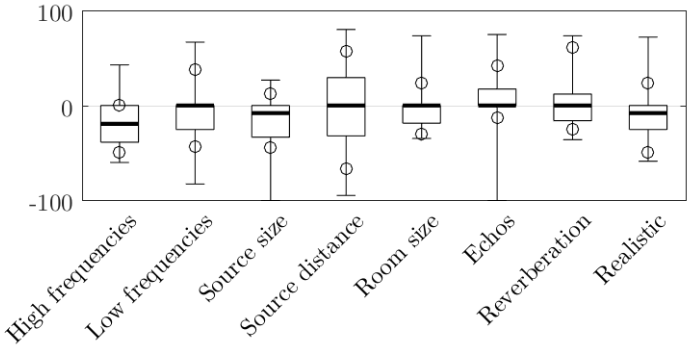


Figure 17: The results from the user study, anechoic conditions with array sampling for the field.

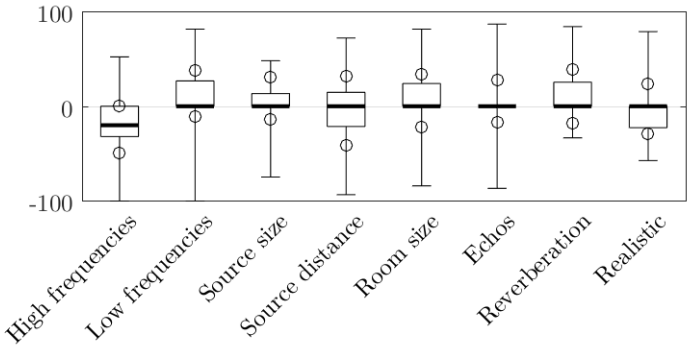


Figure 18: The results from the user study, using the normal expansion methods for both fields. All rooms and audio signals combined.

field. This also indicate some reduction in the higher frequencies and some increase in the lower frequencies. It is possible that this is due to the same effect, but the participants chose to rate using the high frequency sider or the low frequency slider depending on which stimulus was presented first. The results regarding the source and room size are however opposite to those of the anechoic space. In this case the results indicate that the room and source is perceived as larger, but for the anechoic space they were perceived as smaller. There still seems to be something that causes difficulties judging the distance to the source since that is spread over a large range, although a smaller range than for the anechoic space. There is a tendency towards higher perceived reverberation, which is likely to be correlated to the larger room size. The rating of the level of realism of the two stimuli shows a slight tendency towards synthetic, but this is very similar to the two anechoic cases.

Figure 19 shows the ratings for the three rooms using the magnitude-phase separated expansion step for the HRTF field. These results are very similar to the case using the standard SOFiA expansions. The only difference seems to be a slightly smaller rated difference in terms of high frequency content.

Figure 20 shows the results for the case when the RIR field was expanded using the magnitude-phase separated expansion. This is the case that differs the most from the other expansions. The timbre of the stimuli are rated quite similar to the other two expansion methods, with a slight shift towards an increase in the higher frequencies. For this expansion method there is an indication towards a larger source distance, a larger room size, and longer reverberation time. This is also the only method with a significant difference in the perceived echoes. This expansion method is also rated as less realistic than the others. All this could be due to the delayed wavefront seen in section 3.2.

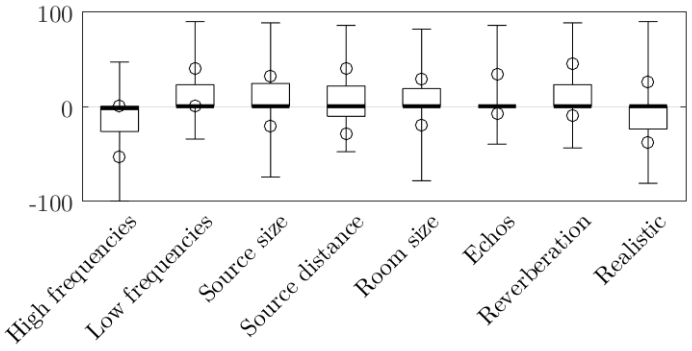


Figure 19: The results from the user study, using the separated expansion methods for the HRTF field. All rooms and audio signals combined.

