Audio Engineering Society

Convention Paper

Presented at the 132nd Convention
2012 April 26–29 Budapest, Hungary

Comparison of modal versus delay-and-sum beamforming in the context of data-based binaural synthesis

Sascha Spors, Hagen Wierstorf and Matthias Geier

Deutsche Telekom Laboratories, Technische Universität Berlin, Ernst-Reuter-Platz 7, 10587 Berlin, Germany

Correspondence should be addressed to Sascha Spors (Sascha.Spors@tu-berlin.de)

ABSTRACT

Several approaches to data-based binaural synthesis have been published that capture a sound field by means of a spherical microphone array. The captured sound field is typically decomposed into plane waves which are then auralized using head-related transfer functions (HRTFs). The decomposition into plane waves is often based on modal beamforming techniques which represent the captured sound field with respect to surface spherical harmonics. An efficient and numerically stable approximation to modal beamforming is the delay-and-sum technique. This paper compares these two beamforming techniques in the context of data-based binaural synthesis. Their frequency- and time-domain properties are investigated, as well as the perceptual properties of the resulting binaural synthesis according to a binaural model.

1. INTRODUCTION

Head-related transfer functions (HRTFs) and their time-domain equivalent head-related impulse responses (HRIRs) capture the acoustic transmission path from an acoustic source to the outer ear. Typically the case of free-field propagation is referred to as HRTFs, while similar transfer functions captured in a room are referred to as binaural room transfer functions (BRTFs) or binaural room impulse responses (BRIRs). Both HRTFs and BRTFs are used for the synthesis of virtual sources in virtual auditory environments, by filtering the (dry) signal of a virtual source with the respective left and right HRTFs/BRTFs. This approach to sound reproduction is called binaural reproduction or synthesis [1]. Typically headphones are used to reproduce the left and right ear signals; however reproduction can also take place over loudspeakers when appropriate transaural techniques are used to cope for the crosstalk between the loudspeaker channels and the ears.

Binaural synthesis using HRTFs/BRTFs is limited to the auralization of a finite number of individual virtual sources. For good results, head-tracking and dynamic exchange of the HRTFs is mandatory. Diffuse sound fields, for instance cafeteria noise, cannot be represented by transfer functions. Hence, a head-tracked auralization of diffuse sound fields cannot be achieved straightforwardly by binaural synthesis. HRTFs vary to some degree among individuals. Hence, the perceived quality can be increased by using individualized HRTF/BRTFs [2]. While the measurement effort for individualized HRTFs is already quite high, similar measurements for BRTFs have to be repeated for all acoustic environments to be auralized. It would be desirable to extract the room effect from the BRTFs, so that it can be added to individualized HRTFs.

These limitations of BRTF-based binaural synthesis can be overcome by combining techniques from sound field analysis with HRTF-based binaural synthesis. The underlying concept is to decompose the sound field into its contributions impinging from different directions, which are auralized by the respective HRTFs. Several approaches have been published that decompose the captured sound field into plane waves [3, 4, 5, 6]. The sound pressure at the left/right ear is given by superposition of the respective (far-field) HRTF filtered by the plane wave expansion coefficients of the captured sound field.

Due to their independence from the incidence direction of sound, spherical microphone arrays are preferred for the spatial analysis of sound fields. It is natural to represent the sound field captured on the surface of a sphere with respect to surface spherical harmonics. The plane wave decomposition can be computed conveniently from the spherical harmonics expansion coefficients of the captured sound field. An efficient alternative to the concept of modal beamforming (MB) is delay-and-sum beamforming (DSB).

Practical implementations of spherical microphone arrays are not capable of capturing the sound field with high spatial accuracy over the full audio frequency range. This holds especially for the lower and higher frequencies. Since we are aiming at binaural synthesis for human listeners the question arises which spatio-temporal accuracy is required to cope for the capabilities of the human ear. First studies have been published with respect to the perceptual impact of spatial bandlimitation in data-based binaural synthesis using modal beamforming [5, 4, 6]. So far these studies focused mainly on localization accuracy for the direction-of-arrival while varying the spatial bandwidth of the plane wave decomposition. In a recent publication we investigated on the perceptual effect of limited spatial bandwidth without considering spatial sampling [7]. This paper extends the published results in various ways.

In this paper a detailed look will be taken on the effects introduced by spatial sampling and the properties of delay-and-sum beamforming. For this we investigate the perceptual properties of data-based binaural synthesis using simulated sound fields and a binaural auditory model. Besides perceptual features linked to the direction-of-arrival we also take spectral features into account.

2. DATA-BASED BINAURAL SYNTHESIS

Within a spherical volume of radius R any propagating sound field $P(\mathbf{x},\omega)$ produced by sources located outside of that volume can be represented in terms of a superposition of plane waves [8, 9]. Denoting the spectrum of a plane wave traveling in direction ϕ, θ by $\bar{P}(\phi,\theta,\omega)$, this superposition reads

$$P(\mathbf{x},\omega) = \frac{1}{4\pi} \int_0^{2\pi}\int_0^{\pi} \bar{P}(\phi,\theta,\omega) e^{-i\mathbf{k}^T\mathbf{x}} \sin\theta \; d\theta d\phi \; , \quad (1)$$

where i denotes the imaginary unit, $\omega = 2\pi f$ the angular frequency, $\mathbf{x} = (x,y,z)^{\mathrm{T}}$ a position in space, $\mathbf{k}$ the wave vector of a particular plane wave with $\mathbf{k} = \frac{\omega}{c}(\cos\phi\sin\theta, \sin\phi\sin\theta, \cos\theta)^{\mathrm{T}}$, ϕ its azimuth and θ its colatitude. The horizontal plane is defined by $\theta = \pi/2$. The wave vector $\mathbf{k}$ points in the traveling direction of the plane wave.

Far-field HRTFs $\bar{H}_{\mathrm{L,R}}(\phi,\theta,\gamma,\delta,\omega)$ represent the acoustic transmission path from a plane wave with incidence angle ϕ, θ to the left and right ear under the head orientation γ, δ. Far-field HRTFs can be extrapolated for instance from HRTFs measured at finite distance [10]. The basic concept of the presented approach to data-based binaural synthesis is to replace the exponential term in (1), representing

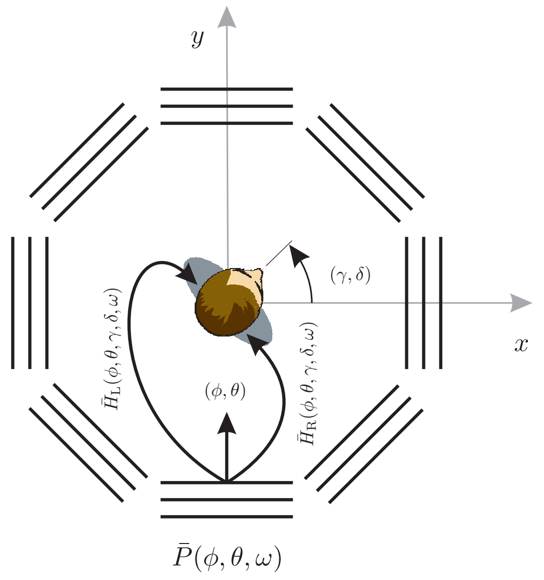

y
x
(γ,δ)
(ϕ,θ)
$\bar{H}_{\mathrm{L}}(\phi,\theta,\gamma,\delta,\omega)$
$\bar{H}_{\mathrm{R}}(\phi,\theta,\gamma,\delta,\omega)$
$\bar{P}(\phi,\theta,\omega)$

Fig. 1: Data-based binaural synthesis using a plane wave expansion of the virtual sound field. For illustration, the filtering of the left/right HRTFs by the plane wave expansion coefficients $\bar{P}(\phi,\theta,\omega)$ is shown only for one particular direction. The z-axis points upwards.

the sound field of a plane wave, by its respective far-field HRTF.

