

Binaural Assessment of Multichannel Reproduction

H. Wierstorf, A. Raake and S. Spors

1 Introduction

Sound reproduction systems have evolved over the years in the direction of including more and more loudspeakers. The goal is to create a sense of auditory *immersion* in the listeners. A first binaural transmission of a concert via telephone was demonstrated in 1881 [9]. Later followed a proposal by Snow and Steinberg [29], aiming at the transmission of entire sound fields. The basic idea of these authors was that a sound field could be captured by an array of microphones and, consequently, be reproduced by replacing the microphones with loudspeakers that are fed with the signals picked up by the microphones. The loudspeaker signals then superimpose and, together, recreate the sound field in a similar manner as described by Huygen's principle [15]. If the number of independent microphones and loudspeakers is restricted to two each, the sound field is recreated correctly only in one specific location, the so-called *sweet spot*. This recording-and-reproduction technique known as stereophony, is still the prominent spatial-audio technique. This is due to its technical simplicity, the wide use of respective audio-mastering chains and the convincing perceptual results. The latter are mainly due to inherent properties of the human auditory system as have extensively been investigated in psychoacoustics—for example, [2].

In parallel to the continuing success of stereophony, the old idea as proposed by Steinberg and Snow has been revisited in recent years. Today's technology allows for the use of several hundreds of loudspeakers, such enabling the reproduction or synthesis of extended sound fields. These novel techniques are termed *sound-field synthesis*, which includes methods like *higher-order ambisonics*, HOA, and *wave-field*

A. Raake (✉) · H. Wierstorf
Assessment of IP-based Applications, Telekom Innovation Laboratories (T-Labs),
Technische Universität Berlin, Berlin, Germany
e-mail: alexander.raake@telekom.de

S. Spors
Signal Theory and Digital Signal Processing, Institute of Communications Engineering,
Universität Rostock, Rostock, Germany

synthesis, WFS. In these methods, the sound field is treated as being spatially-sampled by the loudspeaker array. As these methods target the recreation of a sound field with frequencies up to 20 kHz, the highest audible frequency, loudspeaker spacings below 1 cm are theoretically required to avoid spatial aliasing, which is still impossible. To get around this problem, research in the field has progressed towards the exploitation of psychoacoustics, to the end of synthesizing sound fields with inaudible perceptual errors as compared to natural hearing.

This chapter presents work that investigated how errors in a given sound field, as synthesized with a multichannel loudspeaker array employing the WFS method, influence listeners' localization of virtual sound sources.

Section 2 presents a discussion of what is needed in order to create a convincing reproduction of a given auditory scene, including relevant aspects of the sound-source localization performed by the human auditory system. Then, Sect. 3 provides the required theoretical background regarding WFS and outlines the binaural re-synthesis approach as used here for simulating different WFS setups. In Sect. 4, test results of localization tests with WFS are presented. Finally, in Sect. 5, it is shown how the results of the localization-test can be predicted by means of a model of binaural processing, that is, a so-called *binaural model*.

2 Creating a Convincing Auditory Scene

Anyone who deals with sound reproduction should consider how the reproduced scene is perceived by a listener. In the context of this chapter, an auditory scene is considered to consist of different elements, namely, the underlying auditory events that the listener interactively analyzes in terms of the available auditory information and which leads to the formation of auditory objects. According to [17], an auditory event is characterized by its loudness, its pitch, its perceived duration, its timbre, and spaciousness. Here, spaciousness comprises the perceived location and spatial extent associated with the auditory event [2]. Obviously, localization is just one aspect of auditory scenes. In this section, the perception of the entire scene is discussed, specifically addressing the role that localization plays in it.

Usually, in physical terms, the sound field associated to a reproduced sound scene deviates from that of the intended scene. Prior to reproduction, the intended scene is represented in terms of a *recorded* scene in a specific representation format, for example, created from a given recording using a specific set of source models, or modified based on such recordings and models. When such a stored scene representation is provided as input to a given reproduction method, the result is known as a *virtual sound scene*. A sound-reproduction system is required to present acoustic signals to the listeners' ears in such a way that the corresponding auditory scene matches the desired one as closely as possible, that is, the auditory scene as intended by its creator. The focus of the following considerations will be on system properties that enable a perceptually *authentic* or *plausible* reproduction of a sound scene [3]. Here, *authentic* means indiscernible from an explicit or implicit reference, in other

words true to the original. *Plausible* means that the perceived features of the reproduced scenes show plausible correspondence with the listener's expectations in the given context, without necessarily being authentic. Perceived features in this context are nameable and quantifiable features of the auditory scene and its elements, such as loudness, timbre, localization, and spatial extent.

The totality of features can be considered as the *character* of the auditory scene—perceived or expected. The expected character is often referred to as internal reference. However, for systems such as used for stereo reproduction, the listening experience itself has led to fixed schemata of perception that are linked with internal references of their own kind. The *quality* of a system as perceived by a listener is considered to be the result of *assessing* perceived features with regard to the desired features, that is, the internal reference. With quality being expressed as unidimensional index, systems can be ranked according to their perceived quality, and quality differences can be measured quantitatively.

No comprehensive overview of sound reproduction technology evaluation is available from the literature. A basic concept for evaluation is described in [23], which focuses mainly on the evaluation of multichannel stereophony-based reproduction, and an auditory-scene-based evaluation paradigm. By evaluating various 5.1 surround setups, it was found that the overall quality is composed of timbral and spatial fidelity. The same shows up in first results for typical WFS setups collected in the current research, namely, the timbral fidelity may be of greater importance for overall quality than spatial fidelity. However, the research on this topic is still at an early stage [38].

Timbral and spatial fidelity are perceptual constructs of multidimensional nature. In order to describe their perceptual dimensions that they are composed of, attribute descriptions have been sought for in different studies, employing verbal description and attribute ratings [6, 14, 39]. Multidimensional analysis methods such as multidimensional scaling or the repertory-grid technique followed by attribute scaling are suitable methods when no a-priori knowledge of the perceptual character of the auditory events associated to the stimuli is available. A relevant example is reported in [13], where so-called focused sources in WFS have been assessed. In the current chapter, a special focus is put on *localization* of sound sources reproduced with WFS. In this context, localization is one of the key features associated with spatial fidelity.

2.1 Localization

One basic ability of the human auditory system is the localization of sound. Localization describes the process of assessing the location of auditory events with respect to the positions and other properties of corresponding sound events. Note that the sound events giving rise to one auditory event can be manifold, for example, in the case of classical stereophony the two loudspeakers can create one auditory event at a position located between the two loudspeakers.

For localization, the auditory system evaluates differences between the two ear signals that depend on the position of the sound source. The most prominent cues

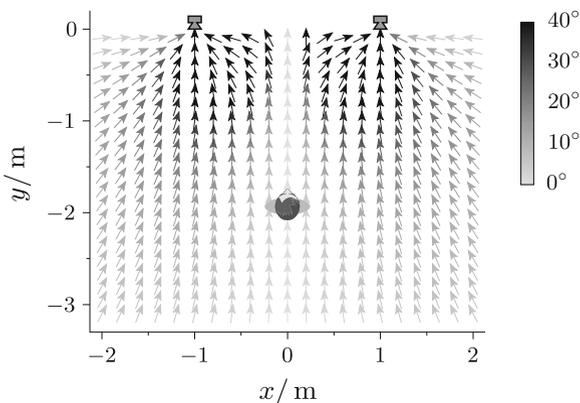
for localization are interaural time differences, ITDs, and interaural level difference, ILDs [2]. In addition, the auditory system estimates the distance and the vertical position of a sound source by interpreting monaural cues, such as the frequency spectrum resulting from a known type of source signal that is transmitted to the eardrums from a distance or from a specific vertical source position, thus yielding direction-specific filtering due to the shape of the outer ear. However note, that in this chapter, only displacements in the horizontal plane are addressed and, hence, only horizontally-oriented loudspeaker setups are used. As a consequence, vertical source displacement and distance are not considered in the localization assessment. Consequently, whenever the term localization is used, it refers to angular displacement in the horizontal plane.

