euronoise 2021
25th to 27th of October
Madeira, Portugal - online

Sound field synthesis: Simulation and evaluation of auralized interaural cues over an extended area

Matthieu Kuntz, Bernhard U. Seeber

Audio Information Processing, Technical University of Munich, Germany
{matthieu.kuntz , seeber@tum.de}

Abstract
Virtual acoustic environments paired with sound field synthesis methods enable the controlled auralization of a wide range of sound scenes, from single sources to a busy environment. Most sound field synthesis methods are mathematically accurate at the centre of a loudspeaker array, their reproduction errors increase as the distance to the centre increases. For a perceptual evaluation of sound field reproduction methods, an analysis of the resulting interaural cues is essential. Level errors in the sound field might be tolerable for evaluating certain aspects of spatial perception as long as the interaural cues match those present in the target sound field. This work evaluates simulated interaural cues in the centre and at off-centre positions in a square loudspeaker array for a sound source auralized with Higher-Order Ambisonics. We show a clear decrease of the consistency of the interaural cues as the distance to the centre increases, and different ILDs for the onset of the stimulus.

Keywords: interaural cues, sound field reproduction, auralization, Ambisonics.

1 Introduction

In this paper, we investigate the interaural cues that result from the playback of a virtual sound source with Higher-Order Ambisonics (HOA) at different positions inside the loudspeaker array. While the sound field reproduction methods are computed to be accurate in the centre of the array, their performance for off-centre listening position is the subject of research. Assessing the reproduction errors that arise when listening conditions are not optimal (*e.g.* when the listener is placed off-centre) is necessary for the verification of any virtual environment used for research purposes.

Evaluations of sound field synthesis methods were mostly simulated and evaluated for simple sound scenes, such as those generated by a plane wave or a point source, which is in the focus of this work as well. Numerical evaluations of the characteristics of reproduced sound fields were carried out to compare different coding strategies [1] or to validate a given setup [2, 3]. Perceptual evaluations of sound field synthesis measured localization errors [4-6], distance perception [7], speech intelligibility [8], or timbre [9].

Some investigation of interaural level differences (ILDs) and interaural time differences (ITDs) can be found in Daniel's work [10, 11], where he showed that the magnitude of ILDs resulting from a HOA reproduction at the centre of the loudspeaker array is smaller than that resulting from the equivalent real source in the free-field. The influence of listener placement on the resulting interaural cues is still a fairly open question. Wierstorf [12] studied localisation performance at different listener positions using a localization model based on the interaural phase difference, using ILDs to unwrap the phase differences. The interaural cues themselves were not investigated further.

With Ambisonics, different frequencies have different *sweet-spot* sizes: the area of good sound field reproduction shrinks with increasing frequency. This introduces a frequency and position dependent reproduction error, which could introduce conflicting interaural cues across auditory filters. Furthermore, at eccentric listener positions, the distance differences between loudspeakers lead to a sound field build-up of a couple milliseconds, until all loudspeaker signals have arrived at the listeners ears. During this onset time, the interaural cues are also susceptible to carry different information about a sound source's position than the cues resulting in the steady-state part of the sound field reproduction. However, due to the precedence effect, there is a particularly high perceptual weight on the onset [13]. This work investigates the distribution of interaural cues across time frames and frequency bands for a single sound source auralized with HOA.

2 Methods

We simulated the 36 horizontally arranged loudspeakers (10° spacing) of the Simulated Open Field Environment (SOFE, [14, 15]) installed in the anechoic chamber at the Technical University of Munich. The loudspeakers are mounted on a custom 4.8 m × 4.8 m square holding frame at a height of 1.4 m. The loudspeaker at 0°, in front of the listener, has a distance of 2.4 m to the centre of the loudspeaker array. The loudspeakers were simulated as perfect point sources. They were virtually equalized by delaying and scaling their playback signal to ensure equal time of arrivals (TOA) in the centre of the loudspeaker array. A grid of 11 × 11 evaluation positions was defined, spanning ±1.6 m from the loudspeaker array centre, two adjacent evaluation positions being 32 cm apart. Figure 1 shows the evaluation area within the loudspeaker setup, to scale. A sound source placed at $\varphi = 13°$ and $r = 4$ m was simulated.

In order to maintain a good balance between onset and steady-state interaural cues, the stimulus used was a sequence of five 50 ms bursts of uniformly exciting noise [16] between 100 Hz and 18 kHz, separated by 50 ms of silence.