The sound pressure at the left/right ear $P_{\mathrm{L,R}}(\gamma,\delta,\omega)$ for a certain head orientation γ,δ is given by superposition of the respective far-field HRTFs $\bar{H}_{\mathrm{L,R}}(\phi,\theta,\gamma,\delta,\omega)$ filtered by the plane wave expansion coefficients $\bar{P}(\phi,\theta,\omega)$

$$P_{\mathrm{L,R}}(\gamma,\delta,\omega) = \frac{1}{4\pi} \times \int_0^{2\pi}\int_0^{\pi} \bar{P}(\phi,\theta,\omega)\bar{H}_{\mathrm{L,R}}(\phi,\theta,\gamma,\delta,\omega)\sin\theta \, d\theta d\phi \ , \quad (2)$$

where γ denotes the azimuth and δ the colatitude. Figure 1 illustrates the superposition of filtered far-field HRTFs.

Using an expansion of the captured sound field with respect to plane waves has a number of benefits in the context of data-based binaural synthesis. Head-tracked binaural synthesis is straightforward to achieve by choosing the appropriate HRTFs in (2) for a given head orientation γ,δ. In the first approximation and under free-field conditions $\bar{H}_{\mathrm{L,R}}$ depends only on the difference between the two angle pairs ϕ,θ and γ,δ. Another benefit is that numerous techniques are known to perform a plane wave decomposition for microphone arrays of different geometries. In this paper two different techniques will be compared to derive $\bar{P}(\phi,\theta,\omega)$ from the sound field captured on a spherical surface. A hemispherical array could also be used [3]. A further benefit of plane waves is that they allow a variable spatial resolution and that horizontal-only synthesis can be realized in a straightforward way.

In practice, only a limited number of plane wave expansion coefficients $\bar{P}(\phi,\theta,\omega)$ can be extracted from the captured sound field. Hence a discretized version of (2) is used in practice. Spatial sampling may lead to sampling artifacts in the synthesized ear signals $P_{\mathrm{L,R}}(\gamma,\delta,\omega)$. The impact of spacial sampling is discussed in more detail in Section 4.

3. PRINCIPLES OF SPHERICAL MICROPHONE ARRAY PROCESSING

The properties of spherical microphone arrays are independent from the incidence direction of sound which is preferable for the analysis of sound fields. In the context of this paper, two different methods to derive the plane wave expansion are briefly reviewed.

3.1. Modal Beamforming

Due to the underlying geometry it is natural to represent the sound field captured on the surface of a sphere with respect to surface spherical harmonics. In this context, spherical harmonics are also known as acoustic modes and the appendant techniques as modal processing or modal beamforming. Various techniques have been published in the past decade for open or rigid spheres equipped with pressure microphones or cardioid microphones [3, 11, 12, 13]. For the analysis of room acoustics, open spheres show the benefit of being acoustically transparent and practically realizable for large apertures. We therefore focus on open sphere designs in the remainder.

The spherical harmonics expansion coefficients $\mathring{S}_n^m(\omega)$ can be derived from the sound field $S(\mathbf{x},\omega)$ captured on an acoustically transparent sphere with radius R as [12]

$$\mathring{S}_n^m(\omega) = \int_0^{2\pi}\int_0^{\pi} S(\mathbf{x},\omega) Y_n^{-m}(\beta,\alpha)\sin\beta \; d\beta d\alpha \; , \quad (3)$$

where $\mathbf{x} = R\,(\cos\alpha\sin\beta, \sin\alpha\sin\beta, \cos\beta)^{\mathrm{T}}$ denotes a position on the sphere with α its azimuth and β its colatitude, $Y_n^m(\cdot)$ the n-th order surface spherical harmonic of m-th degree and $j_n(\cdot)$ the n-th order spherical Bessel function [14]. Note that $S(\mathbf{x},\omega)$ depends on the used microphone type. The expansion of a sound field in terms of plane waves (1) can be linked to the spherical harmonics expansion coefficients as

$$\bar{P}_{\mathrm{MB}}(\phi,\theta,\omega) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} \frac{1}{B_n(\omega)} \mathring{S}_n^m(\omega) Y_n^m(\theta,\phi) \; , \quad (4)$$

where

$$B_n(\omega) = \begin{cases} 4\pi i^n j_n(\frac{\omega}{c}R) & \text{, for pressure} \\ 4\pi i^n (j_n(\frac{\omega}{c}R) - i j_n'(\frac{\omega}{c}R)) & \text{, for cardioid} \end{cases} \quad (5)$$

microphones with R denoting the radius of the sphere. It is evident from (4) and (5) that the use of pressure microphones causes numerical problems due to the zeros of the spherical Bessel function, while this is not the case for cardioid microphones.

Equation (3) together with (4) forms the basis to calculate the plane wave expansion $\bar{P}_{\mathrm{MB}}(\phi,\theta,\omega)$ of a sound field captured by a spherical microphone array.

3.2. Delay-and-Sum Beamforming

The basic concept of delay-and-sum beamforming is – as the name implies – to sum up the signals of the microphones after applying individual delays to them. For a plane wave decomposition of the captured sound field, the delays compensate for the propagation delays between the microphones for a given look direction ϕ, θ. This way the output signal is maximized by constructive interference for a plane wave whose incidence angle coincides with the look direction. When capturing the pressure $P(\mathbf{x},\omega)$ on a open sphere, the plane wave decomposition can be derived by simple geometric considerations as [11]

$$\bar{P}_{\mathrm{DSB}}(\phi,\theta,\omega) = \int_0^{2\pi}\int_0^{\pi} P(\mathbf{x},\omega) e^{-i\mathbf{k}^T\mathbf{x}} \sin\beta \; d\beta d\alpha \; , \quad (6)$$

where $\mathbf{k} = \mathbf{k}(\phi,\theta)$ as defined for (1) and $\mathbf{x} = \mathbf{x}(\alpha,\beta)$ as defined for (3).

4. PRACTICAL ASPECTS OF SPHERICAL MICROPHONE ARRAYS

Practical realizations of spherical microphone arrays are subject to limitations. These are discussed in the following subsections.

4.1. Spatial Dimensionality of Sound Fields

As prerequisite for the following considerations, the concept of spatial dimensionality of multipath sound fields as presented in [15] is reviewed. The spatial dimensionality of a sound field characterizes the finite number of orthogonal components that are required to represent a sound field within a source-free bounded volume with bounded error. Quantitative results have been derived using a representation of the sound field within a spherical volume of radius R in terms of spherical harmonics. The normalized absolute field truncation error ϵ_N for three-dimensional sound fields is bounded as

$$\epsilon_N \leq 0.67848\, e^{-\Delta} \quad \text{for } N > \lceil \frac{e\pi}{c} Rf \rceil + \Delta \; , \quad (7)$$

where $\lceil\cdot\rceil$ denotes the ceiling operation, Δ a natural positive number and N the number of elements the outer sum over n in (4) is truncated to. It can be concluded from (7) that the error decays exponentially with increasing N. Furthermore for fixed error bound, N increases linearly with frequency f and radius R of the considered volume. The dimensionality quantifies the minimum number of sensors/actuators required to capture/synthesize a sound field within the volume $r < R$ under the error bound ϵ_N. For a three-dimensional sound field it is given as $(\lceil \frac{e\pi}{c} Rf \rceil + 1)^2$. For the synthesis of sound fields either point sources or plane waves can be used.

Neglecting the scattering by the human head and considering the typical size of a human head $R = 0.09\,\mathrm{m}$ and full audio bandwidth $f = 20\,\mathrm{kHz}$ it follows from above considerations that $N > 45$. Hence, at least $M = 2116$ sensors would be required to capture the sound field with an absolute field truncation error of $\epsilon_{N>45} \leq 0.67848$. Despite the high technical effort, the resulting error is still quite high. Furthermore, the perceptual impact of the remaining truncation error is not fully investigated at the current state of research.

