Audio Engineering Society

Convention Paper

Presented at the 128th Convention
2010 May 22–25 London, UK

Perceptual Evaluation of Focused Sources in Wave Field Synthesis

Matthias Geier, Hagen Wierstorf, Jens Ahrens, Ina Wechsung, Alexander Raake and Sascha Spors

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

Correspondence should be addressed to Matthias Geier (Matthias.Geier@telekom.de)

ABSTRACT

Wave Field Synthesis provides the possibility to reproduce virtual sound sources located between the loudspeaker array and the listener. Such sources are known as focused sources. A previously published study including an informal listening test has shown that the reproduction of focused sources is subject to audible artifacts, especially for large loudspeaker arrays. The combination of the time-reversal nature of focused sources and spatial sampling leads to pre-echos. The perception of these artifacts is quite different depending on the relative listener position. This paper describes a formal test which was conducted to verify the perceptual relevance of the physical properties found in previous papers.

1. INTRODUCTION

Wave Field Synthesis (WFS) is one of the most prominent high-resolution sound field reproduction methods which are studied and used nowadays [1]. As *Higher-Order Ambisonics* (HOA) and the *Spectral Division Method* (SDM) [2], WFS aims at the physical synthesis of a desired sound field within a given, potentially large, listening area. All of these approaches offer the potential of creating the impression of an acoustic point source located between the loudspeakers and the listener [3, 4, 5]. These virtual sources are termed *focused sources*. In practice, focused sources allow the generation of quite stunning effects, like virtual sources placed within the audience, which are not possible with traditional stereophonic techniques.

The theory of WFS assumes a spatially continuous distribution of appropriately driven acoustic point sources (secondary sources) around the listening area. However, real-life WFS systems are realized with discrete loudspeakers, therefore a spatial sampling of the continuous secondary source distribution occurs. For typical geometries and audio content this may lead to spatial sampling artifacts which

may become audible. The artifacts caused by the spatial sampling process are of special interest for focused sources. For broadband stimuli these sampling artifacts in combination with the time-reversal process used for generating the driving functions may lead to pre-echo artifacts. A previously published informal listening test [6] has revealed that these artifacts are clearly audible and cause different perceptional effects which depend on the listener position. This holds especially for large WFS systems.

This paper investigates the perceptual properties described in the aforementioned paper by performing a formal listening test. To account for the multidimensional nature of the perceptual attributes, the *Repertory Grid Technique* (RGT) was chosen (see section 3). With this method each participant reports her/his own set of attributes in a first pass, which are then rated in a second pass.

The tests were conducted with a "virtual" WFS system realized by dynamic binaural resynthesis and presented to the participants by means of headphones in order to create reproducible test conditions.

2. THEORY AND PRELIMINARY WORK

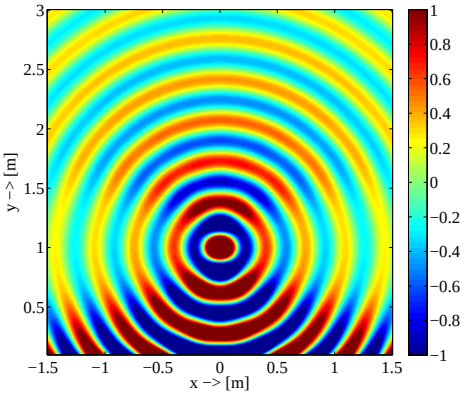
2.1. Wave Field Synthesis

Wave Field Synthesis (WFS) aims at the physical synthesis of a given desired sound field. The theory of WFS was initially derived from the *Rayleigh integrals* which require the employed secondary source distributions to be linear in the two-dimensional case or to be planar in the three-dimensional case. A reformulation of the theory based on the *Kirchhoff-Helmholtz integral* revealed that also arbitrary convex distributions can be employed with only low error [7, 8]. A detailed review of the theory of WFS can be found in the literature such as [9, 10].

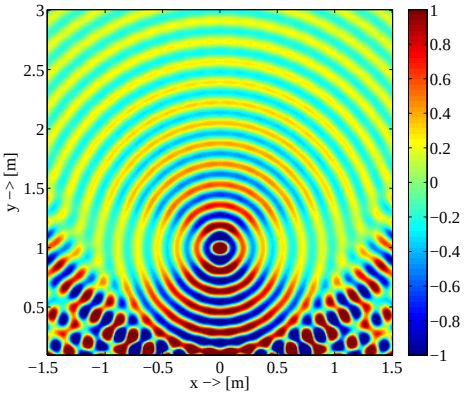
The following two sections summarize selected aspects relevant in the context of the reproduction of focused sources. The topic has been treated in detail in [6].

2.2. Focused Sources

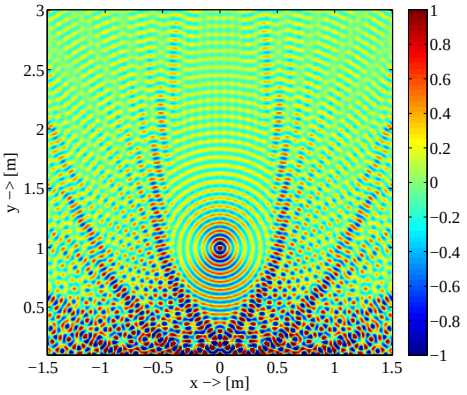
In WFS, typically simple virtual source models like plane and spherical waves are employed which are driven with given input signals like speech or the sound of a musical instrument. The spherical waves



(a) Reproduced sound field $f = 1000\,\mathrm{Hz}$



(b) Reproduced sound field $f = 2000\,\mathrm{Hz}$



(c) Reproduced sound field $f = 5000\,\mathrm{Hz}$

Fig. 1: Reproduction of a monochromatic focused source of different frequencies located at $\mathbf{x}_\mathrm{s} = (0, 1)\,\mathrm{m}$. The loudspeaker spacing is $\Delta x = 0.15\,\mathrm{m}$ which results in a spatial aliasing frequency of $f_\mathrm{al} \approx 1100\,\mathrm{Hz}$.

represent virtual monopole sources which are positioned "behind" the loudspeaker array from the listener's point of view. Focused sources on the other hand evoke the perception of a virtual source "in front of" the loudspeakers for certain listening positions.

