COMPARISON OF A 2D- AND 3D-BASED GRAPHICAL USER INTERFACE FOR LOCALIZATION LISTENING TESTS

Michael Schoeffler, Susanne Westphal, Alexander Adami, Harald Bayerlein, Jürgen Herre

International Audio Laboratories Erlangen
A Joint Institution of Fraunhofer IIS and University of Erlangen-Nürnberg
Erlangen, Germany
michael.schoeffler@audiolabs-erlangen.de

ABSTRACT

Recently, there is a trend in developing new multi-channel formats towards adding additional loudspeakers in elevated positions. While the common 5.1 surround sound system only has loudspeakers in the horizontal plane, more complex systems, such as 10.2 or 22.2, include two or more elevated loudspeakers.

When listening to music using a multi-channel playback system, the audio material has often not been produced for the used system, e.g. listening to 10.2 material while using a 5.1 surround system. In such cases, the audio material has to be down- or upmixed. Compared with listening to the original audio material, down- or upmixing affects the listening experience. The localization of sound sources is one attribute that might be affected by down- or upmixing the audio material.

In the past, some localization listening tests were conducted by using an user interface depicting a two-dimensional representation of the scene. When it comes to elevated loudspeakers, a third dimension also has to be depicted by the user interface. In this work, an experiment was conducted where participants had to locate sound sources by using two different graphical user interfaces (GUIs). The first GUI consisted of two static images of the scene: a top-view and a front-view. The other GUI had a fully adjustable 3D visualization of the scene. The main purpose of the experiment is to investigate the differences between both GUIs. This includes the time participants spend on each GUI and the difference in the responses. This work is a contribution to the development of new evaluation methods for new and existing multi-channel audio formats and renderers.

1. INTRODUCTION

A number of localization experiments were conducted to find out more about the human ability to localize sound sources. In experiments, reporting the perceived location of sound sources by pointing (with or without the extension of a body part) has been found to be the most accurate method [1]. Due to high accuracy, pointing methods were widely used in recent experiments (e.g. [2][3][4]). However, one drawback of pointing methods is that they can only be applied when localizing sound sources in the listener's field of view. As a consequence, pointing methods can not be used when the listener is not allowed to move his head. Such a condition has to be kept when localizing sound sources in the back. One application example of localizing sound sources in the back is the evaluation of down-mixers. A down-mixer is needed when one multi-channel format has to be converted into another multi-channel format with fewer channels. Especially considering down-mixes where the input multi-channel format contains elevated loudspeakers and the output format does not, a method is needed which supports reporting the localization of sound sources in all three dimensions. Such a method becomes even more important if the distance of a sound source has to be evaluated, too. Since pointing methods do not match the afore-mentioned requirements, two graphical user interfaces which enable the listener to indicate sources at any position are compared in this paper.

Our main research questions are: how accurate are the two types of GUIs, how much time is needed for reporting the localized stimuli and which variables influence the accuracy?

2. RELATED WORK

Graphical user interfaces were used in many localization tests before. Wenzel investigated the effect of increasing system latency on localization of virtual sounds by using a graphical response method [5]. The listener's head was displayed from a top view on the left-hand side of the GUI, while a front view was displayed on the right-hand side. Almost the same GUI was used in an experiment conducted by Begault et al [6]. The GUI of these experiments depicted only the listener's head and did not support reporting the distance.

Pernaux et al. tested three different reporting methods [7]. The first one used a 2D visual feedback of the listener's head. The second GUI displayed the listener's head in a three-dimensional view. The third one was similar to the second one apart from using a 3D finger pointing input instead of a computer mouse. They observed significant differences between these three reporting methods. The 2D-based and 3D-based GUIs of their experiments did not support reporting the distance of a sound source.

Martin et al. utilized a GUI displaying a top view of the scene to investigate the localization using a five-channel surround sound reproduction system [8]. The elevation and distance of a sound source was not investigated in their experiment.

