AUDITORY PERCEPTION OF SPATIAL EXTENT IN THE HORIZONTAL AND VERTICAL PLANE

Marian Weger, Georgios Marentakis, Robert Höldrich *

Institute of Electronic Music and Acoustics
University of Music and Performing Arts, Graz, Austria
{weger,marentakis,hoeldrich}@iem.at

ABSTRACT

This article investigates the accuracy with which listeners can identify the spatial extent of distributed sound sources. Either the complementary frequency bands comprising a source signal or the individual grains of a granular synthesis-based stimulus were distributed directly on discrete loudspeakers. Loudspeakers were arranged either on the horizontal or the vertical axis. The algorithms were applied on white noise, an impulse train, and a rain drops stimulus. Absolute judgments of spatial extent were obtained separately for each orientation, algorithm, and stimulus using three different magnitudes of horizontal or vertical extent.

Horizontal spatial extent judgments varied systematically with physical extent for all conditions in the experiment. The correspondence between perceived and actual vertical extent was poor. The time-based synthesis algorithm resulted in significantly larger judgments of spatial extent irrespective of orientation and stimulus compared to the frequency-based algorithm.

1. INTRODUCTION

Perceived spatial extent is a measure of the perceived spatial volume that may be occupied by an auditory event. The term was proposed by [1] and may refer independently to width, height, and potentially also depth. It is important to note here that although most often in the literature the term Auditory or Apparent Source Width (ASW) has been used to refer to the perceived horizontal spatial extent of a sound, we use the term spatial extent here because it allows us to differentiate between spatial extent perception along each of the three axes of the Cartesian coordinate system. Furthermore, in the following, by spatial extent we refer exclusively to the spatial extent of the perceived auditory object and not of the sound producing object.

This study is motivated by the revived interest in algorithms for the representation of auditory spatial extent in the last years. This interest is justifiable if one considers that reliable representation of auditory spatial extent could be useful for both scientific and artistic purposes. Concerning music for example, reliable representations of spatial extent could provide an extra design parameter for composers, sound engineers, and music producers. Concerning interactive systems, such algorithms could improve and augment auditory representations in virtual and mixed reality systems. Importantly, successful representation of horizontal and vertical extent may pave the way for representing more complex shapes with sound, which would be vital for assistive technologies for example. In this article, we focus on the synthesis of relatively small spatial extents, keeping an eye on applications for which the space to deploy loudspeakers is limited. We proceed by first reviewing the literature and then presenting the experiment and their results.

Perceived spatial extent is influenced by both spatial and non-spatial acoustical features. Non-spatial features that affect the perception of spatial extent include loudness, duration, and base frequency of a sound. Increased sound pressure level and duration and lower base frequency are generally associated with larger spatial extent of sound generating sources [2, 3, 4]. Furthermore, decisions about the shape and size of sounding objects can be reached on the basis of spectral cues such as the (sometimes direct) relationship between the modal frequencies of vibrating objects and their geometric shapes and size. In experiments, above chance identification of auditory source shape solely based on spectral cues has been observed [5, 6, 7].

Concerning spatial factors, studies have focused on the perception of horizontal spatial extent (or ASW). This increases in reverse proportion to the interaural cross-correlation coefficient (IACC) [8, 2, 9].

A significant number of studies investigated the horizontal spatial extent of the auditory event that emeges when simultaneous uncorrelated noise sources are distributed directly to individual loudspeakers. Linear or circular loudspeaker arrangements were tested [10, 11, 12, 13]. It was shown that the perceived horizontal extent of such stimuli varies in proportion to the actual spatial extent occupied by the noise sources. The perceived horizontal spatial extent is, however, narrower than the actual spatial extent [12, 11]. Increasing the noise bandwidth or center frequency [11, 14], or the signal duration [13] results in a wider perceived spatial extent. Furthermore, small gaps in the loudspeaker spatial distribution are not easily noticed while the size of large gaps is often exaggerated [12].

In practice, decorrelation techniques for arbitrary monophonic signals are used to recreate a similar effect. A very promising approach works by splitting an auditory signal into a number of unique frequency bands which are then spatialized directly on loudspeakers or as virtual sources [10, 15, 16, 17]. Most often, signals are decomposed in bands whose bandwidth and center frequency correspond to the Equivalent Rectangular Bandwidth (ERB) scale [18, 10]. The way frequency bands are mapped to spatial positions is important as it influences both the center and the spatial extent of the perceived auditory event [10, 19]. Convincing synthesis of horizontal spatial extent has been achieved by using a Halton sequence [20] to map frequency bands to fixed locations [16]. In evaluation studies, this method resulted in perceived horizontal

* Contributions: Marian Weger and Georgios Marentakis designed the experiment, performed the statistical analysis, and authored the article. Marian Weger implemented and executed the evaluation study. Robert Höldrich contributed essential knowledge on acoustic measurements and predictors for apparent source width and provided useful comments and corrections to the article. This work was supported by the Zukunftsfonds Steiermark Klangräume Project (PN:6067) led by Georgios Marentakis.

spatial extent proportional to the physical extent of the distributed sound source. Impressions ranging from a narrow focused source to sounds completely surrounding the listener were obtained using a circular loudspeaker array. Sound quality was however strongly signal dependent [16]; this is a common problem in decorrelation techniques [21, 16, 17].

Another approach, originating in electroacoustic music, is to create spatially extended sound sources using spatialized granular synthesis [22, 23, 24]. Granular synthesis generates sounds by combining short signals (grains) [25, 24]. It can result in a great variety of sounds, including those of everyday events, such as rain, applause, etc. As grains are in general short and may be designed to have steep attacks they can be localized well. Their potential for spatial extent synthesis is therefore high. This hypothesis has however not been tested experimentally.