For broad-band content, the localization is dominated by ITDs [36] of the spectral components below 1 kHz. Moreover, incident wave fronts arriving within a time-delay of around 1 ms after the first wave front are summed up by the auditory system, a phenomenon known as *summing localization*. Delayed wave fronts arriving later, but no later than approximately 50 ms after the first wave front, have no influence on localization at large, an effect known as the *precedence effect* [20]. Considering the delayed playout of sound by the different loudspeakers of a given multichannel loudspeaker setup, the precedence effect has implications with regard to the perceived directions of virtual sources. These are specific for the particular reproduction method and listening position.

In stereophony, the perceived location of an auditory event is caused by the superposition of the wave fronts coming from the two loudspeakers. For example, at low frequencies level differences between the two loudspeakers transform into a corresponding interaural time difference at a central position between the loudspeakers, the so-called *sweet spot*. For positions outside of sweet spot, the superposition is impaired, and the closest loudspeaker dominates localization. This is visualized in Fig. 1. The arrows in the figure point towards the location of the auditory event that a listener perceives when placed at the position of the arrow. The gray-shades of the arrow indicate the deviation from the intended direction, which is, in this example, given by the virtual source to be located in the middle between the loudspeakers, that is, at the x, y -point $(0, 0)$ m. The sweet spot is indicated by the position of the listener placed at $(0, 2)$ m. Calculation of the directions of the individual arrows was performed with the binaural model as described in Sect. 5. Figure 1 is provided for illustrative purposes only. For an overview of methods to predict the sweet spot in stereophony see, for example, [22].

In sound-field synthesis, the physical or authentic reproduction of a given sound field is intended. However, due to the limited number of loudspeakers and respective spatial aliasing, this is only possible up to a specific frequency for a given listening position. Due to this limitation, the localization of a virtual source reproduced with WFS depends on the position of the listener, and on the loudspeaker array configuration. For determining the best possible system layout in practise, it will be helpful to provide a model that predicts localization in the synthesis area of WFS, like the presentation as depicted in Fig. 1 for two-channel stereophony. In the further course of this chapter the development of such a tool will be described. Thereby an

Fig. 1 Sketch of the *sweet spot* phenomenon in stereophony. The *arrows* point into the direction of where the auditory event of a listener appears, if he/she sits at the position of the *arrow*. Increasing *gray-shades* of the *arrows* indicate the deviation from the intended direction which, in this case, is right in the *middle* between the two loudspeakers



existing binaural model will be modified to produce the output needed here, namely, for an application to localization performance analysis. Yet, to be able to specify such a model, the theory of WFS needs to be shortly revisited.

3 Wave-Field Synthesis

Wave-field synthesis, WFS, is a sound-field synthesis method that targets physically accurate synthesis of sound fields over an extended listening or synthesis area, respectively. WFS was formulated in the eighties for linear loudspeaker arrays [1]. In the following, a formulation of WFS is presented that is embedded into the more general framework of sound-field synthesis. Furthermore, restrictions regarding a 2-dimensional only loudspeaker setup are discussed, as well as the usage of loudspeakers with a given fixed inter-loudspeaker spacing. At the end of this section, WFS theory is discussed by means of an example.

3.1 Physical Fundamentals

The sound pressure, $P(\mathbf{x}, \omega)$, at the position, \mathbf{x} , synthesized by a weighted distribution of monopole sources located on the surface, ∂V , of an open area, $V \subset \mathbb{R}^3$, is given as the *single-layer potential*

$$P(\mathbf{x}, \omega) = \oint_{\partial V} G(\mathbf{x}|\mathbf{x}_0, \omega) D(\mathbf{x}_0, \omega) dA(\mathbf{x}_0), \tag{1}$$

where $G(\mathbf{x}|\mathbf{x}_0, \omega)$ denotes the sound field of a monopole source located at $\mathbf{x}_0 \in \partial V$. $D(\mathbf{x}_0, \omega)$ is its weight, usually referred to as the *driving signal*. The geometry of the problem is illustrated in Fig. 2.

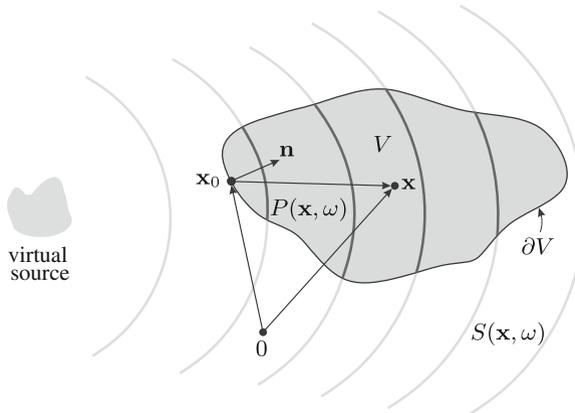


Fig. 2 Illustration of the geometry used to discuss the physical bases of sound-field synthesis and single-layer potential (1)

In sound-field synthesis, the monopole sources are referred to as secondary sources. Under free-field conditions, their sound field, $G(\mathbf{x}|\mathbf{x}_0, \omega)$, is given by the three-dimensional Green's function [37]. The task is to find the appropriate driving signals, $D(\mathbf{x}_0, \omega)$, for the synthesis of a virtual source, $P(\mathbf{x}, \omega) = S(\mathbf{x}, \omega)$, within V . It has been shown that the integral Eq. (1) can be solved under certain reasonable conditions [10].

3.2 Solution of the Single-Layer Potential for WFS

The single-layer potential (1) satisfies the homogeneous Helmholtz equation both in the interior and exterior regions, V and $\bar{V} := \mathbb{R}^3 \setminus (V \cup \partial V)$. If $D(\mathbf{x}_0, \omega)$ is continuous, its pressure value, $P(\mathbf{x}, \omega)$, is continuous when approaching the surface, ∂V , from the inside and outside. Due to the presence of the secondary sources at the surface, ∂V , the gradient of $P(\mathbf{x}, \omega)$ is discontinuous when approaching the surface. As a consequence, ∂V can be interpreted as the boundary of a scattering object with Dirichlet boundary conditions, hence as an object with low acoustic impedance. Considering this *equivalent scattering problem*, the driving signal is given as follows [11].

$$D(\mathbf{x}_0, \omega) = \partial_{\mathbf{n}}P(\mathbf{x}_0, \omega) + \partial_{-\mathbf{n}}P(\mathbf{x}_0, \omega), \quad (2)$$

where $\partial_{\mathbf{n}} := \langle \nabla, \mathbf{n} \rangle$ is the directional gradient in direction \mathbf{n} . Acoustic scattering problems can be solved analytically for simple geometries of the surface ∂V , such as spheres or planes.

The solution for an infinite planar boundary, ∂V , is of special interest. For this specialized geometry and Dirichlet boundary conditions, the driving function is given as

$$D(\mathbf{x}_0, \omega) = 2\partial_{\mathbf{n}}S(\mathbf{x}_0, \omega), \tag{3}$$

since the scattered pressure is the geometrically mirrored interior pressure given by the virtual-source model, $P(\mathbf{x}, \omega) = S(\mathbf{x}, \omega)$, for $\mathbf{x} \in V$. The integral equation resulting from introducing (3) into (1) for a planar boundary, ∂V , is known as *Rayleigh's first integral equation*.