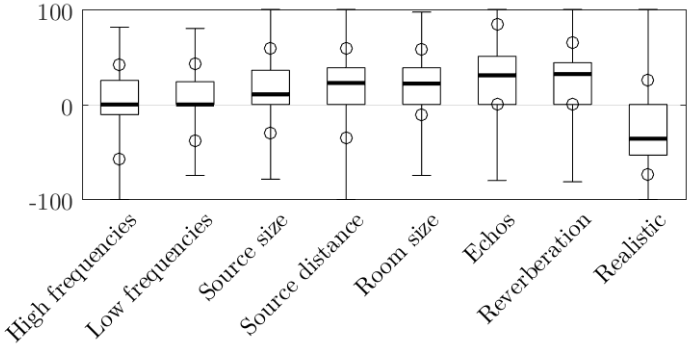


Figure 20: The results from the user study, using the separated expansion methods for the RIR fields. All rooms and audio signals combined.

6 Discussion

6.1 Expansions Methods

From the simulations it is clear that the standard SOFiA expansion method and the normal least-squares fit behave very similar. There are no significant differences in the measured errors in magnitude or phase for expansions of impulse fields or HRIR fields. All the grids used in this work have been structured grids that cover the entire sphere. It is possible that these two expansion methods differ when applied on irregular or incomplete grids. The regularization is shown to have negative effects, mainly on the errors in magnitude. In should be noted again that the grids used in this work cover the entire sphere. One of the reasons for using regularization is to decrease the errors in unsampled areas, at the cost of increased errors in the sampled areas [6].

From the simulations of diffuse fields and the user study it is clear that the magnitude-phase separated expansion method does not work very well for sound fields with diffuse reverberation. The reconstructed impulse responses in the simulations show a delayed and more diffuse version of the direct sound as well as the original direct sound. The user study indicates that this effect could be heard as unnatural echoes. One of the participants described this sound as "A strange surging echo from behind". When the magnitude-phase separated expansion is used for the expansion of the HRTF set the results are more similar to the normal expansion, and no large differences could be measured in the user study. For the HRTF set used in this work, the magnitude-phase separated expansion does not seem to offer any benefits that motivate the additional complexity. Additional testing using other measurements is required to determine if the magnitude-phase separated expansions can be used to increase the robustness of the system.

6.2 User Study

The wide range of ratings in the user study suggests that there are parts of the signal processing that changes the spatial impression of the sound field, particularly the large spread in perceived source distance is an indicator for this. At this stage the cause for this is not known, and has a number of possible explanations. It might be an unfortunate result due to the relatively small sample size in the user study. This does not seem as the most plausible explanation considering the consistent ratings in some of the parameters in the study, e.g. room size or echoes. It is possible that some artifact from the processing is heard by the participants, but interpreted differently. Several of the participants commented that it was difficult to give ratings that described what they heard.

When the data from single participants was analyzed it was clear that the

ratings are not consistent. In many cases the participant gave a high positive score in the first test and a high negative score in the second test of the same stimuli. This could happen if the participant heard a difference but could not connect the subjective difference to the correct parameter in the test. A more likely scenario is that the participant rated the stimuli backwards, as in A compared to B instead of B compared to A. Some of the participants said afterwards that they had realized during the test that they were rating the pair backwards. From their accounts it was more likely to rate A instead of B when A was the "bad sound", i.e. the participants had a tendency to rate the processed stimulus. This could be improved in two ways. Firstly, the dummy head reference could be kept at A all the time, and the processed signal at B. This might reduce the number of A/B confusions since the participant is always rating the processed signal. Secondly, the rating sliders could be disabled when the A stimulus is selected. That would mean that the participant only can rate when listening to the sound that he/she should rate. These two changes should at least reduce the number of A/B confusions which would give a more precise result. Removing a few of the more similar test cases in favor of more repetitions or shorter test length is also possible. The advantage of more repetitions of fewer cases is that the tendencies of each participant should be more clear. The advantage of an overall shorter test is that the participant might focus more on each individual test. These changes in combination with a larger sample would potentially increase the precision of the complete study.