The two aforementioned algorithms, which from here on will be called the frequency-based and the time-based algorithm, are in a sense complementary to each other. While in the frequency-based algorithm, the frequency bands of a monophonic input signal are spatialized independently to yield a coherent sound, in the time-based algorithm this is achieved by using individually spatialized temporal grains. In both cases, it is envisaged that the spatial extent of the auditory event will relate to the spatial distribution of the grains or frequency bands comprising the source. However, both algorithms might be sensitive to the way grains or frequencies are spatialized in addition to the size and geometry of the area within which the signal content is distributed. It is reasonable to hypothesize that the allocation of the individual grains/frequencies to a single auditory event may be infeasible above a certain spatial dispersion.

The above observations motivated us to design and implement an experiment that compares the aforementioned time- and frequency-based spatial extent synthesis algorithms on the basis of their ability to create the impression of spatially distributed sound sources. Our experiment investigates the synthesis of both horizontally and vertically extended sound sources, with extents that are smaller compared to the ones used in the literature. The aim was to understand the relevance of the aforementioned synthesis techniques for fields other than surround music production, e.g., for Human Computer Interaction (HCI) applications.

2. EXPERIMENT

In the experiment, participants performed absolute judgments of perceived spatial extent in conditions that manipulated the spatial extent synthesis algorithm, the type of stimulus used, and the orientation and length of the spatial distribution of the loudspeakers that were used to distribute the stimuli. An overview of the variables is provided in Table 1 and a photo of the experiment setting is provided in Figure 2.

With reference to Figure 2, small, medium, and large spatial distributions were simulated by distributing signals on 3, 7, or 11 adjacent loudspeakers respectively using either the frequency- or the time-based algorithm. Distribution orientation was either horizontal or vertical. As the experiment targeted also the perception of vertical spatial extent, we opted to use discrete loudspeakers instead of phantom sources. This was done specifically because panning algorithms are known to provide weak and inaccurate perception of the vertical location of elevated phantom sources [26, 27, 28, 29].

Table 1: *The independent variables in the experiment.*

Factor	*Levels*
Spatial Distribution	small medium large
Algorithm	frequency-based (FB) time-based (TB)
Stimulus	white noise impulse train rain drops
Orientation	horizontal vertical

2.1. Stimuli

Three different stimuli were used in the experiment. The first two were white noise and an impulse train. These were chosen because they represent optimal scenarios for the frequency- and time-based algorithms, respectively. The third stimulus was designed to create the impression of strong rain and represented a more realistic scenario.

To create this rain drops stimulus 48 different rain drop samples were used. These were extracted from a recording of rain and normalized to the same amplitude. Average duration was 46 ms (standard deviation SD=18 ms) with an approximate attack time[1] of 2.2 ms (SD=1.8 ms). They were combined using a typical granular synthesis algorithm that selected grains by drawing samples from a uniform distribution. The onset of the next event relative to the onset of the current one was sampled from a normal distribution with mean M=10 ms (100 Hz) and SD=3 ms. Occasional negative delays were mirrored around zero to positive ones. Randomizing delay helped to avoid the impression of a pitched sound. The impulse train stimulus was implemented similar to the rain drops stimulus, but with a Dirac impulse as a grain.

While both white noise and impulse train stimuli have a flat frequency spectrum up to half the sampling frequency, the averaged spectral energy of the rain drops stimulus was concentrated primarily in the region between 2 kHz and 7 kHz.

2.2. Algorithms

In case of the frequency-based algorithm, the condition-dependent monophonic input signal was decomposed into frequency-bands whose center frequency and bandwidth corresponded to the Equivalent Rectangular Bandwidth (ERB) scale [30], according to the algorithm proposed in [10] and [19]. 38 ERB-bands with center frequencies from 142.5 Hz to 19.7 kHz were chosen. The complementary ERB-filters were implemented as rectangular windows in the spectral domain using the short-time Fourier transform (STFT) with an FFT size of 1024 samples. Hann-window and 75% overlap were chosen to yield perfect reconstruction[31, p. 113].

To make sure that the ERB-bands are evenly distributed to all active loudspeakers, and to ensure a reproducible distribution, the output channel to which each individual ERB-band was mapped, was chosen by using a Halton sequence [20], as proposed by [16]. In particular, a long (1000 elements) Halton sequence of base 2

[1] In this context the attack time was defined as the time to reach the lowest maximum of all grain's envelopes. The envelope of a signal was computed as the absolute value of its discrete-time analytic signal (Hilbert transform).

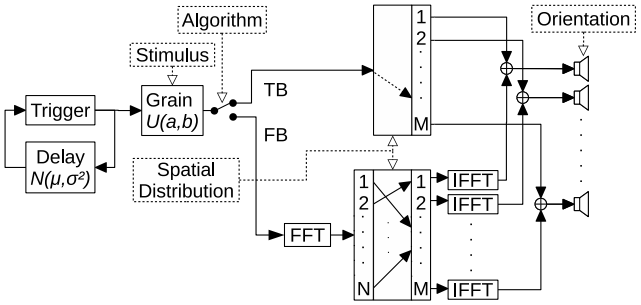

Figure 1: *Simplified block diagram of the signal chain for the granular synthesis-based stimuli (impulse train and rain drops). The independent variables of the experiment are framed in dotted lines. N: number of ERB-bands. M: number of output channels.*