Choisel and Zimmer developed a new pointing method for localizing frontal sources and compared it with a graphical response method [9]. In their experiment, only azimuth angles of sound sources in front of the participants were tested.

Yoo et al. evaluated the localization of sound sources on the horizontal plane for a wave field synthesis system [10]. Listeners reported the sound source positions using an answer sheet which contained a scheme of the scene. Two different listening positions were examined including front and back sound sources. Listeners were found out to have an average localization error of 6.1° azimuth and 9.18° elevation, while the average distance error was 0.5 m and 0.6 m in the horizontal and vertical plane, respectively.

In contrast to many localization tests conducted before, our GUI allows to report the location of a sound source in all three dimensions. By comparing two types of GUIs, the effect size of the GUIs can be measured. Except for the experiment conducted by Pernaux et al., we found no studies which investigated the difference between GUIs for the same experiment setup. Our experiment covers reporting of front, side and back sources with listeners being allowed to move their heads only slightly.

3. METHOD

3.1. Stimuli

Three different signals were used for generating the stimuli. The first signal was pink noise which was faded in and out over 500 ms. The pink noise signal was only played back during the training phase. The second signal was a sine wave with a frequency of 220 Hz and also faded in and out over 500 ms. The steady sine signal represented a sound source which is hard to localize according to Hartmann [11]. The third signal was a castanet recording. In contrast to the sine signal, the castanets recording represented a narrow sound source which is easier to localize due to its transient structure. All signals had a length of 7.8 s and were adjusted to have equal loudness by two expert listeners. The loudness was adjusted using the final experiment setup.

Six loudspeakers were used to reproduce the sine signal and castanets recording. Furthermore, one additional loudspeaker was exclusively used for the training. The loudspeakers, with positions according to Table 1, were placed in 1.5 m distance relative to the participant. We categorized the loudspeaker positions dependent on their azimuth angles to *front*, *side* and *back*. The positions were selected based on previous research about localization and taking into account sound sources in front can be localized more accurately than sources behind the listener[12]. The positions of an established multi-channel system were not used since participants familiar with surround sound might have been biased.

No.	Azimuth	Height	Category
Training	30 °	-1 cm	-
1	10 °	-1 cm	front
2	-55 °	-1 cm	side
3	120 °	-1 cm	back
4	-10 °	33 cm	front
5	55 °	33 cm	side
6	-120 °	33 cm	back

Table 1: Loudspeaker positions are relative to the listener's head ($\text{height} = 120\,\text{cm}$). The height of the loudspeakers was measured from the loudspeakers' center.

Summarized, two different signals were played back from six different loudspeaker positions which results in a total number of twelve stimuli. For the training, a dedicated signal and loudspeaker position was used.

3.2. Participants

Thirty participants including twenty audio professionals took part in the experiment. Most of the participants were employees or students of the International Audio Laboratories Erlangen. Details about the participants are given in Table 2.

Participants		30
Audio professionals		20
Familiar with surround sound		5
Familiar with listening tests		25
Age groups [years]:	$[0-19]$	1
	$[20-29]$	19
	$[30-39]$	5
	$[40-59]$	5

Table 2: Detailed information about the participants.

3.3. Materials and Apparatus

3.3.1. Setup

The experiment took place in a soundproof listening room with room measurements (H x W x D) 256 x 452 x 455 cm. In the middle of the room, a chair and a table were placed for the participants. A 24" widescreen LCD monitor mounted on a small table was placed in front of the chair and table.

The loudspeakers were of type Focal CMS40 with measurements (H x W x D) 23.8 x 15.6 x 15.5 cm. A black-colored 360 ° masking curtain made of deco-molton was installed to veil the loudspeakers. The curtain was fixed to an aluminum ring with a diameter of 2 m which was attached to three truss stands at a height of 212 cm. The lighting in the room was adjusted such that participants could not spot the loudspeakers beyond the curtain. The masking curtain attenuated frequencies above 300 Hz by constantly 2-3 dB.