An approximation of the solution for planar boundaries can be found by applying the *Kirchoff approximation* [7]. Here, it is assumed that a bent surface can be approximated by a set of small planar surfaces for which (3) holds, locally. In general, this will be the case if the wave length is much smaller than the size of a planar surface patch, hence, for high frequencies. In addition, the only part of the surface that is active is the one which is illuminated from the incident field of the virtual source. This also implies that only convex surfaces can be used to avoid contributions from outside of the listening area, V , to re-enter. The outlined principle can be formulated by introducing a window function $w(\mathbf{x}_0)$ into (3), namely,

$$P(\mathbf{x}, \omega) \approx \oint_{\partial V} G(\mathbf{x}|\mathbf{x}_0, \omega) \underbrace{2w(\mathbf{x}_0)\partial_{\mathbf{n}}S(\mathbf{x}_0, \omega)}_{D(\mathbf{x}_0, \omega)} dA(\mathbf{x}_0), \tag{4}$$

where $w(\mathbf{x}_0)$ describes a window function for the selection of the active secondary sources, according to the criterion given above. Equation (4) constitutes an approximation of the Rayleigh integral that forms the basis for WFS-type sound reproduction methods.

3.3 Virtual-Source Models

In WFS, sound fields can be described by using source models to calculate the driving function. The source model is given as $S(\mathbf{x}, \omega)$, and with (3), the driving function can be calculated. Two common source models are point sources and plane waves. For example, point sources can be used to represent the sound field of a human speaker, whereas plane waves could represent room reflections.

The source model for a point source located at \mathbf{x}_s is given as

$$S(\mathbf{x}, \omega) = \hat{S}(\omega) \frac{e^{-i\frac{\omega}{c}|\mathbf{x}-\mathbf{x}_s|}}{|\mathbf{x} - \mathbf{x}_s|}, \tag{5}$$

where \hat{S} is the temporal spectrum of the source signal $\hat{s}(t)$.

The source model for a plane wave with a propagation direction of \mathbf{n}_s is given as

$$S(\mathbf{x}, \omega) = \hat{S}(\omega) e^{-i\frac{\omega}{c}\mathbf{n}_s\mathbf{x}}. \tag{6}$$

3.4 2.5-Dimensional Reproduction

Loudspeaker arrays are often arranged within a two-dimensional space, for example as a linear or circular array. From a theoretical point of view, the characteristics of the secondary sources in such setups should conform to the two-dimensional free-field Green's function. Its sound field can be interpreted as the field produced by a line source. Loudspeakers exhibiting the properties of acoustic line sources are not practical. Real loudspeakers have properties similar to a point source. In this case three-dimensional free-field Green's functions are used as secondary sources for the reproduction in a plane, which results in a dimensionality mismatch. Therefore, such methods are often termed *2.5-dimensional synthesis* techniques. It is well known from WFS, that 2.5-dimensional reproduction techniques suffer from artifacts [30]. Amplitude deviations are most prominent.

3.5 Loudspeakers as Secondary Sources

Theoretically, when an infinitely-long continuous secondary source distribution is used, no errors other than an amplitude mismatch due to 2.5-dimensional synthesis are expected in the sound field.

However, such a continuous distribution cannot be implemented in practice, because a finite number of loudspeakers has to be used. This results in a *spatial sampling* and *spatial truncation* of the secondary source distribution [28, 30]. In principle, both can be described in terms of diffraction theory—see for example [4]. Unfortunately, as a consequence of the size of loudspeaker arrays and the large range of wave lengths in sound as compared to light, most of the assumptions made to solve diffraction problems in optics are not valid in acoustics. To present some of the basic properties for truncated and sampled secondary source distributions, simulations of the sound field are made and interpreted in terms of basic diffraction theory, where possible.

Spatial Sampling

The spatial sampling, which is equivalent to the diffraction by a grating, only has consequences for frequencies greater than the aliasing frequency

$$f_{\text{al}} \geq \frac{c}{2\Delta x_0}, \quad (7)$$

where Δx_0 describes the spacing between the secondary sources [27]. In general, the aliasing frequency is dependent on the listening position \mathbf{x} —compare [28, Eq. 5.17].

For the sound field of a virtual source, the spatial aliasing adds additional wave fronts to the signal. This can be explained as follows. Every single loudspeaker is

sending a signal according to (3). If no spatial aliasing occurs, the signals cancel each other out in the listening area, with the exception of the intended wave front. In the case of spatial aliasing and for frequencies above the aliasing frequency, the cancellation does not occur, and several additional wave fronts reach a given listener position, following the intended wave front. The additional wave fronts also add energy to the signal.

Truncation

The spatial truncation of the loudspeaker array leads to further restrictions. Obviously, the listening area becomes smaller when a smaller array is used.

Another problem is that a smaller loudspeaker array introduces diffraction in the sound field. The loudspeaker array can be seen as a single slit that causes a diffraction of the sound field propagating through it. This can be described in a way equivalent to the phenomenon of edge waves as shown by Sommerfeld and Rubinowicz—see [4] for a summary. The edge waves are two additional spherical waves originating from the edges of the array, which can be softened by applying a tapering window [31].

3.6 Example

For the simulations shown in Fig. 3, a circular loudspeaker array is assumed with a diameter of 3 m, consisting of 56 loudspeakers, which results in a loudspeaker spacing of $\Delta x_0 = 0.17$ m. Note that a circular array constitutes a 2.5-dimensional scenario.

Figure 3 illustrates the reproduced wave field for two different frequencies of the virtual plane wave, and its spatio-temporal impulse response. For 1 kHz, the reproduced wave field shows no obvious artifacts. However, some inaccuracies can be observed close to the secondary sources. This is due to the approximations applied for the derivation of the driving function in WFS. For plane waves with the frequencies of 2 and 5 kHz sampling artifacts are visible, and rather evenly distributed over the listening area. The amplitude decay in the synthesized plane wave due to the 2.5-dimensional approach is clearly visible in Fig. 3a.

The impulse response depicted in Fig. 3d shows that WFS reconstructs the first wave front well, with prominent artifacts following behind. The artifacts consist of additional wave fronts coming from the single loudspeakers. These additional wave fronts would vanish for a loudspeaker array with a loudspeaker spacing smaller than $\lambda_{\min}/2$, where λ_{\min} is the smallest wavelength to be reproduced.

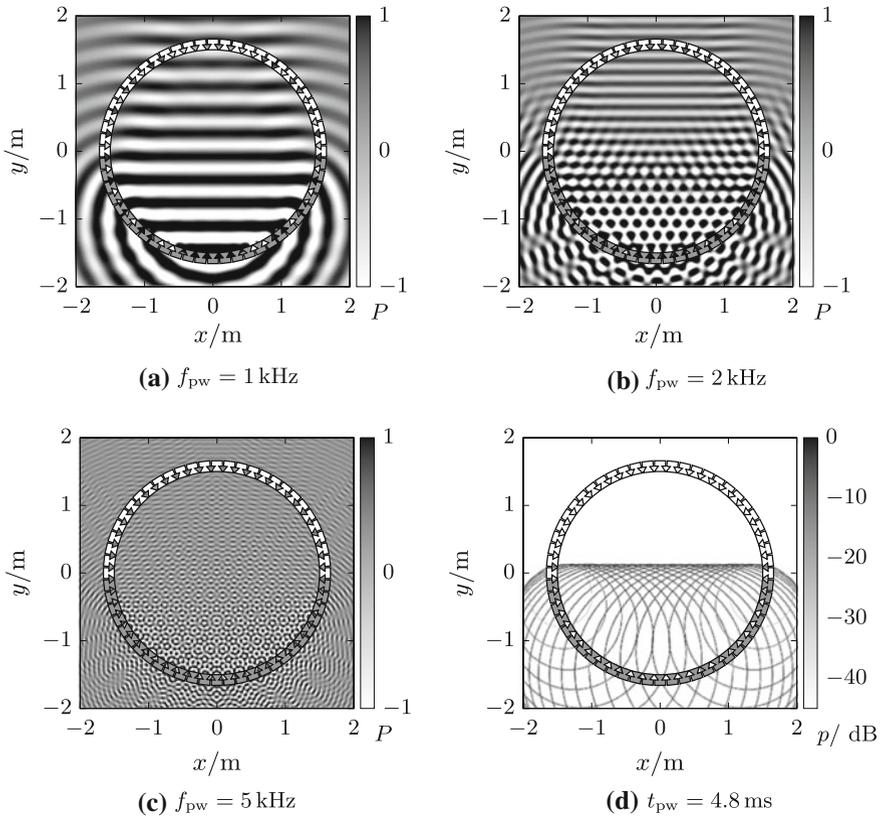


Fig. 3 Snapshot of sound fields synthesized by 2.5-dimensional WFS using a *circular array*, $R = 1.50 \text{ m}$, with 56 loudspeakers. The virtual source constitutes a plane wave with an incidence angle of -90° and the frequency f_{pw} . The *gray shades* denote the acoustic pressure, the active loudspeakers are *filled*. **a–c** Snapshot of $P(\mathbf{x}, \omega)$. **d** Snapshots of broad-band $p(\mathbf{x}, t)$

4 Localization Measurement with Regard to Wave-Field Synthesis

4.1 Binaural Synthesis

As discussed in the last section, spatial aliasing depends on the listening position and the loudspeaker array. In a listening test targeting localization assessment, it is not sufficient anymore to test only one position and one loudspeaker setup. Instead, different listener positions have to be investigated, and different types of loudspeaker setups must be applied, switching configurations more or less instantaneously and without disturbing the listener.

