7 - 12 SEPTEMBER 2014, KRAKÓW
FORUM ACUSTICUM

Predicting localization accuracy for stereophonic downmixes in Wave Field Synthesis

Hagen Wierstorf
Assessment of IP-based Applications, T-Labs, Technische Universität Berlin, Berlin, Germany.

Sascha Spors
Institute of Communications Engineering, Universität Rostock, Rostock, Germany.

Summary
Wave Field Synthesis enables the creation of a correct spatial impression for an extended listening area. However, the synthesized sound scenes are most often created with a model-based rendering approach. This enables, beside other advantages, interactivity of the listener with the sound scene. On the downside it requires anechoic recordings of every sound event of the sound scene, a requirement that most often does not hold for existing music recordings and productions. In this study we investigate methods of reproducing two-channel stereophonic recordings with Wave Field Synthesis. Thereby different virtual stereophonic loudspeaker layouts, like point sources and plane waves are arranged. It is further investigated how these interact with different geometries of the underlying loudspeaker array for Wave Field Synthesis. As typical setups linear, circular and box shaped geometry is employed. A common drawback of stereophonic reproduction is the sweet-spot for localization. Outside of the sweet-spot the listeners start to localize the reproduced sources towards the single loudspeakers. Recent advantages in predicting the localization for Wave Field Synthesis setups with a binaural model will be utilized in this study. Applying the binaural model the influences of the downmixing method and the underlying geometry are investigated. Especially the number of active loudspeakers and the usage of plane waves for stereophonic presentation differ from a typical stereophonic setup and potentially allow to increase the sweet-spot size.

PACS no. 43.66.Qp, 43.60.Sx

1. Introduction

Wave Field Synthesis is a spatial audio presentation technique that uses several loudspeakers to synthesize a desired sound field [1]. It achieves this by driving the single loudspeakers in a way that its signals superimpose to the desired wave fronts in analogy to the Huygens-Fresnel principle [2]. The difference to two-channel stereophony lies in the control of the sound field not only on a single point or a line but in an extended listening area. Up to which frequency the sound field can be controlled depends solely on the number of applied loudspeakers. A common setup with a loudspeaker distance of around 15 cm allows the control of the sound field up to 1.1 kHz. It is obvious that this implies errors in the synthesized sound field for practical applications such as music reproduction that involves frequencies up to 20 kHz. We have shown previously that the errors in the synthesized sound field have very little influence on localizing synthesized sources [3], but can lead to severe deviations of the sound color [4].

This study will not review these perceptional aspects, but will focus on the question of how to create content-rich sound fields with Wave Field Synthesis. A straightforward way could be to record a complete sound field for reproduction, a method called *data-based* rendering [5]. In contrast to recording video images this is not an easy task, because the wave length in the range of several orders have to be recorded. To achieve this microphone arrays can be used, but the recorded sound field has a limited spatial resolution and suffers from spatial aliasing. For stereophony the situation is different for data-based rendering, because not a complete sound field has to be recorded and classical main microphone setups work well. Another way of creating a sound field is to use mathematical models for the field, for example point sources or plane waves. This technique is called *model-based* rendering. In order to reproduce a human speaker, only a dry recording of his voice is needed. In this way the recorded signal has no spatial information at all, and

that information is completely determined by the applied mathematical models. In stereophony the same can be achieved by applying panning to the dry signal.

One of the advantages of model based rendering is its ability to allow an *object-based* representation of a sound field. This implies that the number of stored and transmitted channels becomes independent of the number of loudspeakers of the presentation system which is appreciated especially for Wave Field Synthesis where varying numbers of loudspeaker are applied. For every virtual source of the sound field only the dry audio channel and the information about its source model and position has to be transmitted. Another possibility that comes with object-based audio is interactivity, because the information about its position or loudness can be easily changed without the need to altering the recorded signal. Lately, object-based audio appeared also in the audio industry [6, 7].

Beside all its advantages object-based audio has also a number of disadvantages or challenges. The practical source models are more or less limited to point sources and plane waves which allow not for the presentation of extended or diffuse sources. In addition, for a lot of situations there exist not the possibility to get dry signals for every source of a sound field. In the context of Wave Field Synthesis this has led to the concept of *virtual panning spots* [8] that allow the inclusion of *channel-based* content like stereophony in object-based synthesis. So far this was done by the creation of virtual loudspeakers as panning spots by synthesizing them as virtual point sources applying the stereo-channels as source material. Thereby it was not discussed if the Wave Field Synthesis system with its own physical limitations has an influence on the localization properties of the stereophonic recording. In this study we will apply a binaural model to investigate this question and show further that the localization properties depend also on the applied loudspeaker array and virtual source model of the panning spots.