2.1 Sound field synthesis method

In this work, we used the Ambisonics *basic* decoder. The sound source $s(t)$ is decomposed into a set of basis functions (circular harmonics for a 2D representation of the sound field) to describe the directional information in the sound field. For a single sound source $s(t)$, this results in the encoding vector

$$\boldsymbol{y}(\varphi_S) = \left[1, \sqrt{2}\cos(\varphi_S), \sqrt{2}\sin(\varphi_S), \dots, \sqrt{2}\cos(N\varphi_S), \sqrt{2}\sin(N\varphi_S)\right]^T \tag{1}$$

where φ_S represents the source direction, as described by [10]. This format can be used to adapt the auralization to various loudspeaker setups, which is done in a subsequent decoding stage. The basis functions are sampled at the loudspeaker positions to derive the loudspeaker gains by multiplying the encoded sound source

$$\boldsymbol{\chi}(t) = \boldsymbol{y}(\varphi_S) \times s(t) \tag{2}$$

with a decoding matrix

$$\boldsymbol{D} = \frac{1}{L}\left[\boldsymbol{y}(\varphi_{LS,1}), \dots, \boldsymbol{y}(\varphi_{LS,L})\right]^T. \tag{3}$$

L represents the number of loudspeakers and φ_{LS} their respective azimuth angle. Combining encoding and decoding yields the loudspeaker signals

$$\boldsymbol{S}_{LS,basic}(t) = \boldsymbol{D} \times \boldsymbol{\chi}(t) \tag{4}$$

When auralizing a sound source from 13° and 4 m distance on the simulated loudspeaker array, the loudspeaker with the highest level is at 10°, followed by the loudspeaker at 20°, 7 dB below, followed by the loudspeakers at 0° and 30°, 14 dB below.

2.2 Simulation of ear signals

The HRTF database used was generated by measuring the transfer function from an equalized Dynaudio BM6A mkII loudspeaker (Dynaudio, Skanderborg) at a distance of 2.4 m, to the ears of an artificial head (HMS II.3, Head Acoustics, Aachen) rotating in 0.5° steps. The HRTFs were normalized at the head centre. For evaluating sound field synthesis, a virtual listener facing the 0° direction was assumed. The ear signals computed for each evaluation position by scaling and delaying each loudspeaker signal to account for the distance attenuation and the TOA. The individually scaled and delayed signals were then convolved with the HRTF corresponding to the angle between the 0°-facing virtual listener and the loudspeakers. This procedure considers the azimuthal effect of loudspeaker position on the ear signals, but it does not consider effect of loudspeaker directionality.

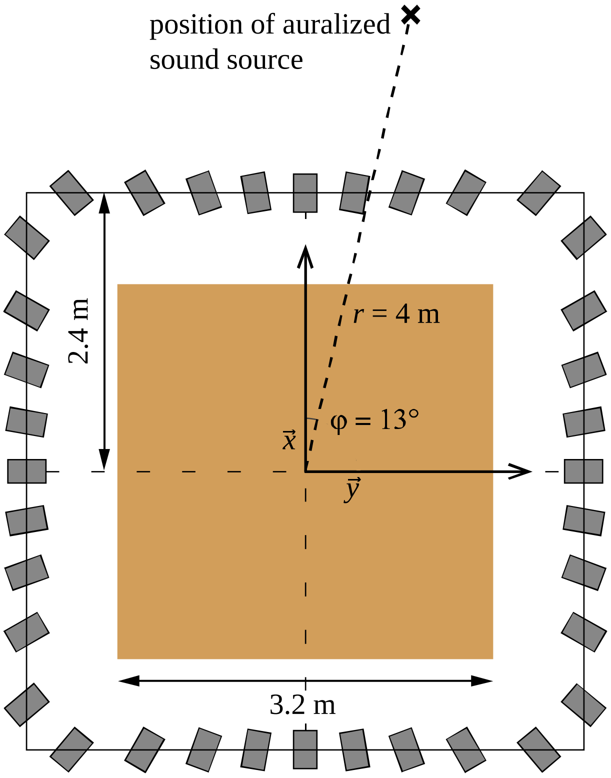

Figure 1 – Simulated loudspeaker array. The loudspeakers are placed on a 4.8 m × 4.8 m frame in 10° azimuth steps and equalized in level, delay and phase. The shaded area indicates the 11 × 11 evaluation grid around the array centre. The origin of the used Cartesian coordinate system is in the middle of the loudspeaker array indicated by the two axes *x* and *y*.

2.3 Analysis of the ear signals

At every evaluation position, ear signals $s_L$and s_R were computed and analysed in Bark bands by filtering them with a Gammatone filterbank. The signals in each critical band b were then processed in time frames by applying 12 ms Hann windows with 50% overlap to represent the short time evaluation ability of the auditory system. The interaural cues were computed in each time frame n.

The interaural level difference (ILD) was computed by subtracting the level of the signal at the left ear from that at the right ear. The interaural time difference (ITD) was computed as the lag of the maximum of the cross-correlation function between both ear signals for Bark bands 1 to 11 (below 1.5 kHz). Positive ILDs and ITDs correspond to a sound source on the right side of the head. The interaural coherence (IC) was defined as the maximum value of the cross-correlation function between the ear signals.