A focused source is essentially a sound field emitted by an ensemble of loudspeakers which converges towards a focus point and diverges after having passed this focus point [6]. The diverging part then resembles the sound field of a monopole sound source located at the focus point. The converging part of the sound field, i.e. that part of the reproduced sound field between the loudspeakers and the focus point, is perceptually meaningless and it should therefore be avoided to expose listeners to this area. The fact that the potential listening area is restricted compared to the reproduction of non-focused sources is one of the essential properties of focused sources. The position of the focus point is referred to as position of the focused source.

Figure 1(a) depicts a cross-section through the horizontal plane of the sound field of a monochromatic focused source located at $\mathbf{x}_\mathrm{s} = (0, 1)\,\mathrm{m}$. The loudspeakers are positioned along the x-axis. The sound field converges for $0 < y < 1\,\mathrm{m}$ towards the position of the focused source and diverges for $y > 1\,\mathrm{m}$ which defines the useful listening area.

2.3. Artifacts in the Reproduction of Focused Sources

Theoretically, when an infinitely long continuous distribution of secondary sources (i.e. loudspeakers) is used, no of the sound reproduction theory restrictions are to be expected. However, such a continuous distribution cannot be implemented in practice because a finite number of loudspeakers has to be used. This circumstance is referred to as *spatial discretization* and *spatial truncation* of the secondary source distribution.

It has been shown in detail in [6] that the spatial discretization leads to so-called pre-echos which arrive from different directions. This means that unlike the reproduction of non-focused sources, artifacts due to discretization precede the desired signal even within the potential listening area. This circumstance is essential in terms of perception since these pre-echos

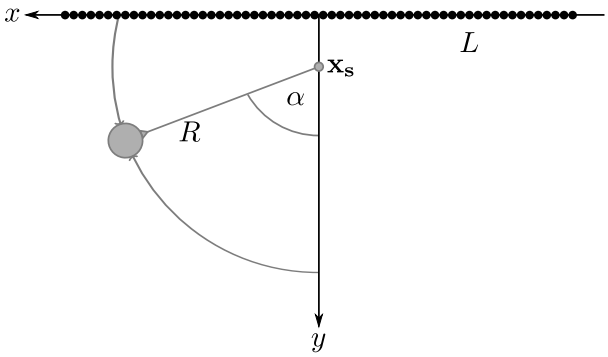


Fig. 2: Geometry of the *virtual* WFS system used in the experiment. The loudspeaker array is located on the x-axis with its center at $(0, 0)$ m. Two array lengths of $L = 4\,\mathrm{m}$ and $L = 10\,\mathrm{m}$ were used. $\mathbf{x_s} = (0, 1)\,\mathrm{m}$ denotes the position of the focused source. The position of the listener is given by the radius $R = 1\,\mathrm{m}$ for the short array and $R = 4\,\mathrm{m}$ for the long array and the angle $\alpha \in [0°, 30°, 60°]$. The head orientation of the listener is always in direction of the focused source.

might trigger the *precedence effect*, which is a fundamental mechanism in spatial hearing [11, 12]. The precedence effect describes the phenomenon that the direction of a perceived sound is not altered by echos of this sound which may arrive from different directions in a time window of 1–40 ms after the leading wave front. In the case of focused sources the possibility hence exists, that the perceived direction of the focused source is determined by the direction of the first pre-echo. On the other hand, the precedence effect only occurs if the relative level of the repetition occurring after the leading wave front is not higher than 10–15 dB. So if the amplitude of the wave front from the focused source is much higher than the amplitudes of the pre-echos, the focused source will be perceived from the intended location. Furthermore this can lead to the perception of a second source, if enough pre-echos arrive from another dominant direction than the wave front of the focused source.

Figure 3 schematically illustrates the appearance of the pre-echos for three different listening positions for the 4 m and the 10 m loudspeaker arrays used in the experiment. Refer to figure 2 and section 3.2 for a description of the geometrical setup. The black arrows in figure 3 denote the direction of incidence

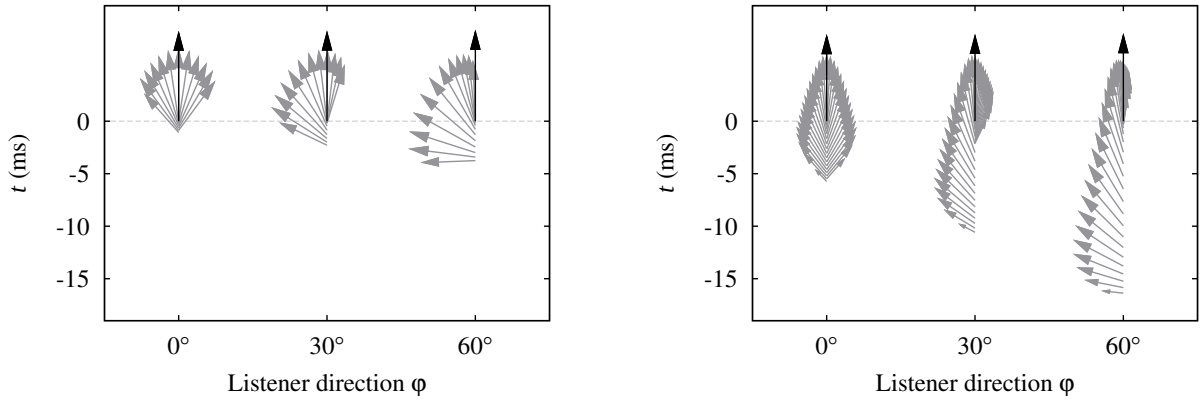


Fig. 3: Direction, amplitude (dB) and time of appearance of the echos for the 4 m (left) and the 10 m (right) loudspeaker array. The amplitude is proportional to the length of the arrows. The position of the arrow-tails on the vertical axis shows the time of arrival at the listener.

