Audio Engineering Society

Convention Paper 9755

Presented at the 142nd Convention
2017 May 20–23, Berlin, Germany

Improvement of the reporting method for closed-loop human localization experiments

Fiete Winter[1], Hagen Wierstorf[2], and Sascha Spors[1]

[1]*Institute of Communications Engineering, University of Rostock, Rostock, D-18119, Germany*
[2]*Audiovisual Technology Group, Technische Universität Ilmenau, Ilmenau, D-98693, Germany*

Correspondence should be addressed to Fiete Winter (fiete.winter@uni-rostock.de)

ABSTRACT

Sound Field Synthesis reproduces a desired sound field within an extended listening area using up to hundreds of loudspeakers. The perceptual evaluation of such methods is challenging, as many degrees of freedom have to be considered. Binaural Synthesis simulating the loudspeakers over headphones is an effective tool for the evaluation. A prior study has investigated, whether non-individual anechoic binaural synthesis is perceptually transparent enough to evaluate human localisation in sound field synthesis. With the used apparatus, an undershoot for lateral sound sources was observed for real loudspeakers and their binaural simulation. This paper reassesses human localisation for the mentioned technique using a slightly modified setup. The results show that the localisation error decreased and no undershoot was observed.

1 Introduction

Contrary to classical, perceptually motivated spatial reproduction techniques such as stereophony, approaches for Sound Field Synthesis (SFS) aim at the physically accurate reconstruction of a desired sound field within a target region. Typically, this region is surrounded by a distribution of loudspeakers, which are driven by individual signals such that the superposition of the emitted sound fields produces the desired sound field. Practical implementations of such techniques cause systematic artefacts in the synthesized sound field that are a consequence of a departure from theoretical requirements. As the most prominent artefact, approximating the continuous secondary source distribution required by theory with a finite number of discrete loudspeakers may introduce spatial aliasing. In practical setups, a perfect physical reconstruction is not possible and SFS relies on the possible masking of the mentioned artefacts by humans. Hence, it is of high interest to evaluate human perception of SFS techniques.

Compared to classical stereophony, SFS offers more flexibility since correct reproduction is pursued within an extended area including many different listening positions. It is also compatible with various reproduction geometries, i.e. number of loudspeakers and shape of the boundary. For the systematic perceptual evaluation this is however challenging as these additional degrees of freedom have to be considered. In the past, dynamic binaural synthesis has emerged as a useful tool to overcome these issues and efficiently evaluate human localisation in SFS. Each loudspeaker is simulated by convolving the Head-Related Impulse Responses (HRIRs) corresponding to the apparent loudspeaker position with the driving signal of the loudspeaker. The superposition of all loudspeakers is played back via

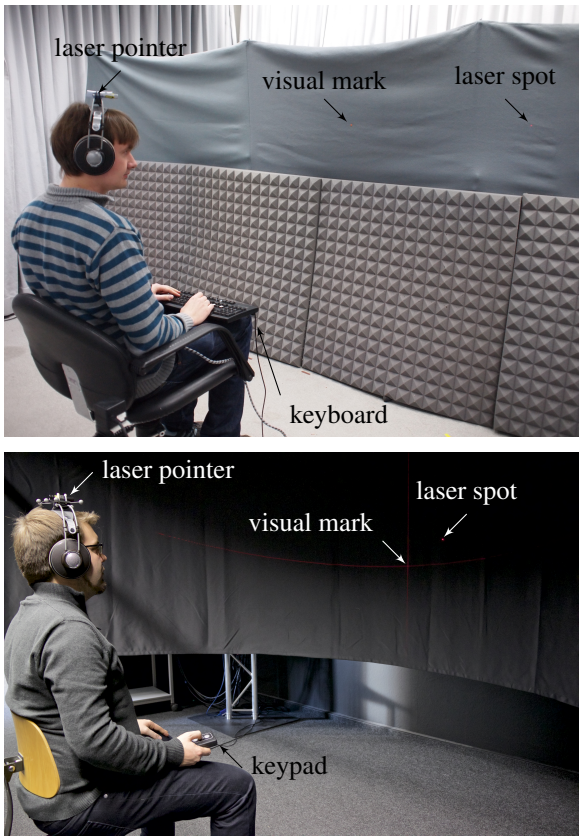

Fig. 1: Listeners during the localisation experiments in 2012 (top) and 2017 (bottom). The rooms were dark during the experiment.

headphones. The orientation of the listener's head is tracked simultaneously and the HRIRs are switched accordingly.