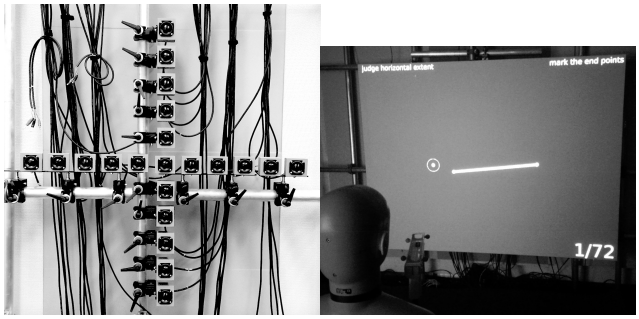

Figure 2: *The planar loudspeaker array (left) and the experiment setting (right).*

(without offset) was pre-computed. First, the resulting Halton sequence elements (that range between 0 and 1) were mapped to the 11 loudspeakers of the linear loudspeaker array (counted from left to right / top to bottom). This sequence was used to generate individual sequences for each condition. For each of those sequences, numbers corresponding to loudspeakers inactive for the corresponding condition were removed. In a second step, direct repetitions (neighboring bands in the same loudspeaker) were removed from the resulting sequence and it was made sure that no loudspeaker received more than one band than the rest to avoid concentrating energy on one loudspeaker. The first 38 sequence numbers of the remaining sequence were used to indicate the channels (from 1 to 11) in which the corresponding ERB-bands would be rendered[2]. The signal to each loudspeaker was calculated using 3, 7, or 11 complementary filters containing the relevant frequency bands. In all conditions, the simulated signal power difference between loudspeakers was less than 2 dBA.

In the case of the time-based algorithm, the individual grains (Dirac impulses or rain drop samples) were simply routed to discrete loudspeakers depending on the spatial distribution in each condition. This was done during sound synthesis which provided access to the individual grains. Each time a grain or impulse was played from a given loudspeaker location, the location of the next grain was selected again using a pre-defined sequence similar to the one in the frequency-based algorithm. However, now the sequence lengths corresponded to the number of active loudspeakers and each channel number was included once.[3]

A simplified overview of the complete signal chain for the granular synthesis-based stimuli (impulse train and rain drops) is illustrated in Figure 1.

When being processed by the frequency-based algorithm both granular stimuli (rain drops and impulse train) were synthesized as monophonic signals and a single monophonic white noise signal was used. The algorithm then did the filtering and the playback from the relevant channels in the same way for all stimuli. When being processed by the time-based algorithm, each grain (rain drop or click) of the two granular stimuli was allocated to its corresponding channel as soon as it was generated. The white noise stimulus was not processed by the time-based algorithm. Instead, each active loudspeaker depending on the condition played statistically independent white noise. This exception served as a control condition in the experiment. It was assigned to the time-based algorithm to simplify the condition names in the experiment.

2.3. Apparatus

Twenty-one custom 2-inch broadband speakers and class-D amplifiers were used [32], connected to Behringer ADA8000 DA converters running at 44.1 kHz / 24 bit. The loudspeakers were arranged on two intersecting lines of 11 loudspeakers each, which were aligned on the horizontal and vertical axis. The center speaker was shared and placed at a height of 120 cm, roughly aligned with the nose of the listeners. There was 10 cm distance between neighboring membrane centers leading to a maximum distance of 1 m between the two outmost speakers. Loudspeakers were aligned to the direction of the listener and hidden behind a 2 by 1.5 m acoustically transparent projection screen[4] installed 10 cm in front of them (see Figure 2). The three different spatial distributions led to physical extents of 25, 65, and 105 cm, or 7.2, 18.5, and 29.4 degrees.

Differences in distance to the listener were compensated by loudspeaker-specific gain and delay corrections, and to protect the loudspeakers all channels in all conditions were high-pass filtered at 200 Hz. Equivalent Continuous Sound Level (Leq) was 55 dBA at the listening position for all stimuli. The experiment took place in an acoustically treated room of 4.3(w) × 6.2(l) × 3.4(h) m size, with reverberation times between 0.15 and 0.22 s in the relevant frequency range above 200 Hz. The direct-to-reverberant ratio (DRR) [33] at the listening position was between 7 and 11 dB for the individual loudspeakers.

Participants performed the experimental task using a toy gun to indicate perceived horizontal or vertical extent. To achieve this, infrared reflective markers were mounted on the gun and the projection screen and tracked using a NaturalPoint OptiTrack optical motion capture system. The experiment and graphics were implemented in Pure Data using the Extended View Toolkit [34] for projection mapping.

[2]Small: 6,5,7,5,6,7,5,6,7,5,7,6,7,5,6,5,7,6,7,5,6,7,5,6,7,5,6,5 7,6,7,5,6,5,7,6,7,5; Medium: 6,4,8,3,7,5,9,3,6,5,8,4,7,9,3,6,4,8,7,5,9,3,7,5,8,4,6,9,3,6,4,8,7,5,9,3,7,5; Large: 6,3,9,2,7,5,10,1,4,8,11,1,6,4,9,2,8,5,10,7,3,11,1,6,3,9,2,8,5,10,7,4,11,1,7,4,9,2

[3]Small: 6,7,5; Medium: 6,3,9,7,5,4,8; Large: 6,3,9,2,7,5,10,1,4,8,11

[4]Gerriets OPERA® white perforated (PVC, 390 g/m^2, 7 percent perforation area)

Table 2: *The results of a four-way (Stimulus × Algorithm × Spatial Distribution × Orientation) repeated measures ANOVA on perceived spatial extent. Non-significant main effects and interactions (p>0.05) were omitted.*