4.2. Spatial Sampling in Modal Processing

In practice it is only possible to measure the pressure on a limited number of positions on the sphere. Therefore, a spatially discretized version of (3) has to be used. Strictly regular sampling of the sphere is provided only by the five platonic solids. Besides these a number of quasi-uniform and non-uniform sampling schemes for the sphere have been proposed [16].

The previous subsection outlined the concept of spatial dimensionality of sound fields and gave the minimum number of sensors required to capture a sound field up to an order N for a given radius R and error bound (7). Depending on the actual spatial sampling scheme, more spatial samples may be required [17]. A Lebedev grid, for instance, requires $M = 1.3\,(N+1)^2$ spatial samples, other grids typically more. Spatial aliasing above the frequency predicted by (7) will occur if less spatial samples are used. The impact of spatial aliasing has been investigated amongst others in [18]. It can be shown that for instance a Gaussian sampling scheme results in spectral repetitions with respect to the order n and degree m of $\mathring{S}_n^m(\omega)$ [19]. Hence, spatial aliasing could only be avoided by a spatial anti-aliasing filter applied before the actual spatial sampling process.

It is common practice in modal array processing to compute the spherical harmonics coefficients $\mathring{S}_n^m(\omega)$ only up to order N for a given number of spatial samples M. As a consequence, the sum in (4) over n will be truncated at N elements. This can be interpreted as a spatial bandlimitation in the spherical harmonics domain. The plane wave decomposition of an incident plane wave with unit amplitude and incidence angle $\phi_{\mathrm{pw}}, \theta_{\mathrm{pw}}$ impinging on an spatially continuous microphone array without sensor noise reads [11]

$$\bar{P}_{\mathrm{MB,cont}}(\phi, \theta, \omega) = \sum_{n=0}^{N} \frac{2n+1}{4\pi} P_n(\cos\Theta) \; , \qquad (8)$$

where $P_n(\cdot)$ denotes the n-th Legendre polynomial and Θ the angle between the incidence angle of the plane wave $\phi_{\mathrm{pw}}, \theta_{\mathrm{pw}}$ and the look direction of the plane wave decomposition ϕ, θ. It is evident from (8) that the continuous array response is frequency-independent. It can furthermore be concluded from [11] that the spatial selectivity of the plane wave decomposition decreases with decreasing order N. Equation (8) allows to investigate the effect of truncation without considering spatial sampling and aliasing.

4.3. Spatial Sampling in Delay-and-Sum Processing

The considerations given above for modal beamforming hold in principle also for delay-and-sum beamforming. The link between both techniques has been established in [11] by formulating both in terms of modal beamforming. In contrast to modal beamforming, spatial bandlimitation (truncation to order N) is not applied in delay-and-sum beamforming when computing the plane wave decomposition. As in (8) the response of a spatially continuous array to an incident plane wave is considered. It is given as [11]

$$\bar{P}_{\mathrm{DSB,cont}}(\phi, \theta, \omega) = \sum_{n=0}^{\infty} \left| 4\pi i^n j_n(\frac{\omega}{c} R) \right|^2 \frac{2n+1}{4\pi} P_n(\cos\Theta) \; . \qquad (9)$$

From the characteristics of the spherical Bessel function can be concluded that for low frequencies the higher orders n are attenuated in comparison to the lower ones. As a consequence, the array response is frequency-dependent, with lower spatial selectivity for low frequencies.

Equations (8) and (9) allow to compare modal beamforming to delay-and-sum beamforming without introducing the additional effects of spatial sampling.

4.4. Equipment Noise

The acoustic quantities captured by microphones are subject to equipment noise. The white noise gain (WNG) characterizes the improvement in signal-to-noise ratio at the beamformer output in comparison to the equipment noise. It has been shown [11] that for the delay-and-sum beamformer the WNG is frequency-independent and is given by $\mathrm{WNG}_{\mathrm{DSB}} =$

M. Hence equipment noise is attenuated by delay-and-sum beamforming.

The situation is different for the modal beamformer. Here the WNG depends on the frequency [11]. For low frequencies equipment noise is amplified while for higher frequencies it is attenuated. For high frequencies the WNG of the modal beamformer approaches the value of the delay-and-sum beamformer. In order to maintain a reasonable white-noise gain (WNG) the order is typically limited for low frequencies in practical modal beamformers.

4.5. Other Aspects

Besides sampling and equipment noise, also sensor mismatch and misplacement play a role in practical applications. For low frequencies modal beamformers are quite sensitive to both due to the underlying differential mechanism [17]. Sensor calibration and precise positioning is therefore mandatory here. Another countermeasure is to limit the order for low frequencies. Delay-and-sum beamforming is in general more robust against sensor mismatch and misplacement than modal beamforming.

5. EVALUATION OF DATA-BASED BINAURAL SYNTHESIS

5.1. Concept

A consequence of the considerations given in Section 4 is that practical implementations of data-based binaural synthesis using modal processing will not be capable to capture the required minimum order of $N = 45$ over the full audio frequency range. This holds especially for the lower frequencies. Delay-and-sum beamforming is more robust but on the other hand has limitations with respect to the achievable directivity. Both techniques differ also in the influence of spatial aliasing due to their different spatial bandwidth.

Since we are aiming at binaural synthesis for human listeners the question arises which technique is better suited and what accuracy is required to cope for the capabilities of the human ear. In order to investigate on this, we first take a look at the properties of modal and delay-and-sum beamforming. This is followed by an analysis of the technical properties of binaural synthesis. Finally, the perceptual properties are investigated by means of a model of human perception. Four scenarios are considered:

1. without spatial sampling
   (a) modal beamforming with different orders
   (b) delay-and-sum beamforming
2. with spatial sampling
   (a) modal beamforming with maximum order
   (b) delay-and-sum beamforming

5.2. Experimental Setup

The evaluation is based on (i) the plane wave expansion coefficients $\bar{P}(\phi,\theta,\omega)$ and (ii) the ear signals $P_{\mathrm{L,R}}(\gamma,\delta,\omega)$ computed for all four considered scenarios. As incident sound field, a unit-amplitude plane wave with $\phi_{\mathrm{pw}} = 0°$ is chosen. The radius of the microphone array is $R = 0.5\,\mathrm{m}$. All results have been derived for a temporal sampling frequency of $f_{\mathrm{s}} = 44.1\,\mathrm{kHz}$. We limit our investigations to the horizontal plane, hence $\theta_{\mathrm{pw}} = 90°$, $\theta = 90°$ and $\delta = 90°$. The reason behind this is that no high-resolution HRTF dataset which covers the entire sphere and no model of human perception covering elevated sources were available to the authors.

Figure 2 illustrates the experimental setup for the four scenarios. The first two considered scenarios exclude the effect of spatial sampling when computing the plane wave decomposition $\bar{P}_{\mathrm{cont}}(\phi,\theta,\omega)$. For this purpose, Equations (8) and (9) were evaluated numerically for modal and delay-and-sum beamforming, respectively. For modal beamforming different orders N have been considered. The third and fourth scenario includes spatial sampling. For modal beamforming the Sound Field Analysis Toolbox (SOFiA) [20] is used to compute the plane wave decomposition. A total of $M = 770$ Cardioid microphones located on a Lebedev grid were simulated. This allows to extract the spherical harmonics coefficients of the incident sound field up to an order of $N = 23$. The plane wave decomposition using delay-and-sum beamforming was derived by numerical evaluation of (6). A total of $M = 770$ pressure microphones distributed on a Lebedev grid were simulated. The delay-and-sum beamformer was implemented in the frequency domain. For all scenarios, the plane wave expansion coefficients have been computed for $\phi = -180°\ldots180°$ in steps of one degree. Equipment noise, microphone mismatch or misplacement have not been considered.

