PERCEPTUAL ASSESSMENT OF SOUND FIELD SYNTHESIS

vorgelegt von
Dipl.-Phys.
HAGEN WIERSTORF
geb. in Rotenburg (Wümme)

von der Fakultät IV – Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
– Dr. rer. nat. –

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr.-Ing. Sebastian Möller
Gutachter:	Prof. Dr.-Ing. Alexander Raake
Gutachter:	Prof. Dr.-Ing. Sascha Spors
Gutachter:	Prof. Dr. Steven van de Par

Tag der wissenschaftlichen Aussprache: 23. September 2014

Berlin 2014
D 83

Zusammenfassung

Die vorliegende Arbeit untersucht die beiden Schallfeldsyntheseverfahren Wellenfeldsynthese und Nahfeld-entzerrtes Ambisonics höherer Ordnung. Sie fasst die Theorie der beiden Verfahren zusammen und stellt eine Software-Umgebung zur Verfügung, um beide Verfahren numerisch zu simulieren. Diskutiert werden mögliche Abweichungen der mit realen Lautsprechergruppen synthetisierten Schallfelder. Dies geschieht sowohl auf theoretischer Basis als auch in einer Reihe von psychoakustischen Experimenten. Die Experimente untersuchen dabei die räumliche und klangliche Treue und zeitlich-spektrale Artefakte der verwendeten Systeme. Systematisch wird dies für eine große Anzahl von verschiedenen Lautsprechergruppen angewendet. Die Experimente werden mit Hilfe von dynamischer binauraler Synthese durchgeführt, damit auch Lautsprechergruppen mit einem Abstand von unter 1 cm zwischen den Lautsprechern untersucht werden können. Die Ergebnisse zeigen, dass räumliche Treue bereits mit einem Lautsprecherabstand von 20 cm erzielt werden kann, während klangliche Treue nur mit Abständen kleiner als 1 cm möglich ist. Zeitlich-spektrale Artefakte treten nur bei der Synthese von fokussierten Quellen auf. Am Ende wird ein binaurales Modell präsentiert, welches in der Lage ist die räumliche Treue für beliebige Lautsprechergruppen vorherzusagen.

Abstract

This thesis investigates the two sound field synthesis methods Wave Field Synthesis and near-field compensated higher order Ambisonics. It summarizes their theory and provides a software toolkit for corresponding numerical simulations. Possible deviations of the synthesized sound field for real loudspeaker arrays and their perceptual relevance are discussed. This is done firstly based on theoretical considerations, and then addressed in several psychoacoustic experiments. These experiments investigate the spatial and timbral fidelity and spectro-temporal-artifacts in a systematic way for a large number of different loudspeaker setups. The experiments are conducted with the help of dynamic binaural synthesis in order to simulate loudspeaker setups with an inter-loudspeaker spacing of under 1 cm. The results show that spatial fidelity can already be achieved with setups having an inter-loudspeaker spacing of 20 cm, whereas timbral fidelity is only possible for setups employing a spacing below 1 cm. Spectro-temporal artifacts are relevant only for the synthesis of focused sources. At the end of the thesis, a binaural auditory model is presented that is able to predict the spatial fidelity for any given loudspeaker setup.

Contents

1
Introduction

LISTENING to music plays an important role in the social and cultural life of human beings. Musical instruments found in archaeological excavations can be dated back as far as 40 000 years in the past.[1] There is a variety of different types of music and the time spent for listening to music is still increasing. One of the preconditions for this increase was the invention of the first electroacoustical transducer – the telephone in 1876. After that it was possible to listen to music without the presence of a musician. The digitalisation and vast availability of music in the last years made it even easier to listen to music.[2]

Due to the importance of communication and music in our everyday life, the electroacoustical presentation of sound has advanced quickly after the invention of the telephone. It was noted that when all sound was presented only by a single transducer the spatial impression of a recorded sound – e.g. an orchestra – is lost. It would enhance the listening experience if the spatial impression of the sound could be recreated during the presentation. This inspired Steinberg and Snow[3] to their idea of recreating a whole sound field in the audience area. But they noted already that they would basically need an infinite number of receivers and transmitters to do this. In an experiment they arranged a setup consisting of only two microphones and two loudspeakers and showed that this low number of loudspeakers "give(s) good auditory perspective".

The ideas and results from Steinberg's and Snow's paper are still part of the main research topics in spatial audio. What is the best way to give a good spatial impression of the presented sound? How many loudspeakers are needed to do this? Is it possible to create a spatial extended sound field in a convincing way? What is the influence of the spatial impression on the overall quality a listener experiences while listening to the played back sound?

The goal of this thesis is to investigate these questions with a focus on sound field synthesis (SFS) techniques and their perception. The main question targets the analysis of which parts of an imperfectly reproduced sound field are perceived as imperfect by the listener and which parts are not. As the work in the field of audio coding has demonstrated, the listener may be insensitive to a large amount of "errors" in a sound signal. In addition, a link to the underlying

[1] For an overview see F. D'Errico et al. "Archaeological Evidence for the Emergence of Language, Symbolism, and Music – An Alternative Multidisciplinary Perspective". *Journal of World Prehistory* 17.1 (2003), pp. 1–70.

[2] J. Sloboda, A. Lamont, and A. Greasly. "Choosing to hear music". In: *The Oxford Handbook of Music Psychology*. Ed. by S. Hallam, I. Cross, and M. Thaut. New York: Oxford University Press, 2009, pp. 431–40.

[3] J. Steinberg and W. B. Snow. "Symposium on wire transmission of symphonic music and its reproduction in auditory perspective: Physical Factors". *Bell System Technical Journal* 13.2 (1934), pp. 245–58.

technical parameters that cause the imperfection of the sound field will be established in order to control the amount of errors in a systematic way. This work is a foundation for the investigation of the larger question of how these systems influence the quality experienced by the listener.

For a more detailed understanding of the research questions the basic principles of the auditory system that contribute to the perception of a spatial sound scene are discussed in this introduction chapter. In addition, a short overview of different spatial sound reproduction techniques is provided, followed by a theoretical framework to establish how to talk about quality of spatial audio and how to investigate it.

In the second chapter the mathematical framework of the different methods for creating a spatial extended sound field is presented and the formulas for the calculations of loudspeaker signals are provided. The methods are restricted to analytical solutions of the integral equation that describes the sound field synthesis problem – namely Wave Field Synthesis (WFS) and Near-Field Compensated Higher Order Ambisonics (NFC-HOA). The presentation of the framework is motivated by the idea of highlighting that both methods have the same foundation and are comparable in several ways.[4] This deviates from classical research papers in the field of WFS and NFC-HOA that employ different mathematical frameworks due to their historical independence.

The limitations of today's hardware in practical implementations – distance between and number of loudspeakers – lead to several deviations in the created sound field from the desired one. The possible implications of these deviations and connections to the underlying hardware and mathematical parameters are discussed in Chapter 3. Considering the functioning of the human auditory system, hypotheses are formulated about the influences of the deviations on their auditory perception. The hypotheses point directly to the research questions that will be dealt with in different psychoacoustic experiments as presented in Chapter 5.

Before that, Chapter 4 discusses the method applied for the psychoacoustic experiments. The challenge of psychoacoustic experiments for spatial audio is the dependence of the results on the position of the listener. It is not only of interest how the perception is at a particular point, but how it is in the whole listening area. Another challenge is to systematically investigate the influence of the technical parameters of the underlying presentation methods on perception. For sound field synthesis methods the number of and distance between the loudspeakers are especially critical parameters and should be adjustable in a wide range from two up to several thousand loudspeakers. Simulating all systems with binaural synthesis is a solution to these problems. The approach can evoke errors, however, and a large part of the fourth chapter investigates if binaural synthesis is suitable as a tool for answering the psychoacoustic

[4] This idea is developed in more detail in J. Ahrens. *Analytic Methods of Sound Field Synthesis*. New York: Springer, 2012.

research questions.

After assessing the influence of the technical parameters on the perception for some setups it is of interest to predict the perception of the listener to investigate every possible setup. Chapter 6 shows this for the case of localization for WFS and NFC-HOA. A binaural auditory model is fed with the signals reaching the ears of the listener and is shown to be able to predict the perceived direction for a given source in the sound field.

The last chapter summarizes the results and discusses the implications of the results for further investigations of the quality of spatial audio systems.

1.1 The Human Auditory System

For human beings hearing means that sound signals arriving at the eardrums of both ears are processed by different mechanisms to finally lead to a corresponding perception – as illustrated by Figure 1.1.

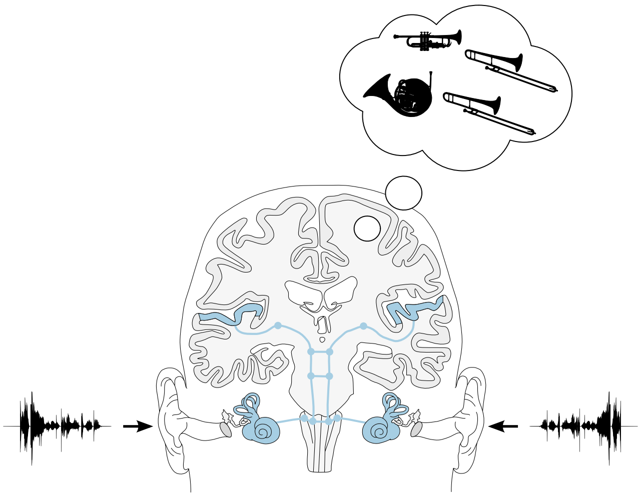


Figure 1.1: The human auditory system. The two ear signals are processed by the outer, middle, and inner ear. In the inner ear they are transformed into neuronal signals which are then analyzed at different interlinked stations in the brain. The interlinking of both ear signals already happens at a low level in the brainstem. In the brain itself a representation of the external ear signals in the form of an auditory scene is the final stage. This figure is based on B. Grothe, M. Pecka, and D. McAlpine. "Mechanisms of Sound Localization in Mammals". *Physiological Reviews* 90 (2010), pp. 983–1012; K. Talbot et al. "Synaptic dysbindin-1 reductions in schizophrenia occur in an isoform-specific manner indicating their subsynaptic location." *PLoS ONE* 6.3 (2011), e16886; L. Chittka and A. Brockmann. "Perception space–the final frontier." *PLoS Biology* 3.4 (2005), e137. ☞

In the discussion of correspondences between physical sound objects and the corresponding perceptional objects the following terms are used. As long as the physical sound signals outside of the listener are considered, the terms *sound event(s)* and *sound scene(s)* describe what is presented to the listener. A sound event corresponds to the signal emitted by a physical sound source such as a human speaker or a loudspeaker. A sound scene is a composition of different sound events. The number of sound events originally involved in the creation of these signals cannot be estimated for all situations, if only the two signals arriving at the ears of the listener are considered. This fact is utilized by sound field synthesis techniques: to generate the sound field corresponding to a single sound event with the help of a very large number of individual sound events.[5]

The superposed signals are transformed into a movement of the eardrums, transformed into neural activity in the inner ear and further interpreted by the brain. The final perceptual output is a com-

[5] This is better known as the Huygens-Fresnel principle, see e.g. C. Huygens. *Treatise on Light*. Ed. by S. P. Thompson. London: Macmillan & Co, 1912

plex *auditory scene* composed of single *auditory events*. Listeners can distinguish single auditory events and steer their auditory attention to one of the events. Furthermore, they are associated with different perceptual attributes such as loudness, pitch, duration, timbre, and spatial features.[6] Together, the different features form the aural character of an auditory event or scene.

As such, the process of hearing generates an auditory scene from the presented sound scene. The extraction of single auditory events from the eardrum signals is referred to as auditory scene analysis. It is obvious that there is no linear function that is able to describe the mapping of sound events to auditory events, because the number of sound and auditory events does not have to match. For example, imagine the following setup: two loudspeakers are placed in an anechoic environment and play exactly the same signal to a listener placed in the middle of the loudspeakers at a distance of a few meters. In that case, the listener perceives only one auditory event in the center of the two loudspeakers, a phenomenon applied in stereophony. Assume now that the same equipment is placed in an echoic room. The room is by itself not a new sound source, but adds reflective elements to the sound scene. Listener perceive the room as additional features to the initial auditory event.

Note that an auditory event is on a lower abstraction level than an *auditory object* that requires even more processing of the brain – including multi-modal processing.[7] For the topics investigated in this thesis it is sufficient to limit the considerations to auditory events.

After introducing the terminology for describing auditory perception, the processing of the sound signals by the auditory system is discussed in more detail. In a first step the sound field created by all sound events is filtered by the outer ears of the listener. Thereby, its frequency content is modified depending on the incidence direction. This provides a relevant cue for the perception of the vertical direction of a sound source.[8] In the ear canal, all signals from the sound events are superimposed and excite the eardrums. The oscillation of the eardrums is amplified by the auditory ossicles in the middle ear and coupled to the inner ear. At this point the mechanical signal is dispersed and converted into neuronal signals. The neuronal signals are able to preserve the temporal information of the sound signals up to a frequency of about 1.5 kHz, for higher frequencies only the temporal information of the envelope can be extracted. The information on higher frequencies is not completely lost in this process because the dispersion of the mechanical signal in the inner ear distributes the signal energy regarding its frequency content and performs a frequency-place-transformation comparable to a Fourier transformation. The neuronal signals are transmitted to different places in the brain starting with the *cochlear nucleus*, the *superior olivary complex*, and the *lateral lemniscus* in the brainstem going further to the *inferiror colliculus* in the midbrain and the *medial geniculate body* in the thalamus before reaching the *primary auditory cortex*.[9]

[6] T. R. Letowski. "Sound quality assessment: concepts and criteria". In: *87th Audio Engineering Society Convention*. 1989, Paper 2825.

[7] A discussion of auditory objects is presented in M. Kubovy and D. V. Van Valkenburg. "Auditory and visual objects." *Cognition* 80.1-2 (2001), pp. 97–126.

[8] See p. 97ff in J. Blauert. *Spatial Hearing*. The MIT Press, 1997.

[9] E.g. B. Grothe, M. Pecka, and D. McAlpine. "Mechanisms of Sound Localization in Mammals". *Physiological Reviews* 90 (2010), pp. 983–1012

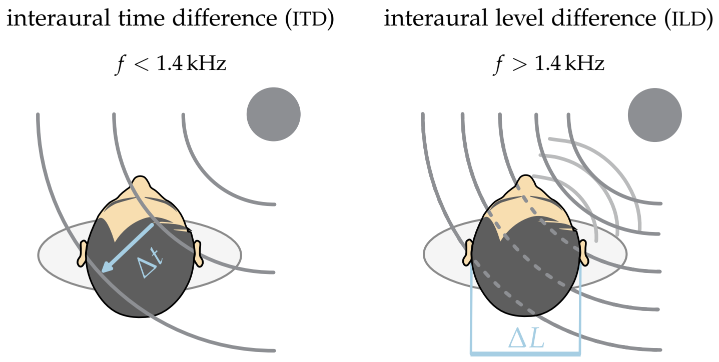


Figure 1.2: Interaural differences occur between both ear signals for sources to the side of the listener. For low frequencies ITDs are the dominant cue, for high frequencies ILDs are more reliable. The figure is based on B. Grothe, M. Pecka, and D. McAlpine. "Mechanisms of Sound Localization in Mammals". *Physiological Reviews* 90 (2010), pp. 983–1012. ☞

An interaction between the neuronal signals of both ears appears already at a very low level, namely the *superior olivary complex*. At this point differences between the two signals can be analyzed that provide evidence for the direction of an auditory event in the horizontal plane. This process of attributing a direction in the horizontal plane to an auditory event is called *localization* in this thesis. Usually localization also includes the vertical direction and the distance of the auditory event, but these aspects are not considered in this work. The two main differences that allow a calculation of a direction of incidence are the interaural time difference (ITD) and the interaural level difference (ILD) of the two ear signals as indicated by Figure 1.2. The ITD exploits the fact that the time of arrival at each ear depends on the direction of the sound. This requires a high temporal accuracy of the auditory system, which is only provided for frequencies up to 1.4 kHz. The frequency limit corresponds roughly to the diameter of a human head that is another natural limit for the usage of the ITD. For higher frequencies more than one wave length of the sound waves lies between the two ears. Thus the ITD becomes an ambiguous cue. For ILD it is the opposite. ILD results from the fact that the human head is not transparent for sound waves and scatters them. That leads to a difference in sound pressure level depending on the incidence angle of the sound. For frequencies below 1.4 kHz the influence of the head can be neglected, whereas for higher frequencies the level differences are relevant. Sound components with wave lengths similar to the head diameter are diffracted around the head and reach the opposite ear. This has motivated the duplex theory[10] of localization, assuming that the ITD cues are used for low frequencies and the ILD cues for high frequencies. How the different cues are combined to form a directional perception is presented in Section 5.1 and 6.1.

The goal of every sound field synthesis technique is to provide the same spatial cues as in the original sound field, as opposed to the sweet-spot phenomenon of stereophony that is explained in the next section. In order to achieve this, highly correlated signals are presented to different loudspeakers. This scenario is roughly equivalent to delaying and summing up the same signal. Thereby a comp-filter

[10] J. W. Strutt. "On our perception of sound direction". *Philosophical Magazine* 13.74 (1907), pp. 214–32.

like amplitude spectrum appears that can introduce severe timbral changes. Besides, it is possible that the creation of completely unnatural signals with SFS lead to the perception of additional spectro-temporal artifacts.

This thesis concentrates on a subset of attributes of auditory events that are relevant for assessing sound field synthesis methods, namely timbre, direction, and technical artifacts. The different attributes will be discussed in detail in Chapter 5, which investigates also the influence of several physical parameters of SFS on those attributes.

In the next section, a short overview of different spatial sound presentation techniques, including stereophony and SFS is presented.

1.2 *Spatial Sound Presentation Techniques*

In this thesis spatial sound presentation refers to all methods that try to recreate some of the spatial aspects of a given sound scene by applying more than one loudspeaker. The first ever practical attempt of spatial sound reproduction dates back to 1881, only five years after the invention of the first monaural transducer. Back then, two parallel telephone channels were used to transmit recorded music to the homes of the listeners.[11] The basic idea was the ability to influence the interaural differences between the two ear signals of the listener. That was achieved by recording a sound scene with two microphones placed at different positions and feeding the recorded signals to the two telephone channels.

Later on the idea advanced to the technique of *binaural presentation* where the basic principle is to recreate the ear signals at both ears as they would appear in reality. This can be achieved by placing two microphones in the ears of the listener for recording and playing the recorded signals back via headphones afterwards. Binaural presentation has advanced in the last decades by measuring the acoustical transmission paths between a source and the ears of a listener, so called head-related transfer functions (HRTFs). Afterwards these can be used to create any sound scene as long as the required HRTFs are available. Due to the good control of the ear signals that arrive at the listener's ears this method is very accurate in creating directional cues and will be applied in this thesis for simulating the ear signals for different loudspeaker setups.

Spatial sound presentation via loudspeakers started in the 1930s, the time when Blumlein invented the stereophonic recording and reproduction[12] and Steinberg and Snow discussed the idea of the acoustical curtain.[13] The original idea of the latter was to create a sound field that mimics the real sound scene. Their practical implementation with two or three loudspeakers was not able to achieve this. With such low numbers of loudspeakers the sound field is only controllable at single points in space. This corresponds to the classical stereophonic setup consisting of a fixed listener position between the two loudspeakers at a distance of around 2 m. The human head has a diameter of around 20 cm and hence only one ear can be placed

[11] T. du Moncel. "The international exhibition and congress of electricity at Paris". *Nature* October 20 (1881), pp. 585–89.

[12] A. D. Blumlein. "Improvements in and relating to Sound-transmission, Sound-recording and Sound-reproducing Systems". *Journal of the Audio Engineering Society* 6.2 (1958), pp. 91–98, 130

[13] Steinberg and Snow, op. cit.

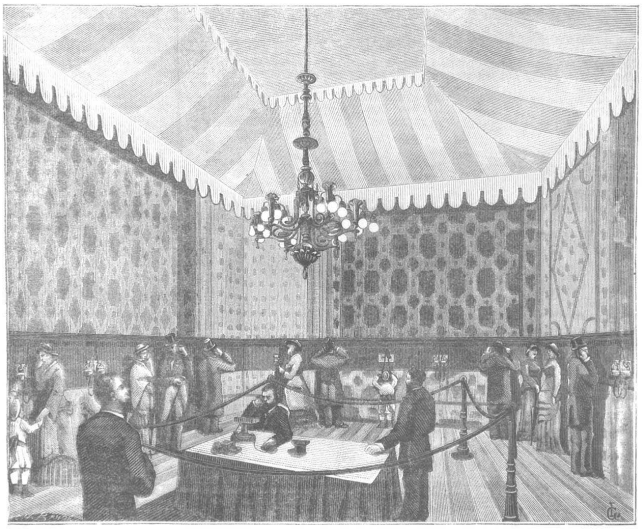


Figure 1.3: Setup of stereophonic telephones at the exhibition 1881 in Paris. Figure from T. du Moncel. "The international exhibition and congress of electricity at Paris". *Nature* October 20 (1881), pp. 585–89 ☞

at the point where the sound field is as desired. But as Steinberg and Snow discovered for their acoustic curtain, the spatial perception of the listener is not disturbed as long as she does not move too far away from a line on which every point has the same distance to both loudspeakers. By staying on that line the listener perceives an auditory event in the center of both loudspeakers, if the same acoustical signal is played through them. If the amplitude of one of the loudspeakers is changed the auditory event is moved between the two speakers. The connection of the amplitude difference between the loudspeakers and the actual position of the auditory event is empirical and is described by so called panning laws.[14] If the listener leaves the central line, the position of the auditory event will always be located at the position of one of the two loudspeakers. The area in which the spatial perception of the auditory scene works without considerable impairments is called the *sweet-spot* of a given loudspeaker setup. It is indicated by the blue color in Figure 1.4. To explain why the spatial perception of the listener is correct at the sweet-spot although the sound field is not, the theory of *summing localization* was introduced by Warncke in 1941.[15]

In the last years the stereophonic setup was expanded to 5.0 surround and even larger setups[16] and the panning laws were formulated in a more general way dealing with setups using multiple loudspeakers.[17] These approaches could not fix the sweet-spot problem, but added a richer spatial impression because sound is no longer restricted to come from the front.

Before 5.0 surround there were other approaches to enhance the spatial impression of stereophony. From the 1970s onwards quadrophony and Ambisonics[18] were developed in order to provide a surround experience with four loudspeakers. The basic idea of Ambisonics is comparable to nowadays NFC-HOA for a larger number of loudspeakers: to describe an extended sound field by spherical basis functions that can be synthesized by any spherical or circular

[14] D. M. Leakey. "Some Measurements on the Effects of Interchannel Intensity and Time Differences in Two Channel Sound Systems". *The Journal of the Acoustical Society of America* 31.7 (1959), pp. 977–86

[15] A discussion is provided in J. Blauert. *Spatial Hearing*. The MIT Press, 1997, p. 204

[16] E.g. K. Hamasaki, K. Hiyama, and H. Okumura. "The 22.2 Multichannel Sound System and Its Application". In: *118th Audio Engineering Society Convention*. 2005, Paper 6406

[17] V. Pulkki. "Virtual Sound Source Positioning Using Vector Base Amplitude Panning". *Journal of the Audio Engineering Society* 45.6 (1997), pp. 456–66

[18] M. A. Gerzon. "Periphony: With-Height Sound Reproduction". *Journal of the Audio Engineering Society* 21.1 (1973), pp. 2–10.

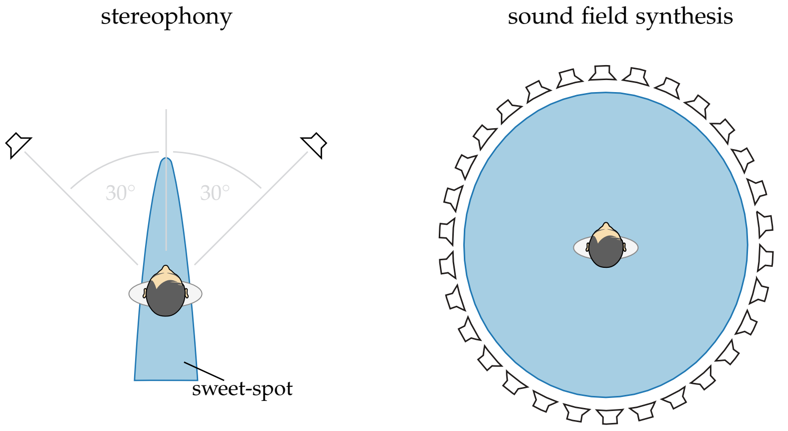


Figure 1.4: Loudspeaker setups for two channel stereophony and sound field synthesis. The area marked in blue describes the positions were the listener can move to and still perceives the same spatial impression. This area is smaller for stereophonic setups and is called the sweet-spot. The figure of the stereophony setup is a modified version of J. Ahrens. *Analytic Methods of Sound Field Synthesis*. New York: Springer, 2012, Fig. 1.1. ☞

loudspeaker setup. In practice, the restriction of the limited number of loudspeakers has led to the usage of only two spherical basis functions. The results are loudspeaker signals that are comparable to the case of panning in stereophony with the difference of more active loudspeakers.[19] If more than four loudspeakers and more than two basis functions are applied the term is changed to Higher Order Ambisonics (HOA) to highlight this fact. For the perceptual side of Ambisonics the sweet-spot problem exists as well. The explanation of this sweet-spot is only partly covered by the theory of summing localization, because that theory is not well investigated for several sound sources coming from all directions. This provoked a high number of different optimizations of the loudspeaker signals by the Ambisonics community.

All of the methods described so far are able to provide a convincing spatial impression at a specific listener position within the loudspeaker setup. That means none of them can handle an equally good spatial impression for a bigger audience.

In the late 1980s the old idea of Steinberg and Snow to reproduce a complete sound field came to new life due to the fact that now arrangements of more than 100 loudspeakers became possible.[20] This high number of loudspeakers is needed: for controlling an extended sound field up to 20 kHz, loudspeaker spacings under 1 cm are required. Small distances like that are not possible in practice. Nonetheless, the experience has shown that even with larger distances reasonable sound field approximations are possible. Some of them provide equal spatial impression in the whole listening area, as indicated by the blue color in Figure 1.4. Methods trying to achieve this goal are summarized under the term sound field synthesis (SFS). This thesis focusses on the two SFS techniques WFS and NFC-HOA that are explained in detail in the next two chapters. The main research goal is to investigate how large the deviations in the synthesized sound field can be without falling back to the sweet-spot phenomenon. Thereby, not only the spatial impression is considered

[19] E.g. M. Frank. "Phantom Sources using Multiple Loudspeakers". PhD thesis. University of Music and Performing Arts Graz, 2013.

[20] A. Berkhout. "A holographic approach to acoustic control". *Journal of the Audio Engineering Society* 36.12 (1988), pp. 977–95.

but also the timbral fidelity of the system and the absence of audible artifacts, meaning all the aspects that contribute to the overall quality perception of the system. The next section discusses the perception of quality in more detail and presents some theoretical considerations for talking about quality in the context of spatial audio presentation.

1.3 *Quality of Spatial Sound*

Listeners are coming to judgements about the *quality* of a presented sound scene by comparing the character of the corresponding auditory scene to the character of a reference.[21] In the case of spatial audio presentation the character of an auditory scene is composed mainly by timbral and spatial features, and by spectro-temporal artifacts introduced by the presentation system.[22] The reference can be explicit by providing a comparison stimulus to the listener. If no explicit reference is presented, the listener compares the auditory scene with her expectations of the character formed by former experiences. In this case, the reference dependents on the individual listeners.

Quality is related to the concepts *authenticity* and *plausibility*. Authenticity deals with the *form-related fidelity* of the auditory character. To test authenticity the listener is asked to judge if the auditory scene is indiscernible from the reference. For example, in the field of audio coding authenticity of the processed sound is the goal, and this is tested by providing the listener with the explicit reference – the unprocessed signal. In the case of spatial audio presentation, experiments are often divided in testing for authenticity of single auditory features independently.[23] The same approach is applied in this thesis by asking for *spatial fidelity* and *timbral fidelity* in different experiments – see Chapter 5. Comparing only single features has the advantage of being able to narrow the reference. For example, the ability to localize a synthesized sound in SFS can be assessed directly by asking the listener for the perceived direction. The test results are then compared to the findings from the literature for real sound sources. Another solution is to work with simplified sound scenes. By looking at the timbral fidelity of a single point source, an explicit reference can easily be generated. On the other hand, by presenting a whole rock concert via SFS to the listener all relevant auditory features would be included and the stimulus would be far more realistic. Nevertheless, the explicit reference can most likely not be presented.

For the presentation of a rock concert the criterion of plausibility seems to be a better concept for the judgement of quality of the sound. Plausibility denotes how believable and credible the correspondence with the listeners expectations is, meaning that plausibility deals only with an implicit reference of the listener.[24] Not only the form of the auditory scene influences plausibility, but to a high degree also its functional aspects of conveying *meaning* to the listener. A good example is the stereophonic presentation of the mentioned rock

[21] J. Blauert and U. Jekosch. "Concepts Behind Sound Quality: Some Basic Considerations". In: *International Congress and Exposition on Noise Control Engineering*. 2003.

[22] F. Rumsey. "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm". *Journal of the Audio Engineering Society* 50.9 (2002), pp. 651–66.

[23] E.g. F. Rumsey et al. "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality". *The Journal of the Acoustical Society of America* 118.2 (2005), pp. 968–76.

[24] A. Raake and J. Blauert. "Comprehensive modeling of the formation process of sound-quality". In: *International Workshop on Quality of Multimedia Experience*. 2013, pp. 76–81.

concert. The sound is recorded, modified and arranged by a mastering engineer, and played back to the audience. Due to the sweet-spot and the fact that the sound can only come from a frontal direction it is very unlikely that the original form of the auditory scene can be preserved. On the other hand, the arrangements of the mastering engineer could enhance the aesthetics of the auditory scene compared to the experience during the live rock concert. If a listener judges the quality of such a presentation she would rely on her internal reference which is formed by her former experiences from concerts and probably even stereophonic presentation techniques in general. A problem related with the plausibility criterion is that it is hard to directly assess in an experiment. Listeners are asked instead to rate their sense of presence or immersion[25] in the auditory scene.

Plausibility is not investigated in this thesis, mainly because of the lack of experiences in mastering sound scenes for SFS systems. That makes it hard to create sound scenes with a similar aesthetic appeal like the ones created for stereophonic presentation. In order to judge plausibility complex auditory scenes are necessary. To rate the degree of immersion for different single point sources does probably not reflect the degree of immersion the different systems are capable of. Hence it has to be considered that even if this thesis shows that SFS systems lead to a better rating regarding authenticity it does not automatically mean that they would be rated as having better quality, too.

1.4 *Reproducible Research*

Like other fields that involve signal processing, the study of SFS implies implementing a multitude of algorithms and running numerical simulations on a computer – compare the figures in Chapter 2 and 3. The same is true for the modeling of the auditory system as it is done in Chapter 6. As a consequence, the outcome of the algorithms are easily vulnerable to implementation errors which cannot completely be avoided.[26]

Beside the software tools, the work presented here relies on measured acoustical data. To ensure that other researchers can test the correctness of results and easily reproduce them, the most straightforward approach is to publish the code together with the measured data. This policy was adapted in the last years by some journals and is known under the term *reproducible research*.[27] It will be adopted for this work as far as possible. All numerical simulations of sound fields or corresponding ear signals are done via the Sound Field Synthesis Toolbox[28] and the modeling of the hearing system via the Auditory Modeling Toolbox[29]. Functions derived in the theoretical chapters that are implemented in one of the toolboxes are accompanied by a link to the corresponding function. All figures in this thesis have a link in the form of ☞ which is a link to a folder containing all the data and scripts in order to reproduce the single figures.

[25] A. Raake and J. Blauert. "Comprehensive modeling of the formation process of sound-quality". In: *International Workshop on Quality of Multimedia Experience*. 2013, pp. 76–81

[26] Compare D. C. Ince, L. Hatton, and J. Graham-Cumming. "The case for open computer programs". *Nature* 482.7386 (2012), pp. 485–88

[27] For one of the pioneers see D. L. Donoho et al. "Reproducible Research in Computational Harmonic Analysis". *Computing in Science & Engineering* 11.1 (2009), pp. 8–18

[28] Sound Field Synthesis Toolbox
version 1.0.0
H. Wierstorf and S. Spors. "Sound Field Synthesis Toolbox". In: *132nd Audio Engineering Society Convention*. 2012, eBrief 50

[29] Auditory Modeling Toolbox
commit aed0198
P. L. Søndergaard and P. Majdak. "The auditory-modeling toolbox". In: *The technology of binaural listening*. Ed. by J. Blauert. New York: Springer, 2013, pp. 33–56

1.5 *Mathematical Definitions*

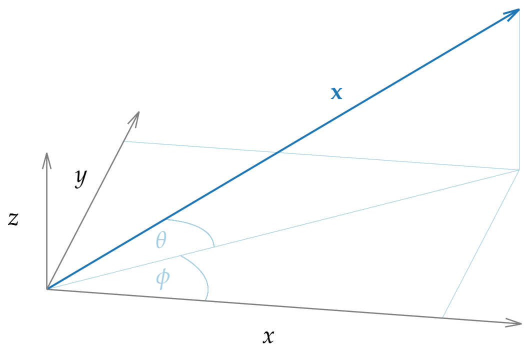


Figure 1.5: Coordinate system used in this thesis. The vector **x** can also be described by its length, its azimuth angle ϕ, and its elevation θ. ☞

Coordinate system Figure 1.5 shows the coordinate system that is used in the following chapters. A vector **x** can be described by its position (x, y, z) in space or by its length, azimuth angle $\phi \in [0, 2\pi[$, and elevation $\theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. The azimuth is measured counterclockwise and elevation is positive for positive z-values.

Fourier transformation Let s be an absolute integrable function, t, ω real numbers, then the temporal Fourier transform is defined as[30]

$$S(\omega) = \mathcal{F}\left\{s(t)\right\} = \int_{-\infty}^{\infty} s(t)\mathrm{e}^{-\mathrm{i}\omega t}\ \mathrm{d}t\ . \tag{1.1}$$

In the same way the inverse temporal Fourier transform is defined as

$$s(t) = \mathcal{F}^{-1}\left\{S(\omega)\right\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)e^{i\omega t}\ d\omega\ . \tag{1.2}$$

[30] R. N. Bracewell. *The Fourier Transform and its Applications*. Boston: McGraw Hill, 2000.

2
Theory of Sound Field Synthesis

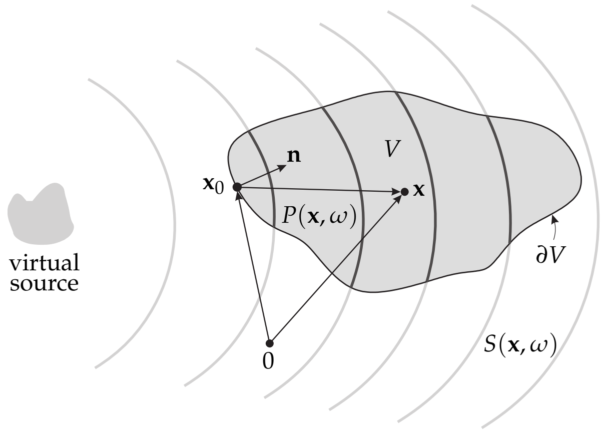


Figure 2.1: Illustration of the geometry used to discuss the physical fundamentals of sound field synthesis and the single-layer potential (2.1). ☞

The problem of sound field synthesis can be formulated as follows.[1] Assume a volume $V \subset \mathbb{R}^n$ which is free of any sources and sinks, surrounded by a distribution of monopole sources on its surface ∂V. The pressure $P(\mathbf{x},\omega)$ at a point $\mathbf{x} \in V$ is then given by the *single-layer potential*

$$P(\mathbf{x},\omega) = \oint_{\partial V} D(\mathbf{x}_0,\omega) G(\mathbf{x}-\mathbf{x}_0,\omega)\, dA(\mathbf{x}_0)\ , \tag{2.1}$$

where $G(\mathbf{x}-\mathbf{x}_0,\omega)$ denotes the sound propagation of the source at location $\mathbf{x}_0 \in \partial V$, and $D(\mathbf{x}_0,\omega)$ its weight, usually referred to as *driving function*. The sources on the surface are called *secondary sources* in sound field synthesis, analogue to the case of acoustical scattering problems. The single-layer potential can be derived from the Kirchhoff-Helmholtz integral.[2] The challenge in sound field synthesis is to solve the integral with respect to $D(\mathbf{x}_0,\omega)$ for a desired sound field $P = S$ in V. It has unique solutions which Zotter and Spors[3] explicitly showed for the spherical case and Fazi[4] for the planar case.

In the following the single-layer potential for different dimensions is discussed. An approach to formulate the desired sound field S is described and finally it is shown how to derive the driving function D.

[1] Small parts of this section are published in H. Wierstorf, A. Raake, and S. Spors. “Binaural assessment of multichannel reproduction”. In: *The technology of binaural listening*. Ed. by J. Blauert. New York: Springer, 2013, pp. 255–78.

[2] E. G. Williams. *Fourier Acoustics*. San Diego: Academic Press, 1999.

[3] F. Zotter and S. Spors. “Is sound field control determined at all frequencies? How is it related to numerical acoustics?” In: *52nd Audio Engineering Society Conference*. 2013, Paper 1.3.

[4] F. M. Fazi. “Sound Field Reproduction”. PhD thesis. University of Southampton, 2010, Chap. 4.3.

2.1 Solution for Special Geometries: Near-Field Compensated Higher Order Ambisonics and Spectral Division Method

The integral equation (2.1) states a Fredholm equation of first kind with a Green's function as kernel. This type of equation can be solved in a straightforward manner for geometries that have a complete set of orthogonal basis functions. Then the involved functions are expanded into the basis functions ψ_n as[5]

$$G(\mathbf{x}-\mathbf{x}_0,\omega) = \sum_{n=1}^{N} \tilde{G}_n(\omega)\, \psi_n^*(\mathbf{x}_0)\psi_n(\mathbf{x}) \tag{2.2}$$

$$D(\mathbf{x}_0,\omega) = \sum_{n=1}^{N} \tilde{D}_n(\omega)\, \psi_n(\mathbf{x}_0) \tag{2.3}$$

$$S(\mathbf{x},\omega) = \sum_{n=1}^{N} \tilde{S}_n(\omega)\, \psi_n(\mathbf{x})\,, \tag{2.4}$$

where $\tilde{G}_n, \tilde{D}_n, \tilde{S}_n$ denote the series expansion coefficients and $\langle \psi_n, \psi_{n'} \rangle = 0$ for $n \neq n'$. Introducing these three equations into (2.1) one gets

$$\tilde{D}_n(\omega) = \frac{\tilde{S}_n(\omega)}{\tilde{G}_n(\omega)}\,. \tag{2.5}$$

This means that the Fredholm equation (2.1) states a convolution. For geometries where the required orthogonal basis functions exist, (2.5) follows directly via the convolution theorem.[6] Due to the division of the desired sound field by the spectrum of the Green's function this kind of approach has been named Spectral Division Method (SDM).[7] For circular and spherical geometries the term Near-Field Compensated Higher Order Ambisonics (NFC-HOA) is more common due to the corresponding basis functions. "Near-field compensated" highlights the usage of point sources as secondary sources in contrast to Ambisonics and Higher Order Ambisonics (HOA) that assume plane waves as secondary sources.

The challenge is to find a set of basis functions for a given geometry. In the following paragraphs three simple geometries and their widely known sets of basis functions will be discussed.

2.1.1 Spherical Geometries

sphharmonics.m
asslegendre.m

The spherical harmonic functions constitute a basis for a spherical secondary source distribution in $\mathbb{R}^3$ and can be defined as[8]

$$Y_n^m(\theta,\phi) = (-1)^m \sqrt{\frac{(2n+1)(n-|m|)!}{4\pi(n+|m|)!}} P_n^{|m|}(\sin\theta)e^{im\phi}\,, \tag{2.6}$$
$$n = 0,1,2,\ldots \qquad m = -n,\ldots,n$$

where $P_n^{|m|}$ are the associated Legendre functions. Note that this function may also be defined in a slightly different way, omitting the $(-1)^m$ factor, see for example Williams.[9]

[5] Compare P. M. Morse and H. Feshbach. *Methods of Theoretical Physics*. Minneapolis: Feshbach Publishing, 1981, p. 940.

[6] Compare G. B. Arfken and H. J. Weber. *Mathematical Methods for Physicists*. Amsterdam: Elsevier, 2005, p. 1013.

[7] J. Ahrens and S. Spors. "Sound Field Reproduction Using Planar and Linear Arrays of Loudspeakers". *IEEE Transactions on Audio, Speech, and Language Processing* 18.8 (2010), pp. 2038–50

[8] N. A. Gumerov and R. Duraiswami. *Fast Multipole Methods for the Helmholtz Equation in Three Dimensions*. Amsterdam: Elsevier, 2004, (12.153), $\sin\theta$ is used here instead of $\cos\theta$ due to the use of another coordinate system, compare Figure 2.1 from Gumerov and Duraiswami and Figure 1.5 in this thesis.

[9] Williams, op. cit., (6.20).

The complex conjugate of Y_n^m is given by negating the degree m as

$$Y_n^m(\theta,\phi)^* = Y_n^{-m}(\theta,\phi)\ . \tag{2.7}$$

For a spherical secondary source distribution with a radius of R_0 the sound field can be calculated by a convolution along the surface. The driving function is then given by a simple division as[10]

$$D_{\text{spherical}}(\theta_0,\phi_0,\omega) = \\ \frac{1}{R_0^2}\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\sqrt{\frac{2n+1}{4\pi}}\frac{\breve{S}_n^m(\theta_\text{s},\phi_\text{s},r_\text{s},\omega)}{\breve{G}_n^0(\frac{\pi}{2},0,\omega)}Y_n^m(\theta_0,\phi_0)\ , \tag{2.8}$$

where $\breve{S}_n^m$ denote the spherical expansion coefficients of the source model, θ_s and ϕ_s its directional dependency, and $\breve{G}_n^0$ the spherical expansion coefficients of a secondary point source that is located at the north pole of the sphere with $\mathbf{x}_0 = (0,0,R_0)$ and is given as[11]

$$\breve{G}_n^0(\tfrac{\pi}{2},0,\omega) = -i\frac{\omega}{c}\sqrt{\frac{2n+1}{4\pi}}\,h_n^{(2)}\left(\frac{\omega}{c}R_0\right)\ , \tag{2.9}$$

where $h_n^{(2)}$ describes the spherical Hankel function of n-th order and second kind.

2.1.2 *Circular Geometries*

The following functions build a basis in $\mathbb{R}^2$ for a circular secondary source distribution[12]

$$\Phi_m(\phi) = e^{im\phi}\ . \tag{2.10}$$

The complex conjugate of Φ_m is given by negating the degree m as

$$\Phi_m(\phi)^* = \Phi_{-m}(\phi)\ . \tag{2.11}$$

For a circular secondary source distribution with a radius of R_0 the driving function can be calculated by a convolution along the surface of the circle as explicitly shown by Ahrens[13] and is then given as

$$D_{\text{circular}}(\phi_0,\omega) = \frac{1}{2\pi R_0}\sum_{m=-\infty}^{\infty}\frac{\breve{S}_m(\phi_\text{s},r_\text{s},\omega)}{\breve{G}_m(0,\omega)}\,\Phi_m(\phi_0)\ , \tag{2.12}$$

where $\breve{S}_m$ denotes the circular expansion coefficients for the source model, ϕ_s its directional dependency, and $\breve{G}_m$ the circular expansion coefficients for a secondary line source with

$$\breve{G}_m(0,\omega) = -\frac{i}{4}\,H_m^{(2)}\left(\frac{\omega}{c}R_0\right)\ , \tag{2.13}$$

where $H_m^{(2)}$ describes the Hankel function of m-th order and second kind.

2.1.3 *Planar Geometries*

The basis functions for a planar secondary source distribution located on the xz-plane in $\mathbb{R}^3$ are given as

$$\Lambda(k_x,k_z,x,z) = e^{-i(k_x x+k_z z)}\ , \tag{2.14}$$

[10] J. Ahrens. *Analytic Methods of Sound Field Synthesis*. New York: Springer, 2012, (3.21). The $\frac{1}{2\pi}$ term is wrong in (3.21) and omitted here, compare the errata and F. Schultz and S. Spors. "Comparing Approaches to the Spherical and Planar Single Layer Potentials for Interior Sound Field Synthesis". *Acta Acustica* 100.5 (2014), pp. 900–11, (24).

[11] F. Schultz and S. Spors. "Comparing Approaches to the Spherical and Planar Single Layer Potentials for Interior Sound Field Synthesis". *Acta Acustica* 100.5 (2014), pp. 900–11, (25).

[12] Williams, op. cit.