In each time frame, the interaural cues were computed only for the frequency bands where the level at both ears was above the threshold in quiet. Hence, the interaural cues were not computed for the inaudible parts of the signal since they are not relevant. To account for the weighting of interaural cues across frequencies, the computed interaural cues were weighted according to the Raatgever frequency weighting, based on the polynomial approximation derived in [17]. The interaural cues were then histogrammed over the length of the signal, resulting in 88 histogram bins. In order to analyse the spread of the histograms, the data were fitted with a Gaussian function by optimizing its amplitude, mean value and standard deviation with a least-squares algorithm.

3 Results

We show here the interaural cues computed on a grid of 11 × 11 evaluation points in a simulated loudspeaker array (Figure 1). The virtual sound source is placed at 13° and 4 m distance, or at the point (x = 3.9 m, y = 0.9 m), which is why we expect a line of 0 dB ILD and 0 µs ITD to run along y = 0.9 m, where the sound source would be directly in front of the listener.

3.1 Interaural level differences

The distribution of interaural level differences across time frames is shown in Figure 2, where each panel corresponds to one evaluation position. While we observe ILDs between -20 dB and 20 dB, the majority of ILD values lay between -5 dB and 10 dB. The high ILD values appear mostly at the sides of the evaluation zone (*e.g.* panels 44, 45, 66). We observe that the distribution of ILDs becomes slightly larger with increasing distance to the centre of the loudspeaker array. Note that the widened distribution does not stem from the variation of ILDs across frequency since only one spectrally weighted ILD was histogrammed per time frame, but from a change of the ILD over the signal's duration. The 0 dB ILD line lies between y = 0.32 m and y = 0.64 m.

3.2 Interaural time differences

The distribution of interaural time differences across time frames is shown in Figure 3, where each panel corresponds to one evaluation position. We observe ITDs between -750 µs and 750 µs, with some outliers indicating ITDs of more than 1 ms. These outliers appear more frequently at the sides of the evaluation zone. The spectrally averaged ITDs around the centre are very consistent, until some secondary peaks appear further away (*e.g.* panels 47, 92). On the right side of the evaluation zone, the ITDs are very inconsistent across the signal's duration and spread out over the whole range. In some cases (*e.g.* panels 44, 120), the outliers account for more than half of the ITD bins. We observe an ITD of 0 µs around panel 62 (y = 0.32 m), which is in accordance with the observed 0 dB ILD between y = 0.32 m and y = 0.62 m.

3.3 Interaural coherence

The distribution of interaural coherence across time frames is shown in Figure 4, where each panel corresponds to one evaluation position. The IC around the centre is very high, as it would be for a single sound source in the free field. As the distance to the centre increases, the IC decreases and the distribution becomes flatter as more ICs move from 0.95 and above to low values. Interestingly, the spread in IC values is not as extreme as the one seen for the ITDs, also indicating that ITDs fluctuate consistently over time, which, on a shorter time scale, leads to reduced and somewhat constant IC.