6.3 Spherical Convolution

Previous works have noted a large difference in subjective overall rating depending on the quadrature grid used to perform the spherical convolution between sound field and HRTF set, described by (8) [1]. The convolution method used in this work does not use any quadrature grid at all in this calculation, but is based purely on the spherical harmonics coefficients. Since no direct comparison between the two methods has been made it is not feasible to give any strong recommendations on the matter. However, it should be noted that if the quadrature grid method does not show clear advantages in form of better subjective appreciation or more robust results, the method proposed in this work has two benefits. The first benefit is that there is no dependence on any quadrature grid, which simplifies implementations since no suitable quadrature grid has to be chosen. The interactions between the measurement grids of the field and the HRIR set and the convolution quadrature grid are not yet fully understood, so simplifying the method by removing one grid entirely would enable more focused research on the other parts of the processing. The other benefit is computational. The quadrature grid method requires four multiplications per grid point, and in most cases more grid points than spherical harmonics coefficients. If the HRTF set has been measured at the exact grid

used as the quadrature grid these measurements can be used as is, otherwise the HRTF set must be expanded to the spherical harmonics domain and resampled at the correct grid points. In contrast, the coefficient multiplication method used in this work requires two multiplications per spherical harmonics coefficient, but the HRTF set must be expanded to the spherical harmonics domain regardless of the sampling points. Assuming identical grids for both methods and that the quadrature grid method is used without resampling the HRTF set, both methods have the same order for the number of calculations ($\mathcal{O}(N^6)$), but the quadrature grid method requires twice as many multiplications. If the quadrature grid method needs HRTF resampling it would require many more calculations ($\mathcal{O}(N^{10})$).

6.4 Further Research

As described in section 3.2 it would be advantageous to expand the field in terms of the phase response and the magnitude response, instead of the real and imaginary parts of the complex response. Viewed in terms of the least-squares expansion method, it is a matrix equation system that is solved in the complex domain. The matrix analogue to the polar form of a complex number ($z = re^{j\varphi}$) is the polar decomposition $A = UP$ where U is a orthogonal matrix and P is positive semi-definite [9]. This could potentially be used as a way to separate the magnitude and the phase and find the spherical harmonics expansion coefficients separately for the two.

It is also possible to unwrap the phase on the spherical surface instead of unwrapping in the frequency domain. This could potentially give a smoother phase response over the sphere for each frequency bin, but the effect might be that the phase response for each receiver is not a smooth function of frequency. It is shown by Zaar [10] that phase unwrapping over the sphere give a better time localization for a measure HRIR. The effects on a set of RIR measured in an reverberant acoustic space is not discussed. This is a possible opportunity to improve the time precision of the processing chain.

7 Conclusion

Using spherical microphone arrays to record sound fields and creating binaural auralizations in headphones is a promising approach. For the method to reach its full potential a better understanding of how the two fields, the sound field and the HRIR field, interacts to form the final result. Different types of processing is important and can be improved, but the actual recording methods and the design of the recording array is essential for the performance of the end result.

The user study performed was far from perfect, with a large spread in the

subjective ratings of processed signals compared to dummy head references. This is believed to be a combination of a too open test design which confused the participants, and that the stimuli are so similar that many did not hear a clearly definable difference. If the test design is improved and a larger sample size is tested, the results are likely more precise.

The magnitude-phase separated expansion method proposed show unexpected and unexplained behavior for sound fields with a reverberant component. A second diffuse wave front appears in the binaural signals, which could be heard by the participants in the listening test. The proposed expansion method is therefore not recommended for use on actual sound fields as it is implemented for this work.

References

[1] Benjamin Bernschütz. "Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording". Doctoral Dissertation. Technical University of Berlin, 2016.