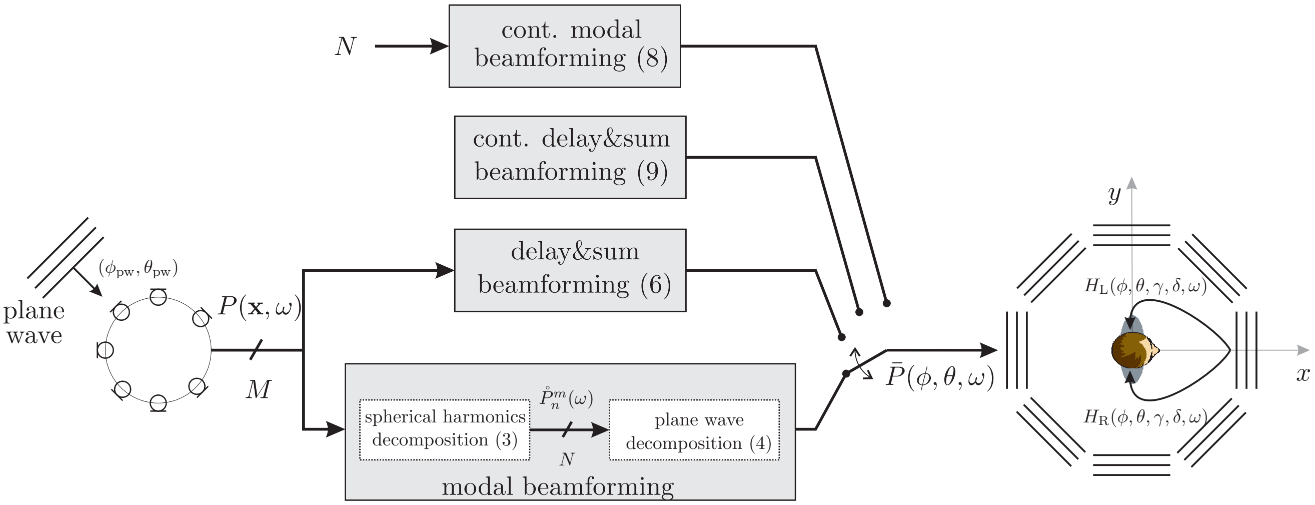

N
cont. modal beamforming (8)
cont. delay&sum beamforming (9)
delay&sum beamforming (6)
$(\phi_{\mathrm{pw}}, \theta_{\mathrm{pw}})$
plane wave
$P(\mathbf{x}, \omega)$
M
spherical harmonics decomposition (3)
$\mathring{P}_n^m(\omega)$
N
plane wave decomposition (4)
modal beamforming
$\bar{P}(\phi, \theta, \omega)$
y
x
$H_{\mathrm{L}}(\phi, \theta, \gamma, \delta, \omega)$
$H_{\mathrm{R}}(\phi, \theta, \gamma, \delta, \omega)$

Fig. 2: Block diagram of experimental scenarios used for the comparison of modal and delay-and-sum beamforming for data-based binaural synthesis.

The left/right ear signals $P_{\mathrm{L,R}}(\gamma, \delta, \omega)$ have been computed by numerical evaluation of (2). Instead of synthetic far-field HRTFs, measured ones at 3 m distance [21] have been used for this purpose. The ear signals have been computed for $\gamma = -90° \dots 90°$ in one degree steps. The overall procedure effectively produces BRTF datasets for an incident plane wave under free-field conditions.

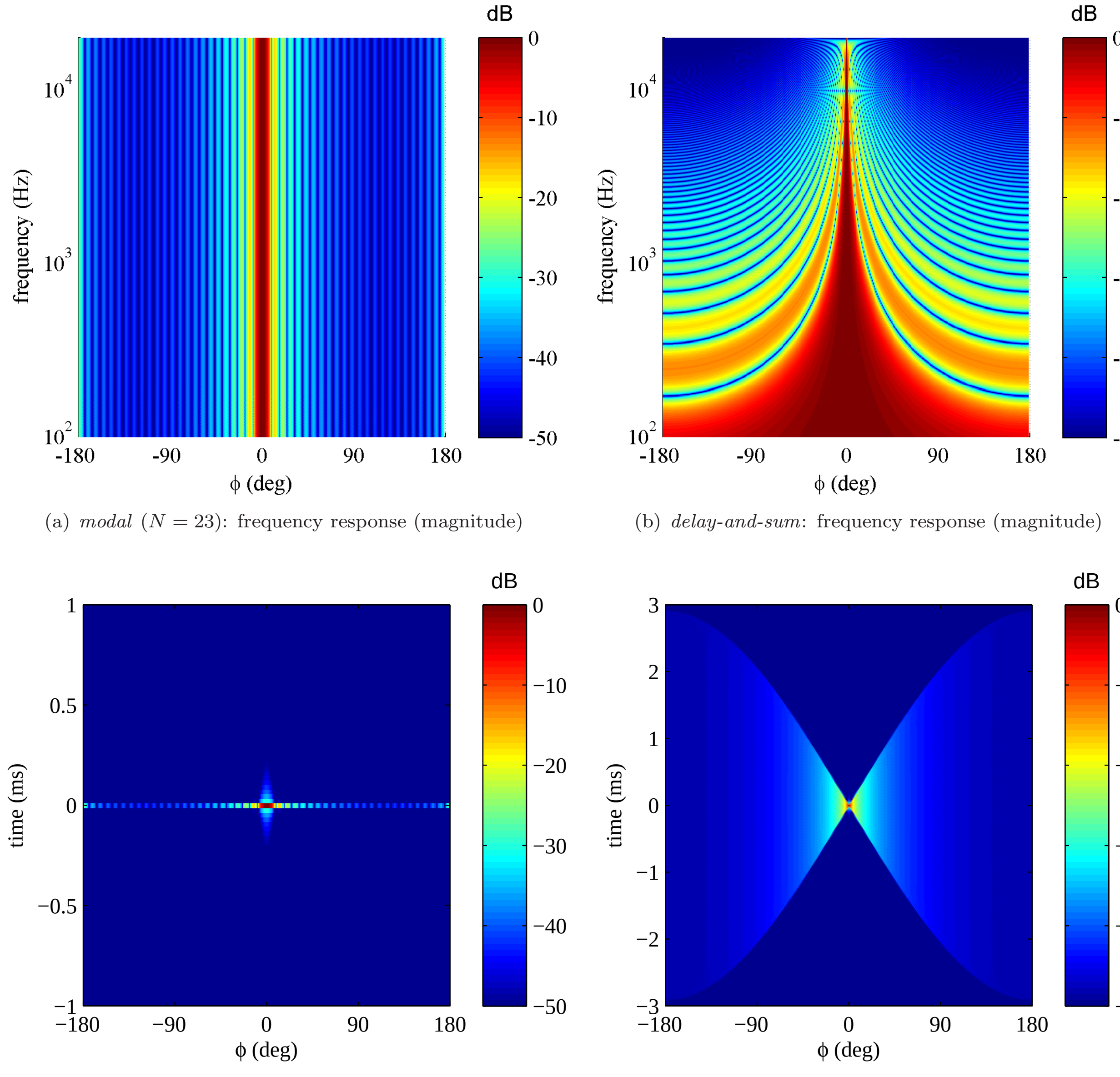
5.3. Modal versus Delay-and-Sum Beamforming

The theoretic plane wave decomposition of an incident plane wave would be given as

$$\bar{P}_{\mathrm{ideal}}(\phi, \theta, \omega) = \delta(\phi - \phi_{\mathrm{pw}})\delta(\theta - \theta_{\mathrm{pw}}) \,. \qquad (10)$$