Stimulus	F(2,34) = 20.749	p < 0.001
Algorithm	F(1,17) = 47.679	p < 0.001
Spatial Distribution	F(2,34) = 19.550	p < 0.001
Orientation	F(1,17) = 16.852	p = 0.001
Algorithm × Spatial Distribution	F(2,34) = 24.634	p < 0.001
Algorithm × Orientation	F(1,17) = 44.511	p < 0.001
Spatial Distribution × Orientation	F(2,34) = 15.171	p < 0.001
Algorithm × Spatial Distribution × Orient.	F(2,34) = 18.897	p < 0.001

2.4. Procedure and Participants

Participants sat on a chair at a distance of 2 m from the loudspeaker array. They went through the trials in a randomized order; there was a 500 ms silence between trials, enough to reset short-term echoic memory [35]. Each stimulus was presented continuously until the trial was over, with a 5 ms linear fade in and fade out. Horizontal and vertical orientation were tested in two separate trial groups presented in a counterbalanced order. There were four repetitions for each combination of algorithm, stimulus, and spatial distribution, leading to a total of 72 stimuli for both orientations.

Participants were instructed to use the toy gun to draw a straight line on the projection screen and to match its horizontal (or vertical) spatial extent to the perceived auditory horizontal (or vertical) spatial extent. They triggered the gun to indicate and adjust the end points of the line and were able to perform corrections before pressing a button on the gun to proceed to the next trial. 18 participants (5 female, M=26.3 years, SD=5.4 years), participated and received a small financial compensation. None of them had prior knowledge or training in the specific task. The task was not restricted in time.

2.5. Results

Perceived vs. physical spatial extent in the different conditions in the experiment are illustrated in Figure 3. It is evident that in most cases perceived extent was narrower in vertical compared to horizontal orientation. In addition, it appears that the time-based algorithm results in broader spatial extent perceptions compared to the frequency-based algorithm. Finally, although results are similar for the white noise and impulse train stimuli, the rain drops stimulus appears to be perceived narrower than the other signals.

The judgments in the different conditions in the experiment were verified to follow the normal distribution using the Lilliefors test [36]. No outliers were detected by Grubbs' test [37]. A four-way (Stimulus × Algorithm × Spatial Distribution × Orientation) repeated measures ANOVA on perceived spatial extent, as indicated by the horizontal or vertical distance between selection endpoints, was used to analyze the results (see Table 2). No violations of sphericity were observed.

2.5.1. Main effects

The main effects of Stimulus, Algorithm, Spatial Distribution, and Orientation were significant. Pairwise t-tests showed that white noise and impulse trains resulted in significantly broader perceived extent in comparison to the rain drops (p<0.001). Pairwise t-tests

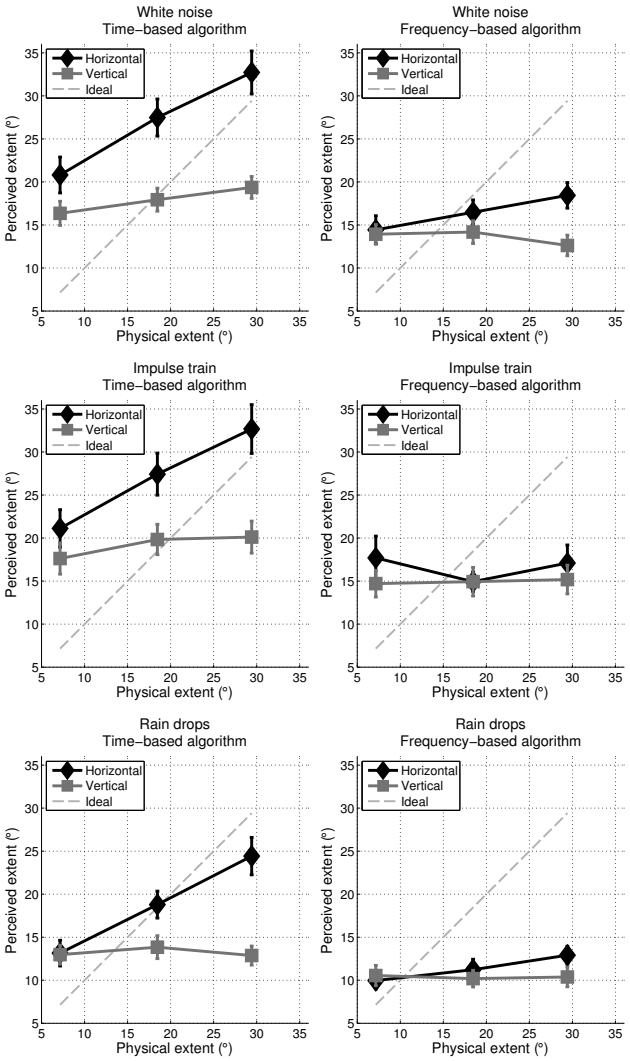

Figure 3: *Perceived vs. physical spatial extent in the different conditions in the experiment. Error bars indicate standard error of the mean.*

showed that the small spatial distribution (3 loudspeakers) was perceived to be significantly narrower than both the medium (7 loudspeakers) and the large (11 loudspeakers) distributions and the medium distribution narrower than the large (p<0.01). Finally, judgments of vertical extent were significantly narrower than those of horizontal extent, and the time-based algorithm resulted in significantly broader judgments.

2.5.2. Algorithm and Spatial Distribution

The interaction between Algorithm and Spatial Distribution was significant. This was because in the case of the time-based algorithm, averaged over orientation and stimulus, perceived extent was significantly different for the three spatial distributions tested in the experiment, e.g., small was perceived as significantly narrower than medium and large spatial distributions, and medium significantly narrower than the large spatial distribution (p<0.001).

This was not the case for the frequency-based algorithm, in which case averaged over stimuli and orientation no statistically significant differences in the perceived extent of different spatial distributions were observed.