Different studies were conducted, that compared the perceived direction of a real loudspeaker and its binaural simulation. As long as head tracking was applied, the localisation errors were usually in the range of 1° to 5°, see for example [1, 2, 3]. One reason for the varying results for the localisation performance found in the literature is the fact that such experiments are critical regarding the utilised pointing method. Due to the fact that the actual localization error can be as small as 1°, the error of the pointing method has to be smaller than 1°, which cannot be achieved with all methods [4, 1].

In a prior study of one of the authors [5], a pointing method similar to [3] was used. Here, the listeners have to point with their head towards the direction of the auditory event, while the sound event is present. This has the advantage that the listener is directly facing the source, a region in which the minimum audible angle is the smallest [6]. If the listeners are pointing their nose in the direction of the source, an estimation error of the sources at the side will occur, due to an interaction with the human sensory-motor system. To overcome this, a visual pointer was added, showing the listeners where their nose is pointing [7]. To realise such a visual pointer a small laser pointer was mounted onto the headphones. With this setup the localisation accuracy was around 1° for real as for the simulated loudspeaker, but only if the loudspeakers where not positioned more than 30° to the side. For loudspeakers positioned further to the side an undershoot in the reported angle occurred in both cases.

In this contribution, the former experiment of [5] is repeated with a slightly modified apparatus. The main goal is to compare the localisation results with the original study and to possibly identify improvements caused by distinct changes in the experimental setup. Special attentions is drawn to the reported undershoot of the original study.

2 Experimental Methods

This section describes the details of the new localization experiment and the one from [5]. The aim was a direct comparison of both studies resulting in a very similar setup. If there were differences between both experiments they are highlighted in the description by the respective years of publication, i.e 2012 for [5] and 2017 for this study. Otherwise the description applies for both experiments. Note, that the aim in 2012 was to compare humans' localisation of real sound sources with their respective simulation via anechoic binaural synthesis or binaural synthesis with room reflections. The focus is now shifted towards localisation in anechoic binaural synthesis only and how the reporting method can be improved compared to 2012. Details of the experiment in 2012 targeting the binaural room simulation or the real loudspeakers are irrelevant for this comparison and are hence omitted in the following explanations.

2.1 Apparatus

The listening test in 2012 took place in a 83 m^3 acoustically damped listening room (room Calypso in the

Telefunken building of TU Berlin). The listeners sat on a heavy rotatable chair, 1.5 m in front of a straight curtain. They wore open headphones (AKG K601) with an attached head tracker (Polhemus Fastrak). The head tracker had an update rate of 120 Hz, but due to further data processing the effective update rate was 60 Hz. Its measured tracking accuracy is around 1°. The listeners had a keyboard on their legs for entering the response, compare Fig. 1. In a separate room, a computer equipped with a multichannel sound card including D/A converters (RME Hammerfall DSP MADI) played back all sounds. The signals travelled through a head phone amplifier (Behringer Powerplay Pro-XL HA 4700) and analogue cable to the head phones in the listening room, a distance of approximately 5 m.