In practice, a real-life physical setup can be approximated by applying dynamic binaural synthesis to simulate the ear signals for the listeners for all needed conditions. Dynamic binaural synthesis simulates a loudspeaker by convolving the *head-related transfer functions*, HRTFs with an intended audio signal, which is played back to a listener via headphones. Simultaneously, the orientation of the head of the listener is tracked, and the HRTFs are exchanged according to the head orientation of the listener. With this dynamic handling included, results from the literature show that the localization performance for a virtual source is equal to the case of a real loudspeaker, provided that individual HRTFs are used. For non-individual HRTFs, the performance can be slightly impaired, and an individual correction of the ITD may be necessary [19]. For the case of a real loudspeaker, the localization performance of listeners in the horizontal plane lies around 1° – 2° . Note that this number only holds for a source located in front of the listener. For sources to the sides of the listener, localization performance can get as bad as 30° , due to the fact that the ITD changes only little for positions to the side of the head. An accuracy of around 1° sets some requirements on the experimental setup. One has to ensure that the employed setup introduces a measurement error that is smaller than the error expected in terms of human localization performance. This is especially difficult for the acquisition of the perceived direction based on the indications/judgments collected from the test listener. A review of different techniques and their advantages and drawbacks can be found in [21, 25].

In general, the WFS simulation based on binaural synthesis can be implemented as follows. For each listener position and individual loudspeaker of the WFS system, a dedicated set of HRTFs is used. The ear signals are constructed from the loudspeaker-driving signals, which are convolved with the respective *head-related impulse responses*, HRIRs, and then superimposed. For the tests and respective setups considered in this chapter, the *SoundScape Renderer* has been used as the framework for implementation [12], as well as the *Sound-Field Synthesis Toolbox* [34]. To simulate loudspeaker setups that deviate from the set of HRIRs, which are typically measured with a loudspeaker at a given radial distance from the dummy head, for different angular loudspeaker positions, the HRIRs are extrapolated using delay and attenuation, according to the propagation delay and respective distance-related attenuation.

4.2 Verification of Pointing Method and Dynamic Binaural Synthesis

For the localization indication, it has been decided to use a method where the listeners have to turn their heads to the direction of the auditory event during sound presentation. This has the advantage that the listener is directly facing the virtual source, a region where the localization performance is at its best. If the listeners point their heads in the direction of the auditory event, an estimation error of the sources at the side will occur, due to an interaction with the motor system. In other words, listeners



Fig. 4 Measurement of HRTFs with a dummy head in an anechoic chamber [35]

do not turn their heads sufficiently far as to indicate the real location. This can be overcome by adding a visual pointer that indicates to the listeners where their noses are pointing at [18].

Before investigating the localization in WFS, a pre-study was conducted [35], where the performance of the pointing method was verified, and it was studied whether the dynamic binaural synthesis introduces errors to the localization of a source. For the binaural synthesis, non-individual HRTFs were used that had been measured with a KEMAR dummy head in an anechoic chamber, as shown in Fig. 4 [32].

For the pre-study, the listeners were seated in an acoustically damped listening room, 1.5 m in front of a loudspeaker array, with an acoustically transparent curtain in between. Eleven of the 19 loudspeakers of the array were used as real sources and also simulated via the dynamic binaural synthesis. The listeners were seated on a heavy chair and were wearing open headphones, AKG K601, both for the loudspeaker and the headphone presentation. A laser pointer and the head tracker sensor, Polhemus Fastrack, were mounted onto the headphones. A visual mark on the curtain was used to calibrate the head-tracker setup at the beginning of each test run. For each trial, the listener was presented with a Gaussian white-noise train, consisting of periods of 700 ms noise and 300 ms silence. The experimenter instructed the listener to look towards the perceived source and to hit a key when the intended direction was correctly indicated by the laser. The conditions in terms of virtual-source directions and loudspeaker-versus-headphone presentation were randomized. The setup is shown in Fig. 5.

Eleven listeners participated in the experiment, and every condition was repeated five times. Figure 6 shows the deviation between the direction of the auditory event

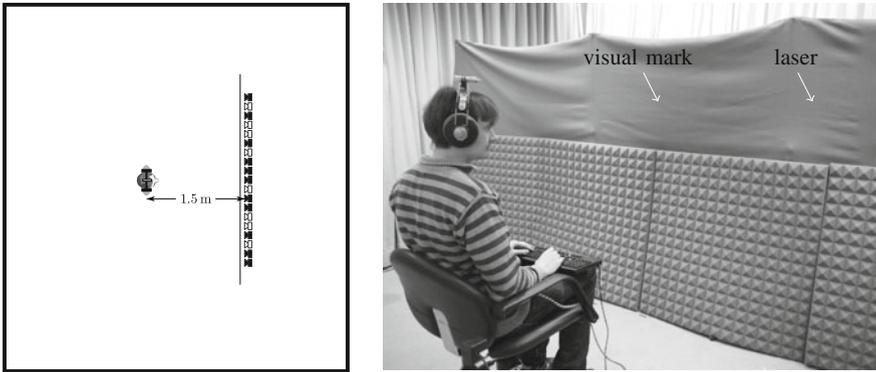


Fig. 5 Sketch of the experimental setup (left) and picture of a listener during the experiment (right), [35]. Only the filled loudspeakers were used in the first experiment. The light in the room was dimmed during all experiments

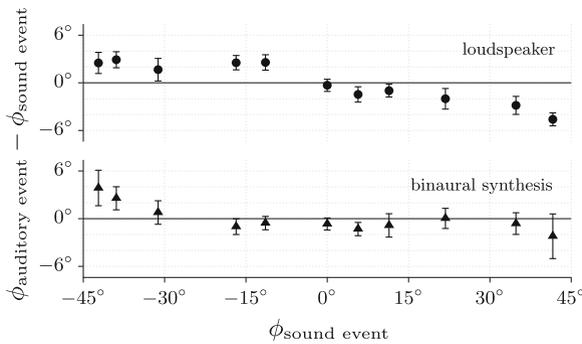


Fig. 6 Deviation $\Delta\phi$ between the position of the auditory event and the position of the sound source. The mean across all listeners and the 95 % confidence intervals are indicated. *Top row* Real loudspeakers. *Bottom row* Binaural synthesis

and the sound event for every single loudspeaker. It can be seen that there are only slight differences between the binaural simulation using headphones and the localization of the noise coming from the real loudspeakers. The mean absolute deviation, $\Delta\phi$, of the direction of the auditory event compared to the position of the sound event together with its confidence interval is $2.4^\circ \pm 0.3^\circ$ for the real loudspeakers and $2.0^\circ \pm 0.4^\circ$ for the binaural synthesis. In both cases, the mean deviation gets higher for sources more than 30° to the side of the listener. For these conditions, the position of the auditory event is underestimated and pulled towards the center. To avoid this kind of error in the examination of localization in WFS, only virtual-source positions within the range of -30° to 30° are be considered in the following. The only differences between simulation and the loudspeakers can be found in the localization blur for individual listeners. The mean standard deviation for a given position is $2.2^\circ \pm 0.2^\circ$ for the loudspeakers and $3.8^\circ \pm 0.3^\circ$ for the binaural synthe-