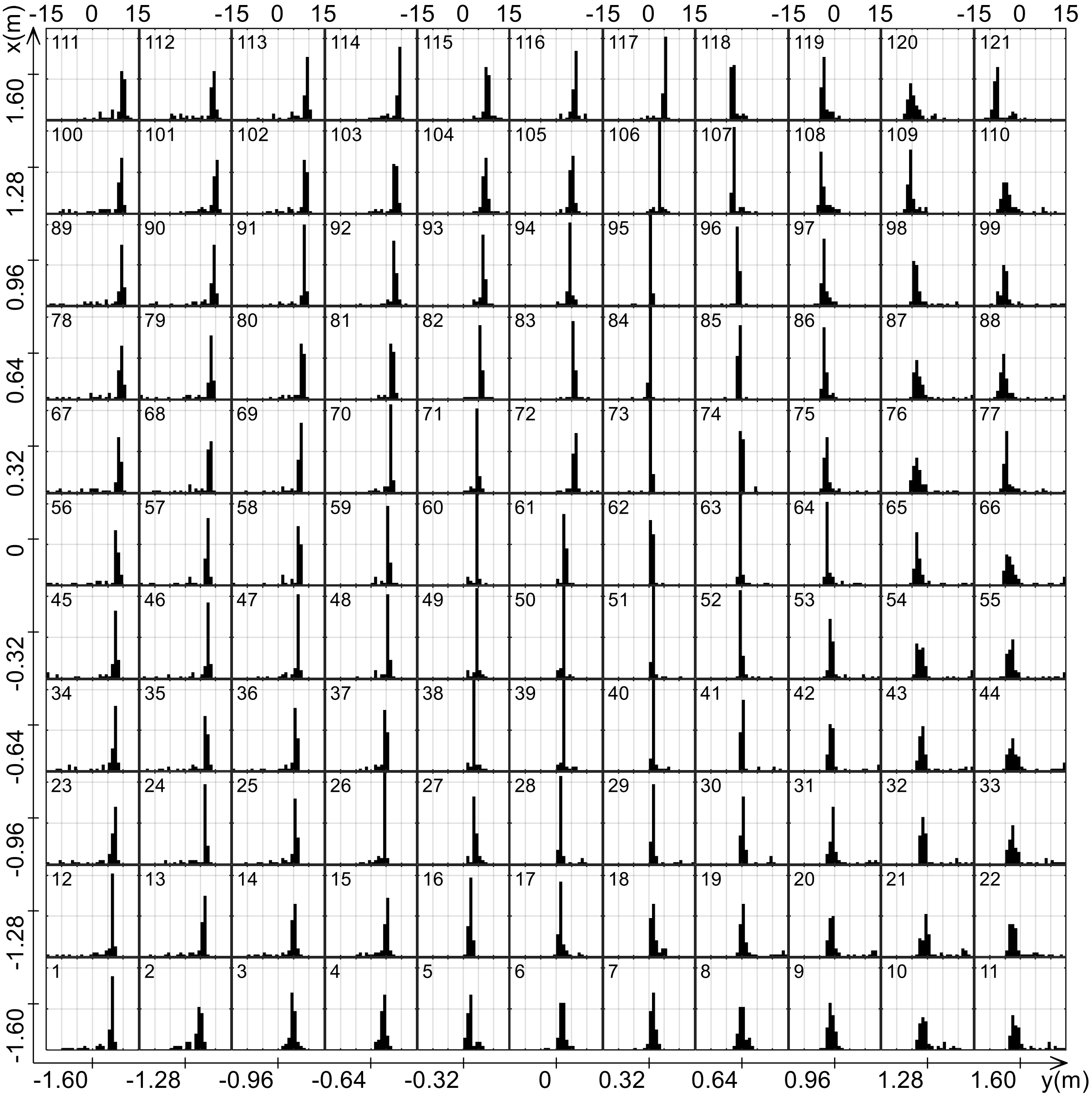

Figure 2 – Histogram of ILDs in dB across time frames. Each panel corresponds to one evaluation point, whose coordinates are indicated on the large axes left and below the panels. The *x*-axis of each panel corresponds to the ILD value grouped in bins of 1 dB between -15 dB and 15 dB. Outliers are added to the first and last bins. The bin height indicates the number of occurrences of each ILD value in the whole signal. The sound source is positioned at 13° and 4 m distance, or *x* = 3.9 m, *y* = 0.9 m.