The listening test in 2017 was conducted in a 86 m^3 acoustically damped room (Audio laboratory at the Institute of Communications Engineering, University of Rostock). The listeners sat on a rotatable chair and were surrounded by a circular curtain with a radius of approximately 1.5 m, see Fig. 2. They wore open headphones (AKG K601) with six optical markers attached to it. The head tracking is achieved with an optical tracking system using eight infra-red cameras (NaturalPoint OptiTrack). The tracking system had an update rate of 120 Hz. The listeners had a keypad in their hands for entering the response, compare Fig. 1. In a separate room, a computer equipped with a stereo sound card (Focusrite Scarlett 2i2, 1st Generation) was used for audio playback. The signals travelled through an analogue cable of approximately 6 m length to the head phones inside the listening room.

2.2 Stimuli: Dynamic Binaural Synthesis

The basic principle of binaural synthesis as a tool for dynamically generating the necessary stimuli is illustrated in Fig. 3. The head tracker provides the horizontal orientation of the listener's head which is fed into the convolution core of the system. Based on the current head orientation the corresponding impulse response is selected from the HRIR dataset. The input signal which is supposed to be emitted by the sound source is convolved with selected impulse response in a block-wise manner. Each block was 1024 samples long. Possible changes in head orientation are handled by convolving the current signal block separately with the old and new impulse response and cross-fading the results within the duration of one block. The SoundScape Renderer [8]

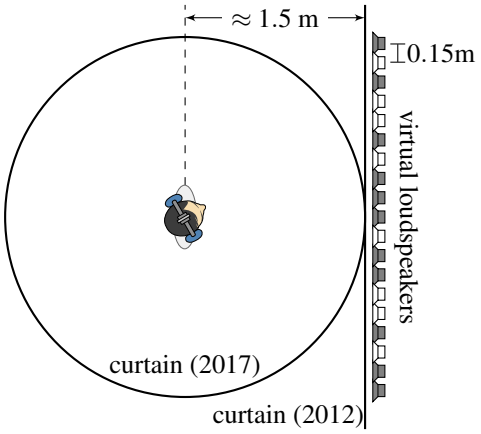

Fig. 2: Sketch of experimental setup and the linear array consisting of the 19 virtual sound sources (loudspeaker symbols) with a spacing of 0.15 m. The eleven source positions used in the experiment are shaded dark.

was utilised as the convolution core. The input signal for SoundScape Renderer was provided by Pure Data [9] which allows to root the dry source signal into different convolution instances of the SoundScape Renderer. Each instance contained the HRIRs corresponding to a specific sound source position. This means that the system was able to instantaneously switch between different conditions without having to restart the playback of the dry audio signal. All components operated at a sampling frequency of 44.1 kHz. More information on the HRIR dataset and the dry source signal is given in the following two sections.

2.2.1 Head-Related Impulse Responses

The used HRIR dataset was measured in an anechoic chamber with an artificial head and a sound source placed in the horizontal plane (at height of the ears) with a distance of 3 m and an azimuth varying from 0° to 359° with 1° resolution. Details about the measurement procedure and involved equipment can be found in [10]. For non-measured source directions, the HRIRs were linearly interpolated using the two nearest measured HRIRs. For distances smaller or larger than the measured 3 m the delay and the amplitude of the HRIRs were adjusted according to the speed of sound and the free-field distance attenuation, respectively.

2.2.2 Dry Source Signal

A Gaussian white noise pulse train of 100 s length was used as the signal emitted by sound source. Each pulse had a duration of 700 ms followed by a pause of 300 ms. The noise signals of each pulse were statistically independent. A cosine-shaped fade-in/fade-out of 20 ms length was applied at the begin/end of each pulse. The signal was bandpass filtered with a fourth order Butterworth filter between 125 Hz and 20000 Hz. In the experiment, the signal was played back in a loop and was convolved with the current HRIR for binaural reproduction.

2.3 Participants

11 listeners were recruited for both experiments. The age of the participants ranged from 21 to 33 years for the 2012 study and from 26 to 60 in 2017 with a respective average of 28.6 and 38 years. 4 and 2 of the listeners had prior experience with listening tests.