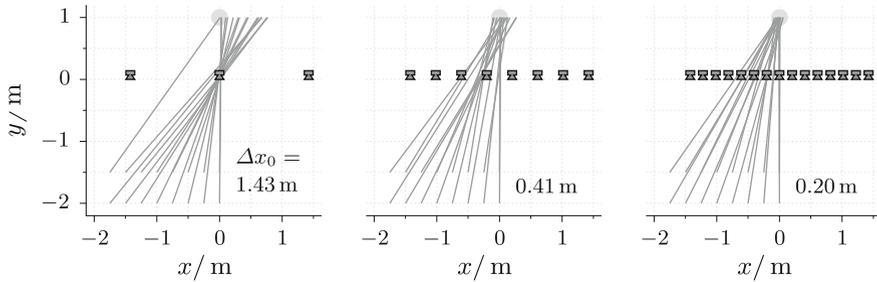


Fig. 7 Average directions that listeners were looking at from the 16 different listener positions evaluated, [33]. Results for the three different loudspeaker spacings. The *gray point* above the loudspeaker array indicates the intended virtual-source position

sis conditions. This is most likely due to higher ease of localization when listening to the real loudspeaker—an interesting issue discussed further in [35].

4.3 Localization Results for Wave-Field Synthesis

The same setup as presented in Sect. 4.2 and shown in Fig. 5 was employed. This time, a virtual source located at $\mathbf{x}_s = (0, 1)$ m was presented via the loudspeaker array, now driven by WFS. Following the descriptions above, the loudspeaker array was simulated using dynamic binaural synthesis. It had a length of 2.85 m, and consisted of 3, 8, or 15 loudspeakers, translating to a loudspeaker spacing of 1.43, 0.41, and 0.20 m. Like in the pre-test, the listeners were seated in the heavy chair in front of the curtain. Now, however, different listener positions of the listeners were introduced via binaural synthesis. The positions were at $x = -1.75$ m up to 0 m in steps of 0.25 m, with $y = -2$ m, and $y = -1.5$ m, leading to a total of 16 positions—compare Fig. 7.

Figure 7 summarizes the results. A line goes from every position of the listener to the direction where the corresponding auditory event was perceived, taking the average over all listeners. The gray point indicates the position of the virtual point source. As can be seen from this figure, the loudspeaker array with 15 loudspeakers leads to high localization accuracy. The intended position of the auditory event is reached with a deviation of only 1.8° . For the arrays with eight and three loudspeakers, the deviations are 2.7° and 6.6° , respectively. For the array with three loudspeakers, a systematic deviation of the perceived direction towards the loudspeaker at $(0, 0)$ m can be observed for all positions except one. For all three array geometries, the mean error is slightly smaller for the listener positions with $y = 2$ m than for that with $y = 1.5$ m. The results for every single position are presented in Fig. 10.

In WFS, the aliasing frequency determines the cut-off frequency up to which the sound field is synthesized correctly. For localization, mainly the frequency content below 1 kHz is important. The aliasing frequencies for the three loudspeaker arrays are 120, 418, 903 Hz, starting from the array with a spacing of 1.43 m between the loudspeakers. Further, the aliasing frequency is position-dependent and can be

higher for certain positions. According to the results for the 15-loudspeaker array, a loudspeaker spacing of around 20 cm seems to be sufficient to yield unimpaired localization for the entire range of listener positions. For a central listening position, a similar result was obtained in other experiments [28, 30, 38]. It was further discovered with the measuring method used here that even for spacings twice as large, the localization is only impaired by 1° . For larger spacings, the behavior tends more towards a stereophonic setup, that is, showing a sweet spot and localization towards the loudspeaker nearest to the sweet spot.

In the next section, a binaural model will be extended to enable predictions of the localization test results found for WFS. It is shown how the model can be used to predict localization maps for the entire listening area, going beyond the set of tested conditions.

5 Predicting Localization in Wave-Field Synthesis

In this section, a binaural model will be extended to enable predictions of the localization test results found for WFS. It is shown how the model can be used to predict localization maps for the entire listening area, going beyond the set of tested conditions.

An important difference of WFS in comparison to stereophony is the feature of uniform localization across an extended listening area. This is in clear contrast to the confined sweet spot of stereophonic systems. The sweet spot phenomenon was illustrated in Fig. 1. It would be of advantage to be able to predict such *localization maps* for further loudspeaker setups and reproduction methods as well, for example, for multichannel loudspeaker arrays and WFS. To this end the binaural model after Dietz [8] was modified and extended to be able to predict the direction of the auditory events for any pair of given ear signals. Specifically, the same ear signals were used as input signals to the binaural model, as have been synthesized for the listening tests by means of binaural synthesis—see Sect. 4.

In the following, the predictions from the binaural model are compared to the actual localization data as obtained in the listening tests. Given that the model provides localization predictions that agree with the listening-test data, it can be used to create localization maps for setups other than those that have been investigated perceptually.

5.1 Modelling the Direction of the Auditory Event

Binaural auditory models as outlined in [16], this volume, typically process the signals present at the right and left ear canal. For example, the model developed by Dietz [8] provides as its output a set of interaural arrival-time-difference values, ITDs, namely, one for every auditory filter. For the prediction of the direction of an auditory event, the ITD values have to be transformed into azimuth values that

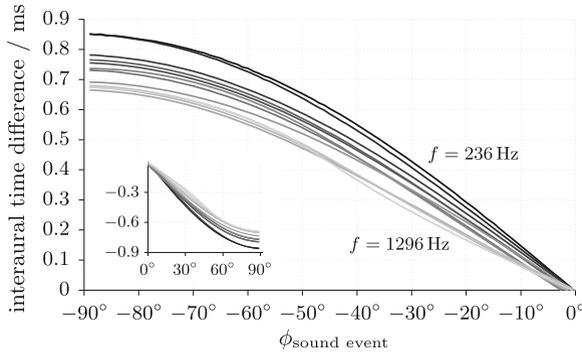


Fig. 8 Lookup table for ITD values and corresponding sound-event directions, shown for the first twelve auditory filters. Data derived with the binaural model of Dietz [8]

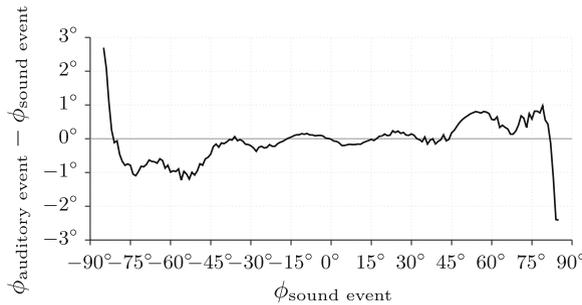


Fig. 9 Deviation of the predicted direction of auditory events from the direction of corresponding sound events

describe the direction of the auditory events. This can be accomplished by means of a lookup table of ITD values and corresponding angles [8].

In the study presented here, such a table was created by convolving a 1-s-long white-noise signal with head-related impulse responses from the same database as has been used for the listening tests presented in Sect. 4. The database has a resolution of 1°. The convolved signals were fitted to match the input format of the binaural model and stored. The result for the first twelve auditory filters are shown in Fig. 8.

For the prediction of the perceived direction belonging to a given stimulus, the binaural model first calculates the ITD values. Then, for each of the twelve auditory channels, the ITD value is transformed into an angle by use of the lookup table—Fig. 8. If the absolute ITD value in an auditory channel turns out to be larger than the natural limit of 1 ms, this channel is disregarded in the following step. Afterwards, the median value across all angles is taken as the predicted direction. If the angle in an auditory filter differs by more than 30° from the median, it is considered an outlier and skipped, and the median is re-calculated.