[13] J. Ahrens and S. Spors. "On the Secondary Source Type Mismatch in Wave Field Synthesis Employing Circular Distributions of Loudspeakers". In: *127th Audio Engineering Society Convention*. 2009, Paper 7952.

where k_x, k_z are entries in the wave vector $\mathbf{k}$ with $k^2 = (\frac{\omega}{c})^2$. The complex conjugate is given by negating k_x and k_z as

$$\Lambda(k_x, k_z, x, z)^* = \Lambda(-k_x, -k_z, x, z)\ . \tag{2.15}$$

For an infinitely long secondary source distribution located on the xz-plane the driving function can be calculated by a two-dimensional convolution along the plane as[14]

$$D_{\text{planar}}(x_0, \omega) = \frac{1}{4\pi^2} \iint_{-\infty}^{\infty} \frac{\breve{S}(k_x, y_\text{s}, k_z, \omega)}{\breve{G}(k_x, 0, k_z, \omega)} \Lambda(k_x, x_0, k_z, z_0)\, dk_x dk_z\ , \tag{2.16}$$

where $\breve{S}$ denotes the planar expansion coefficients for the source model, y_s its positional dependency, and $\breve{G}$ the planar expansion coefficients of a secondary point source with[15]

$$\breve{G}(k_x, 0, k_z, \omega) = -\frac{i}{2} \frac{1}{\sqrt{(\frac{\omega}{c})^2 - k_x^2 - k_z^2}}\ , \tag{2.17}$$

for $(\frac{\omega}{c})^2 > (k_x^2 + k_z^2)$.

For the planar and the following linear geometries the Fredholm equation is solved for a non compact space V, which leads to an infinite and non-denumerable number of basis functions as opposed to the denumerable case for compact spaces.[16]

2.1.4 Linear Geometries

The basis functions for a linear secondary source distribution located on the x-axis are given as

$$\chi(k_x, x) = e^{-ik_x x}\ . \tag{2.18}$$

The complex conjugate is given by negating k_x as

$$\chi(k_x, x)^* = \chi(-k_x, x)\ . \tag{2.19}$$

For an infinitely long secondary source distribution located on the x-axis the driving function for $\mathbb{R}^2$ can be calculated by a convolution along this axis as[17]

$$D_{\text{linear}}(x_0, \omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\breve{S}(k_x, y_\text{s}, \omega)}{\breve{G}(k_x, 0, \omega)} \chi(k_x, x_0)\, dk_x\ , \tag{2.20}$$

where $\breve{S}$ denotes the linear expansion coefficients for the source model, y_s, z_s its positional dependency, and $\breve{G}$ the linear expansion coefficients of a secondary line source with

$$\breve{G}(k_x, 0, \omega) = -\frac{i}{2} \frac{1}{\sqrt{(\frac{\omega}{c})^2 - k_x^2}}\ , \tag{2.21}$$

for $0 < |k_x| < |\frac{\omega}{c}|$.

[14] Ahrens, op. cit., (3.65).

[15] Schultz and Spors, op. cit., (65).

[16] Ibid.

[17] Compare (3.73) in Ahrens, op. cit.

2.2 *High Frequency Approximation: Wave Field Synthesis*

The single-layer potential (2.1) satisfies the homogeneous Helmholtz equation both in the interior and exterior regions V and $V^* := \mathbb{R}^n \setminus (V \cup \partial V)$. If $D(\mathbf{x}_0, \omega)$ is continuous, the pressure $P(\mathbf{x}, \omega)$ 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. The strength of the secondary sources is then given by the differences of the gradients approaching ∂V from both sides as[18]

$$D(\mathbf{x}_0, \omega) = \partial_{\mathbf{n}} P(\mathbf{x}_0, \omega) + \partial_{-\mathbf{n}} P(\mathbf{x}_0, \omega) \,, \tag{2.22}$$

where $\partial_{\mathbf{n}} := \langle \nabla, \mathbf{n} \rangle$ is the directional gradient in direction $\mathbf{n}$ – see Figure 2.1. Due to the symmetry of the problem the solution for an infinite planar boundary ∂V is given as

$$D(\mathbf{x}_0, \omega) = -2\partial_{\mathbf{n}} S(\mathbf{x}_0, \omega) \,, \tag{2.23}$$

where the pressure in the outside region is the mirrored interior pressure given by the source model $S(\mathbf{x}, \omega)$ for $\mathbf{x} \in V$. The integral equation resulting from introducing (2.23) into (2.1) for a planar boundary ∂V is known as *Rayleigh's first integral equation*. This solution is identical to the explicit solution for planar geometries (2.16) in $\mathbb{R}^3$ and for linear geometries (2.20) in $\mathbb{R}^2$.

A solution of (2.22) for arbitrary boundaries can be found by applying the *Kirchhoff* or *physical optics approximation*.[19] In acoustics this is also known as *determining the visible elements* for the high frequency boundary element method.[20] Here, it is assumed that a bent surface can be approximated by a set of small planar surfaces for which (2.23) holds locally. In general, this will be the case if the wave length is much smaller than the size of a planar surface patch and the position of the listener is far away from the secondary sources.[21] Additionally, only one part of the surface is active: the area that is illuminated from the incident field of the source model.

With this approximation also non-convex secondary source distributions can be used with WFS – compare Figure 2.2.[22] This was neglected in most of the literature so far, which postulates convex secondary source distributions.[23]

The outlined approximation can be formulated by introducing a window function $w(\mathbf{x}_0)$ for the selection of the active secondary sources into (2.23) as

$$P(\mathbf{x}, \omega) \approx \oint_{\partial V} G(\mathbf{x}|\mathbf{x}_0, \omega) \underbrace{-2w(\mathbf{x}_0)\partial_{\mathbf{n}} S(\mathbf{x}_0, \omega)}_{D(\mathbf{x}_0, \omega)} \, dA(\mathbf{x}_0) \,. \tag{2.24}$$

One of the advantages of the applied approximation is that due to its local character the solution of the driving function (2.23) does not depend on the geometry of the secondary sources. This dependency applies to the direct solutions presented in Section 2.1.

[18] Compare F. M. Fazi and P. A. Nelson. "Sound field reproduction as an equivalent acoustical scattering problem". *The Journal of the Acoustical Society of America* 134.5 (2013), pp. 3721–9

[19] See D. Colton and R. Kress. *Integral Equation Methods in Scattering Theory*. New York: Wiley, 1983, p. 53–54

[20] E.g. D. W. Herrin et al. "A New Look at the High Frequency Boundary Element and Rayleigh Integral Approximations". In: *Noise & Vibration Conference and Exhibition*. 2003

[21] Compare the two assumptions in S. Spors and F. Zotter. "Spatial Sound Synthesis with Loudspeakers". In: *Cutting Edge in Spatial Audio, EAA Winter School*. 2013, pp. 32–37, made before (15), which lead to the derivation of the same window function in a more explicit way.

[22] See the appendix in M. Lax and H. Feshbach. "On the Radiation Problem at High Frequencies". *The Journal of the Acoustical Society of America* 19.4 (1947), pp. 682–90

[23] E.g. S. Spors, R. Rabenstein, and J. Ahrens. "The Theory of Wave Field Synthesis Revisited". In: *124th Audio Engineering Society Convention*. 2008, Paper 7358

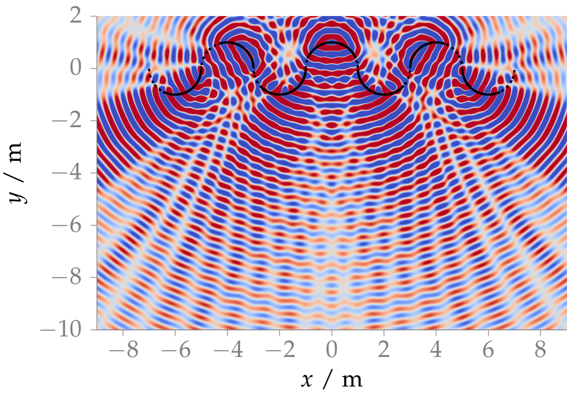


Figure 2.2: Sound pressure of a point source synthesized with WFS (2.62). The secondary source distribution is shown in black, whereby inactive sources are marked with a dashed line. Parameters: $\mathbf{x}_\mathrm{s} = (0, 2.5, 0)\,\mathrm{m}$, $\mathbf{x}_\mathrm{ref} = (0, -3, 0)\,\mathrm{m}$, $f = 700\,\mathrm{Hz}$. ☞

2.3 *Sound Field Dimensionality*

The single-layer potential (2.1) is valid for all $V \subset \mathbb{R}^n$. Consequentially, for practical applications a two-dimensional (2D) as well as a three-dimensional (3D) synthesis is possible. Two-dimensional is not referring to a synthesis in a plane only, but describes a setup that is independent of one dimension. For example, an infinite cylinder is independent of the dimension along its axis. The same is true for secondary source distributions in 2D synthesis. They exhibit line source characteristics and are aligned in parallel to the independent dimension. Typical arrangements of such secondary sources are a circular or a linear setup.

The characteristics of the secondary sources limit the set of possible sources which can be synthesized. For example, when using a 2D secondary source setup it is not possible to synthesize the amplitude decay of a point source.

For a 3D synthesis the involved secondary sources depend on all dimensions and exhibit point source characteristics. In this scenario classical secondary sources setups would be a sphere or a plane.

2.3.1 *2.5D Synthesis*

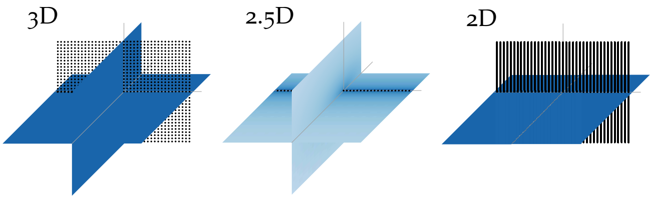


Figure 2.3: Sound pressure in decibel for secondary source distributions with different dimensionality all driven by the same signals. The sound pressure is color coded, lighter color corresponds to lower pressure. In the 3D case a planar distribution of point sources is applied, in the 2.5D case a linear distribution of point sources, and in the 2D case a linear distribution of line sources. ☞

In practice, the most common setups of secondary sources are 2D setups, employing cabinet loudspeakers. A cabinet loudspeaker does not show the characteristics of a line source, but of a point source. This dimensionality mismatch prevents perfect synthesis within the desired plane. The combination of a 2D secondary source setup with secondary sources that exhibit 3D characteristics has led to naming such configurations *2.5D synthesis*.[24] Such scenarios are associated with a wrong amplitude decay due to the inherent mismatch of sec-

[24] E. W. Start. "Direct Sound Enhancement by Wave Field Synthesis". PhD thesis. Technische Universiteit Delft, 1997

ondary sources as is highlighted in Figure 2.3. In general, the amplitude is only correct at a given reference point $\mathbf{x}_{\text{ref}}$.

For a circular secondary source distribution with point source characteristic the 2.5D driving function can be derived by introducing expansion coefficients for the spherical case into the driving function (2.12). The equation is than solved for $\theta = 0°$ and $r_{\text{ref}} = 0$. This results in a 2.5D driving function given in Ahrens[25] as

$$D_{\text{circular,2.5D}}(\phi_0, \omega) = \frac{1}{2\pi R_0} \sum_{m=-\infty}^{\infty} \frac{\breve{S}_{|m|}^{m}(\frac{\pi}{2}, \phi_{\text{s}}, r_{\text{s}}, \omega)}{\breve{G}_{|m|}^{m}(\frac{\pi}{2}, 0, \omega)} \Phi_m(\phi_0) \ . \qquad (2.25)$$

For a linear secondary source distribution with point source characteristics the 2.5D driving function is derived by introducing the linear expansion coefficients for a monopole source (2.38) into the driving function (2.20) and solving the equation for $y = y_{\text{ref}}$ and $z = 0$. This results in a 2.5D driving function given as[26]

$$D_{\text{linear,2.5D}}(x_0, \omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\breve{S}(k_x, y_{\text{ref}}, 0, \omega)}{\breve{G}(k_x, y_{\text{ref}}, 0, \omega)} \chi(k_x, x_0) \, dk_x \ . \qquad (2.26)$$

A driving function for the 2.5D situation in the context of WFS and arbitrary 2D geometries of the secondary source distribution can be achieved by applying the far-field approximation[27] $H_0^{(2)}(\zeta) \approx \sqrt{\frac{2i}{\pi\zeta}} e^{-i\zeta}$ for $\zeta \gg 1$ to the 2D Green's function. Using this the following relationship between the 2D and 3D Green's functions can be established.

$$\underbrace{-\frac{i}{4} \ H_0^{(2)}\left(\frac{\omega}{c}|\mathbf{x} - \mathbf{x}_0|\right)}_{G_{\text{2D}}(\mathbf{x}-\mathbf{x}_0, \omega)} \ \approx \ \sqrt{2\pi \frac{c}{i\omega} |\mathbf{x} - \mathbf{x}_0|} \ \underbrace{\frac{1}{4\pi} \frac{e^{-i\frac{\omega}{c}|\mathbf{x}-\mathbf{x}_0|}}{|\mathbf{x} - \mathbf{x}_0|}}_{G_{\text{3D}}(\mathbf{x}-\mathbf{x}_0, \omega)} \ , \qquad (2.27)$$

where $H_0^{(2)}$ denotes the Hankel function of second kind and zeroth order. Inserting this approximation into the single-layer potential for the 2D case results in

$$P(\mathbf{x}, \omega) = \oint_S \sqrt{2\pi \frac{c}{i\omega} |\mathbf{x} - \mathbf{x}_0|} \ D(\mathbf{x}_0, \omega) G_{\text{3D}}(\mathbf{x} - \mathbf{x}0, \omega) \, dA(\mathbf{x}_0) \ . \qquad (2.28)$$

If the amplitude correction is further restricted to one reference point $\mathbf{x}_{\text{ref}}$, the 2.5D driving function for WFS can be formulated as

$$D_{\text{2.5D}}(\mathbf{x}_0, \omega) = \underbrace{\sqrt{2\pi |\mathbf{x}_{\text{ref}} - \mathbf{x}_0|}}_{g_0} \sqrt{\frac{c}{i\omega}} \, D(\mathbf{x}_0, \omega) \ , \qquad (2.29)$$

where g_0 is independent of $\mathbf{x}$.

[25] Ibid., (3.49).

[26] Ibid., (3.77).

[27] Williams, op. cit., (4.23).

2.4 *Model-Based Rendering*

greens_function_mono.m
greens_function_imp.m

Knowing the pressure field of the desired source $S(\mathbf{x},\omega)$ is required in order to derive the driving signal for the secondary source distribution. It can either be measured, i.e. recorded, or modeled. While the former is known as *data-based rendering*, the latter is known as *model-based rendering*. For data-based rendering, the problem of how to capture a complete sound field still has to be solved. Avni et al. discuss some influences of the recording limitations on the perception of the reproduced sound field.[28] This thesis focusses on the perception of the synthesis part. Therefore it will consider only model-based rendering.

Frequently applied models in model-based rendering are plane waves, point sources, or sources with a prescribed complex directivity. In the following the models used within the Sound Field Synthesis Toolbox are presented.

Plane Wave The source model for a plane wave is given as[29]

$$S(\mathbf{x},\omega) = A(\omega) e^{-i\frac{\omega}{c}\mathbf{n}_k\mathbf{x}} \,, \tag{2.30}$$

where $A(\omega)$ denotes the frequency spectrum of the source and $\mathbf{n}_k$ a unit vector pointing into the direction of the plane wave.

Transformed in the temporal domain this becomes

$$s(\mathbf{x},t) = a(t) * \delta\left(t - \frac{\mathbf{n}_k\mathbf{x}}{c}\right) \,, \tag{2.31}$$

where $a(t)$ is the Fourier transformation of the frequency spectrum $A(\omega)$.

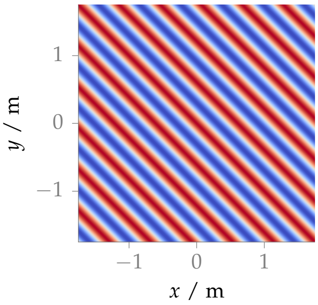


Figure 2.4: Sound pressure for a monochromatic plane wave (2.30) going into the direction $(1,1,0)$. Parameters: $f = 800\,\mathrm{Hz}$. ☞

The expansion coefficients for spherical basis functions are given as[30]

$$\breve{S}_n^m(\theta_k,\phi_k,\omega) = 4\pi i^{-n} Y_n^{-m}(\theta_k,\phi_k) \,, \tag{2.32}$$

where (ϕ_k,θ_k) is the radiating direction of the plane wave.

In a similar manner the expansion coefficients for circular basis functions are given as

$$\breve{S}_m(\phi_\mathrm{s},\omega) = i^{-n}\Phi_{-m}(\phi_\mathrm{s}) \,. \tag{2.33}$$

The expansion coefficients for linear basis functions are given as after Ahrens[31]

$$\breve{S}(k_x,y,\omega) = 2\pi\,\delta(k_x - k_{x,\mathrm{s}})\,\chi(k_{y,\mathrm{s}},y) \,, \tag{2.34}$$

where $(k_{x,\mathrm{s}},k_{y,\mathrm{s}})$ points into the radiating direction of the plane wave.

Point Source The source model for a point source is given by the three dimensional Green's function as[32]

$$S(\mathbf{x},\omega) = A(\omega)\frac{1}{4\pi}\frac{e^{-i\frac{\omega}{c}|\mathbf{x}-\mathbf{x}_\mathrm{s}|}}{|\mathbf{x}-\mathbf{x}_\mathrm{s}|} \,, \tag{2.35}$$

where $\mathbf{x}_\mathrm{s}$ describes the position of the point source.

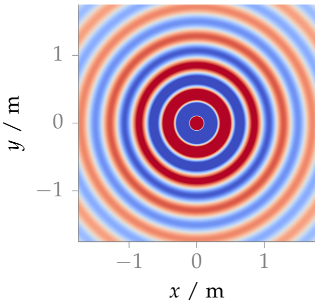


Figure 2.5: Sound pressure for a monochromatic point source (2.35) placed at $(0,0,0)$. Parameters: $f = 800\,\mathrm{Hz}$. ☞

[28] A. Avni et al. "Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution". *The Journal of the Acoustical Society of America* 133.5 (2013), pp. 2711–21.

[29] E. G. Williams. *Fourier Acoustics*. San Diego: Academic Press, 1999, p. 21, (2.24). Williams defines the Fourier transform with transposed signs as $F(\omega) = \int f(t)e^{i\omega t}$. This leads also to changed signs in his definitions of the Green's functions and field expansions.

[30] Ahrens, op. cit., (2.38)

[31] ibid., (C.5)

[32] Williams, op. cit., (6.73)

Transformed to the temporal domain this becomes

$$s(\mathbf{x},t) = a(t) * \frac{1}{4\pi}\frac{1}{|\mathbf{x}-\mathbf{x}_\mathrm{s}|}\,\delta\!\left(t - \frac{|\mathbf{x}-\mathbf{x}_\mathrm{s}|}{c}\right) \ . \tag{2.36}$$

The expansion coefficients for spherical basis functions are given as[33]

$$\breve{S}_n^m(\theta_\mathrm{s},\phi_\mathrm{s},r_\mathrm{s},\omega) = -i\frac{\omega}{c}\,h_n^{(2)}\!\left(\frac{\omega}{c}r_\mathrm{s}\right)Y_n^{-m}(\theta_\mathrm{s},\phi_\mathrm{s}) \ , \tag{2.37}$$

where $(\phi_\mathrm{s},\theta_\mathrm{s},r_\mathrm{s})$ describes the position of the point source.

The expansion coefficients for linear basis functions are given as[34]

$$\breve{S}(k_x,y,\omega) = -\frac{i}{4}\,H_0^{(2)}\!\left(\sqrt{(\tfrac{\omega}{c})^2 - k_x^2}\;|y-y_\mathrm{s}|\right)\chi(-k_x,x_\mathrm{s}) \ , \tag{2.38}$$

for $|k_x| < |\frac{\omega}{c}|$ and with $(x_\mathrm{s},y_\mathrm{s})$ describing the position of the point source.

Line Source The source model for a line source is given by the two dimensional Green's function as[35]

$$S(\mathbf{x},\omega) = -A(\omega)\frac{i}{4}\,H_0^{(2)}\!\left(\frac{\omega}{c}|\mathbf{x}-\mathbf{x}_\mathrm{s}|\right) \ . \tag{2.39}$$

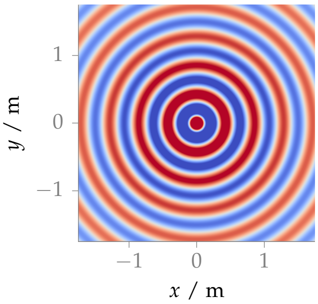


Figure 2.6: Sound pressure for a monochromatic line source (2.39) placed at $(0,0,0)$. Parameters: $f = 800\,\mathrm{Hz}$. ☞

Applying the large argument approximation of the Hankel function[36] and transformed to the temporal domain this becomes

$$s(\mathbf{x},t) = a(t) * \mathcal{F}^{-1}\left\{\sqrt{\frac{c}{i\omega}}\right\} * \sqrt{\frac{1}{8\pi}}\frac{1}{\sqrt{|\mathbf{x}-\mathbf{x}_\mathrm{s}|}}\,\delta\!\left(t - \frac{|\mathbf{x}-\mathbf{x}_\mathrm{s}|}{c}\right) \ . \tag{2.40}$$

The expansion coefficients for circular basis functions are given as

$$\breve{S}_m(\phi_\mathrm{s},r_\mathrm{s},\omega) = -\frac{i}{4}\,H_m^{(2)}\!\left(\frac{\omega}{c}r_\mathrm{s}\right)\Phi_{-m}(\phi_\mathrm{s}) \ . \tag{2.41}$$

The expansion coefficients for linear basis functions are given as

$$\breve{S}(k_x,y_\mathrm{s},\omega) = -\frac{i}{2}\frac{1}{\sqrt{(\frac{\omega}{c})^2 - k_x^2}}\chi(k_y,y_\mathrm{s}) \ . \tag{2.42}$$

2.5 *Driving Functions*

In the following, driving functions for Near-Field Compensated Higher Order Ambisonics, the Spectral Division Method and Wave Field Synthesis are derived for spherical, circular, and linear secondary source distributions. Among the possible combinations of methods and secondary sources not all are meaningful. Hence, only the relevant ones will be presented. The same holds for the introduced source models of plane waves, point sources, line sources and focused sources. Ahrens and Spors[37] in addition have considered Spectral Division Method driving functions for planar secondary source distributions.

The driving functions are given in the temporal-frequency domain. For some of them, especially in the case of WFS an analytic solution in the temporal domain exists and is presented. For NFC-HOA,

[33] Ahrens, op. cit., (2.37).

[34] Ibid., (C.10).

[35] ibid., (8.47)

[36] ibid., (4.23)

[37] Ahrens and Spors, op. cit.

temporal-domain implementations for the 2.5D cases are available for a plane wave and a point source as source models. The derivation of the implementation is not explicitly shown here, but is described in Spors et al.[38]

The 2.5D cases are illustrated in the following by companion figures, because only those cases will be investigated in the remainder of this thesis.

2.5.1 *Near-Field Compensated Higher Order Ambisonics and Spectral Division Method*

`driving_function_mono_nfchoa_pw.m`
`driving_function_mono_sdm_pw.m`
`driving_function_imp_nfchoa_pw.m`

Plane Wave For a spherical secondary source distribution with radius R_0 the spherical expansion coefficients of a plane wave (2.32) and of the Green's function for a point source (2.9) are inserted into (2.8) and yield[39]

$$D_{\text{spherical}}(\theta_0,\phi_0,\omega) = -A(\omega)\frac{4\pi}{R_0^2}\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\frac{i^{-n}Y_n^{-m}(\theta_k,\phi_k)}{i\frac{\omega}{c}h_n^{(2)}\left(\frac{\omega}{c}R_0\right)}Y_n^m(\theta_0,\phi_0)\,. \tag{2.43}$$

For a circular secondary source distribution with radius R_0 the circular expansion coefficients of a plane wave (2.33) and of the Green's function for a line source (2.13) are inserted into (2.12) and yield[40]

$$D_{\text{circular}}(\phi_0,\omega) = -A(\omega)\frac{2i}{\pi R_0}\sum_{m=-\infty}^{\infty}\frac{i^{-m}\Phi_{-m}(\phi_k)}{H_m^{(2)}(\frac{\omega}{c}R_0)}\Phi_m(\phi_0)\,. \tag{2.44}$$

For a circular secondary source distribution with radius R_0 and point source as Green's function the 2.5D driving function is given by inserting the spherical expansion coefficients for a plane wave (2.32) and a point source (2.37) into (2.25) as

$$D_{\text{circular, 2.5D}}(\phi_0,\omega) = -A(\omega)\frac{2}{R_0}\sum_{m=-\infty}^{\infty}\frac{i^{-|m|}\Phi_{-m}(\phi_k)}{i\frac{\omega}{c}h_{|m|}^{(2)}\left(\frac{\omega}{c}R_0\right)}\Phi_m(\phi_0)\,. \tag{2.45}$$

For an infinite linear secondary source distribution located on the x-axis the 2.5D driving function is given by inserting the linear expansion coefficients for a point source as Green's function (2.21) and a plane wave (2.34) into (2.26) and exploiting the fact that $(\frac{\omega}{c})^2 - k_{x_\text{s}}$ is constant. Assuming $0 \leq |k_{x_\text{s}}| \leq |\frac{\omega}{c}|$ this results in[41]

$$D_{\text{linear, 2.5D}}(x_0,\omega) = A(\omega)\frac{4i\chi(k_y,y_\text{ref})}{H_0^{(2)}(k_y y_\text{ref})}\chi(k_x,x_0)\,. \tag{2.46}$$

Transfered to the temporal domain this results in[42]

$$d_{\text{linear, 2.5D}}(x_0,t) = h(t) * a(t - \frac{x_0}{c}\sin\phi_k - \frac{y_\text{ref}}{c} sin\phi_k)\,, \tag{2.47}$$

where ϕ_k denotes the azimuth direction of the plane wave and

$$h(t) = \mathcal{F}^{-1}\left\{\frac{4i}{H_0^{(2)}(k_y y_\text{ref})}\right\}\,. \tag{2.48}$$

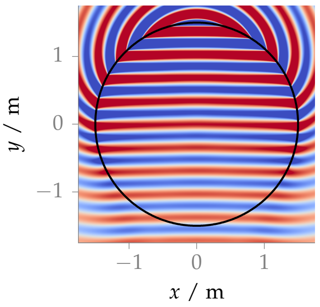


Figure 2.7: Sound pressure of a monochromatic plane wave synthesized with 2.5D NFC-HOA (2.45). Parameters: $\mathbf{n}_k = (0,-1,0)$, $\mathbf{x}_\text{ref} = (0,0,0)$, $f = 1\,\text{kHz}$. ☞

[38] S. Spors, V. Kuscher, and J. Ahrens. "Efficient realization of model-based rendering for 2.5-dimensional near-field compensated higher order Ambisonics". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 2011, pp. 61–64

[39] Schultz and Spors, op. cit., (96).

[40] Compare J. Ahrens and S. Spors. "On the Secondary Source Type Mismatch in Wave Field Synthesis Employing Circular Distributions of Loudspeakers". In: *127th Audio Engineering Society Convention*. 2009, Paper 7952, (16)

[41] J. Ahrens and S. Spors. "Sound Field Reproduction Using Planar and Linear Arrays of Loudspeakers". *IEEE Transactions on Audio, Speech, and Language Processing* 18.8 (2010), pp. 2038–50, (17)

[42] ibid., (18)

The advantage of this result is that it can be implemented by a simple weighting and delaying of the signal, plus one convolution with $h(t)$. The same holds for the driving functions of WFS as presented in the next section.

Point Source For a spherical secondary source distribution with radius R_0 the spherical coefficients of a point source (2.37) and of the Green's function (2.9) are inserted into (2.8) and yield

driving_function_mono_nfchoa_ps.m
driving_function_imp_nfchoa_ps.m

$$D_{\text{spherical}}(\theta_0,\phi_0,\omega) = \\ A(\omega)\frac{1}{R_0^2}\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\frac{h_n^{(2)}(\frac{\omega}{c}r_{\text{s}})Y_n^{-m}(\theta_{\text{s}},\phi_{\text{s}})}{h_n^{(2)}(\frac{\omega}{c}R_0)}Y_n^m(\theta_0,\phi_0)\,. \tag{2.49}$$

For a circular secondary source distribution with radius R_0 and point source as secondary sources the 2.5D driving function is given by inserting the spherical coefficients (2.37) and (2.9) into (2.25) as

$$D_{\text{circular, 2.5D}}(\phi_0,\omega) = A(\omega)\frac{1}{2\pi R_0}\sum_{m=-\infty}^{\infty}\frac{h_{|m|}^{(2)}(\frac{\omega}{c}r_s)\Phi_{-m}(\phi_{\text{s}})}{h_{|m|}^{(2)}(\frac{\omega}{c}R_0)}\Phi_m(\phi_0)\,. \tag{2.50}$$

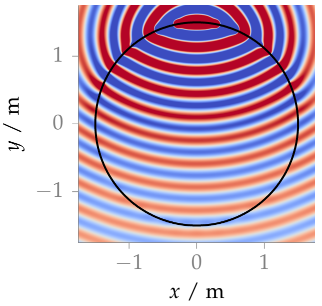


Figure 2.8: Sound pressure for a monochromatic point source synthesized by 2.5D NFC-HOA (2.50). Parameters: $\mathbf{x}_{\text{s}} = (0,2.5,0)\,\text{m}$, $\mathbf{x}_{\text{ref}} = (0,0,0)$, $f = 1\,\text{kHz}$. ☞

For an infinite linear secondary source distribution located on the x-axis and point sources as secondary sources the 2.5D driving function for a point source is given by inserting the corresponding linear expansion coefficients (2.38) and (2.21) into (2.26). Assuming $0 \leq |k_x| < |\frac{\omega}{c}|$ this results in[43]

$$D_{\text{linear, 2.5D}}(x_0,\omega) = \\ A(\omega)\int_{-\infty}^{\infty}\frac{H_0^{(2)}\left(\sqrt{(\frac{\omega}{c})^2-k_x^2}\,(y_{\text{ref}}-y_{\text{s}})\right)\chi(-k_x,x_{\text{s}})}{H_0^{(2)}\left(\sqrt{(\frac{\omega}{c})^2-k_x^2}\,y_{\text{ref}}\right)}\chi(k_x,x_0)\,dk_x\,. \tag{2.51}$$

Line Source For a circular secondary source distribution with radius R_0 and line sources as secondary sources the driving function is given by inserting the circular coefficients (2.41) and (2.13) into (2.12) as

$$D_{\text{circular}}(\phi_0,\omega) = A(\omega)\frac{1}{2\pi R_0}\sum_{m=-\infty}^{\infty}\frac{H_m^{(2)}(\frac{\omega}{c}r_s)\Phi_{-m}(\phi_{\text{s}})}{H_m^{(2)}(\frac{\omega}{c}R_0)}\Phi_m(\phi_0)\,. \tag{2.52}$$

For an infinite linear secondary source distribution located on the x-axis and line sources as secondary sources the driving function is given by inserting the linear coefficients (2.42) and (2.13) into (2.20) as

$$D_{\text{linear}}(x_0,\omega) = A(\omega)\frac{1}{2\pi}\int_{-\infty}^{\infty}\chi(k_y,y_s)\chi(k_x,x_0)\,dk_x\,. \tag{2.53}$$

Focused Source Focused sources mimic point or line sources that are located inside the audience area. For the single-layer potential the

[43] Compare (4.53) in Ahrens, op. cit.

assumption is that the audience area is free from sources and sinks. However, a focused source is neither of them. It represents a sound field that converges towards a focal point and diverges afterwards. This can be achieved by reversing the driving function of a point or line source in time which is known as time reversal focusing.[44]

Nonetheless, the single-layer potential should not be solved for focused sources without any approximation. In the near field of a source, evanescent waves[45] appear for spatial frequencies $k_x > |\frac{\omega}{c}|$. They decay exponentially with the distance from the source. An exact solution for a focused source is supposed to include these evanescent waves around the focal point. That is only possible by applying very large amplitudes to the secondary sources.[46] Since the evanescent waves decay rapidly and are hence not influencing the perception, they can easily be omitted. For corresponding driving functions for focused sources without the evanescent part of the sound field see Spors and Ahrens[47] for SDM and Ahrens and Spors[48] for NFC-HOA.

In this thesis only focused sources in WFS will be considered.

2.5.2 *Wave Field Synthesis*

In the following, the driving functions for WFS in the frequency and temporal domain for selected source models are presented. The temporal domain functions consist of a filtering of the source signal and a weighting and delaying of the individual secondary source signals. This property allows for a very efficient implementation of WFS driving functions in the temporal domain. It is one of the main advantages of WFS in comparison to most of the NFC-HOA/SDM solutions discussed above.

`driving_function_mono_wfs_pw.m`
`driving_function_imp_wfs_pw.m`

Plane Wave By inserting the source model of a plane wave (2.30) into (2.23) and (2.29) it follows

$$D(\mathbf{x}_0,\omega) = 2w(\mathbf{x}_0)A(\omega)i\frac{\omega}{c}\mathbf{n}_k\mathbf{n}_{\mathbf{x}_0}e^{-i\frac{\omega}{c}\mathbf{n}_k\mathbf{x}_0}\,, \tag{2.54}$$

$$D_{2.5\mathrm{D}}(\mathbf{x}_0,\omega) = 2g_0w(\mathbf{x}_0)A(\omega)\sqrt{i\frac{\omega}{c}}\mathbf{n}_k\mathbf{n}_{\mathbf{x}_0}e^{-i\frac{\omega}{c}\mathbf{n}_k\mathbf{x}_0}\,. \tag{2.55}$$

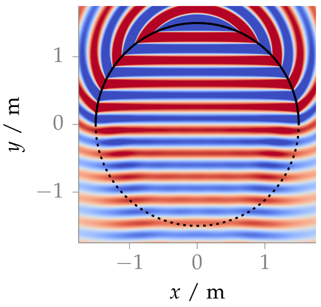


Figure 2.9: Sound pressure for a monochromatic plane wave synthesized by 2.5D WFS (2.55). Parameters: $\mathbf{n}_k = (0,-1,0)$, $\mathbf{x}_{\mathrm{ref}} = (0,0,0)$, $f = 1\,\mathrm{kHz}$. ☞

Transfered to the temporal domain via an inverse Fourier transform (1.2), it follows

$$d(\mathbf{x}_0,t) = 2a(t) * h(t) * w(\mathbf{x}_0)\mathbf{n}_k\mathbf{n}_{\mathbf{x}_0}\,\delta\left(t - \frac{\mathbf{n}_k\mathbf{x}_0}{c}\right)\,, \tag{2.56}$$

$$d_{2.5\mathrm{D}}(\mathbf{x}_0,t) = 2g_0a(t) * h_{2.5\mathrm{D}}(t) * w(\mathbf{x}_0)\mathbf{n}_k\mathbf{n}_{\mathbf{x}_0}\,\delta\left(t - \frac{\mathbf{n}_k\mathbf{x}_0}{c}\right)\,, \tag{2.57}$$

where $h(t) = \mathcal{F}^{-1}\left\{i\frac{\omega}{c}\right\}$ and $h_{2.5\mathrm{D}}(t) = \mathcal{F}^{-1}\left\{\sqrt{i\frac{\omega}{c}}\right\}$ denote the so called pre-equalization filters in WFS.

`wfs_fir_prefilter.m`
`secondary_source_selection.m`

The window function $w(\mathbf{x}_0)$ for a plane wave as source model can be calculated after Spors et al. as[49]

$$w(\mathbf{x}_0) = \begin{cases} 1 & \mathbf{n}_k\mathbf{n}_{\mathbf{x}_0} > 0 \\ 0 & \text{else} \end{cases} \tag{2.58}$$

[44] S. Yon, M. Tanter, and M. Fink. "Sound focusing in rooms: The time-reversal approach". *The Journal of the Acoustical Society of America* 113.3 (2003), pp. 1533–43.

[45] Williams, op. cit., p. 24.

[46] Compare Fig. 2a in S. Spors and J. Ahrens. "Reproduction of Focused Sources by the Spectral Division Method". In: *International Symposium on Communications, Control and Signal Processing*. 2010

[47] ibid.

[48] J. Ahrens and S. Spors. "Spatial encoding and decoding of focused virtual sound sources". In: *International Symposium on Ambisonics and Spherical Acoustics*. 2009

[49] S. Spors, R. Rabenstein, and J. Ahrens. "The Theory of Wave Field Synthesis Revisited". In: *124th Audio Engineering Society Convention*. 2008, Paper 7358.

driving_function_mono_wfs_ps.m
driving_function_imp_wfs_ps.m

Point Source By inserting the source model for a point source (2.35) into (2.23) and (2.29) it follows

$$D(\mathbf{x}_0,\omega) = \frac{1}{2\pi}A(\omega)w(\mathbf{x}_0)\left(i\frac{\omega}{c}+\frac{1}{|\mathbf{x}_0-\mathbf{x}_\text{s}|}\right)\frac{(\mathbf{x}_0-\mathbf{x}_\text{s})\mathbf{n}_{\mathbf{x}_0}}{|\mathbf{x}_0-\mathbf{x}_\text{s}|^2}e^{-i\frac{\omega}{c}|\mathbf{x}_0-\mathbf{x}_\text{s}|}\,, \tag{2.59}$$

$$D_\text{2.5D}(\mathbf{x}_0,\omega) = \frac{g_0}{2\pi}A(\omega)w(\mathbf{x}_0)\sqrt{i\frac{\omega}{c}}\left(1+\frac{1}{i\frac{\omega}{c}|\mathbf{x}_0-\mathbf{x}_\text{s}|}\right)\frac{(\mathbf{x}_0-\mathbf{x}_\text{s})\mathbf{n}_{\mathbf{x}_0}}{|\mathbf{x}_0-\mathbf{x}_\text{s}|^2}e^{-i\frac{\omega}{c}|\mathbf{x}_0-\mathbf{x}_\text{s}|}\,. \tag{2.60}$$

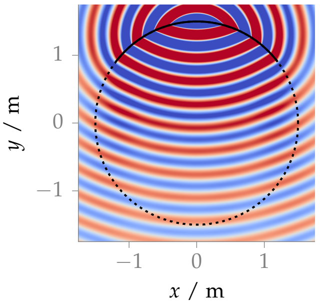


Figure 2.10: Sound pressure for a monochromatic point source synthesized by 2.5D WFS (2.62). Parameters: $\mathbf{x}_\text{s} = (0, 2.5, 0)$ m, $\mathbf{x}_\text{ref} = (0, 0, 0)$, $f = 1$ kHz. ☞

Under the assumption of $|\mathbf{x}_0-\mathbf{x}_\text{s}| \gg 1$ (2.59) and (2.60) can be approximated by

$$D(\mathbf{x}_0,\omega) = \frac{1}{2\pi}A(\omega)w(\mathbf{x}_0)i\frac{\omega}{c}\frac{(\mathbf{x}_0-\mathbf{x}_\text{s})\mathbf{n}_{\mathbf{x}_0}}{|\mathbf{x}_0-\mathbf{x}_\text{s}|^{3/2}}e^{-i\frac{\omega}{c}|\mathbf{x}_0-\mathbf{x}_\text{s}|}\,, \tag{2.61}$$

$$D_\text{2.5D}(\mathbf{x}_0,\omega) = \frac{g_0}{2\pi}A(\omega)w(\mathbf{x}_0)\sqrt{i\frac{\omega}{c}}\frac{(\mathbf{x}_0-\mathbf{x}_\text{s})\mathbf{n}_{\mathbf{x}_0}}{|\mathbf{x}_0-\mathbf{x}_\text{s}|^{3/2}}e^{-i\frac{\omega}{c}|\mathbf{x}_0-\mathbf{x}_\text{s}|}\,, \tag{2.62}$$

which is the traditional formulation of a point source in WFS as given for the 2.5D case in Verheijen.[50] It has the advantage that its temporal domain version could again be implemented as a simple weighting- and delaying-mechanism. This is the default driving function for a point source in the Sound Field Synthesis Toolbox.

Transfered to the temporal domain via an inverse Fourier transform (1.2) it follows

$$d(\mathbf{x}_0,t) = \frac{1}{2\pi}a(t)*h(t)*w(\mathbf{x}_0)\frac{(\mathbf{x}_0-\mathbf{x}_\text{s})\mathbf{n}_{\mathbf{x}_0}}{|\mathbf{x}_0-\mathbf{x}_\text{s}|^{3/2}}\,\delta\left(t-\frac{|\mathbf{x}_0-\mathbf{x}_\text{s}|}{c}\right)\,, \tag{2.63}$$

$$d_\text{2.5D}(\mathbf{x}_0,t) = \frac{g_0}{2\pi}a(t)*h_\text{2.5D}(t)*w(\mathbf{x}_0)\frac{(\mathbf{x}_0-\mathbf{x}_\text{s})\mathbf{n}_{\mathbf{x}_0}}{|\mathbf{x}_0-\mathbf{x}_\text{s}|^{3/2}}\,\delta\left(t-\frac{|\mathbf{x}_0-\mathbf{x}_\text{s}|}{c}\right)\,. \tag{2.64}$$

secondary_source_selection.m

The window function $w(\mathbf{x}_0)$ for a point source as source model can be calculated after Spors at al. as[51]

$$w(\mathbf{x}_0) = \begin{cases} 1 & (\mathbf{x}_0-\mathbf{x}_\text{s})\mathbf{n}_{\mathbf{x}_0} > 0 \\ 0 & \text{else} \end{cases} \tag{2.65}$$

driving_function_mono_wfs_ls.m
driving_function_imp_wfs_ls.m

Line Source By inserting the source model for a line source (2.39) into (2.23) and (2.29) and calculating the derivate of the Hankel function[52] it follows

$$D(\mathbf{x}_0,\omega) = -\frac{1}{2}A(\omega)w(\mathbf{x}_0)i\frac{\omega}{c}\frac{(\mathbf{x}_0-\mathbf{x}_\text{s})\mathbf{n}_{\mathbf{x}_0}}{|\mathbf{x}_0-\mathbf{x}_\text{s}|}\,H_1^{(2)}\left(\frac{\omega}{c}|\mathbf{x}_0-\mathbf{x}_\text{s}|\right)\,, \tag{2.66}$$

[50] E. Verheijen. "Sound Reproduction by Wave Field Synthesis". PhD thesis. Technische Universiteit Delft, 1997, (2.22a), whereby r corresponds to $|\mathbf{x}_0 - \mathbf{x}_\text{s}|$ and $\cos\varphi$ to $\frac{(\mathbf{x}_0-\mathbf{x}_\text{s})\mathbf{n}_{\mathbf{x}_0}}{|\mathbf{x}_0-\mathbf{x}_\text{s}|}$.

[51] Spors, Rabenstein, and Ahrens, op. cit.

[52] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions*. Washington: National Bureau of Standards, 1972, (9.1.30).

$$D_{2.5\mathrm{D}}(\mathbf{x}_0,\omega) = \\ -\frac{1}{2} g_0 A(\omega) w(\mathbf{x}_0) \sqrt{i\frac{\omega}{c}} \frac{(\mathbf{x}_0-\mathbf{x}_\mathrm{s})\mathbf{n}_{\mathbf{x}_0}}{|\mathbf{x}_0-\mathbf{x}_\mathrm{s}|} H_1^{(2)}\left(\frac{\omega}{c}|\mathbf{x}_0-\mathbf{x}_\mathrm{s}|\right) . \tag{2.67}$$

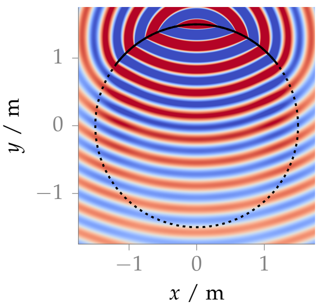


Figure 2.11: Sound pressure for a monochromatic line source synthesized by 2.5D WFS (2.67). Parameters: $\mathbf{x}_\mathrm{s} = (0, 2.5, 0)\,\mathrm{m}$, $\mathbf{x}_\mathrm{ref} = (0,0,0)$, $f = 1\,\mathrm{kHz}$. ☞

Applying $H_1^{(2)}(\zeta) \approx -\sqrt{\frac{2}{\pi i \zeta}} e^{-i\zeta}$ for $z \gg 1$ after Williams[53] and transfered to the temporal domain via an inverse Fourier transform (1.2) it follows

$$d(\mathbf{x}_0,t) = \sqrt{\frac{1}{2\pi}} a(t) * h(t) * w(x0) \frac{(\mathbf{x}_0-\mathbf{x}_\mathrm{s})\mathbf{n}_{\mathbf{x}_0}}{|\mathbf{x}_0-\mathbf{x}_\mathrm{s}|^{3/2}} \delta\left(t - \frac{|\mathbf{x}_0-\mathbf{x}_\mathrm{s}|}{c}\right) , \tag{2.68}$$

$$d_{2.5\mathrm{D}}(\mathbf{x}_0,t) = \\ g_0 \sqrt{\frac{1}{2\pi}} a(t) * \mathcal{F}^{-1}\left\{\sqrt{\frac{c}{i\omega}}\right\} * w(x0) \frac{(\mathbf{x}_0-\mathbf{x}_\mathrm{s})\mathbf{n}_{\mathbf{x}_0}}{|\mathbf{x}_0-\mathbf{x}_\mathrm{s}|^{3/2}} \delta\left(t - \frac{|\mathbf{x}_0-\mathbf{x}_\mathrm{s}|}{c}\right) . \tag{2.69}$$

secondary_source_selection.m

The window function $w(\mathbf{x}_0)$ for a line source as source model can be calculated after Spors et al. as[54]

$$w(\mathbf{x}_0) = \begin{cases} 1 & (\mathbf{x}_0-\mathbf{x}_\mathrm{s})\mathbf{n}_{\mathbf{x}_0} > 0 \\ 0 & \text{else} \end{cases} \tag{2.70}$$

driving_function_mono_wfs_fs.m
driving_function_imp_wfs_fs.m

Focused Source As mentioned before, focused sources exhibit a field that converges in a focal point inside the audience area. After passing the focal point, the field becomes a diverging one as can be seen in Figure 2.12. In order to choose the active secondary sources, especially for circular or spherical geometries, the focused source also needs a direction $\mathbf{n}_\mathrm{s}$.