A face-tracking system was installed to prevent participants from moving their head while locating the stimuli. When participants nodded or turned their head more than 25 °, a warning message popped up and the stimulus stopped playing.

A picture of the experiment setup is shown in Figure 1.

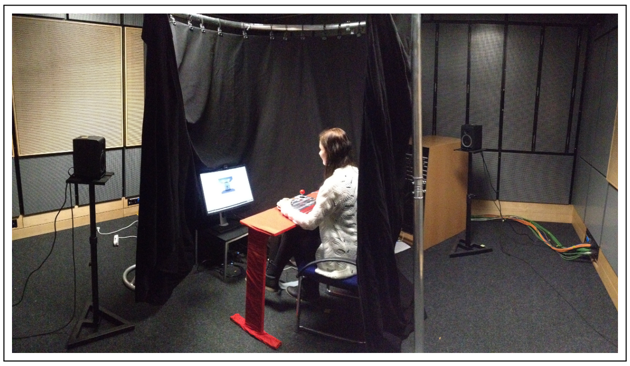
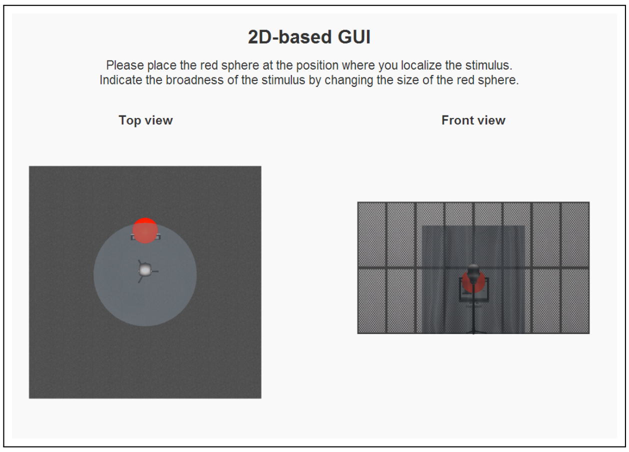


Figure 1: A picture from the setup. The masking curtain was closed during the experiment.

3.3.2. 2D-based GUI

The 2D-based GUI had two orthographic views of the same virtual scene which was a representation of the room the participants were sitting in. On the left-hand side, a top view of the scene was shown whereas a front view was presented on the right-hand side. The virtual scene contained the participant's head, a monitor, the masking curtain and a red sphere. Participants could adjust the position and size of the red sphere and thus indicate where they localized the stimulus. The scene including all modeled objects was true to scale. A screenshot of the 2D-based GUI is depicted in Figure 2.



2D-based GUI

Please place the red sphere at the position where you localize the stimulus.
Indicate the broadness of the stimulus by changing the size of the red sphere.

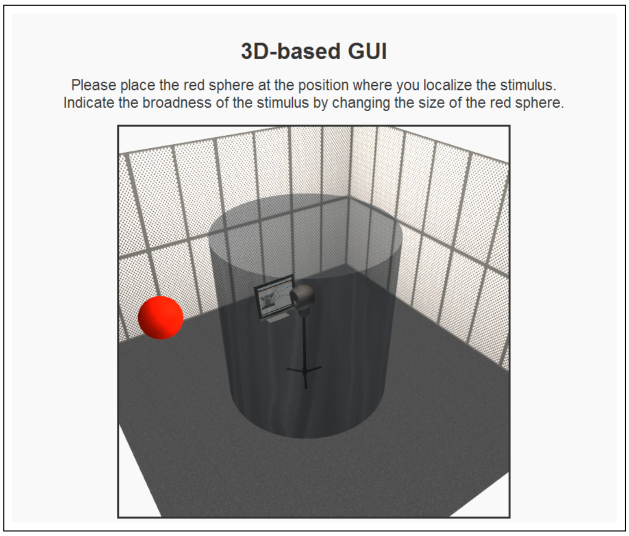
Top view

Front view

Figure 2: Screenshot of the 2D-based GUI.