In the next section we will briefly introduce the binaural model and its ability to predict localization in Wave Field Synthesis. Afterwards we introduce different loudspeaker setups and different source models as virtual panning spots. At the end we analyze the perceptual consequences of these different methods with the help of a binaural model.

2. Predicting localization in Wave Field Synthesis

In order to predict localization with a binaural model the signals at the two ears are needed. These can be gained with the help of binaural synthesis. In this case all loudspeakers of the applied loudspeaker array are simulated via binaural synthesis over headphones. In an anechoic chamber head-related transfer functions with a resolution of 1° were measured [9] and

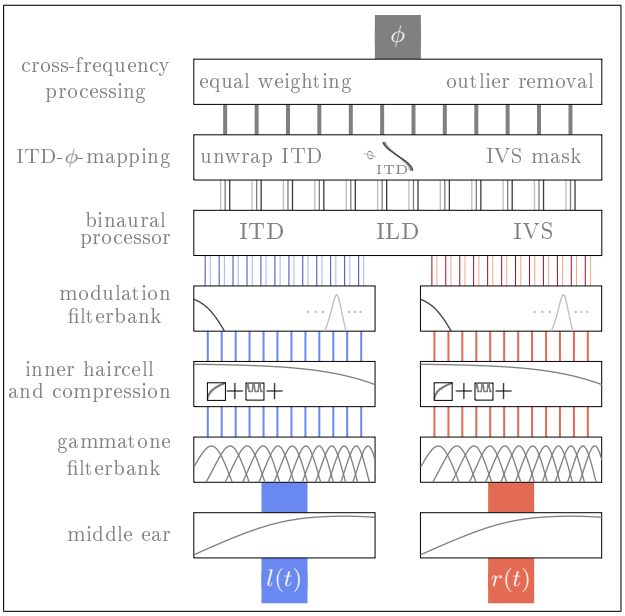

Figure 1. Sketch of the applied binaural model. At the bottom the two ear signals are input to the model. At the end the binaural parameters are mapped to a single direction estimation averaged over the whole time of the input signals.

afterwards extra- and interpolated to allow arbitrary loudspeaker setups. The head-related transfer functions for every single loudspeaker are then weighted in amplitude and delayed in time corresponding to the Wave Field Synthesis driving signals calculated for the given source models and positions.

To verify the validity of the binaural synthesis we compared the localization of a point source presented via a real loudspeaker or simulated via binaural synthesis and found no difference in the perception of its direction [10].

After simulating the two ear signals they are fed into the binaural model. We apply a model that piggybacks on the model presented by Dietz et al. [11]. The different stages of the model are presented in Fig. 1. After a band-pass filter approximated the middle ear transfer function the two input signals are filtered by a gammatone-filterbank [12] into twelve frequency channels in the range of 200 Hz to 1400 Hz. In every frequency channel compression of the cochlea and half-wave rectification together with low-pass filtering representing the hair cells is applied. The next stage involves another filterbank that removes the DC components that are added by the hair cell processing. In addition, the filterbank has a low-pass filter with a cutoff frequency of 30 Hz for smoothing the calculation of the interaural level differences in time. After the filterbank the different binaural parameters interaural time difference (ITD), interaural level difference (ILD), and interaural vector strength (IVS) which is equivalent to the interaural coherence of the signals are calculated. The calculation of the ITD is based on

an estimation of the IPD, see the paper from Dietz et al. [11] for more details on the model.

In the next stage the IVS is applied as a mask hiding ITD values for instances in time where the IVS is below a given threshold. Further, a lookup table is used to map the ITD values of every frequency channel into angle values for the directions. In a last step the direction is averaged in time and over the different frequency channels, resulting in a single value for the estimated perceived direction.