[2] Benjamin Bernschütz et al. "SOFiA Sound Field Analysis Toolbox". In: *International Conference on Spatial Audio, Detmold, Germany*. 2011.

[3] Matthias Geier and Sascha Spors. "Spatial Audio Reproduction with the SoundScape Renderer". In: *27th Tonmeistertagung – VDT International Convention*. 2012.

[4] Andrew Wabnitz et al. "Room acoustics simulation for multichannel microphone arrays". In: *Proceedings of the International Symposium on Room Acoustics, ISRA*. 2010.

[5] George B. Arfken, Hans J. Weber, and Frank E. Harris. *Mathematical Methods for Physicists*. Academic Press, 2013.

[6] Jens Ahrens, Mark R. P. Thomas, and Ivan Tashev. *HRTF Magnitude Modeling Using A Non-regularized Least-squares Fit Of Spherical Harmonics Coefficients On Incomplete Data*. Technical report. Microsoft Research.

[7] Benjamin Bernschütz. "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100". In: *AIA-DAGA 2013 Conference on Acoustics*. 2013.

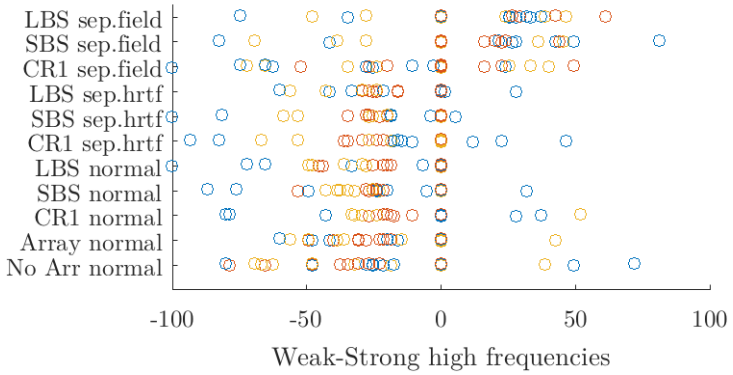
[8] Philipp Stade, Benjamin Bernschütz, and Maximilian Rühl. "A Spatial Audio Impulse Response Compilation Captured at the WDR Broadcast Studios". In: *27th TONMEISTERTAGUNG – VDT INTERNATIONAL CONVENTION*. Nov. 2012.

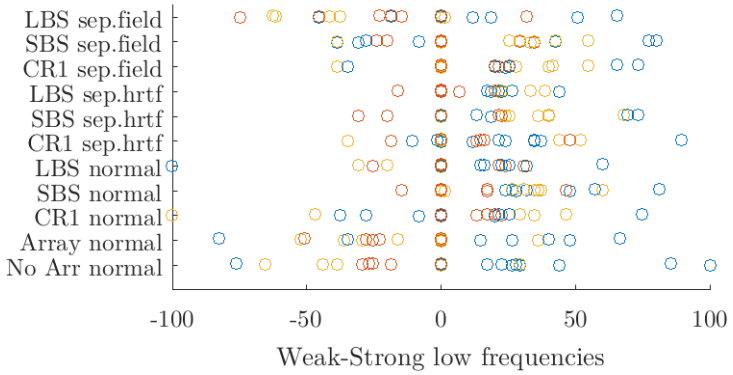
[9] Mohammad Sal Moslehian. *Polar Decomposition*. MathWorld–A Wolfram Web Resource. URL: `http://mathworld.wolfram.com/PolarDecomposition.html` (visited on 2017).