2.4 Procedure

Various trials were presented successively to listeners via the headphones using the dynamic binaural synthesis technique described in Sec. 2.2. The experiment contained 11 unique conditions[1] where a single sound source emitting the source signal described in Sec. 2.2.2 was simulated. The positions of the sound sources are indicated in Fig. 2. Each listener had to pass each condition six times leading to 66 trials in total. The order of presentation was randomised with respect to repetitions and condition, while the first 11 trials where meant for training and contained each unique condition exactly once. In 2012, the remaining 55 trials were split into two sessions with 22 and 33 trials containing each unique condition exactly two and three times, respectively.

The participants were instructed to determine the horizontal direction of the perceived auditory event, while the vertical position should be ignored. A pointing method similar to [3] was used, where the listeners were supposed to point into the direction using the laser pointer. The curtain served as a projection surface for the laser. If the listeners were sure to point into the correct direction, they pressed a key on the input device. The localisation result was calculated as the arithmetic mean of 10 values obtained from the head tracker. For the respective update rate, this corresponds to a time of 167 ms (2012) and 83 ms (2017). After the key press, the next condition started instantaneously.

[1] Due to the additional presentation techniques, i.e. binaural room simulation and real loudspeakers, the experiment in 2012 originally contained 33 different conditions.

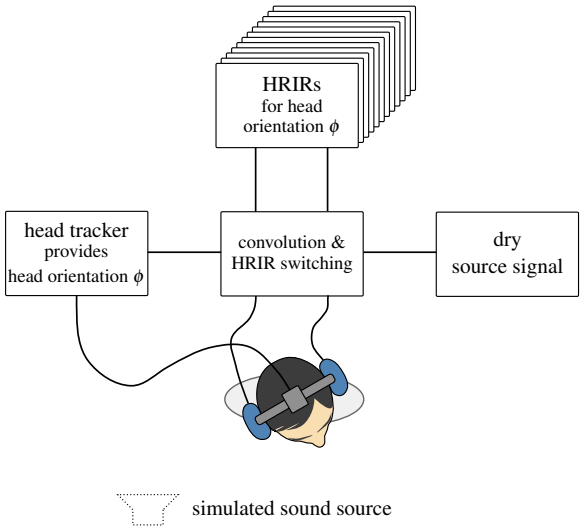

Fig. 3: Basic principle of dynamic binaural synthesis for one simulated sound source.

In an a-priori calibration phase, the listener was indicated to point towards a given visual mark on the curtain. In 2012, a small permanent mark was pasted on the curtain. In 2017, a steady laser cross was projected onto the curtain and switched off after the calibration stage. The room was darkened after calibration.

3 Data Analysis

In following, the statistical methods used to evaluate and compare to the acquired data are presented.

3.1 Perceived Azimuth

As a result of each listening experiment the three-dimensional dataset $\hat{\phi}_l^r(\phi_c)$ describes the perceived azimuths. The index l corresponds to one of the L listeners. The listening condition and respective ground truth source azimuth are denoted by c and ϕ_c, respectively. The total number of conditions is $C = 11$. As each condition is presented R times to each listener, these repetitions are indicated by r. It is assumed, that all samples $\hat{\phi}_l^r(\phi_c)$ are statistically independent due to the randomisation of the presentation order.

3.1.1 Data Correction

It has been already discussed for the original study [5, Sec. 2.6], that the relative location of the laser pointer on the headphones might vary among the listeners (and sessions). This can be caused by e.g. undesired contact, switching on/off the pointer, or changing the batteries. Also the position of the headphones on the head is different for each listener. Consequently, the orientation of the pointing device and the listener's median plane do not necessarily align. This introduces a direction-independent bias to the pointing method, as the listener is forced to point with the laser beam although his/her nose is pointing in a slightly different direction. Under the assumption, that the localisation of each listener is symmetrical to the left and right, this bias is corrected via

$$\tilde{\phi}_l^r(\phi_c) = \hat{\phi}_l^r(\phi_c) - \frac{1}{CR}\sum_{c=1}^{C}\sum_{r=1}^{R}\hat{\phi}_l^r(\phi_c) + \frac{1}{C}\sum_{c=1}^{C}\phi_c\,. \quad (1)$$