The model was already applied to predict the perceived direction of synthesized point sources and plane waves in Wave Field Synthesis. For a linear loudspeaker array with different spacings between the single loudspeakers in the range of 0.2 m to 1.4 m the accuracy of the model compared to the listening test was around 1.5° [3]. This is based on 16 different positions of the listeners in the listening area and a single synthesized point source. For a circular loudspeaker array the model accuracy was 4.1° [13] based on 16 different listener positions and a synthesized point source or plane wave. Due to this accurate results the model will be directly applied in this study without corresponding listening tests.

3. Stereophonic downmixes in Wave Field Synthesis

As mentioned in the introduction one way to create a sound field with Wave Field Synthesis is *model-based* rendering. Here, we assume a mathematical model for the desired sound field S and for this calculate the driving signals D for the loudspeakers. Arbitrary models are possible for the desired sound field, but the most common ones are point sources and plane waves due to their simplicity. They are given by the following two equations.

$$S_{\text{plane wave}}(\mathbf{x}, \omega) = A(\omega) e^{-i \frac{\omega}{c} \mathbf{n}_k \mathbf{x}} , \tag{1}$$

$$S_{\text{point source}}(\mathbf{x}, \omega) = A(\omega) \frac{1}{4\pi} \frac{e^{-i \frac{\omega}{c} |\mathbf{x} - \mathbf{x}_\text{s}|}}{|\mathbf{x} - \mathbf{x}_\text{s}|} , \tag{2}$$

where $\mathbf{x}$ is a position in the sound field, $\mathbf{x}_\text{s}$ the position of the point source, $\mathbf{n}_k$ the direction of the plane wave, ω the circular frequency, c the speed of sound, and A the amplitude spectrum.

The basic idea to include two-channel stereophonic material in Wave Field Synthesis is to arrange two point sources or two plane waves as virtual loudspeakers which are then driven by the stereophonic signals. Figure 2 shows the setup for the investigations in this study. Three different loudspeaker array geometries are used for the synthesis. The virtual panning spots for stereophony are arranged with an opening angle of 60° for a listener position at the center of the listening area which is indicated by the crosses in the figure. The plane waves are arranged that they are traveling into the direction of this central point.

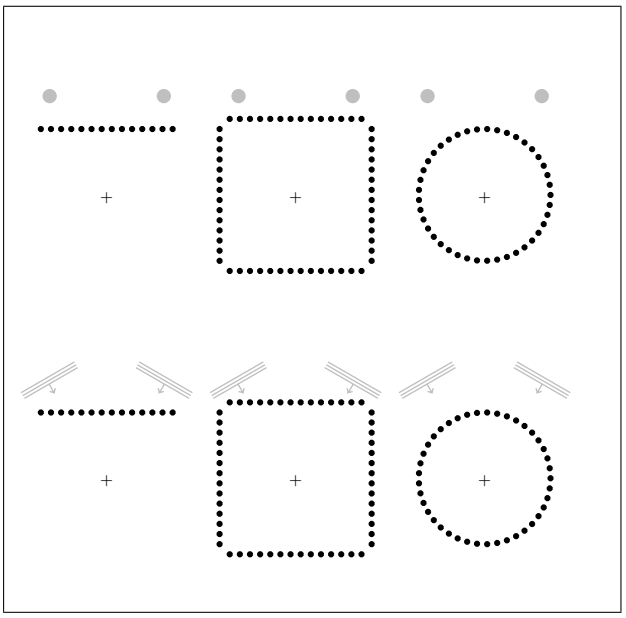

Figure 2. Geometry of the different loudspeaker arrays and virtual panning spots used in the evaluation.

Plane waves as virtual panning spots have the advantage that they are coming always from −30° and 30° independent of the position of the listener. In this way they guarantee a perfect stereophonic setup in the whole listening area. Ideal plane waves have a constant amplitude in the whole listening. This is not possible for a plane wave that is synthesized only via loudspeakers in the horizontal plane. Hence, it is likely that the advantage of the correct incidence direction is eliminated by the wrong amplitude decay.

4. Evaluation of stereophonic downmixes

In the following the stereophonic downmixes are compared to the case of a real two-loudspeaker stereophonic setup. Due to the fact that Wave Field Synthesis suffers from coloration of the synthesized sound field for loudspeaker spacings applied in this paper [4] there will be differences in timbre. There amount will not be further analyzed in this study. The ability to localize synthesized sources in Wave Field Synthesis is on the other hand quite good [13] and it exists the possibility that a stereophonic downmix is able to achieve better results than the real stereophonic setup. The binaural model presented in Section 2 is applied to the different setups and downmixing methods presented in the last section to analyze their localization properties.