Figure 3 shows the resulting plane wave decompositions for modal and delay-and-sum beamforming for a spatially continuous array. In order to support the comparison to the spatially sampled case shown later, the order of the modal beamformer was limited to $N = 23$. Figures 3(a) and 3(c) show the frequency response $\bar{P}_{\mathrm{MB,cont}}(\phi, \theta, \omega)$ and temporal response $\bar{p}_{\mathrm{MB,cont}}(\phi, \theta, t)$ of the modal beamformer. It can be observed that the response is independent from the frequency, as was already concluded in Section 4.2. The frequency response $\bar{P}_{\mathrm{MB,cont}}(\phi, \theta, \omega)$ does not resemble a Dirac pulse with respect to the angle ϕ due to the limited spatial bandwidth. The width of the main lobe at $\phi = 0°$ is directly linked to the spatial bandwidth N; lowering the bandwidth leads to a widening of the main lobe. The temporal response constitutes a Dirac pulse with varying amplitude over the angle ϕ. Figures 3(b) and 3(d) show the frequency response $\bar{P}_{\mathrm{DSB,cont}}(\phi, \theta, \omega)$ and temporal response $\bar{p}_{\mathrm{DSB,cont}}(\phi, \theta, t)$ of the delay-and-sum beamformer. It can be observed that the frequency response is frequency-dependent and especially that the main lobe at $\phi = 0°$ widens for low frequencies. Also the temporal response shows a complex behavior in both dimensions angle and time. Besides a Dirac pulse for $\phi = 0°$, the low-frequency widening in the frequency response leads also to a widening in the temporal response. The temporal extent of this widening is equal to the propagation time of the plane wave through the array, here $1 \text{ m}/c \approx 3\,\text{ms}$.

Figure 4 shows plane wave decompositions for a spatially sampled array. In comparison to Fig. 3 the effects of spatial sampling are clearly visible. For modal beamforming, it can be observed in Fig. 4(a), that additional side lobes are present above 5 kHz. The overall frequency response is now frequency-dependent. As a consequence, the temporal response shown in Fig. 4(c) is now quite complex and also extended into the temporal dimension. For delay-and-sum beamforming, spatial sampling artifacts become prominent above 3 kHz as can be seen in Fig. 4(b). The temporal response in Fig. 4(d) shows in principle a similar behavior as in the continuous case,



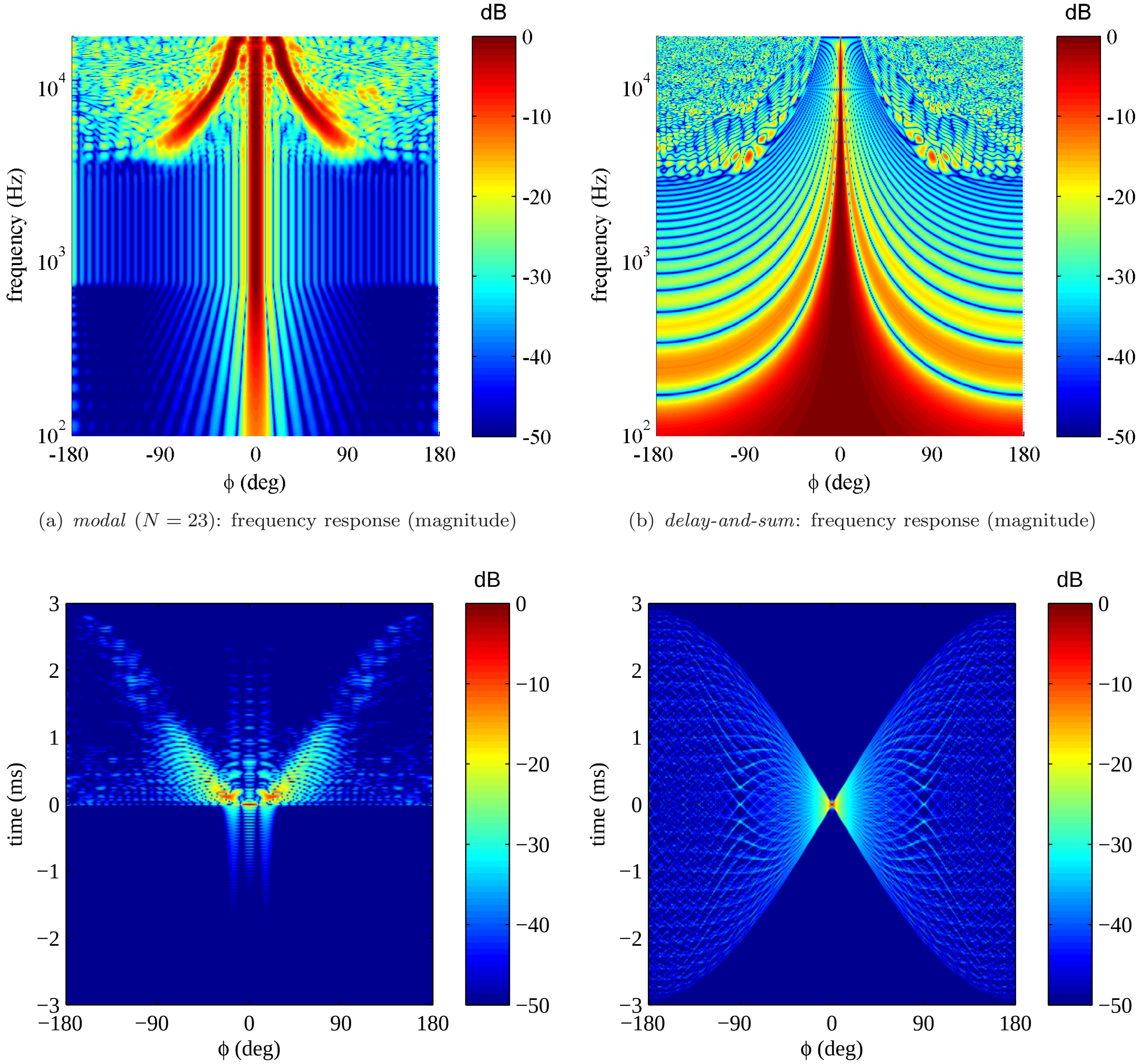
(a) *modal* ($N = 23$): frequency response (magnitude)

(b) *delay-and-sum*: frequency response (magnitude)

(c) *modal* ($N = 23$): temporal response (magnitude)

(d) *delay-and-sum*: temporal response (magnitude)

Fig. 3: Plane wave decompositions using modal and delay-and-sum beamforming of a broadband plane wave with incidence angle $\phi_{\text{pw}} = 0^\circ$ captured spatially continuously on a spherical aperture for $R = 0.5\,\text{m}$. The frequency response $\bar{P}(\phi, \theta, \omega)$ and temporal response $\bar{p}(\phi, \theta, t)$ are shown. The colorscale denotes the magnitude in dB. Note the different time axis in subfigures (c) and (d).



(a) *modal* ($N = 23$): frequency response (magnitude)

(b) *delay-and-sum*: frequency response (magnitude)

(c) *modal* ($N = 23$): temporal response (magnitude)

(d) *delay-and-sum*: temporal response (magnitude)

Fig. 4: Plane wave decompositions using modal and delay-and-sum beamforming of a broadband plane wave with incidence angle $\phi_{\text{pw}} = 0°$ captured by a spatially discrete microphone array with $M = 770$ microphones on a Lebedev grid. The frequency response $\bar{P}_{\text{S}}(\phi, \theta, \omega)$ and temporal response $\bar{p}_{\text{S}}(\phi, \theta, t)$ are shown. The colorscale denotes the magnitude in dB.

however, for $\phi \neq 0°$ additional Dirac-shaped contributions are visible.

So far the properties of modal and delay-and-sum beamforming when used as plane wave decomposition can be summarized as follows: Modal beamforming has the clear advantage of a constant directivity main lobe and high directivity at low frequencies. This comes at the cost of robustness against equipment noise and other practical aspects. The numerical complexity of modal beamforming is higher than for delay-and-sum beamforming. Delay-and-sum beamforming is numerically efficient and stable. A drawback is the frequency-dependency of the main lobe. The appearance of spatial aliasing is quite different in both methods.