The second term on the right-hand side describes the average azimuth for the individual listener over all conditions and repetitions, which is ideally (unbiased case) equal to the arithmetic mean of the ground truth directions ϕ_c (last term). The last term is included since the mean over all ground truth angles is not zero. For the 2012 experiment, the data calibration is performed for each session individually. Hence, $R = 2$ and $R = 3$ in (1) for the first and the second session. The corrected data of the two sessions was pooled afterwards so that $R = 5$ is used for both studies in the following explanations. The azimuthal localisation error is finally given as

$$\Delta_l^r(\phi_c) = \tilde{\phi}_l^r(\phi_c) - \phi_c\,. \quad (2)$$

3.1.2 Descriptive Statistics

The localisation performance of an individual listener w.r.t. a specific condition is evaluated via the sample mean

$$\bar{\Delta}_l(\phi_c) = \frac{1}{R}\sum_{r=1}^{R}\Delta_l^r(\phi_c) \quad (3)$$

and the respective corrected sample standard deviation

$$s_l(\phi_c) = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}\left[\Delta_l^r(\phi_c) - \bar{\Delta}_l(\phi_c)\right]^2}\,. \quad (4)$$

For the average performance for one condition the data is pooled w.r.t. repetitions and listeners. The sample mean is hence defined as

$$\bar{\Delta}(\phi_c) = \frac{1}{LR}\sum_{l=1}^{L}\sum_{r=1}^{R}\Delta_l^r(\phi_c)\,. \quad (5)$$

with its respective corrected sample standard deviation

$$s(\phi_c) = \sqrt{\frac{1}{LR-1}\sum_{l=1}^{L}\sum_{r=1}^{R}\left[\Delta_l^r(\phi_c) - \bar{\Delta}(\phi_c)\right]^2}\,. \quad (6)$$

The overall sample mean, i.e. pooled w.r.t. repetitions, listeners, and conditions, is always zero due to the data correction, cf. (1). The overall sample standard deviation is hence given as

$$s = \sqrt{\frac{1}{CLR-1}\sum_{c=1}^{C}\sum_{l=1}^{L}\sum_{r=1}^{R}\left[\Delta_l^r(\phi_c)\right]^2}\,. \quad (7)$$

and is equivalent to the overall root mean squared error.

3.1.3 Inductive Statistics

For the following explanations it is assumed that the signed localisation error is approximately Gaussian distributed. In order to judge, if the signed localisation error for a specific condition significantly differs from 0°, the confidence interval

$$\left\{\mu(\phi_c)\,\middle|\,\left|\mu(\phi_c) - \bar{\Delta}(\phi_c)\right| \le t_{(1-\alpha/2;LR-1)}\frac{s_l(\phi_c)}{\sqrt{LR}}\right\} \quad (8)$$

of the respective population mean $\mu(\phi_c)$ is used. $t_{(p;\nu)}$ denotes p-Quantile of the Student's t-distribution with ν degrees of freedom. A significant difference is present, if the confidence interval does not include 0°. This procedure is equivalent to a two-sided one-sample t-Test with an significance level of α. The confidence interval for the population standard deviations $\sigma(\phi_c)$ are given as

$$\left\{\sigma(\phi_c)\,\middle|\,\sqrt{\frac{LR-1}{\chi^2_{(1-\alpha/2;LR-1)}}} \le \frac{\sigma(\phi_c)}{s(\phi_c)} \le \sqrt{\frac{LR-1}{\chi^2_{(\alpha/2;LR-1)}}}\right\} \quad (9)$$

with χ^2 denoting the Quantile of Chi-squared distribution with same parametrisation as for the t-distribution. An F-test is used to show, if the population standard

deviation for a specific condition in 2012 is significantly larger than for 2017. The null hypothesis of equal standard deviations is rejected if

$$\frac{s^2_{2012}(\phi_c)}{s^2_{2017}(\phi_c)} > F_{(1-\alpha;LR-1;LR-1)}\,. \tag{10}$$