In order to test whether the predictions depend on the actual method used for determining the look-up table, the head-related impulse responses from the same HRTF database were convolved with another white-noise signal and again fitted to match the model input format. Figure 9 shows the deviation between the predicted direction of the auditory event and the direction of the sound sources for this case. Only for angles of more than $\pm 80^\circ$, the deviation exceeds a value of 1.2° . The deviation is due to the decreasing slope of the ITD for large angles—compare Fig. 8. This effect makes it more difficult to achieve proper fit of the ITDs and their corresponding azimuths.

5.2 Verification of Prediction

The modified binaural model can now be used to *predict* the direction of an auditory event. In this part of the study, the model prediction performance were analyzed in view of the localization results of Sect. 4. Due to limitations of the binaural model used here—for example, the precedence effect is not included—it might well be that it fails to properly predict localization in more complex sound fields, such as those synthesized with WFS. To check on this, the predictions that the model renders for the setups that have been investigated by the listening tests—compare Sect. 4—have been analyzed. See Fig. 10 regarding the results obtained in the localization test and the corresponding model predictions. Open symbols denote a listener distance of 2 m to the loudspeaker array, filled symbols a distance of 1.5 m. The model predictions are presented as dashed lines for the 2 m case and solid lines for 1.5 m case.

For most of the configurations, the model predictions are in agreement with the directions perceived by the listeners. Only for positions far to the side some deviations of up to 7° are visible. The overall prediction error of the model is of 1.3° , ranging from 1.0° for the array with 15 loudspeakers to 2.0° for the array with three loudspeakers. These results indicate that the model is able to predict the perceived direction of a virtual source in WFS almost independently from the listener position and the array geometry.

5.3 Localization Maps

With the method as presented in the previous sections, it is now possible to create a localization map similar to the one shown in Fig. 1. To this end, the ear signals for each intended listener position and loudspeaker array are simulated via binaural synthesis. Then these signals are fed into the binaural model, which delivers the predicted direction for the respective auditory event. In the following, the procedure is illustrated with two different loudspeaker setups. The first one is the same setup as used for the WFS localization test—compare Sect. 4—however, additional listener

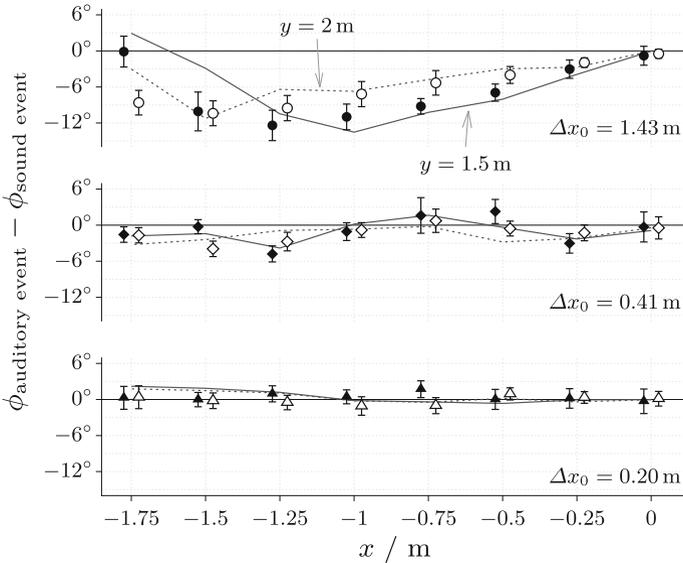


Fig. 10 Means and 95 % confidence intervals of localization errors in WFS dependent on the listening position and the loudspeaker spacing. *Open symbols* Listener positions at $y = 2$ m. *Closed symbols* Listener positions at $y = 1.5$ m. The *lines* denote the model predictions. *Solid line* $y = 1.5$ m. *Dashed line* $y = 2$ m

positions. The second one is a circular loudspeaker array that is installed in the authors’ laboratory.

The following virtual sources were chosen: (i) a point source located either at the center of the array or one meter behind it, (ii) a plane wave traveling into the listening area vertically to the loudspeaker array. Both cases were evaluated separately, because of the expected differences in localization, that is, a point source stays at its position when if the listener moves around in the listening area, but a plane wave moves with the listener.

The resulting localization maps can be presented in the form of arrows pointing into the direction of the auditory event as in Fig. 1. Alternatively, a color can be assigned to each position, denoting the deviation of the perceived direction from the intended one. The latter format renders a better resolution.

Linear Loudspeaker Array

Figure 11 shows localization maps for the three different linear loudspeaker setups as also used in the listening tests of Sect. 4. The first array consists of three loudspeakers with a spacing of 1.43 m between them, the second of eight loudspeakers with a spacing of 0.41 m, and the third of 15 loudspeakers with a spacing of 0.20 m. The localization maps presented at the top of the figure show a sampling of the listening

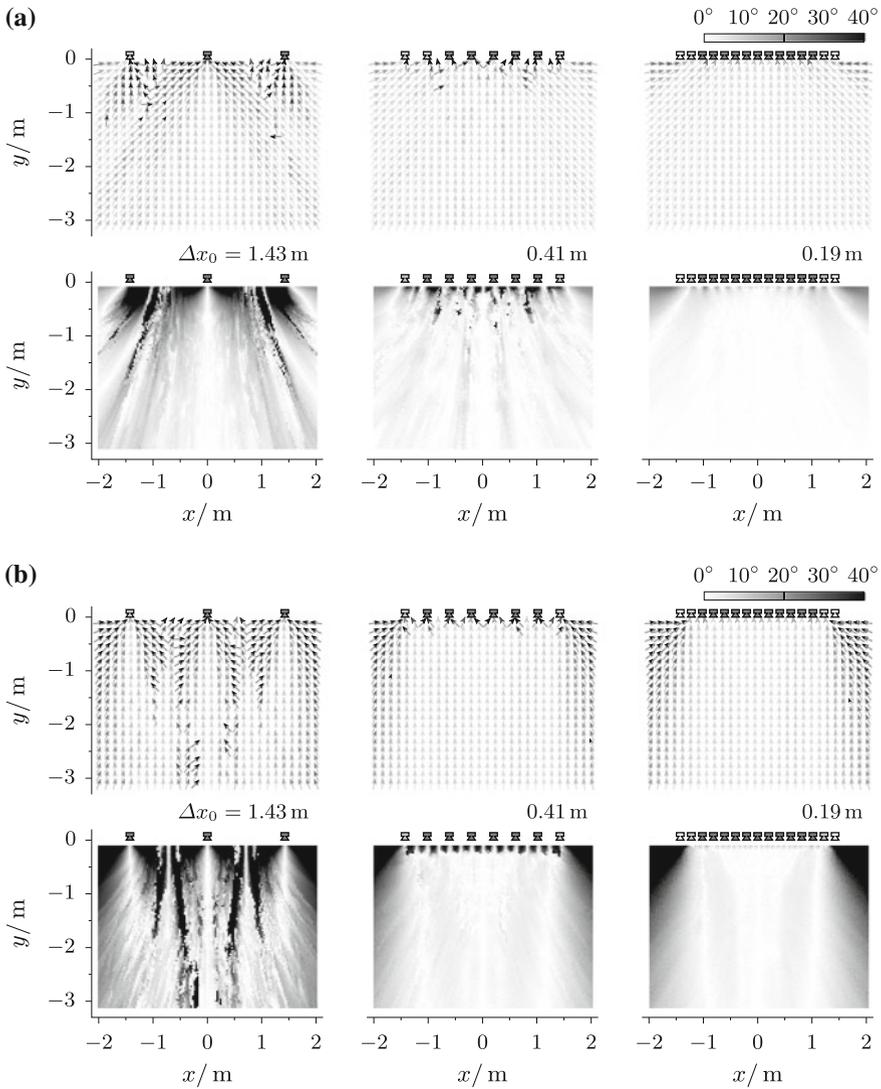


Fig. 11 Localization maps for a linear loudspeaker array driven by WFS for (a) a virtual point source, (b) a plane wave. The *arrows* point into the direction of where the auditory event of a listener appears, if he/she sits at the position of the *arrow*. The *gray-shades* indicate the deviation from the intended direction

area of 21×21 points. The arrows indicate the predicted direction in which the auditory event is predicted to appear as seen from the respective listening position. The localization maps presented at the bottom of the figure show a sampling of the listening area of 135×135 points. The gray-shades of the points indicate the absolute

deviation between the predicted direction of the auditory event and the prospective direction of the virtual sound event. Absolute deviation values are clipped at 40° .