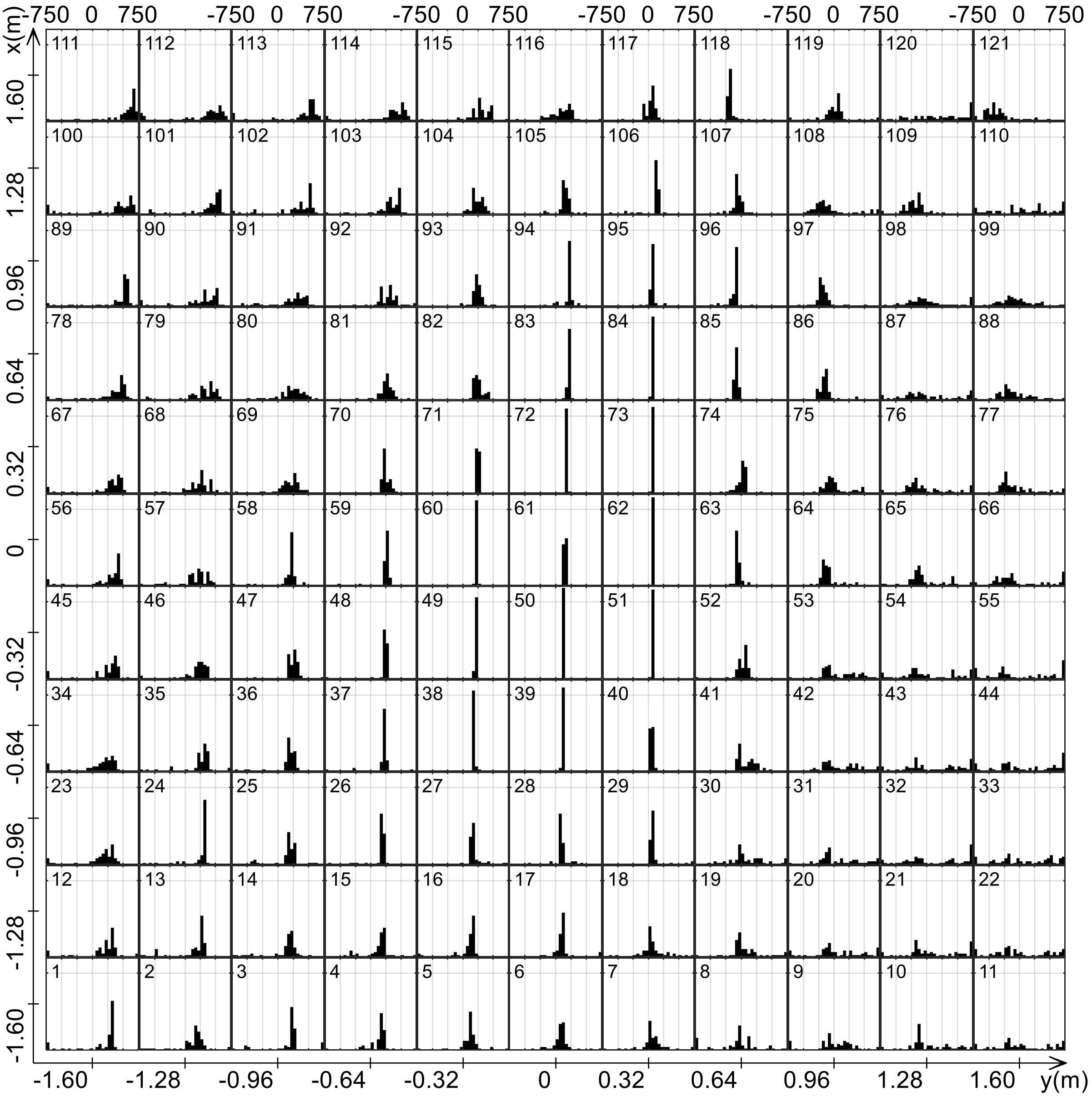

Figure 3 – Histogram of ITDs in µs across time frames. Each panel corresponds to one evaluation point, whose coordinates are indicated on the large axes left and below the panels. The *x*-axis of each panel corresponds to the ITD value grouped in bins of 50 µs, between -750 µs and 750 µs. Outliers are added to the first and last bins. The bin height indicates the number of occurrences of each ITD value in the whole signal. The sound source is positioned at 13° and 4 m distance, or *x* = 3.9 m, *y* = 0.9 m.