2.5.3. Algorithm and Orientation

The interaction between Algorithm and Orientation was significant because in the case of the time-based algorithm, averaged over Stimulus and Spatial Distribution, judgments of the horizontal extent were significantly wider compared to those of vertical extent ($t(17)=5.855$, $p<0.001$), while for the frequency-based algorithm there was no significant difference between the extent of the horizontal and vertical judgments, arguably because they were narrow in both cases.

2.5.4. Spatial Distribution and Orientation

The interaction between Spatial Distribution and Orientation was also significant. This was because of two reasons. The first was that when loudspeakers were arranged horizontally, perceived horizontal extent was significantly influenced from the actual spatial extent, i.e., variations in the spatial distributions resulted in significantly different perceived spatial extent (pairwise-comparisons, at least $p<0.01$). When loudspeakers were aligned vertically, however, the perceived vertical extent varied only little in response to changes in the actual physical extent. Perceived spatial extents for the different spatial distributions were only marginally significantly different to each other, small smaller than medium ($t(17)=-1.839$, $p=0.083$), small smaller than large ($t(17)=-1.865$, $p=0.080$), medium vs. large not significant. The second reason is that for the smallest actual spatial distribution, the perceived horizontal and vertical extent judgments were not different to each other. However, perceived horizontal and vertical extent judgments in the other two spatial distributions were significantly different to each other in the horizontal but not in the vertical orientation.

2.5.5. Algorithm, Spatial Distribution, and Orientation

The three-way interaction between Algorithm, Spatial Distribution, and Orientation was also significant. This was because irrespective of stimulus and for spatial distributions other than the small, the perceived spatial extent was significantly larger in the case of the time-based spatialization for horizontally aligned stimuli in comparison to vertically aligned ones. This interpretation is supported by the observation that the two-way interaction between Orientation and Spatial Distribution was significant for the time-based spatialization algorithm but not for the frequency-based when analyzing the data averaged over stimuli.

3. PREDICTORS FOR APPARENT SOURCE WIDTH

For precise control of perceived spatial extent, predictors which can be calculated on the basis of measurable signal properties are sought. For vertical spatial extent no reliable predictors are known. However, in the case of horizontal spatial extent, i.e., apparent source width, interaural cross-correlation (IACC) and the lateral energy fraction (LF) may provide acceptable results. LF is related to IACC in the sense that a decorrelation between the two ear signals, yielding a low IACC (or high LF), could emerge on the one hand from a spatial distribution of uncorrelated individual sound sources, or on the other hand from room reflections [38]. On the other hand, changes in IACC may not only result from changes in the lateral energy arriving in the ears and may be the result of decorrelation operations in the signals.

3.1. Lateral energy fraction (LF)

Although traditionally used to quantify spaciousness in concert hall acoustics [8, p. 351], it was shown that the lateral energy fraction could also serve as a predictor for auditory source width of loudspeaker signals in short reverberation time environments [39]. The LF describes the ratio of the lateral energy to the total energy, and is computed by contrasting the impulse responses measured by an omni-directional microphone $h_\circ$ with that of a figure-eight microphone h_∞ [40, 41]. Usually, in case of the omni-directional microphone the first 5 ms of the impulse response are omitted [42]. However, for phantom sources in the horizontal plane it was shown that an adapted version, where both impulse responses start from zero (see Equation 1), is a better predictor of auditory source width, at least for pink noise signals [39]. Although this predictor may appear to have limited potential for application in our data, we include it here for completeness.

$$LF = \frac{\int_{0\ ms}^{80\ ms} h_\infty^2 \,\mathrm{d}t}{\int_{0\ ms}^{80\ ms} h_\circ^2 \,\mathrm{d}t} \tag{1}$$

Measurements of the adapted LF for the individual loudspeakers were performed with an NTi M2210 omni-directional microphone and a Schoeps type CMC 5 with MK 8 capsule as a figure-eight microphone. Both microphones were calibrated to compensate sensitivity-differences. The center loudspeaker led to an LF of 0.07, while the LF for the outmost left/right loudspeakers was 0.13. LF for the rest of the loudspeakers were obtained by linear interpolation. As a result, the 3, 7, or 11 simultaneous loudspeakers for the small, medium, or large spatial distribution led to an overall LF of 0.08, 0.09, and 0.10, respectively. These values may be interpreted to show a monotonically increasing LF for increasing physical extent. However, their range is small compared to the literature, e.g., [43] (0.025 compared to 0.15), and even less than one just-noticeable difference (JND) [44]. As already suggested by [43], it therefore appears that the LF is not a suitable predictor for the apparent source width in our experiments, especially as it is computed from impulse responses, ignoring the effects of algorithm and stimulus type.

3.2. Interaural cross-correlation coefficient (IACC)

In previous studies, the IACC, which is the maximum of the cross-correlation between the left and right channel of a binaural recording [8], was shown to be a good predictor for perceived spatial extent in the horizontal plane [2]. To verify this claim, binaural measurements were performed with a head and torso simulator (HATS, B&K type 4128C), which was placed at the listening position. Subsequently the IACC was calculated using the recordings for all combinations of the independent variables Stimulus, Algorithm, Spatial Distribution, and Orientation. While the IACC for vertically extended sound sources was always constantly above 0.8, in horizontal orientation it varied systematically with the apparent source width (see Figure 4).