The driving function for a focused source are given by the time-reversed versions of the driving functions for a point source as

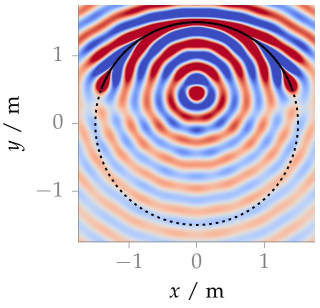


Figure 2.12: Sound pressure of a monofrequent focused source synthesized with 2.5D WFS (2.74). Parameters: $\mathbf{x}_\mathrm{s} = (0, 0.5, 0)\,\mathrm{m}$, $\mathbf{n}_\mathrm{s} = (0,-1,0)$, $\mathbf{x}_\mathrm{ref} = (0,0,0)$, $f = 1\,\mathrm{kHz}$. ☞

$$D(\mathbf{x}_0,\omega) = \\ \frac{1}{2\pi} A(\omega) w(\mathbf{x}_0) \left(i\frac{\omega}{c} + \frac{1}{|\mathbf{x}_0-\mathbf{x}_\mathrm{s}|}\right) \frac{(\mathbf{x}_0-\mathbf{x}_\mathrm{s})\mathbf{n}_{\mathbf{x}_0}}{|\mathbf{x}_0-\mathbf{x}_\mathrm{s}|^2} e^{-i\frac{\omega}{c}|\mathbf{x}_0-\mathbf{x}_\mathrm{s}|} , \tag{2.71}$$

$$D_{2.5\mathrm{D}}(\mathbf{x}_0,\omega) = \\ \frac{g_0}{2\pi} A(\omega) w(\mathbf{x}_0) \sqrt{i\frac{\omega}{c}} \left(1 + \frac{c}{i\omega}\frac{1}{|\mathbf{x}_0-\mathbf{x}_\mathrm{s}|}\right) \frac{(\mathbf{x}_0-\mathbf{x}_\mathrm{s})\mathbf{n}_{\mathbf{x}_0}}{|\mathbf{x}_0-\mathbf{x}_\mathrm{s}|^2} e^{-i\frac{\omega}{c}|\mathbf{x}_0-\mathbf{x}_\mathrm{s}|} , \tag{2.72}$$

or by using an approximated point source as

$$D(\mathbf{x}_0,\omega) = \frac{1}{2\pi} A(\omega) w(\mathbf{x}_0) i\frac{\omega}{c} \frac{(\mathbf{x}_0-\mathbf{x}_\mathrm{s})\mathbf{n}_{\mathbf{x}_0}}{|\mathbf{x}_0-\mathbf{x}_\mathrm{s}|^{3/2}} e^{i\frac{\omega}{c}|\mathbf{x}_0-\mathbf{x}_\mathrm{s}|} , \tag{2.73}$$

$$D_{2.5\mathrm{D}}(\mathbf{x}_0,\omega) = \frac{g_0}{2\pi} A(\omega) w(\mathbf{x}_0) \sqrt{i\frac{\omega}{c}} \frac{(\mathbf{x}_0-\mathbf{x}_\mathrm{s})\mathbf{n}_{\mathbf{x}_0}}{|\mathbf{x}_0-\mathbf{x}_\mathrm{s}|^{3/2}} e^{i\frac{\omega}{c}|\mathbf{x}_0-\mathbf{x}_\mathrm{s}|} . \tag{2.74}$$

As before for other source types, the approximated versions are the default driving functions for a focused source used in this thesis.

[53] Williams, loc. cit.

[54] Spors, Rabenstein, and Ahrens, op. cit.

Transfered to the temporal domain via an inverse Fourier transform (1.2) it follows

$$d(\mathbf{x}_0,t) = \frac{1}{2\pi} a(t) * h(t) * w(\mathbf{x}_0) \frac{(\mathbf{x}_0 - \mathbf{x}_\text{s})\mathbf{n}_{\mathbf{x}_0}}{|\mathbf{x}_0 - \mathbf{x}_\text{s}|^{3/2}} \, \delta\left(t + \frac{|\mathbf{x}_0 - \mathbf{x}_\text{s}|}{c}\right) , \tag{2.75}$$

$$d_\text{2.5D}(\mathbf{x}_0,t) = \frac{g_0}{2\pi} a(t) * h_\text{2.5D}(t) * w(\mathbf{x}_0) \frac{(\mathbf{x}_0 - \mathbf{x}_\text{s})\mathbf{n}_{\mathbf{x}_0}}{|\mathbf{x}_0 - \mathbf{x}_\text{s}|^{3/2}} \, \delta\left(t + \frac{|\mathbf{x}_0 - \mathbf{x}_\text{s}|}{c}\right) . \tag{2.76}$$

In this thesis a focused source always refers to the time-reversed version of a point source, but a focused line source can be defined in the same way starting from (2.66)

$$D(\mathbf{x}_0,\omega) = -\frac{1}{2} A(\omega) w(\mathbf{x}_0) i \frac{\omega}{c} \frac{(\mathbf{x}_0 - \mathbf{x}_\text{s})\mathbf{n}_{\mathbf{x}_0}}{|\mathbf{x}_0 - \mathbf{x}_\text{s}|} \, H_1^{(1)}\left(\frac{\omega}{c} |\mathbf{x}_0 - \mathbf{x}_\text{s}|\right) . \tag{2.77}$$

The window function $w(\mathbf{x}_0)$ for a focused source can be calculated as

secondary_source_selection.m

$$w(\mathbf{x}_0) = \begin{cases} 1 & \mathbf{n}_\text{s}(\mathbf{x}_\text{s} - \mathbf{x}_0) > 0 \\ 0 & \text{else} \end{cases} \tag{2.78}$$

3
Sound Field Errors and their Perceptual Relevance

The theory of sound field synthesis presented so far assumes continuous secondary source distributions. With these, any sound field the dimensionality of the distribution is capable of can be synthesized. The only restriction is that the desired sound field has to be free of both sources and sinks.

Practical setups cannot meet the assumption which underlies these theoretical considerations and will introduce errors in the synthesized sound fields. In this chapter, possible errors will be discussed that are due to different restrictions of secondary source setups. In addition, the perceptual relevance of the different errors will be estimated. For the most relevant errors, perceptual experiments will be carried out, as further described in Chapter 5.

3.1 *Amplitude Errors of 2.5D Synthesis*

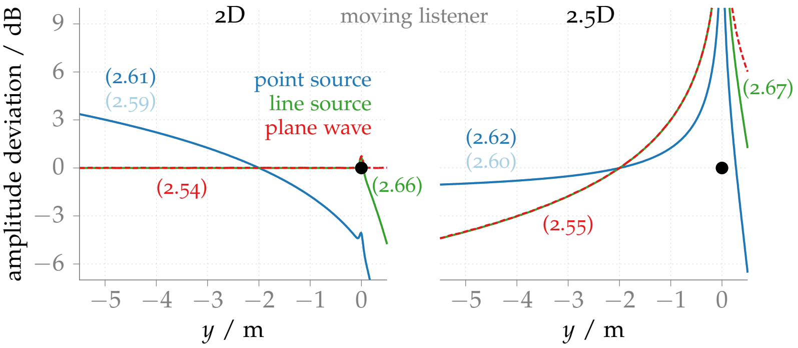


Figure 3.1: Amplitudes of sources synthesized via WFS minus the amplitudes of corresponding real sources dependent on the listener's position along the y-axis. A source is synthesized correctly if its amplitude deviation is 0 dB. An infinite linear secondary source distribution located on the x-axis was used, indicated by the black dot. A comparison between 2D and 2.5D synthesis is shown, with the reference point at $\mathbf{x}_{\text{ref}} = (0, -2, 0)$ m for the 2.5D case. The used driving functions are given within the figure. Parameters: $f = 1$ kHz. ☞

The possible amplitude decays which can be synthesized depend directly on the applied secondary sources. Figure 3.1 shows that a 2D setup is not able to synthesize a point source. The correct amplitude of a point source cannot be synthesized by line sources as secondary sources, as no stronger amplitude decay than that inherent to a line source can be achieved. This solely is a property of the secondary sources and the dimensionality of the setup and thus independent of the applied synthesis method.

Due to the mismatch of the secondary source properties and the dimensionality, deviations of the amplitude are expected in a 2.5D setup for all synthesized sources. Figure 3.1 demonstrates that the deviations for a synthesized line source or plane wave are the strongest ones. Strong deviations for all sources can be observed especially near the position of the secondary sources.

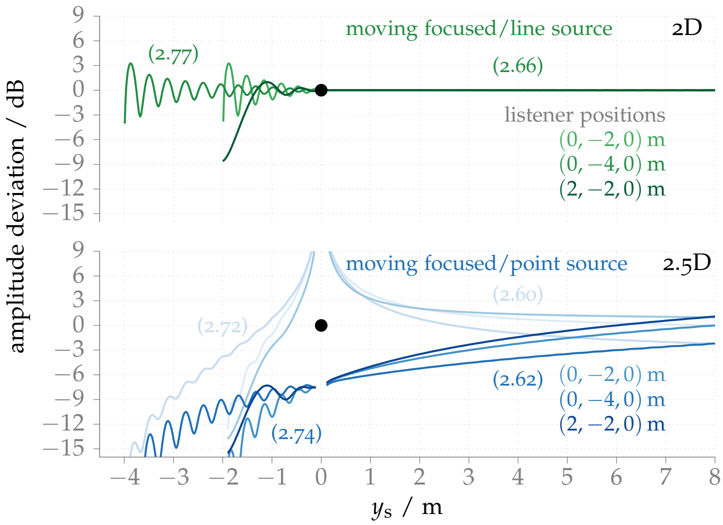


Figure 3.2: Amplitudes of a synthesized point/focused source minus the amplitudes of corresponding real point source located at y_s for three fixed listening positions. The secondary source distribution is located on the x-axis as indicated by the black dot. For positions of the synthesized source with negative y_s values the corresponding focused source models were applied. The used driving functions are indicated within the graphs. For the 2.5D case, two different driving functions are shown whereby the dark blue one is used as default in this thesis. Parameters: $\mathbf{x}_\mathrm{ref} = (0, -2, 0)$ m, $f = 1$ kHz. ☞

In the current considerations, only the listener was moved in the audience area. Another case of amplitude deviations is expected if the listener is at a fixed position and the synthesized source moves. In the 2.5D case it should also matter if the listener is located on the reference point or at another position. Figure 3.2 shows the case of three fixed listener positions in a 2D and a 2.5D setup. The source is positioned 8 m behind the linear secondary source distribution and moves towards the listener until it arrives at the listener position. There are three listener positions, one is at the reference point at $\mathbf{x}_\mathrm{ref} = (0, -2, 0)$ m, one behind that position and one to the left of it. In the 2D case, a line source and the corresponding focused source were synthesized. The line source shows no amplitude deviations. The focused line source exhibits deviations in the form of amplitude ripples. These are not inherent to the 2D setup but originate in the finite length of the secondary source distribution, which had to be used in the numerical simulation.

In the 2.5D case, a point source and the corresponding focused source were synthesized with two different sets of driving functions, as indicated in Figure 3.2. The lighter colors represent the driving functions of a point source and the corresponding focused source and the darker colors the default driving functions for WFS which are approximations under the assumption of $|\mathbf{x}_\mathrm{s} - \mathbf{x}0| \gg 1$. The driving functions without approximation create sources that have large amplitudes near the secondary source distribution and have a strong amplitude decay for focused sources. Völk and Fastl[1] investigated this topic and proposed a correction of the driving function to create a correct amplitude decay at least at the reference point – compare Figure 5 and 6 in their paper. With the default driving functions, as shown by the darker color, the sound field does not exhibit any am-

[1] F. Völk and H. Fastl. "Wave Field Synthesis with Primary Source Correction: Theory, Simulation Results, and Comparison to Earlier Approaches". In: *133rd Audio Engineering Society Convention*. 2012, Paper 8717

plitude increase near the secondary sources. On the other hand, they lead to a sound field that shows an overall decay of the amplitude, which is too strong for an approaching source leading to a deviation of more than $-15\,\mathrm{dB}$ for the simulated geometry. Away from the reference point, the amplitude has a similar behavior except for the small offset.

Distance Perception For 2.5D synthesis amplitude deviations of a synthesized source can have an influence on the perceived distance of the synthesized source for 2.5D synthesis. In the simulation of a moving source, which is approaching the listener, the deviations in amplitude could lead to an impression of a slower approaching source, because the amplitude is not increasing strong enough, the closer the source moves to the listener. In the simulation with a fixed source position and a moving listener, the source could be perceived as also moving due to its wrong change of amplitude.

Völk has asked listeners to judge the distance of synthesized point sources in WFS.[2] The listeners had a fixed position while the synthesized source was moved. He applied a driving function with the same amplitude behavior as (2.62) – see Figure 3.2. The results for the synthesized sources are comparable to the results obtained for real sources showing underestimation of larger distances.[3] His results suggest that while having errors in the produced amplitude in 2.5D sound field synthesis, the perception of the source is not modified by these errors.

For nearby sources located closer than 0.5 m to the head, changes of the ILD are an important distance cue.[4] Kerber et al. have shown that WFS is not able to recreate the ILD cue for nearby sources.[5] Otherwise he stated that this might not be so critical if the listener is allowed to move within the audience area. This was confirmed in an experiment by Müller et al. in which listeners were supposed to move within the audience area, find the position of the focused source and grab it.[6] The results showed that listeners succeed with a precision of around 15 cm.

Besides the distance cues discussed so far, the direct-to-reverberant energy ratio plays a large role for distance perception inside of rooms.[7] Results from the literature indicate that due to the reflections of the signals from the secondary sources, the direct-to-reverberant energy ratio will be different for WFS and possibly influences the perception.[8]

THE MAJORITY OF experiments of distance perception in SFS concentrated on a single position of the listener. One of the goals of this thesis is to investigate the perception in an area as large as possible. This suggests to use the method of binaural synthesis – presented in the next chapter – to investigate the perception of distance. In a critical manner the perceived distance is highly influenced by the binaural simulation.[9] As a consequence, distance perception is not considered in this thesis.

[2] F. Völk. "Psychoakustische Experimente zur Distanz mittels Wellenfeldsynthese erzeugter Hörereignisse". In: *36th German Annual Conference on Acoustics*. 2010, pp. 1065–66.

[3] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst. "Auditory Distance Perception in Humans: A Summary of Past and Present Research". *Acta Acustica united with Acustica* 91 (2005), pp. 409–20.

[4] D. S. Brungart and W. M. Rabinowitz. "Auditory localization of nearby sources. Head-related transfer functions." *The Journal of the Acoustical Society of America* 106.3 Pt 1 (1999), pp. 1465–79.

[5] S. Kerber et al. "Experimental investigations into the distance perception of nearby sound sources: Real vs. WFS virtual nearby sources". In: *Proceedings of the Joint Congress CFA/DAGA*. 2004, pp. 1041–42.

[6] J. Müller et al. "The BoomRoom: Mid-air Direct Interaction with Virtual Sound Sources". In: *Conference on Human Factors in Computing Systems*. 2014.

[7] A. W. Bronkhorst and T. Houtgast. "Auditory distance perception in rooms." *Nature* 397.6719 (1999), pp. 517–20.

[8] F. Völk, M. Straubinger, and H. Fastl. "Psychoacoustical experiments on loudness perception in wave field synthesis". *20th International Congress on Acoustics* (2010).

[9] W. M. Hartmann and A. Wittenberg. "On the externalization of sound images." *The Journal of the Acoustical Society of America* 99.6 (1996), pp. 3678–88

3.2 *Diffraction and Truncation of Secondary Source Distributions*

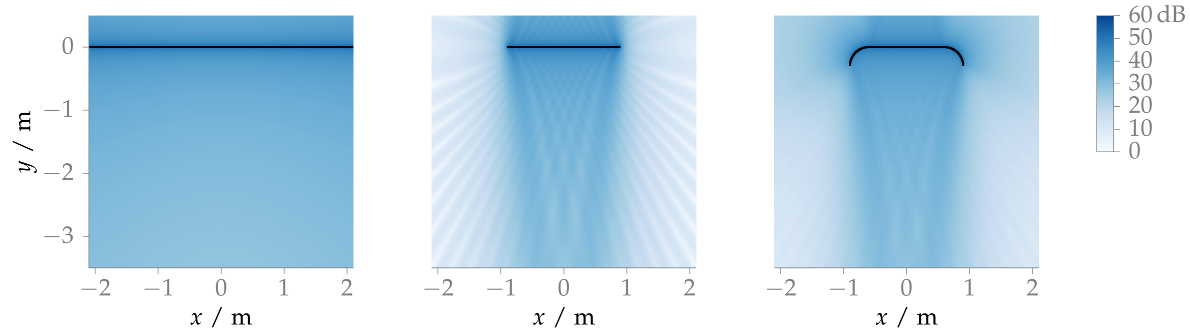


Figure 3.3: Sound pressure in decibel of a plane wave synthesized with 2.5D WFS (2.55). The result of an infinite linear secondary source distribution is compared with two truncated ones. Parameters: $\mathbf{n}_k = (0, -1, 0)$, $\mathbf{x}_{\text{ref}} = (0, -2, 0)$ m $f = 3$ kHz. ☞

The solutions for a linear secondary source distribution assume an infinite length of the distribution, which typically is violated in a real-life setup, where lengths of around 3 m are common. This provokes errors in the synthesized sound field that can be described by diffraction theory. The linear source distribution can be thought of as a slit where a wave field coming from the other side has to go through. The slit can be modelled as a rectangle window, the corresponding diffraction pattern is that of a sinc function as shown in Ahrens.[10] Figure 3.3 displays the sound pressure level of a synthesized plane wave going into the direction $(0, -1, 0)$ for an infinite linear secondary source distribution and a distribution with a length of 1.8 m. The array's form influences the diffraction pattern as can be seen for the secondary source distribution to the right. In this case, it is covering 1.8 m like in the case depicted in the middle, but the edges of the array are bend towards the listening area. In this way the diffraction pattern is pronounced to a lower degree than before, emphasizing that the edges have a large impact on the diffraction.

This can be further highlighted with an equivalent description of the diffraction by so called edge waves. Consider that the length of the secondary source distribution is large compared to the wave length of the sound, but small compared to the distance between the source position $\mathbf{x}_0$ and the receiver position $\mathbf{x}$. In addition, the incidence angle of the sound is approximately vertical to the secondary sources. In this case the problem can be approximated by *Kirchhoff's diffraction theory*.[11] Here, the diffraction can be explained in an equivalent manner by a super-position of the incident sound field and two spherical waves originating from the edges of the array – this is well summarized in Born and Wolf.[12]

The same was found for WFS by Verheijen[13] and by Start[14] – where the diffraction for WFS is treated in more mathematical detail. They introduced a so called tapering window to attenuate the edge waves by reducing the amplitude of the secondary sources at the edges with cosine windows. For investigating the influence of the window function, Ahrens' approach to transform the problem into the k_x-domain can be useful.[15]

[10] Ahrens, op. cit., (3.87).

[11] M. Born et al. *Principles of Optics*. Cambridge University Press, 1999, Sect. 8.3.2.

[12] Ibid., Sect. 8.9.

[13] Verheijen, op. cit.

[14] E. W. Start. "Direct Sound Enhancement by Wave Field Synthesis". PhD thesis. Technische Universiteit Delft, 1997.

[15] Ahrens, op. cit., Sect. 3.7.4.

The edge waves and their reduction by the tapering window are shown in Figure 3.4. There, three cosine shaped pulses are synthesized as plane waves traveling downwards. Looking at the amplitudes for the line at $x = 0\,\mathrm{m}$ parallel to the y-axis the influence of the tapering can be estimated. In the left figure the level of the edge wave is approximately 20 dB lower than that of the desired wave front. In the right figure, including the tapering window, the edge wave is attenuated by 10 dB and thereby is 30 dB below the desired wave front.

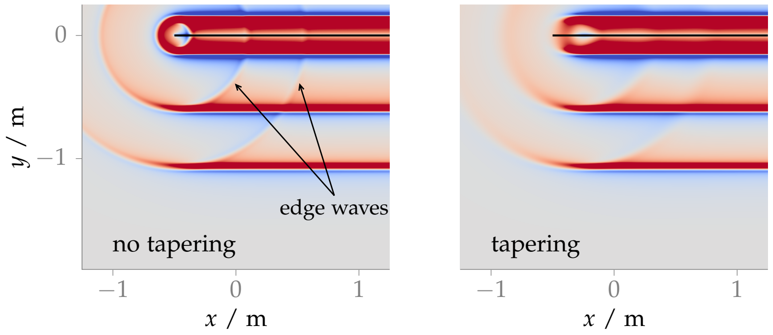


Figure 3.4: Sound pressure of three cosine shaped broad-band pulses synthesized as plane waves with WFS (2.57). Additional edge waves are visible due to diffraction. By applying a tapering window to the last 30 cm of the secondary source distribution the edge waves could be damped, as is shown in the right graph. Parameters: $\mathbf{x}_\mathrm{s} = (0, -1, 0)$, $\mathbf{x}_\mathrm{ref} = (4.5, -2, 0)\,\mathrm{m}$, $t = 4.6\,\mathrm{ms}$. ☞

Diffraction occurs also for non-smooth array contours that include a corner. In this case the level of an additional diffraction wave is negligible, as shown in Verheijen[16] and Ahrens.[17]

Another influence of the truncation is of interest only for focused sources. Size and shape of the secondary source distribution have an influence on the extent of the focal point. The extent of the focal point should be given by the first two minima's distance in the diffraction pattern. With an infinite linear secondary source distribution the size of the focal point is λ.[18] However, for a truncated array the focal point can get larger. Under the assumption of *Fraunhofer diffraction*,[19] the size can be calculated as the distance between the first zeros of the diffraction pattern. The first zeros result for a path difference of λ of the two waves originating from the slit's edges. Figure 3.5 shows the setting for a linear secondary source distribution located on the x-axis. The angle α is obviously given by $\sin\alpha = \lambda/L$, and the focal point size for a linear secondary source distribution is then given as[20]

$$\Delta_\mathrm{s} = \left[2|y_\mathrm{s} - y_0| \tan\left(\sin^{-1}\frac{\lambda}{L}\right), \lambda\right]_\mathrm{max}, \tag{3.1}$$

where L is the length of the truncated secondary source distribution, and $\lambda = c/f$ the wave length. In Figure 3.6 the size of the focal point is shown for different lengths of a linear secondary source distribution and a wave length of $\lambda = 0.172\,\mathrm{m}$. The size of the focal point calculated with (3.1) is 0.47 m, 0.19 m, 0.17 m going from the smallest distribution to the largest one – from left to right. The real sizes that can be examined by analyzing the amplitude distribution in the

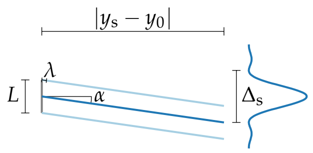


Figure 3.5: Size of the focus point Δ_s as given by Fraunhofer diffraction. ☞

[16] Verheijen, op. cit., Fig. 2.22.

[17] Ahrens, op. cit., Fig. 3.26 and 3.27.

[18] E. Abbe. "III.—Some Remarks on the Apertometer". *Journal of the Royal Microscopical Society* 3 (1880), pp. 20–31, formula on p. 26 with $n = 1$ and $\omega = \pi/2$; the result is $\lambda/2$ for a microscope, because the resolution is defined as the distance between the first maximum and minimum of the diffraction pattern.

[19] Born et al., op. cit., Sect. 8.3, (34). Diffraction for a focal point is equivalent to diffraction for plane waves coming from infinity and focus them by a lens – compare Figure 8.6. For plane waves from infinity eq. 34 is automatically fullfilled.

[20] H. Wierstorf et al. "Perception of Focused Sources in Wave Field Synthesis". *Journal of the Audio Engineering Society* 61.1 (2013), pp. 5–16, (13).

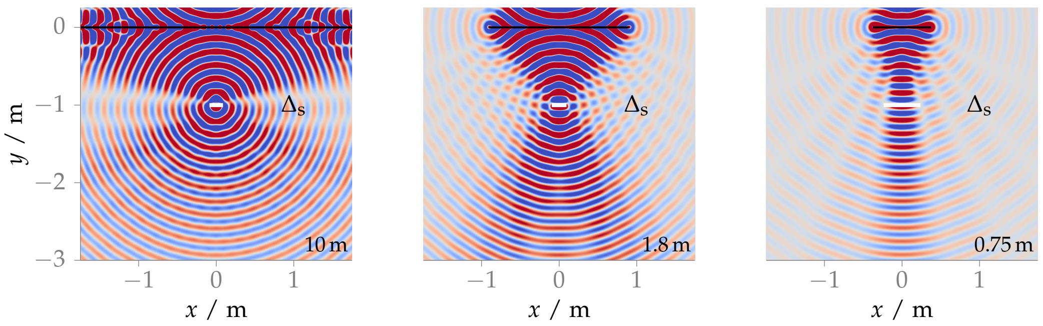


Figure 3.6: Sound pressure of a focused source synthesized with 2.5D WFS (2.74). Three different linear secondary source distributions were applied, ranging from 10 m to 0.75 m. In white the size of the focal point as given by $\Delta_{\text{s, linear}}$ is indicated. Parameters: $\mathbf{x}_\text{s} = (0, -1, 0)$ m, $\mathbf{x}_\text{ref} = (0, -2, 0)$ m, $f = 2000$ Hz. ☞

focal plane of the simulated sound field are 0.50 m, 0.24 m, 0.14 m. The result shows that a smaller secondary source distribution leads to a wider focal point. In addition, the focal point will become even larger for lower frequencies, as can directly be seen from (3.1). Note that the focal point size can also be calculated with (24) from Lucas and Muir[21] which generates similar results.

For a circular secondary source distribution in most cases the focal point size can be approximated by λ. The concave shape of the distribution allows a better focussing of the sound field.

The right graph of Figure 3.6 highlights another phenomenon for small secondary source distributions and low frequencies: the focal point shifts towards the secondary sources. This is discussed in detail by Oldfield,[22] who also proposes a method to correct the shift for the synthesis of focused sources.

Perception After discussing the various physical influences of truncating the secondary source distribution, in the following a short summary of its potential influence on the perception is presented.

The strongest influence of truncating the secondary source distribution is on the size of the listening area which becomes smaller dependent on the exact size of the distribution and the synthesized source – compare Figure 3.3 and 3.6. If the listener is placed at the border of the listening area, the truncation could influence the localization of a synthesized source. Having the center of the head exactly at the border will lead to one ear in the listening area and one ear being out of it, provoking a possibly large ILD between the two ears. Depending on the amount of level difference and uncertainty in the other localization cues, the ILD could dominate the localization perception. The perceived source direction then is very likely to be in another direction than the one desired for the synthesized source. For a mono-frequent source with low frequency a wrong ILD could also be possible within the listening area because of the maxima and minima of the diffraction pattern.

The diffraction due to the truncation by the secondary source distribution can be described in an equivalent way by edge waves starting at the edges of the distribution. For a listener at a certain position within the listening area this means that she will at first hear the desired sound, and shortly afterwards one or two reflections of

[21] B. G. Lucas and T. G. Muir. “The field of a focusing source”. *The Journal of the Acoustical Society of America* 72.4 (1982), pp. 1289–96

[22] R. Oldfield. “The analysis and improvement of focused source reproduction with wave field synthesis”. PhD thesis. University of Salford, 2013, Sect. 4.9.

the same sound coming from the edges of the distribution. If the distribution is not larger than 4 m, this probably does not have any influence on the localization, due to the precedence effect[23] which ensures a domination of the perceived direction by the first arriving sound. However, the additional edge waves will add some sort of coloration to the perceived sound. The edge waves can easily be damped by applying a tapering window. As this effect can easily be avoided, no perceptual experiments were carried out to investigate the influence of the edge waves.

Beside the fact of wrong ILD cues at the margins of the listening area the truncation can have further influence on the localization for focused sources. As shown in Figure 3.6 the size of the focal point depends on the size of the secondary source distribution. A larger focal point will most probably widen the perceived source width. If size of the focal point further increases, it may further be shifting the perceived position of the focused source towards the secondary source distribution.

To investigate the influence of the truncation on the localization of focused sources, a localization experiment for focused sources synthesized with different secondary source distributions will be presented in Chapter. 5.

3.3 *Spatial Aliasing and Discrete Secondary Source Distributions*

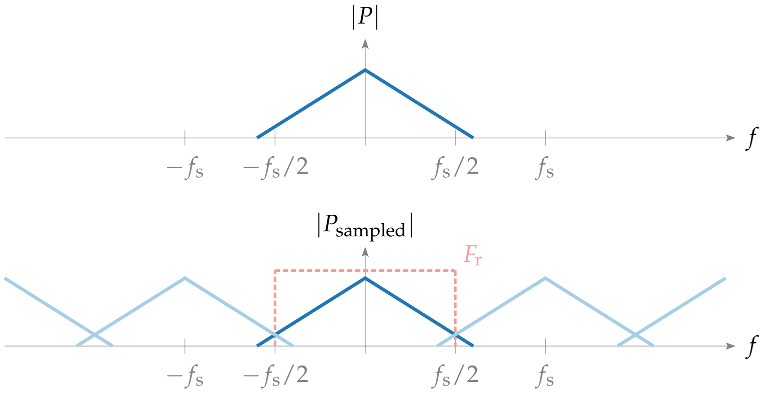


Figure 3.7: Magnitude of a continuous signal P and the same signal sampled with a sampling frequency of f_s. The light blue lines indicate components occurring due to the sampling process. F_r describes an ideal reconstruction filter for the sampled signal. ☞

Up to now, only continuous secondary source distributions have been considered, that can hardly be built up in practice. Normally, an array of loudspeakers is applied which corresponds to a sampling of the secondary source distribution in space. This leads to different impacts on the synthesized sound field.

For a better understanding of the phenomenon a comparison of the temporal sampling of a signal is of interest. Consider a signal $p(t)$ and its Fourier transform $P(\omega)$ with $\omega = 2\pi f$. The signal will be sampled with a sampling frequency of f_s which means that only its values at multiples of $\Delta t = 1/f_s$ are considered. Figure 3.7 illustrates the consequences of the sampling process. The sampled signal

[23] A good summary is provided in R. Y. Litovsky et al. "The precedence effect." *The Journal of the Acoustical Society of America* 106.4 (1999), pp. 1633–54.

P_{sampled} now includes spectral repetitions of the original signal at each multiple of f_s. If the original signal should be recreated from the sampled one, a reconstruction filter F_r as indicated by the red line in Figure 3.7 has to be applied to the signal. If the sampled signal contains frequencies larger than $f_\text{s}/2$, they will overlap and interfere with the base band and the signal will be corrupted by *aliasing*. Thus, the frequency $f_\text{s}/2$ can be defined as the aliasing frequency f_al depending on the distance of the sample points as

$$f_\text{al} = \frac{1}{2\Delta t} \, . \tag{3.2}$$

Now, the same problem with a driving function $D(x_0)$ and a one-dimensional secondary source distribution is considered. In this case the secondary source distribution will be sampled at multiples of $\Delta x_0 = 2\pi/k_x$ and spectral repetitions of the sampled driving function will occur in the k_x domain. Due to the dispersion relation $k^2 = (\omega/c)^2$ a spatial aliasing frequency here can be specified as

$$f_\text{al} = \frac{c}{2\Delta x_0} \, , \tag{3.3}$$

where c is the speed of sound.

The secondary sources themselves can be considered as the reconstruction filter in this case. Because of the dispersion relation only their propagating parts are band-limited in the spatial domain up to a given frequency f. Both spatial aliasing due to interference of repetitions with the base band and reconstruction errors due to the suboptimal suppression of spatial repetitions can occur in the sound field. For some examinations it could be useful to distinguish between these two cases,[24] but for the perceptually relevant aspects discussed in this thesis they will be subsumed jointly under the term spatial aliasing. Hence, the spatial aliasing frequency as used here covers both cases.

In three dimensions with $\mathbf{k} = (k_x, k_y, k_z)$ and $\mathbf{x}_0 = (x_0, y_0, z_0)$ the aliasing frequency specified by (3.3) is only a lower boundary. Now the amount of spatial aliasing will become dependent on the position and on the type and position of the synthesized source. In the following, the dependency on the position of the listener is explained by an example. Consider a linear secondary source distribution and a listener facing the distribution. Assuming the listener is as close as possible to the secondary source distribution standing between two individual secondary sources. In this case, the sound could reach her from directions ranging from $-90°$ to $90°$ relative to her head orientation. If the listener is as far away from the distribution as possible the sound reaching her ears from the same two individual secondary sources arrives from the same direction of $0°$. That means with increasing distance of the listener from the secondary source distribution less spatial frequencies are needed to represent the signal at the listener position.

[24] Ahrens, op. cit., Chap. 4 deals with the discretization of different secondary source distribution in great detail.

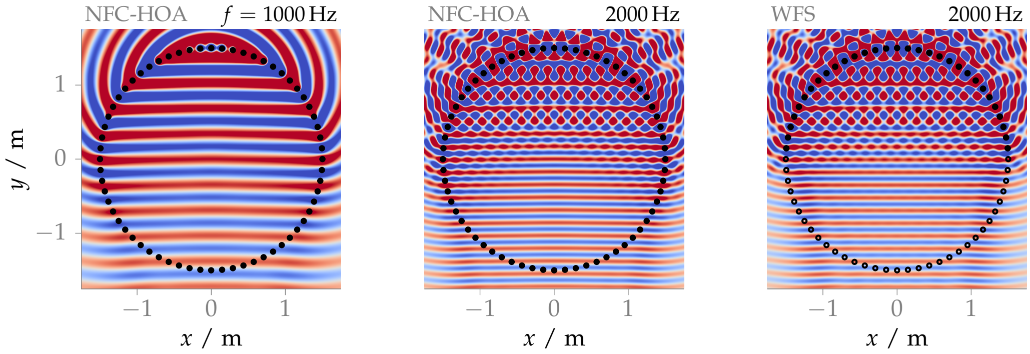


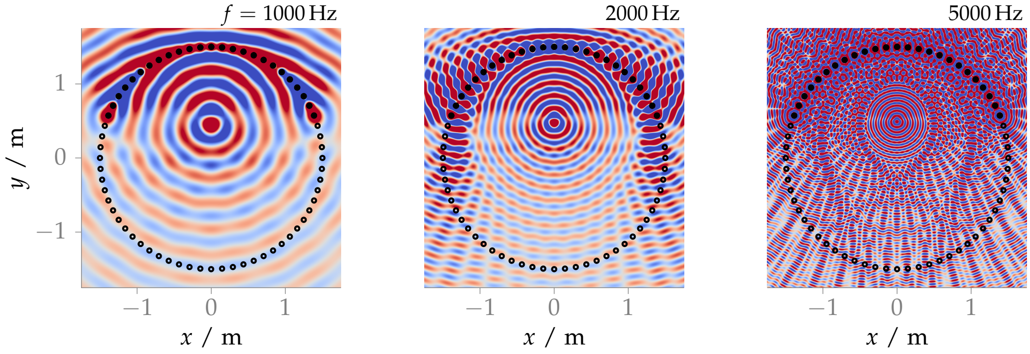
Figure 3.8: Sound pressure of a plane wave synthesized by NFC-HOA (2.45) and WFS (2.55) for different frequencies. For WFS the open circles indicate inactive secondary sources. Parameters: $\mathbf{x}_\text{s} = (0, -1, 0)$, $\mathbf{x}_\text{ref} = (0, 0, 0)$ m, 64 secondary sources. ☞

Figure 3.8 shows aliasing phenomena for NFC-HOA and WFS for a circular secondary source distribution. The secondary source distribution is sampled at 64 discrete points leading to a lower boundary of the aliasing frequency of 1165 Hz. This can be observed in the figure as well. For a frequency of 1000 Hz the synthesized sound field will be free from aliasing, but for a frequency of 2000 Hz aliasing occurs at points near the secondary source distribution. Furthermore, it can be observed that the pattern of aliasing is nearly identical in the case of NFC-HOA and WFS.

As mentioned before the amount of spatial aliasing depends on the type and position of the synthesized source. This is due to the fact that for most of the source types and positions the sampling of the secondary source distribution is irregular in space. In this case nonuniform sampling theory has to be applied to determine the right spatial aliasing frequency. There are suggestions how to calculate it for the synthesis of point sources[25] and for focused sources,[26] but no general solution has been presented so far, and is out of scope for this thesis. If the aliasing frequency is needed for a defined configuration, it will be determined by numerical simulation of the situation and inspection of the spectrum – compare Figure 3.12. However, in most cases it will be sufficient to use (3.3) as an approximation of the aliasing frequency. Only for a synthesized focused source special considerations are necessary.

Figure 3.9 shows the amplitude of the sound field of a focused source placed at $(0, 0.5, 0)$ m synthesized with 2.5D WFS. It indicates that for high frequencies there always is a circular region around the

Figure 3.9: Sound pressure of a focused source synthesized by WFS (2.74) for different frequencies. Parameters: $\mathbf{x}_\text{s} = (0, 0.5, 0)$, $\mathbf{x}_\text{ref} = (0, 0, 0)$ m, 64 secondary sources. ☞



[25] E. Corteel. "On the use of irregularly spaced loudspeaker arrays for wave field synthesis, potential impact on spatial aliasing frequency". In: *International Conference on Digital Audio Effects*. 2006.

[26] Oldfield, op. cit.

focal point where no aliasing occurs. This property of a correct synthesis in a small region with the help of a focused source has already been used in different applications where a small but aliasing-free region should be synthesized.[27] For a linear secondary source distribution located on the x-axis the radius r_{al} of the aliasing-free zone was empirically found in Wierstorf et al[28] as

$$r_{\text{al}} = \frac{y_{\text{s}} c}{f \Delta x_0} \, . \tag{3.4}$$

The idea of a spatially limited but aliasing free region can be applied in NFC-HOA in a more direct way by limiting the spatial bandwidth of the driving function. The lower orders of the Bessel functions contribute to a higher degree to the sound field in the center of the secondary source distribution, whereas higher orders contribute more to positions at distances far from the center.[29] Hence, a correctly synthesized region in the center of the secondary source distribution for a limited order is expected. In order to avoid spatial aliasing the maximum order M should be smaller or equal to π/x_0. For a circular secondary source distribution the maximum order without spatial aliasing is then given as

nfchoa_order.m

$$M \leq \begin{cases} (N_{\text{s}}-1)/2 & \text{for even } N_{\text{s}} \\ N_{\text{s}}/2 & \text{for odd } N_{\text{s}} \, , \end{cases} \tag{3.5}$$

where N_{s} is the number of secondary sources.

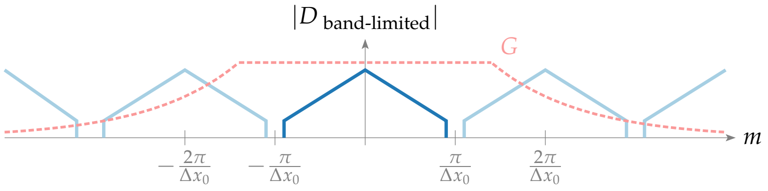


Figure 3.10: Magnitude of a mono-frequent, spatial band-limited NFC-HOA driving function. The light blue lines indicate components occurring due to the sampling process. G describes the reconstruction filter. ☞

Figure 3.10 shows the effect of band-limiting. Here, the spectral repetitions of the driving function no longer interfere with each other. On the other hand, the reconstruction filter – the Green's function G – in general is not band-limited. Above a certain frequency spectral repetitions will be part of the synthesized sound field. Figure 3.11 presents the amplitude distribution of the sound field for

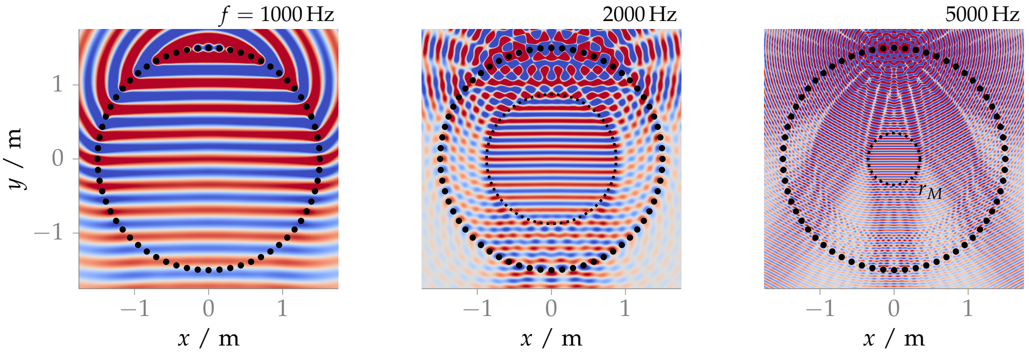


Figure 3.11: Sound pressure of a plane wave synthesized by NFC-HOA (2.45) for different frequencies. The maximum order M was set to be 32 after (3.5). The region of correct synthesis is given by $r_M = M c/\omega$ as indicated by the dotted line. Parameters: $\mathbf{x}_{\text{s}} = (0,-1,0)$, $\mathbf{x}_{\text{ref}} = (0,0,0)$ m, 64 secondary sources. ☞

[27] E.g. S. Spors and J. Ahrens. "Local Sound Field Synthesis by Virtual Secondary Sources". In: *40th Audio Engineering Society Conference*. 2010, Paper 6.3.

[28] Wierstorf, Raake, and Spors, op. cit., (12).

[29] Ahrens, op. cit., Sect. 2.2.2.

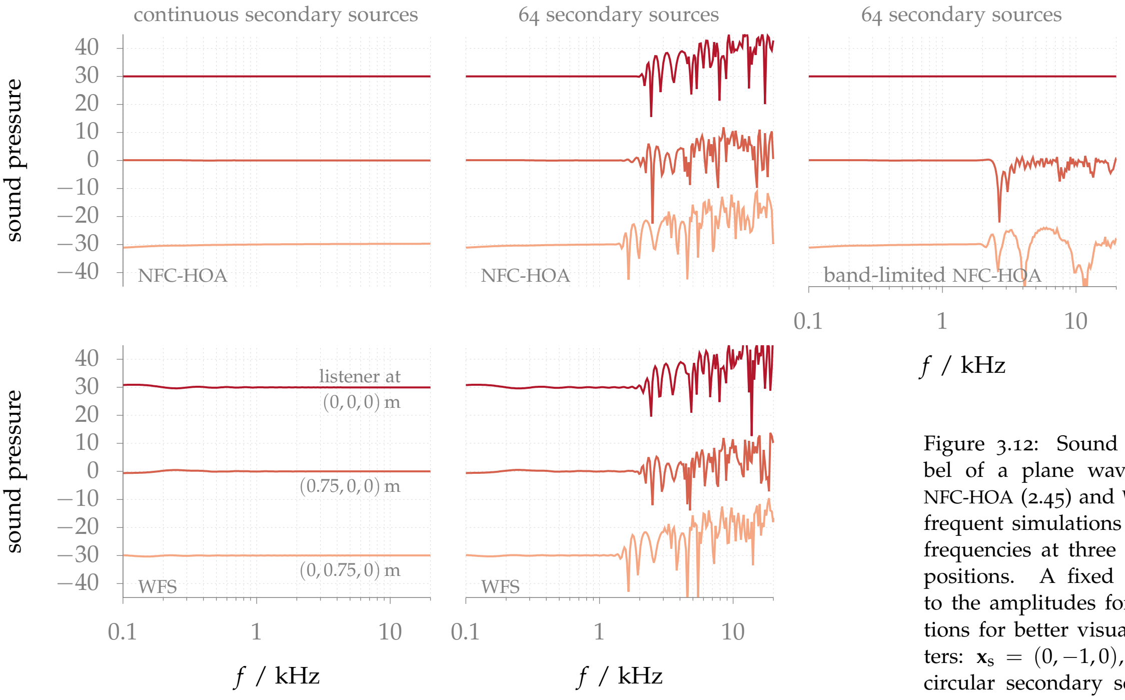


Figure 3.12: Sound pressure in decibel of a plane wave synthesized by NFC-HOA (2.45) and WFS (2.55). Monofrequent simulations were done for all frequencies at three different listening positions. A fixed offset was added to the amplitudes for two of the positions for better visualization. Parameters: $\mathbf{x}_\text{s} = (0, -1, 0)$, $\mathbf{x}_\text{ref} = (0, 0, 0)$ m, circular secondary source distribution with a diameter of 3 m. ☞

a plane wave with direction $(0, -1, 0)$ synthesized by 2.5D band-limited NFC-HOA using a circular secondary source distribution with 64 sources and a radius of 1.5 m. For a frequency of 1 kHz the sound field is synthesized without errors, while for 2 kHz and 5 kHz only a region in the center of the distribution is synthesized correctly. The size of that region is given as[30]

$$r_M = \frac{Mc}{\omega} \,. \tag{3.6}$$

As a result of reconstruction errors outside of the r_M region, spatial aliasing occurs. In contrast to the spatial aliasing in the synthesized sound field for a full-band synthesis as shown in Figure 3.8 the reconstruction errors introduce amplitude fluctuations outside of the r_M region.