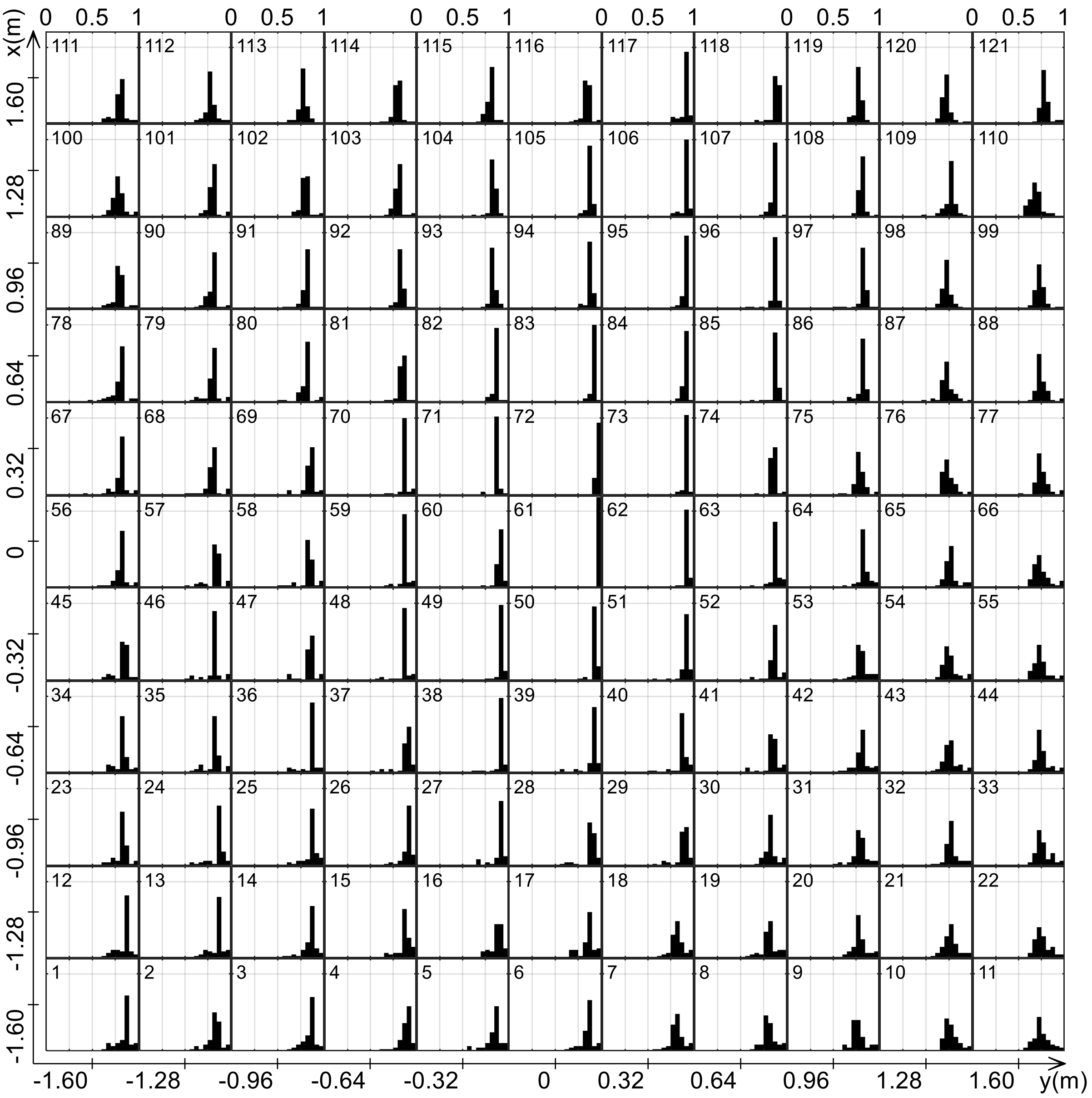

Figure 4 – Histogram of ICs across time frames. Each panel corresponds to one evaluation point, whose coordinates are indicated on the large axes left and below the panels. The *x*-axis of each panel corresponds to the IC value grouped in bins of 0.05, between 0 and 1. The bin height indicates the number of occurrences of each IC value in the whole signal. The sound source is positioned at 13° and 4 m distance, or *x* = 3.9 m, *y* = 0.9 m.

4 Discussion

We simulated the ear signals of a virtual listener inside a loudspeaker array, when playing a noise burst signal from a virtual source at 13° and 4 m distance from the centre using HOA.

The first finding is that the interaural cues are more widely distributed and less consistent across time frames as the distance to the centre of the loudspeaker array increases, which is represented on Figure 2-4.