Figure 3 analyzes if the size of the sweet spot differs between a real stereophonic setup and the different Wave Field Synthesis downmixes. The sweet spot describes the fact that the localization of the virtual

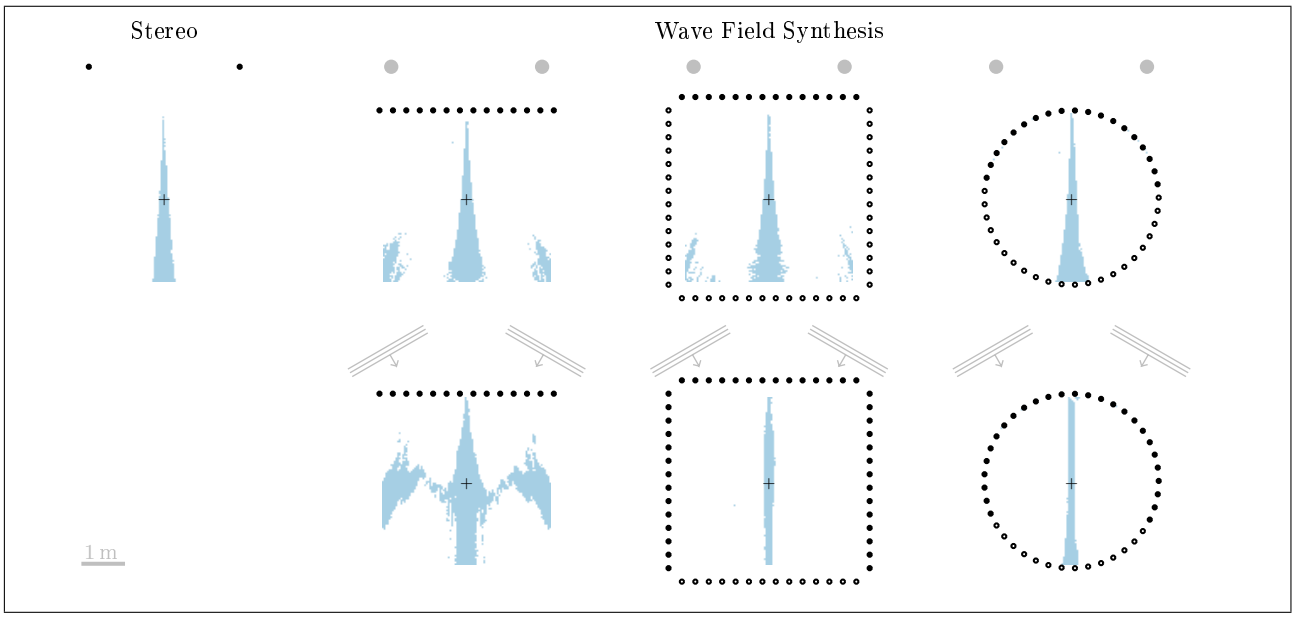

Figure 3. Sweet spot size for a two channel-stereophonic setup and different downmixing methods in Wave Field Synthesis. The source was always panned to 0°. The sweet spot is given by the blue area which highlights positions in the listening area with an absolute localization error of 5° or lower. Inactive loudspeakers are indicated by black open circles, active ones by black filled circles.

source in two-loudspeaker stereophony is only correct at a small line in the center between the two loudspeakers. Outside of this line the localization is more towards the direction of one of the two loudspeakers. To visualize this fact the perceived direction of the virtual source was calculated by the binaural model for several thousand points in the listening area and only points were the perceived direction differs 5° or less from the desired one are highlighted in blue. The left graph in the figure highlights the sweet spot for the two-channel stereophony setup. The region of the sweet spot is relatively narrow and becomes wider for larger distances to the two loudspeakers due to the geometry of the setup. For Wave Field Synthesis and point sources as virtual panning spots the sweet spot is similar at the center but has additional areas at the back of the listening area were the localization error is small. For Wave Field Synthesis and plane waves as virtual panning spots the shape of the sweet spot depends on the loudspeaker array geometry. Now, for a linear loudspeaker array the sweet spot shows a line perpendicular to the array, but also a line parallel to the loudspeaker array at the center of the listening area. For the box-shaped or circular loudspeaker array the sweet spot is only a line perpendicular to the array that has the same extend in the whole listening area.