Rejecting the null hypothesis is equivalent to a significantly higher standard deviation in 2012. $F_{(p;\nu_1;\nu_2)}$ denotes the p-Quantile of the F-distribution with ν_1 and ν_2 degrees of freedom. A analogue F-test can also be formulated for the overall standard deviations (7) with $\nu_1 = \nu_2 = CLR - 1$.

3.2 Elapsed Time per Trial

In addition to the localisation error, the elapsed time $\tau^r_l(\phi_c)$ in each trial was measured. The indices r, l, and c correspond the same dimensions as described in Sec. 3.1. As the focus lies on the localisation accuracy, only simple descriptive statistics such as quantiles and histograms are used to compare both studies. Furthermore, the elapsed time is strictly non-negative and its distribution is likely to be non-Gaussian, which is an essential assumption for the most standard statistical tests.

4 Results and Discussion

It turned out during data analysis, that the standard deviation of the localisation error for one listener in each study was approximately twice as high compared to the maximum among the other participants. These participants were excluded from the analysis resulting into $L = 10$ subjects per study. The histograms for the localisation error are shown in Fig. 4, top row. It can be seen, that the localisation error is approximately Gaussian distributed and hereby fulfils the assumption necessary for statistical test methods. The condensed localisation results are given in Fig. 5: The mean signed error for each condition never exceeds 5° and 3° for 2012 and 2017, respectively. As already mentioned in Sec. 1, an underestimation of the source directions towards the centre can be observed for the lateral sound sources in 2012. In 2017, this phenomenon cannot be observed. For almost every condition, the localisation blur indicated by the standard deviation is significantly higher in 2012 compared to the corresponding value for 2017. The overall standard deviations or root mean squared localisation errors, cf. (7), for 2012 and 2017

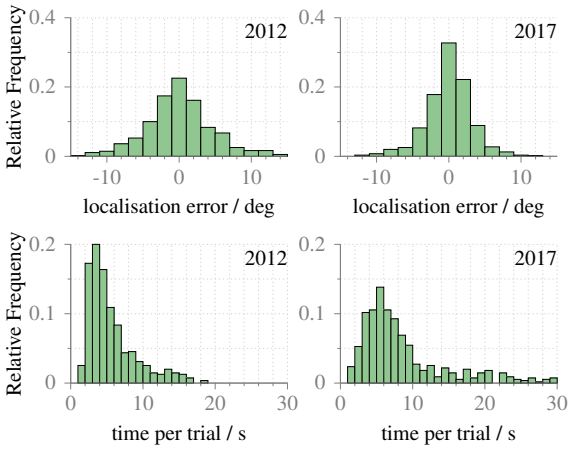

Fig. 4: The top row shows the normalised histograms for the localisation error $\Delta^r_l(\phi_c)$, cf. (2), of 2012 (left) and 2017 (right). The time per trial $\tau^r_l(\phi_c)$ is plotted in the bottom row. All histograms are based on the pooled data over all listeners, conditions, and repetitions. The bin sizes of the histograms are 2° and 1 s, respectively.

are 4.8° and 3.1°, respectively. The former is also significantly higher than the latter (F-test, $\alpha = 5\%$, $\nu_1 = \nu_2 = CLR - 1$).

The main reason for the non-observed localisation undershoot in the study of 2017 is most probably the circular shape of the curtain establishing a close to rotationally invariant projection plane for the pointing method. As depicted in Fig. 1 (top), the ends of the straight curtain in 2012 define a clearly visible limit of projection plane. Even in a dark room these limits are observable due to the change of the reflection pattern of the laser pointer between the curtain and the adjacent wall. Being aware of these limits might have forced the participants to localise towards the centre of the curtain. A further reason for the decrease in localisation blur between 2012 and 2017 might be the increased update rate of head tracker (see Sec. 2.4). As a constant number of values have been captured from the head tracker for averaging, the listeners had to keep their head still for a shorter time frame.