3.3.3. *3D-based GUI*

The 3D-based GUI was almost similar to the 2D-based GUI except for solely a single perspective view was displayed instead of two orthographic views. The virtual camera of this view was controllable by the participants to select the preferred camera views. Figure 3 shows a screenshot of the 3D-based GUI.



3D-based GUI

Please place the red sphere at the position where you localize the stimulus.
Indicate the broadness of the stimulus by changing the size of the red sphere.

Figure 3: Screenshot of the 3D-based GUI.

3.3.4. *Input Controller*

For reporting the position of the red sphere, a custom-made input controller was developed. This controller offered three different types of inputs:

The camera could be moved along a sphere using an analog joystick and was always directed towards the participant's virtual representation. Zooming in and out could be done by pressing the according button next to the camera controlling joystick. Camera controls were only active while using the 3D-based GUI.

The red sphere could be moved on the horizontal plane by using a digital joystick. Each step corresponded to an accuracy of 5 cm. For moving the red sphere up- or downwards, two additional buttons were located next to the digital joystick. By pressing a button once, the red sphere moved 5 cm. The size of the sphere could be adjusted by another two buttons. The sphere controls were active while using the 2D-based as well as the 3D-based GUI.

Furthermore, two buttons for playing back the stimulus (play button) and completing the response (next button) were located below the camera and sphere controls. When a stimulus was already playing, pressing the play button had no effect.

Developing an own custom input controller allowed us to design an individual arrangement of buttons which is easy to understand. E.g. if a keyboard had been used, participants might have spent more time on learning the relevant keys and their function. In Figure 4, a picture of the input controller is shown.

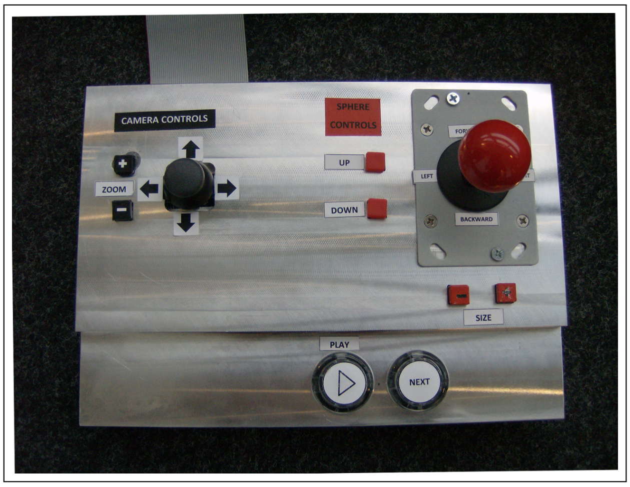


Figure 4: The input controller.

3.4. Procedure

The experiment had a subject-within design. All participants had to localize all twelve stimuli using both GUIs. In total, each participant gave twenty-four responses.

All participants were blindfolded and guided by an experimenter to the chair in the middle of the room. Blindfolding the participants assured that they can not spot loudspeakers while entering the room. After removing the blindfold, participants were instructed to always keep their heads straight towards the monitor during the experiment. Furthermore, they were informed that their faces would be tracked by a face-tracking system to verify that they would be looking towards the monitor. The experimenter then left the room and all subsequent instructions were displayed by the experiment software.

At the beginning of the test, participants had to fill out a questionnaire. They were asked whether they regularly listen to surround sound, whether they are an audio professional, if they are familiar with listening tests and to which age group they belong. The questionnaire was followed by some general instructions: The participants were again reminded that they had to localize stimuli and are not allowed to turn their head. The general instructions announced that a 2D- and a 3D-based GUI would be used for reporting the location of the stimuli.