The results shown so far imply that the plane wave decomposition differs significantly from the theoretic result for both modal and delay-and-sum beamforming techniques when using practical microphone arrays. The responses show contributions from additional incidence angles which may be spread additionally in time.

5.4. Data-based Binaural Resynthesis

The computed plane wave decompositions for all four scenarios have been filtered by the respective HRIRs and summed up for various head orientations, as outline in Section 5.2. The properties of the resulting BRIRs are discussed in this section. A first observation of the authors was that the BRIRs derived by delay-and-sum beamforming require a filtering by a 6 dB per Octave high-pass filter in order to not sound 'muffled'. This finding can be backed by considering the widening of the main lobe for low frequencies in delay-and-sum beamforming (see Figures 3(b) and 4(b)). Interestingly a 6 dB per Octave high-pass filter is also required for three-dimensional Wave Field Synthesis (WFS) [22].

Figure 5 shows the BRIRs and the original HRIRs for the considered four scenarios and for two different head orientations $\gamma = 0°$ and $\gamma = 90°$. At the far right the graph shows the impulse responses of the original HRIRs, to their left the BRIRs for the delay-and-sum beamformer and further to the left the BRIRs for the modal beamformer. The grey and black signals indicate spatially sampled and spatially continuous data, respectively.

The continuous data for the modal beamformer is shown for different orders N. At an order of $N = 23$ no obvious deviations from the impulse response of the original HRIR are visible, but for lower orders the impulse response differs significantly from the original one. The spatially continuous delay-and-sum beamformer adds two additional small peaks to the impulse response, one before the original peak and one afterwards. This is a consequence of the temporal widening that can be observed in Figure 3(d).

Spatial sampling adds further peaks – before and after the original peak – to the BRIRs derived by delay-and-sum beamforming. This can be especially observed for the contralateral (right) ear for $\gamma = 90°$. Again, this is a consequence of the additional peaks that can be observed in Figure 4(d). For the spatially sampled modal beamformer additional contributions can be observed after the original peak. This is a consequence of the properties shown in Figure 4(c).

The additional contributions could be a problem for the perception of the direction-of-arrival for the data-based BRIRs. The next section therefore investigates on the directional perception.

5.5. Localization

The human auditory system has a remarkable performance in estimating the localization of a sound source even in the presence of diffuse background noise or in reverberant situations. This ability is achieved by exploring different characteristics of the sound field present at the two ears. Besides non-acoustical cues like vision or the change of the sound field with head movements, spectral cues and interaural differences are used by the auditory system [1]. Therefore it is important that the computed BRIRs preserve these characteristics and deviate only in an inaudible range. This section deals with the perception of the direction-of-arrival of a given source signal convolved with the calculated BRIRs.

The two most important features for the estimation of the direction-of-arrival of a sound in the horizontal plane are interaural time differences (ITDs), and interaural level differences (ILDs). To investigate the influence of the different beamforming techniques a binaural model after [23] is applied. The Auditory Modeling Toolbox (AMT) [24] was used for this purpose. A binaural model simulates the behavior of

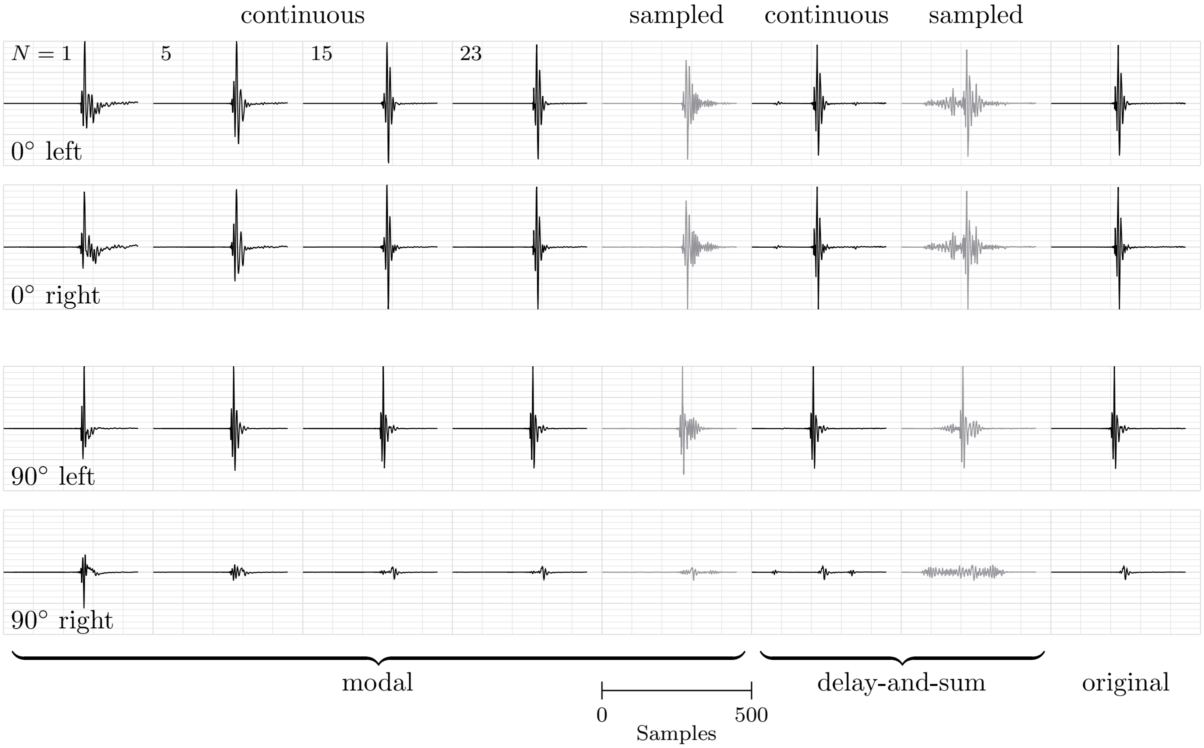
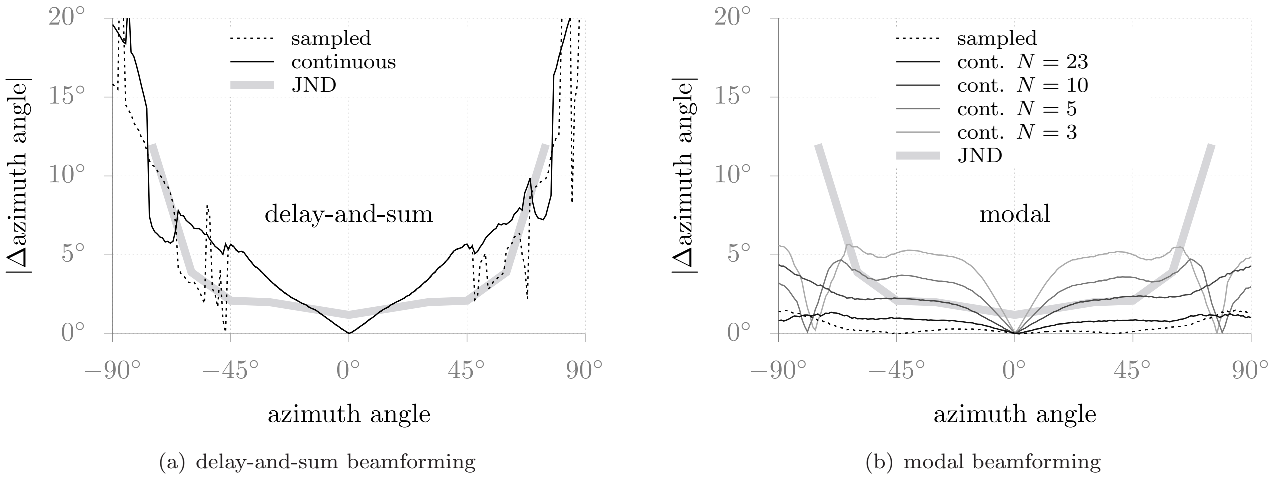


Fig. 5: Data-based BRIRs for azimuth angles of $\gamma = 0°$ and $\gamma = 90°$. Outmost to the right are the original HRIRs, followed by delay-and-sum beamforming BRIRs and modal-beamforming BRIRs to the right. HRTF is shown.

the two ears by applying a gammatone filterbank to the input signals of the left and right ear. In the frequency bands between 200 Hz and 1300 Hz the interaural phase difference (IPD) and the ILD are calculated. The IPD is ambiguous for frequencies greater than 700 Hz. In addition, the ILD has its maximum around 60° and is ambiguous for higher angles. On the other hand the sign of the ILD can be easily used to overcome the ambiguity of the IPD and to calculate the real ITD for a given stimulus. At the same time this mechanism accounts for the dominance of the perceived direction by the ITD for the considered frequency range [25].