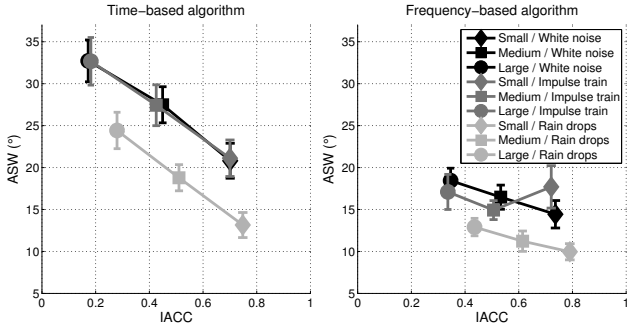

Figure 4: *Apparent source width (ASW) as a function of the IACC for the different conditions of the experiment. Error bars indicate standard error of the mean.*

It is evident in Figure 4 that the majority of the findings reported earlier can be explained on the basis of the IACC algorithms. In particular, larger spatial distributions led to a lower IACC than smaller ones, and the time-based algorithm resulted in lower IACC values and a smaller range of the values compared to the frequency-based algorithm. Furthermore, the rain drops led to always higher IACC than both white noise and impulse train stimuli which produced similar values, which explains why this was always judged to be narrower than the other two stimuli. The correlation coefficient between IACC values for each spatial distribution and the resulting perceived horizontal extent (ordinate and abscissa in Figure 4) was calculated. In the case of the time-based algorithm, a value of at least -0.996 was obtained when considering all three stimuli. In the case of the frequency based algorithm, a value of at least -0.997 was obtained for the noise and the rain stimuli, however the correlation for the impulse train was poor (0.27). On a closer inspection, this relates to an unexpected trend in the ASW value in the small spatial distribution for this condition and stimulus combination (see also the behavioral data in Figure 3).

Concluding, the IACC was found to be a good predictor for perceived spatial extent in the horizontal plane, as it is highly correlated with the absolute judgments of perceived spatial extent.

4. DISCUSSION

The results of the experiments and the acoustic measurements allow certain conclusions to be made with respect to the possibility of eliciting the perception of either vertically or horizontally extended sounds. In summary, even within the relatively small spatial distributions used in this experiment, it was possible to create the impression of horizontally extended sounds. This was not the case for vertical extent. The algorithms used here can only partially create the impression of vertical extent within the range of spatial extents used in the experiments. Finally, irrespective of orientation or stimulus type, the time-based algorithms resulted in significantly larger perceptions of both horizontal and vertical extent.

4.1. Time-based algorithm

Perceived horizontal extent created by the time-based algorithm varied systematically with the actual extent of the spatial distribution irrespective of stimulus type as evidenced by the fact that the different horizontal extents used in the study resulted in significantly different distributions of perceived spatial extent. Interestingly, actual horizontal extent was overestimated in the perceptual judgments, especially at the smaller actual spatial extents. In the vertical direction, however, perceived spatial extent varied less systematically with the actual one. Although a significant increase in the perceived vertical spatial extent with increased actual vertical spatial extent appeared for the white noise and the Dirac impulses, the difference was consistently significant only when comparing the smallest with largest displacement ($t(17)=3.46$, $p=0.003$ for white noise and $t(17)=2.43$, $p=0.047$ for impulses) and no significant differences in perceived vertical extent when using the rain drops stimulus were observed. In addition, judgments of perceived vertical extent underestimated actual extent by far pointing to limited applicability in real-world applications.

4.2. Frequency-based algorithm

Concerning the frequency-based algorithm, although in general perceived horizontal extent increased in proportion to the actual horizontal extent, as a rule judgments underestimated the actual horizontal extent of the spatial distribution and were significantly narrower than the ones obtained by the time-based algorithm. In addition, the algorithm failed to represent vertically extended sound sources. This could be attributed to the different mechanisms that operate and determine azimuth and elevation perception. While azimuth perception operates on the basis of interaural time differences, spectral cues and familiarity with source spectrum are mainly responsible for elevation perception [45, 8]. It appears therefore that while the combination of information from frequencies at different azimuths to yield the impression of coherent spatially extended auditory sources provides a functional basis for the creation of horizontally extended sources, this mechanism fails for vertically extended sources. This may be explained by the fact that presenting signal frequencies at different elevations destroys the consistency with which the signal spectrum is filtered by the outer ear to result in the perception of elevation. This is a fundamental problem when it comes to representing vertical extent by distributing signal frequencies in elevation that might be difficult to overcome.

4.3. Stimuli

Performance for the white noise and the impulse train stimulus was similar, both for horizontally and for vertically extended sources. The rain drops were perceived to be consistently narrower. In addition, although differences in spatial extent represented with this stimulus were well identified in the horizontal orientation, this was not the case in the vertical one. The aforementioned difficulties could arguably relate to the bandwidth of the rain drops stimulus, which was smaller compared to other two. The difficulties in vertical extent perception might relate to sound design issues that need to be investigated further, such as optimization of the grains to yield as good localization as possible.

4.4. Sound design

An aspect worth considering further is the overestimation of the actual spatial extent that occurred for all stimuli in the horizontal orientation in the case of the time-based and to a lesser extent in the case of the frequency-based algorithm. This may be attributed to

non-spatial factors pertaining to source-size perception, that may confound spatial extent judgments. It appears that the creation of predetermined spatial extent impressions requires the simultaneous calibration of both spatial and non-spatial factors. A general solution to provide a specific spatial extent that is applicable to all signals may therefore be difficult to achieve and perceptual calibration might be necessary in order to improve the match between actual and perceived spatial extent.

4.5. Predictors for apparent source width

The perceived spatial extent judgments in the horizontal orientation in the experiments could be explained on the basis of the IACC. LF values did not correlate with spatial extent measurements. This could be expected as room acoustics were the same throughout the experiment and the distance between active loudspeakers in the experiment was small. Listeners responded therefore on the basis of IACC and not LF. LF may be interpreted to indicate the contribution of the interaction between loudspeaker positioning and room on the judgments of spatial extent. The observed values show that this contribution is negligible.