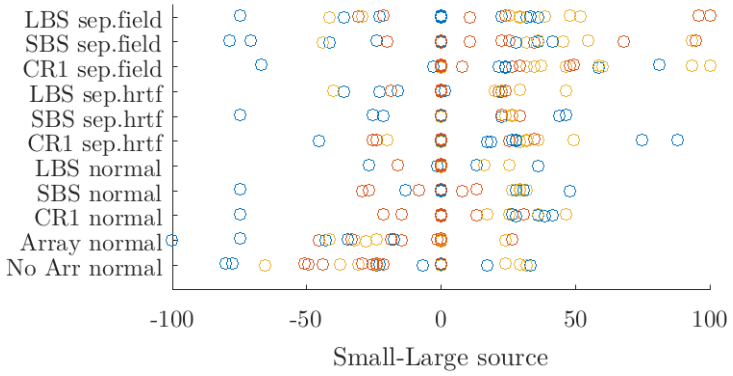
[10] Johannes Zaar. “Phase Unwrapping for Spherical Interpolation of Head-Related Transfer Functions”. Diploma Thesis. Institute of Electronic Music and Acoustics, University of Music and Performing Arts, Graz, 2011.

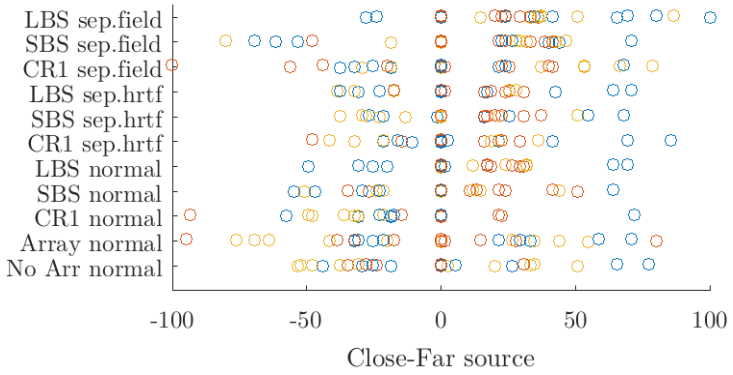
A Results from User Study

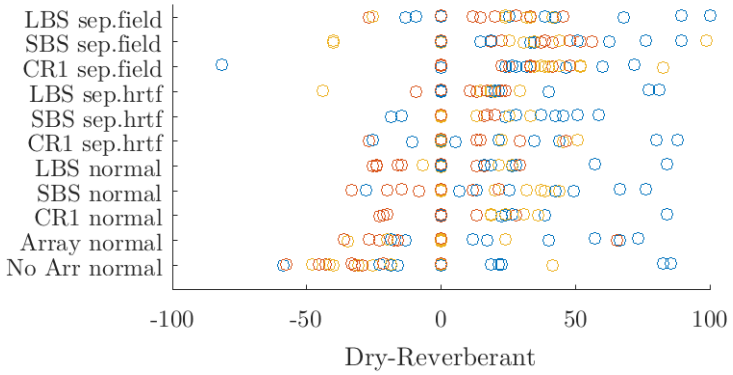
This section contains all the responses from the user study. In these figures, the blue marks correspond to tests using the saxophone audio signal, the red marks correspond to the drums audio signal, and the yellow marks correspond to the speech audio signal. The labels on the left hand side show which room and method that correspond to that line. The labels have been abbreviated as follows: The two anechoic spaces are labeled ”Array” and ”No Arr” respectively, the small control room is labeled ”CR1”, the medium performance room is labeled ”SBS” (Small Broadcasting Studio), and the large performance room is labeled ”LBS” (Large Broadcasting Studio). The expansion methods have been abbreviated as; ”normal” for the Standard SOFiA expansion applied on both fields, ”sep.hrtf” where the separation step have been used on the HRTF field, and ”sep.field” where the separation have been used on the RIR field. The labels below each axis show which parameter the data corresponds to.