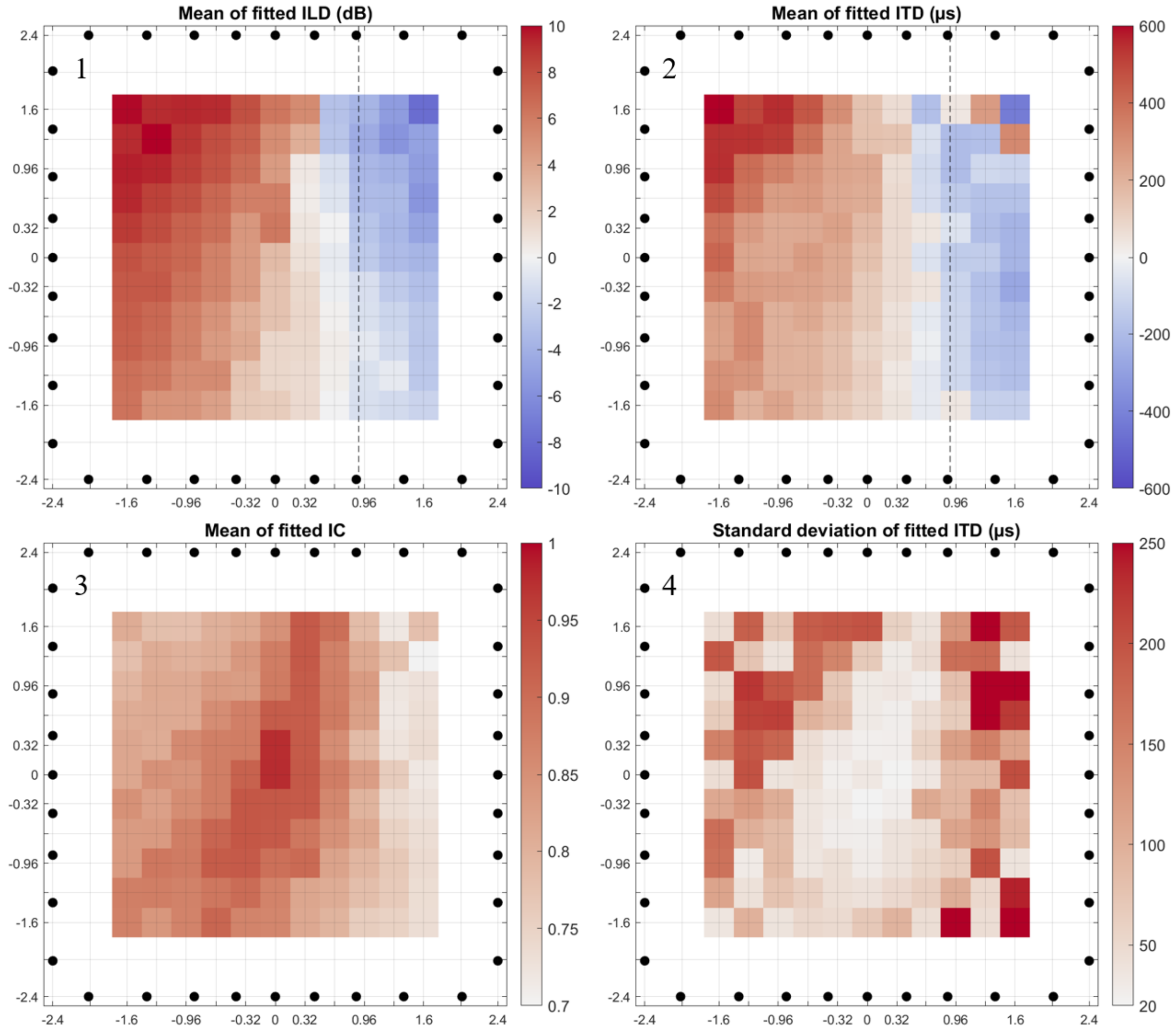

Figure 5 – Distribution of interaural cues across the evaluation zone in the loudspeaker array. The centre of each colored square represents the evaluation position; the loudspeakers are indicated as black dots. Top row: The dotted line shows the expected position of 0 interaural cues.

We also observe a clear shift in parallax. The ILDs and ITDs are null at evaluation positions between y = 0.32 m and y = 0.64 m, as shown in panels 1 and 2 of Figure 5, while the virtual source would have been in front of the listener at y = 0.9 m. This is explained by looking at the position of the main loudspeaker, placed at y = 0.42 m. Even though our virtual sound source is placed 4 m away, the physical sound source generating the sound field is closer. This is a well-known issue with Ambisonics and its plane wave based sound field model, where the distance of the virtual sound source and the distance of the loudspeakers are not considered. To correct this, Daniel introduced Near-Field Compensated Higher-Order Ambisonics (NFC-HOA), including distance information of the source and the loudspeakers [1]. Panel 4 of Figure 5 shows the standard deviation of the fitted ITDs. While the mean value agree qualitatively with our expectations, pointing towards the main loudspeaker, the standard deviation indicates that the mean value probably does not meaningfully represent the ITD perception at those evaluation positions.

The outliers observed in the results occur at the onset of the signal. Figure 6 shows the frequency-weighted ILD across time for all evaluation positions. It is clearly seen that the ILD magnitude at the onset of the stimulus is much higher and settles after the first time frames. This is due to the difference in TOA between the loudspeakers for off-centre evaluation positions. At the sides, this difference is large enough for the onsets of individual loudspeaker signals to fall into different time frames, resulting in strong onset cues. When considering the strong perceptual onset weighting, this illustrates the high probability of mislocalisation and split images perceived at the edges of the reproduction zone and corresponds qualitatively to the experimental data from [18]. The lower interaural coherence found on the sides, shown in panel 3 of Figure 5, indicates that the sound field will be perceived as more diffuse, with an increased apparent source width, which also impairs localisation ability. Indicated by the consistently low IC over time, this likely stems from ongoing phase fluctuations due to the non-ideal superposition of loudspeaker signals at off-centre positions.

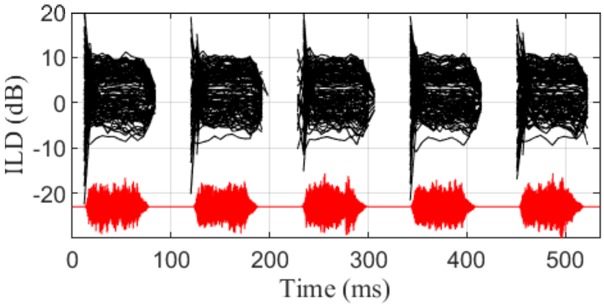

Figure 6 – Variation of ILDs in dB across time frames and evaluation positions. The black lines indicate the ILD at a every evaluation position. The red line shows the ear signals at every evaluation position, offset for better readability.

An asymmetric behaviour of the IC becomes apparent, also seen in panel 3 of Figure 5, where the evaluation points on the right side of the evaluation zone (y > 1.25 m) show lower IC values. This is probably due to the proximity of loudspeakers playing at relatively high level, compared to the loudspeakers diametrically opposed to the sound source direction (at the bottom left, around 225°). The main acoustic energy received covers a wider azimuthal range for a listener placed at the right hand side than the bottom left corner of the evaluation zone, decreasing the interaural coherence.