4.6. Loudspeaker array design

In the experiment, adjacent loudspeakers were used to create the impression of spatially extended sources. This may appear uneconomical as an auditory source width of at least 10 degrees for a single loudspeaker emitting noise, depending on room acoustics and loudspeaker model has been observed [39]. Furthermore, gaps of up to 15 degrees in noise emitting loudspeaker arrays were found to be difficult to notice [10, 13]. The loudspeakers we have used, however, were smaller than the ones used in the aforementioned studies. We opted out from introducing gaps in the distribution in order to exclude the possibility of perceptual discontinuities in the perceived auditory event. It may however well be that the results of this experiment could be replicated with even less loudspeakers than used here, given appropriate calibration.

It may be worth noting that the difficulties with vertical perception in the experiment may originate in the small range of spatial extents used. It would therefore be interesting to replicate this study using larger spatial distributions in vertical orientation in order to understand whether the limitations observed here reflect a limitation in the algorithms used or a limitation in the perception of vertical extent in the auditory system. Larger spatial distributions and listener training may be interesting factors to vary in future experiments targeting this aspect.

4.7. Applications

Concerning the auditory representation of horizontal extent the results are very promising. Participants could differentiate well even in response to the small spatial distributions tested here. Designers may therefore start to integrate horizontally extended sounds in virtual and mixed reality applications. It also appears that musical compositions in which the spatial extent of sounds is explicitly manipulated will become commonplace in the future. The results of this study show that when extents of small magnitude need to be used, time-based extent synthesis algorithms are preferable, as they yield larger impressions of horizontal extent. The use of granular synthesis in this context appears to be a far reaching solution for sound and interface designers.

5. CONCLUSION

We presented a study that investigated the perception of auditory spatial extent using two spatial extent synthesis algorithms. The algorithms aimed to create the impression of spatially extended objects by either distributing the frequencies or the grains comprising a sound source in space. In a controlled experiment, the ability of the algorithms to create spatially extended sound sources as a function of Spatial Distribution, Orientation, and Stimulus type was tested. It was found that while both algorithms were successful in generating the impression of horizontally extended sound sources, the time-based algorithm resulted in broader perceptions of spatial extent irrespective of Stimulus, Orientation, or actual Spatial Distribution. Furthermore, for similar spatial distributions, judgments of horizontal extent were significantly larger than these of vertical extent. Finally, judgments of horizontal extent overestimated the physical extent, while judgments of vertical extent underestimated the physical extent. Results could be explained on the basis of measurements of the interaural cross-correlation in the different conditions in the experiment.

6. REFERENCES

[1] Jens Ahrens and Sascha Spors, "Two physical models for spatially extended virtual sound sources," in *Audio Engineering Society Convention 131*, Oct 2011.

[2] R. Mason, T. Brookes, and F. Rumsey, "Frequency dependency of the relationship between perceived auditory source width and the interaural cross-correlation coefficient for time-invariant stimuli," *J. Acoust. Soc. Am.*, vol. 117, no. 3 Pt 1, pp. 1337–1350, Mar 2005.

[3] D. R. Perrott and T. N. Buell, "Judgments of sound volume: effects of signal duration, level, and interaural characteristics on the perceived extensity of broadband noise," *J. Acoust. Soc. Am.*, vol. 72, no. 5, pp. 1413–1417, Nov 1982.

[4] Densil Cabrera and Steven Tilley, "Parameters for auditory display of height and size," in *Proceedings of the 9th International Conference on Auditory Display (ICAD)*, Eoin Brazil and Barbara Shinn-Cunningham, Eds., Boston, MA, July 2003, International Community for Auditory Display, Georgia Institute of Technology.

[5] A. J. Kunkler-Peck and M. T. Turvey, "Hearing shape," *J Exp Psychol Hum Percept Perform*, vol. 26, no. 1, pp. 279–294, Feb 2000.

[6] Claudia Carello, Krista L. Anderson, and Andrew J. Kunkler-Peck, "Perception of object length by sound," *Psychological Science*, vol. 9, no. 3, pp. 211–214, May 1998.

[7] Mark Kac, "Can one hear the shape of a drum?," *American Mathematical Monthly*, vol. 73, no. 4/2, pp. 1–23, Apr. 1966.

[8] Jens Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press, revised edition, 1997.

[9] J. Blauert and W. Lindemann, "Spatial mapping of intracranial auditory events for various degrees of interaural coherence," *The Journal of the Acoustical Society of America*, vol. 79, no. 3, pp. 806–813, 1986.

[10] Toni Hirvonen and Ville Pulkki, "Center and spatial extent of auditory events as caused by multiple sound sources in frequency-dependent directions," *Acta Acoustica united with Acoustica*, vol. 92, pp. 320–330, 2006.

[11] Olli Santala and Ville Pulkki, "Directional perception of distributed sound sources," *The Journal of the Acoustical Society of America*, vol. 129, no. 3, pp. 1522–1530, 2011.

[12] Olli Santala and Ville Pulkki, "Resolution of spatial distribution perception with distributed sound source in anechoic conditions," in *Audio Engineering Society Convention 126*, May 2009.

[13] Toni Hirvonen and Ville Pulkki, "Perceived spatial distribution and width of horizontal ensemble of independent noise signals as function of waveform and sample length," in *Audio Engineering Society Convention 124*, May 2008.

[14] Koichiro Hiyama, Setsu Komiyama, and Kimio Hamasaki, "The minimum number of loudspeakers and its arrangement for reproducing the spatial impression of diffuse sound field," in *Audio Engineering Society Convention 113*, Oct 2002.