of the desired sound field, i.e. the focused source, the gray arrows denote the direction of incidence of the pre-echos. The position of the root of an arrow with respect to the vertical axis represents the arrival time. Thus, for the 60° listener position in figure 3, the pre-echos arrive significantly earlier in time than e.g. for the 0° position, especially when the 10 m array is used.

Note that pre-echos do only occur above the *spatial aliasing frequency* of a given loudspeaker system [6]. Typically, this spatial aliasing frequency is between 1 kHz and 2 kHz with practical systems. This circumstance is illustrated in figures 1(b) and 1(c). It can be seen that increasing the frequency of the focused source results in substantial artifacts, and the artifact-free region around the focused source becomes smaller with increasing frequency. This property is in contrast to the reproduction of non-focused sources, where such an artifact-free region does typically not evolve above the spatial aliasing frequency. The pre-echos cannot be identified in figures 1(b) and 1(c) due to the fact that these figures depict a steady-state scenario. Another fundamental aspect is the fact that the time at which the first echo precedes the desired signal is proportional to the length of the loudspeaker array [6], so that shorter arrays are less likely to produce strong artifacts (compare the left and right images in figure 3).

2.4. Pre-equalization

The reproduction of simple source models like plane and spherical waves as well as focused sources can be implemented as a delaying and weighting of the input signal and a filtering operation. This filtering operation is termed *pre-equalization* or *pre-filtering* [13]. In order to minimize systematic coloration, the pre-filter is only applied below the spatial aliasing frequency. For non-focused virtual sources, the spatial aliasing frequency varies only little with the receiver position [14] so that it is straightforward to determine the operating frequency range of the pre-filter.
As discussed in section 2.3, the spatial aliasing frequency of focused sources strongly depends on the position of the listener. This requires that pre-equalization is optimized for selected listening positions. The results at other listening positions can not be controlled. In order to avoid a systematic coloration in the presented experiment due to an improper pre-filter, the pre-filter was optimized for each simulated listening position separately.
The calculation of the loudspeaker driving signals is typically implemented as a *driving function* which can be represented by an impulse response with which the input signal is convolved.

2.5. Preliminary Experiment

In [6], the results of an informal listening test based on dynamic binaural re-synthesis of a virtual linear loudspeaker array have been presented. These results are to be confirmed and extended in the ex-

periment presented in this paper.

Informal reporting by the test subjects revealed that the artifacts discussed in [6] were perceived as a high-pass filtered and sometimes distorted version of the original source signal. Audible distortions were comb filtering, smearing of transients, even chirping and whistling sounds. The artifacts were in many cases perceived as arriving from other directions than the desired focused source, thus by the majority of test participants perceived as a separate source, and by a few participants as a contribution to the acoustic room impression.

3. METHOD

The challenge about an experiment regarding focused sources in WFS is that the various artifacts in the reproduced wave field (mainly caused by spatial sampling, but also truncation and amplitude errors) address several different, multidimensional attributes in the perceptional domain such as coloration, smearing of transients or split images. For unexperienced listeners, the perceived artifacts are often hard to describe, partly because pre-echos do not occur in natural situations. Additionally, the errors vary heavily depending on the position of the focused source and the listener and on the chosen source signal emitted from the focused source. The listener positions used in the experiment are shown in figure 2 and explained in section 3.2. To account for different source signals, all tests were done with both a speech recording and a recording of castanets as source input. Audio examples are available at `http://audio.qu.tu-berlin.de/?p=128`.

To account for the multidimensional nature of the perceptual attributes, the *Repertory Grid Technique* (RGT) was chosen. It was developed in the 1950s by Kelly in the context of personal construct psychology [15] and introduced to the field of spatial audio perception by Berg and Rumsey [16]. With this method each participant creates her/his own set of attributes and uses them subsequently for rating a set of stimuli. No attributes are provided by the experimenter, and thus the test subject has complete freedom in the choice of attributes.

The RGT procedure consists of two parts, the *elicitation phase* and the *rating phase*. In the elicitation phase, groups of three stimuli (*triads*) are presented to the test subject. For each triad, the subject has to decide which two of the three stimuli are more similar, and she/he has to describe the property which makes them similar, and in which characteristic they are different from the third stimulus (which should be the opposite of the first property). In the rating phase, the subject rates all stimuli on scales defined by her/his own attributes.

3.1. 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 (3 female, 9 male; between 21 and 33 years old). The participants had at least 5 years (and up to 20 years) of musical education and all of them had experience with listening tests.

Each of the subjects participated in two sessions of the experiment, which were essentially the same except for different source material (speech, castanets) used in the stimuli.

Performing an audiometry on the test subjects was not necessary, because all *Tonmeister* students have to pass an audiometric test before they even are admitted to the entrance examination. It can be assumed that all participants have a very good hearing ability. The participants were financially compensated for their effort.

3.2. Apparatus

The tests were conducted with a "virtual" WFS system realized by dynamic binaural re-synthesis and presented to the test subjects by means of headphones. See figure 2 for a sketch of the geometry of the virtual WFS arrays. Two linear loudspeaker arrays with a length of $L = 4\,\mathrm{m}$ and $L = 10\,\mathrm{m}$, respectively, and a spacing of $0.15\,\mathrm{m}$ were used/synthesized. The transfer functions of the individual virtual loudspeakers were obtained by interpolating a database of anechoic *Head-Related Transfer Functions* (HRTFs) of the FABIAN mannequin [17] to the required directions and applying further weighting and delaying in order to account for the virtual loudspeakers' distances.