To investigate why the sizes and positions of the sweet spot differ for different downmixes it is also of interest from what direction listeners perceive the virtual source outside of the sweet spot. Figure 4 illustrates this in more detail. Here, the perceived direction of a virtual source with a panning angle of 0° is indicated by the arrows for different listening positions. The arrows are centered at the corresponding listening position and are pointing towards the perceived direction. The color of the arrows indicates the deviation from the desired direction of the panned source. The more red the arrow the larger the deviation. For the two-loudspeaker stereophonic setup it is obvious that the listener localizes the nearest loudspeaker outside of the sweet spot. The same holds for the downmixing method using point sources as virtual panning spots. The only difference is that for listening positions in the back, especially at the sides the localization is more to the center again. This is due to the limited listening area of the applied loudspeaker arrays. The virtual panning spots are placed near the edges of the active loudspeakers which are not able to generate the desired sound field correctly at those positions in the back at the sides. In our case this brings an advantage, because the perceived direction is intended to be towards the center.

For the downmixing method applying plane waves the results look different. Starting with the linear array, the sweet spot is more shaped like a cross than a line. For frontal positions outside the sweet spot the listeners again localized towards the direction of the nearest virtual panning spot. For positions in the back outside of the sweet spot the situation is different and the listeners localize the panning spot from the opposite side. This explains also the line in the center were the sweet spot has a large extend to the sides, because here we find the transition area between these two different zones. The explanation for this behavior is given again by the limited extend of the linear

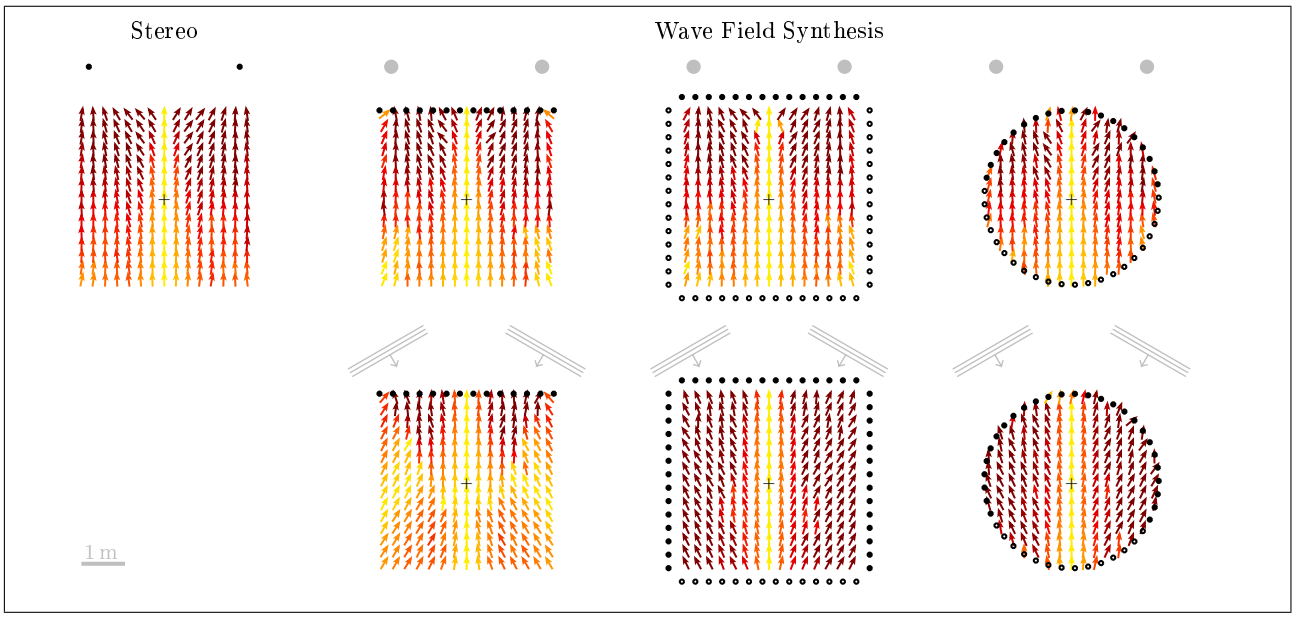

Figure 4. Perceived direction of the virtual source for different positions in the listening area. The source was always panned to 0°. The direction is given by the direction the arrows are pointing to. The arrows are centered at the corresponding listening positions and their color indicates the deviation from the desired direction with larger deviation indicated by red. Inactive loudspeakers are indicated by black open circles, active ones by black filled circles.