In Fig. 11a, the sound event corresponds to a point source located at (0, 1) m. The same source configuration was used for the listening experiment in Sect. 4. The predicted results fit very well with the results from the experiment, as already shown in Sect. 5. For a spacing of 0.20 m, a large region with no deviations can be seen across the listening area. Only towards the edges of the loudspeaker array are large deviations between intended and predicted directions visible. For loudspeaker arrays with fewer loudspeakers, the deviations of the direction are spread across the listening area, but are worse in the close to the loudspeakers. This is obviously a general trend for all arrays. The larger the distance of the listener to the array in y -direction, the smaller is the intended direction from the predicted perceived one.

In Fig. 11b, the sound event is a plane wave impinging parallel to the y -direction onto the listener. The pattern of results is similar to the one for the point source, but the deviations are larger for the case of only three loudspeakers. This is mainly due to the fact that the auditory event is bound towards the single loudspeaker which, in the case of a plane wave, leads to larger deviations in the whole listening area. For the loudspeaker array with 15 loudspeakers, the deviation-free region is slightly smaller as for the point source. Deviations due to the edges of the array are more visible.

Circular Loudspeaker Array

In addition to the linear loudspeaker arrays used in the listening experiment, localization maps were derived for circular arrays with a geometry similar to the one available in the authors' laboratory at the TU Berlin. Three configurations were considered, consisting of 14, 28, or 56 loudspeakers. These numbers correspond to an inter-loudspeaker spacing of 0.67, 0.34, and 0.17 m, respectively. All three configurations have a diameter of 3 m. Again, a point source located 1 m behind the array, and a plane wave traveling parallel to the y -direction were used as virtual sources.

The results are shown in Fig. 12. They are very similar to the case of a linear loudspeaker array. For the plane wave, the deviations increase with increasing listener distance to the loudspeaker array, as was also observed for the linear array, but to a smaller degree.

For listener positions in the near-field of the loudspeakers, the predicted direction of the auditory event deviates in most cases toward the direction of the corresponding loudspeaker. This seems to be a plausible result, but it should be mentioned that the model used in the current study is not optimally prepared for the near-field case. Particularly, a HRTF dataset with a distance of 3 m between source and dummy head has been used. It is well known from literature that for distances under 1 m, the interaural level differences, ILDs, vary with distance [5, 32]. Hence, the model predictions could probably be enhanced for the small-distance cases by using appropriate HRTF datasets.

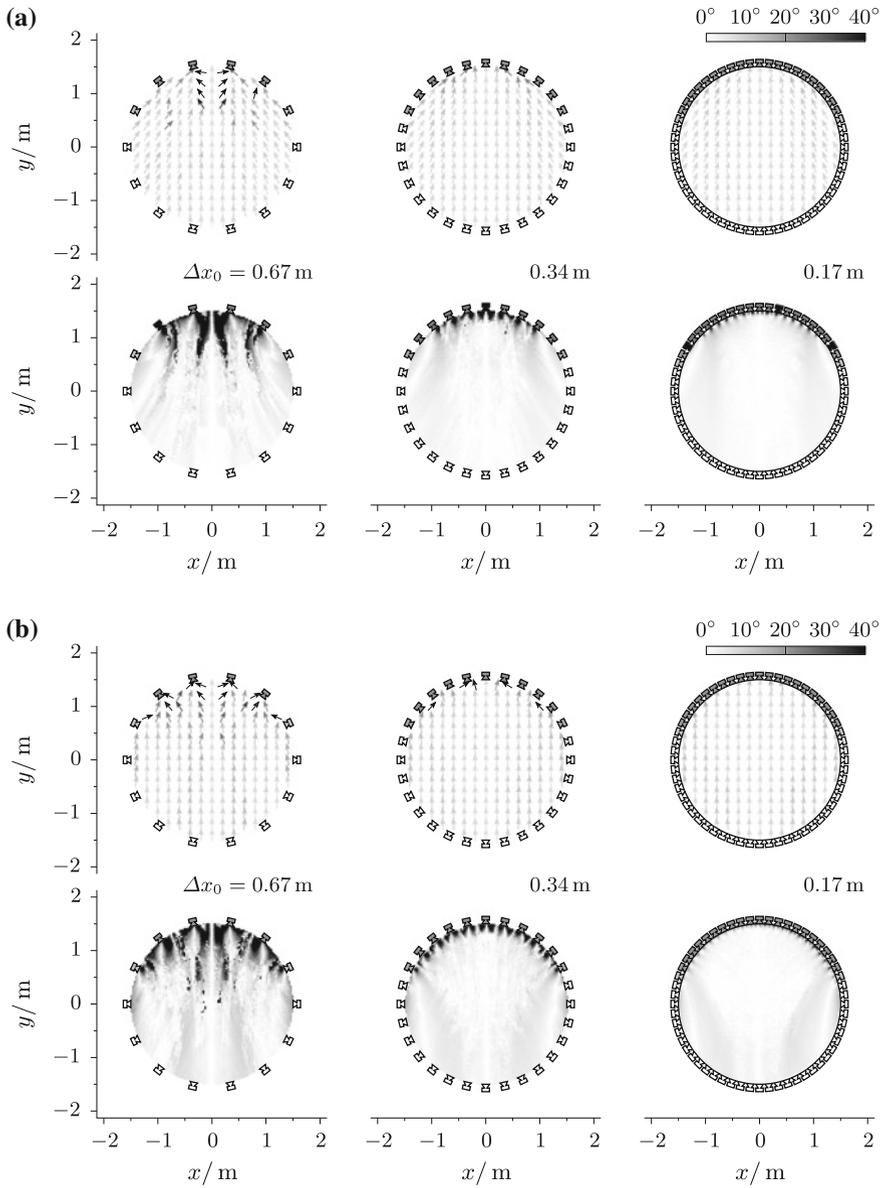


Fig. 12 Localization maps for a circular loudspeaker array driven by WFS for (a) a virtual point source, (b) a plane wave. The *arrows* point into the direction of where the auditory event of a listener appears, if he/she sits at the position of the *arrow*. The *gray-shades* indicate the deviation from the intended direction

6 Conclusion

In sound-field synthesis such as *wave-field synthesis*, WFS, it is of great interest to evaluate perceptual dimensions such as localization and/or coloration not only at one listener position but also for an extended listening area. As concerns localization, this chapter has provided relevant results along these lines by employing binaural synthesis. This method was applied to generate the ear signals at each listener position for headphone presentation. This approach allows further to feed these ear signals into an auditory model which then predicts the localization at all simulated listening positions. To achieve this, the binaural model by Dietz has been extended by a stage that transforms the interaural time differences provided by the model into azimuths corresponding to the sound-source directions. By combining these predicted angular positions with the binaural simulations, localization maps for different loudspeaker setups and WFS configurations were predicted. With an accompanying listening test, the model results for linear loudspeaker arrays were verified. The results showed that the localization in WFS is not distorted as long as the inter-loudspeaker spacing is below 0.2 m. For larger spacings, small deviations between the intended and perceived source locations occur. In practice, one has to specify the localization accuracy that is needed for the intended application of a given WFS system. The predicted localization maps are a valuable aid when planning the task-required loudspeaker setup. However, for practical applications of WFS, it is not only the localization accuracy which is important. WFS may also be affected by the localization blur, for example, indicated by the standard deviation of the localization. In order to investigate the localization blur via binaural synthesis, one has to account for the localization blur as already contributed by binaural synthesis [35]. Further, beside these spatial-fidelity features, coloration or timbral fidelity is of high relevance, as reported by Rumsey [24], who found by comparison of different stereophonic 5.1 surround setups that the overall quality is composed of timbral and spatial fidelity, whereby, according to Rumsey, timbral fidelity explained approximately 70 % of the variance the overall quality ratings while spatial fidelity explained only 30 %. Hence, to provide further components of a model of integral WFS quality, ongoing work by the authors addresses the prediction of coloration resulting from different WFS system and listener configurations.