It was randomly chosen which GUI was initially presented to the participants. Before the participants could report the stimuli locations, they had to read the detailed instructions. These contained a brief description of the GUI, how the input controller worked and that they are asked to localize the stimuli. For the 3D-based GUI, additional information about moving the camera was included in the detailed instructions. After reading the instructions, the participants had to undertake training in which they were asked to place the red sphere at the position where they localize the stimulus. Additionally, they were asked to indicate the broadness of the stimulus by changing the size of the red sphere. The tutorial could only be finished if every control element was used at least once (play button, position and size of the red sphere). In the tutorial for the 3D-based GUI, participants also had to move the camera. Afterwards, participants had to localize twelve stimuli using the present GUI. When they finished reporting all stimuli locations, the same procedure was applied for the second GUI.

At the end of the experiment, the participants had to fill out a questionnaire about how they got along with each GUI.

4. RESULTS

Completing the experiment took 19 minutes ($SD^1 = 6$) on average. To analyze the accuracy of both GUIs, we define the localization error as the euclidean distance between the reported position $\mathbf{r}$ and the actual loudspeaker position $\mathbf{l}$:

$$LocError = ||\mathbf{r} - \mathbf{l}||_2, \tag{1}$$

where bold-faced letters represent vectors with $\mathbf{v} = [v_x, v_y, v_z]^T$. The mean of $LocError$ was 82 cm ($SD = 68$ cm) for all stimuli. The 2D-based GUI had a $LocError$ mean of 83 cm ($SD = 68$ cm) for all stimuli. The 3D-based GUI had a mean $LocError$ of 82 cm ($SD = 68$ cm). The effect of the GUI on $LocError$ was not significant at the $p < .05$ level [$F(1, 717) = 0.094$, $p = .759$]. Levene's test indicated equal variances for $LocError$ ($F = 0.20, p = .655$). In Table 3, detailed information about $LocError$ is given.

	2D		**3D**		**both**	
	M	SD	M	SD	M	SD
front	55	52	60	53	58	52
side	66	47	71	65	68	57
back	129	76	114	71	121	74
all	83	68	82	68	82	68

Table 3: Average Localization errors in cm for both stimuli. The table is segmented by the GUI type and the loudspeaker category.

The distance of a stimulus is dependent, among other factors, on its loudness [13]. As the loudness was only subjectively adjusted by two expert listeners, a normalized localization error is calculated. The normalized localization error excludes the depth distance of the distance between reported position and loudspeaker position. The normalized reported position is defined as:

$$\mathbf{r}_{\text{Norm}} = \frac{\mathbf{r}}{|\mathbf{r}|}|\mathbf{l}|. \tag{2}$$

The normalized reported position has the same direct distance to the listener as the loudspeaker and is used for calculating the normalized localization error:

$$LocError_{\text{Norm}} = ||\mathbf{r}_{\text{Norm}} - \mathbf{l}||_2. \tag{3}$$

The mean of the normalized localization error was 70 cm ($SD = 62$). The 3D-based GUI ($M = 67$ cm, $SD = 59$) turned out to be more accurate than the 2D-based GUI ($M = 73$ cm, $SD = 65$). According to a repeated measures ANOVA, the difference between the two GUIs was not statistically significant at the $p <$ .05 level [$F(1, 717) = 2.295, p = .13$]. Levene's test indicated equal variances for $LocError$Norm ($F = 1.710, p = .192$). In Table 4, all values for the normalized localization error can be found.

	2D		**3D**		**both**	
	M	SD	M	SD	M	SD
front	46	55	40	49	43	52
side	57	42	57	42	57	42
back	119	69	104	66	111	68
all	74	65	67	59	70	62

Table 4: Normalized localization errors in cm for both stimuli. The table is segmented by the GUI type and the loudspeaker category.

A linear regression model with mixed effects was calculated to analyze the influences on the normalized localization error in more detail. The participants were defined as random factor (*participants_id*). The fixed factors were the type of GUI (*GUI*), the loudspeaker category (*category*), the signal (*signal*), the total time spent on the GUI type (*time_GUI*), the time spent on the training for the GUI type (*time_training*) and if the response was given when the GUI was chosen last. (*GUI_last*). The results of the fitted model are described in Table 5.