THE INFLUENCE of the spatial aliasing on the signals at a given listener position can further be analyzed by the corresponding temporal-frequency spectrum at a given position. Figure 3.12 shows the magnitude of the spectrum at three different listener positions. Synthesizing a plane wave with a circular secondary source distribution with a radius of 1.5 m in all cases, a continuous secondary source distribution or a sampled one with 64 sources was driven by WFS, NFC-HOA or band-limited NFC-HOA. For the continuous distribution for all listener positions the spectrum is flat for NFC-HOA and WFS. Only for frequencies below 400 Hz some slight deviations are visible for WFS. That is not surprising as WFS is a high-frequency approximation of NFC-HOA.

The sampling with only 64 secondary sources introduces aliasing artifacts at all listener positions. The figure indicates that these

[30] Compare (9.1.31) and Fig. 9.5 of Gumerov and Duraiswami, op. cit.

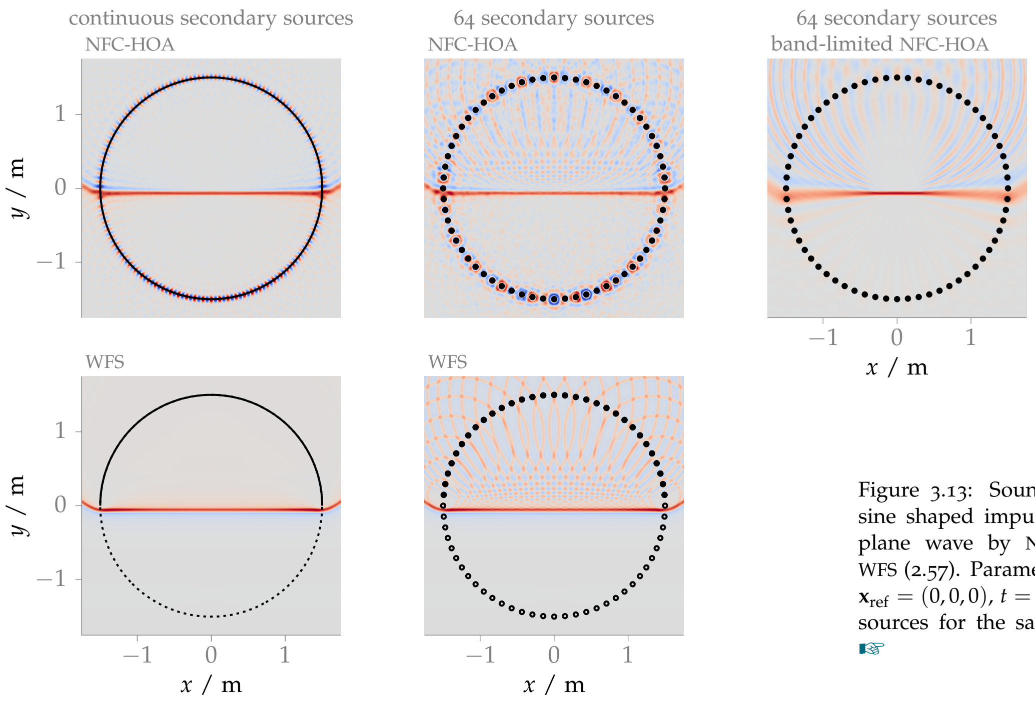


Figure 3.13: Sound pressure of a cosine shaped impulse synthesized as a plane wave by NFC-HOA (2.45) and WFS (2.57). Parameters: $\mathbf{x}_\mathrm{s} = (0, -1, 0)$, $\mathbf{x}_\mathrm{ref} = (0, 0, 0)$, $t = 4.6\,\mathrm{ms}$, 64 secondary sources for the sampled distributions. ☞

artifacts appear only above a given aliasing frequency and are position dependent. The lower limit of the aliasing frequency can be calculated with (3.3) as 1165 Hz which is exceeded by the aliasing frequency observed in Figure 3.12 at all three of the investigated listener positions. The aliasing not only introduces fluctuations to the magnitude spectrum but in addition adds energy to the spectrum with a slope of 3 dB per octave. While this slope is identical to the slope of the pre-equalization filter (2.57) it is common practice in WFS to apply this filter only up to the aliasing frequency. The experiments presented in this thesis also use this practice.

In addition to NFC-HOA, the sampled secondary sources were also driven by band-limited NFC-HOA. As expected, this results in no aliasing for the central listening position. Outside of the center the problem remains, although the pattern of deviations in the spectrum differs and the 3 dB slope is missing completely.

The spatial aliasing artifacts are not only visible in the spectrum of the synthesized sound field, but also manifest themselves in the temporal signals at a fixed listener position.

The temporal signals of the sound field provide additional insights on the properties of the spatial aliasing artifacts. A broad band pulse is synthesized and after some time t the sound field is frozen and plotted. Figure 3.13 demonstrates the synthesized sound field of a ten samples long, Hann-window shaped pulse after $t = 4.6\,\mathrm{ms}$ for different setups. The top row represents 2.5D NFC-HOA for a continuous secondary source distribution and a discrete one with 64 sources. For the latter in addition the result for NFC-HOA is shown in the graph to the right. The bottom row shows the same sound fields synthesized with 2.5D WFS. Band-limited WFS has not been investigated so far and will be omitted in this thesis.

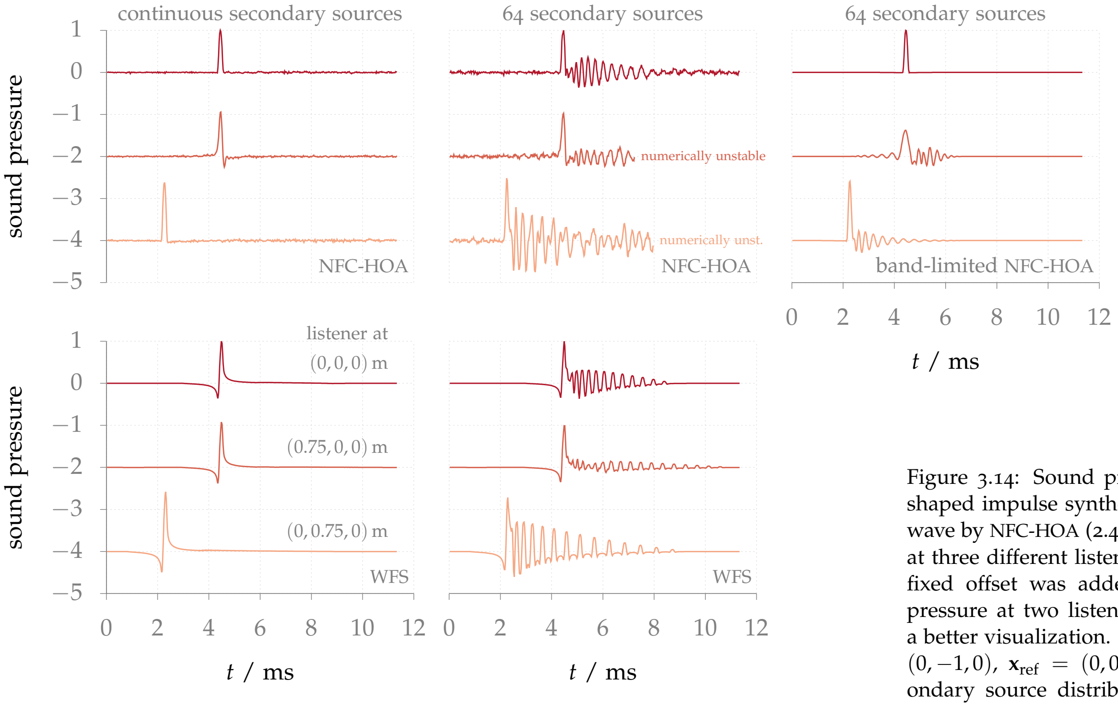


Figure 3.14: Sound pressure of cosine shaped impulse synthesized as a plane wave by NFC-HOA (2.45) and WFS (2.57) at three different listening positions. A fixed offset was added to the sound pressure at two listening positions for a better visualization. Parameters: $\mathbf{x}_\text{s} = (0, -1, 0)$, $\mathbf{x}_\text{ref} = (0, 0, 0)$, circular secondary source distribution with a diameter of 3 m. ☞

The results for the full-band cases are similar between NFC-HOA and WFS. For a continuous secondary source distribution the pulse is synthesized without errors. For discrete secondary sources additional wave fronts are visible which arrive after the desired pulse and have a similar pattern for both methods. The large magnitudes near the secondary sources in the case of NFC-HOA are most likely due to numerical problems. These are inherent to the calculation of higher order components,[31] which should be most prominent near the secondary sources. For the band-limited NFC-HOA case the pulse is synthesized correctly in the center of the distribution, but again has additional signal components outside of the center.

A further investigation of the exact temporal pattern of the wave fronts at three different listener positions is presented in Figure 3.14. Here, the following positions were chosen: one at the center $(0, 0, 0)$ m, one in the frontal part of the audience area $(0, 0.75, 0)$ m, and one in the right part of the audience area $(0.75, 0, 0)$ m. The figure reveals only few errors in the temporal pattern for the synthesis of the plane wave with a continuous secondary source distribution. Additionally in this case, for all three positions the Hann-window shaped pulse is synthesized correctly. In the case of NFC-HOA the amplitude shows more noise than for WFS, again likely due to numerical limitations in the calculation of the NFC-HOA driving functions. WFS constitutes a high-frequency approximation of the exact SFS solution, which explains the slight undershoots at the beginning of the impulse for WFS.

For a sampled secondary source distribution, errors are visible in all cases. Now additional positive and negative pulses are evident after the first one. Again the pattern is very similar for WFS and NFC-HOA. For the latter, numerical instabilities could not be solved for parts of the signal. Hence, parts of it are omitted in the figure.

[31] In order to calculate orders higher than 85 at all the Multiprecision Computing Toolbox for Matlab was applied for finding the zeros of the Bessel function – compare `sphbesselh_zeros.m`

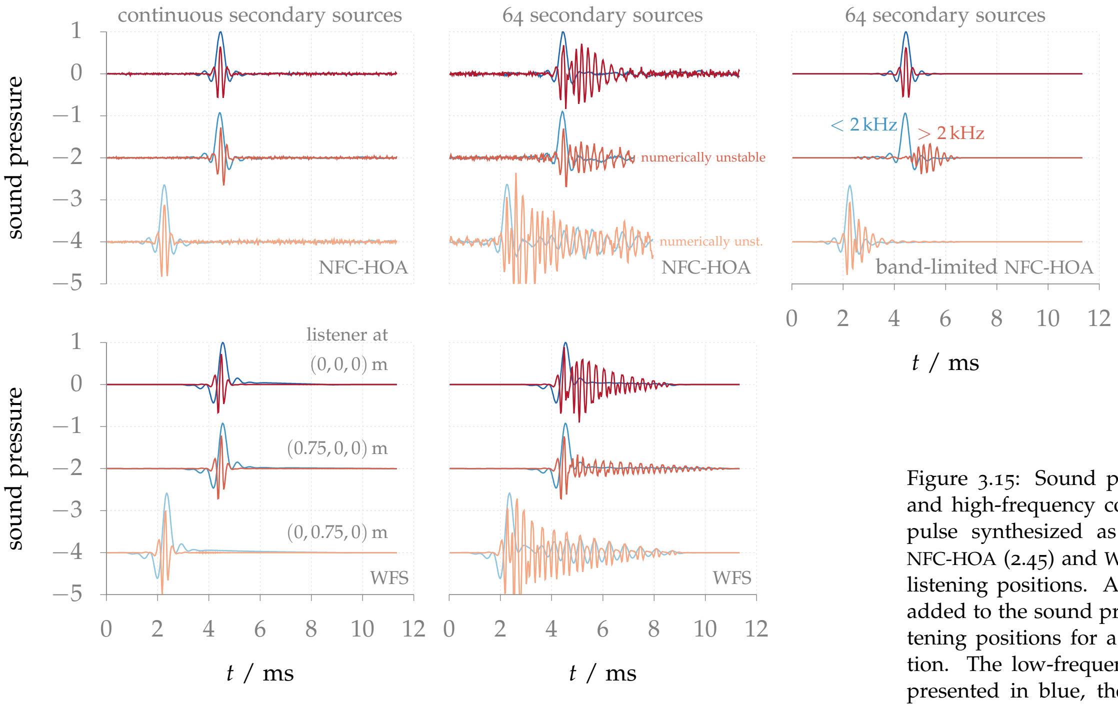


Figure 3.15: Sound pressure of a low and high-frequency cosine shaped impulse synthesized as plane wave by NFC-HOA (2.45) and WFS (2.57) at three listening positions. A fixed offset was added to the sound pressure at two listening positions for a better visualization. The low-frequency impulses are presented in blue, the high-frequency impulses in red. Parameters: $\mathbf{x}_\mathrm{s} = (0, -1, 0)$, $\mathbf{x}_\mathrm{ref} = (0, 0, 0)\,\mathrm{m}$, circular secondary source distribution with a diameter of 3 m. ☞

The time window during which the additional wave fronts arrive at the listener is clearly dependent on the position of the listener. They arrive in a time frame of 4 ms for the central position, increasing to 6 ms for the frontal position. Further simulations, not shown here, reveal that the number of additional wave fronts is directly proportional to the number of secondary sources involved. This is not surprising, because every single loudspeaker is emitting one of the wave fronts. However, when the number of sources is increased up to approaching the continuous case, the individual supplementary pulses become increasingly smaller. The length of the time window during which the additional wave fronts arrive depends on the geometry of the secondary source distribution. The larger the employed distribution, the longer the time window will be – compare Figure 5.15.

For band-limited NFC-HOA the synthesis is correct at the center position, as expected by inspecting Figure 3.11. At the frontal and side position, small errors are visible mainly after the desired wave front.

An interesting question is how different frequencies are distributed in the unwanted wave fronts. Investigating this in Figure 3.15, a low frequency pulse with $f < 2\,\mathrm{kHz}$ and a high frequency pulse with $f > 2\,\mathrm{kHz}$ are synthesized. It is evident that for the case of a continuous secondary source distribution, both the low and the high frequencies are synthesized at the same time. For the sampled secondary sources and the case of WFS and NFC-HOA, a common onset of both pulses is visible, but the additional wave fronts contain only high frequencies. The explanation is that frequencies below the aliasing frequency are synthesized correctly, and errors in the synthesized sound field are expected only for frequencies above the aliasing

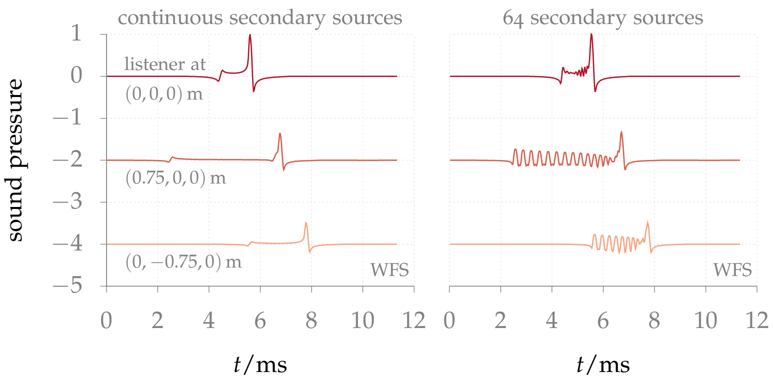


Figure 3.16: Sound pressure of a cosine shaped impulse synthesized as focused source by WFS (2.76) at three listening positions. A fixed offset was added to the sound pressure at two listening positions for a better visualization. Parameters: $\mathbf{x}_s = (0, 0.5, 0)$ m, $\mathbf{n}_s = (0, -1, 0)$, circular secondary source distribution with a diameter of 3 m. ☞

frequencies. In the case of band-limited NFC-HOA, for the center position the low- and high-frequency pulses arrive at the same time. On the other hand, for both positions out of the center the low and high frequencies disintegrate. At first only low frequencies arrive at the listener's position, thereafter only high frequencies.

Focused sources have similar aliasing properties as band-limited NFC-HOA. The difference is that the aliasing-free region is not around the center of a circular secondary source distribution but around the focused source position. However, beside this they have some special properties that are obvious when analyzing their time signal at a non aliasing free listener position. As Figure 3.16 indicates the additional spatial aliasing wave fronts arrive before the desired one. This is caused by the time reversal approach[32] that is applied to produce the field of a focused source and cannot be overcome – compare (2.76).

So far this section has summarized the errors in the sound field due to the spatial sampling process of the secondary sources. In the following paragraph the possible influence of these errors on the perception of the sound field will be discussed.

Perception The discretization of the secondary source distribution may have a huge impact on the perception of the synthesized sound field, as a result of the large amount of errors which are introduced by the spatial sampling process.

Three perceptual features will strongly be influenced by the errors, namely the localization of the synthesized sound, its perceived timbre, and the perception of spectro-temporal artifacts. These features will at first be discussed in the following and then analyzed with dedicated experiments in Chapter 5.

For localization it is worthwhile to analyze the influence of the spatial aliasing on the ITD of the sound field. It is especially interesting for frequencies below 1.4 kHz, because for broad-band signals localization in the horizontal plane is dominated by the ITDs of low frequencies.[33] Figure 3.17 shows the ITD for synthesized signals of a point source for three different listening positions with the head always oriented to the front. The ITDs were calculated applying the

[32] Yon, Tanter, and Fink, op. cit.

[33] 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 (1992), pp. 1648–61

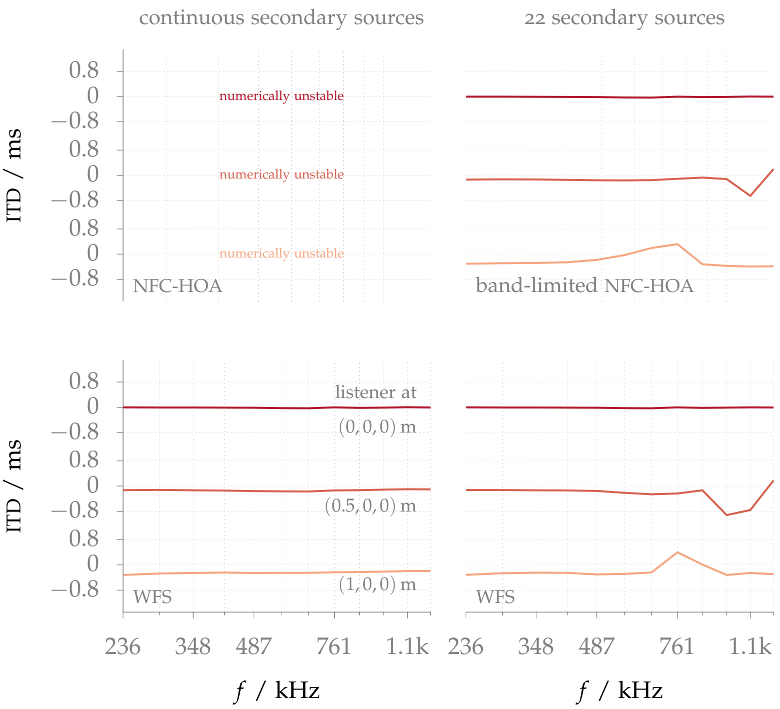


Figure 3.17: ITDs for a pink noise signal synthesized as a point source by NFC-HOA (2.45) and WFS (2.57) at three listening positions, with a head orientation of 90°. Parameters: $\mathbf{x}_\mathrm{s} = (0, 2.5, 0)$, $\mathbf{x}_\mathrm{ref} = (0, 0, 0)$, circular secondary source distribution with a diameter of 3 m. ☞

binaural model that is introduced in Chapter 6. In case of a continuous secondary source distribution, the ITDs are correct, because the sound field is not distinguishable from the desired one. For the sampled secondary source distribution consisting of 22 sources in this analysis, deviations of the ITD above 500 Hz are visible. The 500 Hz roughly corresponds to the aliasing frequency. By applying (3.3) it can be calculated that a distance of 12 cm between the secondary sources would be sufficient to achieve correct ITDs below 1.4 kHz.

For WFS and NFC-HOA the first wave front which arrives at the listener position is synthesized correctly. Subsequently, additional wave fronts with disturbing directional information are arriving at the listener position – compare Figure 3.15. The additional wave fronts are carrying the directional information of the corresponding secondary source they are coming from. This case is comparable to the localization in closed spaces. There, after the direct wave front additional reflectional wave fronts from the walls follow. The localization is mostly dominated by the first wave front, a phenomenon known as the precedence effect.[34] This leads to the hypothesis that localization for the case of SFS will be comparable to the case of a real sound field. If the number of secondary sources is becoming so low that the first wave front can no longer be synthesized correctly in the whole audience area, impairments of localization are likely. This is the case for stereophony.

For band-limited NFC-HOA this effect could also be critical for positions outside of the center, because here the first wave front is only correct for low frequencies. In contrast, wave fronts with high frequencies are coming from a different direction – compare again Figure 3.15. In this situation more than one source will probably be perceived.

[34] Litovsky et al., op. cit.

Another case where the localization could be largely affected by the synthesis errors is the synthesis of focused sources. Here the critical aspect is that the correct wave front arrives after – not before – all the additional high frequency wave fronts that arrive from false directions.

In Section 5.1 the localization in case of different sound field synthesis methods and for different synthesized sound sources like a point source or a focused source is investigated in great detail. Afterwards, in Chapter 6, an auditory model is used to predict the localization results and to predict the localization in the entire audience area.

Coloration is the other perceptual feature that is affected by synthesis errors. Here, it is not as straightforward as for the perception of the direction of the auditory event. There is a correlation between the amount of perceived coloration and the comb-filter like deviations in the magnitude of the spectrum resulting from sound field synthesis – compare Figure 3.12. Similar deviations of the spectrum occur also for stereophonic presentation and for auditory perception in closed spaces, where room reflections could create a comb-filter like spectrum. However, for stereophony, coloration plays only a minor role.[35] Further, in the case of closed spaces, not a colored auditory event is perceived, but the coloration is perceived as separate information about the acoustic properties of the room. The perceived character of the space in terms of an independent auditory event can then also be perceived as colored. For example, the human listener seems to be highly trained for natural patterns of reflections. Simulating a room with a simple model including only the first reflections will sound very unnatural, and different perceptual dimensions of coloration will be affected.

In the context of SFS both deviations in the magnitude spectrum and several reflection-like wave fronts will occur. Either of these might contribute to the perceived coloration, which may correspond to an additional unnatural room impression. The amount of deviations and number of additional wave fronts depend on the number and distance between the secondary sources. A continuous secondary source distribution will not lead to coloration, whereas a lower number might. Section 5.2 investigates these questions, and the coloration of different WFS systems will be compared to that of a two-channel stereophony setup.

For cases without a spatio-temporally correct first wave front, the errors in the synthesized sound field will also induce coloration, but it is not straightforward to predict the degree of coloration. It depends on how many auditory events the listener will perceive, and on the frequency content of the different sources.

Spectro-temporal artifacts may become audible if a technical system manipulates a given signal in an unnatural way, as it may happen for audio codecs. These codecs can add artifacts to the orig-

[35] V. Pulkki. "Coloration of Amplitude-Panned Virtual Sources". In: *110th Audio Engineering Society Convention*. 2001, Paper 5402.

inal signal that are, for example, perceivable as additional clicks. Especially pre-echoes are very likely to become perceivable as additional auditory events. Since the synthesized signal of a focused source has additional wave fronts arriving as pre-echoes, it is likely that these wave fronts are heard as independent auditory events – compare Figure 3.16. In Section 5.3, this effect is investigated for different secondary source setups and different listening positions. Additionally, an approach is presented to avoid spectro-temporal artifacts for the synthesis of focused sources.

To investigate the localization and coloration of synthesized sources, the number of applied secondary sources has to be varied in a large range going from only two or three up to a continuous secondary source distribution. This could be as much as 7000 sources for reaching an upper frequency of 20 kHz. Beside the varying number of secondary sources, results for different listening positions within the audience area are of interest: A similar impression of the auditory scene in an extended audience area is one of the claimed benefits of sound field synthesis.

It is obvious that these requirements represent a big challenge for the experimenter because it is not possible to practically build secondary source distributions with up to 7000 sources and inter-loudspeaker distances of below 1 cm. Even the task of positioning a listener reliably at different positions within the audience area is not straightforward for practical evaluation tests. To overcome these problems, all the different secondary source distributions and listener positions are simulated for this thesis via dynamic binaural synthesis using headphones.

In the next section, the method of dynamic binaural synthesis is explained in detail altogether with the experimental setups. Binaural synthesis is not absolutely transparent and its limitations and restrictions are also investigated and discussed.

4 Binaural Synthesis and Experimental Setup

The auditory perception of an acoustic event is largely triggered by the input signals to the ear drums. Other influences like multi-modal interactions from the visual or tactile senses have contributions to the auditory perception but will be neglected here as a first approximation. Hence, it is possible to simulate any acoustical event by synthesizing the corresponding signals at the ear drum.[1] One solution to achieve this is the application of binaural synthesis. For binaural synthesis the transfer functions of an acoustic source to the two ears of the listener are measured in the desired environment. Afterwards, any audio signal can be convolved with the time signal corresponding to the transfer function and then played back to the listener over headphones. If the headphones are considered as an acoustically transparent system the signals of the ear drums of the listener correspond to the ones from the acoustic source for which the transfer function was measured. As this thesis is only interested in the investigation of sound field synthesis methods without the influence of the reproduction room, an anechoic chamber is chosen as environment. The transfer functions are then called *head-related transfer functions (HRTFs)*.[2]

One of the drawbacks of static binaural synthesis is the assumption of a static listener which can move neither head nor body. This can be overcome by measuring transfer functions for different positions and head-orientations of the listener and the usage of a head tracker system at reproduction time in order to switch the transfer functions accordingly. The binaural synthesis is then termed dynamic binaural synthesis or binaural room scanning (BRS)[3] – compare Fig. 4.1.

A corresponding problem is the large number of measurements that are required. Ideally, different positions and head-orientations for every single listener have to be measured. In order to limit the number of measurement points, non-individual transfer functions are often recorded with a dummy head that mimics a common listener. Another restriction applied here is to limit the possible movements of the listener to head rotations in the horizontal plane only. With such restrictions dynamic binaural synthesis is already possible with 360 or less measurements in the horizontal plane, one per

[1] H. Møller. "Fundamentals of Binaural Technology". *Applied Acoustics* 36 (1992), pp. 171–218.

[2] To simplify the syntax transfer functions measured in a room will also be called HRTFs in this thesis and not binaural room impulse responses (BRTFs) as it is often the case in the literature.

[3] U. Horbach et al. "Design and Applications of a Data-based Auralization System for Surround Sound". In: *106th Audio Engineering Society Convention*. 1999, Paper 4976.

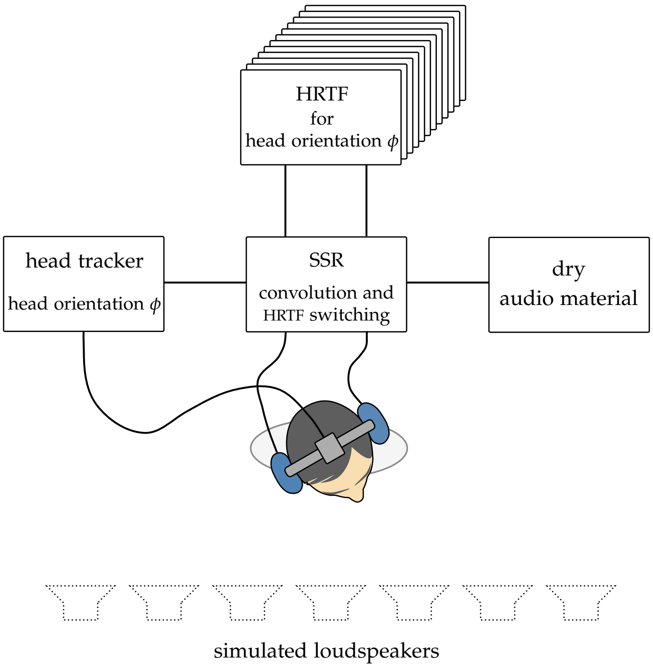


Figure 4.1: Functional principle of dynamic binaural synthesis. The listener is wearing headphones and a head tracker. The audio material is convolved with the HRTF that incorporates all simulated loudspeakers for the corresponding listener orientation. ☞

degree.[4]

The approximation by non-individual transfer functions leads to a binaural synthesis which cannot be considered as transparent. Deviations in the magnitude of the transfer function may be perceived as a change in timbre of the signal.[5] A mismatch in the ear distance of the dummy head and the listener can lead to a slight moving of the sound source with head movements of the listener.[6]

The usage of a dummy head still implies a lot of measurements, if not only head movements but different source directions are considered. For all source directions a measurement with 360 head-orientations will sum up to a measurement with 129 600 points. This can be simplified by the assumption that the difference between a movement of the head relative to the torso or the movement of the head together with the torso makes no perceptible difference in an anechoic chamber.[7]

The usage of headphones for synthesizing the ear signals adds another non-transparent component. Headphones have their own transfer function which can be compensated only in a limited way due to the restrictions that exist in filter design. Different labs have spent a large amount of time to come up with systems that are transparent at least for some signals. For instance, they built extraaural headphones[8] or developed techniques for fast individual HRTF measurements and individual headphone compensations.[9]

Simplifying the prerequisites of this thesis, non-individual transfer functions and commercially available headphones with a custom compensation filter have been used. In order to nevertheless get meaningful results, additional experiments have been carried out to investigate the influence of the non-transparent binaural synthesis on the different sound field synthesis feature addressed in this thesis

[4] A. Lindau, H.-J. Maempel, and S. Weinzierl. "Minimum BRIR grid resolution for dynamic binaural synthesis". *The Journal of the Acoustical Society of America* 123.5 (2008), pp. 3851–56

[5] M. Takanen, M. Hiipakka, and V. Pulkki. "Audibility of coloration artifacts in HRTF filter designs". In: *45th Audio Engineering Society Conference*. 2012, Paper 3.3.

[6] V. R. Algazi, C. Avendano, and R. O. Duda. "Estimation of a Spherical-Head Model from Anthropometry". *Journal of the Audio Engineering Society* 49.6 (2001), pp. 472–79.

[7] R. Popko. *Zur Hörbarkeit und Interpolation von Kopf-über-Torso-Orientierungen bei Aufnahmen mit einem Kopf-und-Torso-Simulator*. Technische Universität Berlin, 2013.

[8] V. Erbes et al. "An extraaural headphone system for optimized binaural reproduction". In: *39th German Annual Conference on Acoustics*. 2013.

[9] B. S. Masiero. "Individualized Binaural Technology". PhD thesis. RWTH Aachen, 2012.

such as the localization and coloration of the created auditory events.

In the following section, the experimental setup common to all experiments of this thesis is presented. Thereafter, experiments verifying the adequateness of using binaural synthesis to investigate different sound field synthesis issues are carried out.

4.1 Experimental Setup of Binaural Synthesis

4.1.1 Head-Related Transfer Function

The HRTFs used for most of the binaural simulations are part of a larger measurement conducted in the anechoic chamber of the Technische Universität (TU) Berlin.[10,11] The set used in this thesis for the experiments was measured in the horizontal plane only. It has a distance of 3 m between the loudspeaker (Genelec 8030A) and the dummy head (KEMAR, type 45BA) and a resolution of 1°. For non-measured directions, HRTFs were calculated by linear interpolation. For distances smaller or larger than 3 m the HRTF was adapted by delaying and weighting accordingly. Loudspeaker arrays were created by a super-position of the HRTFs corresponding to the single loudspeakers.

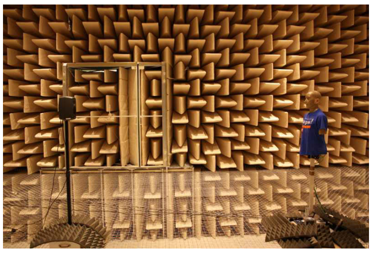


Figure 4.2: Measurements with the artificial head in the anechoic chamber. ☞ `intpol_ir.m` `get_ir.m`

4.1.2 Apparatus

Stimuli were digitally generated at a sampling rate of 44.1 kHz. A computer was used to generate impulse responses for the binaural synthesis. This could be an HRTF representing just one loudspeaker up to an array of loudspeakers driven by signals calculated by one of the sound field synthesis methods, depending on the experiment. The convolution of the time signal of the HRTFs with the audio signal was performed using the SoundScape Renderer.[12] The audio signals were fed into the SoundScape Renderer via Pure Data.[13] The advantage of this setup is that Pure Data easily allows to switch the output to another convolution instance of the SoundScape Renderer including a pair of HRTF representing another condition of the experiment. This means that the listener can switch between different conditions within the audio signal in contrast to the case where the audio signal starts playing from the beginning for every switch. The PC was equipped with an RME HDSP MADI card and for the digital-to-analog conversion CreamWare A16 converters were used. The listeners wore AKG K601 headphones and a corresponding headphone compensation filter was applied to the signals. The head movements of the listeners were tracked by a Fastrak Polhemus head tracker with a resolution of around 1°, and the tracking data were passed to the SoundScape Renderer. The SoundScape Renderer was then switching the HRTFs for the dynamic binaural synthesis, according to the orientation of the listener given by the head tracker data. Figure 4.1 illustrates the setup.

[10] The HRTF set is freely available, and is described in H. Wierstorf et al. "A Free Database of Head-Related Impulse Response Measurements in the Horizontal Plane with Multiple Distances". In: *130th Audio Engineering Society Convention*. 2011, eBrief 6.

[11] The author would like to recommend the Spatial Oriented Audio File (SOFA) format for the reader that is interested in HRTFs. It is a joined effort between different labs to define a common file format for exchanging HRTFs and other spatial oriented acoustical measurements. It is described in P. Majdak et al. "Spatially Oriented Format for Acoustics: A Data Exchange Format Representing Head-Related Transfer Functions". In: *134th Audio Engineering Society Convention*. 2013, Paper 8880.

[12] The SoundScape Renderer is an open source software and is described in M. Geier, J. Ahrens, and S. Spors. "The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods". In: *124th Audio Engineering Society Convention*. 2008, Paper 7330.

[13] Pure Data is an open source software and first described in M. S. Puckette. "Pure Data: another integrated computer music environment". In: *Second Intercollege Computer Music Concerts*. 1996, pp. 37–41.

4.2 *Verifying Binaural Synthesis for Localization Experiments*[14]

The thesis aims to investigate localization in the audience area for different sound field synthesis methods based on listening tests. To accomplish this, the ear signals are simulated with dynamic binaural synthesis as outlined above. One requirement for this kind of investigation is that the localization results will not be influenced by the binaural synthesis. This will be verified in the following.

Experiments from the literature show that binaural synthesis has only little influence on the results as long as its dynamic implementation is applied. In this case, front-back confusions are avoided. The localization error for real sources, that is the absolute difference between the direction of a real loudspeaker and direction of the auditory event lies between 2° to 5°.[15] If individual HRTFs of the listeners were used, no difference between localization of real and virtual speakers were found. For non-individual HRTFs deviations around 1° were found.[16]

One reason for the varying results for the localization performance in the literature is the fact that localization experiments are critical regarding the used pointing method. Due to the fact that the actual localization error can be as small as 1°, the error of the pointing method has to be smaller than 1°, which cannot be achieved with all methods.[17]

In order to test the influence of the pointing method and the dynamic binaural synthesis with non-individual HRTFs on human localization a listening test was conducted. Here, the localization of different loudspeakers representing real sound sources and of the same sources simulated via dynamic binaural synthesis were tested. If the localization results for the real sources are comparable to the results from the literature, it indicates that the accuracy of the pointing method is sufficient. If the localization results for the simulated sources are equal to the ones of the real sources it proves that the dynamic binaural synthesis has no influence on the localization results. If both conditions are fulfilled the presented method is suitable for investigating the localization in sound field synthesis.

First, the applied pointing method is introduced, followed by a description of the experiment and the results.

4.2.1 *Pointing Method*

This thesis applies a pointing method that Makous and Middlebrooks[18] used in a similar way. Here, the listener has to point with her head towards the direction of the auditory event, while the sound event is present. This has the advantage that the listener is directly facing the source, a region in which the minimum audible angle is the smallest.[19] If the listener is pointing her nose in the direction of the source, an estimation error of the sources at the side will occur, due to an interaction with the human sensory-motor system. To

[14] Parts of this section are published in H. Wierstorf, S. Spors, and A. Raake. "Perception and evaluation of sound fields". In: *59th Open Seminar on Acoustics*. 2012; Wierstorf, Raake, and Spors, op. cit.

[15] B. U. Seeber. "Untersuchung der auditiven Lokalisation mit einer Lichtzeigermethode". PhD thesis. 2003; A. W. Bronkhorst. "Localization of real and virtual sound sources". *The Journal of the Acoustical Society of America* 98.5 (1995), pp. 2542–53; W. Hess. "Influence of head-tracking on spatial perception". In: *117th Audio Engineering Society Convention*. 2004, Paper 6288; J. C. Makous and J. C. Middlebrooks. "Two-dimensional sound localization by human listeners." *The Journal of the Acoustical Society of America* 87.5 (1990), pp. 2188–200.

[16] Seeber, op. cit.

[17] P. Majdak et al. "The Accuracy of Localizing Virtual Sound Sources: Effects of Pointing Method and Visual Environment". In: *124th Audio Engineering Society Convention*. 2008, Paper 7407; Seeber, op. cit.

[18] Makous and Middlebrooks, op. cit.

[19] A. W. Mills. "On the minimum audible angle". *The Journal of the Acoustical Society of America* 30.4 (1958), pp. 237–46.

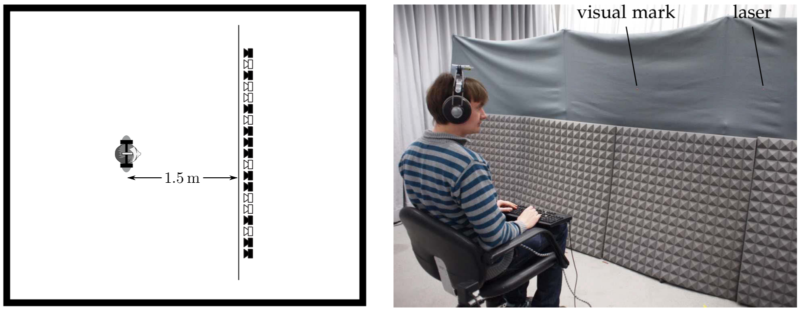


Figure 4.3: Sketch of the apparatus in the listening room (left) and a listener during the experiment (right). Only the loudspeakers marked in black were used in the experiment. Note that the room was dark during the experiment. ☞

overcome this, a visual pointer is added, showing the listener what her nose it pointing at.[20] To add such a visual pointer a small laser pointer was mounted onto the headphones – compare Fig. 4.3.

4.2.2 *Apparatus*

For the binaural synthesis the apparatus described in Sec. 4.1 is applied. In addition 19 Fostex PM0.4 loudspeakers were placed in an acoustically damped listening room (room *Calypso* in the Telefunken building of the TU Berlin). The room has a volume of 83 m^3 and a reverberation time RT_{60} of 0.17 s at a frequency of 1 kHz. The loudspeakers were arranged as a linear array with a spacing of 0.15 m between them. Only the eleven loudspeakers highlighted in Fig. 4.3 were involved in the experiment. The listener was positioned in a heavy chair, 1.5 m in front of the loudspeaker array, with an acoustically transparent curtain in between. A sketch of the setup and a picture is shown in Fig. 4.3. The orientation and position of the listeners during the experiment was recorded with the same head tracker that provides the data for the dynamic binaural synthesis.

4.2.3 *Listeners*

Eleven adult listeners were recruited to conduct both parts of the experiment – aged 21 to 33 years. Four of them had prior experiences with psychoacoustic testing. The listeners were financially compensated for their effort.

4.2.4 *Stimuli*

As audio material, Gaussian white noise pulses with a duration of 700 ms and a pause of 300 ms between them were applied. The single pulses were windowed with a Hanning window of 20 ms length at the start and the end. The signal was band-pass filtered with a fourth order butterworth filter with its pass-band between 125 Hz and 20000 Hz. The signal with a total length of 100 s was stored and played back in a loop during the experiment. The single pulses of this signal were independent white noise signals. For the headphone reproduction the noise file was convolved with the time signal of the corresponding HRTF.

[20] J. Lewald, G. J. Dörrscheidt, and W. H. Ehrenstein. "Sound localization with eccentric head position." *Behavioural Brain Research* 108.2 (2000), pp. 105–25.

4.2.5 *Procedure*

The listeners sat on a chair, wearing headphones with the mounted laser pointer and had a keyboard on their knees – see Fig. 4.3. They were instructed to use the point for pointing into the direction from where they perceived the auditory event. The test participants were informed that the vertical direction should be ignored. After they made sure to point into the right direction, they were asked to hit the enter key. The listeners' head orientation was calculated as the mean over the following 10 values obtained from the head tracker, which corresponds to a time of 90 ms. After the key press, the next trial started instantaneously, which implied that the listener always started the localization from the last position. The listeners were instructed that they could turn their head if they were unsure about the direction of the sound.

There were three conditions in the experiment, *loudspeaker*, *binaural synthesis*, and *binaural synthesis with room reflections*. In this thesis, only the first and second will be discussed. The full experiment is presented in Wierstorf et al.[21]

For the first one, the noise pulses were played through any of the eleven loudspeakers. For the other two conditions the sound was played via headphones. Three different conditions and eleven different loudspeakers led to 33 trials. Every listener had to pass all 33 trials six times. The first 33 trials were for training, thereafter a session with 66 trials and one with 99 trials was passed. The order of the conditions and presented loudspeakers was randomized. In average the listeners needed 15 minutes to complete the experiment excluding the training.

At the beginning of each session, a calibration was carried out. First, the loudspeaker at 0° was active, and the listener had to look into the respective direction in order to calibrate the head tracker. In a second step, the listener was indicated to point towards a given visual mark on the curtain. The second step formed a connection between the head tracker orientation and the room. After the calibration, the room was darkened and the experiment started.

4.2.6 *Data Analysis*

The listener was able to turn her head, and move the head in a translatory way. For the conditions employing headphone reproduction, this had no influence on the results for the perceived direction, because the dynamic binaural synthesis compensated only for the angle of the head, not for its absolute position. Hence, the virtual source was moving with the listener in case of translational movements. For the loudspeaker condition, this is no longer true, and the perceived angle between a loudspeaker and the head of the listener is changing with possible translatory head movements. To calculate the direction of the auditory event, the data was compensated for these head movements, which were acquired by the head tracker, by the

[21] Wierstorf, Spors, and Raake, op. cit.

following formula.

$$\phi' = \tan^{-1}\left(\left[(1.5 - y)\tan\phi - x\right]/1.5\right) \tag{4.1}$$

Here ϕ is the measured head orientation, x, y are the measured coordinates of the head tracker, assuming that the origin of the coordinate system is at the center of the chair, and ϕ' is the final value for the direction of the auditory event. In an additional step, the measured orientation of the head had to be connected to the orientation of the listener within the room. This step is needed because the orientation of the head tracker is not an absolute value and was chosen anew in every session. In practice, this was solved by compensating the measured head orientation data with the position of a tiny visual mark on the curtain. Its position in the current head tracker orientation coordinate system was measured in the calibration step.