For both arrays, three different listener positions on a given radius around the focused source were used. The radius was $R_{4\mathrm{m}} = 1\,\mathrm{m}$ for the short array and

$R_{10\mathrm{m}} = 4\,\mathrm{m}$ for the long array. Three different listener angles of $\alpha = 0°$, $30°$ and $60°$ were applied for both array lengths. These six conditions shall henceforth be called $0°_{4\mathrm{m}}$, $30°_{4\mathrm{m}}$, $60°_{4\mathrm{m}}$, $0°_{10\mathrm{m}}$, $30°_{10\mathrm{m}}$ and $60°_{10\mathrm{m}}$. The initial head orientation was always pointing towards the focused source, as shown in figure 2. As seventh condition, a reference stimulus ("ref.") was created, which consisted of a single sound source straight in front of the listener. This was realized by directly using the corresponding HRTFs from the database. In all conditions, the focused source was located directly in front of the listener.

As discussed in section 2.4, the WFS pre-filter was optimized separately for each simulated listening position. Systematic coloration by an improper choice of pre-filter was not part of the investigation and should be avoided.

For each head orientation, the driving function of each virtual loudspeaker (refer to section 2.4) was convolved with that pair of HRTFs representing the given combination of loudspeaker and head orientation and the result was added for all loudspeakers. Each stimulus was thus represented by a pair of impulse responses (left and right ear) which in turn represent the spatio-temporal transfer function of the loudspeaker system driven with the given configuration to the ears of the mannequin for a given head orientation [18]. This type of spatio-temporal transfer function is then typically referred to as *Binaural Room Transfer Function* or *Binaural Room Impulse Response* (BRIR) when represented in time domain. The BRIRs were calculated for all possible head orientations. The headphone signal was then obtained by convolving a given input signal with the BRIRs representing the entire loudspeaker system as described above.

In order avoid biases in the subjects' responses due to different levels of the different stimuli, all BRIRs were normalized in amplitude based on the frontal direction.

As mentioned before, two different input signals were used – speech and castanets. The speech signal was chosen because it contains both periodic and aperiodic components and it is a very common and familiar type of signal. The castanets sample was chosen because it contains very strong transients which emphasize potential pre-echo artifacts.

The flow of the test phases and the interaction with the test subjects was controlled with a graphical user interface (GUI) programmed in *Python* using *GTK+* and the *Glade* user interface designer. Two screenshots can be seen in figures 4 and 5.

The real-time convolution was performed using the *SoundScape Renderer* (SSR) [19, 20], an open-source software environment for spatial sound reproduction, running in *binaural room scanning* (BRS) mode. The BRIR sets of all stimuli were loaded into memory and an input port was created for each BRIR set. The GUI indicated the current audio file and test condition as text messages sent via a TCP/IP socket to Pure Data [21], which in turn routed the audio signal to the corresponding input of the SSR. Due to the internal processing in the SSR the switching between between different audio inputs leads to a smooth cross-fade with raised-cosine shaped ramps. The SSR convolved the input signal in realtime with that pair of impulse responses corresponding to the instantaneous head orientation of the test subject as indicated by a Polhemus Fastrack tracking system. AKG K601 headphones were used with a compensation of the transfer function applied [22].

3.3. Procedure

The participants got written instructions explaining their tasks in the two phases of the experiment. In order not to influence the choice of attributes, examples were given from a different domain. The elicitation procedure was explained by means of three images of animals and similar and differentiating attributes were given as example for these images. Questions could be addressed to the experimenter at any time.

The main screen of the elicitation GUI is shown in figure 4. With the *A,B,C* buttons, the three conditions of the current triad can be selected. Once a button is pressed, the audio sample is played and looped and further key-presses only switch between conditions. The duration of the sound files is 8 and 7 seconds for speech and castanets, respectively. The participants could listen to each stimulus as long as they wanted to. At any time, the *Pause* button could be used to stop playback. Once all three stimuli have been listened to, the test subject had to choose which two of the stimuli are similar and therefore different from the third. If there were competing

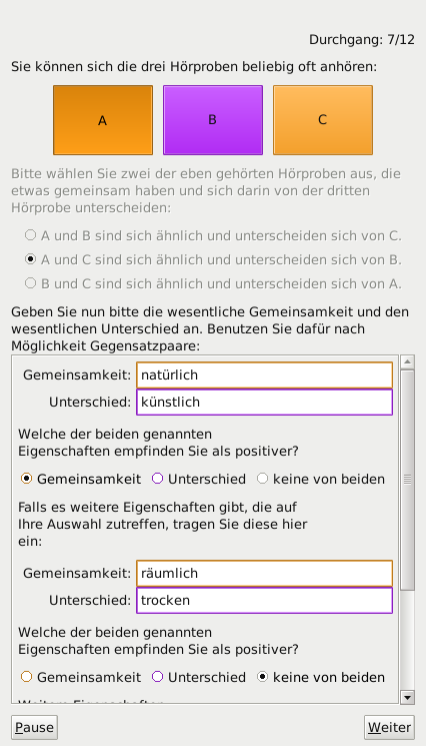


Fig. 4: GUI screenshot of the *elicitation* phase. The conditions can be switched with the *A, B, C* buttons at the top. Below, the similar pair can be selected. In the lower part, up to three attribute pairs can be specified.

aspects, only the strongest one should be taken into account. After selecting the two similar stimuli, a bipolar attribute pair had to be given by specifying the attribute that makes the two similar stimuli similar, and the opposite attribute which sets them apart from the third stimulus. One attribute pair per triad had to be specified and two more could be given optionally if the test subject perceived several different properties.