loudspeaker array. The plane wave coming from the left emits very few energy to the positions in the back-left of the listening area and vice versa for the plane wave coming from the right. The situation is quite different for the other two loudspeaker setups. In the case of the plane waves as virtual panning spots a lot more loudspeakers are active than for the case of point sources as virtual panning spots. By supplying energy also by the loudspeakers at the side the setups are able to synthesize both plane waves correctly in the whole listening area. This leads to a small sweet spot like in the case of the two-channel stereophony setup and a localization towards the direction of the nearest plane wave outside of the sweet spot. Unfortunately, this implies that the deviation from the desired direction is large for all positions outside of the sweet spot, even at the back as compared to the case of point sources as virtual panning spots.

So far only a virtual source panned to the center was investigated. Figure 5 shows the results for a virtual source panned to an angle of 15° with amplitude panning applying the following tangential law [14]

$$\frac{\tan(\phi)}{\tan(30°)} = \frac{\text{gain}_{\text{right}} - \text{gain}_{\text{left}}}{\text{gain}_{\text{right}} + \text{gain}_{\text{left}}} , \qquad (3)$$

where ϕ is the desired panning angle of the virtual source and $\text{gain}_{\text{left}}$ and $\text{gain}_{\text{right}}$ the two amplitude factors which are multiplied with the signal of the corresponding stereo channel.

The results show again that the downmixing method using point sources as virtual panning spots delivers the same localization in the whole listening area as it would be the case for a real two-channel stereophonic setup. For plane waves as virtual panning spots the result shows the same behavior as in Figure 4 and differs between the linear loudspeaker array and the two others.

Considering only the localization properties the evaluation results show that the usage of point sources as virtual panning spots delivers the same spatial experience as a real two-channel stereophonic setup would provide. This is also independent of the loudspeaker array geometry used for Wave Field synthesis. By applying plane waves as virtual panning spots the sweet spot for stereophony can be enlarged from a line to a cross shaped area. This depends on the usage of a linear loudspeaker array and will not happen for a circular or box shaped array. For the last two cases listener will localize the active loudspeakers at the side outside of the sweet spot.

5. Summary

In Wave Field Synthesis often model-based rendering is used to synthesize sound fields. This has the disadvantage that dry recordings of the sound sources of a given scene are needed, which is not always possible. In addition, most of the produced content nowadays is available for stereophonic presentation and stored in a channel-based manner. In this paper we reviewed the usage of virtual panning spots to reproduce stereophonic recordings in Wave Field Synthesis. To evaluate different virtual panning spots and loudspeaker arrays a binaural model was used to evaluate the localization of the reproduced source in the listening area.