Materials

The algorithm of the binaural model is included in the *AMToolbox* described in [26], this volume. The function for the prediction of the direction of the auditory event is `estimate_azimuth`. In addition, all other software tools and data are also available as open source items. The *Sound Field Synthesis Toolbox* [34], which was used to generate the binaural simulation for WFS, can be downloaded from <https://dev.qu.tu-berlin.de/projects/sfs-toolbox/files>. The version used in this chapter is 0.2.1. The HRTF data set [32] is part of a larger set available for down-

load from <https://dev.qu.tu-berlin.de/projects/measurements/wiki/2010-11-kemar-anechoic>. The set that has been used here is the one with a distance of 3 m. The *SoundScape Renderer* [12] that was employed as the convolution engine for the dynamic binaural synthesis is available as open source as well. It can be downloaded from <https://dev.qu.tu-berlin.de/projects/ssr/files>.

Acknowledgement

The authors are obliged to M. Geier, who is the driving force behind the development of the SoundScape Renderer, and who helped with the setup of the experiments. Thanks are also due to two anonymous reviewers for their valuable input. This work was funded by DFG–RA 2044/1-1.

References

1. A. Berkhout. A holographic approach to acoustic control. *J. Audio Eng. Soc.*, 36:977–995, 1988.
2. J. Blauert. *Spatial Hearing*. The MIT Press, 1997.
3. J. Blauert and U. Jekosch. Concepts Behind Sound Quality: Some Basic Considerations. In *Proc. 32nd Intl. Congr. Expos. Noise, Control*, 2003.
4. M. Born, E. Wolf, A. Bhatia, P. Clemmow, D. Gabor, A. Stokes, A. Taylor, P. Wayman, and W. Wilcock. *Principles of Optics*. Cambridge University Press, 1999.
5. D. S. Brungart and W. M. Rabinowitz. Auditory localization of nearby sources. Head-related transfer functions. *J. Acoust. Soc. Am.*, 106:1465–79, 1999.
6. S. Choisel and F. Wickelmaier. Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference. *J. Acoust. Soc. Am.*, 121:388, 2007.
7. D. Colton and R. Kress. *Integral Equation Methods in Scattering Theory*. Wiley, New York, 1983.
8. M. Dietz, S. D. Ewert, and V. Hohmann. Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Commun.*, 53:592–605, 2011.
9. T. du Moncel. The Telephone at the Paris Opera. *Scientific America*, pp. 422–23, 1881.
10. F. M. Fazi. *Sound Field Reproduction*. PhD thesis, University of Southampton, 2010.
11. F. M. Fazi, P. A. Nelson, and R. Pothast. Analogies and differences between three methods for sound field reproduction. In *Proc. Ambis. Sym.*, 2009.
12. M. Geier, J. Ahrens, and S. Spors. The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods. In *Proc. 124th Conv. Audio Eng. Soc.*, 2008.
13. M. Geier, H. Wierstorf, J. Ahrens, I. Wechsung, A. Raake, and S. Spors. Perceptual Evaluation of Focused Sources in Wave Field Synthesis. In *Proc. 128th Conv. Audio Eng. Soc.*, 2010.
14. C. Guastavino and B. F. G. Katz. Perceptual evaluation of multi-dimensional spatial audio reproduction. *J. Acoust. Soc. Am.*, 116(2):1105, 2004.
15. C. Huygens. *Treatise on Light (English translation by S. P. Thompson)*. Macmillan & Co, London, 1912.
16. A. Kohlrausch, J. Braasch, D. Kolossa, and J. Blauert. An introduction to binaural processing. In J. Blauert, editor, *The technology of binaural listening*, chapter 1. Springer, Berlin-Heidelberg-New York NY, 2013.
17. T. Letowski. Sound Quality Assessment: Concepts and Criteria. In *Proc. 89th Conv. Audio Eng. Soc.*, 1989.

18. J. Lewald, G. J. Dörrscheidt, and W. H. Ehrenstein. Sound localization with eccentric head position. *Behav. Brain Res.*, 108(2):105–25, 2000.
19. A. Lindau, J. Estrella, and S. Weinzierl. Individualization of dynamic binaural synthesis by real time manipulation of the ITD. In *Proc. 128th Conv. Audio Eng. Soc.*, 2010.
20. R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman. The precedence effect. *J. Acoust. Soc. Am.*, 106:1633–54, 1999.
21. P. Majdak, B. Laback, M. Goupell, and M. Mihocic. The Accuracy of Localizing Virtual Sound Sources: Effects of Pointing Method and Visual Environment. In *Proc. 124th Conv. Audio Eng. Soc.*, 2008.
22. S. Merchel and S. Groth. Adaptively Adjusting the Stereophonic Sweet Spot to the Listeners Position. *J. Audio Eng. Soc.*, 58:809–817, 2010.
23. F. Rumsey. Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm. *J. Audio Eng. Soc.*, 50:651–666, 2002.
24. F. Rumsey, S. Zielinski, R. Kassier, and S. Bech. On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality. *J. Acoust. Soc. Am.*, 118:968–976, 2005.
25. B. U. Seeber. *Untersuchung der auditiven Lokalisation mit einer Lichtzeigermethode*. PhD thesis, 2003.
26. P. Søndergaard and P. Majdak. The auditory modeling toolbox. In J. Blauert, editor, *The technology of binaural listening*, chapter 2. Springer, Berlin-Heidelberg-New York NY, 2013.
27. S. Spors and J. Ahrens. Spatial Sampling Artifacts of Wave Field Synthesis for the Reproduction of Virtual Point Sources. In *Proc. 126th Conv. Audio Eng. Soc.*, 2009.
28. E. Start. *Direct Sound Enhancement by Wave Field Synthesis*. PhD thesis, Technische Universiteit Delft, 1997.
29. J. Steinberg and W. B. Snow. Symposium on wire transmission of symphonic music and its reproduction in auditory perspective: Physical Factors. *AT&T Tech. J.*, 13:245–258, 1934.
30. E. Verheijen. *Sound Reproduction by Wave Field Synthesis*. PhD thesis, Technische Universiteit Delft, 1997.
31. P. Vogel. *Application of Wave Field Synthesis in Room Acoustics*. PhD thesis, Technische Universiteit Delft, 1993.
32. 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 *Proc. 130th Conv. Audio Eng. Soc.*, 2011.
33. H. Wierstorf, A. Raake, and S. Spors. Localization of a virtual point source within the listening area for Wave Field Synthesis. In *Proc. 133rd Conv. Audio Eng. Soc.*, 2012.
34. H. Wierstorf and S. Spors. Sound Field Synthesis Toolbox. In *Proc. 132nd Conv. Audio Eng. Soc.*, 2012.
35. H. Wierstorf, S. Spors, and A. Raake. Perception and evaluation of sound fields. In *Proc. 59th Open Sem. Acoust.*, 2012.
36. F. L. Wightman and D. J. Kistler. The dominant role of low-frequency interaural time differences in sound localization. *J. Acoust. Soc. Am.*, 91(3):1648–61, 1992.
37. E. G. Williams. *Fourier Acoustics*. Academic Press, San Diego, 1999.
38. H. Wittek. *Perceptual differences between wavefield synthesis and stereophony*. PhD thesis, University of Surrey, 2007.
39. N. Zacharov and K. Koivuniemi. Audio descriptive analysis & mapping of spatial sound displays. In *Proc. 7th Intl. Conf. Audit. Display*, pages 95–104, 2001.