Coefficient	Value	Std. Error	t-value	p-value
Fixed Effects:				
(Intercept)	34.25	7.9	4.34	.000
GUI = 3D	-3.43	4.58	-0.75	.455
category = side	13.49	4.6	2.92	.004
category = back	68.05	4.6	14.71	.000
signal = sine	36.32	3.78	9.61	.000
time_GUI	-1.61	1.00	-1.62	.105
time_training	4.34	2.57	1.69	.091
GUI_last = true	-5.87	4.29	-1.37	.172
Random Effects:				
participants_id				
	(Intercept)	Residual		
StdDev:	8.52	50.69		

Table 5: Linear regression model with mixed effects for the normalized localization error. The table is segmented by the GUI type and the loudspeaker category.

The normalized localization error can also be expressed as the

[1] M = mean, SD = standard deviation.

elevation and azimuth error of the normalized reported position ($EleError_{\mathrm{Norm}}$ and $AziError_{\mathrm{Norm}}$). The mean of $EleError_{\mathrm{Norm}}$ was $10\,^{\circ}$ ($SD = 34$). As indicated by the linear regression model of the normalized localization error, the signal type had an influence on the elevation and azimuth errors. The average normalized elevation error of the castanets recording was $9\,^{\circ}$ ($SD = 7$). When sine wave was played back the average normalized elevation error increased ($M = 11\,^{\circ}, SD = 8$). If the normalized elevation errors are analyzed for each loudspeaker category, the castanets recording resulted in lower errors for each category (front: $M = 8\,^{\circ}, SD = 6$; side: $M = 9\,^{\circ}, SD = 6$; back: $M = 10\,^{\circ}, SD = 8$). The sine wave resulted in much higher average normalized elevation errors (front: $M = 11\,^{\circ}, SD = 9$; side: $M = 11\,^{\circ}, SD = 7$; back: $M = 10\,^{\circ}, SD = 7$). All normalized elevation errors for both stimuli are described in Table 6.

	2D		**3D**		**both**	
	M	SD	M	SD	M	SD
front	10	8	9	7	10	8
side	10	7	9	7	10	7
back	11	7	9	8	10	8
all	11	8	9	7	10	7

Table 6: Elevation errors of the normalized reported position in degrees. The table is segmented by the GUI type and the loudspeaker category.

The mean of $AziError_{\mathrm{Norm}}$ was $24\,^{\circ}$ ($SD = 30$). As expected, the average normalized azimuth errors widely differed for each signal type. The average normalized azimuth error of the castanets recording was $16\,^{\circ}$ ($SD = 21$). The average normalized azimuth error increased when the sine wave was played back ($M = 32\,^{\circ}, SD = 34$). The castanets recording resulted in smaller errors for all categories (front: $M = 5\,^{\circ}, SD = 4$; side: $M = 13\,^{\circ}, SD = 13$; back: $M = 30\,^{\circ}, SD = 29$). As expected, the sine wave resulted in larger average normalized azimuth errors (front: $M = 21\,^{\circ}, SD = 38$; side: $M = 22\,^{\circ}, SD = 19$; back: $M = 54\,^{\circ}, SD = 31$). All normalized azimuth errors are described in Table 7 for both stimuli.

	2D		**3D**		**both**	
	M	SD	M	SD	M	SD
front	14	29	12	27	13	28
side	18	17	18	17	18	17
back	46	33	39	31	42	32
all	26	31	23	28	24	30

Table 7: Azimuth errors of the normalized reported position in degrees. The table is segmented by the GUI type and the loudspeaker category.

The differences of localization errors between the sine wave and the castanets recording are confirmed by the reported broadness of the two stimuli. The average radius of the red sphere was 23 cm($SD = 7$) when the castanets recording was played back. For the sine wave stimulus, the average reported radius of the sphere was $M = 45\,cm$ ($SD = 29$).

Reporting twelve stimuli for the 2D-based GUI took the participants in average 5 minutes ($SD = 2$). For the 3D-based GUI 8 minutes ($SD = 3$). According to a repeated measures ANOVA, the difference of the time participants spent on each GUI was statistically significant at the $p < .05$ level [$F(1, 57) = 17.99$, $p = .000$]. Levene's test indicated equal variances for $LocError$ (F = 4.01, p = .04997).