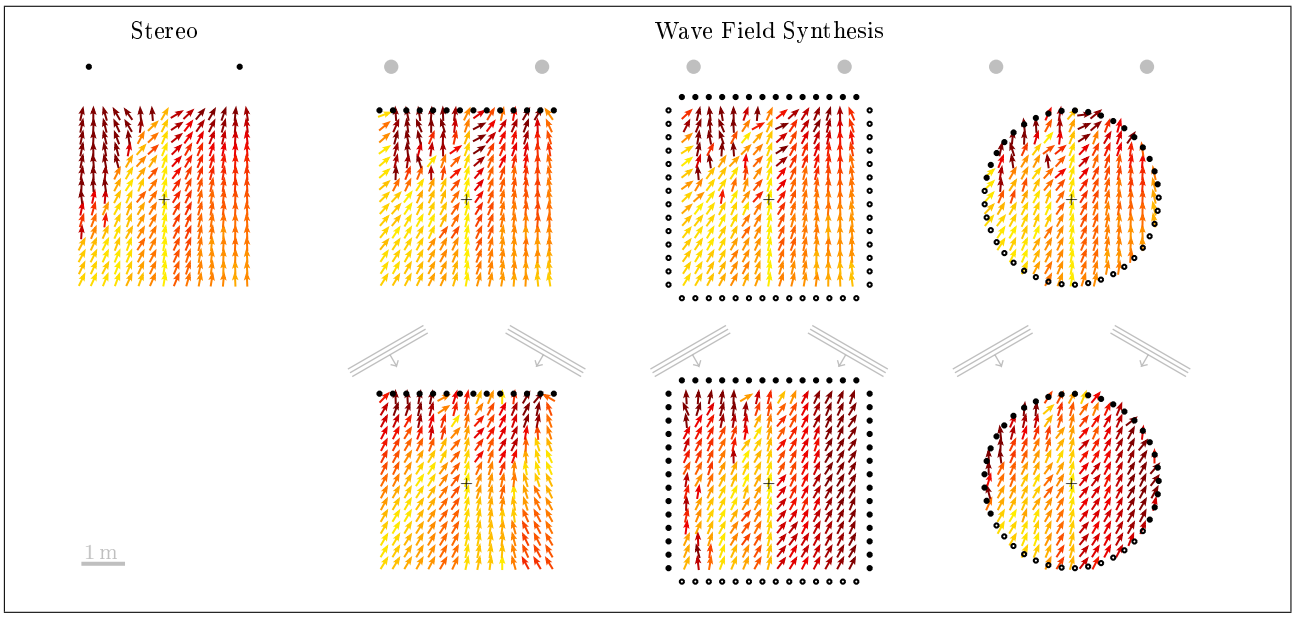

Figure 5. Perceived direction of the virtual source for different positions in the listening area. The source was always panned 15° to the right. The direction is given by the direction the arrows are pointing to. The arrows are centered at the corresponding listening positions and their color indicates the deviation from the desired direction with larger deviation indicated by red. Inactive loudspeakers are indicated by black open circles, active ones by black filled circles.

The model was able to show that the usage of point sources as virtual panning spots lead to the same spatial experience as a real two-loudspeaker stereophonic setup would provide. By the combination of a linear loudspeaker array and plane waves as virtual panning spots the sweet spot of stereophony could even be increased. Beside the good match of spatial impression there will be a difference between a real stereophonic setup and a downmixed version in Wave Field Synthesis due to coloration which is inherent to most of the Wave Field Synthesis systems [4].

Acknowledgement

This research has been supported by EU FET grant Two!Ears, ICT-618075.

References

[1] A. Berkhout: A holographic approach to acoustic control. Journal of the Audio Engineering Society **36** (1988) 977-95.

[2] C. Huygens: Treatise on Light. S. P. Thompson (ed.). Macmillan & Co, London, 1912.

[3] H. Wierstorf, A. Raake, S. Spors: Binaural Assessment of Multichannel Reproduction. J. Blauert (ed.). Springer, Heidelberg, 2013.

[4] H. Wierstorf, C. Hohnerlein, S. Spors, A. Raake: Coloration in Wave Field Synthesis. Proc. AES 55th International Conference, 2014.

[5] M. Geier, J. Ahrens, S. Spors: Object-based Audio Reproduction and the Audio Scene Description Format. Organised Sound **15** (2010) 219-27.

[6] C. Q. Robinson, S. Mehta, N. Tsingos: Scalable Format and Tools to Extend the Possibilities of Cinema Audio. SMPTE Motion Image Journal **121** (2012) 63-69.

[7] M. Mann, A. Churnside, A. Bonney, F. Melchior: Object-Based Audio Applied to Football Broadcasts. Proc. 2013 ACM international workshop on Immersive media experiences, 13-16.

[8] G. Theile, H. Wittek, M. Reisinger: Potential wavefield synthesis applications in the multichannel stereophonic world. Proc. AES 24th International Conference, 2003.

[9] H. Wierstorf, M. Geier, A. Raake, S. Spors: A Free Database of Head-Related Impulse Response Measurements in the Horizontal Plane with Multiple Distances. Proc. 130th AES Convention, eBrief 6, 2011.

[10] H. Wierstorf, S. Spors, A. Raake: Perception and evaluation of sound fields. Proc. 59th Open Seminar on Acoustics, 263-68, 2012.

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

[12] V. Hohmann: Frequency analysis and synthesis using a Gammatone filterbank. Acta Acustica united with Acustica **88** (2002) 433-42.

[13] H. Wierstorf: Perceptual Assessment of sound field synthesis. PhD thesis, Technische Universität Berlin, to appear.

[14] V. Pulkki, M. Karjalainen: Localization of amplitude-panned virtual sources I: stereophonic panning. Journal of the Audio Engineering Society **49** (2001) 739-52.