[15] Mikko-Ville Laitinen, Tapani Pihlajamäki, Cumhur Erkut, and Ville Pulkki, "Parametric time-frequency representation of spatial sound in virtual worlds," *ACM Trans. Appl. Percept.*, vol. 9, no. 2, pp. 8:1–8:20, June 2012.

[16] Tapani Pihlajamaki, Olli Santala, and Ville Pulkki, "Synthesis of spatially extended virtual source with time-frequency decomposition of mono signals," *J. Audio Eng. Soc*, vol. 62, no. 7/8, pp. 467–484, 2014.

[17] Franz Zotter, Matthias Frank, Georgios Marentakis, and Alois Sontacchi, "Phantom source widening with deterministic frequency dependent time delays," in *Proc. of the 14th International Conference on Digital Audio Effects (DAFx-11)*, Paris, France, Sept. 2011, pp. 307–312.

[18] B. C. J. Moore and B. R. Glasberg, "A revision of Zwicker's loudness model," *Acustica United with Acta Acustica*, vol. 82, no. 2, pp. 335–345, 1996.

[19] Toni Hirvonen and Ville Pulkki, "Perception and analysis of selected auditory events with frequency-dependent directions," *J. Audio Eng. Soc*, vol. 54, no. 9, pp. 803–814, 2006.

[20] J. H. Halton, "Algorithm 247: Radical-inverse quasi-random point sequence," *Commun. ACM*, vol. 7, no. 12, pp. 701–702, Dec. 1964.

[21] Michael A. Gerzon, "Signal processing for simulating realistic stereo images," in *Audio Engineering Society Convention 93*, Oct 1992.

[22] Barry Truax, "Composition and diffusion: space in sound in space," *Organised Sound*, vol. 3, pp. 141–146, 8 1998.

[23] Natasha Barrett, "Spatio-musical composition strategies," *Organised Sound*, vol. 7, pp. 313–323, 12 2002.

[24] Etienne Deleflie and Greg Schiemer, "Spatial grains: Imbuing granular particles with spatial-domain information," in *Proceedings of the Australasian Computer Music Conference ACMC09*, July 2009.

[25] Curtis Roads, *Microsound*, The MIT Press, 2004.

[26] Rory Wallis and Hyunkook Lee, "The effect of interchannel time difference on localization in vertical stereophony," *J. Audio Eng. Soc.*, vol. 63, no. 10, pp. 767–776, October 2015.

[27] Hyunkook Lee, "Investigation on the phantom image elevation effect," in *139th Audio Engineering Society Convention*, October 2015, This is Author Accepted Manuscript for Green Open Access.

[28] Ville Pulkki, "Localization of amplitude-panned virtual sources ii: Two- and three-dimensional panning," *J. Audio Eng. Soc*, vol. 49, no. 9, pp. 753–767, 2001.

[29] James L. Barbour, "Elevation perception: Phantom images in the vertical hemi-sphere," in *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*, Jun 2003.

[30] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, 1990.

[31] D. Rocchesso, *Introduction to Sound Processing*, Mondo estremo, 2003.

[32] Sebastian Blamberger, "Development of a modular system for speaker array prototyping," Tech. Rep., Institute of Electronic Music and Acoustics, Graz University of Music and Performing Arts, 2011.

[33] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *J. Acoust. Soc. Am.*, vol. 112, no. 5 Pt 1, pp. 2110–2117, Nov 2002.

[34] Peter Venus, Marian Weger, Cyrille Henry, and Winfried Ritsch, "Extended view toolkit," in *Proceedings of the 4th Pure Data Convention*, 2011, pp. 161–167.

[35] Nelson Cowan, "On short and long auditory stores," *Psychological Bulletin*, vol. 96, no. 2, pp. 341–370, Sep 1984.

[36] Hubert W. Lilliefors, "On the kolmogorov-smirnov test for normality with mean and variance unknown," *Journal of the American Statistical Association*, vol. 62, no. 318, pp. 399–402, 1967.

[37] Frank E. Grubbs, "Sample criteria for testing outlying observations," *Ann. Math. Statist.*, vol. 21, no. 1, pp. 27–58, 03 1950.

[38] Johannes Käsbach, Marton Marschall, Bastian Epp, and Torsten Dau, "The relation between perceived apparent source width and interaural cross-correlation in sound reproduction spaces with low reverberation," in *Proceedings of DAGA 2013*, 2013.

[39] Matthias Frank, "Source width of frontal phantom sources: Perception, measurement, and modeling," *Archives of Acoustics*, vol. 38, no. 3, pp. 311–319, 10 2013.

[40] Barron M and Marshall AH, "Spatial impression due to early lateral reflections in concert halls: the derivation of a physical measure," *Journal of Sound & Vibration*, vol. 77, no. 2, pp. 211–232, 1981.

[41] Trevor J. Cox, W. J. Davies, and Yiu W. Lam, "The Sensitivity of Listeners to Early Sound Field Changes in Auditoria," *Acustica*, vol. 79, no. 1, pp. 27–41, 1993.

[42] ISO, "3382-1:2009: Acoustics - measurement of room acoustic parameters - part 1: Performance spaces.," 2009.

[43] Matthias Blau, "Correlation of apparent source width with objective measures in synthetic sound fields," *Acta Acustica united with Acustica*, vol. 90, no. 4, pp. 720–730, 2004-07-01T00:00:00.

[44] Matthias Blau, "Difference limens for measures of apparent source width," in *Forum Acusticum*, Sevilla, Spain, 2002.

[45] F. Wightman and D. Kistler, "Factors affecting the relative salience of sound localization cues," *Binaural and spatial hearing in real and virtual environments*, vol. 1, pp. 1–23, 1997.