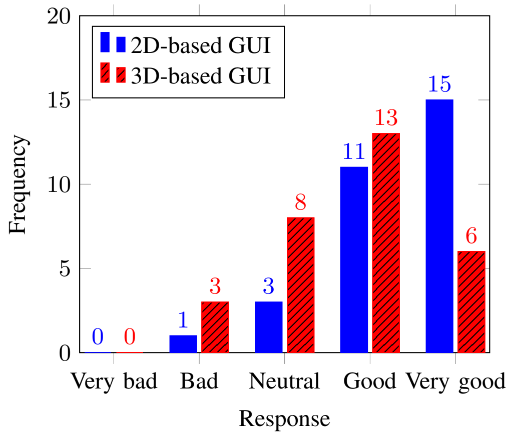


Figure 5: Frequencies of participants' responses about they got along with each GUI.

At the end of the experiment, the participants were asked how they got along with each GUI. The possible answers were: "Very bad" (=1), Bad (=2), "Neutral" (=3), "Good" (=4) and "Very good" (=5). In Figure 5, the frequency of the answers are shown. The mode of the 2D-based GUI was Very Good and for the 3D-based it was Good. A cumulative link model (Table 8 supported the information provided by the listeners that they got along better with the 2D-based GUI. The 3D-based GUI type (*GUI*) had a significant negative effect on the participants' answers ($F = -1.34, p = .008$). If a participant used a type of GUI last (*last*), it had a non-significant positive effect on the answer ($F = 0.46, p = .344$).

Coefficient	Estimate	Std. Error	z-value	p-value
GUI = 3D	-1.34	0.51	-2.64	.008
last = true	0.46	0.49	0.95	0.344
Threshold coefficients:				
	Estimate	Std. Error	z-value	
$Bad\|Neutral$	-3.28	0.67	-4.91	
$Neutral\|Good$	-1.67	0.50	-3.37	
$Good\|Very\ good$	0.25	0.44	0.57	
Number of observations: 60				
Cragg and Uhler's pseudo R^2: 0.14				

Table 8: Logit cumulative link model of the response about how participants got along with each GUI.

5. DISCUSSION

Participants were much faster in reporting the location when using the 2D-based GUI. This was expected since the 2D-based GUI showed two views at the same time. To have a similar perspective with the 3D-based GUI, the camera had to be moved which took some time. Some participants reported the two views of the 2D-based GUI were perfectly good for them. When using the 3D-based GUI, they consecutively moved the camera such that a front and top view were shown. Another reason for the time differences might be due to the 3D-based GUI was something new and fun to use as some participants reported. These participants spent a bit more time on the 3D-based GUI just for playing around with the camera. Therefore, for listening tests with many items, a 2D-based GUI with multiple views should be used instead of a 3D-based GUI with a virtual camera.

Participants reported getting along better with the 2D-based GUI. This was expected since the 3D-based GUI needs the camera to be adjusted which increases the complexity. The 2D-based GUI already showed two views which enabled the user to monitor all three dimensions. The red sphere was controlled by a joystick relative to the virtual participant, even if the camera was moved. Some participants reported that they would have expected the red sphere to move relative to the camera (e.g. pressing the joystick forward moves the red sphere away from the camera).

Localization errors of the participants' responses were slightly smaller when the 3D-based GUI was used. However, the differences in the localization error and normalized localization error were not significant. Nevertheless, the similar results are interesting, considering that participants reported that they got along much better with the 2D-based GUI. By the linear regression model, it could be revealed that for predicting the localization error other effects are much more relevant than the GUI. The effect size of the GUI type was small in the model and also not significant. Loudspeaker position and the signal type influenced the localization most. This was expected since these effects are known from established research. There are non-significant indications that training is important for reporting the localization by graphical user interfaces. When the GUI was last used, the localization error was reduced according to the model.