In order to obtain an estimation of the azimuth for a calculated ITD, a lookup table is required that maps calculated ITDs on corresponding azimuth angles. Such a table was created from the HRIR data set described in Section 5.2 for azimuth angles $\gamma = -90° \ldots 90°$ in one-degree steps. A white noise signal of 1 s length was convolved with each set of the computed BRIRs for the prediction of the perceived direction. The resulting ear signals were then fed into the binaural model together with the lookup table to generate a prediction of the perceived azimuth angle. This procedure was repeated for all scenarios, additionally for various orders N of the continuous model beamformer and the available azimuth angles (see Section 5.2). The same procedure was applied to the original HRIR data set to obtain the desired perceived azimuth direction. The deviation of the azimuth $\Delta\gamma$ of the computed BRIRs is given by the absolute difference between the estimated azimuth of the BRIR and the HRIR. The result is presented in Fig. 6.

In addition to the azimuth deviation the just notable difference (JND) for localization in the horizontal plane [26] is indicated by the thick grey line. In the case of the delay-and-sum beamformer, the predicted deviation $\Delta\gamma$ of the perceived direction-of-arrival increases approximately linearly with in-



(a) delay-and-sum beamforming

(b) modal beamforming

Fig. 6: Deviation of the perceived direction of the computed BRTFs from the original HRTFs for delay-and-sum beamforming and modal beamforming. The spatial sampled beamforming is shown as dotted lines. The spatial continuous sampling for different orders N for the modal beamforming.

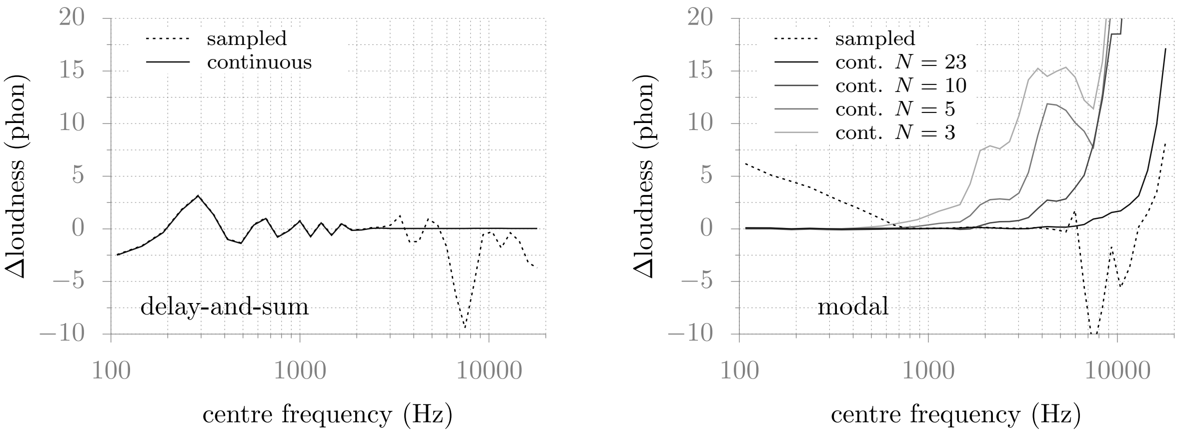
creasing azimuth angle. This corresponds to the properties of the time-domain plane wave responses shown in Fig. 4(d) and Fig. 3(d). The results for the spatially sampled data is very similar to the continuous data, which is not obvious when investigating the BRIRs in Fig. 5.

For the modal beamformer, the deviation $\Delta\gamma$ of the perceived direction-of-arrival is less pronounced for larger angles than for the delay-and-sum beamformer. The deviation for the spatially sampled data is completely below the JND, and it is therefore very likely that this deviation in terms of perceived angle is not audible. For the continuous sampling the deviation is plotted dependent on the order N. For lower orders the deviation gets larger. If the order is larger than $N = 10$, the deviation may become audible.

5.6. Coloration and Distance Perception

Another perceptual dimension which is influenced by spectral changes is coloration. The coloration of a stimulus in comparison to a baseline condition can be audible if changes in the spectrum exceed 1 dB. To investigate the influence of the beamforming techniques on the spectral properties a 1 s white noise signal was convolved with the left ear BRIRs and the original HRIRs. The deviation in loudness between the such derived stimuli were computed for an azimuth angle of $\gamma = 0°$. To account for the spectral resolution of the human auditory system, the spectrum was calculated in auditory filters ranging from 100 Hz to 20 kHz and applying a loudness compression to the power of 0.54 to the sound pressure [27]. The result is presented in Fig. 7. For spatially continuous delay-and-sum beamforming deviations can be observed in the low-frequency region up to 2 kHz in Figure 7(a). Interestingly these deviations have a similar shape as deviations observed for low frequencies in 2.5-dimensional WFS [28]. Hence, similar equalization techniques as proposed for WFS could be used to decrease the deviations. Spatial sampling introduces additional artifacts for frequencies above 4 kHz which can be accounted to spatial aliasing.

Figure 7(b) shows the spectral deviations for the continuous modal beamformer for different orders N. Deviations can be observed for the mid to high frequencies, while these deviations start at higher frequencies for higher orders N. Even for $N = 23$ prominent deviations are present for frequencies above 10 kHz. The spatially sampled modal beamformer behaves quite different in comparison to the continuous counterpart for $N = 23$. Deviations can be observed for low frequencies which are due to the limitation of the modal amplification applied in practice. This can also be observed in Figure 4(a) by the widening of the main lobe. The deviations



(a) delay-and-sum beamforming

(b) modal beamforming

Fig. 7: Deviation of the magnitude of the computed BRTFs from the original HRTFs for delay-and-sum beamforming and modal beamforming for an azimuth of $\gamma = 0°$.

above 5 kHz can be accounted to spatial aliasing, as for the delay-and-sum beamformer. The results given in Figure 7 imply that practical data-based binaural synthesis might be subject to coloration if no countermeasures are taken.

In addition to the perceived direction of a source, its localization is determined by its perceived distance, which depends on the amplitude of the source and on the naturalness of the given BRIRs. Changes of the spectrum or the interaural differences of the BRIR can lead to in-head localization of the auditory event [29]. A quantitative explanation of in-head localization is not possible at the moment. Results are available which show no strong dependency on the spectrum [30]. Hartmann [29], however found a stronger dependency on the spectrum than on the interaural parameters. As a consequence, we briefly review some findings from informal listening in the next section.

5.7. Informal Listening

In order to get more insight into the perceptual properties of the investigated techniques, listening examples have been generated by convolving an input signal with the respective left and right BRIRs. As input signals speech, music and a noise pulse train have been used. The listening examples have been computed for a range of different azimuth angles γ. We briefly report on the findings we have derived from informal listening. The reader is invited to download the examples [1].

For the spatially continuous delay-and-sum and modal ($N = 23$) beamformer hardly any differences between the original HRIRs and the data-based BRIRs can be observed for the speech and music samples. However, for the noise pulse train some slight coloration changes can be identified between the different techniques.