As soon as the first attribute was typed in, the selection of the two similar stimuli was disabled (see figure 4). It was still possible to listen to and switch between the stimuli.

After a short training phase using a different input signal, 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

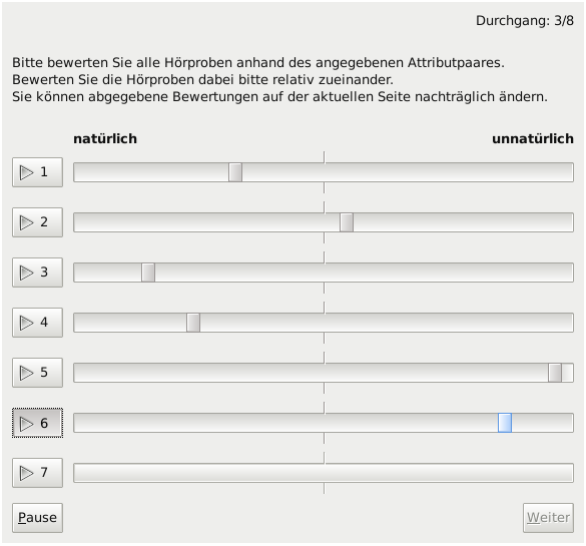


Fig. 5: GUI screenshot of the *rating* phase. At the top left and right, one of the previously elicited bipolar attribute pairs is displayed. Below, there are seven sliders (one per condition) and for each slider there is a button to switch to the corresponding condition.

five conditions ref., $30°_{4m}$, $60°_{4m}$, $30°_{10m}$ and $60°_{10m}$. The other two additional triads were ref., $0°_{4m}$, $0°_{10m}$ and $0°_{4m}$, $30°_{4m}$, $0°_{10m}$. These two have been chosen to have the two additional conditions in triads with very similar conditions in order 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 had 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*.

After the elicitation phase, the participants took a break and in the meantime repeatedly given attribute pairs were removed for the list used in the next phase.

The GUI of the rating phase is shown in figure 5. At each screen one previously elicited attribute pair is displayed on the top left and right. Below, the seven stimuli can be played back and they have to be rated on the corresponding sliders. Test subjects

ref.	-						
$0°_{4m}$	0	-					
$30°_{4m}$	1	1	-				
$60°_{4m}$	3	0	2	-			
$0°_{10m}$	1	2	0	0	-		
$30°_{10m}$	2	0	0	2	0	-	
$60°_{10m}$	4	0	2	1	0	3	-
	ref.	0°	30°	60°	0°	30°	60°
speech		4 m			10 m		

ref.	-						
$0°_{4m}$	0	-					
$30°_{4m}$	1	1	-				
$60°_{4m}$	2	0	2	-			
$0°_{10m}$	1	1	1	0	-		
$30°_{10m}$	1	0	2	3	0	-	
$60°_{10m}$	3	0	3	3	0	0	-
	ref.	0°	30°	60°	0°	30°	60°
castanets		4 m			10 m		

Table 1: Frequency of dissimilarity pairs for speech (left) and castanets (right) for one test subject. Conditions $0°_{4m}$ and $0°_{10m}$ were not used for the MDS.

could give a continuous rating which was internally saved in a range from −1.0 to 1.0. Once all stimuli received a rating, the test subject could switch to the next screen. The number of repetitions depended on the number of attribute pairs elicited in the first phase (duplicates removed). Before the actual test, another 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 other input sound. Half of the subjects were confronted with speech samples in the first session and castanets samples in the second session and vice versa for the other half.

The elicitation phase took between 16 and 45 minutes in the first session, the rating phase took between 7 and 25 minutes. In the second session, the subjects needed from 11 to 44 minutes and from 5 to 26 minutes for the elicitation and rating phases, respectively.

4. RESULTS/DISCUSSION

There is a multitude of possibilities for analyzing and interpreting the data generated by the experiment; first results are presented in the following subsections.

4.1. Attributes/Constructs

One of the main results of the experiment are the elicited attribute pairs. They reflect the range of perceptual similarities and differences among the conditions. Some of the most prominent properties were coloration (e. g. *original* vs. *filtered*, *balanced* vs. *unbalanced frequency response*), localization (e. g. *center* vs. *off-center*, *close* vs. *far*), artifacts (e. g. *clean sound* vs. *chirpy, squeaky, phasey sound*) and reverberation (e. g. *dry* vs. *reverberant*), just to name a few. Many attributes describe the distorted sound and artifacts which was also experienced in preliminary tests [6] (section 2.5). All elicited attributes were originally generated in German and were translated to English for this paper.

The elicitation phase yielded 12 to 33 attribute pairs per participant for the speech stimuli and 13 to 34 pairs for the castanets stimuli. After removing duplicates, the rating phase was done with 6 to 16 attribute pairs per participant for the speech stimuli and 8 to 17 pairs for the castanets stimuli.

4.2. (Dis-)Similarity Ratings

In the elicitation phase – as described in section 3.3 – the test subjects had to select two stimuli of a triad which sounded more similar than the third stimulus. Based on these decisions, the frequency of dissimilar pairs was determined. The quantity of dissimilar pairs can be seen for one test subject in table 1.

A multidimensional scaling (MDS) was applied using these dissimilarity responses. The MDS was calculated for only five of the seven conditions, because $0°_{4m}$ and $0°_{4m}$ had not a complete set of dissimilarity ratings because they appeared in only 2 of the 12 triads (see section 3.3). The results for a single participant and for all participants are shown in figure 6.