After the data calibration, the results from both sessions were pooled for every listener and the mean and standard deviation were calculated. The average over all listeners together with the confidence interval[22] was then calculated using these data.

4.2.7 Results

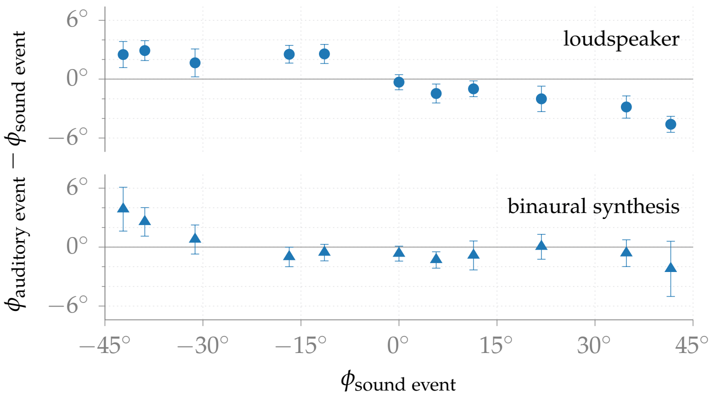


Figure 4.4: Difference between the direction of the auditory event and the sound event for loudspeakers and the binaural simulation of the loudspeakers. Average over all listeners together with the 95% confidence interval is shown. ☞

One listener had a standard deviation that was twice as high as that of the other listeners. The measurements from this listener were excluded from the results. Figure 4.4 shows the average over the listeners, together with the 95% confidence intervals. The difference between the direction of the auditory event and the direction of the sound event is shown. It states that it never exceeds 5° for any condition and loudspeaker, but a slight underestimation of the loudspeakers at the sides can be observed for both conditions. As another measure the mean of the standard deviations of the listeners was calculated. The loudspeaker condition has an average standard deviation of 2.2° ± 0.2°. For the binaural synthesis condition the standard deviation is slightly higher at 3.8° ± 0.3°.

[22] G. Cumming, F. Fidler, and D. L. Vaux. "Error bars in experimental biology." *The Journal of Cell Biology* 177.1 (2007), pp. 7–11.

4.2.8 Discussion

Exclusively considering the results for $-30°$ to $30°$ of the direction of the sound event, the localization error for the loudspeakers is around $1°$-$2°$ which is in agreement with the literature and indicates that the resolution of the pointing method is sufficient. For sound event positions closer to the side a slight undershoot of the direction of the auditory event is visible. This means that the listener is not looking far enough to the side. This effect is known from pointing methods without visible feedback, which should be compensated for with the laser pointer giving visual feedback. To overcome this problem, which is also prominent in the results for the binaural synthesis condition, only sound events in the range from $-30°$ to $30°$ will be investigated in further experiments.

The next question to answer is the influence of the binaural simulation on the localization accuracy in comparison to the case of real loudspeakers. The average localization error together with its confidence interval is $2.4° \pm 0.3°$ for the loudspeaker condition and $2.0° \pm 0.4°$ for the binaural synthesis condition. This allows the conclusion that the simulation of the loudspeakers by dynamic binaural synthesis has no influence on the localization accuracy that can be achieved. That means binaural simulations can be used to study localization for different loudspeaker setups as they are needed for sound field synthesis.

Interestingly the average standard deviation is significant higher for the binaural synthesis condition than for the loudspeaker condition. This correlates with the finding that the listeners needed longer for answering in the case of the binaural synthesis. The average time after the start of the stimulus and the pressing of the answer key was 3.5 s ±0.7 s for the loudspeaker condition and 5.5 s ±1.7 s in the case of the binaural simulation. This indicates that even though the average localization accuracy is not affected by the binaural simulation, it takes more effort for the listener to find the position of the auditory event in the case of dynamic binaural synthesis.

4.2.9 Conclusion

It was found that the accuracy of the applied pointing method is sufficient as long as the sound event is not placed more than $\pm 30°$ to the side of the listener. In addition, the binaural simulation of the loudspeakers has a negligible influence on the localization accuracy of the test participants. Only the time and the certainty with which the listeners localize the sound is slightly degraded. It can be concluded that the dynamic binaural synthesis method is a proper tool to investigate localization in SFS.

4.3 *Verifying Binaural Synthesis for Simulation of Loudspeaker Arrays*[23]

Up to here only the binaural simulation of a single loudspeaker was considered. In order to investigate sound field synthesis methods, arrays of loudspeakers are of interest that can have up to thousands of loudspeakers. To handle such setups it is practically preferable if the complete loudspeaker array can be simulated by a HRTF set measured only for one loudspeaker. In this section it will be discussed how this can be achieved and what differences can be expected in comparison to the case of having HRTFs for the whole loudspeaker array.

get_ir.m
intpol_ir.m

If a HRTF set for a single loudspeaker is measured for placements of the loudspeaker all around the dummy head, HRTF sets for loudspeaker arrays can be created by applying interpolation, extrapolation and superposition to the HRTFs from the set of the single loudspeaker. The straightforward solution is to apply a linear interpolation and a time delay and amplitude weight for the extrapolation.

One of the differences of simulating the whole array with an HRTF from a single loudspeaker is the absence of the other loudspeakers during the HRTF measurement. These other loudspeakers – if active – can influence the impedance of the measured one. In addition the body of the loudspeakers can change the magnitude of the transfer function by adding reflections. Völk et al.[24] have investigated both effects. The influence of the impedance is negligible, but the change in magnitude of the transfer function can reach 4 dB.

Another influencing factor is the loudspeaker's directivity. In the case of an HRTF measurement of a single loudspeaker and a common setup, that loudspeaker is always pointing towards the listener. Whereas in an HRTF measurement of a linear loudspeaker array, the single loudspeakers all are pointing in the same direction and not towards the point where the listener is sitting. To investigate the degree of difference due to the directivity, an HRTF measurement of a linear loudspeaker array was performed in an anechoic chamber.

4.3.1 *Method*

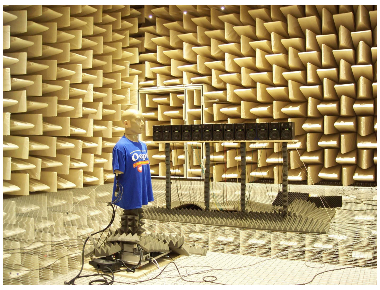


Figure 4.5: HRTF measurement of a loudspeaker array in the anechoic chamber of the TU Berlin. ☞

HRTFs of an loudspeaker array were measured in the anechoic chamber of the TU Berlin. The same dummy head and hardware as described in Wierstorf et al.[25] was used for the recording. The loudspeaker array itself consisted of 13 Fostex PM0.4 loudspeakers placed with a distance of 15 cm between them and a distance of 2 m between the center loudspeaker and the dummy head. In order to get results for larger arrays, the 13 loudspeakers were moved to the right and the left whereby the last two loudspeakers of the array were placed at the same positions as the two last loudspeakers of the central array on both sides. This means from the center array the signals of 11 loudspeakers were recorded and for the side arrays the signals of 12 loudspeakers, each. This corresponds to the size of the whole array

[23] Parts of this section are published in H. Wierstorf, A. Raake, and S. Spors. "Psychoakustik der Wellenfeldsynthese: Vor- und Nachteile binauraler Simulation". In: *38th German Annual Conference on Acoustics*. 2012.

[24] F. Völk, E. Faccinelli, and H. Fastl. "Überlegungen zu Möglichkeiten und Grenzen virtueller Wellenfeldsynthese". In: *36th German Annual Conference on Acoustics*. 2010, pp. 1069–70.

[25] Wierstorf et al., op. cit.

of 35 loudspeakers or an array length of 5.1 m.

To investigate the influence of the directivity the same loudspeaker array was simulated via binaural synthesis by applying inter- and extrapolation to the HRTF of the single loudspeaker placed directly in front of the dummy head. The setup is illustrated in Fig. 4.6. Afterwards, for every loudspeaker of both arrays the corresponding HRTF was convolved with a 1 s long white noise signal and fed into an auditory filterbank realized by gammatone filters with a distance of 1 ERB. The magnitude was calculated by building the root mean square value in every frequency channel.

auditoryfilterbank.m

4.3.2 *Results and Discussion*

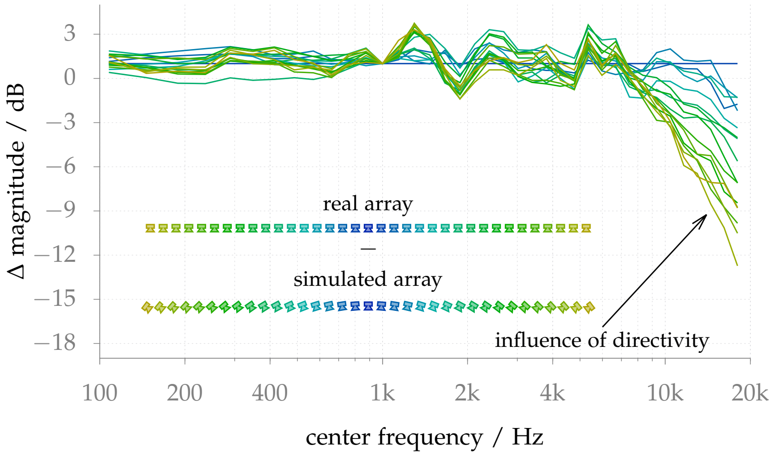


Figure 4.6: Difference between the magnitude of the simulated and real loudspeaker array dependent on the center frequency of the auditory filter bank. ☞

The difference of the output signals in each frequency band was calculated for the loudspeaker arrays. Figure 4.6 presents the results. It can be seen that for frequencies greater than 7 kHz the signal from the measured loudspeaker array is getting extenuated compared to the simulated loudspeaker array. This extenuation is getting larger for higher frequencies.

The main difference between the simulated and the measured array is the orientation of the loudspeakers. The difference in magnitude for high frequencies can be explained by a non-omnidirectional directivity pattern for high frequencies, which is the case for most loudspeakers.

4.3.3 *Conclusion*

A binaural simulation of a loudspeaker array can be implemented in two ways. One way is to build the desired loudspeaker array in the anechoic chamber and then measure the HRTF set. The other way is to measure the HRTF for a single loudspeaker and create the array by superposition and inter- and extrapolation of the HRTF of the single loudspeaker to come up with the same array.

Both results will be slightly different. If a loudspeaker array is measured in an anechoic chamber the other loudspeakers add additional reflections which will alter the transfer function of the measured one up to 4 dB.[26] The loudspeakers of a linear array are all pointing towards one direction whereas in the case of the simulation by a single loudspeaker the loudspeakers are always pointing towards the listener. The directivity pattern of the loudspeakers will lead to differences in the transfer functions. Those differences will not be present for a circular loudspeaker array.

For the investigation of sound field synthesis methods the differences are negligible or could be an advantage. The case of the binaural simulation using a single loudspeaker allows to build a virtual loudspeaker array that better fits the assumption made in the theory, namely that the loudspeakers are monopoles and have no directivity and no influence upon another. Hence, in this thesis binaural simulations of loudspeaker arrays will be produced with the HRTF of a single loudspeaker.

4.4 *Verifying Binaural Synthesis for Coloration Experiments*

The investigation of changes in timbre with binaural synthesis is possible only to some extent. The biggest problem is, that the synthesis itself introduces changes in timbre, which can only be compensated for to some degree by using individual HRTFs and individual headphone compensations.[27]

That means that an investigation of absolute coloration judgements will not be possible with binaural synthesis, because the measured coloration could be due to the synthesis process itself or due to the system under investigation and there is no way to distinguish between both cases. If the absolute coloration due to the binaural synthesis could be limited, the differences in coloration between different simulated systems could be investigated under the assumption that the binaural synthesis has the same influence on coloration for all systems.

One promising result from the literature is the study by Olive et al.[28] who found no difference in the preference ratings for four different loudspeakers between the measurement with real loudspeakers and their binaural simulations. They applied non-individual HRTFs and non-individual headphone compensation filters. This was further supported by a study presented in Wittek[29] that found the same amount of intra-system coloration for a stereophonic setup realized by real or simulated loudspeakers. The simulation was done via dynamic binaural synthesis.

In this section the amount of deviation of the amplitude spectrum due to binaural synthesis will be quantified to some extent. In addition, it will be shown that the deviation of the spectrum will be the same, independent on the simulated system. This is accomplished by recording both the ear signals for the binaural simulation and the ear signals for the reproduction with real loudspeakers each by using a dummy head. To consider the fact of non-individual HRTFs record-

[26] Völk, Faccinelli, and Fastl, op. cit.

[27] Compare for example Masiero, op. cit.

[28] S. E. Olive, T. Welti, and W. L. Martens. "Listener Loudspeaker Preference Ratings Obtained in situ Match Those Obtained via a Binaural Room Scanning Measurement and Playback System". In: *122nd Audio Engineering Society Convention*. 2007, Paper 7034.

[29] H. Wittek. "Perceptual differences between wavefield synthesis and stereophony". PhD thesis. University of Surrey, 2007, Fig. 8.4.

ing is done with a different dummy head than the measurement of the HRTFs that are used for the binaural synthesis.

4.4.1 *Method*

In room *Pinta* in the Telefunken building of TU Berlin a circular loudspeaker array with 56 loudspeakers is installed. The room has a volume of 54 m^3 and is acoustically damped with a reverberation time RT_{60} of 0.81 s at a frequency of 1 kHz. For all installed loudspeakers, HRTFs were measured with the FABIAN dummy head.[30] The dummy head was wearing AKG K601 headphones during the measurement. Wearing of open headphones during the measurement is often done in HRTF measurements for binaural simulations. It allows the direct comparison of a reproduction by the real or binaurally simulated loudspeaker by a listener without the need to remove the headphone from her head in the experiment.[31]

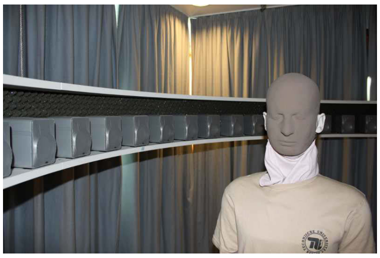


Figure 4.7: Recording of HRTFs with FABIAN in room *Pinta*. Note that this picture is from a recording session where FABIAN was placed out of the center and did not wear headphones. ☞

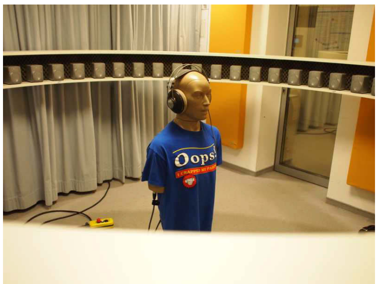


Figure 4.8: Recording of the ear signals for binaural synthesis and real loudspeakers in room *Pinta*. ☞

For the measurement the KEMAR dummy head was placed at the same position in the center of the loudspeaker array in *Pinta* as FABIAN have been placed. Again open AKG K601 headphones were placed on its head. In this way, the dummy head was able to record sound coming from the loudspeakers going through the open headphones. The HRTFs measured with FABIAN as described above were used for a binaural simulation which was then recorded by the KEMAR dummy head, too.

Four different conditions were compared. For the condition *mono* a single loudspeaker was active. The other three conditions involved all the synthesis of a point source placed 1 m behind the array applying a circular loudspeaker array with 14, 28, or 56 loudspeakers and WFS. Note that this corresponded to 4, 9, or 17 active loudspeakers due to the secondary source selection – compare (2.65).

For all four loudspeaker configurations the source was placed in the front, left, right, or back of the dummy head, leading altogether to 16 different conditions. The source consisted of a continuous white noise signal, that was recorded for 24 s for every condition and position.

4.4.2 *Results and Discussion*

Figure 4.9 shows the differences in magnitude between the real loudspeakers and their binaural simulations for each condition. The magnitude was calculated by processing the recorded noise signal with an auditory filterbank and calculating the root mean square of the signal per band. Afterwards, the magnitude of the binaurally simulated loudspeakers was subtracted from the magnitude of the real loudspeakers.

`auditoryfilterbank.m`

For all four positions, the deviations in magnitude are within ±5 dB up to 5 kHz. For higher frequencies larger dips and peaks are visible going up to a difference in magnitude of 15 dB for the single loudspeaker in the back. The degree of deviation is not dependent on the condition type, and there is no systematic change in

[30] The measurements were performed by Alexander Lindau, the used dummy head is described in A. Lindau, T. Hohn, and S. Weinzierl. "Binaural resynthesis for comparative studies of acoustical environments". In: *122nd Audio Engineering Society Convention*. 2007, Paper 7032.

[31] An example of verifying binaural synthesis is presented in A. Lindau and S. Weinzierl. "Assessing the Plausibility of Virtual Acoustic Environments". *Acta Acustica united with Acustica* 98.5 (2012), pp. 804–10.

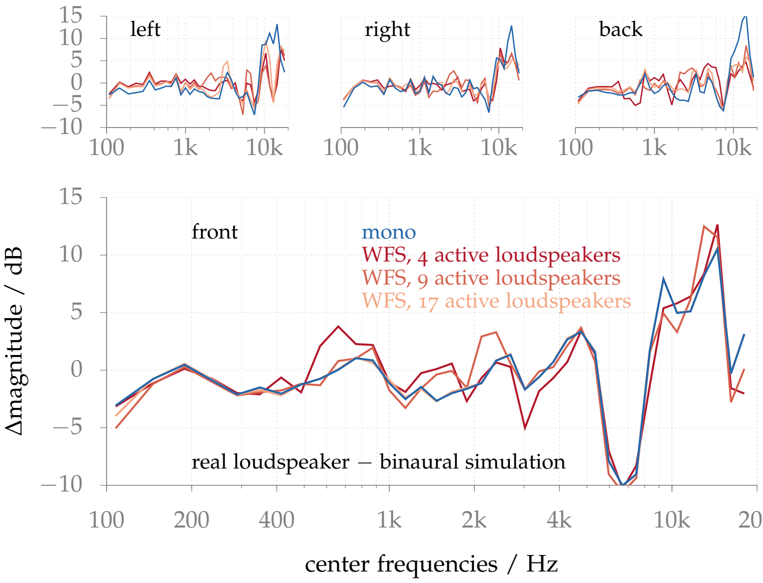


Figure 4.9: Differences in magnitude between real loudspeaker setups and their binaural simulation. A single loudspeaker (mono) or a circular loudspeaker array driven by WFS was applied to synthesize a point source in the front, left, right, or back of the listener. ☞

deviation visible due to the number of used loudspeakers. The only exception is the magnitude of the single loudspeaker for a frequency around 15 kHz. At this frequency the deviation between the real and simulated loudspeaker is stronger for all non-frontal source positions than it is for all WFS conditions.

The measurements showed that the deviation is independent of the number of loudspeakers in the case of WFS. The deviation between all WFS systems and the monaural source for some source positions at high frequencies is still an open question and could not be solved with the conducted experiments.

4.4.3 *Conclusion*

In this thesis, experiments to quantify the coloration in sound field synthesis methods were carried out. In order to investigate the usage of non-individual HRTFs and a non-individual headphone compensation, dummy head recordings were carried out. The fact of non-individual HRTFs was considered by using two different dummy heads for the recording of the HRTFs that were used for the binaural simulation and for the recording of the simulated results.

The results show that deviations up to 5 dB are present for frequency channels below 5 kHz. For higher frequencies the deviation in magnitude between a simulated and a real loudspeaker can be up to 15 dB. Further, the measurements underlines the view that the offset in magnitude is a linear process and it is independent of the number of simulated loudspeakers. This assumption is necessary to compare different loudspeaker setups to each other in a coloration experiment.

These results in combination with similarly supportive conclusions from the literature are the basis for applying the binaural synthesis to investigating coloration in WFS as described in Sec. 5.2.

5
Psychoacoustics of Sound Field Synthesis

THE PERCEPTION of synthesized sound fields could be highly affected by the errors that are introduced into the sound field by practical setups, as discussed in Chapter 3. In this chapter, different experiments are presented that investigate the influence of those errors on the perception and if an authentic synthesis is possible at all. The investigation is split up in single experiments for different perceptual attributes. It will start with measuring the localization accuracy as an indicator for spatial fidelity for WFS and NFC-HOA and different secondary source distributions. This approach will be repeated for timbral fidelity and WFS. The last section deals with the special case of focused sources in WFS for which spectro-temporal artifacts are another perceptual attribute in addition to coloration and localization accuracy.

5.1 Spatial Fidelity[1]

The human auditory system has the remarkable ability to detect the horizontal direction of a sound source up to a accuracy of 1°. This imposes strict requirements on a spatial audio system, if the system tries to achieve authenticity compared with the real world. In this section the localization accuracy of the listener for different sound field synthesis systems is investigated in a systematic way. It is shown which distance of the loudspeakers is required to achieve authenticity and what happens for larger inter-loudspeaker distances. In the next step, the properties of the synthesized sound fields that allow or hinder the localization will be discussed.

For the simplest possible spatial audio system – the stereophonic setup – it is well known that the localization is only correct inside a small area which is called the sweet-spot. If the listener is standing outside the sweet-spot, the localization is dominated by the position of the nearest loudspeaker. For sound field synthesis methods, on the other hand, it is assumed that they are able to provide an equally good localization in the whole listening area. A feature that is especially claimed for WFS. But a transition from the sweet-spot-like behavior of stereophony to sound field synthesis has to take place for SFS setups applying a low number of loudspeakers. A good example is band-pass limited NFC-HOA, for which also a pronounced

[1] Parts of this section are published in H. Wierstorf, A. Raake, and S. Spors. "Localization of a virtual point source within the listening area for Wave Field Synthesis". In: *133rd Audio Engineering Society Convention*. 2012, Paper 8743; H. Wierstorf, A. Raake, and S. Spors. "Localization in Wave Field Synthesis and higher order Ambisonics at different positions within the listening area". In: *39th German Annual Conference on Acoustics*. 2013.

sweet-spot exists in the center of the loudspeaker array – compare Figure 3.11.

Localization was investigated for different sound field synthesis setups in the last years, but in most of the publications only a central listening position was considered.

Test results for WFS show that the localization at a central listening position is not or only slightly impaired for loudspeaker spacings less than 25 cm.[2] All studies included a synthesized point source and different linear loudspeaker arrays. The experiments were performed directly with a localization test or indirectly with a minimum audible angle experiment.[3] No differences were found between the results for broadband stimuli and low-pass stimuli, containing only energy below the aliasing frequency. This indicates that the distortions for ITDs and ILDs at high frequencies due to the spatial aliasing artifacts apparently does not influence the localization accuracy. Verheijen has carried out localization tests for point sources and focused sources placed at different positions.[4] For a loudspeaker spacing of 11 cm he found no difference in localization compared to a real source. For a spacing of 22 cm the localization blur increased by 0.5°. If an infinitely long linear array is applied, the localization impact due to a change of the source position would be equivalent to that due to the change of the listener position. The length of the array in Verheijen's experiment was 2.53 m, which is too short to apply this equivalence.

For NFC-HOA no localization results are available. For HOA experiments were carried out for a central and in some cases one off-center listening position.[5] In all of the studies a maximum Ambisonics order of five was investigated. Systems with low orders like these are more equivalent to stereophonic panning approaches than sound field synthesis methods. This implies that for off-center listening positions the synthesized source will be localized towards the nearest loudspeaker – compare Figure 7 in Spors et al.[6] For the highest order of five the localization was no longer strictly bound to the nearest loudspeaker and a localization accuracy around 3° could be achieved.[7]

In this thesis, the focus of the localization experiments lies on two aspects. At first, the accuracy in the whole audience area should be assessed. This was ensured by applying 16 different listener positions, equally distributed in the audience area. In addition, the dependence of the localization accuracy on the distance between adjacent loudspeakers was investigated. To achieve this, three different loudspeaker distances were tested for a linear and a circular loudspeaker array. For NFC-HOA another dependency was added by varying the order of the spherical harmonics using the same loudspeaker distance.

The tests were split in four different experiments, which will be introduced in the following.

[2] P. Vogel. "Application of Wave Field Synthesis in Room Acoustics". PhD thesis. Technische Universiteit Delft, 1993; Start, op. cit.; Wittek, op. cit.

[3] Vogel, op. cit.

[4] Verheijen, op. cit.

[5] E.g. S. Bertet et al. "Investigation on Localisation Accuracy for First and Higher Order Ambisonics Reproduced Sound Sources". *Acta Acustica* 99.4 (2013), pp. 642–57.

[6] S. Spors et al. "Spatial Sound With Loudspeakers and Its Perception: A Review of the Current State". *Proceedings of the IEEE* 101.9 (2013), pp. 1920–38.

[7] M. Frank, F. Zotter, and A. Sontacchi. "Localization Experiments Using Different 2D Ambisonics Decoders". In: *VDT International Convention*. 2008.

5.1.1 Method

The different loudspeaker arrays were simulated via the dynamic binaural synthesis system described in Section 4.1. The test procedure was the same as described in Section 4.2, including the pointing method, test setup, and white noise pulse. That means that the listener listened to a binaural simulation of a sound field synthesis system synthesizing a white noise pulse. Thereafter, the listener was to look into the direction from which she perceived the noise and press a key, with the laser pointer mounted on the headphones, providing her visual feedback about her viewing direction.

In the following, the different sound field synthesis conditions, loudspeaker setups, and test participants will be described for every experiment. The test participants were financially compensated for their effort. All of them had self-reported normal hearing. The presentation of the conditions was randomized, with the exception that – due to limited computing power – all conditions belonging to one source model were presented together. The order of the source models for every listener was again randomized.

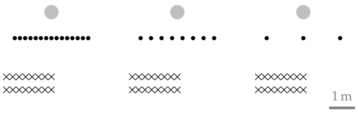


Figure 5.1: Setup for Experiment 1. The position of the synthesized source is indicated by the grey point. The position of the listener by black crosses and secondary sources by black dots. ☞

Experiment 1: WFS, Linear Loudspeaker Array Three different linear loudspeaker setups were considered in the experiment. The length of the loudspeaker array was always 2.85 m, measured from the center of each edge loudspeaker. The center of the loudspeaker array was placed at $(0, 0, 0)\,m$. The number of loudspeakers varied, including 3, 8, and 15 loudspeakers. This corresponds to a distance of adjacent loudspeakers of 1.43 m, 0.41 m, and 0.20 m. For each loudspeaker setup a point source was synthesized at $(0, 1, 0)$ m with WFS using (2.64) to calculate the driving functions.

The listeners were placed at 16 different positions: $(0, -1.5, 0)$ m, $(-0.25, -1.5, 0)$ m, $(-0.5, -1.5, 0)$ m, $(-0.75, -1.5, 0)$ m, $(-1, -1.5, 0)$ m, $(-1.25, -1.5, 0)$ m, $(-1.5, -1.5, 0)$ m, $(-1.75, -1.5, 0)$ m, $(0, -2, 0)$ m, $(-0.25, -2, 0)$ m, $(-0.5, -2, 0)$ m, $(-0.75, -2, 0)$ m, $(-1, -2, 0)$ m, $(-1.25, -2, 0)$ m, $(-1.5, -2, 0)$ m, $(-1.75, -2, 0)$ m – compare Figure5.1. Only positions in the left half of the listening area were considered due to the symmetry of the problem. Because everything was simulated via binaural synthesis the listener was able to switch instantaneously between the different positions – see Chapter 4.

Three different loudspeaker setups and 16 different listening positions led to a total of 48 conditions, which were presented five times to every listener. The listening experiment was split into two sessions to avoid fatigue: one session for the listener positions with a y-position of -1.5 m and the other for a y-position of -2 m. Additionally, each session included ten presentations of a real loudspeaker at an azimuth of $-5.7°$. For the array with 8 loudspeakers the test participant's viewpoint in the simulation was rotated by $35°$, and for the array with 15 speakers by $17.5°$. This was done to ensure an evenly distribution of the virtual source positions to the left/right of the listener.

11 listeners were recruited for the experiment – aged 21 to 33 years. Four of them had prior experiences with psychoacoustic testing and WFS. One test participant was removed from the analysis, because the standard deviation for the repeated conditions was approximately three times as high as for the average listener.

Experiment 2:* WFS*, Circular Loudspeaker Array The experiment consisted of three different circular loudspeaker setups. The diameter of the loudspeaker array was 3 m. The center of the loudspeaker array was placed at $(0,0,0)$ m. The number of loudspeakers varied between 14, 28, and 56 loudspeakers. This corresponds to a distance of adjacent loudspeakers of 0.67 m, 0.34 m, 0.17 m. For every loudspeaker setup a point source placed at $(0,2.5,0)$ m, a plane wave traveling into the direction $(0,-1,0)$, and a focused source placed at $(0,0.5,0)$ m were synthesized with WFS using (2.64), (2.57) and (2.76) to calculate the driving functions.

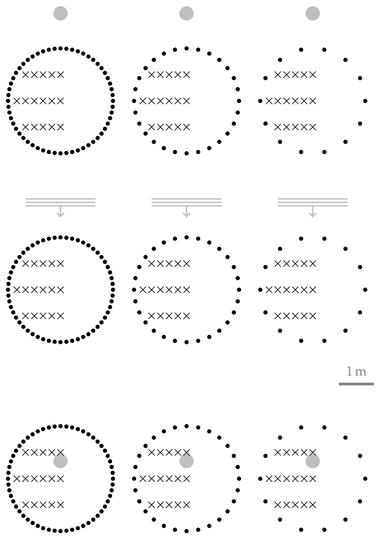


Figure 5.2: Setup for Experiment 2. The position of the synthesized source is indicated by the grey point. The position of the listener by black crosses and secondary sources by black dots. ☞

The test participants were placed at 16 different positions: $(0,0.75,0)$ m, $(-0.25,0.75,0)$ m, $(-0.5,0.75,0)$ m $(-0.75,0.75,0)$ m, $(-1,0.75,0)$ m, $(0,0,0)$ m, $(-0.25,0,0)$ m, $(-0.5,0,0)$ m, $(-0.75,0,0)$ m $(-1,0,0)$ m, $(-1.25,0,0)$ m, $(0,-0.75,0)$ m, $(-0.25,-0.75,0)$ m, $(-0.5,-0.75,0)$ m, $(-0.75,-0.75,0)$ m, $(-1,-0.75,0)$ m – compare Figure 5.4.

Three different loudspeaker setups, 16 different listening positions, and three different source types resulted in a total of 144 conditions, which were presented five times to every listener. The measurement was split up in two days, one session lasted approximately 45 minutes.

To ensure a more equal distribution of the presented locations of the sound events a pseudo-randomized jitter was added to the listener's viewpoint. They were chosen in a way that the position of the sound event always was within the boundary of ±30°.

12 listeners were recruited for the experiment – aged 23 to 33 years. One of them had prior experiences with psychoacoustic testing and WFS.

Experiment 3:* NFC-HOA*, Circular Loudspeaker Array Here, the same circular loudspeaker setups and listening positions as described for Experiment 2 were applied. For every loudspeaker setup a point source placed at $(0,2.5,0)$ m and a plane wave traveling into the direction $(0,-1,0)$ were synthesized with band-limited NFC-HOA using (2.50) and (2.45) to calculate the driving functions. The time domain implementations of the driving functions were realized as filters. For the loudspeaker setup with 14 loudspeakers, both sources were also synthesized with NFC-HOA using an order of $M = 28$.

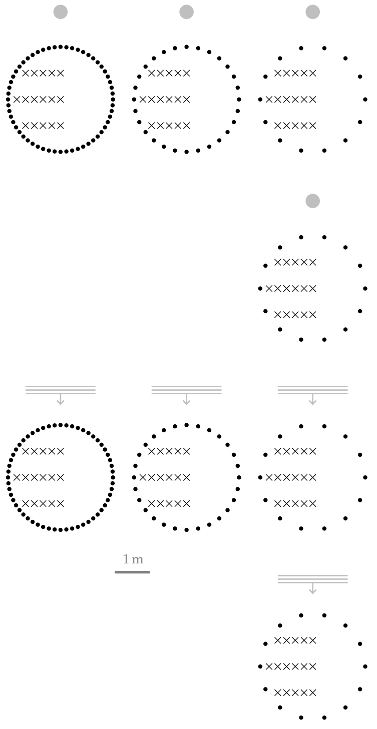


Figure 5.3: Setup for Experiment 3. The position of the synthesized source is indicated by the grey point. The position of the listener by black crosses and secondary sources by black dots. ☞

Four different loudspeaker setups and orders of spherical harmonics combinations, 16 different listening positions, and two different source types resulted in a total number of 128 conditions, which were presented five times to every listener. The measurement was split up in two days, one session lasted approximately 40 minutes.

The positions of the sound events were again jittered as described for Experiment 2.

A pre-test showed that the synthesis of one point source or one plane wave could lead to two auditory events coming from different directions that could be quite far away from each other. Therefore, in this experiment it could not be assured for all conditions that the auditory event was perceived in a region ±30°. This is important, because the pointing method has some deviations for angles outside of this region – compare Figure 4.4. In addition, the instruction for the test participants were slightly updated and they were told to look into the direction of the more pronounced source, if they heard more than one. In cases where they were not able to state which was more pronounced, they were instructed to randomly choose one of the sources.

12 listeners were recruited for the experiment – aged 24 to 35 years. Three of them had prior experience with psychoacoustic testing and sound field synthesis. One of the listeners completed only the condition with plane wave as source model and one completed only the condition with point source as source model. Two test participants were excluded from the analysis, because their standard deviation for the five repetitions was more than twice as large as for the other participants.

Experiment 4:* NFC-HOA*, Number of Sources Due to the fact that for some conditions in Experiment 3 more than one auditory event was audible, a post-test was conducted. The listeners were asked to indicate on a keyboard how many sources they heard: one or two? The same conditions as in Experiment 3 were used, but this time each condition was only repeated four times. All conditions were presented in one session lasting approximately 40 minutes.

7 listeners were recruited for the experiment – aged 23 to 33 years. Three of them had prior experience with psychoacoustic testing and sound field synthesis.

5.1.2 *Results*

Figure 5.4 summarizes the results of all four experiments. For every sound field synthesis method, the used loudspeaker setups are drawn as black dots and the synthesized sources are indicated by the grey symbols. At every listener position an arrow is pointing towards the average direction from which the listeners perceived the corresponding auditory event. The added random jitter of the head orientation of the listener at every position is already compensated in the presented arrows. The color of each arrow displays the localization error, which is defined as the absolute deviation between the desired sound event direction and the direction of the auditory event. The absolute deviation is represented by the color, ranging from light yellow for 0° to dark red for values of 40° or higher. For the condition of the synthesized point source, the perceived direction

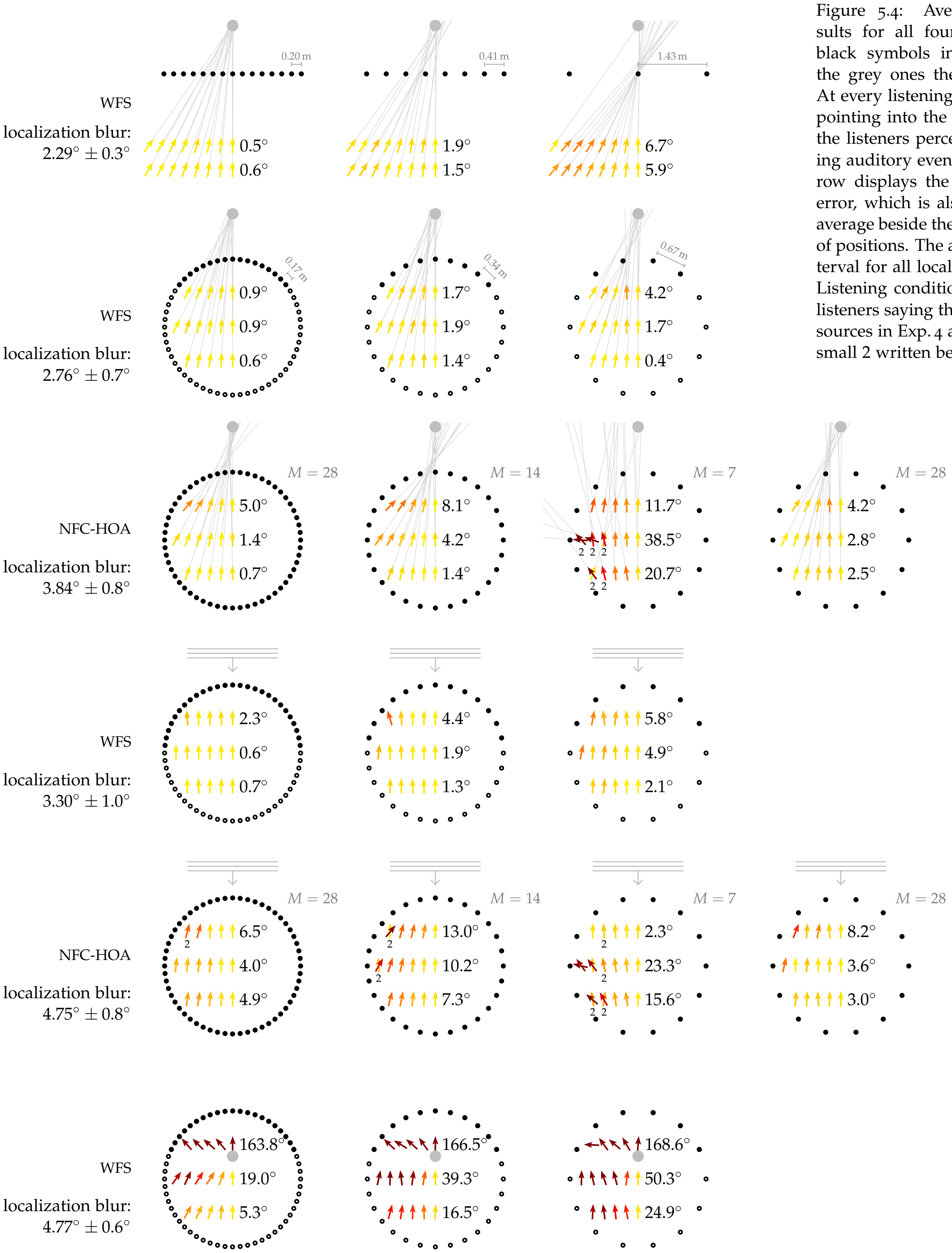


Figure 5.4: Average localization results for all four experiments. The black symbols indicate loudspeakers, the grey ones the synthesized source. At every listening position, an arrow is pointing into the direction from which the listeners perceived the corresponding auditory event. The color of the arrow displays the absolute localization error, which is also summarized as an average beside the arrows for every row of positions. The average confidence interval for all localization results is 2.3°. Listening conditions which resulted in listeners saying that they perceived two sources in Exp. 4 are highlighted with a small 2 written below the position. ☞

of the listener is additionally highlighted by a small grey line going into this direction. The results from the fourth experiment, where no direction but only the number of perceived sources is the outcome, are included by adding a small "2" below the position of all conditions where two sources were perceived. For no condition more then two sources were perceived.

The localization error for WFS synthesizing a point source or a plane wave is approximately 0.9° in the case of loudspeaker spacings around 20 cm. Only the position $(-1, 0.75, 0)$ m for the synthesis of a plane wave deviates from this pattern and yielded an error of around 5°. For the synthesis of a plane wave with WFS the dependency of the localization error on the position is more pronounced for larger loudspeaker spacings. Here, especially the positions to the side near the loudspeakers lead to larger localization errors than in the case of the point source conditions. For a loudspeaker spacing around 40 cm the localization error increases only slightly to an average of 2°. For larger loudspeaker spacings the localization error increases and varies for different listening positions. In addition, the listeners start to look into the direction of the nearest loudspeaker instead of the direction of the synthesized point source. This is most prominent for the linear loudspeaker array with only three loudspeakers and a loudspeaker spacing of 1.43 m.

The localization error for band-limited NFC-HOA synthesizing the same point source is larger at all positions, starting at 3.8° for a loudspeaker spacing of 17 cm and 7.4° for a spacing of 34 cm. The results are more dependent on the listening position as for the WFS conditions, showing stronger errors for positions to the side. In the case of the loudspeaker array with 14 loudspeakers, the localization error for the point source condition is larger than 10° for most of the positions to the side. In addition, for five positions to the side the listeners reported that they heard more than one auditory event.

For the case of a loudspeaker array with 14 loudspeakers, NFC-HOA up to an order of $M = 28$ was also tested. An order of 28 corresponds the order of band-limited NFC-HOA for the loudspeaker array with 56 loudspeakers. In this case, the results are very similar to the ones of the WFS conditions for 14 loudspeakers. The overall localization error is slightly larger than for WFS. In contrast the pattern is very similar for the point source as well as for the plane wave conditions, meaning that the localization error now has similar values for all positions across the listening area and only one auditory event is perceived at all positions.

In order to analyze the influence of more than one auditory event on the localization ratings, the distributions of the reported directions of auditory events from all listeners were visually proofed for normal distribution. An example is presented in Figure 5.5 for the point source condition at the listening position $(-1, -0.75, 0)$ m. The distributions of ratings of 9 listeners are shown in comparison for WFS and NFC-HOA with an order of 28 and of 7. For the case of WFS

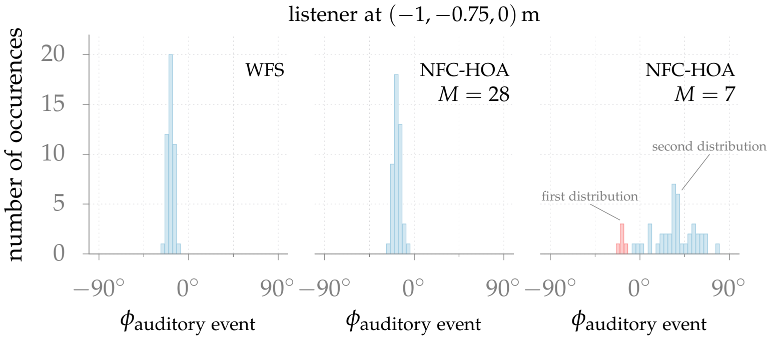


Figure 5.5: Distributions of the directions of the auditory event as rated by the listeners at the position $(-1, -0.75, 0)$ m for the loudspeaker array with 14 loudspeakers. The results for a synthesized point source for WFS and NFC-HOA for different orders M are shown. For WFS only the results for the first 9 listeners were analyzed to have an equal number of answers as in the case for NFC-HOA. ☞

and NFC-HOA with an order of 28, a normal distribution is visible. On the other hand, for the case of NFC-HOA with an order of 7 the distribution is far more spread and the data are most probably characterized by more than one normal distribution. In all cases where visual inspection leads to the assumption of two underlying normal distributions, a Gaussian mixture model was applied to estimate what data point belongs to what distribution. This was repeated for 100 iterations for each position. Afterwards the data points were assigned to their belonging average. The two distributions and their corresponding data points are indicated by two different colors in Figure 5.5. After the assignment to a particular distribution, the average direction was calculated for every distribution. In Figure 5.4 two arrows, one for each corresponding direction were drawn for all positions where more than one normal distribution was found.

For WFS, the localization of focused sources placed in the audience area is investigated in the following. The results are shown at the bottom row of Figure 5.4. The focused source was placed at $(0, 0.5, 0)$ m and was heading towards $(0, -1, 0)$ m, which means that the five listener positions with $y = 0.75$ m were placed between the focused source and the active loudspeakers. For these positions, the listeners were not looking into the direction of the focused source but into the direction of the active loudspeakers, which leads to large localization errors, because the loudspeakers are up to 180° opposite of the focused source direction. In addition, it could be observed that for the loudspeaker arrays with 14 and 28 loudspeakers only a small region with low localization error exists around the central listening positions. For positions to the side, the listeners were again pointing more into the direction of the active loudspeakers. Only for the loudspeaker array with 17 cm spacing between the loudspeakers, a triangle-shaped listening area can be identified where the localization error is around or less than 10°.

5.1.3 Discussion

The ability to localize a sound that is synthesized by WFS is very good in the whole audience area, independent of the form of the secondary sources. The localization error for a synthesized point source and a synthesized plane wave is below 5° on average and is only degraded for positions in the proximity of around 20 cm to the loudspeakers. Only if a small number of loudspeakers is employed, the localization error will become the same as for the case of stereophony, meaning that the direction of a single loudspeaker rather than of the desired source will be perceived. This was the case for the linear loudspeaker array with 3 loudspeakers in the presented tests.