To compute the ear signals at the evaluation positions, a single set of far-field HRTFs was used, measured at a distance of 2.4 m. As long as the distances between loudspeakers and the virtual listener do not deviate substantially from this value, the resulting ear signals should be close to the real ones. When moving closer to the loudspeaker, up to 0.80 m in our case, the far-field assumption becomes less valid. Close sound sources at the sides introduce stronger ILDs than far-field sources at the same azimuth angle [19], which should increase the ILD spread on the edge positions of the evaluation zone even further.

5 Conclusion

Ear signals resulting from a sound source played back via Ambisonics were simulated across an extended area inside a loudspeaker array. A series of noise bursts was used as stimulus to balance the onset and steady-state part of the sound field build-up. A time-frame and frequency band based analysis of the interaural cues was carried out to visualise the repartition of the cues across time frames. We observed an increasing variance in the cues over the duration of the signal with increasing distance from the centre of the loudspeaker array. The ITD distribution becomes surprisingly wide and it is unclear whether the ITDs convey information that can be used for localisation at those evaluation positions. Further work will compare the simulated interaural cues to measured ones and relate the observed errors to more perceptual measures.

Acknowledgements

The SOFE system was funded by the BMBF Bernstein Center for Computational Neuroscience, 01GQ1004B.

References

1. Daniel, J. *Spatial Sound Encoding Including Near Field Effect: Introducing Distance Coding Filters and a Viable, New Ambisonic Format*. in *AES 23rd International Conference*. 2003. Copenhagen, Denmark.
2. Favrot, S., *A Loudspeaker-Based Room Auralization System for Auditory Research*, in *Centre for Applied Hearing Research*. 2010, Technical University of Denmark.
3. Grandjean, P., *Captation 3-D de Métriques Directionnelles et Reproduction de Champs Sonores par la Combinaision d'Ambisonies Circulaires et Sphériques*, in *Département de génie mécanique*. 2021, Université de Sherbrooke.
4. Frank, M., F. Zotter, and A. Sontacchi. *Localization experiments using different 2D Ambisonics decoders*. in *25th Tonmeistertagung - VDT international convention*. 2008.
5. Stitt, P., S. Bertet, and M. Van Walstijn. *Perceptual investigation of image placement with ambisonics for non-centred listeners*. in *DAFx 2013 - 16th International Conference on Digital Audio Effects*. 2013. Maynooth, Ireland.
6. Stitt, P., S. Bertet, and M. Van Walstijn, *Off-centre localisation performance of ambisonics and HOA for large and small loudspeaker array radii.* Acta Acustica united with Acustica, 2014. **100**(5): p. 937-944.
7. Favrot, S. and J. Buchholz. *Distance perception in loudspeaker-based room auralization*. in *127th AES Convention*. 2009. New York.
8. Ahrens, A., *Characterizing auditory and audio-visual perception in virtual environments*, in *DTU Health Technology*. 2019, Technical University of Denmark.
9. Yang, L. and X. Bosun. *Subjective Evaluation on the Timbre of Horizontal Ambisonics Reproduction*. in *International Conference on Audio, Language and Image Processing*. 2014. IEEE.
10. Daniel, J., *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. 2001, Université Paris 6.
11. Daniel, J., J.-B. Rault, and J.-D. Polack. *Ambisonics Encoding of Other Audio Formats for Multiple Listening Conditions*. in *105th AES Convention*. 1998.
12. Wierstorf, H., *Perceptual assessment of sound field synthesis*, in *Fakultät IV – Elektrotechnik und Informatik*. 2014, Technische Universität Berlin.
13. Seeber, B.U. and S. Clapp, *Auditory Room Learning and Adaptation to Sound Reflections*, in *The Technology of Binaural Understanding*, J. Blauert and J. Braasch, Editors. 2020, Springer International Publishing: Cham. p. 203-222.
14. Seeber, B. and S. Clapp, *Interactive simulation and free-field auralization of acoustic space with the rtSOFE.* The Journal of the Acoustical Society of America, 2017. **141**(5): p. 3974.
15. Seeber, B.U., S. Kerber, and E.R. Hafter, *A system to simulate and reproduce audio-visual environments for spatial hearing research.* Hearing Research, 2010. **260**(1-2): p. 1--10.
16. Fastl, H. and E. Zwicker, *Psychoacoustics, Facts and Models*. 3 ed. 2007: Springer-Verlag Berlin Heidelberg. XII, 463.
17. Stern, R.M., A.S. Zeiberg, and C. Trahiotis, *Lateralization of complex binaural stimuli: A weighted-image model.* The Journal of the Acoustical Society of America, 1988. **84**(1): p. 156-65.
18. Wierstorf, H., A. Raake, and S. Spors, *Assessing localization accuracy in sound field synthesis.* J Acoust Soc Am, 2017. **141**(2): p. 1111.
19. Brungart, D.S. and W.M. Rabinowitz, *Auditory localization of nearby sources. Head-related transfer functions.* The Journal of the Acoustical Society of America, 1999. **106**(3): p. 1465-1479.