The histogram of the elapsed time per trial is shown in Fig. 4, bottom row: the assumption about its non-Gaussian distribution is confirmed. In can be seen, that the elapsed time in the 2017 reaches higher values than in 2012. The quantiles and arithmetic means shown

in Tab. 1 also attest that the participants in 2017 had a higher response time than in 2012. Although a comparison of the number of participants with prior experience or of the average age of the participants (see Sec.2.3) immediately suggest a connection between the listener composition and elapsed time, no further evidence for that could be found by the authors.

The measured data is available for 2012 under [11] and for 2017 under [12].

elapsed time	2012	2017
arithmetic mean	5.6	9.0
5%-quantile	2.2	2.6
median	4.6	6.8
95%-quantile	13.6	24.2

Table 1: Comparison of the elapsed time per trial given in seconds.

5 Conclusion

In this study, an experiment to evaluate the human localisation in anechoic binaural synthesis [5] has been rerun with some important modifications of the apparatus. As the main difference, a circular curtain as the projection surface for the pointing method was used instead of a straight curtain. Besides the rotational invariance, this setup also allows for listening test with sound sources located 360° around the listener. The analysis of the results revealed that the localisation error and localisation blur have decreased compared to the original study which is mainly due to changes in the apparatus. Hence, it is possible to use this experimental setup to evaluate human localisation in Sound Field Synthesis without the restrictions with respect to the sound source location reported of the original study.

To further improve the method, better calibration mechanisms eliminating the bias caused by the orientation offset between the listener's head and the laser pointer remain as future work. These would allow to study possibly existing phenomena which became unobservable due to the data correction. A more robust mounting of laser pointer on the headphones decreasing the impact of undesired or unavoidable contact is also beneficial for this purpose.

As a drawback, the average time which was needed by a participant to assess one trial increased compared to the original study. Clear reasons for this could however not be identified. Future research has to answer the question, whether the changed apparatus caused this increase. This is of importance for the practicability and scalability of the experiment.

References

[1] Seeber, B. U., *Untersuchung der auditiven Lokalisation mit einer Lichtzeigermethode*, Ph.D. thesis, 2003.

[2] Bronkhorst, A. W., "Localization of real and virtual sound sources," *The Journal of the Acoustical Society of America*, 98(5), pp. 2542–53, 1995.

[3] Makous, J. C. and Middlebrooks, J. C., "Two-dimensional sound localization by human listeners," *The Journal of the Acoustical Society of America*, 87(5), pp. 2188–2200, 1990, doi: 10.1121/1.399186.

[4] Majdak, P., Laback, B., Goupell, M., and Mihocic, M., "The Accuracy of Localizing Virtual Sound Sources: Effects of Pointing Method and Visual Environment," in *124th Audio Engineering Society Convention*, p. Paper 7407, 2008.

[5] Wierstorf, H., Spors, S., and Raake, A., "Perception and evaluation of sound fields," in *Open Seminar on Acoustics*, Boszkowo, Poland, 2012.

[6] Mills, A. W., "On the minimum audible angle," *The Journal of the Acoustical Society of America*, 30(4), pp. 237–46, 1958.

[7] Lewald, J., Dörrscheidt, G. J., and Ehrenstein, W. H., "Sound localization with eccentric head position." *Behavioural Brain Research*, 108(2), pp. 105–25, 2000.

[8] Geier, M. and Spors, S., "Spatial Audio with the SoundScape Renderer," in *27th Tonmeistertagung – VDT International Convention*, Cologne, Germany, 2012.

[9] Puckette, M. et al., "Pure Data: another integrated computer music environment," *Proceedings of the second intercollege computer music concerts*, pp. 37–41, 1996.