The localization accuracy for the same secondary source setups driven by band-limited NFC-HOA is inferior to that of WFS. Only the secondary source distribution employing 56 sources is capable of providing a localization error smaller than 5° in most of the audience area. For fewer secondary sources, large localization errors occur outside the center of the audience area. In the case of 14 sources, the listeners start to perceive more than one source and hear single loudspeakers. If in the case of 14 sources the order of NFC-HOA is chosen not to be band-limited to 7, but going up to 28 as it would be the case for 56 sources, the localization accuracy is comparable to that of band-limited NFC-HOA with the same order and 56 sources. This highlights that in the case of NFC-HOA the applied order is very critical for the localization accuracy in the whole audience area. If the order is reasonably high, localization performance will be identical to hat of WFS. Otherwise it will be impaired outside of the center of the audience area. This is not surprising, remembering Figure 3.13 where the impulse responses of different SFS systems were shown. WFS and NFC-HOA behave similarly if the order of NFC-HOA is high enough, in opposition to the splitting of the impulse for different frequency regions outside of the center for band-limited NFC-HOA – compare also Figure 3.15.

By comparing the localization blur it can be seen that for NFC-HOA the synthesized sources will be perceived on average to have a slightly larger spatial extent than for WFS.

FOR THE SPECIAL CASE of the synthesis of a focused source with WFS, the localization accuracy depends strongly on the secondary source distribution. In contrast to the case of a plane wave or point source, a synthesized focused source achieves an extended listening area only for the setup with 56 sources. The localization blur further indicates that focused sources are perceived to be the widest synthesized sources in all of the experiments. This can be explained by the fact that the size of the focal point is limited by the wavelength of the sound, leading to a large focal point size especially for low frequencies.

5.1.4 Conclusion

If a sound field synthesis system is desired that is authentic with regard to the localization accuracy, WFS and a distance between the loudspeakers of at least 20 cm can already lead to satisfactory results. If a larger distance of loudspeakers is applied, WFS still leads to a very high localization accuracy in the whole audience area. Especially in these cases it can deliver a superior localization accuracy compared to band-limited NFC-HOA which has large localization errors outside of the sweet-spot. Inside the sweet-spot the localization accuracy is authentic in terms of the desired result as well. But outside, more than one source could be perceived.

If NFC-HOA is desired for synthesis and good localization accuracy should be achieved in the whole audience area, the Ambisonics order should be increased to a reasonable number. In that case, NFC-HOA becomes comparable to WFS, which constitutes a high-frequency approximation of NFC-HOA with infinite order.

5.2 *Timbral Fidelity*[8]

Section 3.3 showed that a limited number of secondary sources leads to a repetition of parts of the desired synthesized signal. In the last section, the influences on localization have been demonstrated. At this point, the influences of sound field errors on the timbre of the corresponding auditory event will be discussed. As already seen in Figure 3.12 the repetitions will change the spectrum of the synthesized signal, and hence the timbre of the corresponding auditory event is expected to change. But repetitions can also occur in closed spaces. Hence, it is possible that the auditory event related with the synthesized sound could create the impression of being situated in a room.

First, the definitions and perceptual features of timbre and coloration will be discussed. They will be considered especially for cases of repeated sound events, and their connection to the perceptual features of a room is outlined. Afterwards, listening test results for coloration in WFS are shown and supplemented by an experiment investigating the dependence of the perceived coloration on the number of secondary sources.

All definitions of timbre are "negative", they state what timbre is not . This leads to the circumstance that the definition of timbre already has a big influence on the resulting research questions. Timbre is most often defined as "that attribute of auditory sensation which enables a listener to judge that two nonidentical sounds, similarly presented and having the same loudness and pitch, are dissimilar".[9] Plomp added "same duration" to the list of properties of the nonidentical sounds.[10] To highlight the wide range of features that are included in such a definition, Patel[11] provides an analogy. Timbre is similar as if describing "looks" of human faces, where "looks" is that attribute which enables an observer to judge that two nonidentical faces with the same height, width and complexion, are dissimilar. It is obvious that timbre is a multidimensional percept and the number of dimensions that can be detected in an experiment depends highly on the used stimuli.

If the difference of two points in the timbral space is assessed, it

[8] This experiment was published in a slightly modified way in H. Wierstorf et al. "Coloration in Wave Field Synthesis". In: *55th Audio Engineering Society Conference*. 2014, Paper 5.3, the presented experiment was carried out in coorperation with Christoph Hohnerlein as part of his Bachelor thesis.

[9] ANSI. *American National Standard Acoustical Terminology, ANSI S1.1-1994*. New York, 1994.

[10] B. C. J. Moore. *An Introduction to the Psychology of Hearing*. Bingley: Emerald, 2012, p. 285.

[11] A. D. Patel. *Music, Language And The Brain*. New York: Oxford University Press, 2010.

is described as *coloration*, whereby one of the points is considered as the uncolored reference and the other point is considered as colored. The reference point can explicitly be presented to a listener, or it is implicitly known to the listener due to her experience. The latter has lead to the formation of an internal reference. One of the complicating aspects of coloration is that the metric of the timbral space is not known and it could be non-trivial. In the literature an euclidean metric[12] or a weighted euclidean metric[13] is commonly assumed, but cannot be assured. Another questionable assumption that is often made is the negative connotation of coloration. For example Brüggen[14] defined the reference as the desirable point and coloration as the move in timbral space to an adverse point. This statement makes the implicit assumption that there is only one point in timbral space that corresponds with a high perceived sound quality and that the reference should always be placed at this point.

One problem of the above definition of timbre is that only three perceptual aspects are directly named that should be constant between different stimuli. Whereas it is not specified which other dimensions the phrase "similarly presented" should include. For example, is it still similar if one stimulus is presented in an anechoic chamber and the other in an office? In order to clarify this situation some authors have included more aspects in the indirect definition of timbre. Letowski[15] gives a definition of timbre that explicitly adds spatial perception to the list of attributes no covered by timbre. Emiroglu[16] has a similar approach stating: "The label timbre combines all auditory object attributes other than pitch, loudness, duration, spatial location and reverberation environment."

As mentioned at the beginning of this section, this thesis is especially interested in the influence of repetitions or reflections of the sound signal on its perceived timbre. Hence, the influence of the room perception on timbre has to explicitly be considered. In the literature, there are mainly two phenomena investigated in this context. One is the influence of different rooms on the coloration of an auditory event. The other deals with the fact that the coloration of an auditory event placed in a room is different depending on the listener's usage of only one or both of her ears when listening to the sound. The second one is summarized under the term *binaural decoloration*.[17]

A straightforward explanation of both phenomena is presented by Brüggen.[18] He defined timbre after the ANSI definition[19] and subsumed any kind of spatial impression due to the room under coloration. In the choice of attributes he followed Berkley[20] who found the dimensions *echo* (related to reverberation) and *color* (related to spectral deviation) using a multidimensional scaling method for sound events with reflections. Here, the echo dimension is mainly influenced by late reflections and the color dimension by early reflections. The binaural decoloration phenomenon is then explained via a blind system identification that tries to identify the part of the spec-

[12] See for example R. Plomp, L. C. W. Pols, and J. P. van de Geer. "Dimensional Analysis of Vowel Spectra". *The Journal of the Acoustical Society of America* 41.3 (1967), pp. 707–12.

[13] An example and discussion of different metrices is presented in S. McAdams et al. "Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes." *Psychological Research* 58.3 (1995), pp. 177–92.

[14] M. Brüggen. "Klangverfärbungen durch Rückwürfe und ihre auditive und instrumentelle Kompensation". PhD thesis. Ruhr-Universität Bochum, 2001, p. 8; note that on p. 13 he relativates his opinion by stating that for performances such as music played in a concert hall the coloration due to the room is a desired one and the perceived quality of the sound is better for the colored case.

[15] Letowski, op. cit.

[16] S. S. Emiroglu. "Timbre perception and object separation with normal and impaired hearing". PhD thesis. Carl-von-Ossietzky-Universität Oldenburg, 2007, p. 89.

[17] It was first reported by W. Koenig. "Subjective Effects in Binaural Hearing". *The Journal of the Acoustical Society of America* 22.1 (1950), pp. 61–62.

[18] Brüggen, op. cit.

[19] ANSI, op. cit.

[20] D. A. Berkley. "Hearing in rooms". In: *Directional Hearing*. Ed. by W. A. Yost and G. Gourevitch. New York: Springer, 1987, pp. 249–60.

trum that has been contributed by the room, and removes that from the spectrum of the sound source placed in that room. With this mechanism, Brüggen was able to explain his results regarding the coloration of a sound source placed in different rooms.[21]

To explain the binaural decoloration phenomenon for stereophony, Theile[22] proposed his association model. The model says that a listener associates a single location to two sound events if they are characterized by the same signal. In the case of stereophony the listener is then able to carry out a binaural decoloration of the corresponding perceived single auditory event. Obviously the association model has problems to predict the same amount of binaural decoloration for a sound source placed in a room, where the number of different locations of the sound events due to reflections is higher than two.

A shortcoming of the proposed binaural decoloration mechanism is its independence from the task or context of the listener. For example, Olive et al.[23] published a study where they showed an influence of room acoustics on absolute quality ratings of loudspeakers via dynamic binaural synthesis.

For sound field synthesis, only a few investigations of the coloration properties are available, although timbral fidelity seems to be one of the most important parts for rating the quality of a spatial audio system.[24] Wittek[25] has investigated the differences in intra-system coloration between WFS and stereophony, using loudspeaker arrays with different spacings. He asked the listeners if they perceive a timbral difference between a reference source coming from 5° and the given test stimuli coming from other directions. The reference source and the test stimuli were always presented by the same system, leading to an assessment of the coloration differences that is inherent to each system. These differences were rated on a scale ranging from *no difference* to *extremely different*. The listeners were centrally seated at a distance of 1.5 m from the array, and pink noise bursts were presented as source signal. The test stimuli were generated via dynamic binaural synthesis. Figure 5.6 summarizes the results. For a loudspeaker spacing of 3 cm, the intra-system coloration of WFS was comparable to the case of stereophony and single loudspeakers. For larger loudspeaker spacings ranging from 12 cm to 48 cm, the intra-system coloration was perceived as being stronger but independent of the different loudspeaker spacings.

De Bruijn[26] investigated the variation of timbre for WFS within the listening area for linear loudspeaker arrays with different spacings. He found large differences in terms of coloration for loudspeaker spacings of 0.5 m and negligible differences for a spacing of 0.125 m. As source stimulus, speech shaped noise was applied. This choice of stimulus explains why he observed less coloration for larger spacings than Wittek.

For NFC-HOA no results are available. For HOA with different orders, Solvang[27] showed that there will be stronger coloration near the sweet-spot if too many loudspeakers are used for a given order.

[21] Brüggen, op. cit., Fig. 5.11.

[22] G. Theile. "Über die Lokalisation im überlagerten Schallfeld". PhD thesis. Technische Universität Berlin, 1980.

[23] S. E. Olive et al. "The Variability of Loudspeaker Sound Quality Among Four Domestic-Sized Rooms". In: *99th Audio Engineering Society Convention*. 1995, Paper 4092.

[24] Rumsey et al., op. cit.

[25] Wittek, op. cit.

[26] W. P. J. de Bruijn. "Application of Wave Field Synthesis in Videoconferencing". PhD thesis. Technische Universiteit Delft, 2004.

[27] A. Solvang. "Spectral Impairment for Two-Dimensional Higher Order Ambisonics". *Journal of the Audio Engineering Society* 56.4 (2008), pp. 267–79

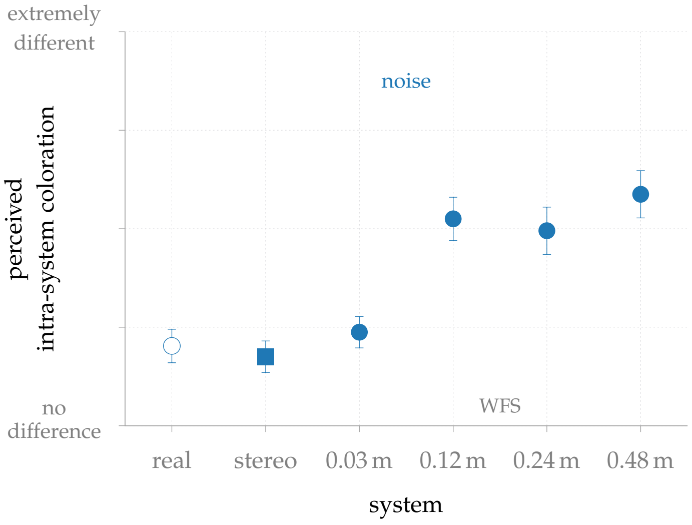


Figure 5.6: Average results with confidence intervals for the following question: Is there a timbral difference between the reference and the stimulus? Whereby the reference and the other stimuli were presented by the same system each time, leading to the measurement of intra-system coloration. The average is calculated over all subjects and the different positions of the sources. All loudspeakers, including real, stereo, and WFS, were simulated via binaural synthesis. The results are replotted from H. Wittek. "Perceptual differences between wavefield synthesis and stereophony". PhD thesis. University of Surrey, 2007, Fig. 8.6. ☞

In the following an experiment is described that compares the coloration of several WFS setups and a stereophonic setup to the reference case of a single loudspeaker.

5.2.1 *Method*

The experiment was performed with the binaural simulation method as described in Section 4. The only difference is that the dynamic head-tracking part was disabled to exclude time-varying coloration due to head movements.

Stimuli In order to ask the listeners to judge changes in timbre, a point source placed at $(0, 2.5, 0)$ m was chosen as a reference stimulus, which was realized by using a single HRTF. The same point source was synthesized with WFS using (2.64) for several circular secondary source distributions. Each distribution had the same geometry with a radius of 3 m with its center at $(0, 0, 0)$ m, but different numbers of secondary sources, namely 14, 28, 56, 112, 224, 448, 896, 1 792, 3 584. For the distribution with 14 secondary sources, this corresponds to a distance of 67 cm between the individual secondary sources going down to 0.3 cm for the distribution with 3 584 sources. In addition, a stereophonic setup with two loudspeakers placed at $(1.4, 2.5, 0)$ m and $(-1.4, 2.5, 0)$ m was included leading to a total number of 10 different conditions, not counting the reference. All impulse responses were normalized to the same maximum absolute amplitude before convolving them with the audio material during the experiment.

Three different audio source materials were used. A pulsed pink noise train composed of 800 ms noise bursts with 50 ms windowing at the beginning and end and a pause of 500 ms between the bursts. This stimulus was also used by Wittek.[28] As a second stimulus, a twelve second clip from the electronic song "Luv deluxe" by "Cinnamon Chasers" was chosen. It is an instrumental song including

[28] Wittek, op. cit., Sec. 8.2.

cymbals and subtle white noise which may help revealing coloration to a similar degree as the pink noise stimulus does. The third stimulus was an eight second long female speech sample.

Procedure The listeners were asked to rate the difference in timbre between the reference stimulus and the other conditions on a continuous scale with the attribute pair *no difference* and *very different* at its end-points. This was accomplished with a MUSHRA test design, including a hidden reference and a lower anchor. The low anchor was created by high-pass filtering the reference condition with a second order Butterworth filter with a cutoff frequency of 5 kHz. The listeners were instructed to rate the coloration and not the differences in loudness or perceived externalization of the stimuli. They started with one training run before the real experiment began. The training consisted of a run with a central listening position, varying numbers of secondary sources and a different music track.

During a single run in the experiment, the participants had to rate all 10 different conditions, the hidden reference and the lower anchor for one given audio material. The stimuli were looped during the experiment and the listener could switch instantaneously between the conditions as often as she liked.

The listeners were placed at two positions in the audience area, at $(0, 0, 0)$ m and $(-1, 0, 0)$ m – see Figure 5.7. The central listening position was repeated two times, resulting in a total of nine runs.

To investigate only the influence of the position of the listener on coloration, another three runs were added. Here, the secondary source distribution with 56 sources was used and tested for the 11 different listening positions at $(0, 0, 0)$ m, $(-0.25, 0, 0)$ m, $(-0.5, 0, 0)$ m, $(-0.75, 0, 0)$ m, $(-1, 0, 0)$ m, $(-1.25, 0, 0)$ m, $(0, -0.5, 0)$ m, $(-0.25, -0.5, 0)$ m, $(-0.5, -0.5, 0)$ m, $(-0.75, -0.5, 0)$ m, $(-1, -0.5, 0)$ m, $(-1.25, -0.5, 0)$ m. The head of the listener was always orientated towards the source at all positions, to exclude a change of the direction the synthesized source was presented from. The synthesized point source for the listening position at $(0, 0, 0)$ m was used as the reference stimulus, which was also included as a hidden reference. In contrast to the other runs, no low anchor was included.

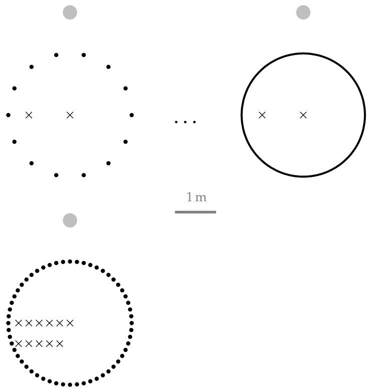


Figure 5.7: Experimental setup for coloration experiment. ☞

Participants 15 normal hearing listeners were recruited for the experiment – aged 23 to 29 years. None of them had prior experience with psychoacoustic tests.

Sound Field Synthesis Settings As mentioned in Section 3.3, the pre-equalization filter in WFS should only be applied up to the aliasing frequency. This was done for the different secondary source setups by investigating the amplitude spectrum at the position $(0, 0, 0)$ m and adjusting the lower and upper frequency limit of the pre-equalization filter to create an amplitude spectrum that is as flat as possible. As the aliasing frequency is dependent on the listener position, the optimization for $(0, 0, 0)$ m can lead to slight deviations at other

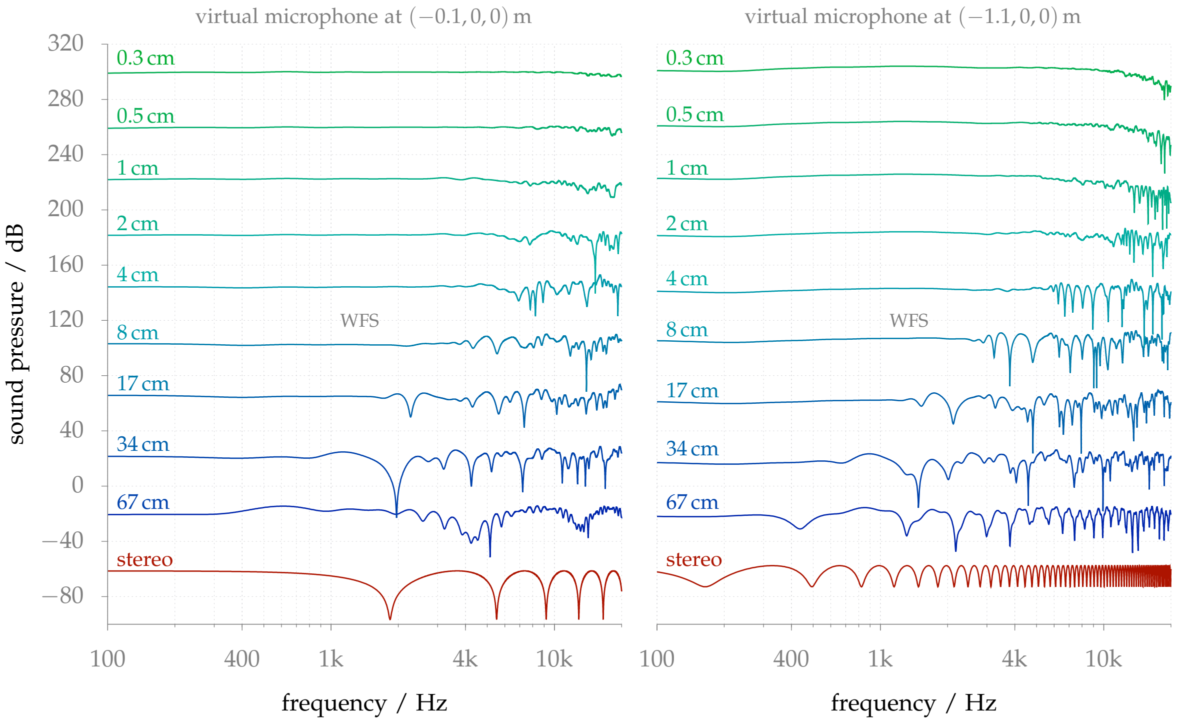


Figure 5.8: Amplitude spectra for the varying secondary source distribution conditions. The spectra was simulated for the place of the left ear of the listener. The left graph shows the spectra for the central listening position, the right for the off-center position. The distance between the secondary sources is given for all WFS spectra. The spectra are shifted in absolute magnitude in order to meaningfully display them. Parameters: $\mathbf{x}_\mathrm{s} = (0, 2.5, 0)$, $\mathbf{x}_\mathrm{ref} = (0, 0, 0)$ m, circular secondary source distribution with a diameter of 3 m. ☞

listener positions. Figure 5.8 shows the amplitude spectra of the impulse responses for the different secondary source distributions synthesizing a point source. The impulse responses were calculated for positions of the left ear of the test participants – excluding any HRTF. The calculation is identical to placing microphones at these positions and measuring the impulse responses. This has been done for the stereophonic setup as well.

The amplitude spectra highlight that the secondary source setup with 3 584 sources and a corresponding distance of 0.3 cm between them has a more or less flat frequency spectrum, whereas for lower numbers of secondary sources comp-filter like deviations in the spectrum occur. The lower the number of sources, the earlier these deviations are occurring, starting around 400 Hz for 67 cm. The deviations of the amplitude spectra for the listener position at $(-1, 0, 0)$ cm tend to start at lower frequencies compared to the ones at the central listening position. The stereophonic amplitude spectrum has a more regular comb-filter structure due to the involvement of only two loudspeakers. For the central position, deviations of the spectrum in the form of large dips occur slightly below 2 kHz. For the off-center listening position the deviations are spread along all frequencies and the number of dips are more than four times as large. On the other hand, the dips are no longer as deep as at the central position.

Figure 5.9 provides an overview of the amplitude spectra for the WFS conditions applying a secondary source distribution with 56 sources and a corresponding distance of 17 cm between them. The spectra are plotted for twelve different listening positions, as indi-

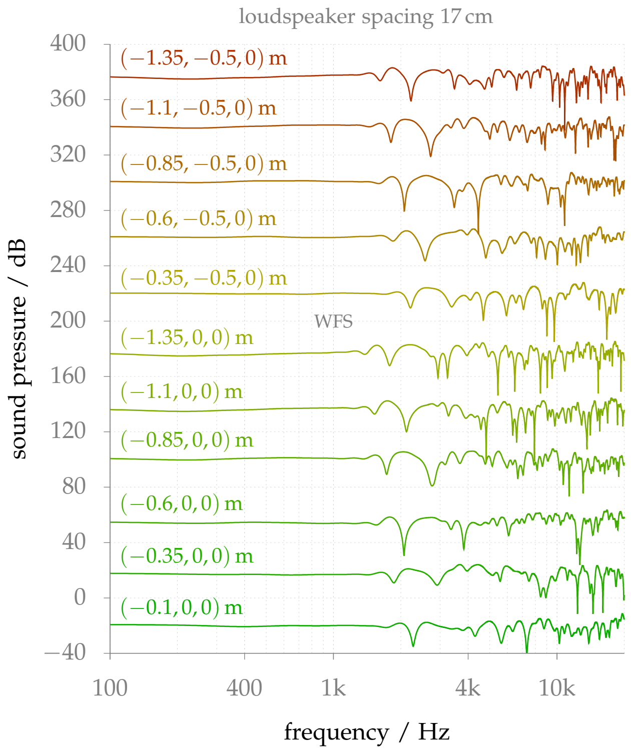


Figure 5.9: Amplitude spectra for WFS and a fixed secondary source distribution. The spectra are simulated for different positions as indicated by the colored labels in the figure. The spectra are shifted in absolute magnitude in order to display them. Parameters: $\mathbf{x}_\mathrm{s} = (0, 2.5, 0)$, $\mathbf{x}_\mathrm{ref} = (0, 0, 0)$, circular secondary source distribution with a diameter of 3 m. ☞

cated in the figure. The further the listener will move to the left of the audience area, the slightly lower the spatial aliasing frequency, which is visible in the form of an earlier start of the spectral deviations. By comparing the first dips of the spectra it can also be observed that the dips are shifted to higher frequencies for listener positions further to the back of the audience area.

5.2.2 *Results*

Figure 5.10 summarizes the results for the nine runs of the experiment, where the number of secondary sources was varied and the listener was positioned at $(0, 0, 0)$ m or $(-1, 0, 0)$ m. Only the results for pink noise and speech as stimuli are presented. The results for music were only significantly ($p < 0.05$) different from the ones for noise at two conditions, as indicated by an independent-samples Mann-Whitney U test. The results for the two center position runs are summarized by calculating the average for every listener before calculating the mean over all listeners. An independent-samples Mann-Whitney U test showed that the results of the repeated measurements were not significantly different from each other ($p < 0.05$), highlighting that the listeners were able to answer the task in a reliable way. The test participants rated the hidden reference as not different from the reference and the lower anchor as being very different from the reference. The overall ratings for the WFS stimuli show a clear dependency of the perceived coloration on the distance between the secondary sources. The system with the lowest distance was rated to be only slightly colored, whereas the system with the largest inter-

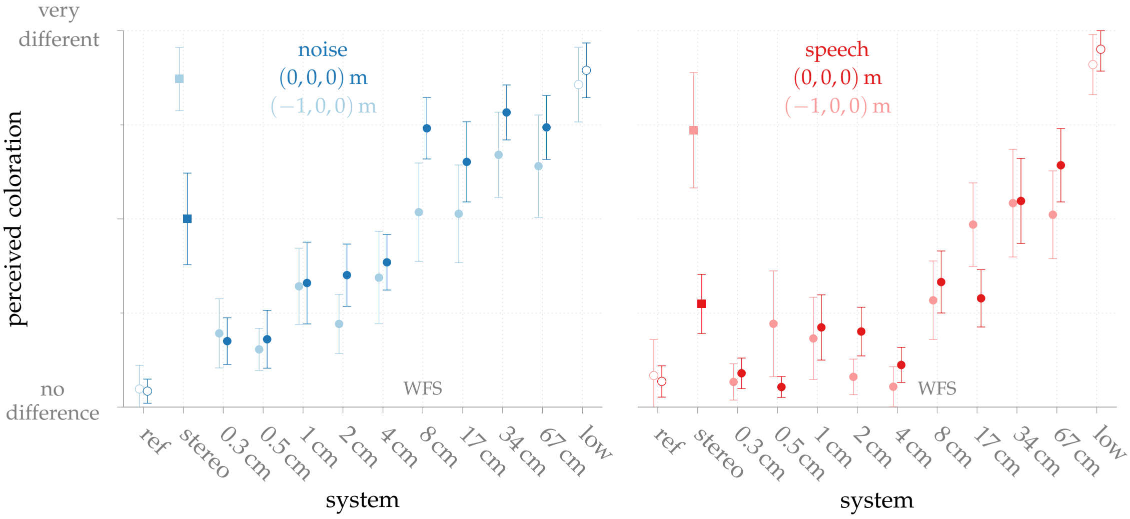


Figure 5.10: Average results with confidence intervals for the perceived coloration. Dark colors show results for the central listening position, lighter colors for the off-center position. ☞

loudspeaker distance was rated the most colored WFS systems. The listener position at $(-1, 0, 0)$ m exhibits a very similar pattern as the central listening position for WFS. For the stereophonic presentation, the perceived coloration is considerably more dependent on the listener position. The off-center position is rated as the most colored off all systems. In contrast, its perceived coloration at the central position is rated as being low to medium. A similar coloration was achieved with an inter-loudspeaker spacing of 8 cm for WFS.

When using the speech stimuli, coloration was consistently rated lower in comparison to the case of using noise stimuli. A WFS system with a inter-loudspeaker distance of 4 cm already achieved a transparent presentation for the speech stimulus in terms of coloration.

The other three runs of the experiment investigated the perceived coloration at different positions in the audience area for a WFS system with 56 secondary sources with a corresponding distance of 17 cm between the sources. Figure 5.11 summarizes the results. The condition with the listener at the center was the hidden reference and was not rated as being different. Most of the other positions were rated to be equally colored, where the noise stimuli were rated to be more

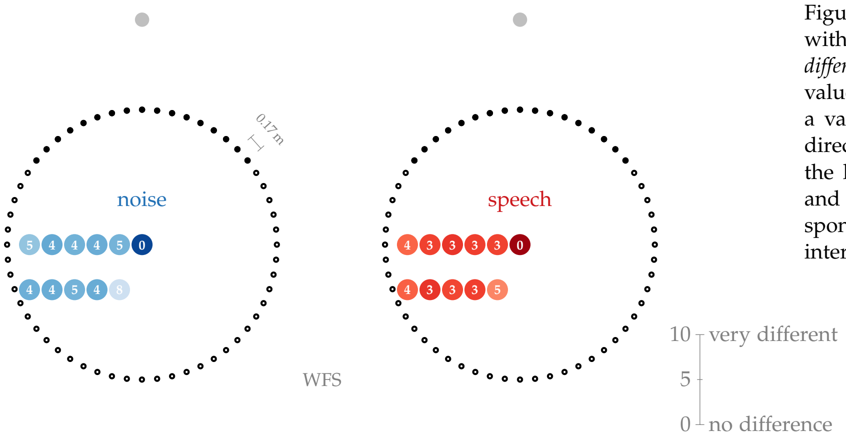


Figure 5.11: Perceived coloration rated with the attribute pair *very different*, *no difference*. The latter corresponds to a value of 0 in the figure, the former to a value of 10. The values are written directly at the listening position where the listener had to rate the coloration, and are further highlighted by a corresponding color. The average confidence interval is 1.2 over all positions. ☞

colored than the other two. Only the position at $(-0.25, -0.5, 0)$ m is deviating from that pattern by being perceived as more colored than all other positions.

5.2.3 *Discussion*

The results indicate that the number of secondary sources have a large influence on the perceived coloration for WFS. This is not a surprising result, reconsidering the magnitude spectra of the different systems as shown in Figure 5.8. Here, it is obvious that the spectrum deviates from a desired flat frequency response for frequencies above the aliasing frequency, which is directly dependent on the distance between adjacent secondary sources. In contrast to the localization results, where a distance of 17 cm already resulted in an authentic localization accuracy, the perceived coloration never vanishes for WFS and pink noise as stimulus. Even for an inter-loudspeaker spacing of 0.3 cm, slight coloration is perceived. Only for the speech stimulus and the inter-loudspeaker spacing of 0.3 cm the perceived coloration was indistinguishable from the hidden reference for both the central and off-center listening positions.

The results for stereophony suggest that sources presented by that method exhibit coloration, meaning that binaural decoloration is not able to suppress it completely. If the amplitude spectrum in Figure 5.8 is compared to the ones for WFS it could be concluded that the binaural decoloration has a larger impact on stereophony, because the perceived amount of coloration seems to be less than what may be predicted by the position of the first dip in the amplitude spectrum. Another possibility might be that the dips in the spectrum for stereophony are more smeared out by the auditory filters as it is the case for WFS.

The coloration ratings for WFS with 56 secondary sources at different listening positions revealed a more or less equal coloration to that obtained at the central listening position. However, this conclusion has to be relativized due to the multi-dimensionality of timbre. The fact that conditions are rated to have the same coloration compared to a reference condition does not necessarily imply that they have no relative coloration between each other.

By averaging the coloration results for noise from Figure 5.11, a value of 4.7 on a scale from 0 to 10 is found. This result is identical to the one Wittek[29] obtained for loudspeaker arrays with a distance between the loudspeaker of 12 cm and 24 cm, which were presented in Figure 5.6. The distance for the 56 secondary sources in the current experiment is 17 cm.

AN INSPECTION of the actual root mean square value of the presented signals revealed that there were fluctuations of up to 3 dB between the single conditions. Therefore, it could be that the listeners have included a loudness-related cue in their coloration rating even if they were advised not to do so. To further analyze this, the corre-

[29] Ibid.

lation between the actual root mean square values and the coloration ratings were calculated using the average for the speech and noise stimuli, which is 0.6 for the runs with the central listening position. This indicates that the loudness was not the main cue for the given coloration ratings. For the off-center listening position the correlation is 0.8. This highlights that the loudness could have a large contribution on the coloration ratings for this case. The correlation was also calculated for the conditions with different listening positions and a fixed number of loudspeakers. When averaging over speech and noise, the correlation results to 0.2. This indicates that the loudness did not have a major influence on the coloration ratings for these conditions. A more precise inspection of the listening position that was rated to be most colored, namely $(-0.25, -0.5, 0)$ m, nonetheless revealed that other factors could have influenced the coloration ratings. The position $(-0.25, -0.5, 0)$ m had the loudest signal and was reported as being less externalized compared to all other conditions by two test participants after the experiment.

The music and pink noise stimuli show no significant differences in all but two positions. This indicates that even the usage of music alone might be suitable to investigate the perceived coloration. That is of advantage, because most listeners considered the noise stimulus as unpleasant, as revealed by informal reports after the tests.

5.2.4 *Conclusion*

The results show a clear dependency of the perceived coloration of a synthesized point source from the given loudspeaker setup. The higher the inter-loudspeaker spacing, the more coloration will be perceived. This direct relation is due to the connection between the aliasing frequency and the distance between the secondary sources. The aliasing frequency specifies from which frequency onwards deviations in the amplitude spectrum of the synthesized source will appear, which seems to be a good measure for the perceived coloration of the synthesized source.

The aliasing frequency changes only to a small extent at nearby positions in the audience area, which seems to correspond with the results showing that the perceived coloration is similar at different positions in the audience area for WFS and 56 secondary sources.

For stereophony, the amount of coloration seems to be less than for a WFS system with a similar position of the first dip in the amplitude spectrum. This indicates that binaural decoloration may be more pronounced for stereophony than for WFS.

5.3 *Spectro-Temporal Artifacts*

In Section 3.3 the influence of discrete secondary source distributions on the perception of synthesized sound fields was discussed. For the synthesis of focused sources with WFS corresponding time signals were presented in Figure 3.16. It can be seen that for focused sources additional wave fronts arrive at the listener position before the desired wave front. This could have several implications for the perception of focused sources, because the auditory system is optimized for the opposite case of additional reflective wave fronts after a desired one, as it happens for example in rooms. This will be illustrated by the following example applying a loudspeaker array and Wave Field Synthesis.

Assume the synthesis of a point source placed at $(0, 1, 0)$ m behind a linear loudspeaker array composed of 34 secondary sources with a size of 20 m placed on the x-axis. If an impulse is played back as audio signal through this system and the ear signal at the right ear of a listener placed at $(4, -4, 0)$ m is recorded it will look like the left time signal in Figure 5.12. The direct sound and a bunch of early repetitions have the highest magnitude. Later repetitions are arriving up to 50 ms after the first wave front, but are lower in magnitude the later they arrive. In the upper part of the figure the corresponding frequency response is presented. Clear peaks and dips due to the repetitions are visible. Now assume a system with the same magnitude spectrum, but a complete different time pattern. This can be achieved by time reversing the impulse response as shown in the bottom right of Figure 5.12. In reality this kind of impulse responses can occur for a WFS system synthesizing a focused source due to the time reversing technique that is involved to achieve the desired source model. If the impulse response from the WFS point source is convolved with a speech signal it adds some coloration to this signal. If the time reversed version of the same impulse response is convolved with the same speech signal two additional auditory events occur. One colored version of the original one coming from another direction and an additional auditory event consisting of spectro-temporal artifacts coming from the same direction as one of the speech signals. The two speech signals can be downloaded and listened to via headphones by clicking on the two *listen* links.

Because the spectro-temporal artifacts are not occurring in a natural environment and their perception is likely multi-dimensionally a first listening test was conducted to identify important perceptual dimensions for focused sources. A second experiment investigated the influence of the size of the secondary source distribution on these perceptual dimensions.

5.3.1 Experiment 1: Perceptual Dimensions of Focused Sources in WFS[30]

To accommodate the perceptual multi-dimensionality the repetory grid technique (RGT) was used in a first experiment to identify relevant perceptual attributes.[31] With this method, in a first step each participant creates her own set of attributes and in a second step uses respective attribute scales for rating her perception. No attributes are

[30] This experiment was done in collaboration with Matthias Geier and parts of this section are published in M. Geier et al. "Perceptual Evaluation of Focused Sources in Wave Field Synthesis". In: *128th Audio Engineering Society Convention*. 2010, Paper 8069.

[31] G. A. Kelly. *The Psychology of Personal Constructs*. New York: Norton, 1955.

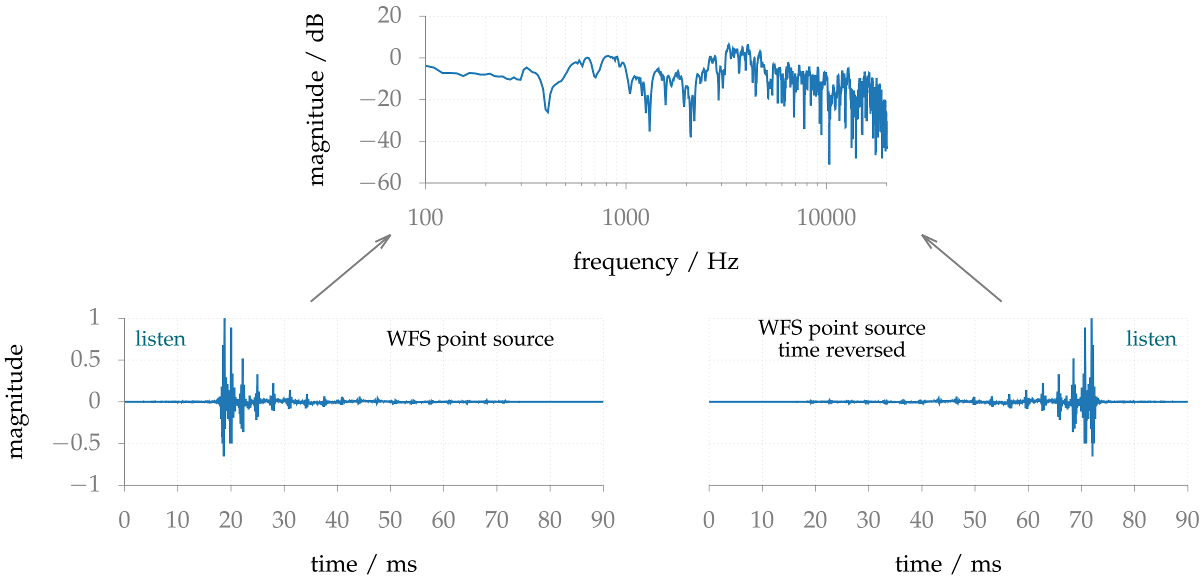


Figure 5.12: Impulse response and amplitude spectrum of a point source synthesized by WFS (2.64). Beside the impulse response its time reversed version is shown. Both impulse responses were convolved with a speech signal which can be downloaded via the *listen* links. Parameters: $\mathbf{x}_\mathrm{s} = (0, 1, 0)$, $\mathbf{x}_\mathrm{ref} = (4, -4, 0)$ m, linear secondary source distribution with a length of 20 m and 34 sources. ☞

provided by the experimenter, and, thus, the test subject has complete freedom in the choice of attributes. Berg and Rumsey were the first to apply the RGT towards perception in spatial audio.[32]

Stimuli The test was conducted via the dynamic binaural synthesis system including binaural simulation of the secondary sources as presented in Chapter 4. Two linear secondary source distributions with a length L of 4 m and 10 m, and a loudspeaker spacing of $\Delta x_0 = 0.15$ m were synthesized – compare Figure 5.13. For the binaural simulation of the loudspeakers HRTFs of the FABIAN manikin[33] measured in the anechoic chamber of Technical University Berlin were applied. The focused source was placed at $(0, -1, 0)$ m in front of the secondary sources. It was synthesized with 2.5D WFS by applying the driving function 2.76.

As discussed in Section 3.3, the aliasing frequency f_al depends on the listener position, therefore the WFS pre-equalisation filter was calculated separately for each simulated listening position. Coloration introduced by an improper choice of the pre-equalisation filter was not part of the investigation and should be avoided.

For both arrays, three different listener positions on a given circle around the focused source were used. The radius was $R = 1$ m for the short array and 4 m for the long array. Three different listener angles of $\phi = 0°$, $30°$ and $60°$ were applied for both array lengths – see Figure 5.13. These result in the following six listener positions: $(0, -2, 0)$ m, $(-0.5, -1.9, 0)$ m, $(-0.9, -1.5, 0)$ m, $(0, -5, 0)$ m, $(-2, -4.5, 0)$ m, $(-3.5, -3, 0)$ m. These six configurations will be referred to as $0^\circ_{4\,\mathrm{m}}$, $30^\circ_{4\,\mathrm{m}}$, $60^\circ_{4\,\mathrm{m}}$, $0^\circ_{10\,\mathrm{m}}$, $30^\circ_{10\,\mathrm{m}}$, $60^\circ_{10\,\mathrm{m}}$. In all conditions, the listener was always looking into the direction of the focused source. A seventh, reference condition ("ref") was created, which consisted of a single sound source located at the position of the focused source. This was realized by directly using the corresponding HRTF from the

[32] J. Berg and F. Rumsey. "Spatial Attribute Identification and Scaling by Repertory Grid Technique and other methods". In: *16th Audio Engineering Society Conference*. 1999, pp. 51–66

[33] A. Lindau and S. Weinzierl. "FABIAN – An instrument for software-based measurement of binaural room impulse responses in multiple degrees of freedom". In: *Tonmeister Tagung*. November. 2006

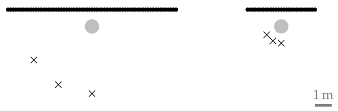


Figure 5.13: Setup for Experiment 1. The position of the synthesized focused source is indicated by the grey point. The position of the listener by black crosses and secondary sources by black dots. ☞

database.

As audio source signals, anechoic recordings of speech and of castanets were chosen.[34] The speech signal was an 8 s sequence of three different sentences uttered by a female speaker. The castanets recording was 7 s long. The levels of the stimuli were normalized to the same loudness by informal listening for all conditions.

Participants In order to generate a large amount of meaningful attributes, test subjects with experience in analytically listening to audio recordings were recruited. The experiment was conducted with 12 *Tonmeister* students – aged 21 to 33 years. The participants had between 5 years and 20 years of musical education, and all of them had experience with listening tests. They had normal hearing levels, and were financially compensated for their effort.

Procedure The participants received written instructions explaining their tasks in the two phases of the experiment.

The RGT procedure consisted of two parts, the *elicitation phase* and the *rating phase*. In the elicitation phase, groups of three conditions (*triads*) were presented to the test subject. The subjects were able to switch between them by pressing a corresponding button, and could listen to each stimulus as long as they wanted. For each triad, the subject had to decide which two of the three stimuli were more similar, and had to describe the characteristic which made them similar, and in which characteristic they were different from the third stimulus (which should be the opposite of the first property). If there were competing aspects, only the strongest one should be taken into account. One attribute pair per triad had to be specified, and two more could optionally be given if the test subject perceived several different properties. A screenshot of the used test GUI is shown in Geier et al.[35]

After a short training phase, every participant had to execute this procedure 12 times, using 12 different triads. 10 of the 12 triads resulted from a complete set of triads from the five conditions ref, $30^\circ_{4\,\mathrm{m}}$, $60^\circ_{4\,\mathrm{m}}$, $30^\circ_{10\,\mathrm{m}}$ and $60^\circ_{10\,\mathrm{m}}$. The two additional triads were (ref, $0^\circ_{4\,\mathrm{m}}$, $0^\circ_{10\,\mathrm{m}}$) and ($0^\circ_{4\,\mathrm{m}}$, $30^\circ_{4\,\mathrm{m}}$, $0^\circ_{10\,\mathrm{m}}$). These two have been chosen in order to consider the additional, very similar conditions together, to get attributes for the small differences between them. Complete triads for only five conditions have been chosen because of the time-consuming procedure – a complete set of triads for 7 conditions would have resulted in 35 triads.

The presented triads were the same for all participants, however, the order of the triads and the order of conditions within a triad was alternated over all participants based on a *Latin Square* design.

After the elicitation phase, the participants took a break. During this time, the test supervisor removed repetitions of attribute pairs for constructing the attribute list used in the second RGT test phase.

For this rating phase in each trial one previously elicited attribute pair was displayed on top of the screen. Below, the seven conditions

[34] Audio examples are available as supplementary material.

[35] Geier et al., op. cit.

could be played back and had to be rated on corresponding continuous sliders. Once a rating was collected for all conditions, the test subject was able to switch to the next screen, a procedure repeated until all elicited attribute pairs were used. Before the actual test, a training phase had to be completed for two rating screens.

In the second session, which was in the most cases done on another day, the elicitation and rating phase was repeated with the respective other source stimulus. Half of the subjects were presented with the speech sample in the first session and the castanets in the second session, and vice versa for the other half.