Already for a two-dimensional solution, the MDS shows reasonable stress values, therefore it can be

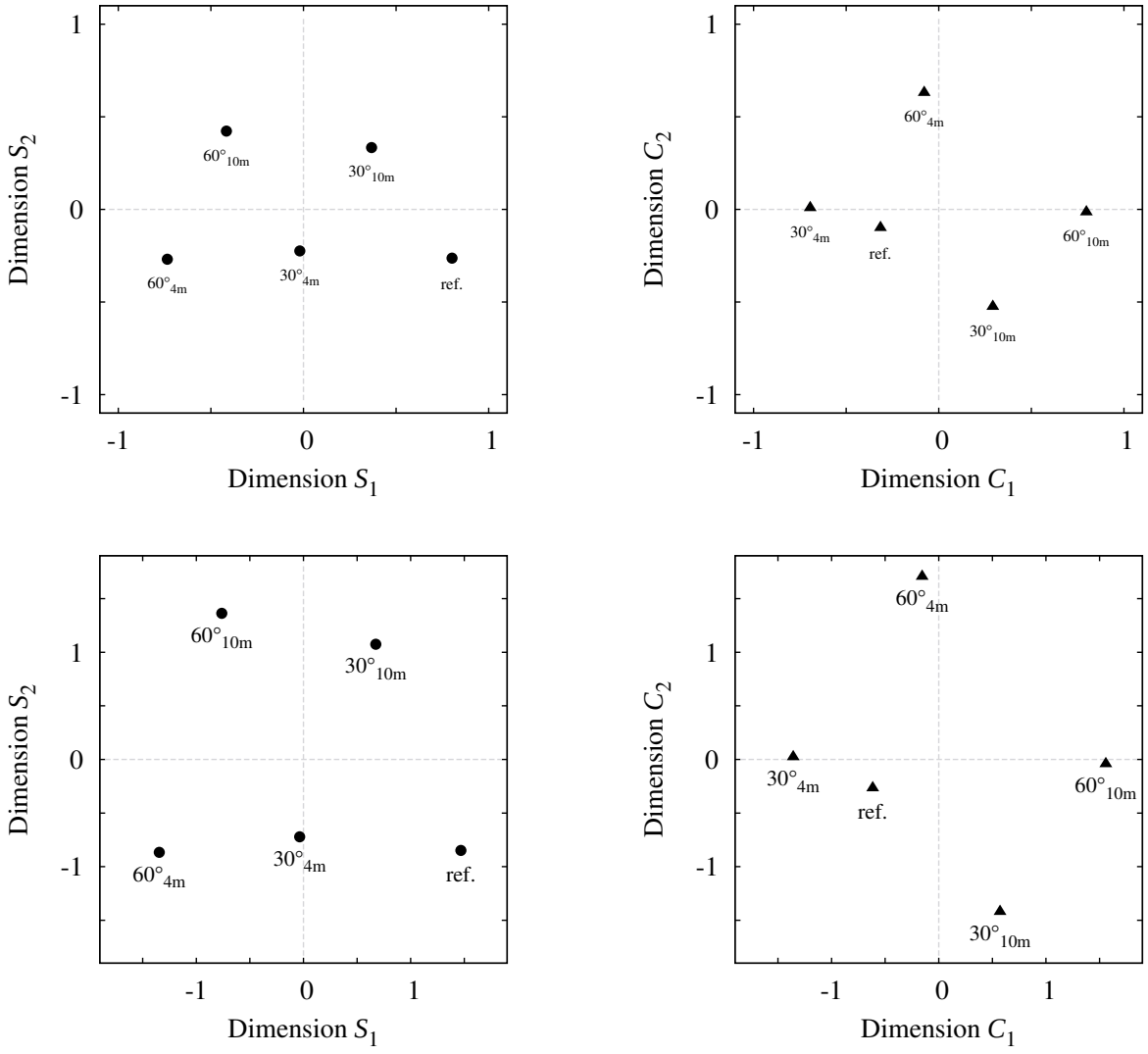


Fig. 6: Results of an MDS (PROXSCAL) with weighted euclidian distances for speech (left) and castanets (right) for a single participant (top row) and all participants (bottom row). Conditions $0°_{4m}$ and $0°_{10m}$ were not used for the MDS.

	mean	std. deviation
ref.	0.748	0.435
$0°_{4m}$	0.662	0.462
$0°_{10m}$	0.217	0.492
$30°_{4m}$	-0.001	0.506
$30°_{10m}$	-0.028	0.561
$60°_{4m}$	-0.225	0.580
$60°_{10m}$	-0.493	0.631

	mean	std. deviation
ref.	0.630	0.460
$0°_{4m}$	0.624	0.498
$30°_{4m}$	0.145	0.615
$0°_{10m}$	0.015	0.621
$60°_{4m}$	-0.074	0.754
$30°_{10m}$	-0.333	0.584
$60°_{10m}$	-0.472	0.711

Table 2: Ranking over all test subjects based on the positive/negative poles of attribute pairs in combination with the respective ratings. Ratings for speech in the left table, ratings for castanets in the right table. Neutral attribute pairs were not taken into account. Minimum value: -1.0, maximum value: 1.0. Number of attribute pairs: 111 (speech), 112 (castanets).

assumed that all of the participants show a very homogeneous perception. The value of *stress-1* was 0.105 for the speech conditions and 0.098 for the castanets conditions. The similarity between the result for a single test subject and for all subjects can be seen in figure 6.

For the speech input, the resulting dimension 1 can be interpreted as *left* vs. *center*. The reference, positioned on the very right, is perceived correctly from in front of the listener. Further to the left, the stimuli are perceived from the left side. Dimension 2 could represent the amount of artifacts and comb filtering. The reference and the stimuli of the 4 m array exhibit no artifacts, the stimuli of the 10 m array contain clearly audible artifacts.