[10] Wierstorf, H., Geier, M., and Spors, S., "A Free Database of Head Related Impulse Response Measurements in the Horizontal Plane with Multiple

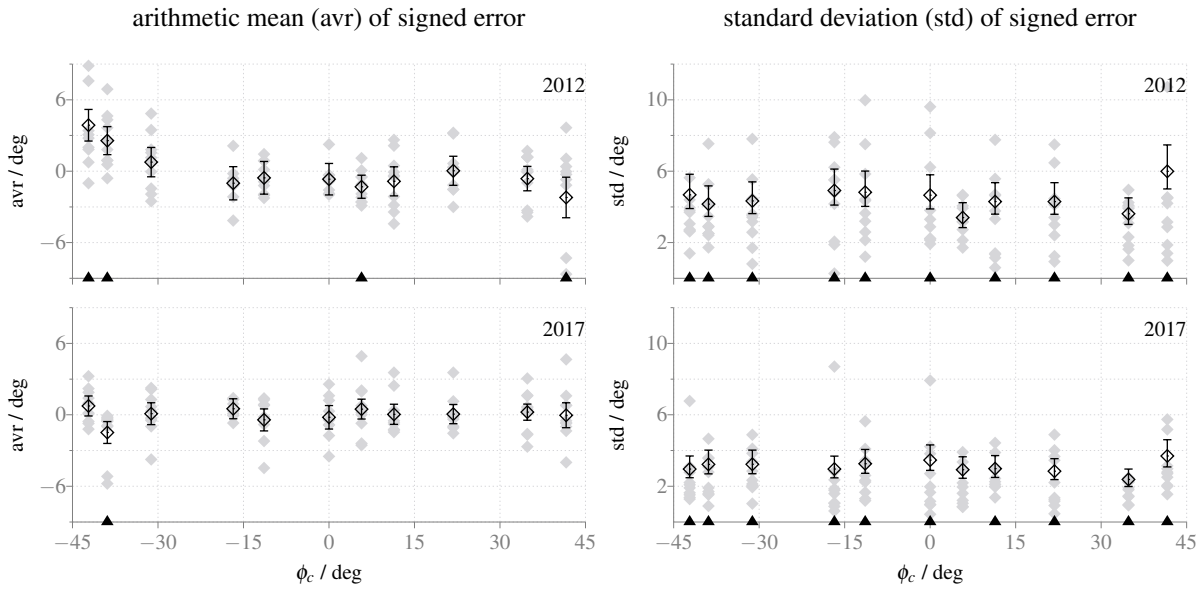

Fig. 5: The top and the bottom row show the results for the studies in 2012 and 2017, respectively. The left column shows the mean of the localisation error $\bar{\Delta}(\phi_c)$ for each condition, cf. (5), together with the $(1-\alpha) = 95\%$-confidence interval of the population mean, cf. (8). The black triangles at the bottom of each plot indicate a significant difference (t-test, $\alpha = 5\%$) of the population mean from zero for the respective condition. The sample standard deviation $s(\phi_c)$ for each condition, cf. (6), together with the 95%-confidence interval of the population standard deviation, cf. (9), are shown in the right column. The black triangles at the bottom of each plot indicate a significantly (F-test, $\alpha = 5\%$) higher population standard deviation in 2012 than 2017 for the respective condition. The individual subject's mean $\bar{\Delta}_l(\phi_c)$ (left) and standard deviation $s_l(\phi_c)$ (right) are shown in grey.

Distances," in *Proc. of 130th Aud. Eng. Soc. Conv.*, London, UK, 2011.

[11] Wierstorf, H., "Localisation of a real vs. binaural simulated point source – data," 2016, doi:10.5281/zenodo.164616.

[12] Winter, F., Wierstorf, H., and Spors, S., "Improvement of reporting method for closed-loop human localization experiments – Data," 2017, doi:10.5281/zenodo.245826.