Results One of the main results of the experiment were the elicited attribute pairs. They reflect the range of perceptual similarities and differences among the conditions. Their number was different between subjects, ranging from 6 to 17 pairs for individual subjects. The most prominent choices were artifacts (e.g. *clean sound* vs. *chirpy, squeaky, unnatural sound*) and localization (*left* vs. *center*). For the latter, it has to be noted that the focused source was always positioned straight in front of the listener. Attributes describing artifacts were provided by 10 of the 12 subjects for castanets, and by 9 subjects for speech. Localization-related attributes were given by 7 subjects for castanets, and 5 subjects for speech. Other common attributes were related to coloration (*original* vs. *filtered*, *balanced* vs. *unbalanced frequency response*), distance (*far* vs. *close*) and reverberation (*dry* vs. *reverberant*). All elicited attributes were originally collected in German.

The ratings of the attributes can be used to identify the underlying dimensions which best describe the perception of focused sources. This was done using a principal component analysis (PCA) for individual subjects. For all subjects, two principal components could be identified as the main dimensions of the perceptual space. These dimensions can explain 90% of the variance for castanets and 97% for speech, respectively.

This also allows to determine the positions of the different conditions in the resulting perceptual space. Figure 5.14 shows the PCA results for one individual subject for the speech and castanets, respectively. The PCA results for another subject can be found in Geier et al.[36] The blue and red dots represent the different conditions in this two-dimensional perceptual space. The gray lines show the arrangement of elicited attribute pairs in this space. From Figure 5.14 it can be seen that for both castanets and speech the first principal component C_1 resp. S_1 can be interpreted as a mixture of the amount of artifacts and the distance, and the second principal component C_2 resp. S_2 as the localization of the source. Considering individual conditions, it can be observed that the 10 m loudspeaker array was rated to produce artifacts in the perception of the focused source, while the artifact-related ratings for the 4 m array are more or less the same as for the reference condition. For the longer array, the amount of artifacts depends on the listener position, with the high-

[36] Ibid.

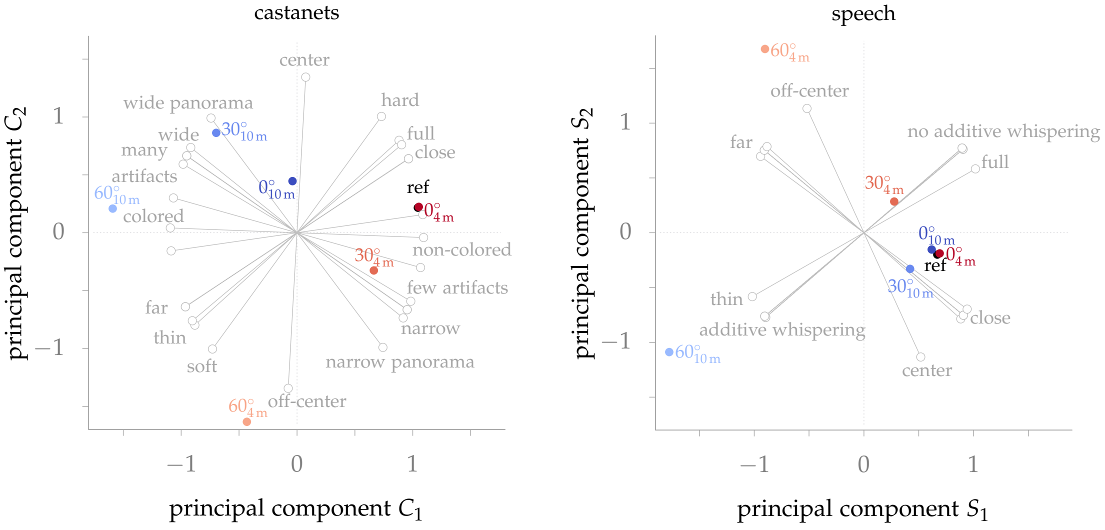


Figure 5.14: Principal component analysis for castanets (left) and speech (right) for one single subject. The blue, red and black points indicate the position of the conditions given in the two-dimensional space determined by the two given components for each stimulus type. The gray lines show the arrangement of the attribute pairs in these two dimensions. ☞

est rating of artifacts at the lateral position $60^\circ_{10\,\mathrm{m}}$. The perception of a wrong direction is most distinct for the lateral positions of the shorter array, with the condition $60^\circ_{4\,\mathrm{m}}$ as the most prominent case. Both lateral positions ($\phi = 60^\circ$) were perceived as more off-center than the other ones. Furthermore, it can be noted that the perceptual deviation from the reference condition occurs for more conditions for the castanets than for the speech stimuli.

Discussion The results show that the amount of perceived artifacts depends on the length of the loudspeaker array and the position of the listener, being worse for a larger loudspeaker array and a more lateral position of the listener. This is can be explained by the higher number of additional wave fronts for a larger loudspeaker array and a longer time between the first additional wave front the desired one for a lateral position. Figure 5.15 illustrates this effect. There the assumption is made that every single loudspeaker contribute an additional wave front with an amplitude that is only influenced by the distance of the loudspeaker to the listener. The direction of incidence of the single wave fronts is indicated by the direction the arrows point to. The starting point on the y-axis of an arrow indicates the position in time of the wave front, and the length and color of the arrow is proportional to its amplitude in dB. It is obvious that the larger the used loudspeaker array, the earlier the occurrence of additional wave fronts, and the higher their amplitude. This is due to the fact, that every single loudspeaker adds a wave front. For a given array, the number of wave fronts will be the same regardless of the lateral listener position, but the time of arrival of the first wave front will be earlier. This can be explained by the fact that the listener is positioned closer to one end of the loudspeaker array in this case. The loudspeakers at the ends of the array had to be driven as the first ones in order to create a focused source in the middle of the loudspeaker array, resulting in the significantly earlier incidence of the wave fronts from the loudspeakers close to the listener.

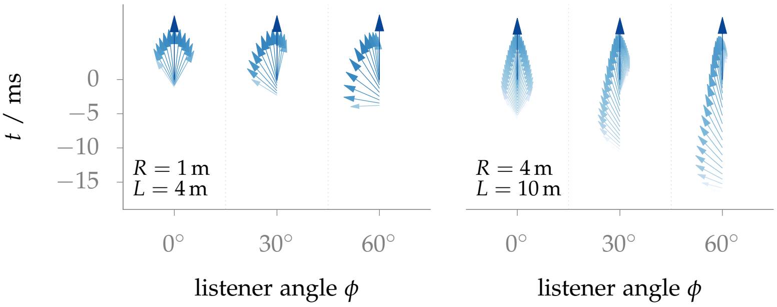


Figure 5.15: Direction, amplitude and time of appearance of wave fronts for the 4 m loudspeaker array (left) and the 10 m array (right). The results are shown for different angles ϕ at a radius of 1 m and 4 m, respectively. The arrows are pointing towards the direction from which the wave fronts arrive. The time of appearance is given by the starting point of the arrow. The length and color of the arrow is proportional to the amplitude of the wave front in dB. The dark blue arrows indicate the desired wave fronts. ☞

The results show a dependency of the perceived direction on the listener position and the array size. The condition $60^\circ_{4\,\mathrm{m}}$ was perceived as most from the left. The perceived direction can be explained by the additional wave fronts, too. The conditions with $\phi = 0^\circ$ were perceived from the same direction for both array lengths as the reference condition in front of the listener. For these conditions, the additional wave fronts have no effect on the perceived direction, because they arrive at the listener position symmetrically from all directions – compare Figure 5.15. For the lateral conditions, the first wave front will come mainly from the left side of the listener. Due to the precedence effect this can lead to localization of the sound to the direction of the (first) wave front. For the 10 m array, the perceived direction is different from that of the shorter array. Most of the subjects localized the sound in the same direction as the reference. However, a few subjects indicated that they had heard more than one sound source – one high-frequency chirping source from the left and a cleaner source in front of them. This can be explained with the echo threshold related with the precedence effect, which means that further wave fronts which follow the first one with a lag larger than the echo threshold are perceived as an echo.[37]

In order to verify this hypothesis, an experiment has been performed to examine the localization dominance for this kind of time-delayed wave front pattern.[38] Here, an approximated time of 8 ms between the first wave front and the desired one has been identified to be the threshold until which the perceived direction is dominated by the first wave front. This is in conformance with the results for the large array.

5.3.2 *Experiment 2: Influence of the Secondary Source Geometry on the Perception of Focused Sources in* WFS[39]

In a second experiment the main two attributes elicited in the first experiment, namely *artifacts* and *direction* were investigated in more

[37] E.g. Blauert, op. cit.

[38] H. Wierstorf and S. Spors. "Die Rolle des Präzedenzeffektes bei der Wahrnehmung von räumlichen Aliasingartefakten bei der Wellenfeldsynthese". In: *36th German Annual Conference on Acoustics*. 2010.

[39] Parts of this section are published in Wierstorf, Raake, and Spors, op. cit.

detail. The goal of this experiment is to highlight the connection between these two attributes and the geometry of the secondary source distribution.

As mentioned in Chapter 3, truncation of a sampled secondary source distribution leads to two opposite effects. On the one hand, a smaller distribution leads to fewer additional wave fronts and reduces the perception of artifacts as shown in the first experiment. On the other hand, a smaller distribution is linked to stronger diffraction of the sound field and therefore a smaller possible audience area as well as larger focal points – as discussed in Figure 3.6. In addition, the maxima and minima of the diffraction pattern could introduce wrong ILDs and the additional wave fronts could trigger a wrong direction due to the precedence effect.

To verify if there is an array length for which the artifacts are not audible, and the wrong binaural cues are negligible as well, a listening test was conducted that included three shorter array lengths together with the two array lengths used in the first experiment.

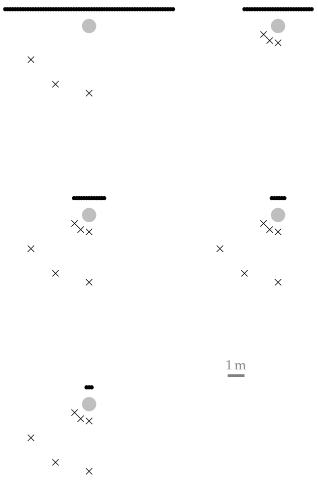


Figure 5.16: Setup for Experiment 2. The position of the synthesized focused source is indicated by the grey point. The position of the listener by black crosses and secondary sources by black dots. ☞

Stimuli The experiment was conducted with a similar geometry and the same source materials as described in Section 5.3.1. The same listener positions were used, including now the array sizes of $L = 10\,\mathrm{m}$, $4\,\mathrm{m}$, $1.8\,\mathrm{m}$, $0.75\,\mathrm{m}$ and $0.3\,\mathrm{m}$. For the three shortest arrays the listener were placed at all six positions. Figure 5.16 summarizes the experimental setup.[40]

Participants Six test subjects participated in the test. All of them were members of the Audio Group at TU Berlin and were normal hearing.

Procedure After an introduction and a short training phase with a violin piece as source material, one half of the participants started the first session presenting speech, the other half presenting castanets. In a second session, the speech and castanets source materials were switched between the groups. The subjects were presented with a screen containing nine sliders representing nine different conditions. At the top of the screen, one of the two attribute pairs *few artifacts* vs. *many artifacts* and *left* vs. *right* were presented. After a subject had rated all conditions, the next attribute pair was presented for the same conditions. Thereby the order of the conditions attached to the slider and the appearance of the attribute pairs was randomized. This procedure was repeated three times, once for all the array conditions assessed in case of each listening angle ϕ. For the listening angle of $0°$, the attribute pair *left* vs. *right* was omitted.

Results The left part of Figure 5.17 presents the mean ratings over all subjects, all listener positions and both source materials (speech and castanets) for the attribute pair *few artifacts* vs. *many artifacts*. Hence, the only independent variable is the length of the secondary source distribution plotted on the x-axis. The $0°$ position for the

[40] Audio examples are available as supplementary material.

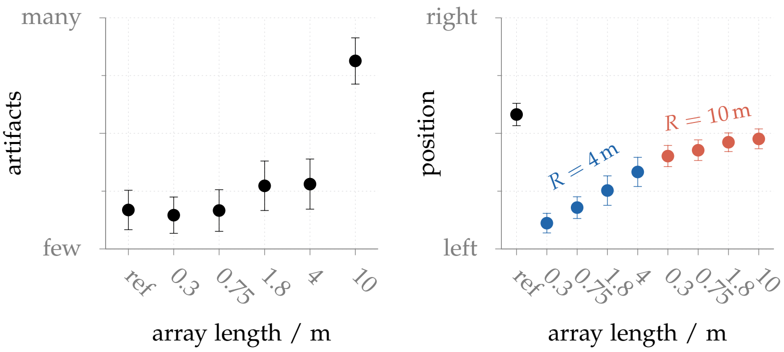


Figure 5.17: Mean and standard error for the rating of the attribute pair *few artifacts* vs. *many artifacts* plotted over the condition. The mean is calculated over all subjects, source materials and the different listener positions. ☞

speech material resulted as an outlier, and was not considered for the plot. At this position and with speech as source material, artifacts are only little audible. On the other hand, there is the coloration introduced by the spatial sampling, and independent of the fact that focused sources were realized. An interview with the subjects revealed, that four of them have rated this coloration rather than the targeted audible artifacts. It can be seen in the figure that the results for the different loudspeaker arrays build three different groups. The two shortest arrays resulted in as few artifacts as the reference condition. The 10 m array was found to lead to strong artifacts, as it was expected from the previous experiment. The amount of artifacts caused by the 1.8 m and the 4 m array are positioned between these two groups. A one-way ANOVA shows that the mentioned three groups are statistically different ($p < 0.05$) from each other, and not different within each group.

In the right part of Figure 5.17, the results for the attribute pair *left* vs. *right* are presented. The means for the arrays were calculated over the 30° and 60° conditions, but once for each radius indicated by the two different shades of gray. For a listener angle of 0° the rating of this attribute pair was omitted. It can be seen that the reference condition (arriving from straight ahead of the listener) was rated to come slightly from the right side. All other conditions came from the left side, where shorter arrays and smaller radii lead to a rating further to the left.

The two different source materials speech and castanets showed significant differences only for the 10 m array and the 30° and 60° positions, with more artifacts perceivable for the castanets stimuli.

Discussion As shown already in the first experiment, the appearance of additional wave fronts due to spatial aliasing leads to strong artifacts for focused sources. The arrival time of the first wave front at the listener position can be reduced by using a shorter loudspeaker array. This leads to a reduction of audible artifacts, as shown by the results for the attribute pair *few artifacts* vs. *many artifacts*. The two smallest arrays with a length of 0.3 m and 0.75 m are rated to have the same amount of artifacts as the single loudspeaker reference.

All three loudspeaker arrays with a length of $L < 2$ m have arrival times of the first wave front of below 5 ms. This means that they fall in a time window in which the precedence effect should work, and no echo should be audible. The artifacts audible for the array with $L = 1.8\,\text{m}$ are therefore due to a comb-filter shaped ripple in the frequency spectrum of the signal, as a result of the temporal delay and superposition procedure of the loudspeakers, see (2.76).

However, there are other problems related with a shorter array. The main problem is the localization of the focused source. Figure 5.17 shows a relation between array length and localization: the shorter the array, the further left the focused source is perceived. This result implies that the precedence effect cannot be the only reason for the wrong perception of the location. For a shorter array, too, the first wave front arrives from the loudspeaker at the edge of the array. This loudspeaker will be positioned less far to the left for a shorter array than for a longer array. Therefore, it is likely that the diffraction due to the short array length introduces wrong binaural cues, namely a wrong ILD.

5.3.3 *Conclusion*

Sound field synthesis allows for the synthesis of focused sources placed directly in the audience area, a feature that makes it distinct from all stereophonic presentation techniques. The problematic aspect of focused sources is that the additional wave fronts due to spatial sampling appear not after but before the desired wave front. This is inherent for focused sources due to the time reversal technique employed to create them.

Experiments were carried out that investigated the influence of these special situation in the perception of focused sources. For loudspeaker arrays larger than 2 m spectro-temporal artifacts were perceivable in addition to coloration of the synthesized source. By applying smaller arrays or smaller parts of a large array these artifacts can be eliminated. On the other hand by using smaller arrays the localization of the focused source is impaired.

6
Prediction of Spatial Fidelity

It is a challenging task to investigate the localization accuracy in an audience area with a listening test, even with binaural simulations of the desired setup. To already indicate in the planing phase of a loudspeaker array the achievable localization accuracy of such a setup would be helpful. Using binaural simulations ear signals at every point of the audience area of the desired loudspeaker setup are available, at least if an anechoic chamber is assumed as a first approximation. What is required is then a model that is able to predict the localization of the two ear signals by a human listener.

The estimation of a sound source direction for two given ear signals is well known in different applications. For example it is applied in binaural hearing aid algorithms for self-steering beamformers,[1] for human speaker separation and recognition,[2] or generally in the context of computational auditory scene analysis.[3] These more technical motivated approaches apply in most cases a cross-correlation in the frequency domain.[4] Normally they have access to all features of the signal which is not the case for the human auditory system that has a restricted timing resolution. In order to predict the direction a human listener will hear a synthesized source from it seems to be more appropriate to only use such cues for direction estimation that are available to the human brain as well. Models that mimic the auditory system in estimating the direction of a sound are called binaural models and will be introduced in the next section. Afterwards, an existing binaural model is slightly modified and applied to the binaural signals of the localization experiments presented in Section 5.1. It is verified that the model is able to predict the perceived direction correctly in most of the cases. Eventually the model is used to analyze the localization accuracy in a wide range of different loudspeaker setups and SFS techniques for getting a systematic overview of the dependency of the accuracy on the applied technique and spacing between the loudspeakers.

6.1 Binaural Models

Section 5.1 discussed the binaural cues a human listener utilizes in order to localize a sound. For a broad-band sound the perceived direction is dominated by the ITD between the left and right ear sig-

[1] T. Rohdenburg et al. "Objective perceptual quality assessment for self-steering binaural hearing aid microphone arrays". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. 2008, pp. 2449–52.

[2] T. May, S. van de Par, and A. Kohlrausch. "A Binaural Scene Analyzer for Joint Localization and Recognition of Speakers in the Presence of Interfering Noise Sources and Reverberation". *IEEE Transactions on Audio, Speech, and Language Processing* 20.7 (2012), pp. 2016–30.

[3] D. Wang and G. J. Brown, eds. *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Hoboken: Wiley, 2006.

[4] E.g. C. K. Knapp and G. C. Carter. "The Generalized Correlation Method for Estimation of Time Delay". *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24.4 (1976), pp. 320–27.

nals. For frequencies above 1.4 kHz the human auditory system is no longer able to decode the time differences since the signals are changing too fast. Now, ILD cues together with the ITD of the signal's envelope are cues for the perceived direction.

Most of the existing binaural models focus on the extraction of the ITD as a cue for the direction in the horizontal plane of the sound. Jeffress was one of the first to propose a mechanism of the hearing system that is able to extract the ITD.[5] His model applies two delay lines connected together with coincidence detectors. The delay lines compensate for the external delay of the two ear signals. Thus the ITD is indicated by the position of the most active coincidence detector. Such a process describes a neuronal place mapping of the ITD. It can be mathematically described by a cross-correlation between the left and right ear signal. Models relying on similar principles as the delay line model by Jeffress are referred to as cross-correlation models. Later versions include contralateral inhibition[6] and are able to predict not only localization phenomena but even binaural masking.[7]

Although the cross-correlation models are wide spread and applied in technical direction estimation algorithms, there is ongoing discussion about the physiology basis of the human localization. Delay lines similar to the ones proposed by Jeffress were found in the barn owl,[8] but were not unambiguously found in mammals as the result for guinea pigs highlights.[9]

An alternative approach is that the ITD is coded by the firing rate of a given neuron population. In gerbils neurons were detected that directly encoded ITDs with their firing rate.[10] The peak of the firing rate function for the ITD depended on the best frequency of the neuron. However, the interaural phase difference (IPD) where the peak occurred was constant for all neurons and restricted to half a cycle which is referred to as the π-limit. Grothe et al.[11] provide a compelling summary of the evolution, physiology, and functionality of sound localization in different species, discussing the delay line versus rate coding approach.

These discoveries inspired new models that try to implement the physiology of the mammalian ear in more detail and base their estimation of the perceived direction on the IPD and rate coding.[12] This thesis modifies Dietz' et al.[13] model in order to predict the perceived direction of synthesized source in SFS. In the next section the modified model will be introduced.

6.2 *Applied Binaural Model*

The binaural model applied in this thesis piggybacks on the model presented by Dietz et al.[14] that bases its procedure on the rate coding of the IPD. First, the detailed processing steps applied in this thesis will be described. Afterwards differences to the original model are discussed. The general structure of the model consists of the following steps and is depicted in Figure 6.1:

[5] L. A. Jeffress. "A place theory of sound localization." *Journal of Comparative and Physiological Psychology* 41.1 (1948), pp. 35–9.

[6] E.g. W. Lindemann. "Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals." *The Journal of the Acoustical Society of America* 80.6 (1986), pp. 1608–22

[7] J. Breebaart, S. van de Par, and A. Kohlrausch. "Binaural processing model based on contralateral inhibition. I. Model structure". *The Journal of the Acoustical Society of America* 110.2 (2001), pp. 1074–88.

[8] C. E. Carr and M. Konishi. "A circuit for detection of interaural time differences in the brain stem of the barn owl." *The Journal of Neuroscience* 10.10 (1990), pp. 3227–46

[9] D. McAlpine, D. Jiang, and a. R. Palmer. "A neural code for low-frequency sound localization in mammals." *Nature Neuroscience* 4.4 (2001), pp. 396–401.

[10] A. Brand et al. "Precise inhibition is essential for microsecond interaural time difference coding". *Nature* 417.6888 (2002), pp. 543–47.

[11] B. Grothe, M. Pecka, and D. McAlpine. "Mechanisms of Sound Localization in Mammals". *Physiological Reviews* 90 (2010), pp. 983–1012.

[12] M. Dietz et al. "Coding of temporally fluctuating interaural timing disparities in a binaural processing model based on phase differences." *Brain Research* 1220 (2008), pp. 234–45; M. Takanen, O. Santala, and V. Pulkki. "Visualization of functional count-comparison-based binaural auditory model output". *Hearing Research* 309 (2014), pp. 147–163

[13] M. Dietz, S. D. Ewert, and V. Hohmann. "Auditory model based direction estimation of concurrent speakers from binaural signals". *Speech Communication* 53.5 (2011), pp. 592–605.

[14] Ibid.

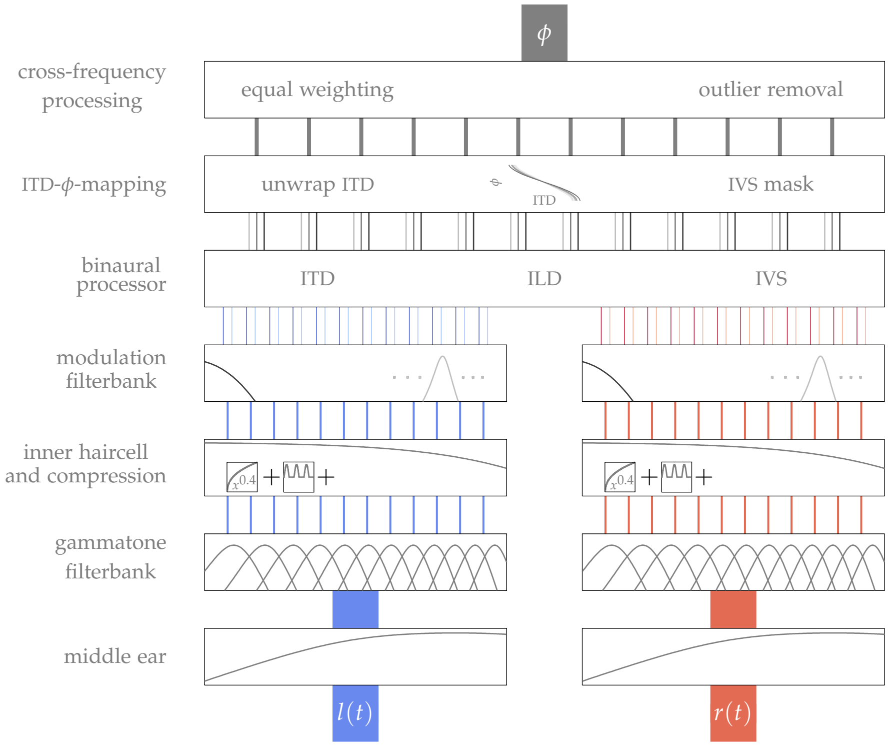


Figure 6.1: Sketch of the applied binaural model. At the bottom the two time signals $l(t)$ and $r(t)$ are the input to the model. After that identical monaural preprocessing is applied. The binaural processor calculates the ITD, ILD, and interaural vector strength (IVS) based on the arriving input signals for every frequency channel. At the end the binaural parameter is mapped to a single direction estimation averaged over the whole time of the input signals. ☞

dietz2011.m

wierstorf2013estimateazimuth.m

- The middle ear transfer characteristic is approximated by a first-order band-pass filter between 500 Hz and 2 kHz.[15]
- The auditory filters present at the basilar membrane are modeled with a fourth-order gammatone-filterbank.[16] Twelve filter bands were applied in the range of 200 Hz to 1400 Hz with a spacing of 1 ERB.
- Compression of the cochlea was modelled as to the power of 0.4.[17] The transduction process of the inner hair-cells was implemented with a half-wave rectification followed by a successive 770 Hz fifth-order low-pass filter.[18]
- The signals leaving the haircell stage have a DC component and are broaden in frequency. To derive a meaningful phase difference a second stage of bandpass filtering has to be applied. This is done by a gammatone filterbank with second-order filters with an attenuation of 10 dB. Every filter is centered at the corresponding center frequency of the twelve frequency channels. In addition a second-order low-pass filter with a cutoff frequency of 30 Hz is applied to every frequency channel in order to calculate the ILD.
- At the output of the band-pass filter the interaural transfer function is calculated in order to derive the IPD from it. The IPD is then divided by the instantaneous frequency for getting the ITD. To be able to estimate the reliability of the binaural parameters at certain time-frequency steps the IVS is derived as an approximation of the interaural coherence.[19]

[15] S. Puria. "Measurements of human middle ear forward and reverse acoustics: Implications for otoacoustic emissions". *The Journal of the Acoustical Society of America* 113.5 (2003), pp. 2773–89.

[16] V. Hohmann. "Frequency analysis and synthesis using a Gammatone filterbank". *Acta Acustica united with Acustica* 88.3 (2002), pp. 433–42.

[17] S. D. Ewert. "Characterizing frequency selectivity for envelope fluctuations". *The Journal of the Acoustical Society of America* 108.3 (2000), pp. 1181–96; M. A. Ruggero et al. "Basilar-membrane responses to tones at the base of the chinchilla cochlea". *The Journal of the Acoustical Society of America* 101.4 (1997), pp. 2151–63.

[18] Breebaart, Par, and Kohlrausch, op. cit.

[19] Compare C. Faller and J. Merimaa. "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence". *The Journal of the Acoustical Society of America* 116.5 (2004), pp. 3075–89; M. J. Goupell and W. M. Hartmann. "Interaural fluctuations and the detection of interaural incoherence: Bandwidth effects". *The Journal of the Acoustical Society of America* 119.6 (2006), pp. 3971–86

- The sign of the ILD in each frequency channel is a way to extend the π-limit of the IPD to 2π which gives the opportunity to get the natural range of ITDs between -700 ms and $+700$ ms. This way center frequencies up to 1.4 kHz can be handled. This process is called *unwrapping* the ITD. For details see Figure 2 in Dietz et al.[20] and the corresponding discussion.
- A lookup table with stored ITD values and the corresponding azimuth angles converts the ITD in the perceived direction in each frequency channel and for every time sample. The lookup table was generated with the same HRTF set the binaural synthesis utilizes and has a resolution of 1°.
- An IVS binary mask is created by setting the threshold of the IVS value to 0.98 and demanding a rising slope condition with $\frac{\mathrm{dIVS}(t)}{\mathrm{d}t} \geq 0$. The median over time of the direction is then calculated in every frequency channel by using only those time steps that are fullfilling the conditions of the binary IVS mask.
- Eventually the median over the direction in the single frequency channels is calculated. If the azimuth of single frequency bands deviate more than 30° from the median these frequency bands are removed and the median is recalculated. The median azimuth is then the provided estimation for the direction of the auditory event corresponding to the two ear signals from the beginning.

For the last step of the model a weighting of the directions in the different frequency bands is possible. For example, Raatgever[21] has found large differences in the degree single frequency bands dominate the perceived lateralization in the presence of conflicting cues in different frequency bands. Dietz et al.[22] applied a weighting accordingly to the magnitude of the signal in the different frequency bands. In this thesis both weighting methods were tested, but both impaired the results compared to an equal weighting scheme. Especially the strong weighting after the data by Raatgever leads to large deviations of the model predictions compared to the localization results from the listening experiments. Therefore, an equal weighting of the different frequency bands was adopted for this thesis. This is also in accordance with the results for modelling listening data from stereophonic experiments presented in Park.[23]

For predicting the directions of up to five different speakers the study by Dietz et al.[24] also utilizes the IPDs of the envelope signal for frequency channels above 1.4 kHz. They found that these IPD cues were not as salient as the ones derived from the fine structure at lower frequency channels and were not able to improve the results by including them. In a similar way the performance of the binaural model could not be enhanced by including these envelope IPDs for the prediction of the localization of synthesized sources in this thesis. Therefore, they will be considered neither in the following nor included in the description of the applied binaural model.

[20] Dietz, Ewert, and Hohmann, op. cit.

[21] J. Raatgever. "On the binaural processing of stimuli with different interaural phase relations". PhD thesis. Technische Universiteit Delft, 1980, Sec. 3.4.

[22] Dietz, Ewert, and Hohmann, op. cit.

[23] M. Park. "Models of binaural hearing for sound lateralisation and localisation". PhD thesis. University of Southampton, 2007, Fig. 5.10.

[24] Dietz, Ewert, and Hohmann, op. cit.

6.3 *Validation of the Binaural Model*

The binaural model described in the last section was validated by all of the data from the listening experiments presented in Section 5.1. In addition to the perceived direction the localization blur was estimated by the standard deviation of the azimuth value of the model over time. The five repetitions of every condition for each listener were simulated by applying five different noise bursts to the model. Afterwards the mean about these five stimuli were calculated for the predicted direction and its corresponding localization blur. In addition, eleven listeners were simulated by applying different head orientations of the binaural model. For the first listener the model was facing the secondary sources forward at 0°. The second listener was facing towards 1° and so forth. The offset in the perceived angle due to the head orientation was compensated for in the estimation of the direction of the synthesized source. Afterwards the mean and confidence values were calculated above the eleven different head orientations for both the direction and localization blur, given by the standard deviation for single head orientation.

In exactly the same way as for the data from the listening experiment, the prediction data of the binaural model was inspected for deviations from a single Gaussian distribution. For cases were the data could not be explained by a single distribution a Gaussian mixture model was applied to separate the data into two distributions – compare Figure 6.2 and 5.5.

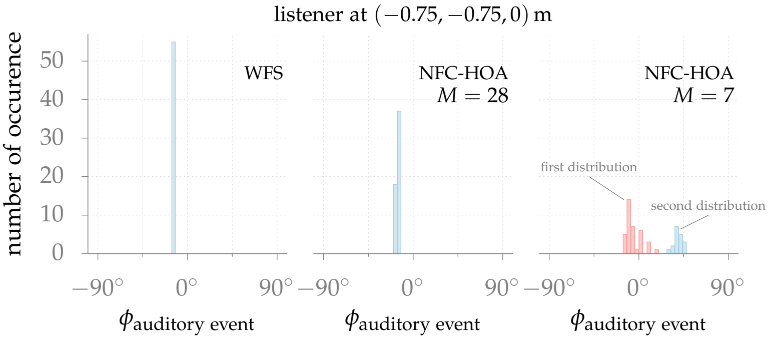


Figure 6.2: Distributions of the auditory event's directions as predicted by the binaural model for the listening position $(-0.75, -0.75, 0)$ m and the circular secondary source distribution with 14 sources. The results for a synthesized point source for WFS and NFC-HOA up to different orders M are shown. ☞

All model predictions are compared with all localization results from the experiments in Figure 6.3b. The figure visualizes that the model predictions are very accurate, especially for the case of WFS and a synthesized point source or plane wave. By averaging the results for all these conditions the deviation of the model results from the experimental data is 1.8° with the largest deviation being 14° for 14 secondary sources and a synthesized plane wave at the listener position $(0.75, -0.5, 0)$ m.

The problem of more than one perceived source was ignored for testing the model performance by assuming a single perceived source at all positions. Therefore, the mean value of both directions for these cases was used. The accuracy of the model predictions is then

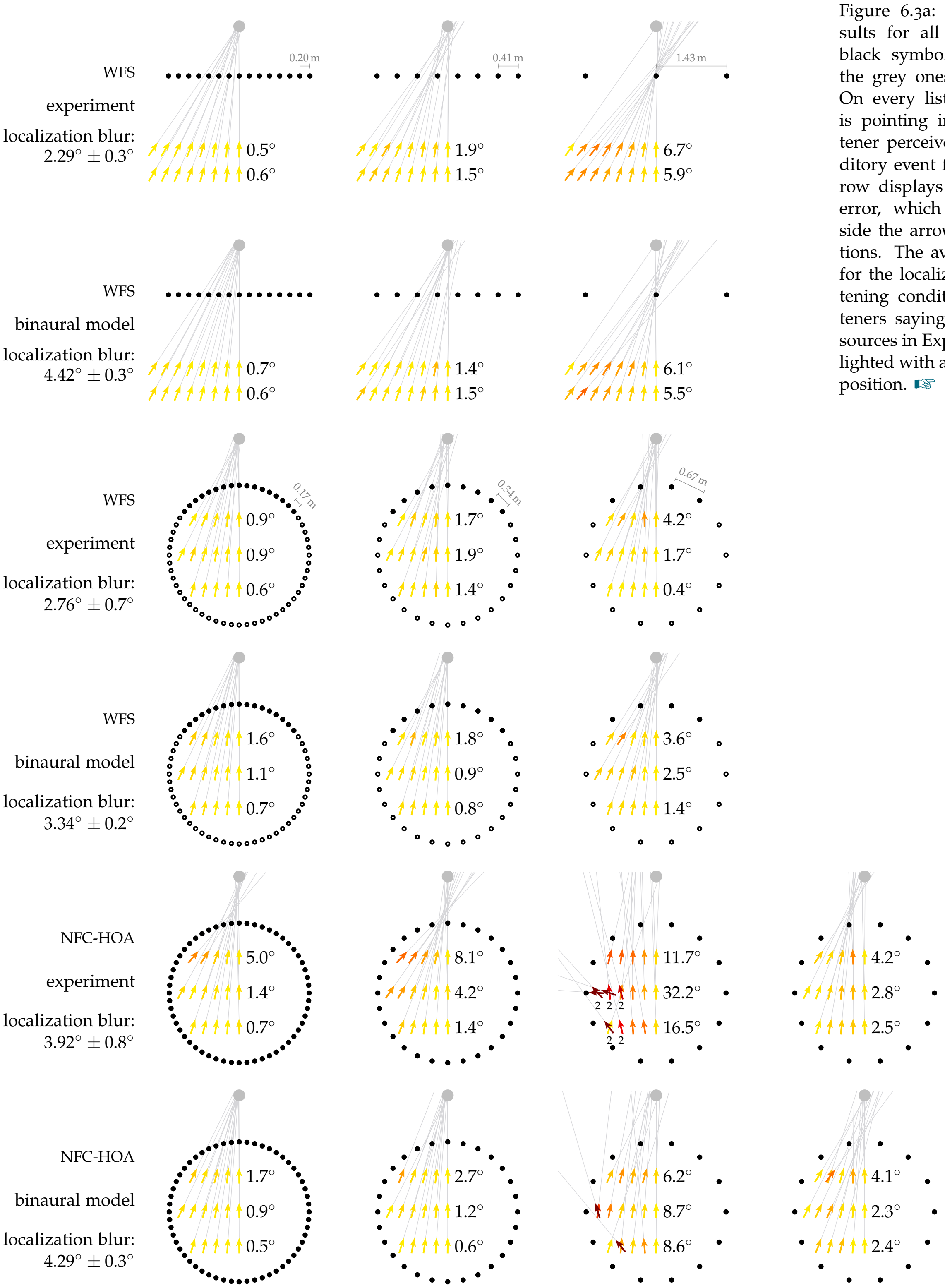


Figure 6.3a: Average localization results for all four experiments. The black symbols indicate loudspeakers, the grey ones the synthesized source. On every listening position an arrow is pointing into the direction the listener perceived the corresponding auditory event from. The color of the arrow displays the absolute localization error, which is also summarized beside the arrows for every row of positions. The average confidence interval for the localization results is 2.3°. Listening conditions that resulted in listeners saying that they perceived two sources in Exp. 4 of Section 5.1 are highlighted with a small 2 written below the position. ☞

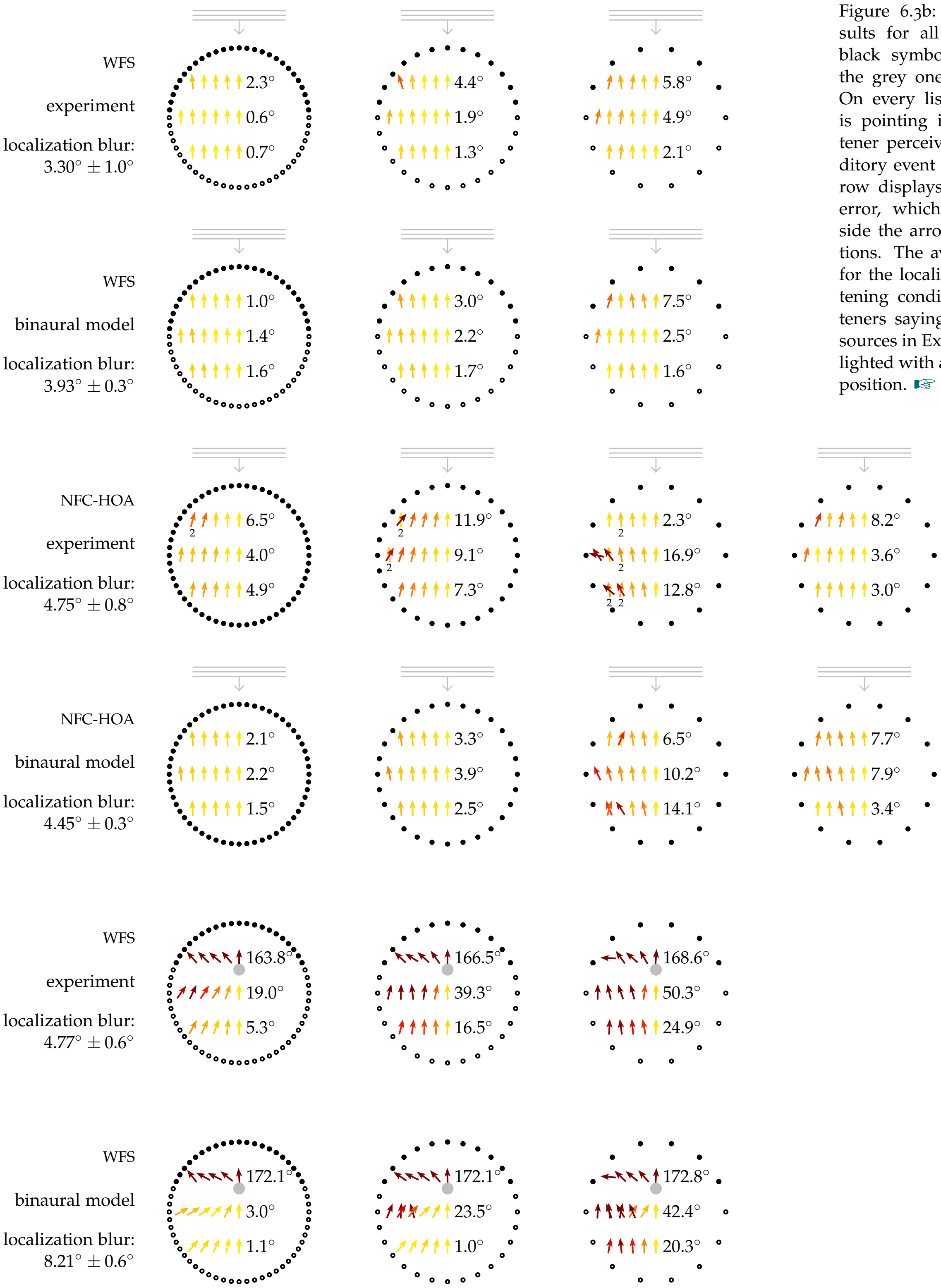


Figure 6.3b: Average localization results for all four experiments. The black symbols indicate loudspeakers, the grey ones the synthesized source. On every listening position an arrow is pointing into the direction the listener perceived the corresponding auditory event from. The color of the arrow displays the absolute localization error, which is also summarized beside the arrows for every row of positions. The average confidence interval for the localization results is 2.3°. Listening conditions that resulted in listeners saying that they perceived two sources in Exp. 4 of Section 5.1 are highlighted with a small 2 written below the position. ☞

in average 8° for all NFC-HOA conditions. Whereby the model performance is slightly better for a synthesized point source and higher orders. The maximum deviation is 40° for a secondary source distribution with 14 sources and an order of 7 for the spherical harmonics at the listener position of $(-1.25, 0, 0)$ m. For the focused source condition and WFS the model predicted a correct localization of the synthesized source even for several positions were the listener had large deviations. That leads to the largest deviations of the model prediction from the experimental data for focused sources of 11° in average. The maximum deviation was 40° for a secondary source distribution of 28 sources and the listener position $(-0.5, 0, 0)$ m.

Beside the perceived direction the localization blur was modeled by calculating the standard deviation of the predicted direction over time. This method could predict the localization blur adequately when averaging over the different loudspeaker arrays. Looking at the single loudspeaker array the localization blur of the model depended on the number of used loudspeaker, having a larger localization blur for the arrays with fewer loudspeakers. In the listening experiment no such dependency could be observed.

THE MOST INTERESTING conditions are those where the model is not as accurate as for the other ones. Especially the case for focused sources is challenging. For a closer inspection of the model behavior for these cases the ITD values for different frequencies channels over time will be analyzed. Figure 6.4 visualizes them for a listener position of $(-0.75, 0, 0)$ m. The top row presents the situation of WFS and a focused source. The ITDs of three adjacent frequency channels are plotted in the same color. ITDs for frequencies around 200 Hz are marked as blue points, ITDs for frequencies around 500 Hz as green, 800 Hz as yellow, and 1200 Hz as red. For a secondary source distribution with 56 sources the ITDs are clustered around −0.6 ms, whereby the lower frequencies are spread in a larger region ranging from −0.2 ms to 1 ms. This is not astonishing, because the minimal size of the focal point is determined by its wave length. For example, a frequency of 240 Hz corresponds to a size of 1.4 m of the focal point. For the two secondary source distributions with less sources the ITDs for lower frequencies are still spread out in the same area, but those for higher frequencies are vanishing or have negative values. Comparing this with the listener results it is of interest that the listeners perceived the focused source to come from the left for the secondary source distributions with 14 and 28 sources, whereby the model predicted them to come from the right and the left. The latter was a result of including different head orientations. Figure 6.4 shows only the head orientation of 0° for which the perceived direction would be predicted to come from the right for 14 and 28 sources.