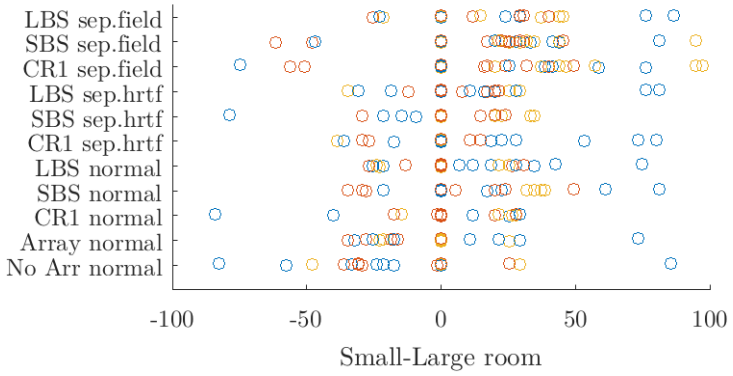


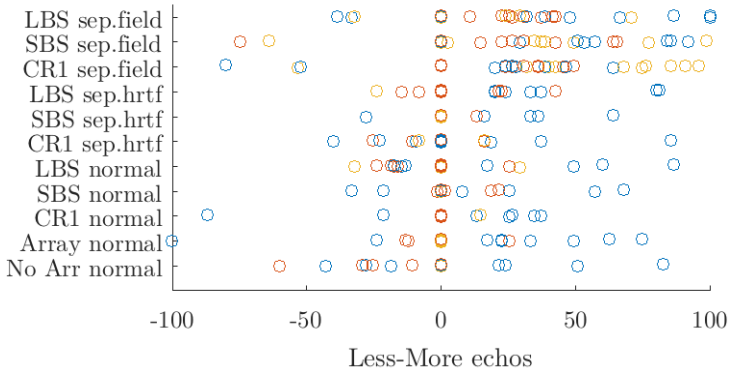


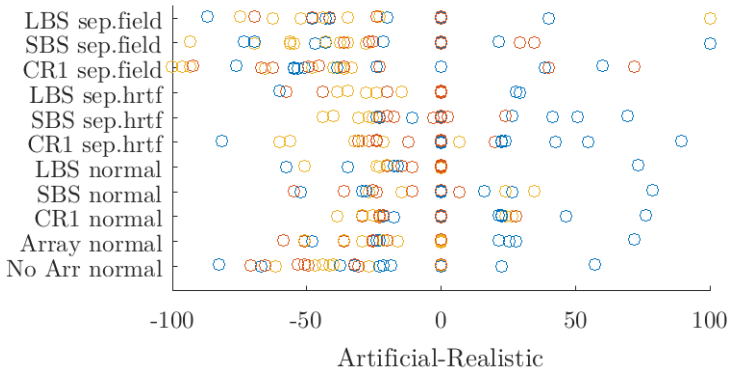

LBS sep.field
SBS sep.field
CR1 sep.field
LBS sep.hrtf
SBS sep.hrtf
CR1 sep.hrtf
LBS normal
SBS normal
CR1 normal
Array normal
No Arr normal
-100
-50
0
50
100
Small-Large source


LBS sep.field
SBS sep.field
CR1 sep.field
LBS sep.hrtf
SBS sep.hrtf
CR1 sep.hrtf
LBS normal
SBS normal
CR1 normal
Array normal
No Arr normal
-100
-50
0
50
100
Close-Far source


LBS sep.field
SBS sep.field
CR1 sep.field
LBS sep.hrtf
SBS sep.hrtf
CR1 sep.hrtf
LBS normal
SBS normal
CR1 normal
Array normal
No Arr normal
-100
-50
0
50
100
Dry-Reverberant


LBS sep.field
SBS sep.field
CR1 sep.field
LBS sep.hrtf
SBS sep.hrtf
CR1 sep.hrtf
LBS normal
SBS normal
CR1 normal
Array normal
No Arr normal
-100
-50
0
50
100
Small-Large room


LBS sep.field
SBS sep.field
CR1 sep.field
LBS sep.hrtf
SBS sep.hrtf
CR1 sep.hrtf
LBS normal
SBS normal
CR1 normal
Array normal
No Arr normal
-100
-50
0
50
100
Less-More echos


LBS sep.field
SBS sep.field
CR1 sep.field
LBS sep.hrtf
SBS sep.hrtf
CR1 sep.hrtf
LBS normal
SBS normal
CR1 normal
Array normal
No Arr normal
-100
-50
0
50
100
Artificial-Realistic