The reported azimuth angle of the normalized location error turned out to be accurate when the castanets recording was played back by the front loudspeakers. The average normalized azimuth error was $5\,^\circ$ which is close to results achieved by other localization methods. E.g. Haber et al. reported that the average localization errors of nine different methods ranged from $+3.5\,^\circ$ to $-5.2\,^\circ$ for front loudspeaker positions[1]. In the experiment of Yoo et al. the average azimuth error was $-5\,^\circ$ and $2.3\,^\circ$ for two tested front loudspeaker positions ($-30\,^\circ$ and $+30\,^\circ$)[10].

6. CONCLUSION

A method for reporting the location of sound sources in all three dimensions was presented. The method was evaluated by conducting an experiment with two different types of GUIs: A 2D-based GUI and a 3D-based GUI. The 2D-based GUI had an average localization error of 83 cm and turned out to be the less time-consuming and more convenient choice. The 3D-based GUI was slightly more accurate and had an average localization error of 82 cm. The analysis of the experiment results revealed that the used GUI had only a small effect on the localization error. The signal type and loudspeaker position played a much more important role. For front loudspeaker positions, both GUIs resulted in an average normalized azimuth error of $5\,^\circ$ when a castanets recording was played back.

7. REFERENCES

[1] L. Haber, R. N. Haber, S. Penningroth, K. Novak, and H. Radgowski, "Comparison of nine methods of indicating the direction to objects: data from blind adults.," *Perception*, vol. 22, no. 1, pp. 35–47, Jan. 1993.

[2] M. Frank, L. Mohr, A. Sontacchi, and F. Zotter, "Flexible and Intuitive Pointing Method for 3-D Auditory Localization Experiments," in *Audio Engineering Society Conference: 38th International Conference: Sound Quality Evaluation*, 2010.

[3] H. Wierstorf, A. Raake, and S. Spors, "Localization of a Virtual Point Source within the Listening Area for Wave Field Synthesis," in *Audio Engineering Society Convention 133*, 2012.

[4] T. Ashby, R. Mason, and T. Brookes, "Head Movements in Three-Dimensional Localization," in *Audio Engineering Society Convention 134*, 2013.

[5] E. M. Wenzel, "Effect of increasing system latency on localization of virtual sounds," in *Proceedings of the Audio Engineering Society 16th International Conference on Spatial Sound Reproduction*, 1999, pp. 42–50.

[6] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source.," *Journal of the Audio Engineering Society. Audio Engineering Society*, vol. 49, no. 10, pp. 904–916, Oct. 2001.

[7] J. Pernaux, M. Emerit, and R. Nicol, "Perceptual Evaluation of Binaural Sound Synthesis: the Problem of Reporting Localization Judgments," in *Audio Engineering Society Convention 114*, 2003.

[8] G. Martin, W. Woszczyk, J. Corey, and R. Quesnel, "Sound Source Localization in a Five-Channel Surround Sound Reproduction System," in *Audio Engineering Society Convention 107*, 1999.

[9] S. Choisel and K. Zimmer, "A pointing Technique with Visual Feedback for Sound Source Localization Experiments," in *Audio Engineering Society Convention 115*, 2003.

[10] J. Yoo, J. Seo, H. Shim, H. Chung, K. Sung, and K. Kang, "Subjective Listening Experiments on a Front and Rear Array-Based WFS System," *ETRI Journal*, vol. 33, no. 6, pp. 977–980, 2011.

[11] W. M. Hartmann, "Localization of sound in rooms," *The Journal of the Acoustical Society of America*, vol. 74, no. November 1983, pp. 1380–1391, 1983.

[12] S. Carlile, P. Leong, and S. Hyams, "The nature and distribution of errors in sound localization by human listeners.," *Hearing research*, vol. 114, no. 1-2, pp. 179–96, Dec. 1997.

[13] G. von Békésy, "The moon illusion and similar auditory phenomena," *The American journal of psychology*, vol. 62, no. 4, pp. 540–552, 1949.