Beside ITDs for focused sources, Figure 6.4 furthermore compares the ITDs for a point source synthesized with WFS and NFC-HOA. The localization accuracy of the listener was around 1° for WFS in the listening experiment at the corresponding listening position of

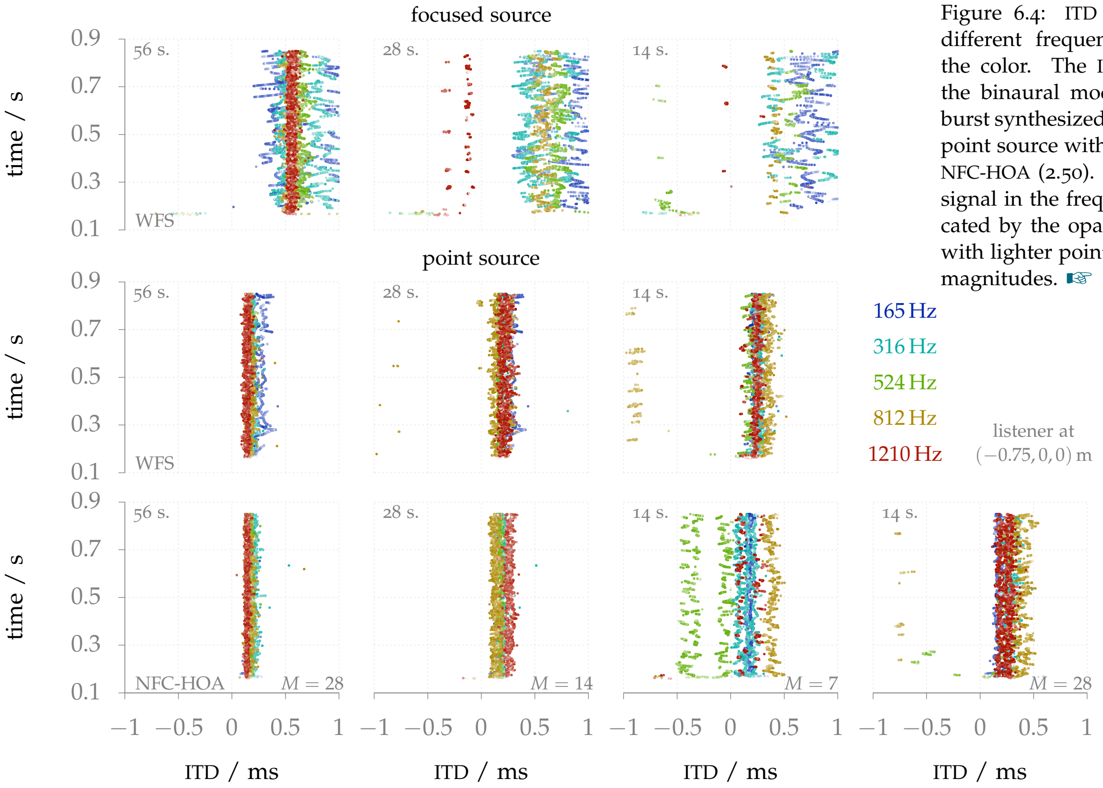


Figure 6.4: ITD values over time for different frequencies as indicated by the color. The ITD was calculated by the binaural model for a single noise burst synthesized as a focused source or point source with WFS (2.76), (2.64) and NFC-HOA (2.50). The magnitude of the signal in the frequency channel is indicated by the opacity of the ITD points, with lighter points correspond to lower magnitudes. ☞

$(-0.75, 0, 0)$ m. Whereas, for NFC-HOA the accuracy was around 3° for all conditions with $M > 7$. For an order of $M = 7$ and 14 secondary sources the listener reported to perceive two sources. The binaural model was not able to predict the two sources for that condition. However, the corresponding ITD values indicate at least that the localization should be significantly different from the corresponding WFS condition as they are spread for some frequencies in the negative and positive region. These differences point out that the performance of the binaural model probably could be improved for the challenging conditions.

6.4 *Example Applications of the Binaural Model in Sound Field Synthesis*

The last section validated the predictions of the binaural model and highlighted its limits. This section will present some example applications of the binaural model in the context of sound field synthesis. It will be predicted how the localization is in the audience area. Another question is for which setups the listener will localize the synthesized sources towards the direction of the nearest loudspeaker. In addition, the model can be employed as a planing tool by specifying the size of the sweet-spot that a setup should achieve.

6.4.1 *Prediction of Localization for Sound Field Synthesis and Arbitrary Secondary Source Distributions*

The validation of the binaural model has shown that the model is especially able to predict the localization for WFS with high accuracy. Hence, the model can be used to predict the localization accuracy for WFS in the whole audience area. Figure 6.5 shows the estimations of the perceived direction for the same setup that was applied in the experiment in Section 5.1. However, this time the audience area was sampled with a higher resolution of around 15 cm. The model predictions support the results of the listening test. The localization is not impaired by WFS and a distance between the secondary sources of 20 cm. In this case the absolute localization error averaged over all listening positions is 1.6°. By doubling the distance the localization accuracy slightly becomes degraded, but still is relatively equal in the whole audience area. The average absolute localization error is now 3.4°. Whereas the setup with 3 secondary sources leads to an average absolute localization error of 10.3° and the localization of the nearest loudspeaker at several listener positions. The ratio between localising single loudspeaker or the synthesized source is further analyzed in the next section.

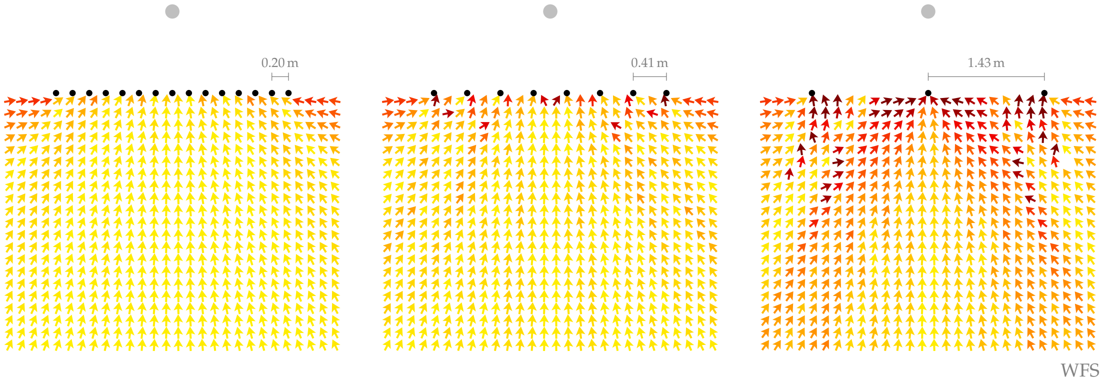


Figure 6.5: Model predictions of the perceived directions for a synthesized point source in the audience area. The three different linear secondary source distributions were all driven by WFS (2.64). ☞

The model furthermore gives the opportunity to test the influence of different secondary source geometries on the localization results. Figure 6.6 summarizes the result for two different geometries. In the left of the figure a box shaped secondary source distribution was applied. This geometry is of special interest, because it can easily be installed in rooms and was among the first ever build WFS setups.[25] If a source is coming from a direction orthogonal to one of the four linear parts of the distribution the perception will be similar to a linear source distribution. Of high interest are those synthesized sources coming from the direction of the array's edges. In Figure 6.6 a plane wave with an incidence direction of $(-1, -1, 0)$ is synthesized. The model predictions indicate that the localization accuracy in this case is only impaired directly in the edges of the array and comparable

[25] For an overview of existing WFS systems see D. de Vries. *Wave Field Synthesis*. New York: Audio Engineering Society, 2009

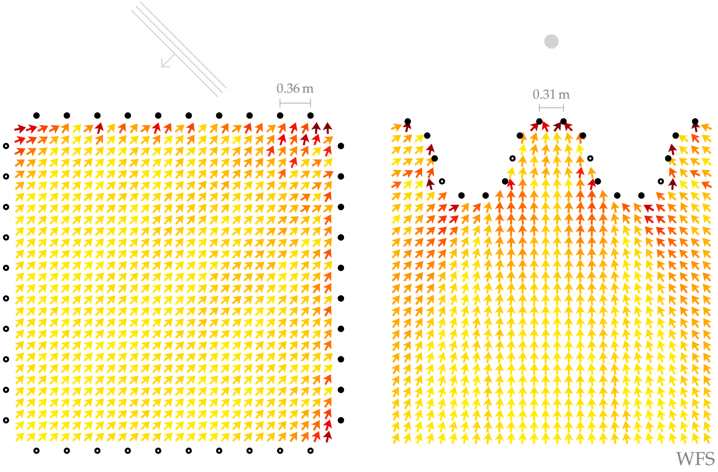


Figure 6.6: Model predictions of the perceived direction for a synthesized sources in the audience area. Both secondary source distributions were driven by WFS with (2.57) for the plane wave and (2.64) for the point source. ☞

to the one of a linear array at all other positions within the audience area.

The second example employs a smaller version of the concave secondary source distribution from Figure 2.2. The prediction results of the binaural model show that the concave parts introduce an impairment of the localization accuracy especially in the proximity of the secondary sources.

Comparing the model predictions the accuracy for NFC-HOA was not as convincing as for WFS. Nonetheless, as long as the order of the NFC-HOA systems was 14 or above the results were in fair agreement with the listening tests. Therefore, the binaural model will only be applied for NFC-HOA with orders of 14 in the following. An interesting question in the context of NFC-HOA is the vulnerability of the system to variations of the secondary source distribution from a perfect circle as the underlying theory assumes. Figure 6.7 shows results for a perfect circular geometry, an impaired one that has different angles between the single sources, and an impaired one that has a random jitter of up to 7 cm on its source positions. For the different angles a circular secondary source distribution with 56 sources was created and randomly 28 of those sources were chosen. The results for the regular distribution show very high localization accuracy. In fact, compared to the results from the listening test in Figure 6.3 the model probably predict a too high localization accuracy. If the angles between the single secondary sources are varied the overall localization accuracy drops enormously. Further investigation reveals that the position of the synthesized source is shifted to the left. Considering this shift the localization accuracy is only slightly impaired. Such a shift can occur if the secondary sources are not equally distributed on the circle as the underlying driving functions assume. In the given example they are more dense at the right bottom and left top of the distribution. The results for the randomly jittered secondary sources are presented in the right graph of Figure 6.7. They show that slight variations of the secondary source positions can influence

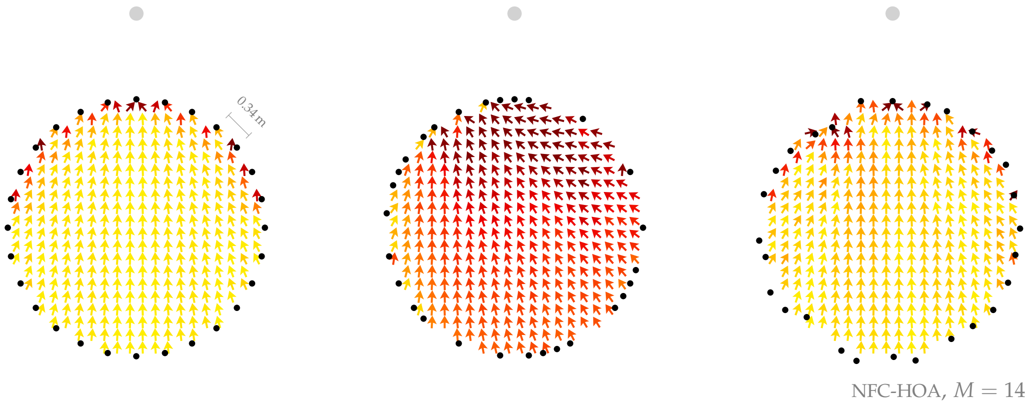


Figure 6.7: Model predictions of the perceived direction for a synthesized point source in the audience area. All three secondary source distributions were driven by NFC-HOA (2.50) with an order of 14. For both distributions to the right the positions of the secondary sources were changed. ☞

the localization of the synthesized source. At listening positions to the side the localization is now more towards the frontal sources of the distribution than towards the synthesized source.

6.4.2 *Determining the Border of Single Loudspeaker Localization*

The sweet-spot phenomenon for stereophony implicates that outside of the sweet-spot listener get the impression that the auditory events come from the nearest loudspeaker. An unsettled question is how many loudspeakers are needed for a linear secondary source distribution in WFS in order to avoid the localization of single loudspeakers as it is for example the case for a loudspeaker array with 3 sources as presented in Figure 6.5. One way of measuring is to predict the perceived direction with the binaural model at every listener position similar to Figure 6.5 and afterwards calculate if the perceived direction is closer to the actual position of the synthesized source or the nearest loudspeaker. After performing the prediction for each position the ratio between loudspeaker and synthesized source localization could be calculated over all listener positions.

The calculation conducted for the same WFS configuration as shown in Figure 6.5 for 2, 3, 4, up to 16 secondary sources. For only 2 secondary sources the listener localizes towards the nearest secondary source at 90% of all positions in the audience area. Using 3 secondary sources this is only the case for 40% of all positions, using 4 secondary sources for 10% of all positions and for 8 secondary sources it vanishes completely.

6.4.3 *Estimating the Size of the Sweet-Spot*

In the introduction of this thesis the phenomenon of the sweet-spot in stereophony was discussed and sketched in Figure 1.4. Normally in stereophony this term combines two facts: the existence of a small area in which the localization behaves as wanted and localization of the nearest loudspeaker outside of this area. As discussed in the last section the localization of single loudspeaker in WFS only happens for secondary source distributions consisting of less than 4 loudspeakers. Extending the idea of the sweet-spot to SFS it is defined as a first approximation to be that part of the audience area where the localization accuracy is equal or better than a given value.

The localization accuracy was calculated by the binaural model for different SFS setups and stereophony. Figure 6.8 summarizes the results by coloring all parts of the audience area that give a localization accuracy of 5° or better in blue. The result replicates the small sweet-spot of stereophony. In addition, the sweet-spot extends towards the whole audience area for WFS, achieving the desired effect shown in the introduction in Figure 1.4. It further demonstrates that a WFS system and a NFC-HOA system without band-limitation yield an identical spatial impression.

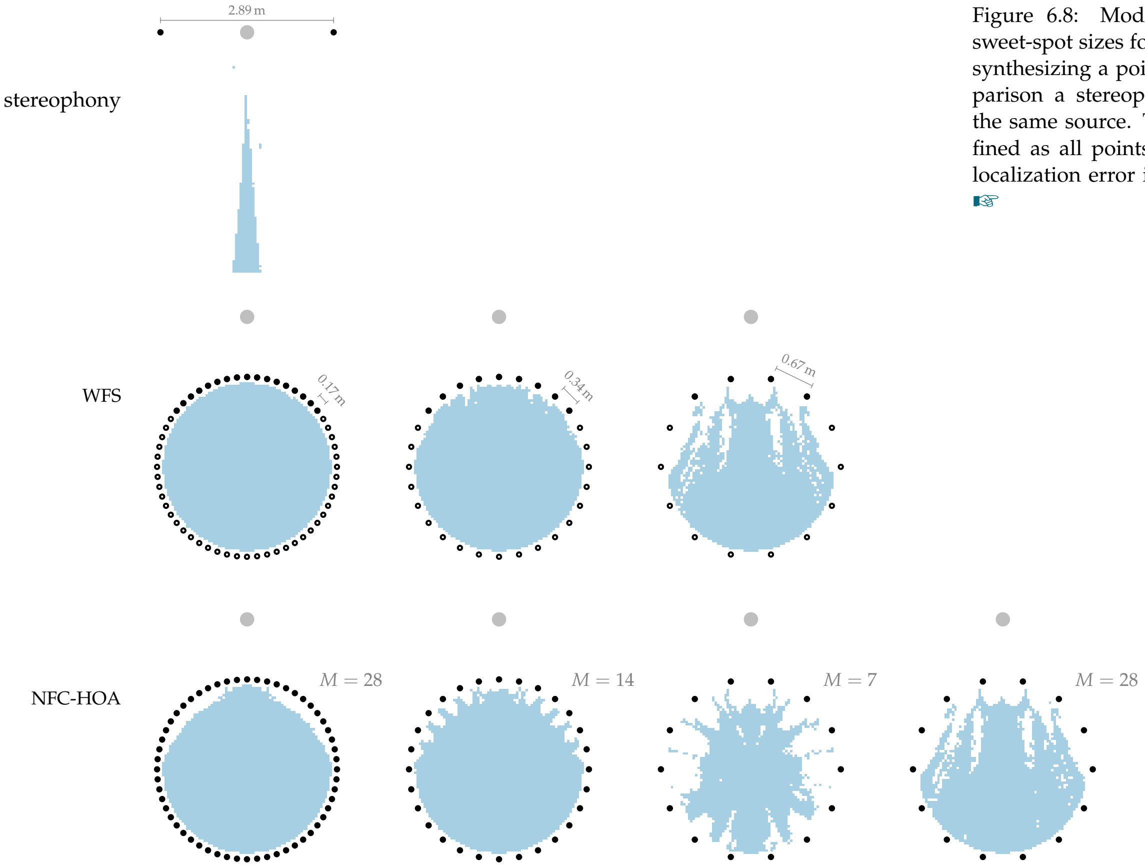


Figure 6.8: Model prediction of the sweet-spot sizes for different SFS setups synthesizing a point source. As a comparison a stereophony setup presents the same source. The sweet-spot is defined as all points where the absolute localization error is less or equal to 5°. ☞

For band-limited NFC-HOA the spread of the sweet-spot for the two distributions with less secondary sources shows that the localization is getting wrong in the proximity of the single loudspeakers. What cannot be directly concluded from the sweet-spot sketch is to which extent the localization will be impaired outside of the sweet-spot. The results of the listening experiment have shown that severe problems can occur for band-limited NFC-HOA. In that case even more than one auditory event could be perceived outside of the sweet-spot.

6.5 *Summary*

A binaural model after Dietz et al.[26] was modified and validated against the localization results from Section 5.1. The model showed very good agreement with the test results, especially in the case of WFS and NFC-HOA with high orders. For band-limited NFC-HOA the model had to be extended, because listeners reported perceiving more than one auditory event outside the sweet-spot. This was accomplished by the usage of different head-orientations during the prediction of the localization. Afterwards, a Gaussian mixture model helped to identify the different directions.

For focused sources synthesized with WFS the model is not able to predict the localization of focused sources in the whole listening area with a high accuracy. It performed better in the localization task of the synthesized source than the listeners. By inspecting ITDs for different frequency channels the model still allows interesting discernments.

Towards the end of the chapter the validated model was used for some example applications like predicting the size of a sweet-spot. These applications highlight the value a binaural model could have for planing new loudspeaker array setups for SFS or comparing different SFS approaches.

[26] Ibid.

7
Conclusions

Sound field synthesis allows to create controllable sound fields in an extended volume. This emphasizes the usage of such methods in spatial sound presentation in order to convey a rich audio experience to the listener. The shortcoming of SFS is its underlying theoretical assumption of continuous loudspeaker distributions that are not possible to build in practice. Nonetheless, experiences with build sound field synthesis setups showed that a spatial convincing presentation is nonetheless possible, by accepting slight degradations of the perceived timbre.

The goal of this thesis was to investigate in more detail the employment of sound field synthesis for spatial sound presentations. In the introduction a couple of corresponding research questions were formulated that are repeated here and tried to answered in the light of the achieved results. *What is the best way to give a good spatial impression of the presented sound? How many loudspeakers are needed to do this? Is it possible with the current hardware limitations to create a whole sound field in a convincing way? What is the influence of the spatial impression on the overall quality a listener experiences while listening to the played back sound?*

The results from the spatial fidelity experiments in Section 5.1 indicate that especially WFS gives a good spatial impression in the whole audience area. This was achieved even for the relatively low number of 28 loudspeakers for a circular loudspeaker array with a diameter of 3 m and a corresponding distance between them of 34 cm. Further WFS can be implemented in an efficient way and arbitrary geometries of loudspeaker arrays can be used.

Considering not only the spatial but the overall impression of the synthesized sound field, results become more subtle. Rumsey et al.[1] conducted an experiment in which listeners rated the spatial fidelity, timbral fidelity, and the perceived overall quality of stereophonic surround setups. The results show that the contribution of spatial fidelity to the overall quality is only 30%. Timbral fidelity accounts for the rest. Their findings demonstrate that for SFS timbral fidelity might have the same importance for the perceived quality. The results from the experiments in Section 5.2 show that timbral fidelity could only be achieved by employing more than 3 000 loudspeakers for a circular loudspeaker array with a diameter of 3 m. However,

[1] Rumsey et al., op. cit.

practical setups with the same size will seldom have more than 64 loudspeakers. In this case timbral fidelity cannot be achieved. On the other hand the same is true for two-channel stereophony, without large impairments of the perceived quality.

Further investigations of sound field synthesis in comparison to stereophonic techniques should be carried out to assess its perceived quality. Therefore, more complex sound scenes should be synthesized and listener should rate the plausibility of the corresponding auditory scene. Another research topic is of interest in order to understand the influence of sound field errors on its perceived quality. The experimental results from this thesis point to a close connection between the perception of coloration and the precedence effect and one hypothesis is that both are part of the same mechanism in the brain.

By limiting the audience area to a region of approximately the size of a human head, band-limited NFC-HOA is an interesting approach that can provide a convincing sound field. The limitation to a small region enables the synthesis of an error free sound field which achieves spatial and timbral fidelity in this region.

Further Resources

This section is devoted to the acknowledgements of further literature that had an impact on the writing of this thesis but was not mentioned so far.

The underlying design principals of the figures in this thesis are motivated by Tuftes work.[1] The colormap used for plotting all of the numerical sound field simulations is given by Moreland.[2] Most of the colors in the other figures are based on colormaps published by Brewer.[3]

The idea of predicting the localization in the whole audience area has its origin in a paper by Merchel et al.[4]

[1] E.g. E. R. Tufte. *Envisioning Information*. Cheshire: Graphics Press LLC, 2011

[2] K. Moreland. "Diverging Color Maps for Scientific Visualization". In: *International Symposium on Visual Computing*. 2009, pp. 92–103

[3] colorbrewer2.org
C. A. Brewer. *Designing Better Maps: A Guide for GIS Users*. New York: ESRI Press, 2005

[4] S. Merchel and S. Groth. "Adaptively Adjusting the Stereophonic Sweet Spot to the Listener's Position". *Journal of the Audio Engineering Society* 58.10 (2010), pp. 809–17.

Bibliography

Abbe, E. “III.—Some Remarks on the Apertometer”. *Journal of the Royal Microscopical Society* 3 (1880), pp. 20–31.

Abramowitz, M. and I. A. Stegun. *Handbook of Mathematical Functions*. Washington: National Bureau of Standards, 1972.

Ahrens, J. *Analytic Methods of Sound Field Synthesis*. New York: Springer, 2012.

Ahrens, J. and S. Spors. “On the Secondary Source Type Mismatch in Wave Field Synthesis Employing Circular Distributions of Loudspeakers”. In: *127th Audio Engineering Society Convention*. 2009, Paper 7952.

– “Sound Field Reproduction Using Planar and Linear Arrays of Loudspeakers”. *IEEE Transactions on Audio, Speech, and Language Processing* 18.8 (2010), pp. 2038–50.

– “Spatial encoding and decoding of focused virtual sound sources”. In: *International Symposium on Ambisonics and Spherical Acoustics*. 2009.

Algazi, V. R., C. Avendano, and R. O. Duda. “Estimation of a Spherical-Head Model from Anthropometry”. *Journal of the Audio Engineering Society* 49.6 (2001), pp. 472–79.

ANSI. *American National Standard Acoustical Terminology, ANSI S1.1-1994*. New York, 1994.

Arfken, G. B. and H. J. Weber. *Mathematical Methods for Physicists*. Amsterdam: Elsevier, 2005.

Avni, A. et al. “Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution”. *The Journal of the Acoustical Society of America* 133.5 (2013), pp. 2711–21.

Berg, J. and F. Rumsey. “Spatial Attribute Identification and Scaling by Repertory Grid Technique and other methods”. In: *16th Audio Engineering Society Conference*. 1999, pp. 51–66.

Berkhout, A. “A holographic approach to acoustic control”. *Journal of the Audio Engineering Society* 36.12 (1988), pp. 977–95.

Berkley, D. A. “Hearing in rooms”. In: *Directional Hearing*. Ed. by W. A. Yost and G. Gourevitch. New York: Springer, 1987, pp. 249–60.

Bertet, S. et al. “Investigation on Localisation Accuracy for First and Higher Order Ambisonics Reproduced Sound Sources”. *Acta Acustica* 99.4 (2013), pp. 642–57.

Blauert, J. *Spatial Hearing*. The MIT Press, 1997.

Blauert, J. and U. Jekosch. "Concepts Behind Sound Quality: Some Basic Considerations". In: *International Congress and Exposition on Noise Control Engineering*. 2003.

Blumlein, A. D. "Improvements in and relating to Sound-transmission, Sound-recording and Sound-reproducing Systems". *Journal of the Audio Engineering Society* 6.2 (1958), pp. 91–98, 130.

Born, M. et al. *Principles of Optics*. Cambridge University Press, 1999.

Bracewell, R. N. *The Fourier Transform and its Applications*. Boston: McGraw Hill, 2000.

Brand, A. et al. "Precise inhibition is essential for microsecond interaural time difference coding". *Nature* 417.6888 (2002), pp. 543–47.

Breebaart, J., S. van de Par, and A. Kohlrausch. "Binaural processing model based on contralateral inhibition. I. Model structure". *The Journal of the Acoustical Society of America* 110.2 (2001), pp. 1074–88.

Brewer, C. A. *Designing Better Maps: A Guide for GIS Users*. New York: ESRI Press, 2005.

Bronkhorst, A. W. "Localization of real and virtual sound sources". *The Journal of the Acoustical Society of America* 98.5 (1995), pp. 2542–53.

Bronkhorst, A. W. and T. Houtgast. "Auditory distance perception in rooms." *Nature* 397.6719 (1999), pp. 517–20.

Brüggen, M. "Klangverfärbungen durch Rückwürfe und ihre auditive und instrumentelle Kompensation". PhD thesis. Ruhr-Universität Bochum, 2001.

Bruijn, W. P. J. de. "Application of Wave Field Synthesis in Videoconferencing". PhD thesis. Technische Universiteit Delft, 2004.

Brungart, D. S. and W. M. Rabinowitz. "Auditory localization of nearby sources. Head-related transfer functions." *The Journal of the Acoustical Society of America* 106.3 Pt 1 (1999), pp. 1465–79.

Carr, C. E. and M. Konishi. "A circuit for detection of interaural time differences in the brain stem of the barn owl." *The Journal of Neuroscience* 10.10 (1990), pp. 3227–46.

Chittka, L. and A. Brockmann. "Perception space–the final frontier." *PLoS Biology* 3.4 (2005), e137.

Colton, D. and R. Kress. *Integral Equation Methods in Scattering Theory*. New York: Wiley, 1983.

Corteel, E. "On the use of irregularly spaced loudspeaker arrays for wave field synthesis, potential impact on spatial aliasing frequency". In: *International Conference on Digital Audio Effects*. 2006.

Cumming, G., F. Fidler, and D. L. Vaux. "Error bars in experimental biology." *The Journal of Cell Biology* 177.1 (2007), pp. 7–11.

D'Errico, F. et al. "Archaeological Evidence for the Emergence of Language, Symbolism, and Music – An Alternative Multidisciplinary Perspective". *Journal of World Prehistory* 17.1 (2003), pp. 1–70.

Dietz, M., S. D. Ewert, and V. Hohmann. "Auditory model based direction estimation of concurrent speakers from binaural signals". *Speech Communication* 53.5 (2011), pp. 592–605.

Dietz, M. et al. "Coding of temporally fluctuating interaural timing disparities in a binaural processing model based on phase differences." *Brain Research* 1220 (2008), pp. 234–45.

Donoho, D. L. et al. "Reproducible Research in Computational Harmonic Analysis". *Computing in Science & Engineering* 11.1 (2009), pp. 8–18.

Emiroglu, S. S. "Timbre perception and object separation with normal and impaired hearing". PhD thesis. Carl-von-Ossietzky-Universität Oldenburg, 2007.

Erbes, V. et al. "An extraaural headphone system for optimized binaural reproduction". In: *39th German Annual Conference on Acoustics*. 2013.

Ewert, S. D. "Characterizing frequency selectivity for envelope fluctuations". *The Journal of the Acoustical Society of America* 108.3 (2000), pp. 1181–96.

Faller, C. and J. Merimaa. "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence". *The Journal of the Acoustical Society of America* 116.5 (2004), pp. 3075–89.

Fazi, F. M. "Sound Field Reproduction". PhD thesis. University of Southampton, 2010.

Fazi, F. M. and P. A. Nelson. "Sound field reproduction as an equivalent acoustical scattering problem". *The Journal of the Acoustical Society of America* 134.5 (2013), pp. 3721–9.

Frank, M. "Phantom Sources using Multiple Loudspeakers". PhD thesis. University of Music and Performing Arts Graz, 2013.

Frank, M., F. Zotter, and A. Sontacchi. "Localization Experiments Using Different 2D Ambisonics Decoders". In: *VDT International Convention*. 2008.

Geier, M., J. Ahrens, and S. Spors. "The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods". In: *124th Audio Engineering Society Convention*. 2008, Paper 7330.

Geier, M. et al. "Perceptual Evaluation of Focused Sources in Wave Field Synthesis". In: *128th Audio Engineering Society Convention*. 2010, Paper 8069.

Gerzon, M. A. "Periphony: With-Height Sound Reproduction". *Journal of the Audio Engineering Society* 21.1 (1973), pp. 2–10.

Goupell, M. J. and W. M. Hartmann. "Interaural fluctuations and the detection of interaural incoherence: Bandwidth effects". *The Journal of the Acoustical Society of America* 119.6 (2006), pp. 3971–86.

Grothe, B., M. Pecka, and D. McAlpine. "Mechanisms of Sound Localization in Mammals". *Physiological Reviews* 90 (2010), pp. 983–1012.

Gumerov, N. A. and R. Duraiswami. *Fast Multipole Methods for the Helmholtz Equation in Three Dimensions*. Amsterdam: Elsevier, 2004.

Hamasaki, K., K. Hiyama, and H. Okumura. "The 22.2 Multichannel Sound System and Its Application". In: *118th Audio Engineering Society Convention*. 2005, Paper 6406.

Hartmann, W. M. and A. Wittenberg. “On the externalization of sound images.” *The Journal of the Acoustical Society of America* 99.6 (1996), pp. 3678–88.

Herrin, D. W. et al. “A New Look at the High Frequency Boundary Element and Rayleigh Integral Approximations”. In: *Noise & Vibration Conference and Exhibition*. 2003.

Hess, W. “Influence of head-tracking on spatial perception”. In: *117th Audio Engineering Society Convention*. 2004, Paper 6288.

Hohmann, V. “Frequency analysis and synthesis using a Gammatone filterbank”. *Acta Acustica united with Acustica* 88.3 (2002), pp. 433–42.

Horbach, U. et al. “Design and Applications of a Data-based Auralization System for Surround Sound”. In: *106th Audio Engineering Society Convention*. 1999, Paper 4976.

Huygens, C. *Treatise on Light*. Ed. by S. P. Thompson. London: Macmillan & Co, 1912.

Ince, D. C., L. Hatton, and J. Graham-Cumming. “The case for open computer programs”. *Nature* 482.7386 (2012), pp. 485–88.

Jeffress, L. A. “A place theory of sound localization.” *Journal of Comparative and Physiological Psychology* 41.1 (1948), pp. 35–9.

Kelly, G. A. *The Psychology of Personal Constructs*. New York: Norton, 1955.

Kerber, S. et al. “Experimental investigations into the distance perception of nearby sound sources: Real vs. WFS virtual nearby sources”. In: *Proceedings of the Joint Congress CFA/DAGA*. 2004, pp. 1041–42.

Knapp, C. K. and G. C. Carter. “The Generalized Correlation Method for Estimation of Time Delay”. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24.4 (1976), pp. 320–27.

Koenig, W. “Subjective Effects in Binaural Hearing”. *The Journal of the Acoustical Society of America* 22.1 (1950), pp. 61–62.

Kubovy, M. and D. V. Van Valkenburg. “Auditory and visual objects.” *Cognition* 80.1-2 (2001), pp. 97–126.

Lax, M. and H. Feshbach. “On the Radiation Problem at High Frequencies”. *The Journal of the Acoustical Society of America* 19.4 (1947), pp. 682–90.

Leakey, D. M. “Some Measurements on the Effects of Interchannel Intensity and Time Differences in Two Channel Sound Systems”. *The Journal of the Acoustical Society of America* 31.7 (1959), pp. 977–86.

Letowski, T. R. “Sound quality assessment: concepts and criteria”. In: *87th Audio Engineering Society Convention*. 1989, Paper 2825.

Lewald, J., G. J. Dörrscheidt, and W. H. Ehrenstein. “Sound localization with eccentric head position.” *Behavioural Brain Research* 108.2 (2000), pp. 105–25.

Lindau, A., T. Hohn, and S. Weinzierl. “Binaural resynthesis for comparative studies of acoustical environments”. In: *122nd Audio Engineering Society Convention*. 2007, Paper 7032.

Lindau, A., H.-J. Maempel, and S. Weinzierl. “Minimum BRIR grid resolution for dynamic binaural synthesis”. *The Journal of the Acoustical Society of America* 123.5 (2008), pp. 3851–56.

Lindau, A. and S. Weinzierl. “Assessing the Plausibility of Virtual Acoustic Environments”. *Acta Acustica united with Acustica* 98.5 (2012), pp. 804–10.

– “FABIAN – An instrument for software-based measurement of binaural room impulse responses in multiple degrees of freedom”. In: *Tonmeister Tagung*. November. 2006.

Lindemann, W. “Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals.” *The Journal of the Acoustical Society of America* 80.6 (1986), pp. 1608–22.

Litovsky, R. Y. et al. “The precedence effect.” *The Journal of the Acoustical Society of America* 106.4 (1999), pp. 1633–54.

Lucas, B. G. and T. G. Muir. “The field of a focusing source”. *The Journal of the Acoustical Society of America* 72.4 (1982), pp. 1289–96.

Majdak, P. et al. “Spatially Oriented Format for Acoustics: A Data Exchange Format Representing Head-Related Transfer Functions”. In: *134th Audio Engineering Society Convention*. 2013, Paper 8880.

Majdak, P. et al. “The Accuracy of Localizing Virtual Sound Sources: Effects of Pointing Method and Visual Environment”. In: *124th Audio Engineering Society Convention*. 2008, Paper 7407.

Makous, J. C. and J. C. Middlebrooks. “Two-dimensional sound localization by human listeners.” *The Journal of the Acoustical Society of America* 87.5 (1990), pp. 2188–200.

Masiero, B. S. “Individualized Binaural Technology”. PhD thesis. RWTH Aachen, 2012.

May, T., S. van de Par, and A. Kohlrausch. “A Binaural Scene Analyzer for Joint Localization and Recognition of Speakers in the Presence of Interfering Noise Sources and Reverberation”. *IEEE Transactions on Audio, Speech, and Language Processing* 20.7 (2012), pp. 2016–30.

McAdams, S. et al. “Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes.” *Psychological Research* 58.3 (1995), pp. 177–92.

McAlpine, D., D. Jiang, and a. R. Palmer. “A neural code for low-frequency sound localization in mammals.” *Nature Neuroscience* 4.4 (2001), pp. 396–401.

Merchel, S. and S. Groth. “Adaptively Adjusting the Stereophonic Sweet Spot to the Listener’s Position”. *Journal of the Audio Engineering Society* 58.10 (2010), pp. 809–17.

Mills, A. W. “On the minimum audible angle”. *The Journal of the Acoustical Society of America* 30.4 (1958), pp. 237–46.

Møller, H. “Fundamentals of Binaural Technology”. *Applied Acoustics* 36 (1992), pp. 171–218.

Moncel, T. du. “The international exhibition and congress of electricity at Paris”. *Nature* October 20 (1881), pp. 585–89.

Moore, B. C. J. *An Introduction to the Psychology of Hearing*. Bingley: Emerald, 2012.

Moreland, K. "Diverging Color Maps for Scientific Visualization". In: *International Symposium on Visual Computing*. 2009, pp. 92–103.

Morse, P. M. and H. Feshbach. *Methods of Theoretical Physics*. Minneapolis: Feshbach Publishing, 1981.

Müller, J. et al. "The BoomRoom: Mid-air Direct Interaction with Virtual Sound Sources". In: *Conference on Human Factors in Computing Systems*. 2014.

Oldfield, R. "The analysis and improvement of focused source reproduction with wave field synthesis". PhD thesis. University of Salford, 2013.

Olive, S. E., T. Welti, and W. L. Martens. "Listener Loudspeaker Preference Ratings Obtained in situ Match Those Obtained via a Binaural Room Scanning Measurement and Playback System". In: *122nd Audio Engineering Society Convention*. 2007, Paper 7034.

Olive, S. E. et al. "The Variability of Loudspeaker Sound Quality Among Four Domestic-Sized Rooms". In: *99th Audio Engineering Society Convention*. 1995, Paper 4092.

Park, M. "Models of binaural hearing for sound lateralisation and localisation". PhD thesis. University of Southampton, 2007.

Patel, A. D. *Music, Language And The Brain*. New York: Oxford University Press, 2010.

Plomp, R., L. C. W. Pols, and J. P. van de Geer. "Dimensional Analysis of Vowel Spectra". *The Journal of the Acoustical Society of America* 41.3 (1967), pp. 707–12.

Popko, R. *Zur Hörbarkeit und Interpolation von Kopf-über-Torso-Orientierungen bei Aufnahmen mit einem Kopf-und-Torso-Simulator*. Technische Universität Berlin, 2013.

Puckette, M. S. "Pure Data: another integrated computer music environment". In: *Second Intercollege Computer Music Concerts*. 1996, pp. 37–41.

Pulkki, V. "Coloration of Amplitude-Panned Virtual Sources". In: *110th Audio Engineering Society Convention*. 2001, Paper 5402.

– "Virtual Sound Source Positioning Using Vector Base Amplitude Panning". *Journal of the Audio Engineering Society* 45.6 (1997), pp. 456–66.

Puria, S. "Measurements of human middle ear forward and reverse acoustics: Implications for otoacoustic emissions". *The Journal of the Acoustical Society of America* 113.5 (2003), pp. 2773–89.

Raake, A. and J. Blauert. "Comprehensive modeling of the formation process of sound-quality". In: *International Workshop on Quality of Multimedia Experience*. 2013, pp. 76–81.

Raatgever, J. "On the binaural processing of stimuli with different interaural phase relations". PhD thesis. Technische Universiteit Delft, 1980.

Rohdenburg, T. et al. "Objective perceptual quality assessment for self-steering binaural hearing aid microphone arrays". In: *IEEE*

International Conference on Acoustics, Speech, and Signal Processing. 2008, pp. 2449–52.

Ruggero, M. A. et al. “Basilar-membrane responses to tones at the base of the chinchilla cochlea”. *The Journal of the Acoustical Society of America* 101.4 (1997), pp. 2151–63.

Rumsey, F. “Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm”. *Journal of the Audio Engineering Society* 50.9 (2002), pp. 651–66.

Rumsey, F. et al. “On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality”. *The Journal of the Acoustical Society of America* 118.2 (2005), pp. 968–76.

Schultz, F. and S. Spors. “Comparing Approaches to the Spherical and Planar Single Layer Potentials for Interior Sound Field Synthesis”. *Acta Acustica* 100.5 (2014), pp. 900–11.

Seeber, B. U. “Untersuchung der auditiven Lokalisation mit einer Lichtzeigermethode”. PhD thesis. 2003.

Sloboda, J., A. Lamont, and A. Greasly. “Choosing to hear music”. In: *The Oxford Handbook of Music Psychology*. Ed. by S. Hallam, I. Cross, and M. Thaut. New York: Oxford University Press, 2009, pp. 431–40.

Solvang, A. “Spectral Impairment for Two-Dimensional Higher Order Ambisonics”. *Journal of the Audio Engineering Society* 56.4 (2008), pp. 267–79.

Søndergaard, P. L. and P. Majdak. “The auditory-modeling toolbox”. In: *The technology of binaural listening*. Ed. by J. Blauert. New York: Springer, 2013, pp. 33–56.

Spors, S. and J. Ahrens. “Local Sound Field Synthesis by Virtual Secondary Sources”. In: *40th Audio Engineering Society Conference*. 2010, Paper 6.3.

– “Reproduction of Focused Sources by the Spectral Division Method”. In: *International Symposium on Communications, Control and Signal Processing*. 2010.

Spors, S., V. Kuscher, and J. Ahrens. “Efficient realization of model-based rendering for 2.5-dimensional near-field compensated higher order Ambisonics”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 2011, pp. 61–64.

Spors, S., R. Rabenstein, and J. Ahrens. “The Theory of Wave Field Synthesis Revisited”. In: *124th Audio Engineering Society Convention*. 2008, Paper 7358.

Spors, S. and F. Zotter. “Spatial Sound Synthesis with Loudspeakers”. In: *Cutting Edge in Spatial Audio, EAA Winter School*. 2013, pp. 32–37.

Spors, S. et al. “Spatial Sound With Loudspeakers and Its Perception: A Review of the Current State”. *Proceedings of the IEEE* 101.9 (2013), pp. 1920–38.

Start, E. W. “Direct Sound Enhancement by Wave Field Synthesis”. PhD thesis. Technische Universiteit Delft, 1997.

Steinberg, J. and W. B. Snow. "Symposium on wire transmission of symphonic music and its reproduction in auditory perspective: Physical Factors". *Bell System Technical Journal* 13.2 (1934), pp. 245–58.

Strutt, J. W. "On our perception of sound direction". *Philosophical Magazine* 13.74 (1907), pp. 214–32.

Takanen, M., M. Hiipakka, and V. Pulkki. "Audibility of coloration artifacts in HRTF filter designs". In: *45th Audio Engineering Society Conference*. 2012, Paper 3.3.

Takanen, M., O. Santala, and V. Pulkki. "Visualization of functional count-comparison-based binaural auditory model output". *Hearing Research* 309 (2014), pp. 147–163.

Talbot, K. et al. "Synaptic dysbindin-1 reductions in schizophrenia occur in an isoform-specific manner indicating their subsynaptic location." *PLoS ONE* 6.3 (2011), e16886.

Theile, G. "Über die Lokalisation im überlagerten Schallfeld". PhD thesis. Technische Universität Berlin, 1980.

Tufte, E. R. *Envisioning Information*. Cheshire: Graphics Press LLC, 2011.

Verheijen, E. "Sound Reproduction by Wave Field Synthesis". PhD thesis. Technische Universiteit Delft, 1997.

Vogel, P. "Application of Wave Field Synthesis in Room Acoustics". PhD thesis. Technische Universiteit Delft, 1993.

Völk, F. "Psychoakustische Experimente zur Distanz mittels Wellenfeldsynthese erzeugter Hörereignisse". In: *36th German Annual Conference on Acoustics*. 2010, pp. 1065–66.

Völk, F., E. Faccinelli, and H. Fastl. "Überlegungen zu Möglichkeiten und Grenzen virtueller Wellenfeldsynthese". In: *36th German Annual Conference on Acoustics*. 2010, pp. 1069–70.

Völk, F. and H. Fastl. "Wave Field Synthesis with Primary Source Correction: Theory, Simulation Results, and Comparison to Earlier Approaches". In: *133rd Audio Engineering Society Convention*. 2012, Paper 8717.

Völk, F., M. Straubinger, and H. Fastl. "Psychoacoustical experiments on loudness perception in wave field synthesis". *20th International Congress on Acoustics* (2010).

Vries, D. de. *Wave Field Synthesis*. New York: Audio Engineering Society, 2009.

Wang, D. and G. J. Brown, eds. *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Hoboken: Wiley, 2006.

Wierstorf, H., A. Raake, and S. Spors. "Binaural assessment of multichannel reproduction". In: *The technology of binaural listening*. Ed. by J. Blauert. New York: Springer, 2013, pp. 255–78.

– "Localization in Wave Field Synthesis and higher order Ambisonics at different positions within the listening area". In: *39th German Annual Conference on Acoustics*. 2013.

– "Localization of a virtual point source within the listening area for Wave Field Synthesis". In: *133rd Audio Engineering Society Convention*. 2012, Paper 8743.

– "Psychoakustik der Wellenfeldsynthese: Vor- und Nachteile binauraler Simulation". In: *38th German Annual Conference on Acoustics*. 2012.

Wierstorf, H. and S. Spors. "Die Rolle des Präzedenzeffektes bei der Wahrnehmung von räumlichen Aliasingartefakten bei der Wellenfeldsynthese". In: *36th German Annual Conference on Acoustics*. 2010.

– "Sound Field Synthesis Toolbox". In: *132nd Audio Engineering Society Convention*. 2012, eBrief 50.

Wierstorf, H., S. Spors, and A. Raake. "Perception and evaluation of sound fields". In: *59th Open Seminar on Acoustics*. 2012.

Wierstorf, H. et al. "A Free Database of Head-Related Impulse Response Measurements in the Horizontal Plane with Multiple Distances". In: *130th Audio Engineering Society Convention*. 2011, eBrief 6.

Wierstorf, H. et al. "Coloration in Wave Field Synthesis". In: *55th Audio Engineering Society Conference*. 2014, Paper 5.3.

Wierstorf, H. et al. "Perception of Focused Sources in Wave Field Synthesis". *Journal of the Audio Engineering Society* 61.1 (2013), pp. 5–16.

Wightman, F. L. 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 (1992), pp. 1648–61.

Williams, E. G. *Fourier Acoustics*. San Diego: Academic Press, 1999.

Wittek, H. "Perceptual differences between wavefield synthesis and stereophony". PhD thesis. University of Surrey, 2007.

Yon, S., M. Tanter, and M. Fink. "Sound focusing in rooms: The time-reversal approach". *The Journal of the Acoustical Society of America* 113.3 (2003), pp. 1533–43.

Zahorik, P., D. S. Brungart, and A. W. Bronkhorst. "Auditory Distance Perception in Humans: A Summary of Past and Present Research". *Acta Acustica united with Acustica* 91 (2005), pp. 409–20.

Zotter, F. and S. Spors. "Is sound field control determined at all frequencies? How is it related to numerical acoustics?" In: *52nd Audio Engineering Society Conference*. 2013, Paper 1.3.