The interpretation is harder in the case of castanets. The meaning of dimension 1 may be *clean signal* vs. *artifacts/chirping noise*. It is not clear why the reference is between $30°_{4m}$ and $60°_{10m}$. Dimension 2 possibly means *dry* vs. *reverberant*, but this is not quite obvious.

4.3. Positive/Negative Poles

In the elicitation phase, each attribute had to be specified if it is positive, negative or neutral compared to the second attribute. This could be selected in the GUI shown in figure 4. Of the 281 rated constructs (134 for speech, 147 for castanets; including many similar ones), 224 (111 for speech, 112 for castanets) were specified by the participants as having a positive and a negative pole. The ratings of all attributes with positive or negative connotation were averaged, resulting in the rankings shown in table 2.

As expected, the reference leads the ranking, followed by the positions at the center of the array and then the positions with larger angles. In case of the castanets stimuli, the condition $30°_{4m}$ has a higher ranking than $0°_{10m}$ which suggests that the castanets stimulus is more sensitive to the on-axis sound impairments. The values are calculated over all participants regardless of the dimension of the ratings. Therefore, the standard deviation is rather high.

4.4. Grids

A grid contains the attribute pairs chosen by the subject and the ratings for all conditions belonging to those pairs. Table 3 shows the grid for one subject for speech and castanets. The interval scale used on the rating sliders has been transformed in a 7-point ordinal scale with values from 1 to 7 by dividing the sliders' range from -1.0 to 1.0 into 7 equal intervals.

It can be seen that the reference is mostly rated at the positive side of the attribute pair, but it is not the best condition for all attribute pairs. The condition $60°_{4m}$ is rated as the most off-centered condition, whereas the $60°_{10m}$ is the second one. This can be explained by the results from figure 3. There the condition $60°_{4m}$ has a strong first wave front from the left and only 4 ms before the desired wave front from the focused source arrives. The condition $60°_{10m}$ has a longer time between its weaker first echo arriving from the left at -16 ms and the desired focused

speech:	ref.	$0°_{4m}$	$30°_{4m}$	$60°_{4m}$	$0°_{10m}$	$30°_{10m}$	$60°_{10m}$	
off-center	7	7	5	1	7	6	3	center
phasey	7	7	6	1	5	4	2	non-phasey
few artifacts	3	1	2	4	3	5	7	many artifacts
unnatural	7	7	4	3	3	2	1	natural
little coloration	1	1	2	3	2	5	6	much coloration
little comb filter	1	1	2	4	3	5	7	much comb filter

castanets:	ref.	$0°_{4m}$	$30°_{4m}$	$60°_{4m}$	$0°_{10m}$	$30°_{10m}$	$60°_{10m}$	
center	1	1	2	7	1	1	4	off-center
little coloration	3	3	2	1	4	6	7	much coloration
unnatural	4	6	5	7	3	1	1	natural
non-phasey	3	1	7	1	2	5	7	phasey
front	2	2	5	6	1	4	5	back
reverberant	5	6	6	7	3	3	3	dry
dark	3	3	5	2	7	6	5	bright
non-localizable	6	6	1	7	5	3	2	localizable

Table 3: Grid of the rating results of one subject. The ratings were converted to a 7-point scale. A value of "1" means the corresponding condition was rated as the corresponding attribute on the left, a value of "7" means the other extreme on the right side.

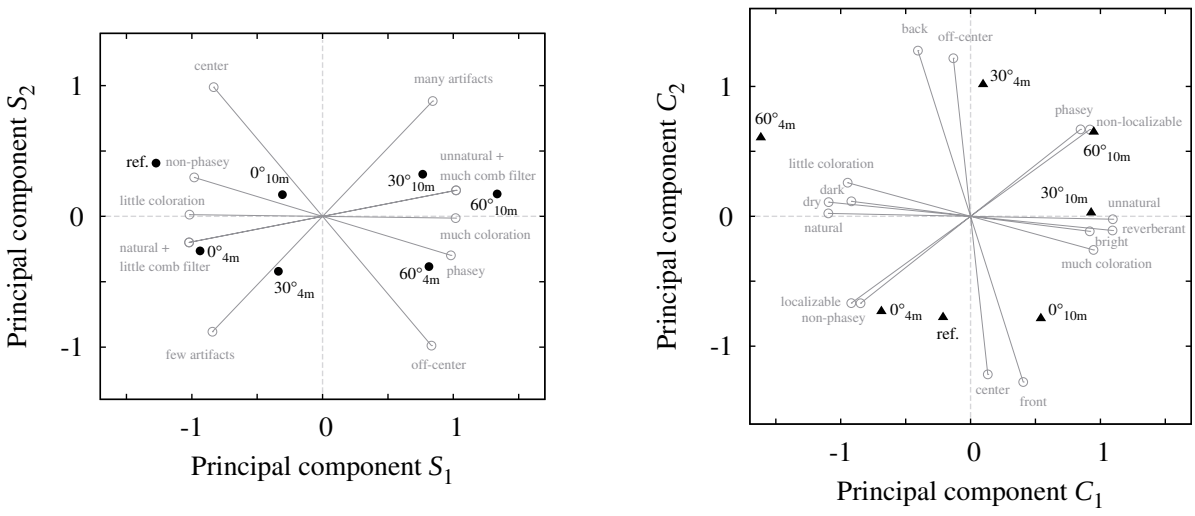


Fig. 7: Results of a PCA for one subject – speech (left), castanets (right)

source wavefront, which may lead to the perception of two sources, one in the center and one on the left.