For the spatially sampled case and modal beamforming ($N = 23$), the localization deviations shown in Figure 6(b) can be confirmed, as well as the spectral deviations shown in Figure 7(b). However, the latter are not so prominent as one would assume from Figure 7(b). Again the spectral deviations can be heard best for the noise pulse train.

For the spatially sampled delay-and-sum beamforming technique spatial artifacts have been observed. The direct sound can be localized well from the expected direction. However, additional contributions were heard at the contralateral ear for azimuth angles $|\gamma| > 30°$. The cause for the undesired contributions seem to be the additional Dirac-shaped contributions that can be seen in Figure 4(d). Note this has not been predicted by the binaural model.

[1] http://audio.qu.tu-berlin.de/?p=808

Another interesting observation has been found from informal listening to the modal beamforming examples. For spatially continuous modal beamforming, the perceived distance changed with the order N. While for order $N = 0$ in-head localization occurred, the perceived distance increased with increasing order. Around $N = 15$, the perceived distance saturated towards the distance of the HRIR dataset used.

6. SUMMARY AND CONCLUSIONS

This paper compares modal versus delay-and-sum beamforming techniques as plane wave decomposition for data-based binaural synthesis. Results have been presented considering the localization and coloration of one single far-field source located in the horizontal plane. For practical microphone arrays, modal and delay-and-sum beamforming have comparable properties regarding localization and coloration. However, informal listening revealed undesired spatial contributions for delay-and-sum beamforming. Hence in the considered scenario, modal beamforming techniques seem to produce better results in the context of data-based binaural synthesis. Although quite some effort has been spent for the present study, it is hard to give a final conclusion regarding the favorable beamforming technique. Equipment noise has a major impact on the performance of practical modal beamforming techniques. So far equipment noise has not been considered. One reason is that the authors wanted to systematically introduce the different degradations one after each other in order to investigate their influence. This comprehensive study focused on two different beamforming techniques and the influence of spatial sampling. Another reason is that the resulting WNG is frequency-dependent for modal beamforming. As a consequence, low-frequency noise will be present in the computed BRIRs. At the current state not much is known about the perceptual influence of (spectrally shaped) noise in the context of BRIRs. Hence, further studies including listening experiments have to be conducted on the influence of the WNG in data-based binaural synthesis.

For a final conclusion also other aspects should be investigated. These are, for instance, the localization properties of elevated sources and the performance under natural (multipath) sound fields.

7. REFERENCES

[1] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization.* MIT Press, 1996.

[2] E.M. Wenzel, M. Arruda, D.J. Kistler, and F.L. Wightman. Localization using nonindividualized head-related transfer functions. *Journal of the Acoustic Society of America*, 94(1):111–123, July 1993.

[3] R. Duraiswami, D.N. Zotkin, Z. Li, E. Grassi, N.A. Gumerov, and L.S. Davis. System for capturing of high order spatial audio using spherical microphone array and binaural head-tracked playback over headphones with head related transfer function cue. In *119th AES Convention*, New York, USA, October 2005. Audio Engineering Society (AES).

[4] B. Rafaely and A. Avni. Interaural cross correlation in a sound field represented by spherical harmonics. *Journal of the Acoustic Society of America*, 127(2):823–828, February 2010.

[5] A. Avni and B. Rafaely. Sound localization in a sound field represented by spherical harmonics. In *International Symposium on Ambisonics and Spherical Acoustics*, Paris, France, May 2010.

[6] F. Melchior, O. Tiergart, G. Del Galdo, D. de Vries, and S. Brix. Dual radius spherical cardoid microphone arrays for binaural auralization. In *127th AES Convention*, New York, USA, October 2009. Audio Engineering Society (AES).

[7] S. Spors and H. Wierstorf. Evaluation of perceptual properties of phase-mode beamforming in the context of data-based binaural synthesis. In *IEEE-EURASIP International Symposium on Control, Communications, and Signal Processing*, pages 1–4, May 2012.

[8] N.A. Gumerov and R. Duraiswami. *Fast Multipole Methods for the Helmholtz Equation in three Dimensions.* Elsevier, 2004.

[9] E.G. Williams. *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography.* Academic Press, 1999.

[10] S. Spors and J. Ahrens. Generation of far-field head-related transfer functions using sound field synthesis. In *German Annual Conference on Acoustics (DAGA)*, March 2011.

[11] B. Rafaely. Phase-mode versus delay-and-sum spherical microphone array processing. *IEEE Signal Processing Letters*, 12(10):713–716, 2005.

[12] I. Balmages and B. Rafaely. Open-sphere designs for spherical microphone arrays. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2):727–732, 2007.

[13] B. Rafaely. Plane-wave decomposition of the sound field on a sphere by spherical convolution. *Journal of the Acoustical Society of America*, 116(4):2149–2157, 2004.

[14] Digital library of mathematical functions. `http://dlmf.nist.gov/`.

[15] R.A. Kennedy, P. Sadeghi, T.D. Abhayapala, and H.M. Jones. Intrinsic limits of dimensionality and richness in random multipath fields. *IEEE Transactions on Signal Processing*, 55(6):2542–2556, 2007.

[16] F. Zotter. *Analysis and Synthesis of Sound-Radiation with Spherical Arrays.* PhD thesis, Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, 2009.

[17] B. Rafaely. Analysis and design of spherical microphone arrays. *IEEE Transactions on Audio, Speech, and Language Processing*, 13(1):135–143, January 2005.

[18] B. Rafaely, B. Weiss, and E. Bachmat. Spatial aliasing in spherical microphone arrays. *IEEE Transactions on Signal Processing*, 55(3):1003–1010, 2007.

[19] J. Ahrens and S. Spors. A modal analysis of spatial discretization in spherical loudspeaker arrays used for sound field synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 2012. To appear.

[20] B. Bernschütz, C. Pörschmann, S. Spors, and S. Weinzierl. SOFiA - Sound field analysis toolbox. In *International Conference on Spatial Audio*, Detmold, Germany, November 2011.

[21] H. Wierstorf, M. Geier, A. Raake, and S. Spors. A free database of head related impulse response measurements in the horizontal plane with multiple distances. In *130th AES Convention*. Audio Engineering Society (AES), May 2011.

[22] S. Spors, R. Rabenstein, and J. Ahrens. The theory of wave field synthesis revisited. In *124th AES Convention*. Audio Engineering Society (AES), May 2008.

[23] M. Dietz, S.D. Ewert, and V. Hohmann. Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Communication*, 53(5):592–605, May 2011.

[24] Peter L. Søndergaard, John F. Culling, Torsten Dau, Nicolas Le Goff, Morten L. Jepsen, Piotr Majdak, and Hagen Wierstorf. Towards a binaural modelling toolbox. In *Proceedings of the Forum Acousticum 2011*, 2011.

[25] F.L. Wightman and D.J. Kistler. The dominant role of low-frequency interaural time differences in sound localization. *The Journal of the Acoustical Society of America*, 91(3):1648–61, Mar 1992.

[26] A.W. Mills. On the minimum audible angle. *The Journal of the Acoustical Society of America*, 30(4):237–246, May 1958.

[27] M. Epstein and J. Marozeau. *Loudness and intensity coding*, pages 45–69. Oxford University Press, 2010.

[28] S. Spors and J. Ahrens. Analysis and improvement of pre-equalization in 2.5-dimensional wave field synthesis. In *128th AES Convention*, pages 1–17, London, UK, May 2010. Audio Engineering Society (AES).

[29] M. Hartmann and A. Wittenberg. On the externalization of sound images. *The Journal of the Acoustical Society of America*, 99(6):3678–88, Jun 1996.

[30] A. Kulkarni and H.S. Colburn. Role of spectral detail in sound-source localization. *Nature*, 396(6713):747–9, 1998.