The attributes *few artifacts* vs. *many artifacts*, *unnatural* vs. *natural*, *little coloration* vs. *much coloration* and *little comb filter* vs. *much comb filter*, show that the conditions from the 4 m array have fewer artifacts than the conditions from the 10 m array. The selected test subject (see table 3) has rated the condition $0°_{4\mathrm{m}}$ as more natural than the reference condition for the castanets stimuli. This may be due to the fact that the pre-echos create a room impression that may be perceived as more natural than the dry reference stimulus.

Using the grids, a *Principal Component Analysis* (PCA) has been calculated for single subjects. The result for one subject is presented in figure 7. The principal component C_1 for the castanets correlates with the dimension *coloration* and C_2 with the dimension *position*. For the speech stimuli one principal component also correlates with *coloration* (S_1), but the second one is harder to interpret, because it is a superposition of *position* and *artifacts*.

5. CONCLUSIONS/FURTHER WORK

This paper presented a formal evaluation of the perceptual attributes of focused sources in WFS. The stimuli were presented using a binaural resynthesis of a linear WFS system. this procedure allows seamless switching between different listener positions and constitutes the only possibility to assure reproducible conditions. It was shown that focused sources exhibit various artifacts that can be accounted to spatial sampling, spatial truncation and pre-equalization. Most prominent is the occurrence of pre-echos before the desired wave front emerging from the focused source position. Complex pre-echos do not occur in natural situations and due to their unfamiliarity unexperienced listeners have certain difficulties in describing them. Therefore, the *Repertory Grid Technique* has been used where the test subjects create individual attribute pairs which are used afterwards for a rating of stimuli. The results revealed a number of interesting perceptual effects such as strong coloration or even chirpy sounds, incorrect localization and the perception of one or more (distorted) copies of the virtual sound source.

Although the results give no direct indication on the quality degradation caused by the artifacts, the authors would like to emphasize that the artifacts are clearly audible for certain listening positions. We have published some of the stimuli at `http://audio.qu.tu-berlin.de/?p=128` in order to provide the reader the possibility to collect her/his own impressions. Additional coloration of the source signal has to be expected in practice, since the pre-equalization filter was optimized for the investigated listener positions.

The auralization of focused sources using (large) WFS systems will be subject to severe artifacts. In the current stage, focused sources should be used with care and as an effect rather than for high-quality auralization of single sources. The presented analysis could provide the basis to design improved driving functions for focused sources.

Further work includes a more detailed analysis of the results, e.g. by performing a cluster analysis of the elicited attribute pairs of single subjects and over all subjects. Furthermore, we aim at deriving insights into the perception of complex pre-echos.

6. REFERENCES

[1] D. de Vries. *Wave Field Synthesis*. AES Monograph. AES, New York, 2009.

[2] J. Ahrens and S. Spors. Sound field reproduction using planar and linear arrays of loudspeakers. *IEEE Trans. on Sp. and Audio Proc.*, 2010. In press.

[3] E. N. G. Verheijen. *Sound Reproduction by Wave Field Synthesis*. Ph.D. thesis, Delft University of Technology, 1997.

[4] J. Ahrens and S. Spors. Spatial encoding and decoding of focused virtual sound sources. In *Ambisonics Symposium*. Graz, Austria, June 2009.

[5] S. Spors and J. Ahrens. Reproduction of focused sources by the spectral division method. In *IEEE Int. Symp. on Comm., Control, and Sig. Proc. (ISCCSP)*. Limassol, Cyprus, March 2010.

[6] S. Spors, H. Wierstorf, M. Geier and J. Ahrens. Physical and perceptual properties of focused sources in Wave Field Synthesis. In *127th AES Convention*. October 2009.

[7] E. W. Start. Application of curved arrays in Wave Field Synthesis. In *100th AES Convention*. May 1996.

[8] J. Ahrens and S. Spors. On the secondary source type mismatch in Wave Field Synthesis employing circular distributions of loudspeakers. In *127th AES Convention*. October 2009.

[9] A. Berkhout, D. de Vries and P. Vogel. Acoustic control by Wave Field Synthesis. *JASA*, 93(5):2764–2778, May 1993.

[10] S. Spors, R. Rabenstein and J. Ahrens. The theory of Wave Field Synthesis revisited. In *124th AES Convention*. May 2008.

[11] H. Wallach, E. B. Newman and M. R. Rosenzweig. The precedence effect in sound localization. *American Journal of Psychology*, 57:315–336, 1949.

[12] W. Haas. The influence of a single echo on the audibility of speech. *Acustica*, 1:49–58, 1951.

[13] S. Spors and J. Ahrens. Analysis and improvement of pre-equalization in 2.5-dimensional Wave Field Synthesis. In *128th AES Convention*. May 2010.

[14] S. Spors and J. Ahrens. A comparison of Wave Field Synthesis and Higher-Order Ambisonics with respect to physical properties and spatial sampling. In *125th AES Convention*. October 2008.

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

[16] J. Berg and F. Rumsey. Spatial attribute identification and scaling by Repertory Grid Technique and other methods. In *16th AES Conference*. March 1999.

[17] A. Lindau and S. Weinzierl. FABIAN – An instrument for the software-based measurement of binaural room impulse responses in multiple degrees of freedom. In *24. Tonmeistertagung (VDT International Convention)*. November 2006.

[18] M. Geier, J. Ahrens and S. Spors. Binaural monitoring of massive multichannel sound reproduction systems using model-based rendering. In *NAG/DAGA International Conference on Acoustics*. March 2009.

[19] The SoundScape Renderer. `http://tu-berlin.de/?id=ssr`.

[20] M. Geier, J. Ahrens and S. Spors. The SoundScape Renderer: A unified spatial audio reproduction framework for arbitrary rendering methods. In *124th AES Convention*. May 2008.

[21] M. S. Puckette et al. Pure Data. `http://puredata.info`.

[22] Z. Schärer and A. Lindau. Evaluation of equalization methods for binaural signals. In *126th AES Convention*. May 2009.